VACCINE

Visual Analytics for Command, Control and Interoperability Environments
A U.S. Department of Homeland Security
Science and Technology Center of Excellence

VACCINE ANNUAL REPORT – YEAR 5

Addendum A - Presentations

July 1, 2013 – June 30, 2014

Cooperative Agreement No. 2009-ST-061-CI0001

PURDUE UNIVERSITY

HOMELAND SECURITY UNIVERSITY PROGRAMS
TODAY'S RESEARCH & EDUCATION, TOMORROW'S SECURITY

Addendum A – Presentations

Table of Contents

ARIZONA STATE UNIVERSITY

Abstracting Attribute Space for Transfer Function Exploration and Design 4

Analytical Brushing in Spatiotemporal Analysis 21

Box-Office VAST – VADER 60

Bristle Maps: A Multivariate Abstraction Technique for Geovisualization 91

Enabling Predictive Analytics Through Visualization 126

Integrating Visual Analytics Curriculum Into Your Own Classroom 178

Visual Analytic Tools for Geo-Genealogy 234

Visual Analytics Summer School 2013: Fundamentals of Visual Analytics 254

GEORGIA INSTITUTE OF TECHNOLOGY

Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw 435

The Value of Visualization and Why Interaction Matters 454

The Value of Visualization for Exploring and Understanding Data 553

The Value of Visualization for Exploring and Understanding Data 571

MORGAN STATE UNIVERSITY

iLEAPS Radio Interview Presentation 597

PENNSYLVANIA STATE UNIVERSITY

Coordinating Views within Interactive Web Geovisualization 603

Designing a Web Service to Geo-Locate Subjects of Volunteered, Textual Geographic Information 604

Exploratory Visualization of Movement Patterns in Twitter Data 623

GeoTxt: A Web API to Leverage Place References in Text 637

Geo-Visual Analytics & 'Big' Data: Leveraging Geo-Information in Unstructured Data 639

Moving forward: Reasoning about movement and other things dynamic 647

Questions Facing Map Design in the Age of Mobility and Siri 655

Symbol Sharing Tools for Emergency Management Cartography at NASA 673

PURDUE UNIVERSITY

Big Data, Visualization, and Risk Based Decision Making 690

COAST Case Study – DHS Data Workshop 703

Empowering Discovery and Decision Making Through Visual Analytics 719

Empowering Effective Decision Making Through Visual Analytics 773

Evaluation: Multistep and Multi-Level 808

First Responder Technology Demonstration 813

VACCINE and Visual Analytics Solutions 817

Visual Analytics: The Power to Solve Big Data Challenges Through Interaction and Visual Reasoning 852

Visual Analytics for CyberSecurity 884

Visual Analytics for Investigation and Law Enforcement 929

Visual Analytics for Security, Safety, and Privacy: Approaches, Lessons Learned, Opportunities, and Challenges 966

Visual Analytics For Shared, Synchronized Situational Awareness 1001

Visual Analytics: Overcoming Challenges to Create Tools for Decision Making 1036

Visual Analytics Solutions for Data-Driven Decision Making 1057

Visual Analytics: The Right Solution for Big Data and Global Challenges 1078

Visual Analytics: The Right Solution for Big Data and Global Challenges 1159

Visualization Literacy: Research Challenges Do You See What I See? 1227

UNIVERSITY OF OXFORD

An Information Theoretic Approach to the Flight Data Problem 1231

An Information Theoretic Approach to the Flight Data Problem 1268

UKVAC II JRP WP1: Information –Theoretic Framework 1321

UKVAC II JRP WP21: Information-Theoretic Framework Part 2 1346

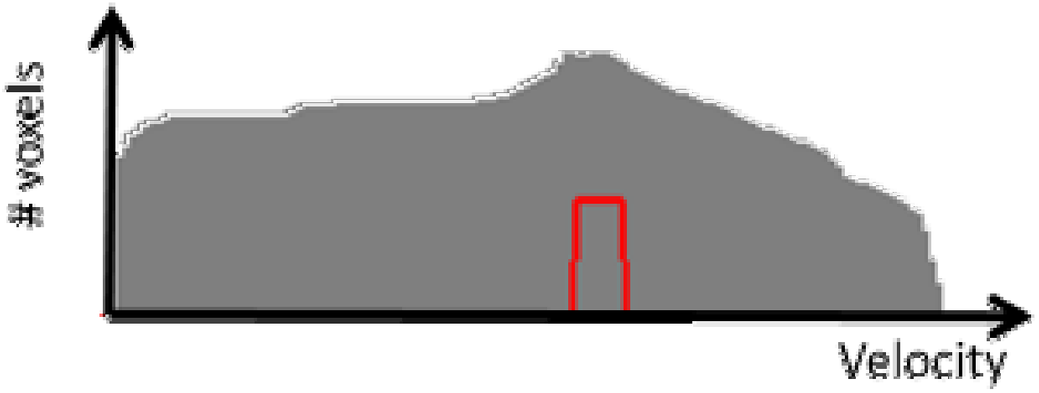
UKVAC II JRP Problem 2: Flight Data (Overview, part II) 1356

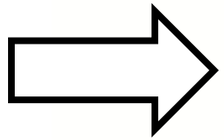


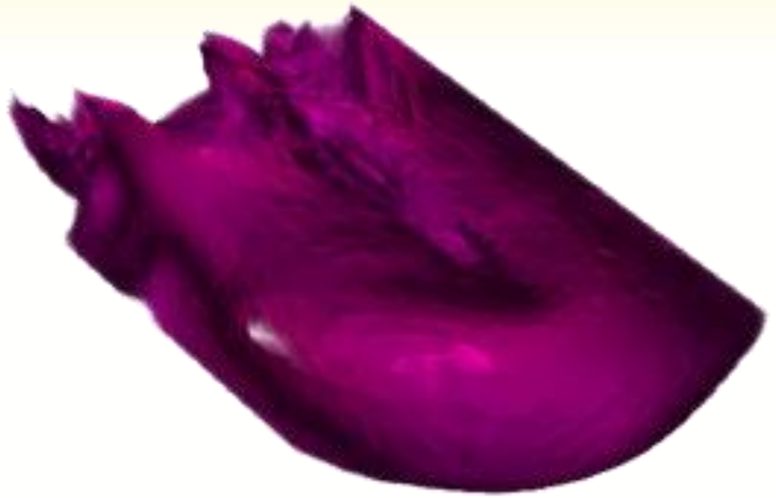
Abstracting Attribute Space for Transfer Function Exploration and Design

Ross Maciejewski, Yun Jang, Insoo Woo, Heike Janicke
Kelly P. Gaither, David S. Ebert

Designing a transfer function is intuitive, Right?



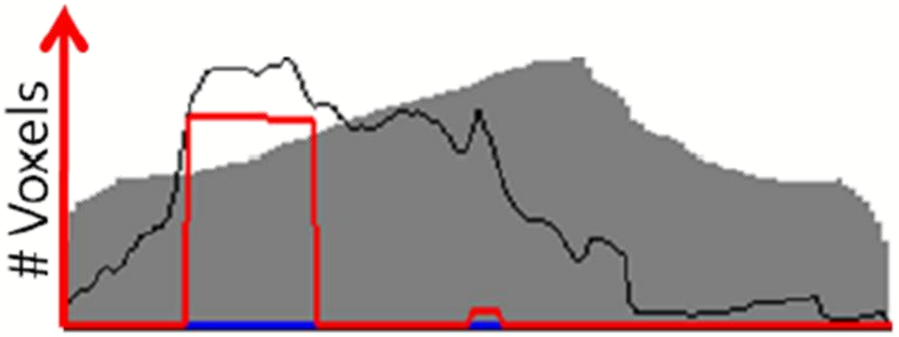


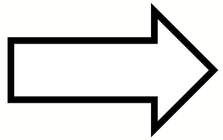


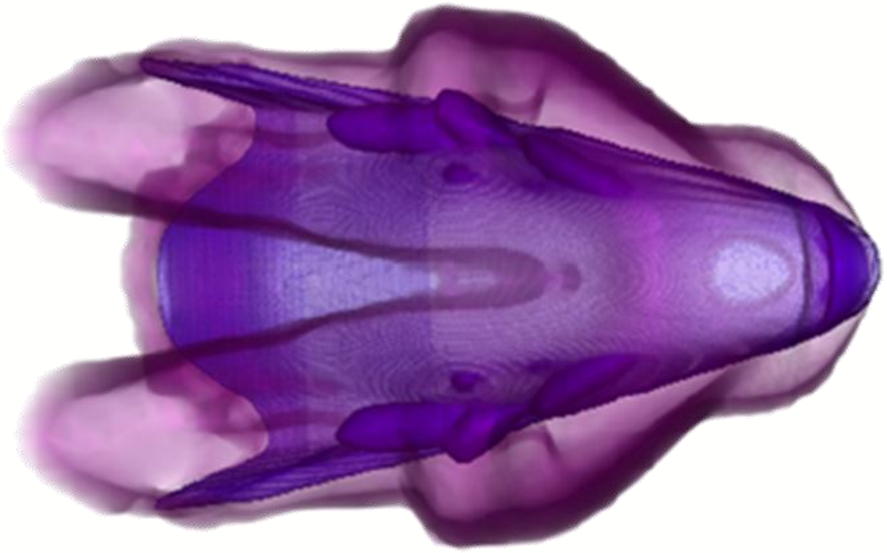
Perfect!

VIS 2013 VAST • INFOVIS • SCIVIS BIOVIS • LDAV

What if there was some extra information in the data to give you a hint?

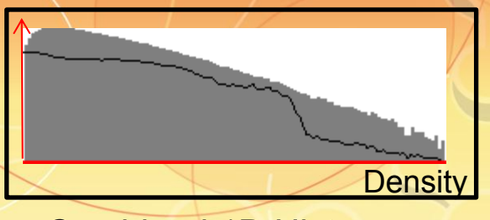






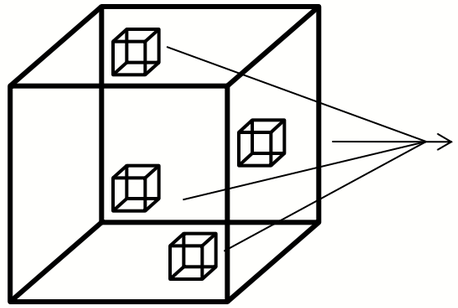
VIS 2013
VAST • INFOVIS • SCIVIS
BIOVIS • LDAV

So what was that black line?



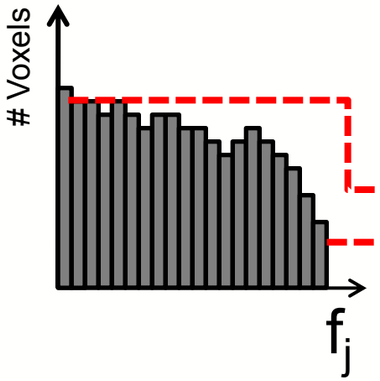
Combined 1D Histogram

Calculate information metric (i.e., entropy, skewness) based on attribute f_k

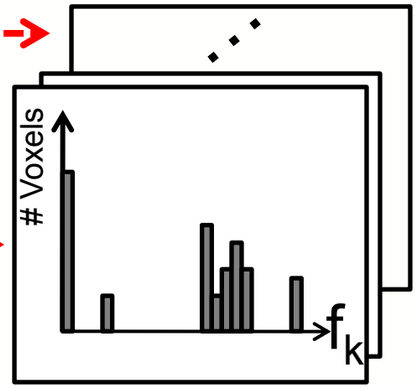


Each voxel has a set of attribute values

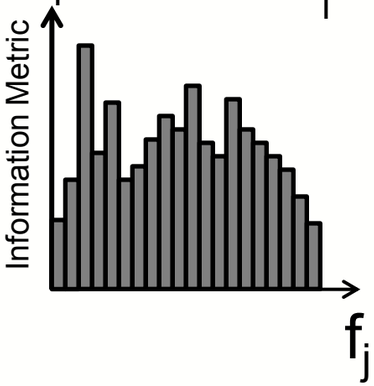
$= \{f_1, f_2, f_3, \ldots f_N\}$



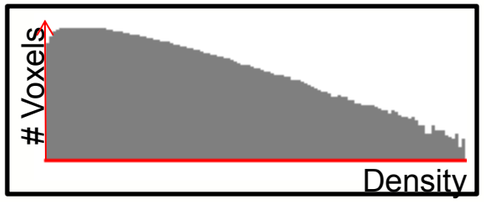
The attribute space (i.e., density) histogram is plotted versus the number of voxels that map to a given attribute bin value



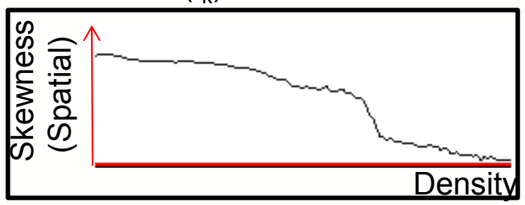
The voxels that map to a given bin also have a probability distribution p(x) with respect to a secondary attribute (i.e., temperature)



The attribute space histogram can now be plotted such that the y-axis is the statistical measure of the secondary attribute (f_k).



Traditional 1D Histogram



Abstracted 1D Histogram

VIS 2013 VAST · INFOVIS · SCIVIS BIOVIS · LDAV

Information Metrics

- Mean

$$\mu(S_V) = \frac{1}{|S_V|} \sum_{x \in S_V} f_k(x)$$

- Standard Deviation

$$\sigma(S_V) = \sqrt{\left(\frac{1}{|S_V|} \sum_{x \in S_V} (f_k(x) - \mu(S_V))^2\right)}$$

- Skewness

$$\gamma(S_V) = \frac{\frac{1}{|S_V|} \sum_{x \in S_V} (f_k(x) - \mu(S_V))^3}{\left(\frac{1}{|S_V|} \sum_{x \in S_V} (f_k(x) - \mu(S_V))^2\right)^{\frac{3}{2}}}$$

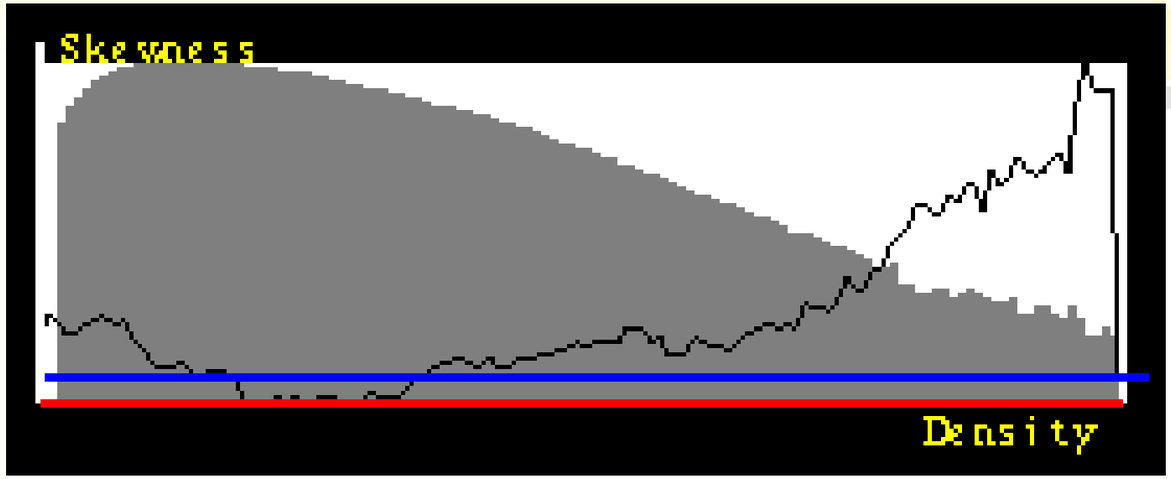
- Entropy

$$H(S_V) = -\sum_{x \in S_V} p(f_k(x)) \ln(p(f_k(x)))$$

A Comparison of Some Metrics

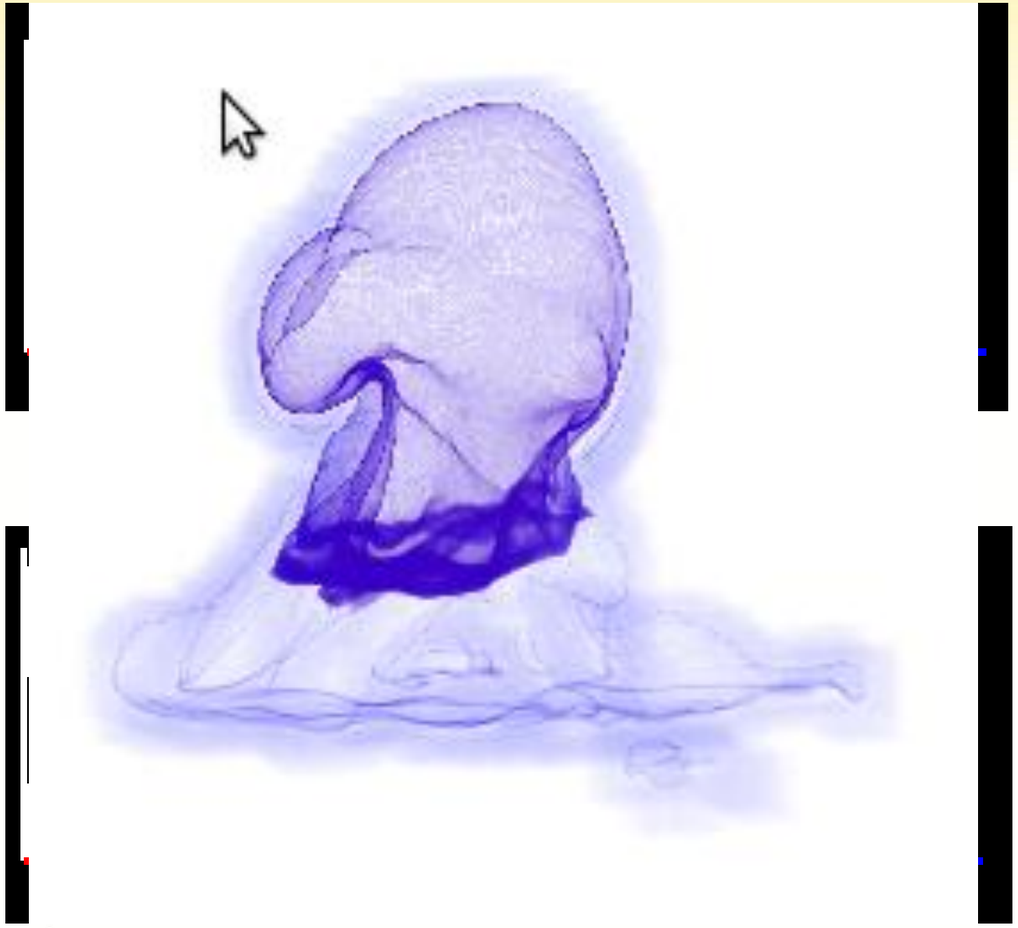


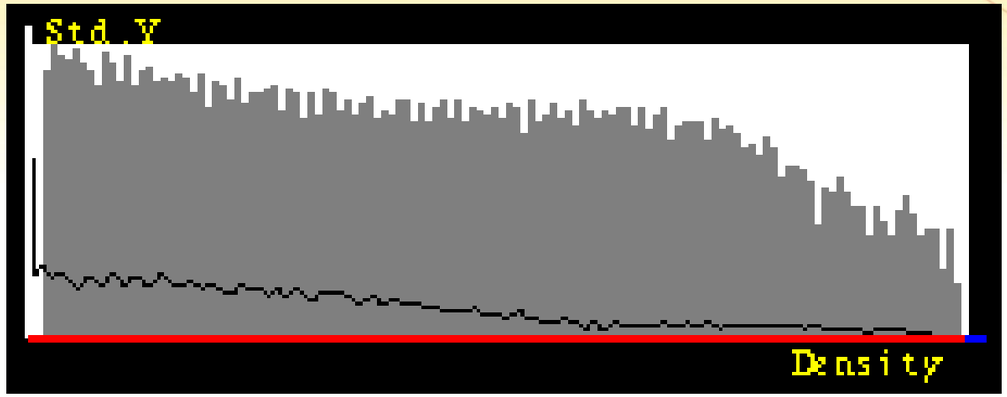
Standard Deviation Along x-axis

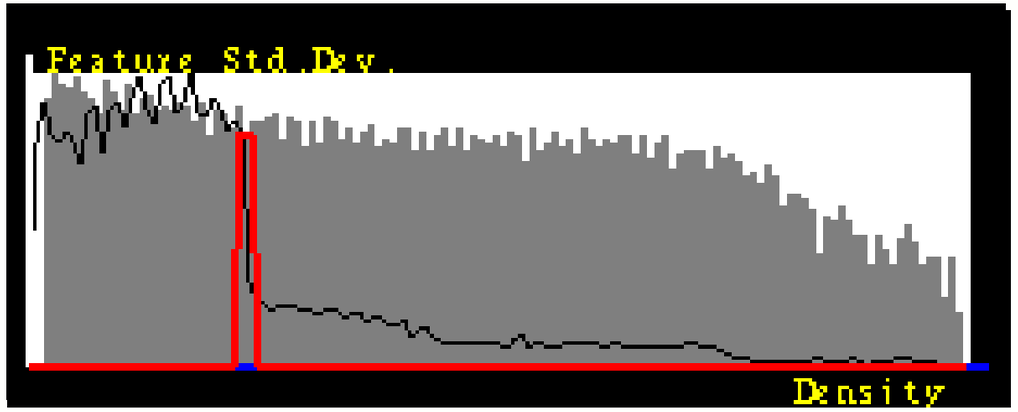


Skewness Along x-axis

So What's This Look Like?

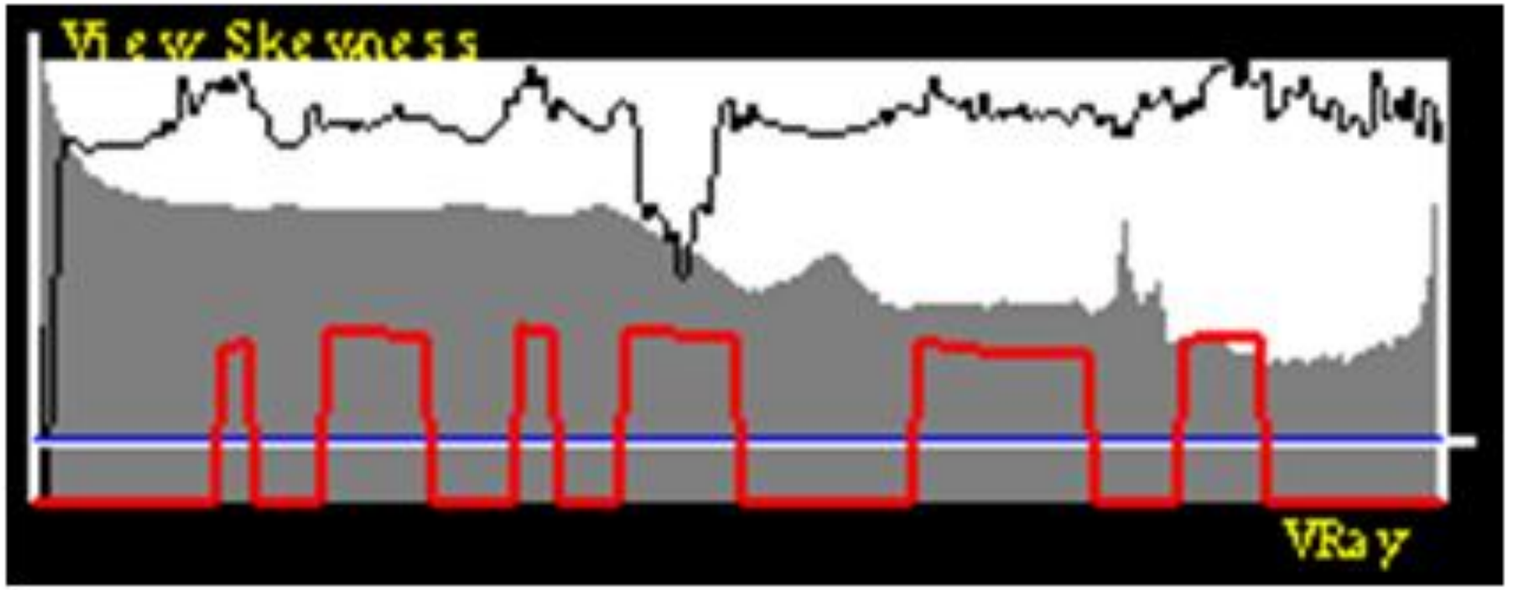


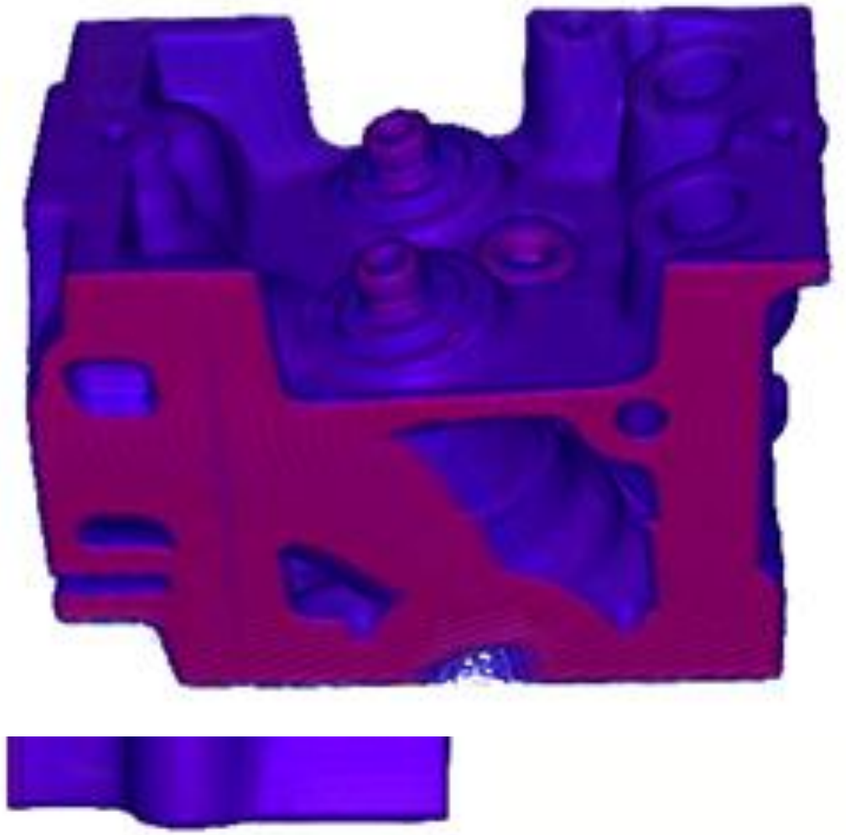






Non-Axis Aligned Spatial Features

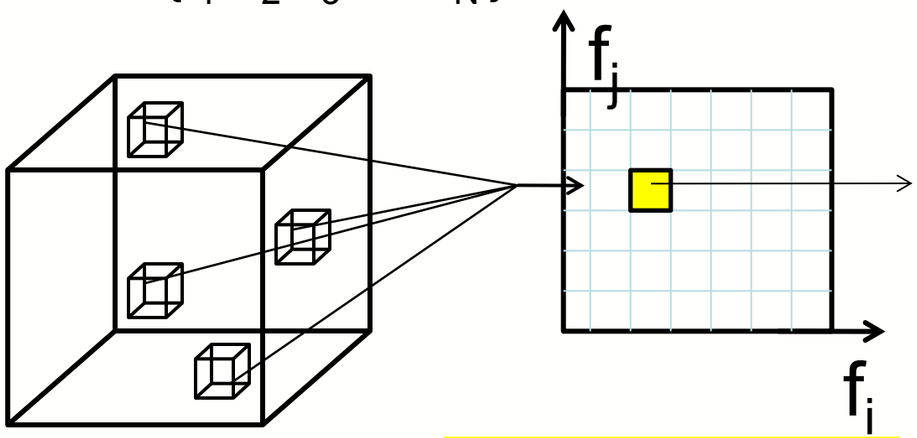




2D Transfer Functions

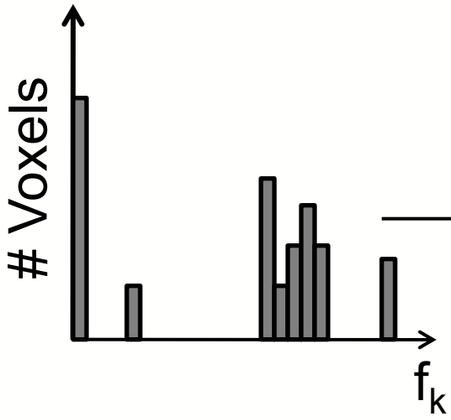
Volumetric Space

= $\{f_1, f_2, f_3, \ldots f_N\}$

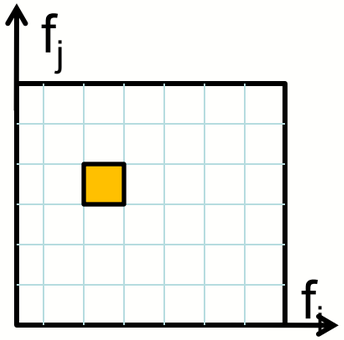


2D Feature Space where entry (i,j) is the number of voxels

Probability Space of feature k with respect to voxels at (i,j)



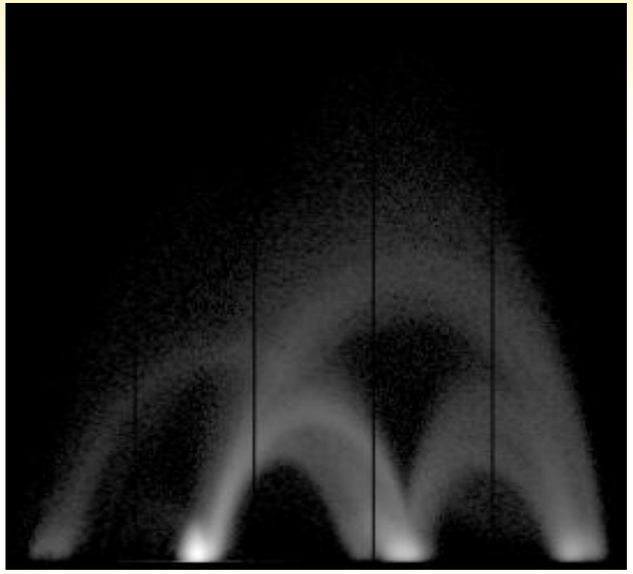
Compute Info Metric



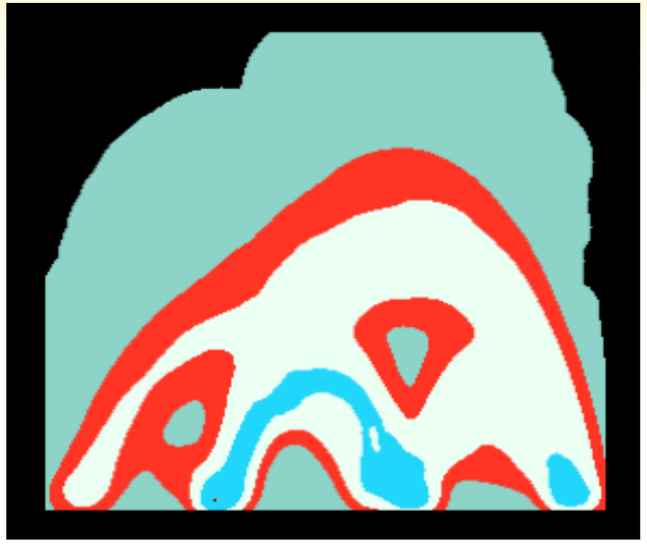
2D Abstracted Feature Space where entry (i,j) is the information metric

VIS 2013 VAST • INFOVIS • SCIVIS BIOVIS • LDAV

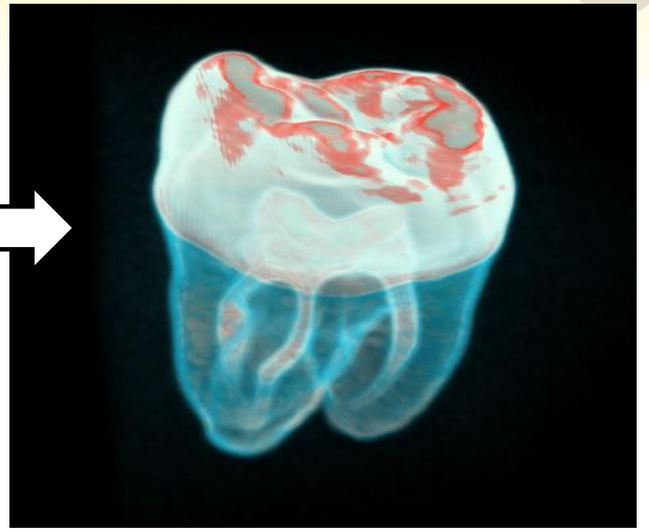
Enhancing Transfer Function Design



Density vs. Density Gradient Normal Histogram Feature Space

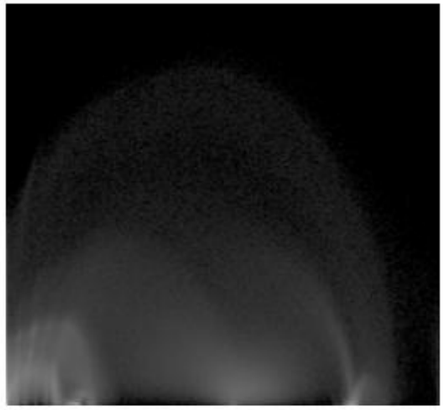


Color Transfer Function Via Density Estimation Grouping



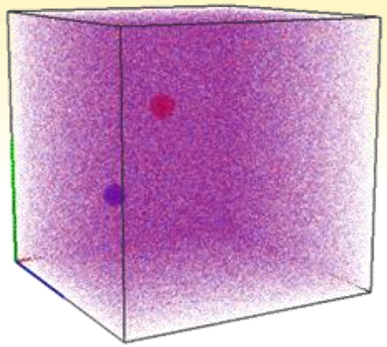
Application of Transfer Function to Volumetric Data

Ross Maciejewski, Wei Chen, Insoo Woo, David S. Ebert. Structuring Feature Space - A Non-Parametric Method for Volumetric Transfer Function Generation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6): 1473-1480, Nov./Dec. 2009.

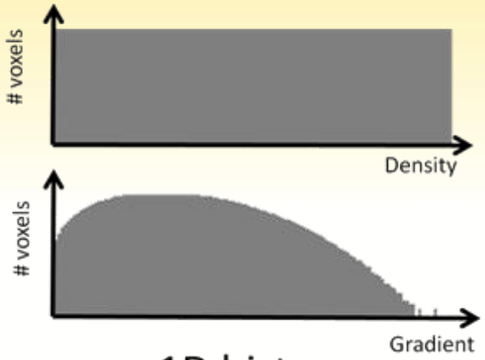


Voxel Magnitude Feature Space

Noisy Data



Input data



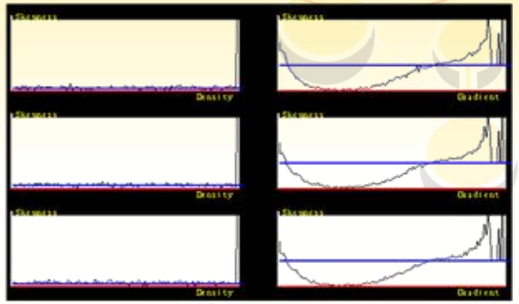
1D histogram



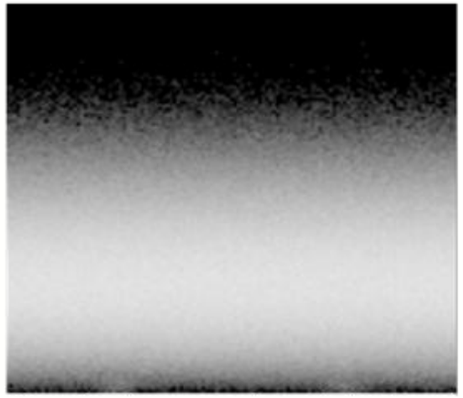
Entropy



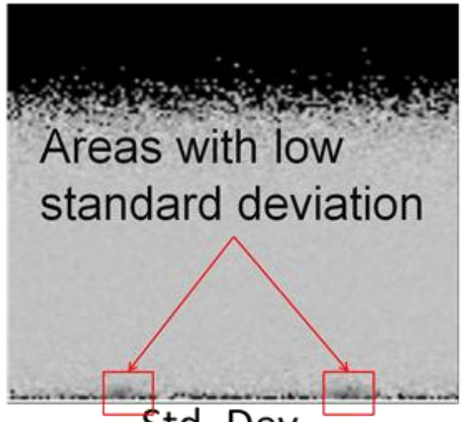
Std. Dev.



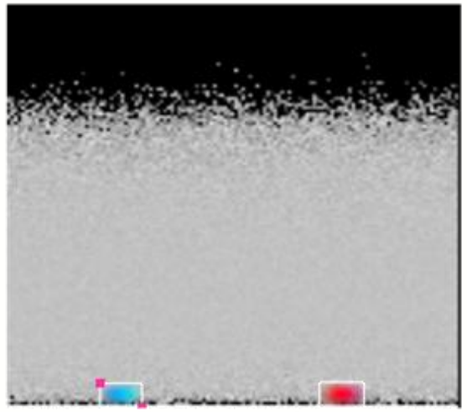
Skewness



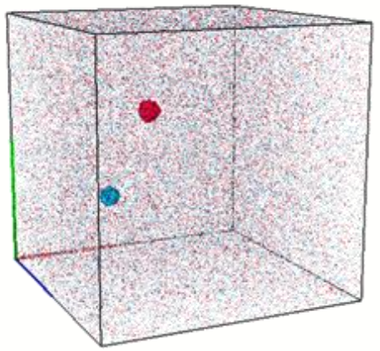
2D histogram



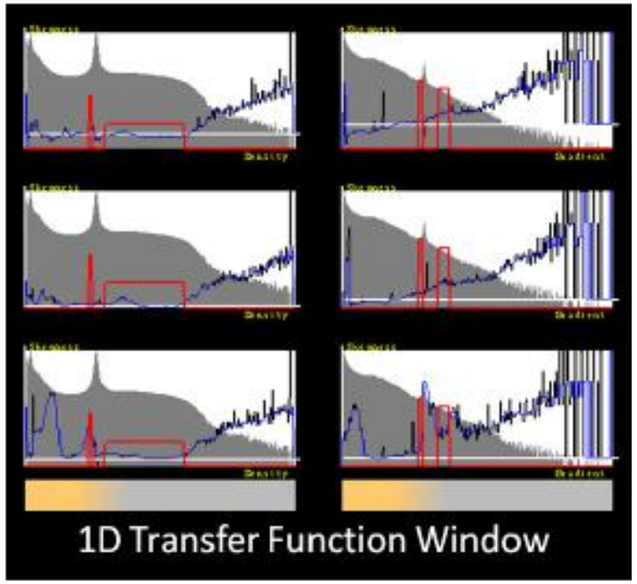
Std. Dev.



User selection (Std. Dev.)



Linking 1D and 2D Transfer Functions



Purdue University is Hiring!

- The School of Electrical and Computer Engineering at Purdue University invites applications for a faculty position at any level in human-centered computing, including but not limited to visualization, visual analytics, human-computer interaction (HCI), imaging, and graphics.
- For more information contact ebertd@purdue.edu (post-doc positions open too!)



VIS 2013 VAST • INFOVIS • SCIVIS BIOVIS • LDAV

Arizona State Is Hiring!

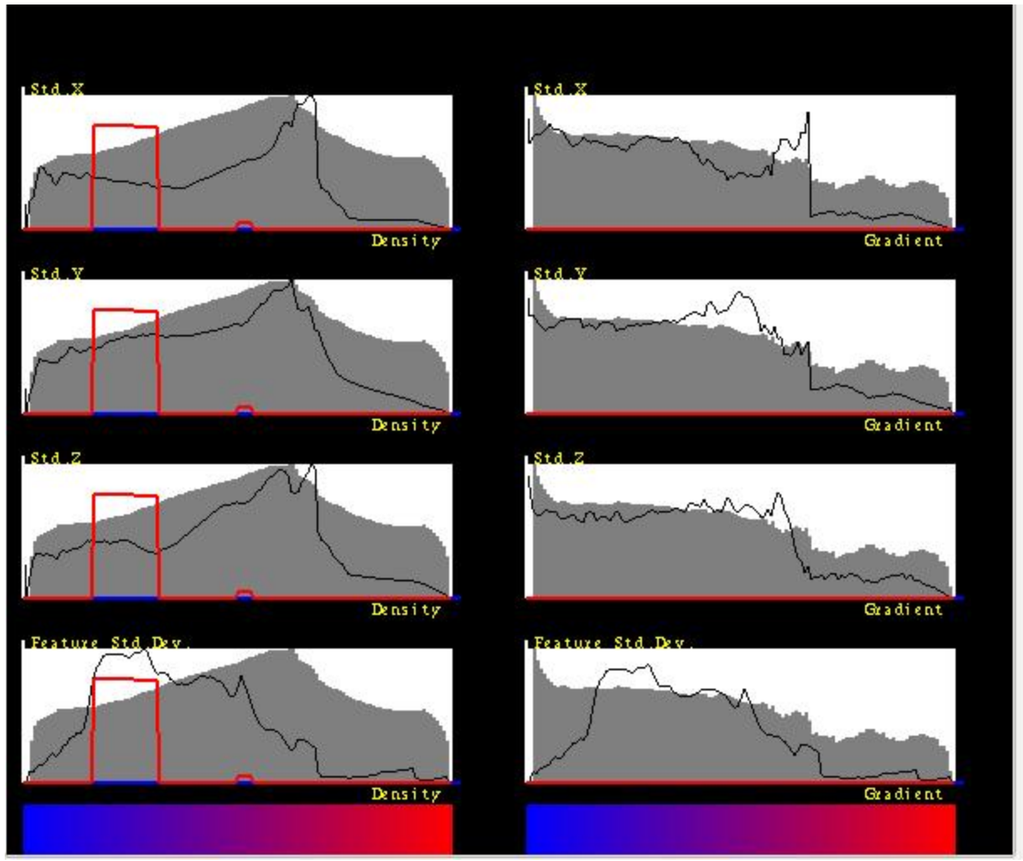
- Arizona State University is hiring in Computer Science
  - Big Data (Job #10518)
  - Cyber Security (Job #10519)
  - Next Gen Computing (Job #10523)
  - Software Engineering (Job #10524)
  - Human Centered Computing (Job #10555)
  - Emerging Computational Technologies for Learning (Job #10556)
- http://engineering.asu.edu/hiring
- I have lots of fliers if you want one!



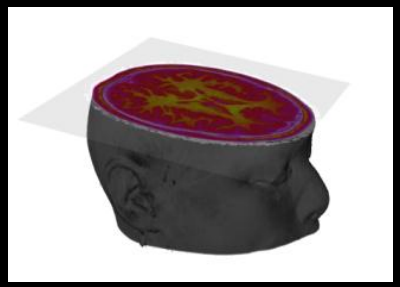
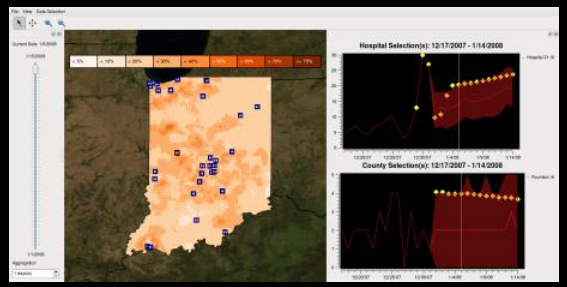
VIS 2013 VAST • INFOVIS • SCIVIS BIOVIS • LDAV

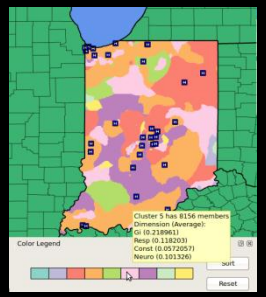
Questions?

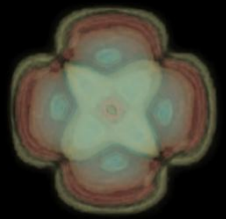
- This work has been supported by the U.S. Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001 and the US National Science Foundation (NSF) under Grants 0328984 and 0121288.







Analytical Brushing in Spatiotemporal Analysis

Ross Maciejewski



Ross Maciejewski
E-mail: rmacieje@asu.edu
http://rmaciejewski.faculty.asu.edu/



Spatiotemporal Analysis

Much of the data being collected and analyzed today is spatiotemporal

- Twitter[1,2]
- Emergency Departments[3]
- Law Enforcement[4]
- Mobile Phones[5,6]
- Animal Migration[7]
- Flow Simulations[8]

1 - A. M. MacEachren, A. Jaiswal, A. C. Robinson, S. Pezanowski, A. Savelyev, P. Mitra, X. Zhang, and J. Blanford. "Senseplace2: Geotwitter analytics support for situational awareness." In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, 2011.

2 - Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D. S., Ertl, T., "Spatiotemporal Social Media Analytics for Abnormal Event Detection Using Seasonal-Trend Decomposition," *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012.

3 - Maciejewski, R., Rudolph, S., Hafen, R., Mitchell, M. A., Larew, S. G., Cleveland, W., Ebert, D., "Forecasting Hotspot – A Predictive Analytics Approach," *IEEE Transactions on Visualization and Computer Graphics*, 17(4): 440-453, 2011.

4 - Malik, A., Maciejewski, R., Collins, T., Ebert, D., T., "Visual Analytics Law Enforcement Toolkit," *IEEE International Conference on Technologies for Homeland Security*, 2010.

5 – Z. Shen and K.-L. Ma. Mobivis: A visualization system for exploring mobile data. In *IEEE Pacific Visualization*, pages 175–182, 2008.

6 - G. L. Andrienko, N. V. Andrienko, C. Hurter, S. Rinzivillo, and S. Wrobel. "From movement tracks through events to places: Extracting and characterizing significant places from mobility data." In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, 2011.

7 – N. Ferreira, L. D. Lins, D. Fink, S. Kelling, C. Wood, J. Freire, C. T. Silva, "BirdVis: Visualizing and Understanding Bird Populations." IEEE Trans. Vis. Comput. Graph. 17(12): 2374-2383 (2011)

8 – J. Woodring, H.-W. Shen, "Semi-Automatic Time-Series Transfer Functions via Temporal Clustering and Sequencing." Computer Graphics Forum 28(3): 791-798 (2009)

Carrier 10:11 AM 100%

Coast Guard Search and Rescue Visual Analytics

View Settings Visualize Heatmap Data Risk Assessment Boat Registration Help

Max. Days: 7

2011-2012	M	T	W	R	F	S	S	Total
7/24-7/24							11	11
7/25-7/31	2	4	5	3	3	18	15	51
8/1-8/7	5	11	7	4	8	6	4	45
8/8-8/14	6	8	6	4	9	15	6	54
8/15-8/21	2	3	6	10	3	10	12	46
8/22-8/28	8	4	6	4	2	4	7	35
8/29-9/4	9	3	5	4	5	10	11	47
9/5-9/11	4	3	2	5	1	2	11	28
9/12-9/18	6	6	6	3	1	1	2	25
9/19-9/25	2	3	3	4	3	5	7	27
9/26-10/2	4	4	2	3	5	6	3	27
10/3-10/9	1	4	3	1	3	4	9	25
10/10-10/16	3	4	3	5	1	3	6	25
10/17-10/23	1	2	2	4	2	2	5	18
10/24-10/30	3	7	4	1	1	3	2	21
10/31-11/6	2	3	2	2	5	4	3	21
11/7-11/13	1	4	2	3	1	2	7	20
11/14-11/20	2	0	1	1	2	4	0	10
11/21-11/27	1	4	1	1	2	1	6	16
11/28-12/4	4	1	2	1	1	0	2	11
12/5-12/11	1	2	0	1	2	4	3	13
12/12-12/18	1	1	1	0	0	2	1	6
12/19-12/25	0	0	2	1	0	3	1	7
12/26-1/1	1	3	1	2	0	0	0	7
1/2-1/8	1	1	0	1	1	1	1	6
1/9-1/15	0	1	0	3	3	1	2	10
1/16-1/22	1	1	1	2	2	2	2	11
1/23-1/29	2	2	2	0	0	1	0	7
1/30-2/5	0	1	1	1	0	0	1	4
2/6-2/12	1	0	0	1	1	3	0	6
2/13-2/19	1	0	2	2	0	2	1	8
2/20-2/26	1	1	1	2	1	1	0	7
2/27-3/4	2	2	0	1	1	2	0	8
3/5-3/11	2	1	2	2	2	3	0	12
3/12-3/18	1	0	1	2	4	2	2	12
3/19-3/25	2	4	1	2	0	0	1	10
3/26-4/1	0	0	0	0	0	0	0	0
4/2-4/8	0	1	0	2	0	0	0	3
4/9-4/15	0	0	1	0	1	1	3	6
4/16-4/22	0	0	1	1	0	0	4	6
4/23-4/29	1	0	0	0	0	1	0	2
4/30-5/6	0	0	0	2	2	0	2	6
	84	99	85	91	78	130	153	

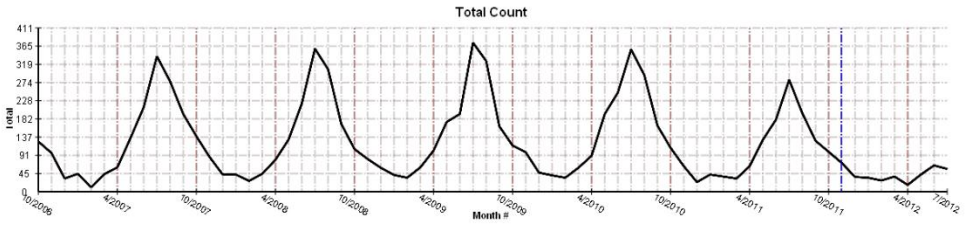
Month 11 Year 2011 Play

10/2006 — 7/2012

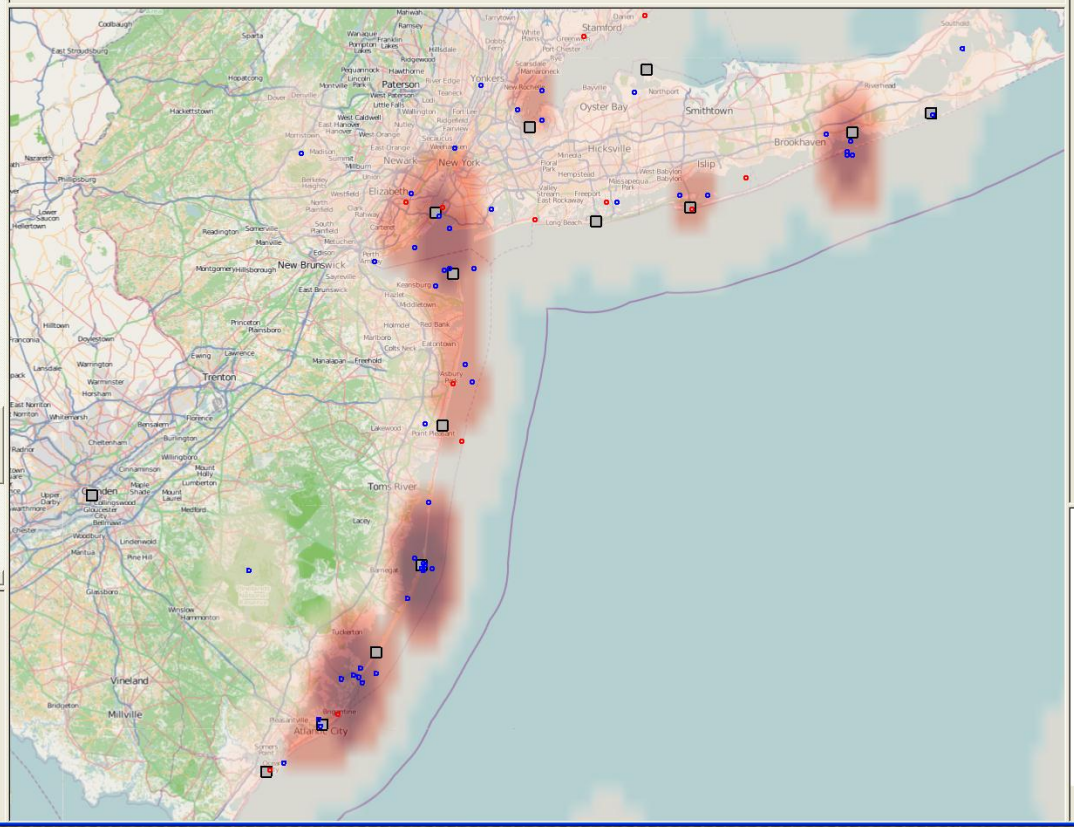
Day Week Month Year Fiscal Year

Aggregate: 10/ 2/2006 to 7/23/2012

Apply Boating Season Non Boating Season



Change tick frequency: 10 Line Graph Bar Graph Current Value: 72 Separate Total COTP Distress Type Unit Type Grid Lines



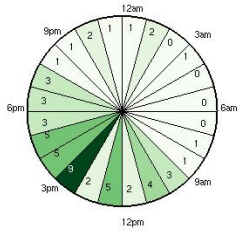
406 MHZ EPIRB Abandoned/Derelict Adrift (Unmanned) Aground
Capsized Vessel Collision **Disabled Vessel** Disoriented Vessel
Distress Alert - situation unknown Diving Accident Flare (Red/Orange) MEDEVAC MEDICO
Person in Water (Ice) Person in Water (PIW) Special Operation
Stranded (on island) **Taking on Water (TOW)**
Uncorrelated MAYDAY Unreported Vessel

Distress Type | Unit Selection | Initiation Type | Case Status | COTP | District Tree

SAR Select all Select none

Search: Clear

- 121.5 MHZ EPIRB
- 243.0 MHZ ELT
- 406 MHZ EPIRB
- 406 MHZ PLB
- Abandoned/Derelict
- Adrift (Unmanned)
- Aground
- Aircraft Crash
- Aircraft Emergency
- Alleged Violation of Law/Regulation
- Allision
- Anchored (Unmanned)
- Assist Other Agency
- Asylum Requests
- Atlantic Right Whale Sighting
- Beset by Weather
- Boating Under the Influence
- Boating Under the Influence/Cite and Release
- Boating Under the Influence/Detaining for Local Authorities
- Bomb Threat
- Bomb Threat Against Commercial Vessel/Facility
- Bomb Threat Against Military Vessel/Facility
- Breakaway
- Bridge Closures
- Bridge Jumper
- Building Intruder and Duress Alarm
- Capsize
- Capsized Vessel
- Cargo/Container Inspection
- Casualties to CG Cutters & Aircraft
- Civil Disasters / Disturbances
- Collision
- Commercial Fishing Vessel Safety
- Commercial Vessel Safety Enforcement
- Decedent Affairs
- Dependent Emergencies
- Disabled Vessel



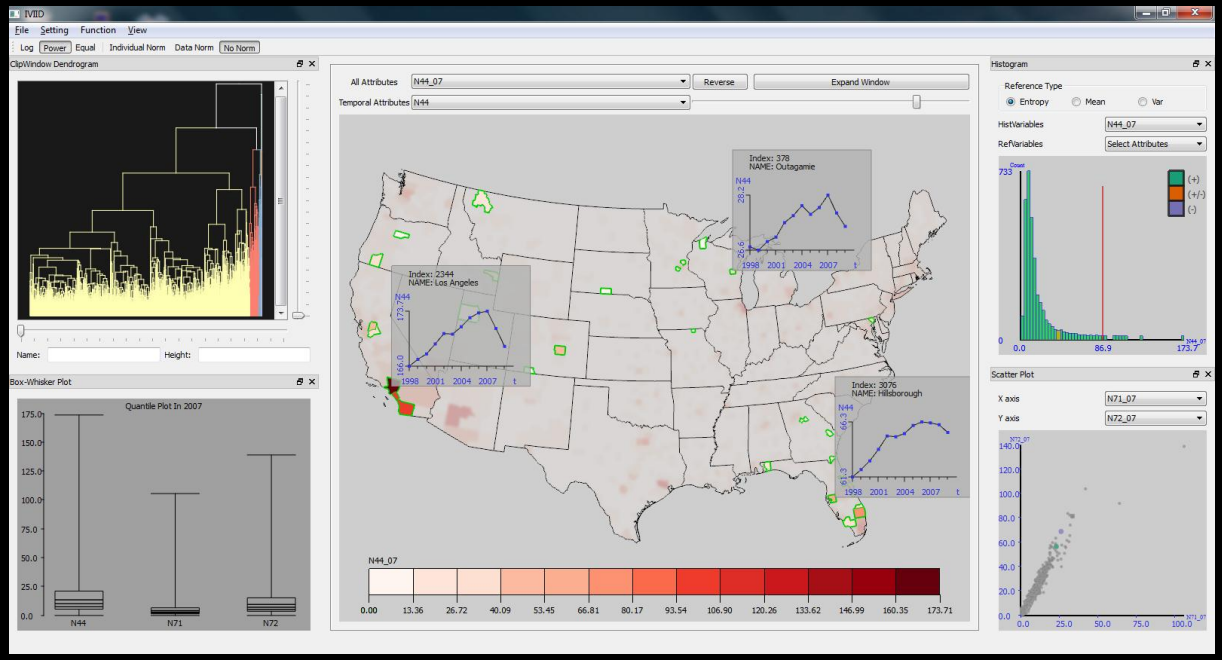
Accuracy : 76%

2010-08-14 to 2010-08-20

Coordinated Multiple Views in Spatiotemporal Analysis

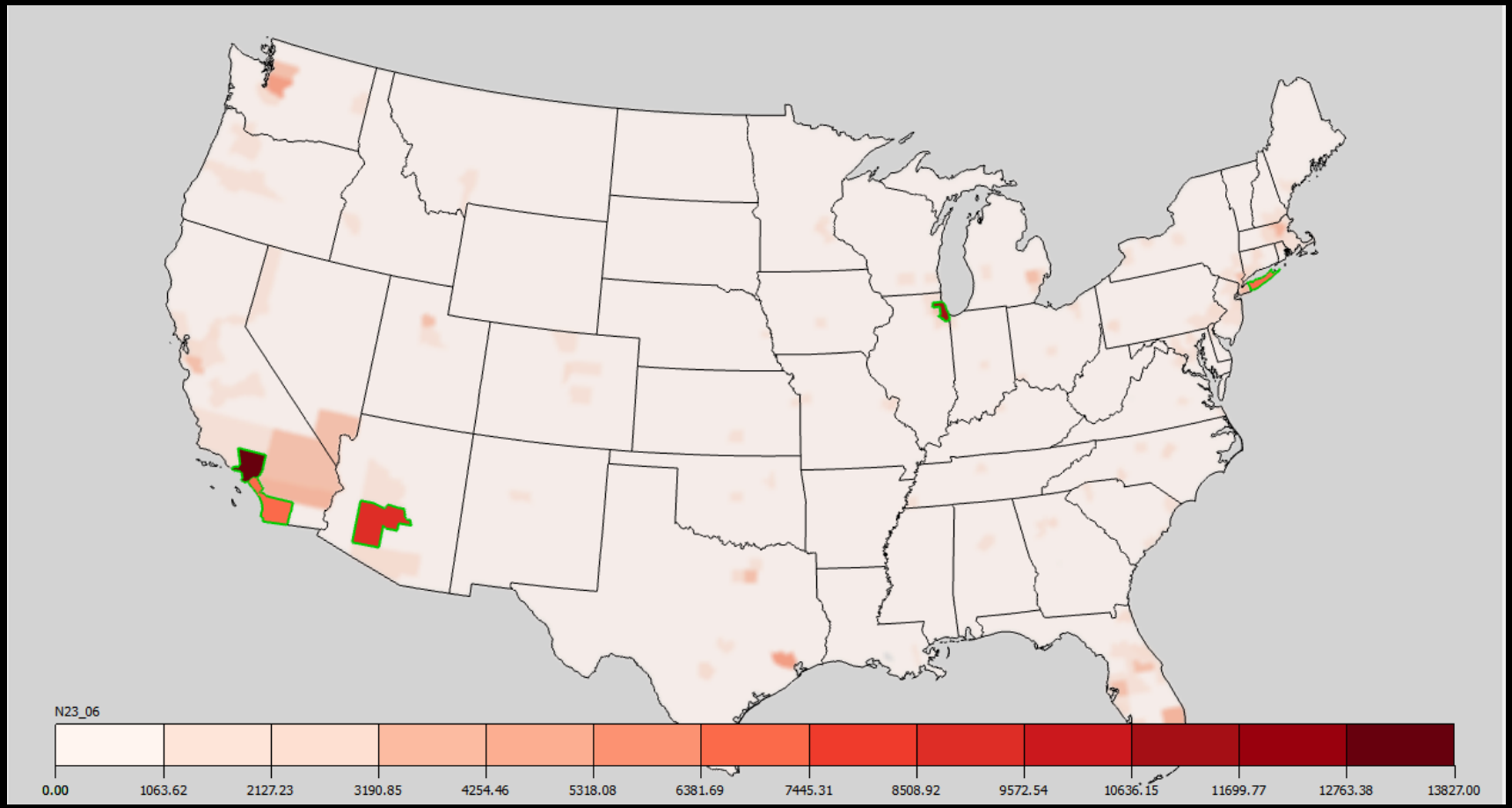
To facilitate answering these questions, we typically rely on interactive coordinated multiple views[1]

- Geographic Maps
- Scatterplots
- Parallel Coordinate Plots
- Dendogram
- Histogram
- Time Series Plots
- Tabular View
- Box-and-Whisker

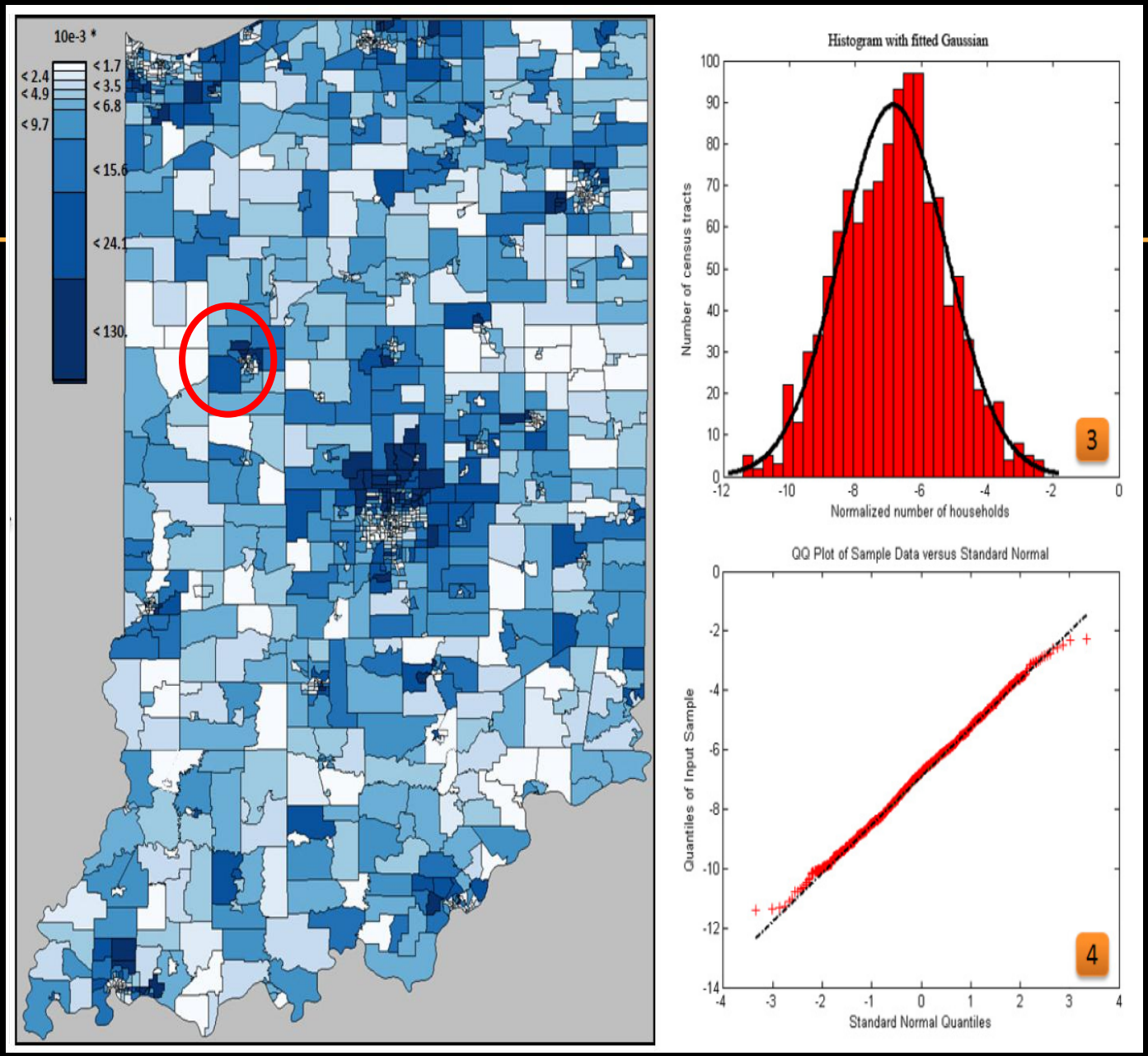


1 – M. Q. W. Baldonado, A. Woodruff, and A. Kuchinsky, "Guidelines for Using Multiple Views in Information Visualization," in *Proc. of Advanced Visual Interfaces (AVI 2000)*, pages 110-119. ACM Press, 2000

Facilitating Spatial Analysis: Focus+Context Brushing



1 - J. W. Tukey. *Exploratory Data Analysis, 1977*. Addison-Wesley
2 – N. Andrienko, G.Andrienko, and P.Gatalsky, "Exploratory Spatio-Temporal Visualization: an Analytical Review *Journal of Visual Languages and Computing*," special issue on Visual Data Mining December 2003, v.14 (6), pp. 5403-5413
3 - A. Robinson. Highlighting in geovisualization. *Cartography and Geographic Information Science*, In Press.



Maciejewski, R., Pattah, A., Ko., S., Hafen, R., Cleveland, W. S., Ebert, D. S., "Automated Box-Cox Transformations for Improved Visual Encoding," *IEEE Transactions on Visualization and Computer Graphics*, 19(1): 130-140, 2013.

Visual Analytics

Visualization is good for exploring data, but we can do more than just explore

"Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces"[1]

Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets.[2]

"A graphic display has many purposes but it achieves its highest value when it forces us to see what we were not expecting."[3]

1 - J. J. Thomas and K. A. Cook (Ed.). *Illuminating the Path: The R&D Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.
2 – Keim et al. chapter in *Information Visualization: Human-Centered Issues and Perspectives*, 2008.
3 – H. Wainer. *A Graphic Discovery: A Trout in the Milk and Other Visual Adventures*.

Spatiotemporal Analysis

Key spatial questions tend to be about causal relationships

- Do neighboring spatial regions have similar values?
- Do neighboring spatial regions have similar temporal trajectories?
- Are disparate spatial regions showing similar trajectories or values?
- Are temporal trends lagging or leading in neighboring spatial regions?

Key Temporal questions tend to focus on change[1]

- What is the magnitude of the change?
- What is the shape of the changes?
- How fast is the change occurring (what is the velocity of the change)?
- What is the direction of the change?

1 - S. Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

Challenges in Spatiotemporal Analysis

Exploring trends and patterns in spatiotemporal data

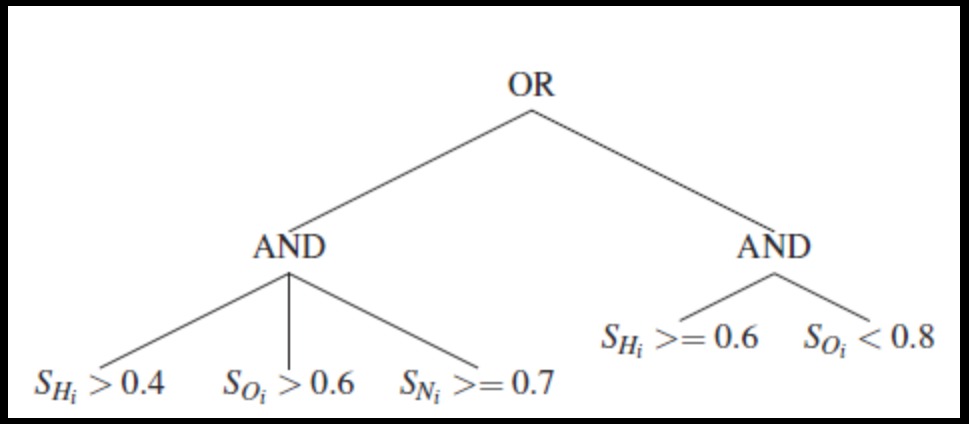
- Correlations among multi-variate data
- Explore possible predictive links among different variables

Challenges:

- End users understanding underlying statistical algorithms
- Applying algorithms at appropriate spatial and temporal scales
- Dealing with noisiness of the real-world data

Spatial Analysis: Linked Multivariate Filtering

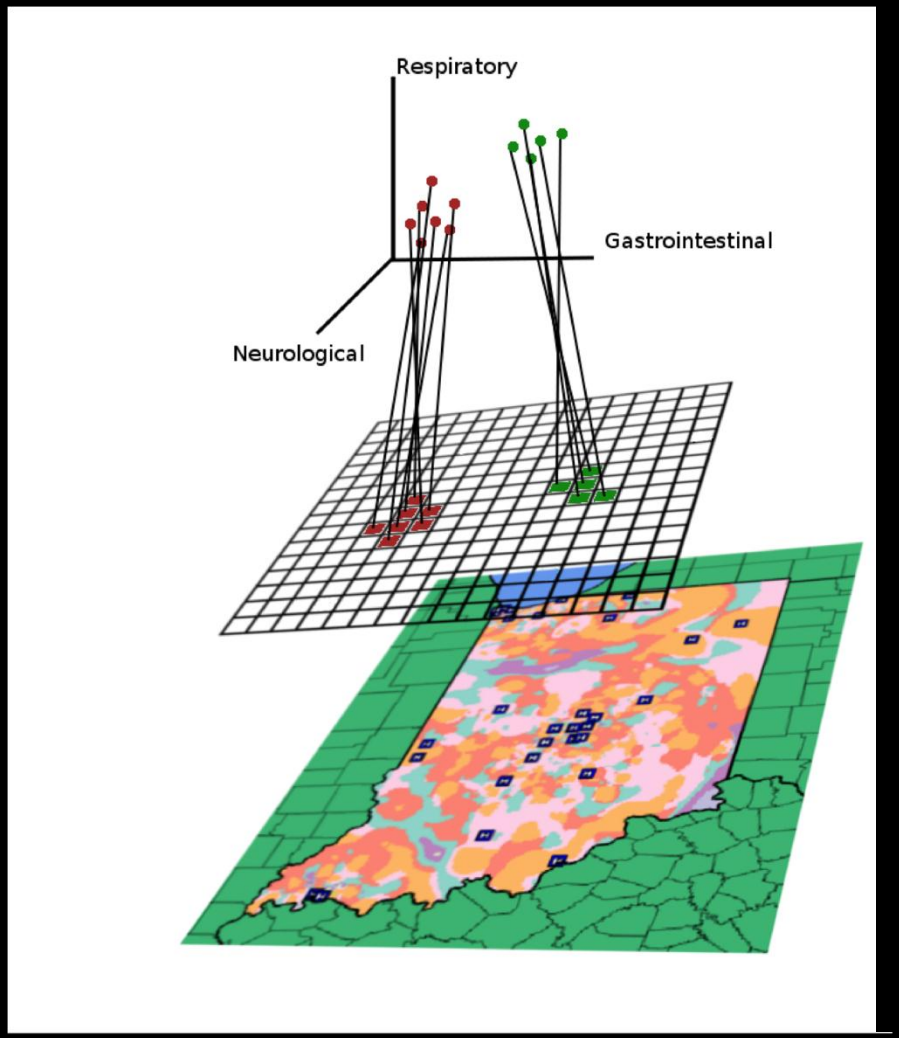
Link complex data queries directly to an interactive logic tree to facilitate multivariate queries

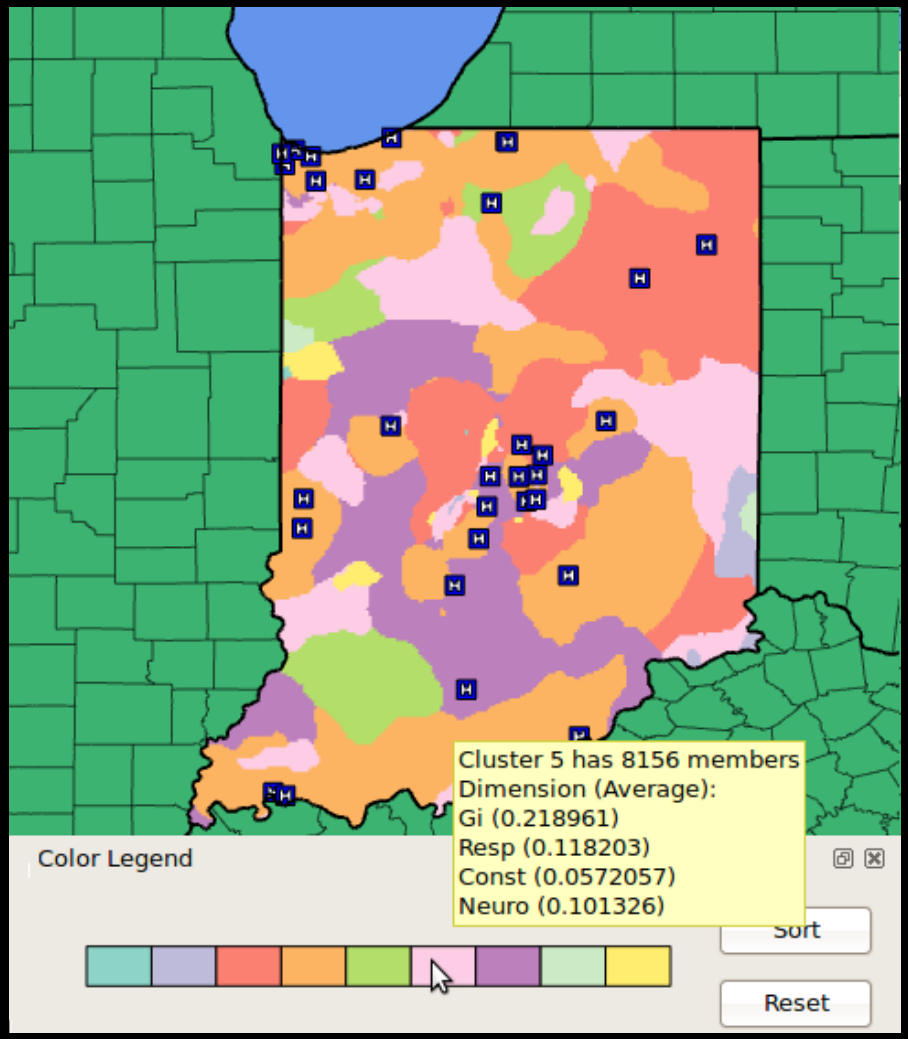


Current Variable	Comparing	Relation
Or		
And		
...	N23	< 2.77...
...	N22	< 65.3
...	N21	< 287
And		
...	N52	> 8.13...
...	N11	> 162

Problem is this is time consuming and tree gets complex quickly as the number of variables and constraints increase

Spatial Analysis: Multivariate Clustering





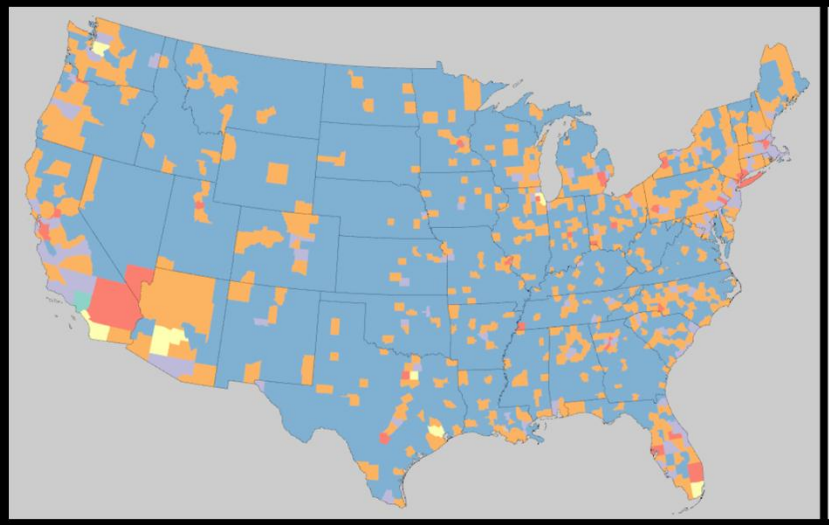
Spatial Analysis: Multivariate Clustering

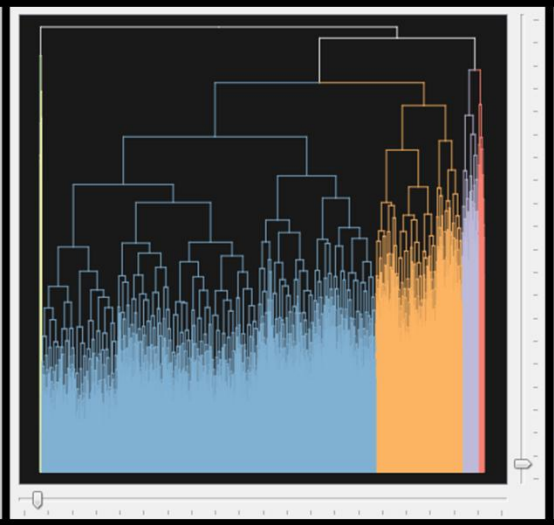
Hierarchical Clustering

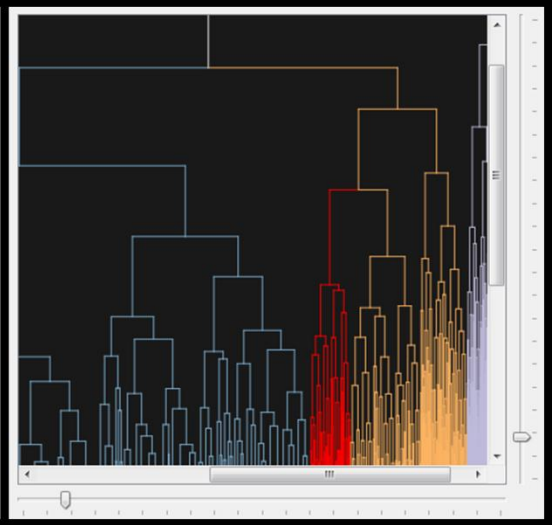
- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram

No assumptions on the number of clusters

- Any desired number of clusters can be obtained by ‘cutting’ the dendogram at the proper level







Spatial Analysis: Multivariate Distance Function

In these high-dimensional clusters, we don't necessarily have a concept of how far away two points are

We want to know how spatial units are related in high-dimensional space and observe these patterns in the geographic projection

We can directly link a distance function to the brush tip

Metric Names	Formula
Euclidean	$dist = \sqrt{\sum_i (v_i^1 - v_i^2)^2}$
Manhattan	$dist = \sum_i \lvert v_i^1 - v_i^2 \rvert$
Canberra	$dist = \sum_i \frac{\lvert v_i^1 - v_i^2 \rvert}{\lvert v_i^1 \rvert + \lvert v_i^2 \rvert}$
Binary	$dist = \sum_i v_i^1 = v_i^2$

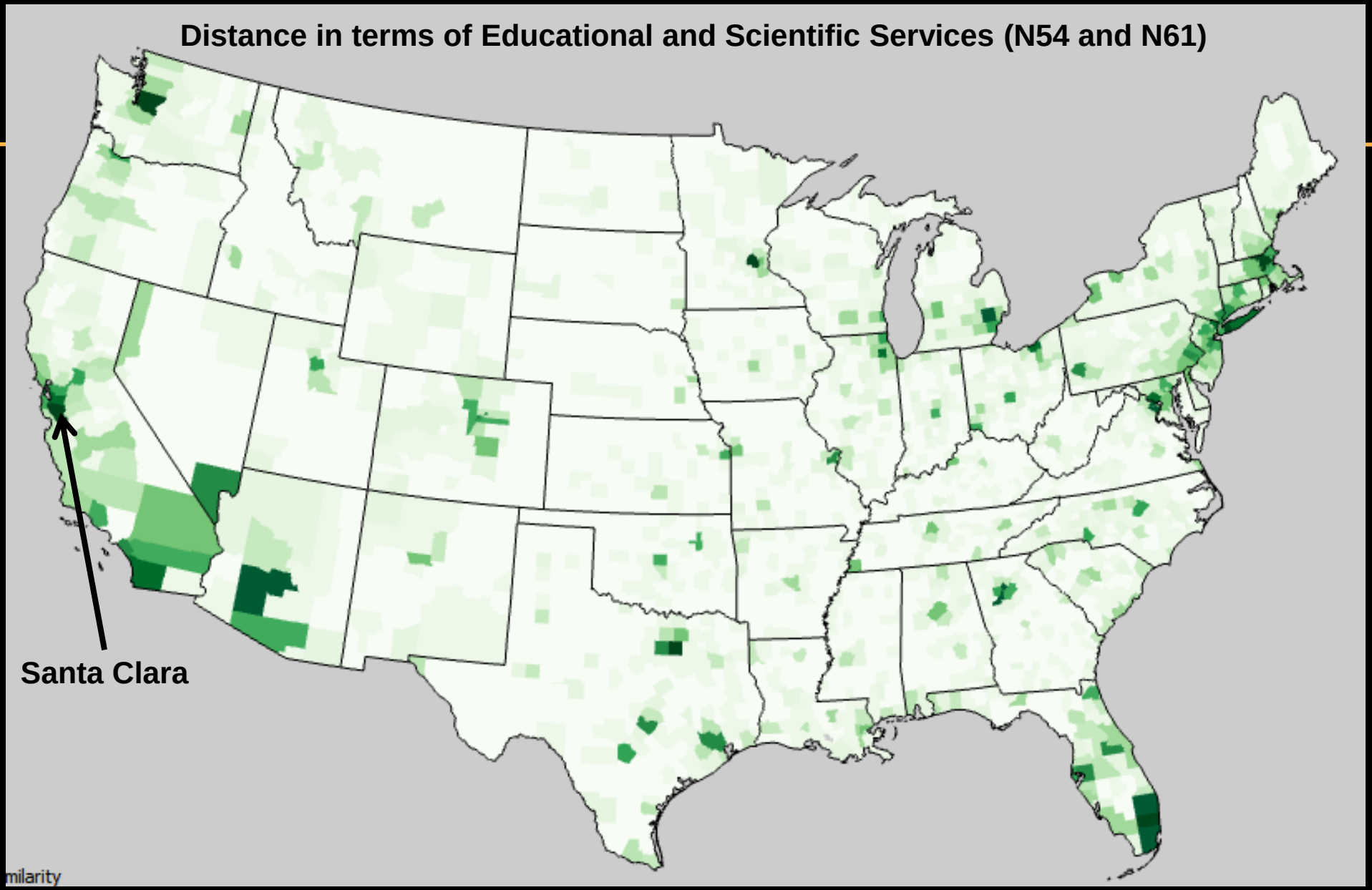
Indices of Industrial Diversity

Indices of industrial diversity measure the relative concentration of a particular economic variable across a specified number of industries

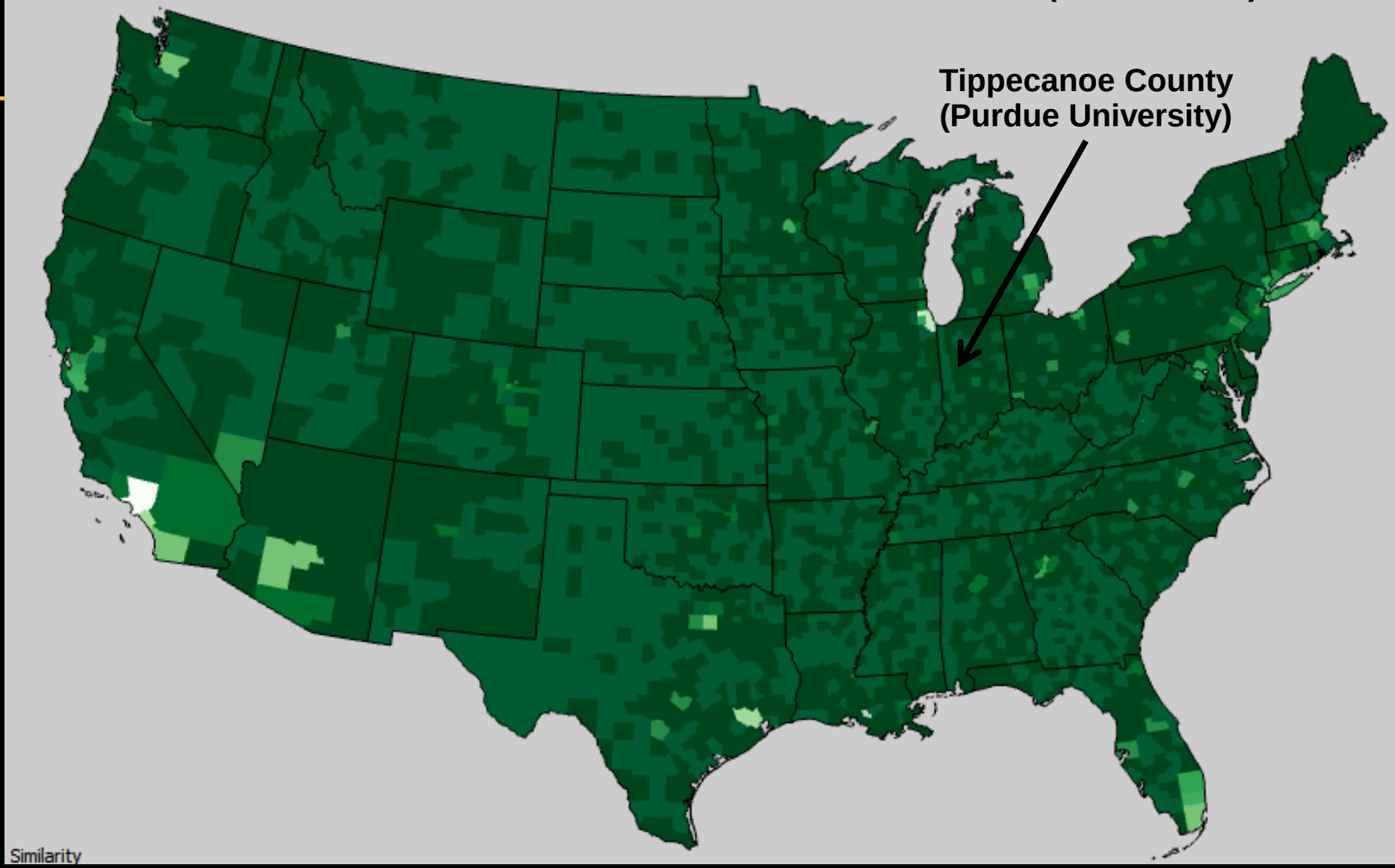
Industrial diversity is measured by the number of firms as reported by the North American Industrial Classification System (NAICS)

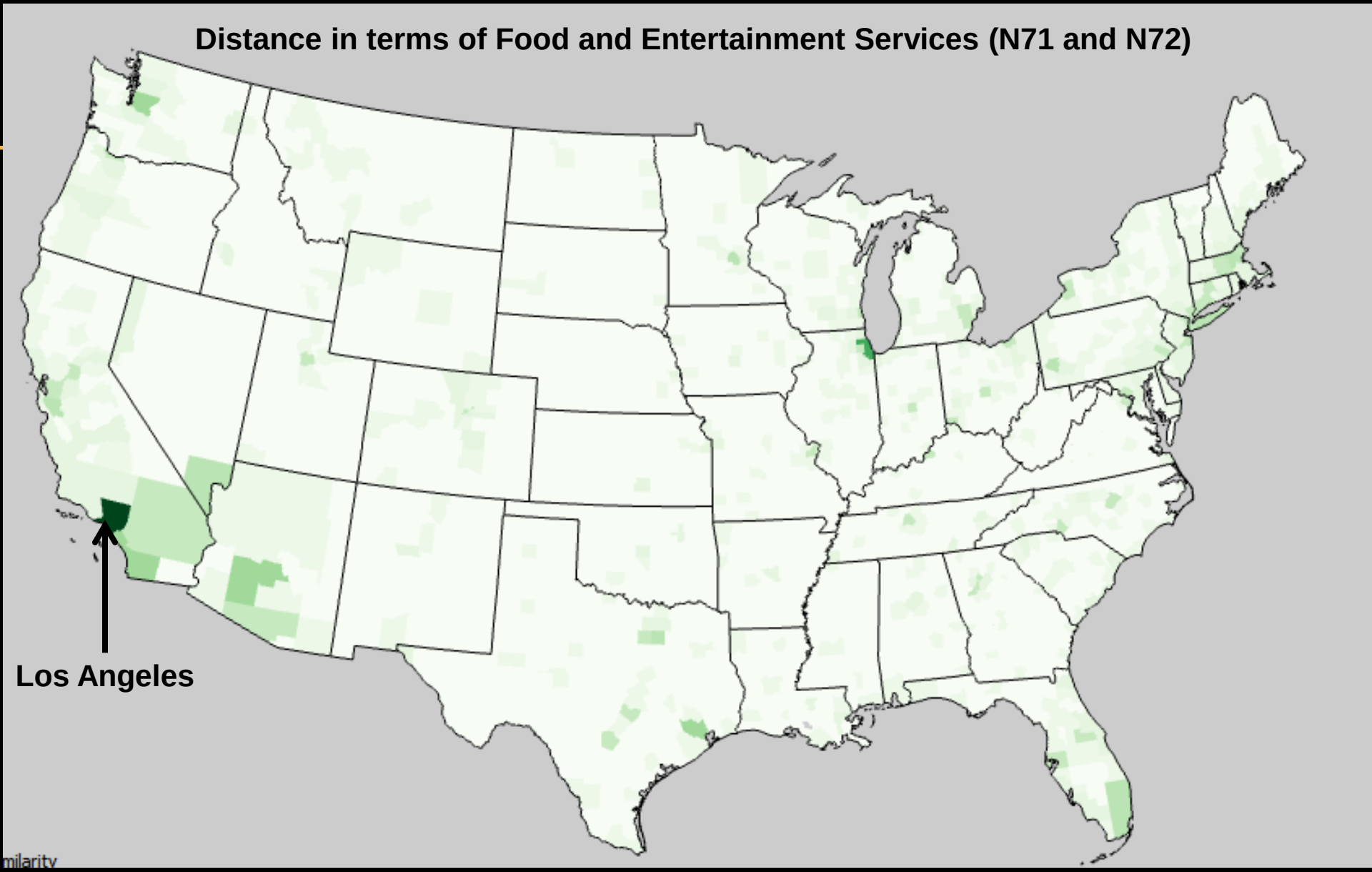
N11	Agriculture
N21	Mining
N22	Utilities
N23	Construction
N31	Manufacturing
N42	Wholesale Trade
N44	Retail Trade
N48	Transportation and Warehousing
N62	Health Care

N51	Information
N52	Finance and Insurance
N53	Real Estate
N54	Scientific and Technical Service
N55	Management of Companies
N56	Administrative Waste Services
N61	Educational Services
N71	Arts and Recreation
N72	Accommodation and Food Service

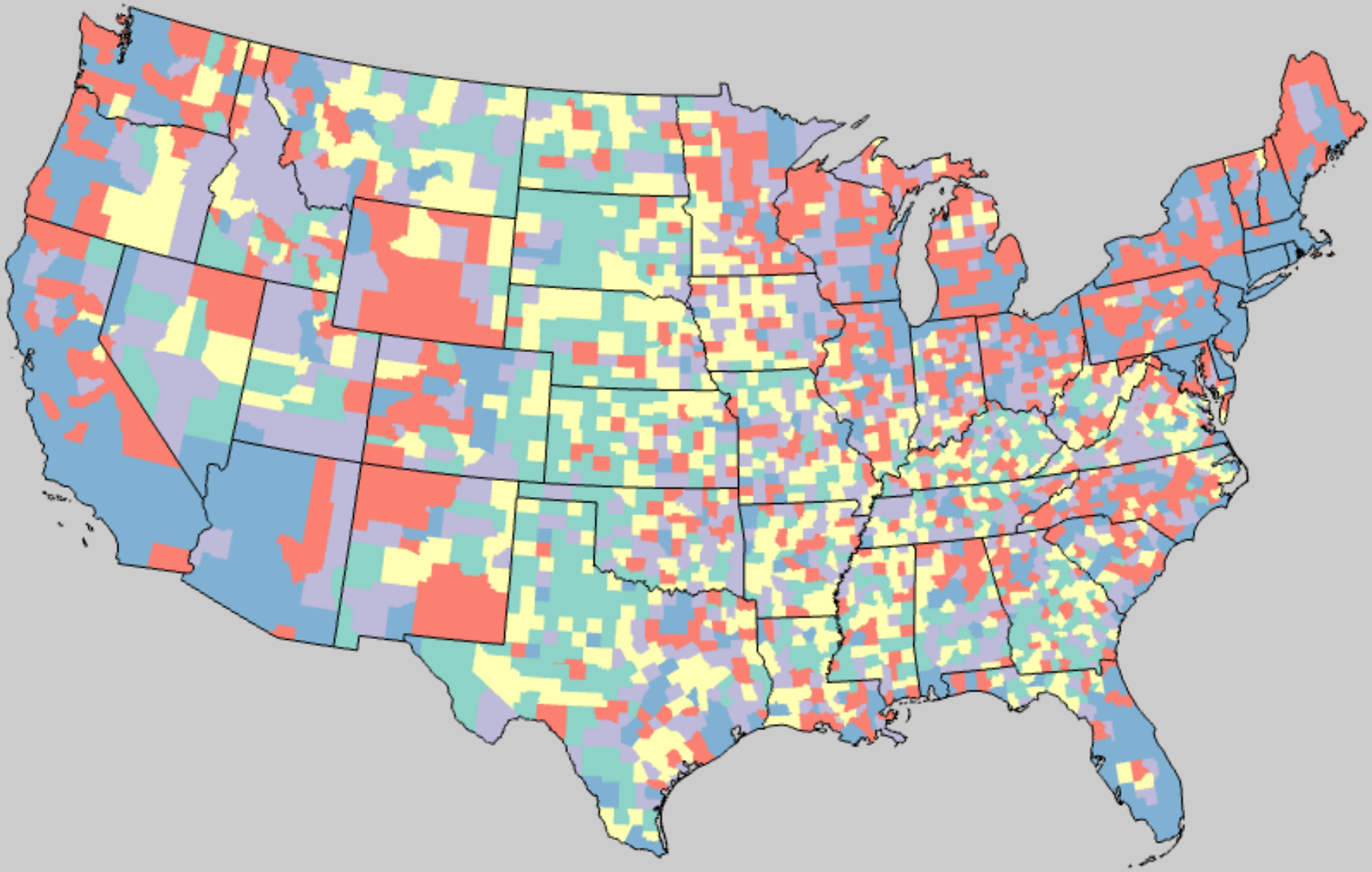

Distance in terms of Educational and Scientific Services (N54 and N61)
Santa Clara
milarity

Distance in terms of Educational and Scientific Services (N54 and N61)




Distance in terms of Food and Entertainment Services (N71 and N72)
Los Angeles
milarity

K-means clustering (k=5) in terms of Food and Entertainment Services (N71 and N72)



Spatial Analysis: Interactive Clustering

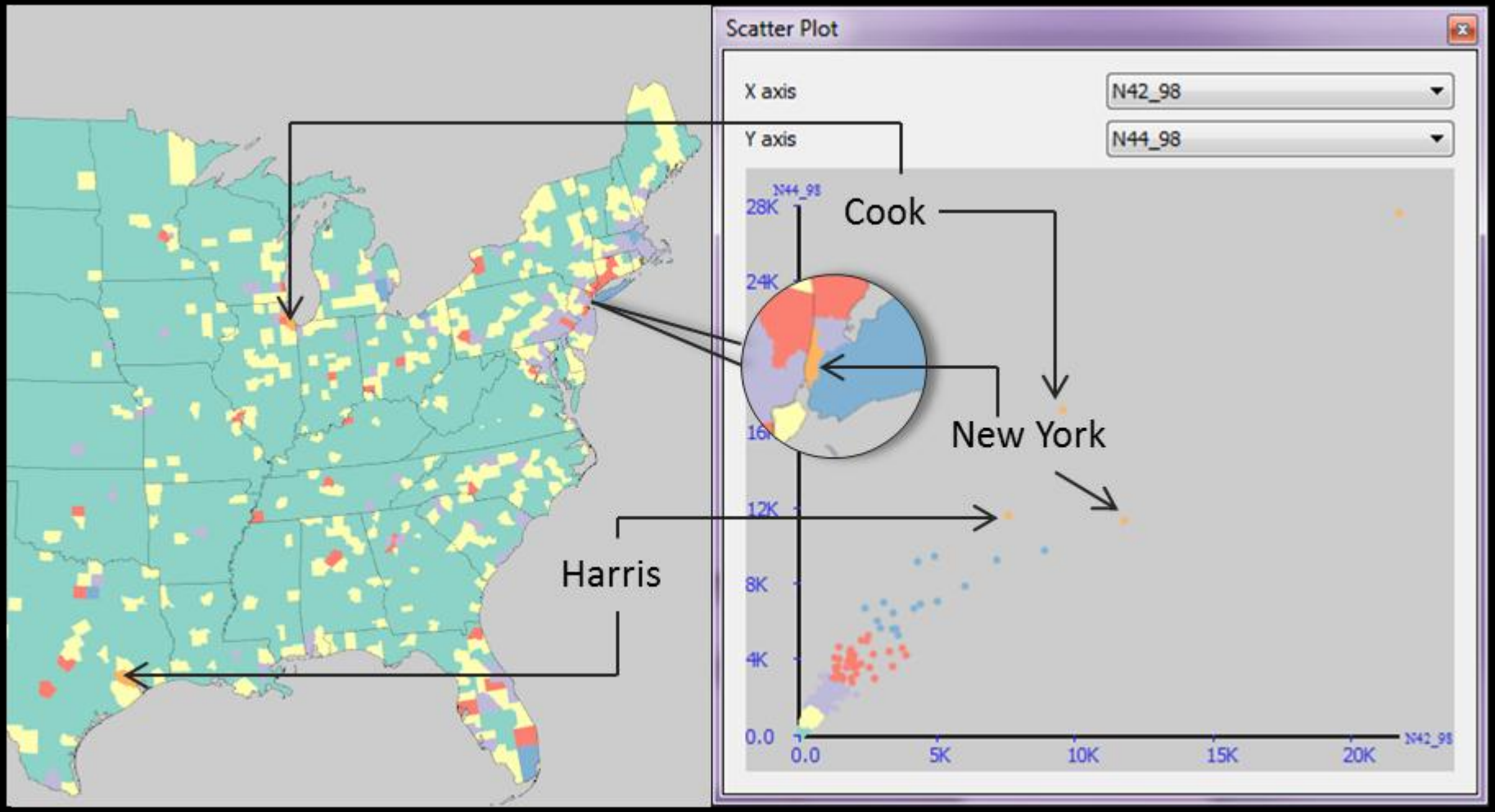
Adapt k-means so that each spatial unit has a unique weight (initially set so all weights are equal)

Analysts may now modify weights based on original k-means output, this is done by assigning spatial units to different clusters

Upon cluster reassignment, the weight of that spatial unit is updated to be $w_i = 1 + \sqrt{s_j}$ *where* s_j *is the size of cluster j*

Now the cluster centroids are recomputed and k-means rerun such that cluster centroids are $c_j = \frac{\sum_{k \in J} w_k u_k}{\sum_{k \in J} w_k}$

Spatial Analysis: Interactive Clustering



Spatiotemporal Analysis

Key spatial questions tend to be about causal relationships

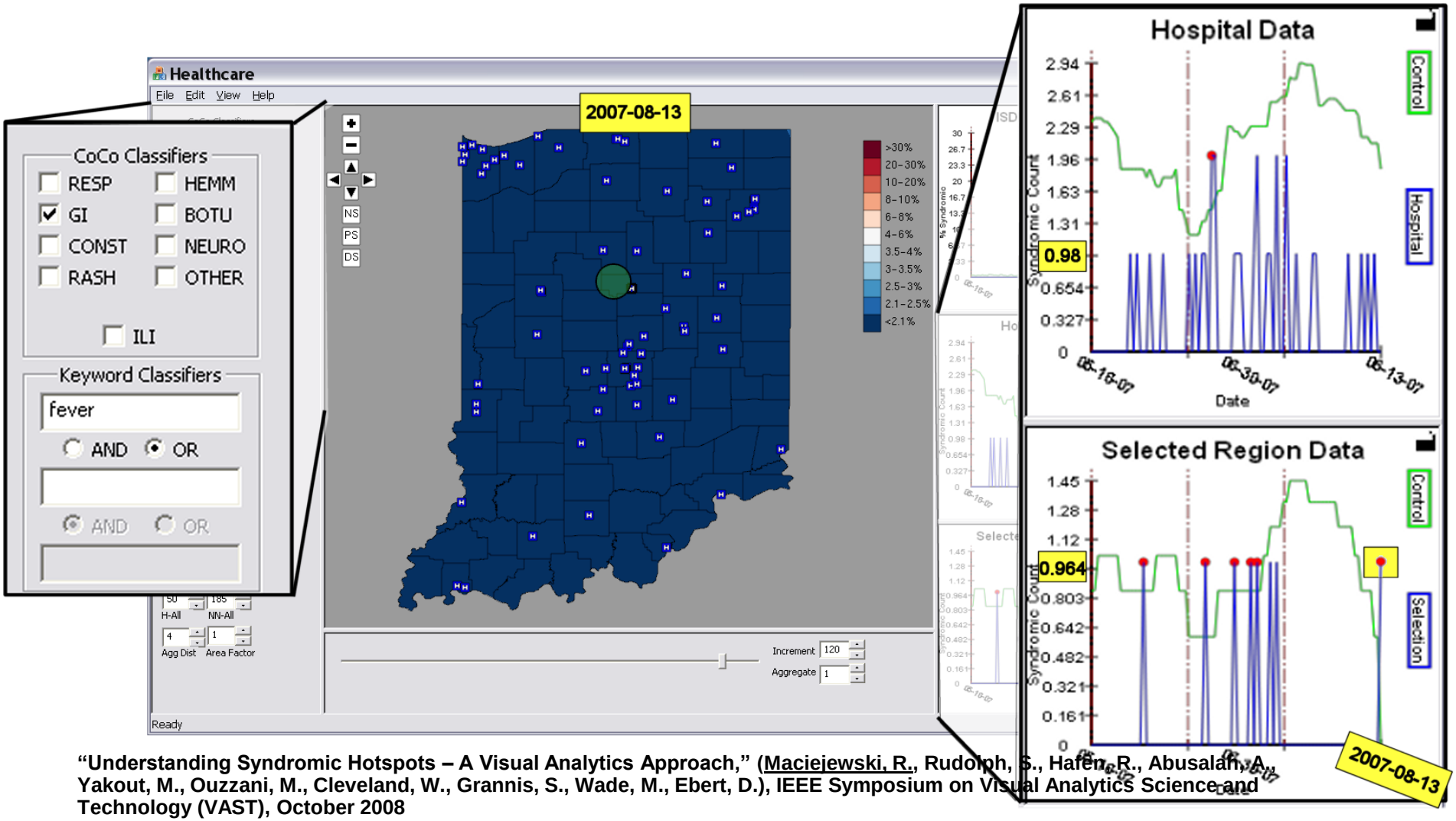
- Do neighboring spatial regions have similar values?
- Do neighboring spatial regions have similar temporal trajectories?
- Are disparate spatial regions showing similar trajectories or values?
- Are temporal trends lagging or leading in neighboring spatial regions?

Key Temporal questions tend to focus on change[1]

- What is the magnitude of the change?
- What is the shape of the changes?
- How fast is the change occurring (what is the velocity of the change)?
- What is the direction of the change?

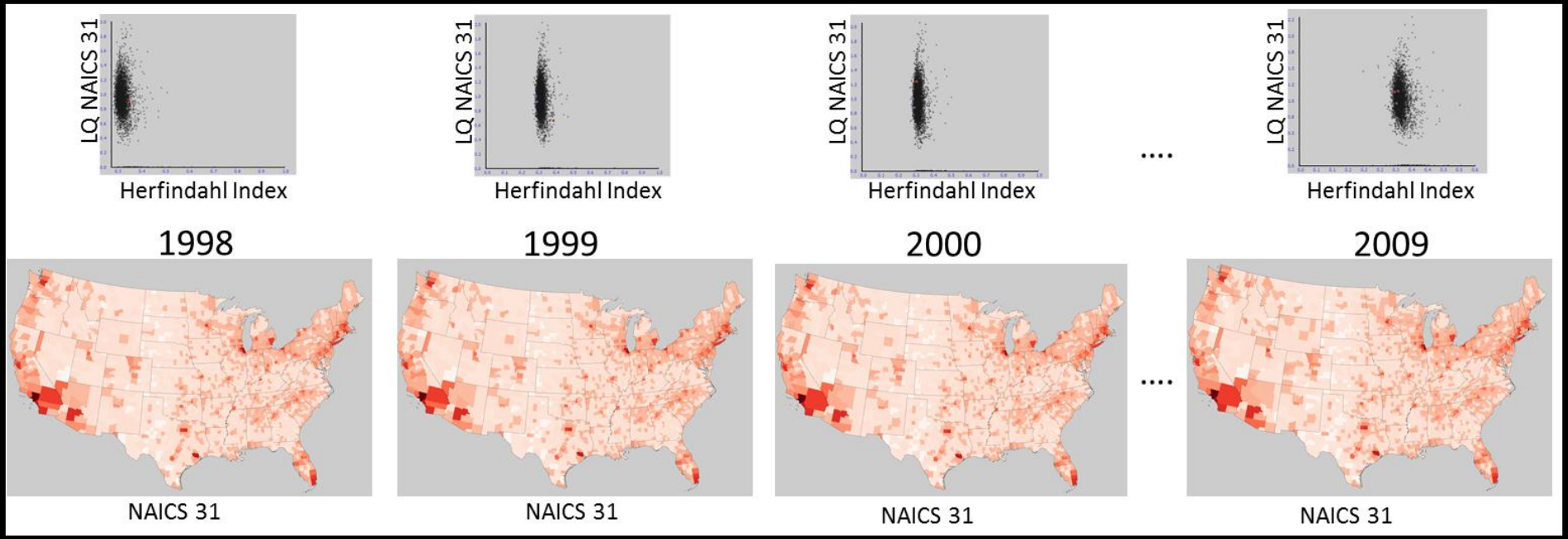
1 - S. Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

Facilitating Spatiotemporal Analysis: Linked Time Series



"Understanding Syndromic Hotspots – A Visual Analytics Approach," (Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W., Grannis, S., Wade, M., Ebert, D.), IEEE Symposium on Visual Analytics Science and Technology (VAST), October 2008

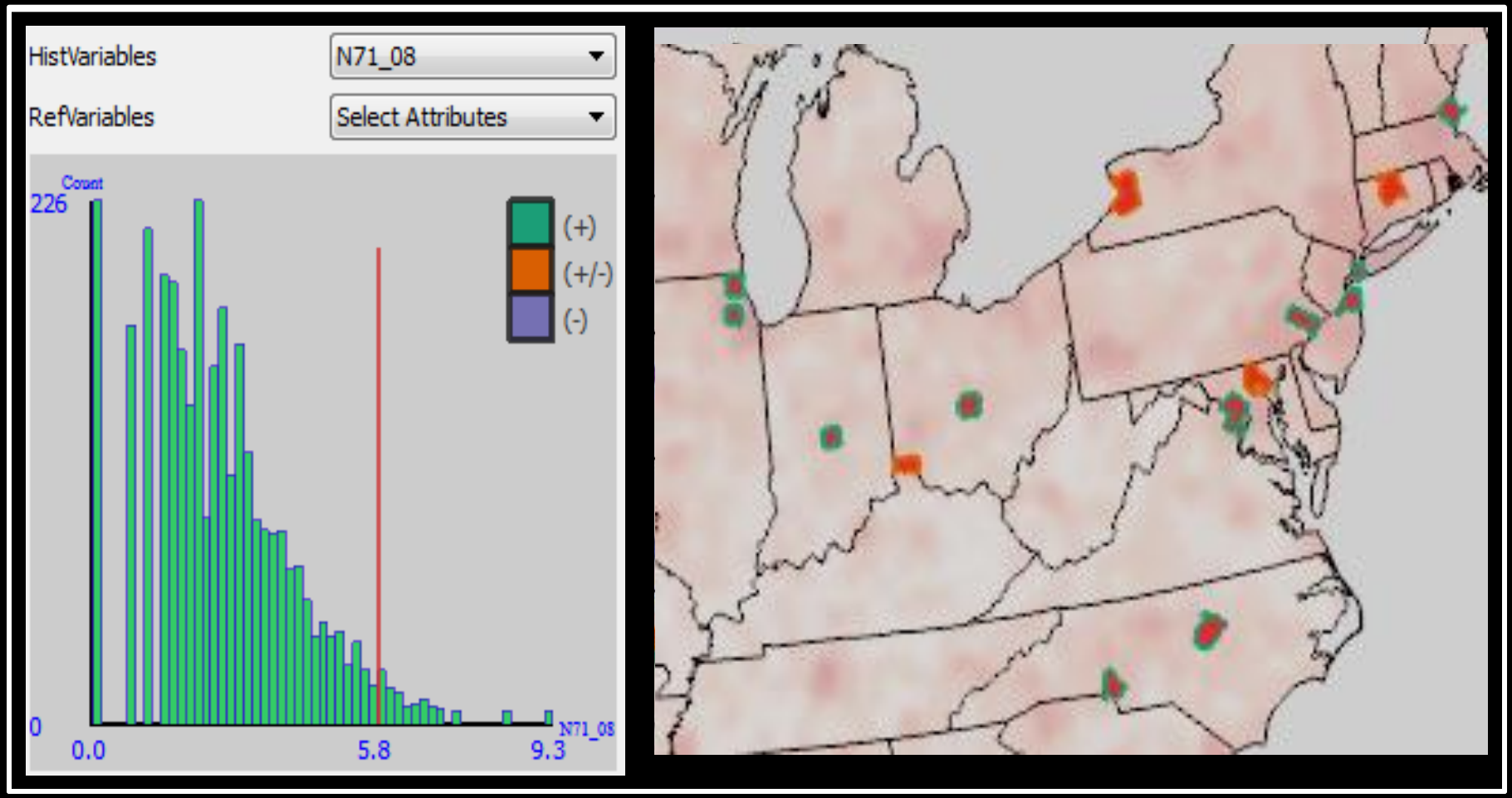
Facilitating Spatiotemporal Analysis: Animation and Small Multiples



1 – B. Tversky, J. B. Morrison, and M. Bertrancourt. Animation: Can it facilitate? *International Journal of Human-Computer Studies*, 57:247– 262, 2002.

2 – M. Harrower. The cognitive limits of animated maps. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 42(4):349–357, 2009.

Facilitating Spatiotemporal Analysis: Identifying Temporal Thresholds



Facilitating Spatiotemporal Analysis: Temporal Similarity Brushing

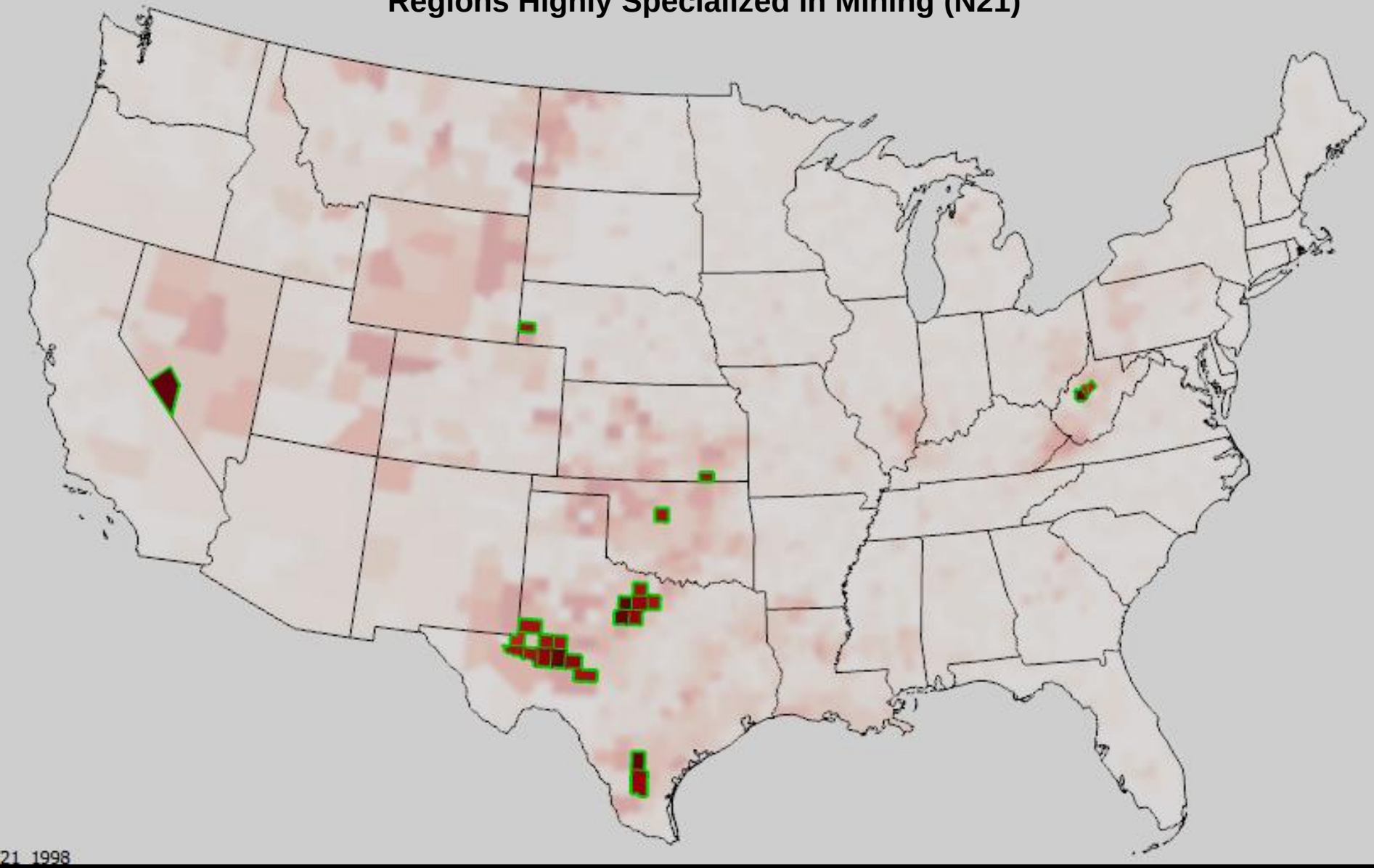
Many of these questions depend on temporal trajectories

We can apply time series similarity work as a means of determining which regions have similar trajectories

One common metric is the sequence normalized metric which will focus on matching trends as opposed to exact values

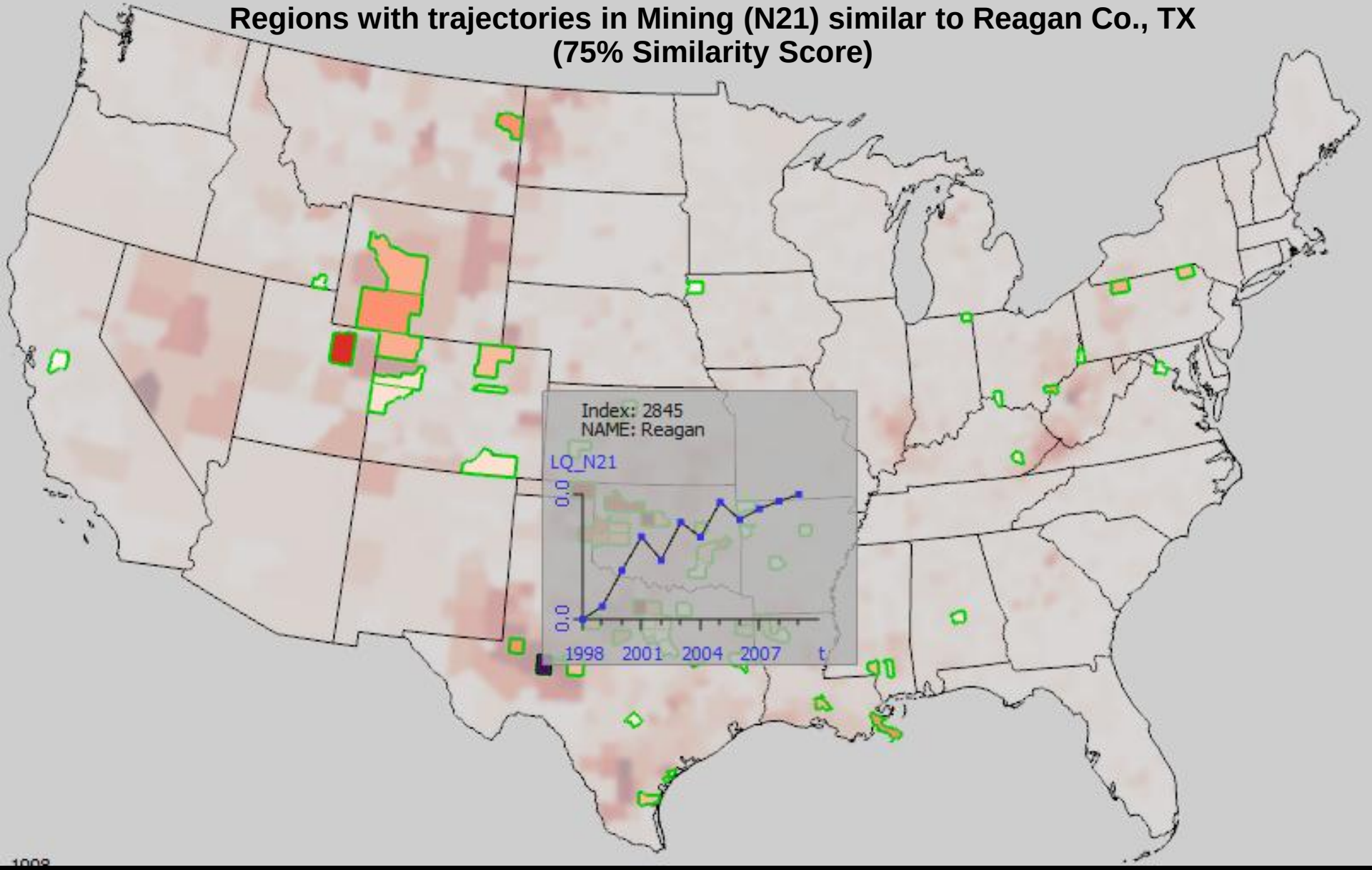
$$dist_{SNE}(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}\left(\frac{q_i-\mu_q}{\sigma_q}-\frac{p_i-\mu_p}{\sigma_p}\right)^2}.$$

Regions Highly Specialized in Mining (N21)

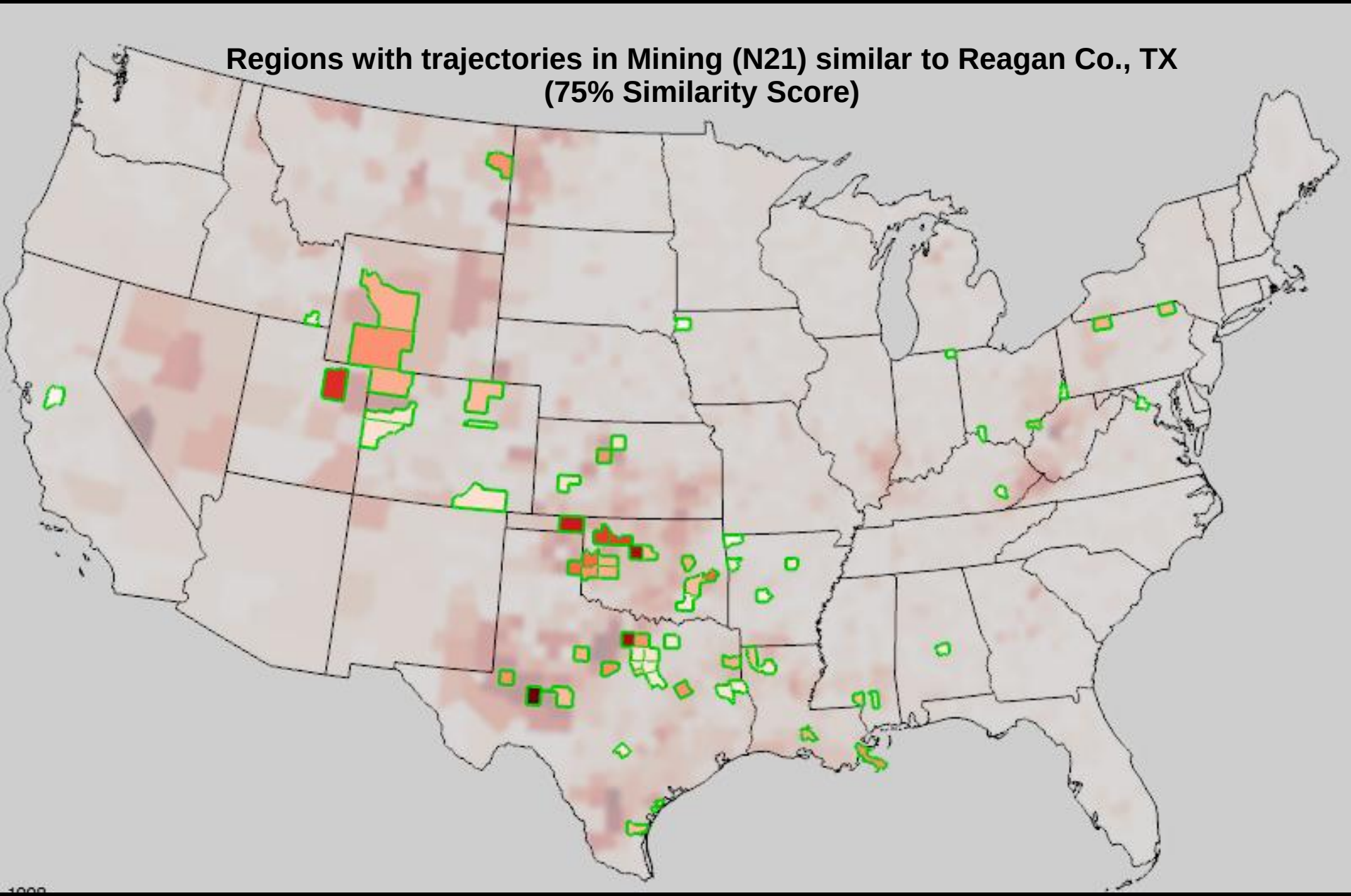


21 1998

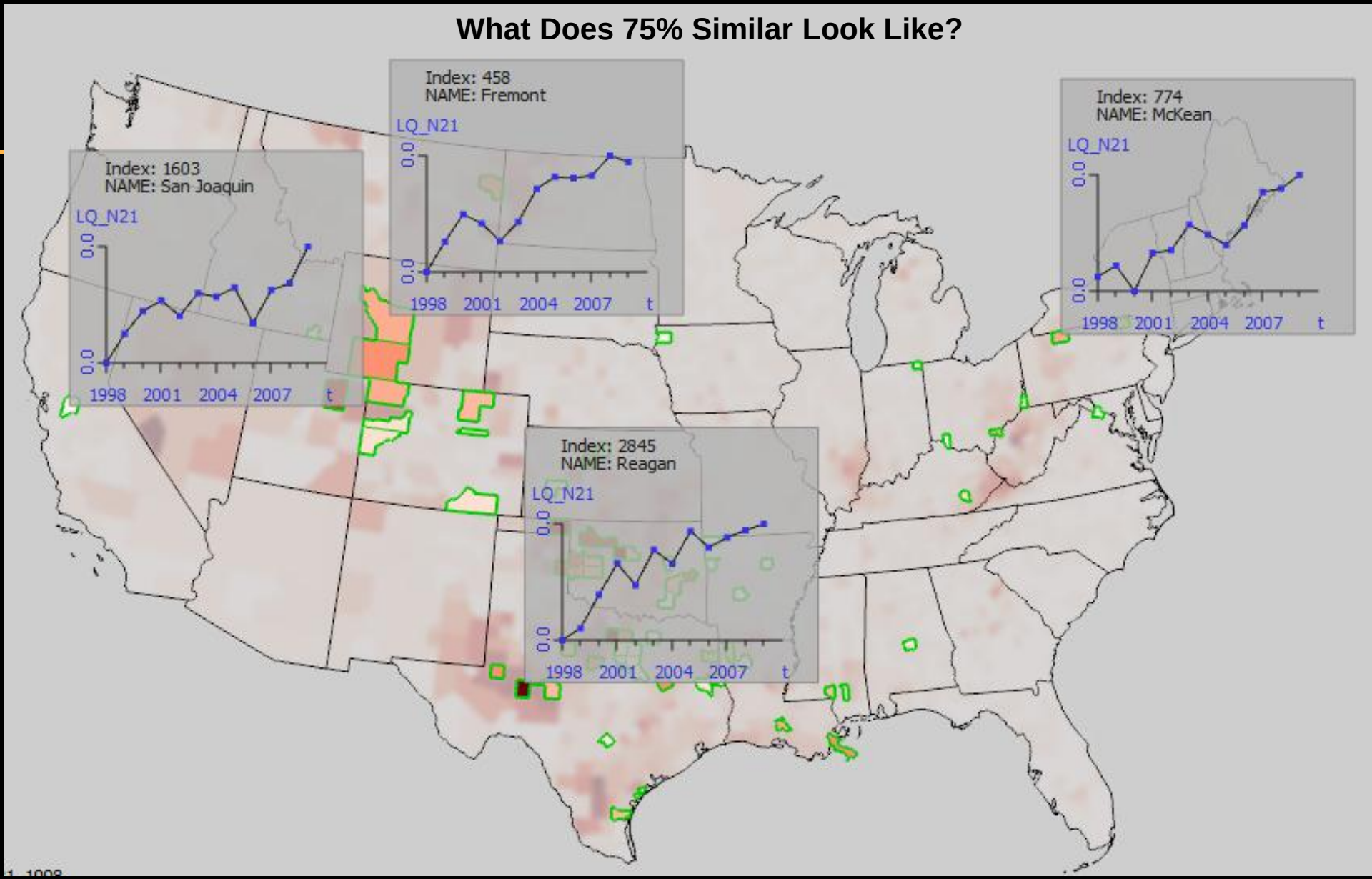
Regions with trajectories in Mining (N21) similar to Reagan Co., TX (75% Similarity Score)



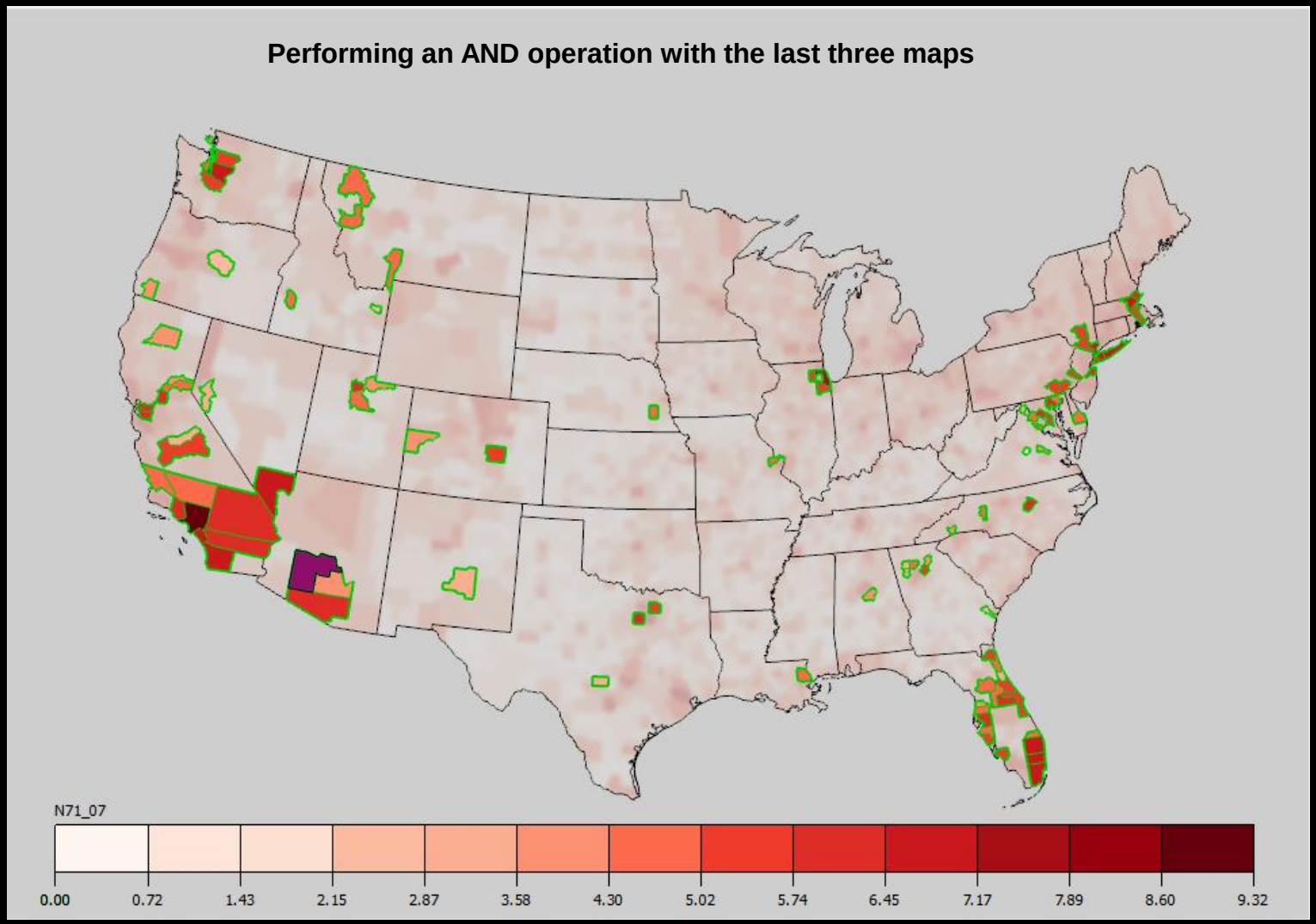
Regions with trajectories in Mining (N21) similar to Reagan Co., TX
(75% Similarity Score)



What Does 75% Similar Look Like?



Applied Analytical Brushing:
A Case Study in Industrial Diversity

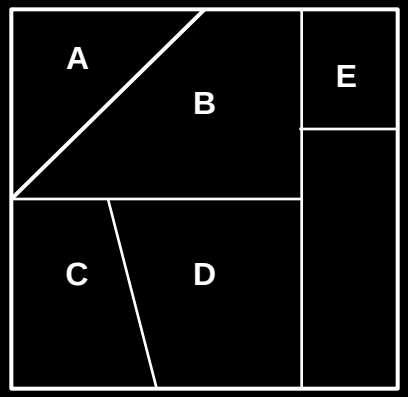


Temporally Coherent Spatial Clusters

What is the best way to cluster spatiotemporal data?

- ***Allow time to be a vector in the cluster algorithm***
  - Now spatial units can only belong to one cluster
- ***Cluster data per time step and then link clusters across time step by space***
  - Now spatial units can belong to multiple clusters, but need to maintain coherence between cluster labels at each time step

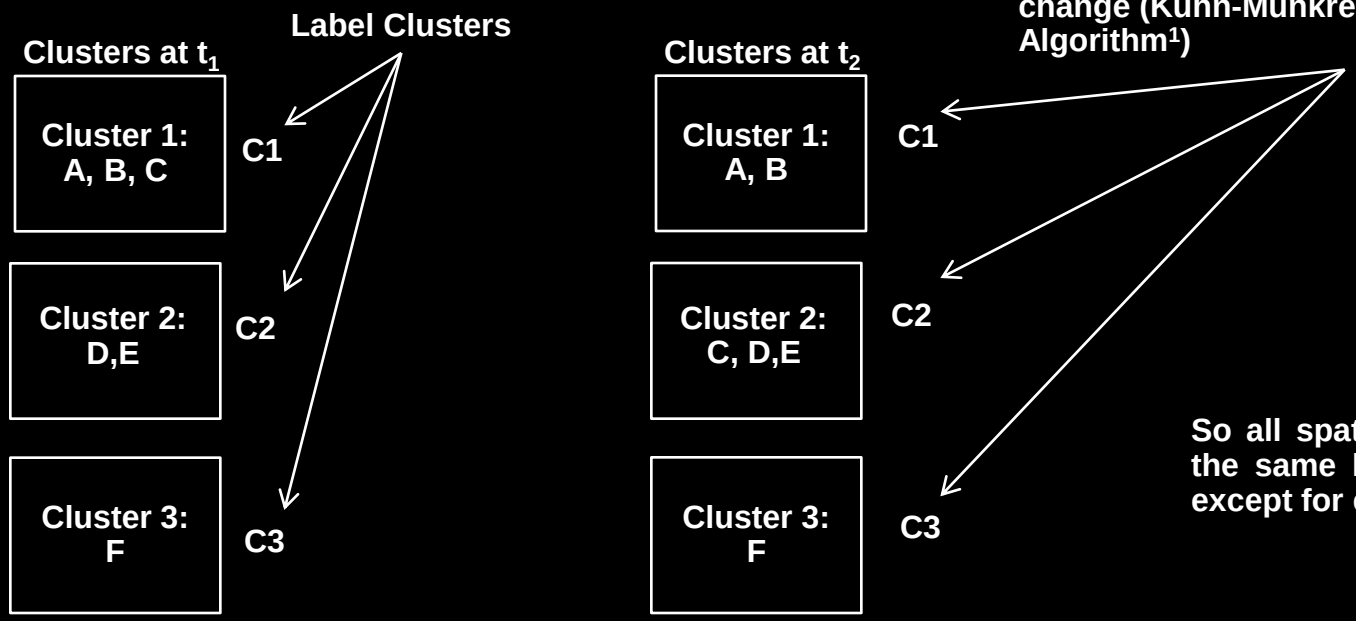
Temporally Coherent Spatial Clusters



Assume that we have some spatial arrangement of data with multivariate measures per time step

For each time step, we perform k-means clustering

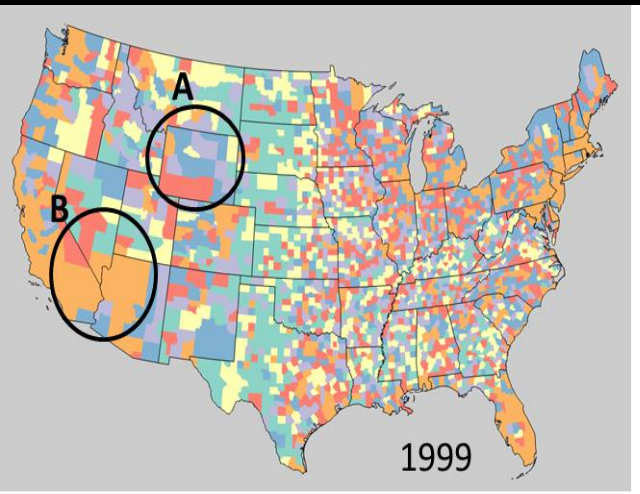
Label Clusters, but use information from previous time step, minimize the number of elements for which the cluster label will change (Kuhn-Munkres (KM) Algorithm[1])

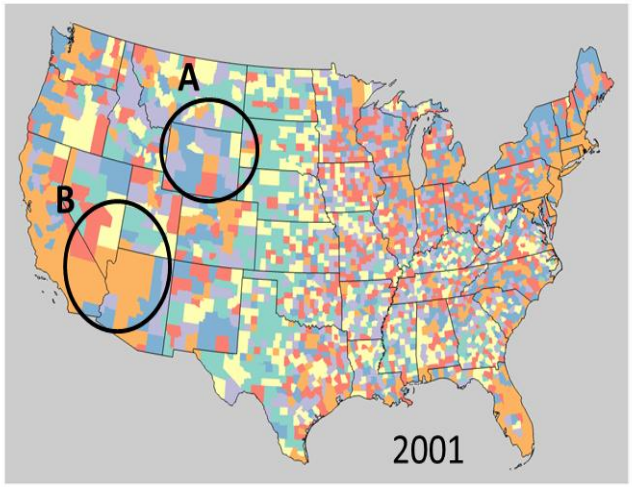


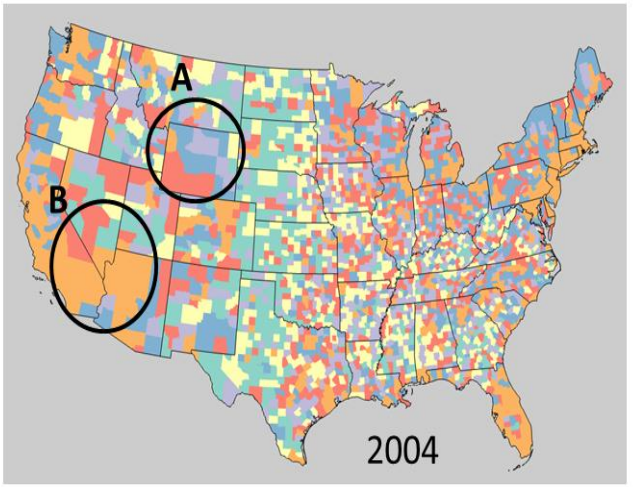
So all spatial units maintain the same label from t_1 to t_2 except for element C

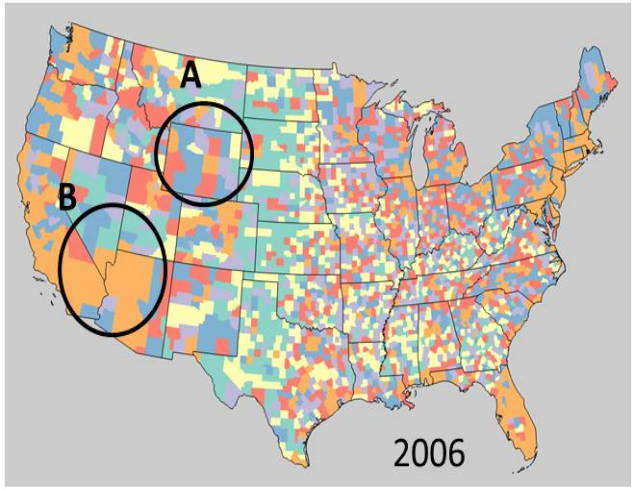
1 - J. Munkres. Algorithms for the assignment and transportation problems, Journal of the Society for Industrial & Applied Mathematics, 5(1):32–38, 1957.

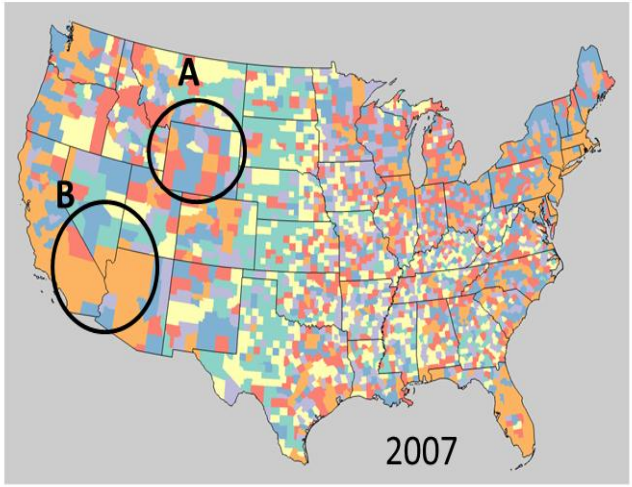
Temporally Coherent Spatial Clusters

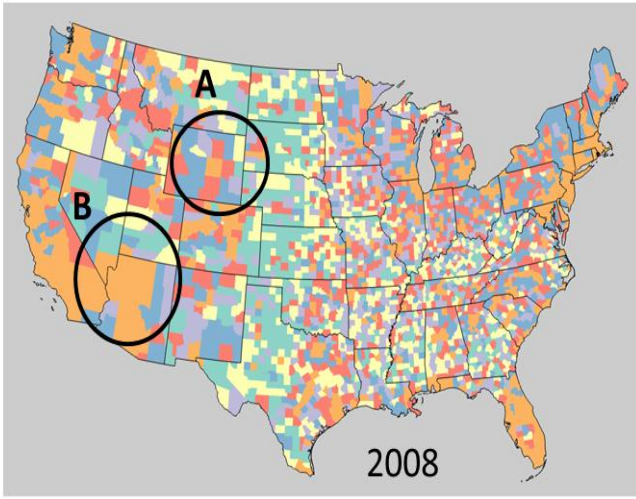








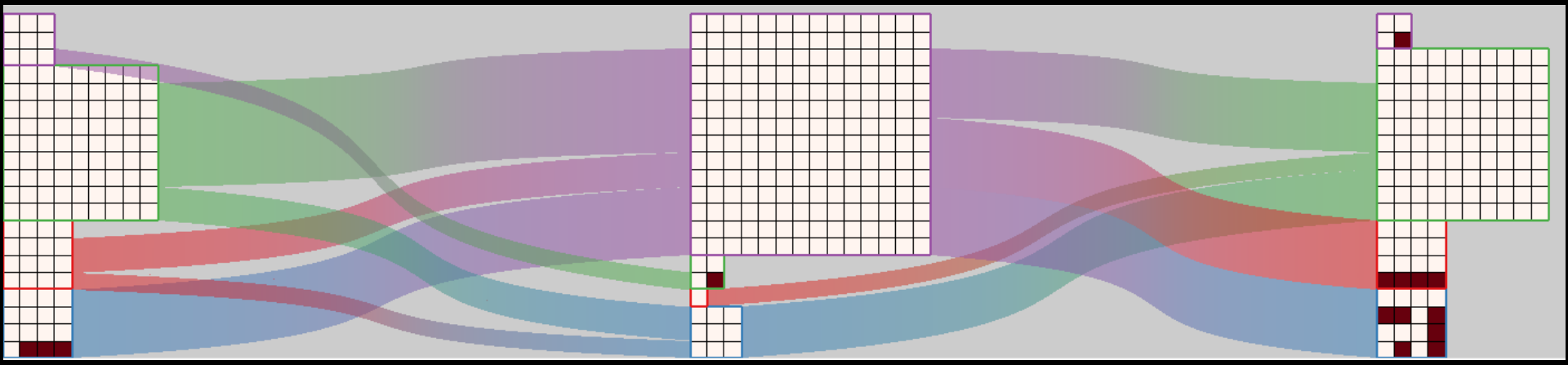




Temporally Coherent Spatial Clusters

What if we ignore the space and instead visualize how the element cluster labels change?

- What time steps should we present?
- Can we globally minimize the label changes?
- Measures of stability based on label changes?



Challenges

Scalability

- ***Unfortunately, as dimensionality increases, performance of similarity metrics decreases[1]***
  - Is the use of fractional distance metrics intuitive for users?
  - Should dimensional reduction/clustering be done spatially, temporally or spatiotemporally?

How do we move from hypothesis generation to hypothesis testing?

- ***Traditional techniques such as multiple regression do not scale well to high-dimensions***

1 – Charu C. Aggarwal, Alexander Hinneburg, and Daniel A. Keim. 2001. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. In *Proceedings of the 8th International Conference on Database Theory* (ICDT '01), Jan Van den Bussche and Victor Vianu (Eds.). Springer-Verlag, London, UK, 420-434.

Facilitating Spatial Analysis: Multivariate Representations

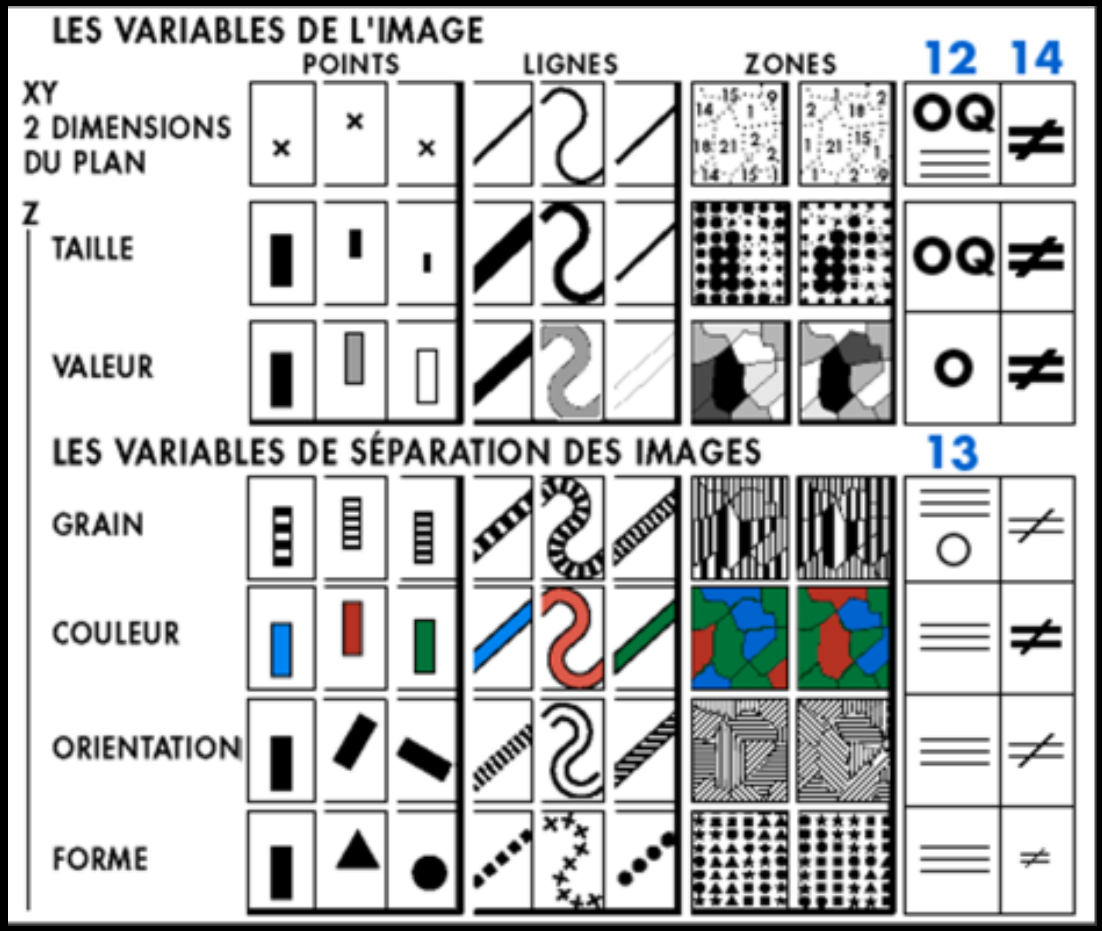
Visualization is concerned primarily with a mapping to visual form

[x,y]

- ~~Position~~

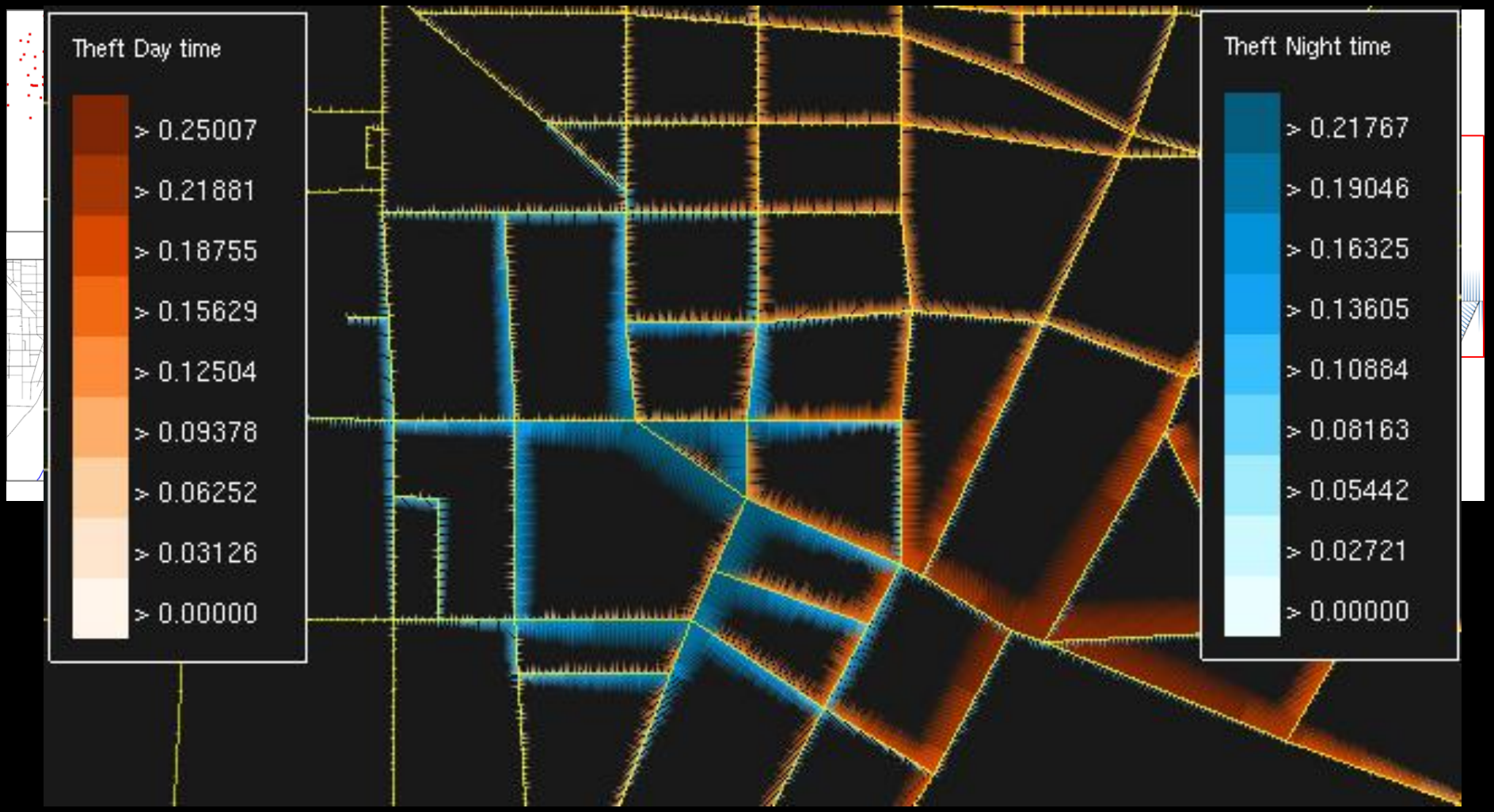
[z]

- ~~Size (Taille)~~
- Value (Valeur)
- Color (Couleur)
- Texture (Grain)
- ~~Orientation~~
- ~~Shape (Forme)~~



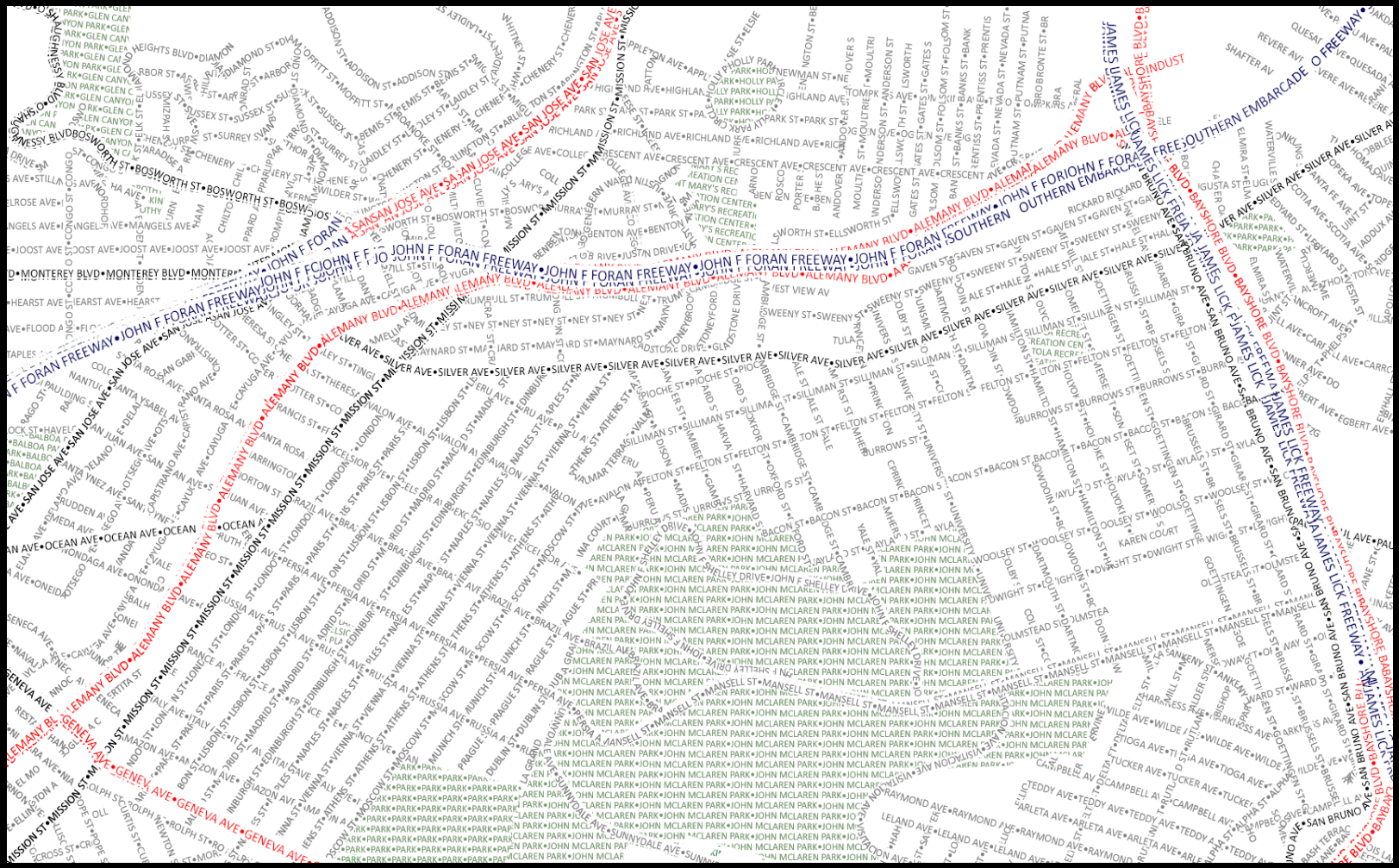
J Bertin (1967), *The Semiology of Graphics*

Bristle Maps: A Multivariate Abstraction Technique for Geovisualization



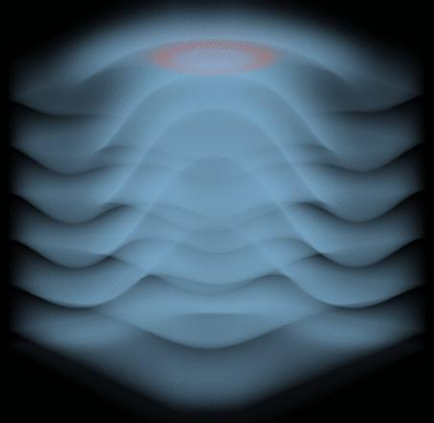
Kim, S., Maciejewski, R., Malik, A., Jang, Y., Ebert, D. S., Isenberg, T., "Bristle Maps: A Multivariate Abstraction Technique for Geovisualization," *IEEE Transactions on Visualization and Computer Graphics*, (To appear)

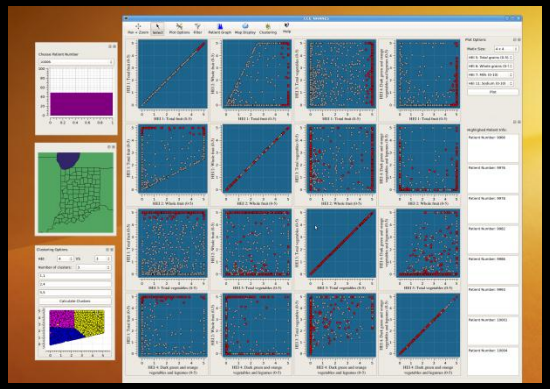
Typographic Maps: An Illustrative Abstraction Technique

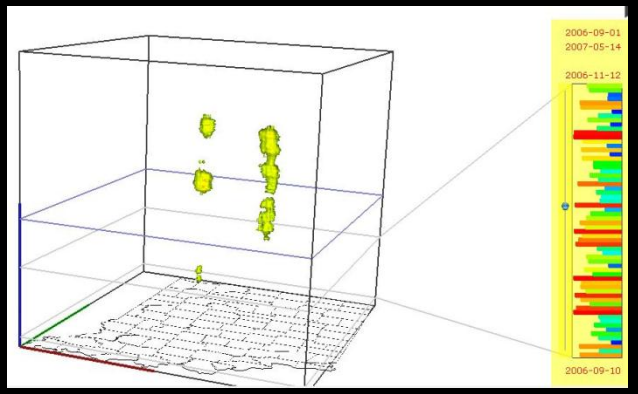


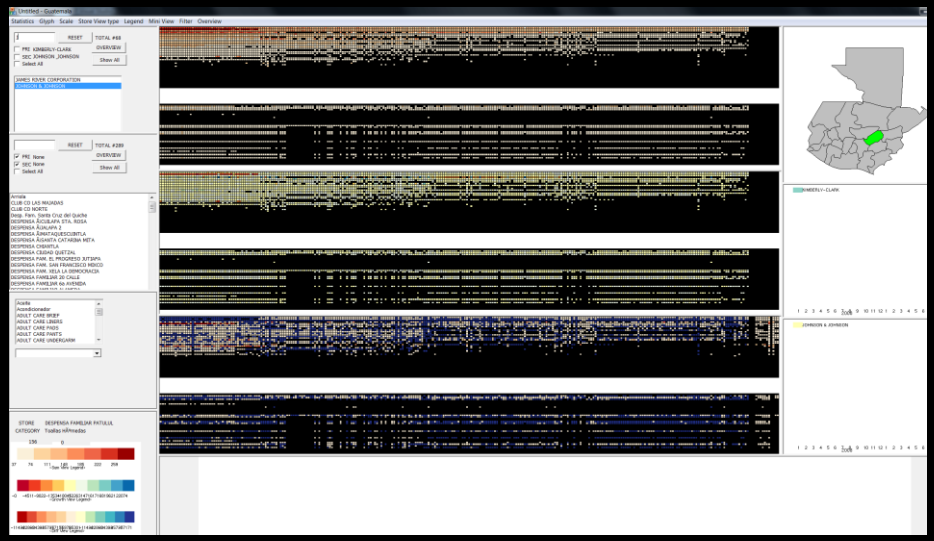
1 – Afzal, S., Maciejewski, R., Jang, Y., Elmqvist, N., Ebert. D. S., "Spatial Text Visualization Using Automatic Typographic Maps," *IEEE Transactions on Visualization and Computer Graphics*, 18(12): 2556-2564, 2012.

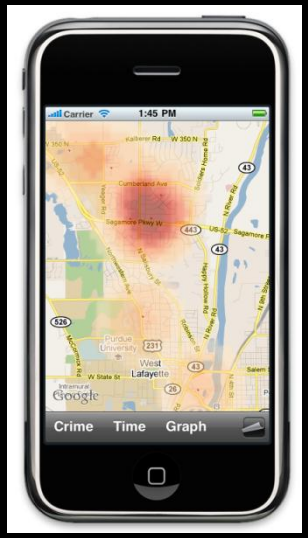
Questions?






2006-09-01
2007-05-14
2006-11-12
2006-09-10




Carrier
1:45 PM
Crime
Time
Graph



Box-Office VAST --- VADER

by Yafeng Lu, Feng Wang, Ross Maciejewski

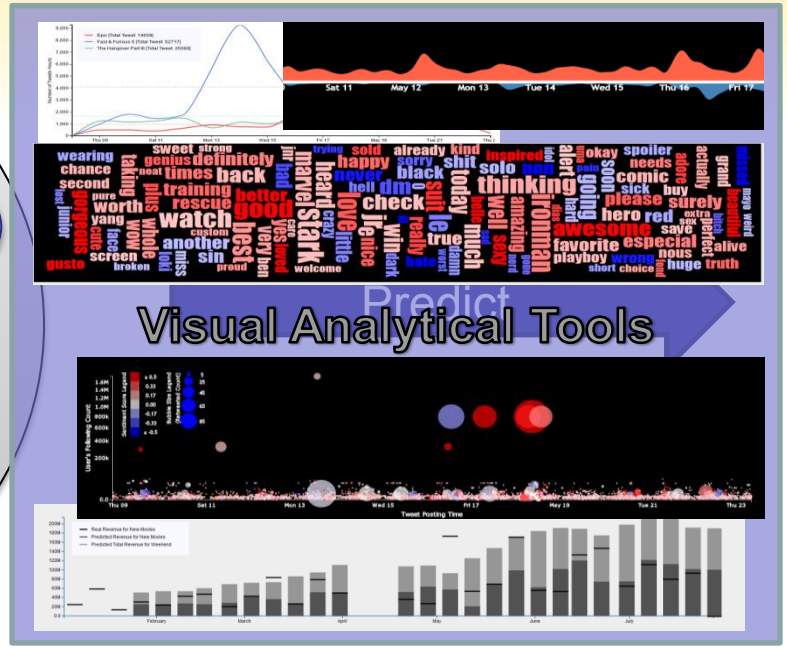
VAST Challenge Description

Twitter indices

Bitly Links

64 Movies
from Jan. to Jul.

IMDB

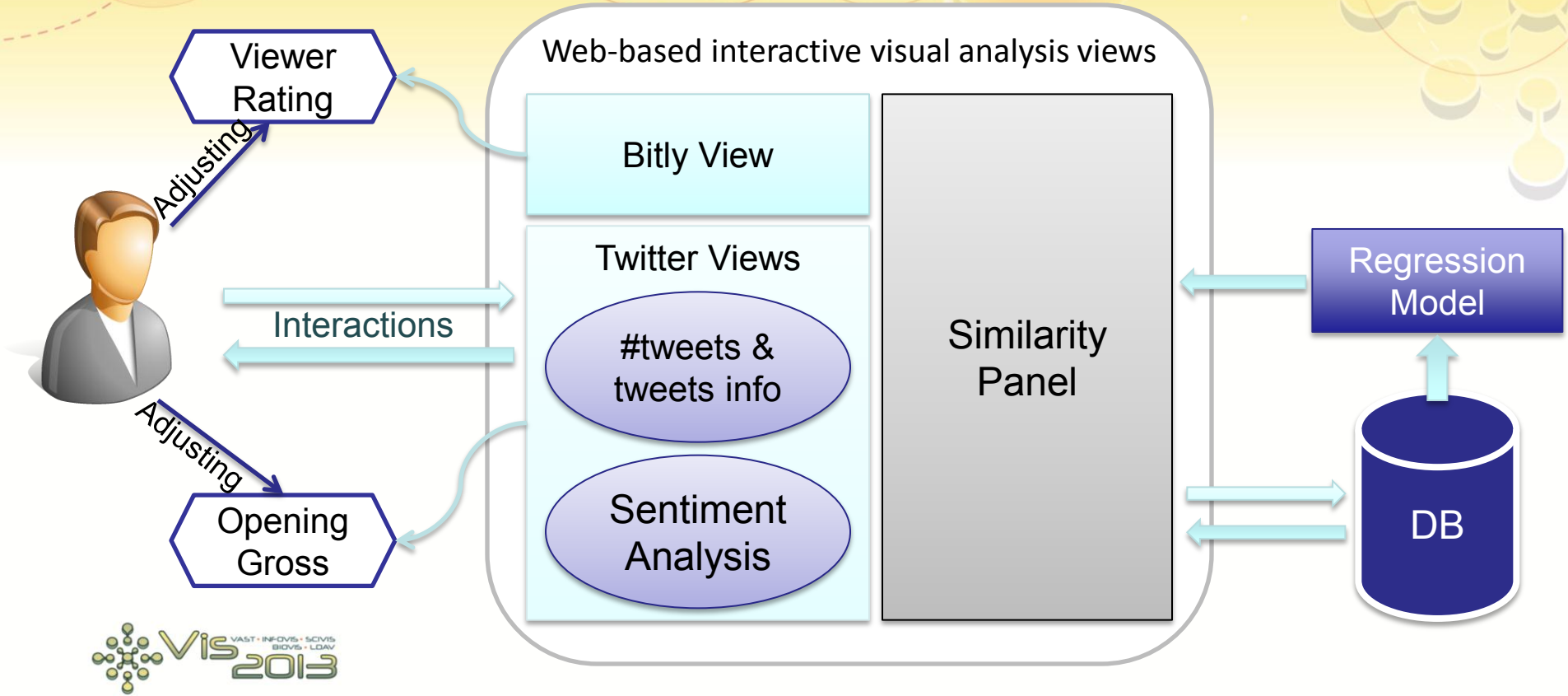


Opening Weekend Gross

Viewer Rating

VIS 2013

Overview of Our System

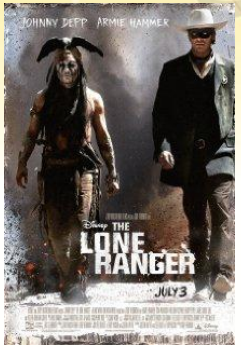


VIS 2013

DM2 and Lone Ranger OW Prediction

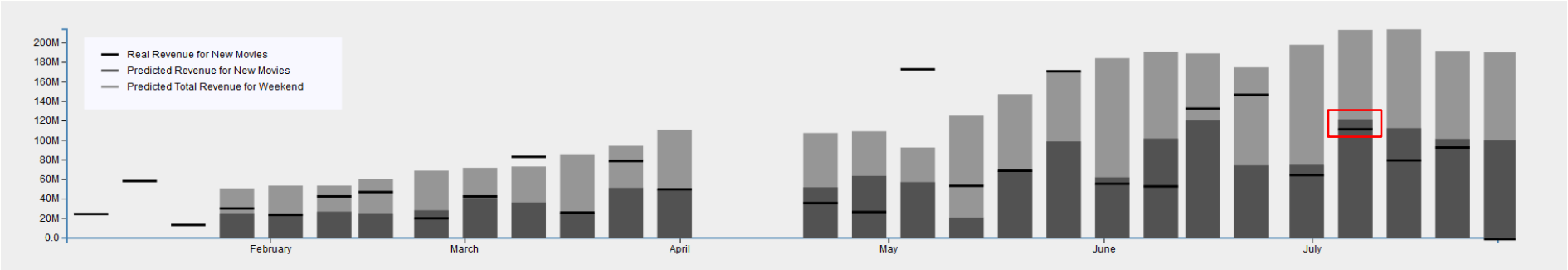
- Total weekend revenue
- Individual box-office take
  - Multiple linear regression model
  - Adjustment
    - Exclude non-related tweets
    - Correct miscalculated sentiment scores
  - Similarity
- Viewer rating prediction
- Prediction finalization





DM2 and Lone Ranger OW Prediction

- Two movies released on same weekend ---- total gross
  - 3-day weekend total gross prediction: $121M
  - 5-day long weekend: ~$175M



DM2 and Lone Ranger OW Prediction

- Individual movie OW box-office prediction ---- Multiple Linear Regression Model

Response 

Regressors 

Variable	Description
MPAA rating	G, PG, PG-13, R, NC-17
Genre	Animation, Action, Thriller, Comedy, etc.
OW	3-day Opening Weekend Gross
TBD	Average daily #tweets
Budget	Approximate movie budget from IMDB
TSS	Tweet sentiment score
MSS	Movie sentiment score
MSP	Movie Star Power

DM2 and Lone Ranger OW Prediction

- Individual movie OW box-office prediction ---- Multiple Linear Regression Model

$$OW = \beta_0 + \beta_1 TBD + \beta_2 Budget$$

$$\beta_0 \approx 4.9 \times 10^3$$
$$\beta_1 \approx 4462$$
$$\beta_2 \approx 0.23$$

$$adj\,R^2 \approx 0.60, \qquad P\,value < 0.05$$

- Long Ranger $85M +/- $13M
- DM2 $76M +/- $13M

From January to July, each week model parameters were updated with newly collected data.

VAST · INFOVIS · SCIVIS · BIOVIS · LDAV
VIS 2013

DM2 and Lone Ranger OW Prediction

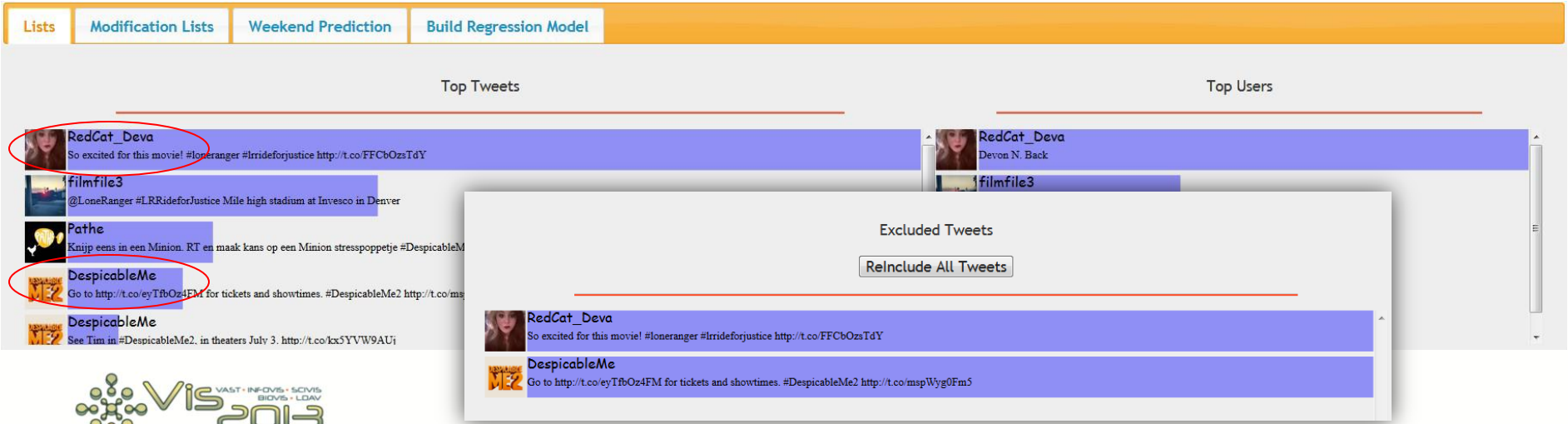
- Individual movie OW box-office prediction ---- Model Adjustment

$$adj\ R^2 \approx 0.60, \qquad P\ value < 0.05$$

  - Exclude non-related tweets
  - Correct miscalculated sentiment scores

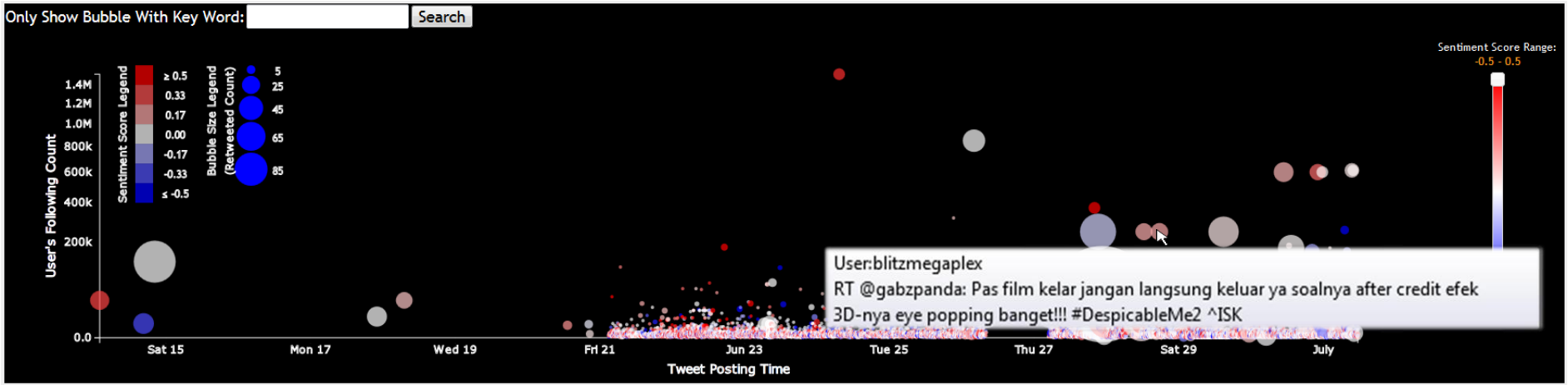
DM2 and Lone Ranger OW Prediction

- Individual movie OW box-office prediction ---- Model Adjustment
  - Exclude non-related tweets
    - From tweet/user list or From sentiment analysis view



DM2 and Lone Ranger OW Prediction

- Individual movie OW box-office prediction ---- Model Adjustment
  - Exclude non-related tweets
    - From tweet/user list or From sentiment analysis view

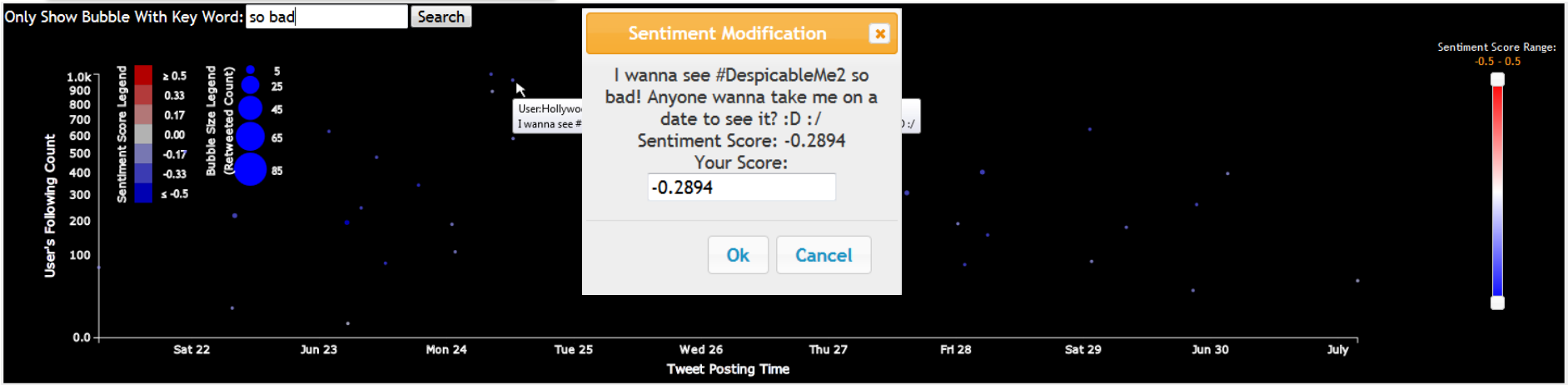


VAST · INFOVIS · SCIVIS
BIOVIS · LDAV
VIS 2013

DM2 and Lone Ranger OW Prediction

- Individual movie OW box-office prediction ---- Model Adjustment
  - Correct miscalculated sentiment scores

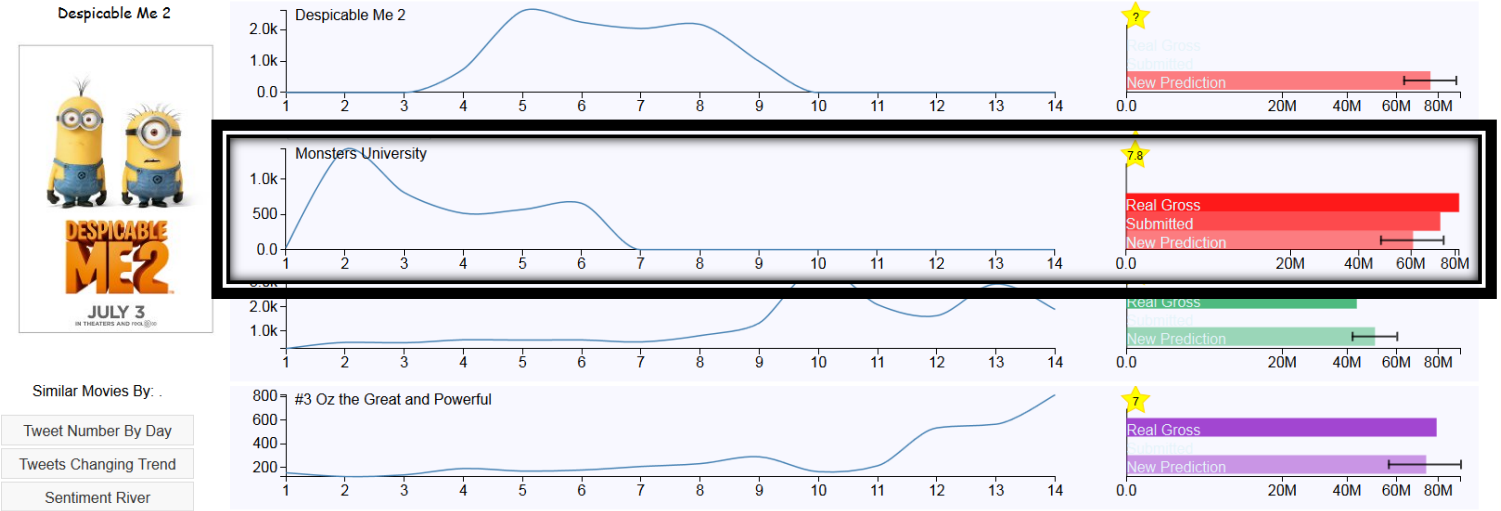
"I want to see this so bad." ⇨ TSS > 0



DM2 and Lone Ranger OW Prediction

- Individual movie OW box-office prediction ---- Comparison with similar movies

3-day weekend: ~$82M



DM2 and Lone Ranger OW Prediction

- Individual movie OW box-office prediction ---- Comparison with similar movies



Compared with OZ and MU:

3-day weekend ~$82M

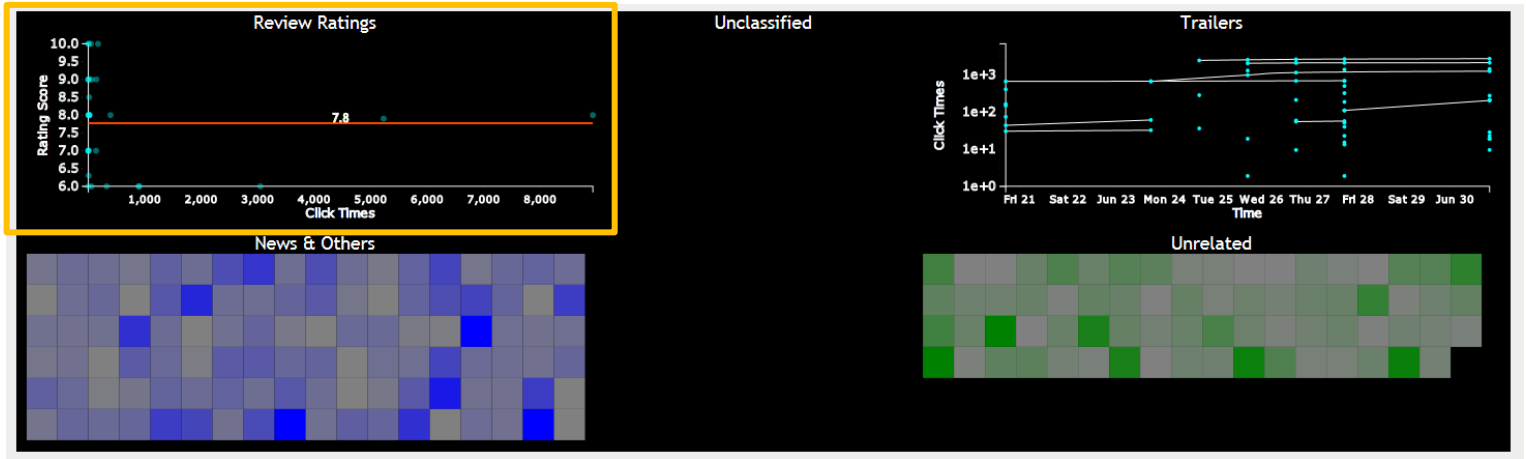


Compared with WWZ:

3-day weekend < $66M

DM2 and Lone Ranger OW Prediction

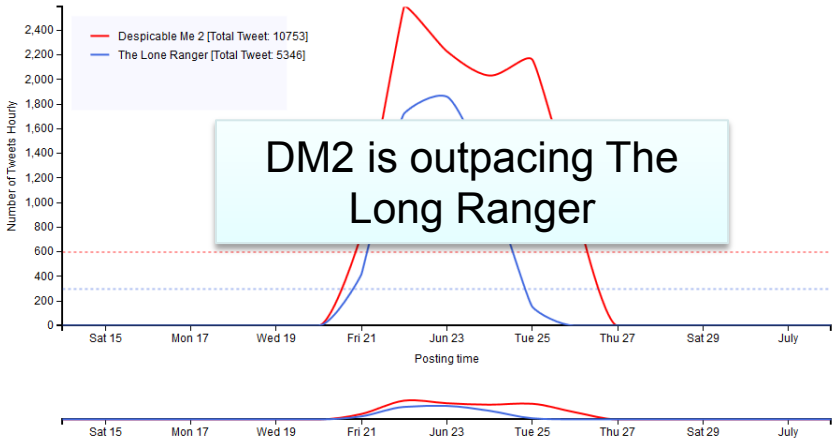
- Viewer Rating Prediction
  - DM2: 7.8/10.0
  - Long Ranger: 6.4/10.0



$$Viewer\ Rating = \frac{1}{k}\sum_{i=1}^{k} score(i)$$

DM2 and Lone Ranger OW Prediction

- Finalize box-office prediction
  - Overall view of two movies by line chart







3-day OW	Viewer rating	5-day OW
~$82M	7.8/10.0	?
~$39M	6.4/10.0	
Total ~$121M		Total ~$175M

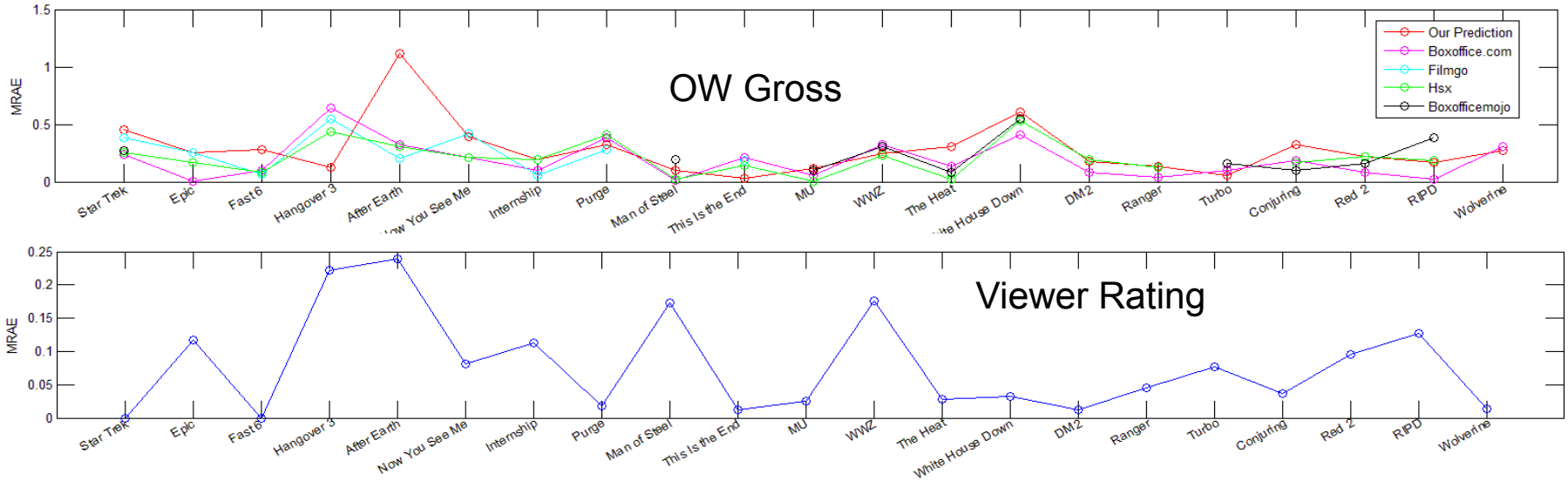
DM2 and Lone Ranger OW Prediction

- Predictions vs. Real OW

	Opening weekend gross		Viewer Rating	
	prediction	real	prediction	real
DM2 (3-day)	$82M	$83.5M	7.8	7.9
DM2 (5-day)	$117M	$143M		
The Lone Ranger (3-day)	$39M	$29M	6.4	6.8
The Lone Ranger (5-day)	$55M	$48.7M		

Performance Evaluation

- Mean Relative Absolute Error (MRAE)
  - Compared with 4 professional website predictions



Performance Evaluation

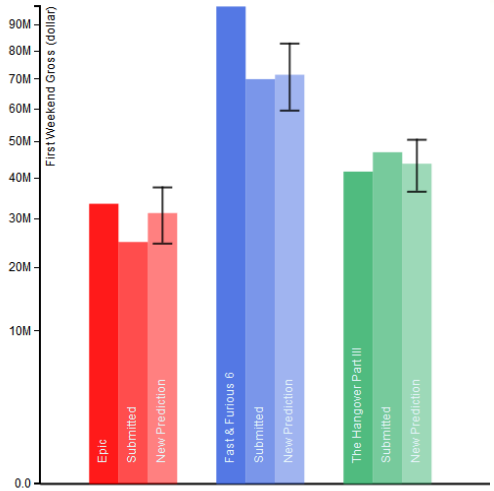
Prediction Source	Average MRAE
Our Submitted OW Prediction	0.281 With interaction
Regression Model OW Prediction	0.447 No interaction
boxoffice OW Prediction	0.183
filmgo OW Prediction	0.265
hsx OW Prediction	0.205
boxofficemojo OW Prediction	0.230
Our Submitted Viewer Rating Prediction	0.078

Questions?

Thank You!

Model for Gross Prediction

- Multiple Linear Regression Modeling



Variable	Description
MPAA rating	G, PG, PG-13, R, NC-17
Genre	Animation, Action, Thriller, Comedy, etc.
OW	3-day Opening Weekend Gross
TBD	Average daily #tweets
Budget	Approximate movie budget from IMDB
TSS	Tweet sentiment score
MSS	Movie sentiment score
MSP	Movie Star Power

VAST · INFOVIS · SCIVIS
BIOVIS · LDAV
VIS 2013

Model for Gross Prediction

- Multiple Linear Regression Modeling

Response 

Variable	Description
MPAA rating	G, PG, PG-13, R, NC-17
Genre	Animation, Action, Thriller, Comedy, etc.
OW	**3-day Opening Weekend Gross**
TBD	Average daily #tweets
Budget	Approximate movie budget from IMDB
TSS	Tweet sentiment score
MSS	Movie sentiment score
MSP	Movie Star Power

VIS 2013 VAST · INFOVIS · SCIVIS BIOVIS · LDAV

Model for Gross Prediction

- Multiple Linear Regression Modeling

Response
Regressor_1
Regressor_2

Variable	Description
MPAA rating	G, PG, PG-13, R, NC-17
Genre	Animation, Action, Thriller, Comedy, etc.
OW	**3-day Opening Weekend Gross**
TBD	Average daily #tweets
Budget	Approximate movie budget from IMDB
TSS	Tweet sentiment score
MSS	Movie sentiment score
MSP	Movie Star Power

Model for Gross Prediction

- Weekend Gross Prediction
  - We hypothesized that movies would run for three weeks following their opening weekend, and each weekend their box office take would reduced by 50%.

$$Weekend\ Gross(t) = \sum_{\forall_i} OW_i(t) + \sum_{\forall_i, j=1}^{j=3} 0.5^j OW_i(t-j)$$

t is the current weekend, i is the index to a movie that exists at time t.

  - Three-day moving average

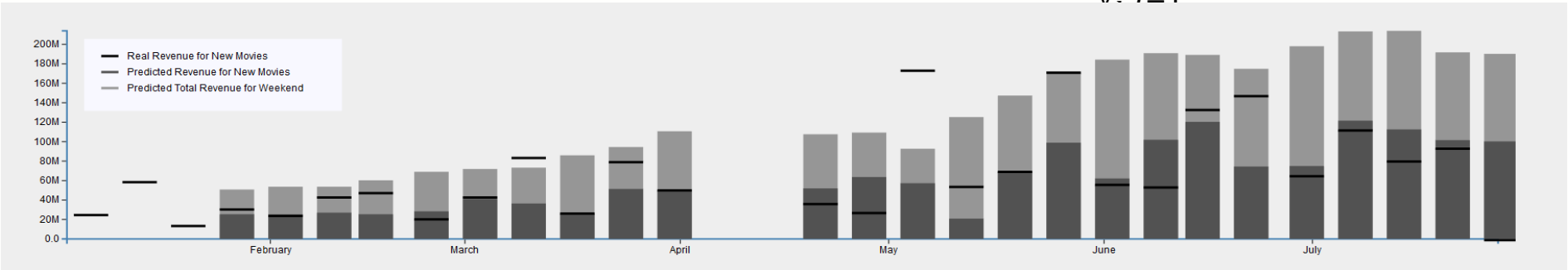
$$M\ Weekend\ Gross(t+1) = \frac{1}{3}\sum_{j=0}^{j=2} Weekend\ Gross(t-j)$$

VAST · INFOVIS · SCIVIS
BIOVIS · LDAV
VIS 2013

Model for Gross Prediction

- Weekend Gross Prediction
  - Available revenue for new movies

$$New\ Movie\ Gross(t+1) = M\ Weekend\ Gross(t+1) - \sum_{\forall i,\ j=1}^{j=3} 0.5^j OW_i(t+1-j)$$



Model for Viewer Rating Prediction



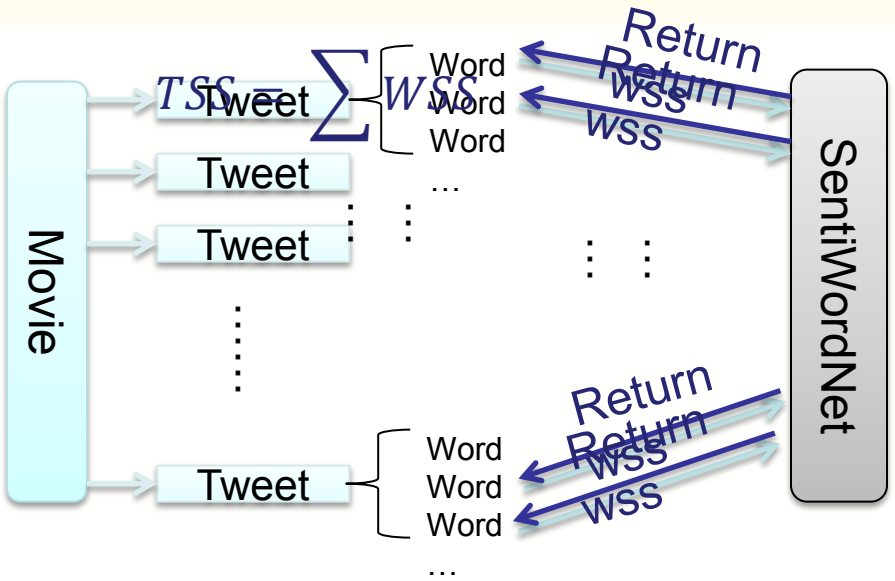
$$Viewer\ Rating = \frac{1}{k}\sum_{i=1}^{k} score(i)$$

VAST · INFOVIS · SCIVIS
BIOVIS · LDAV
VIS 2013

Tweet Sentiment Mining

- Process of calculating MSS

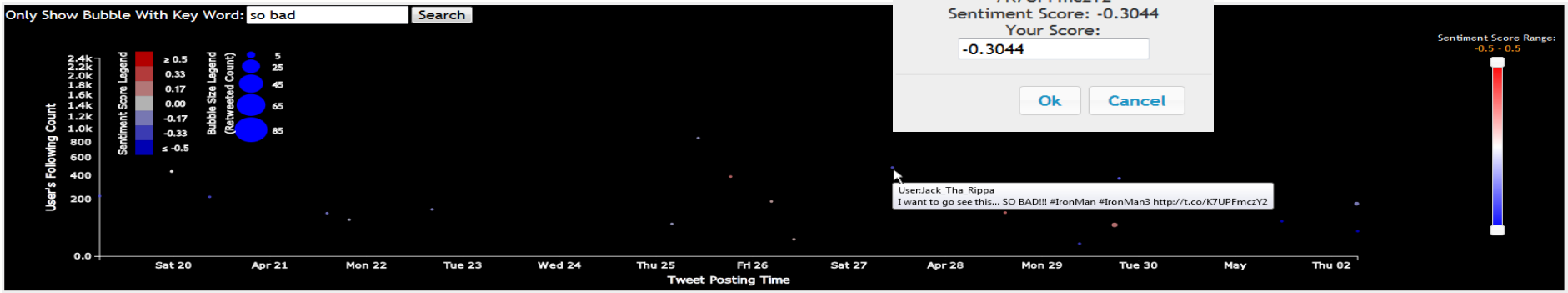
$$MSS = \frac{\sum_{\forall tweet, TSS>0} TSS}{\sum_{\forall tweet} |TSS|}$$



Adjustment with Visualization Tools

- Tweet Bubble Chart

"I want to see this so bad." ⇨ TSS > 0

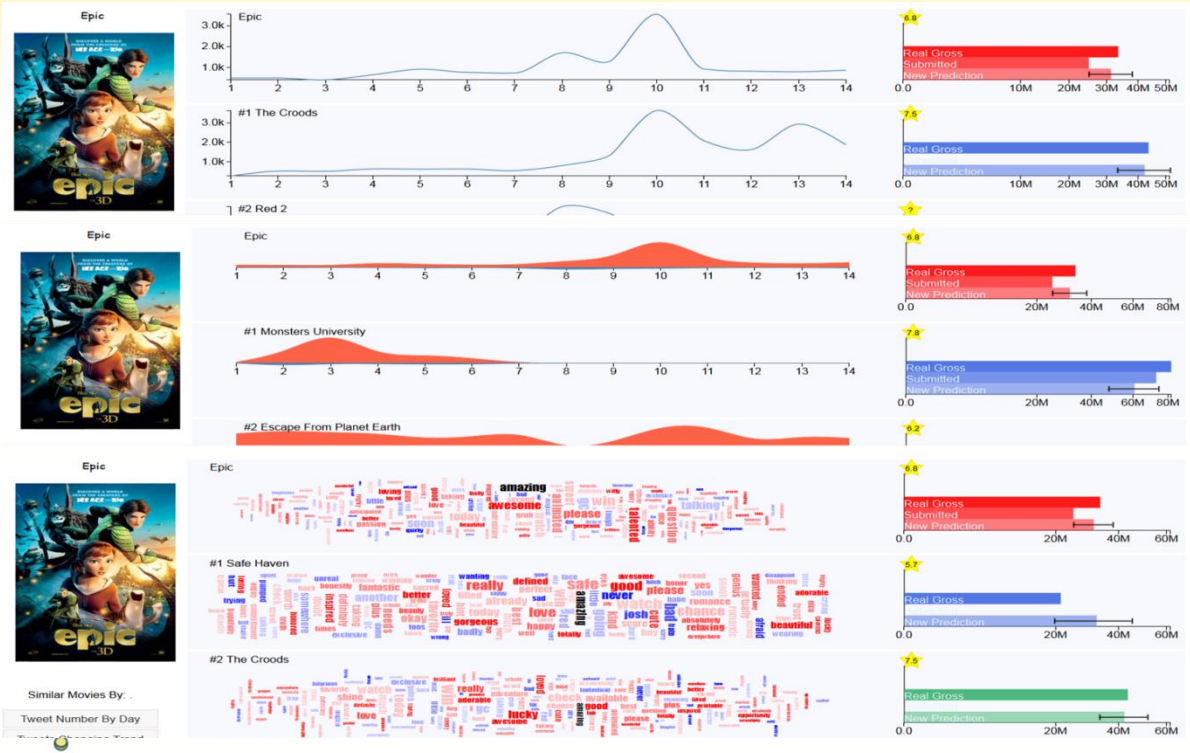


Adjustment with Visualization Tools

- Sentiment Wordle & Theme River



Adjustment with Visualization Tools

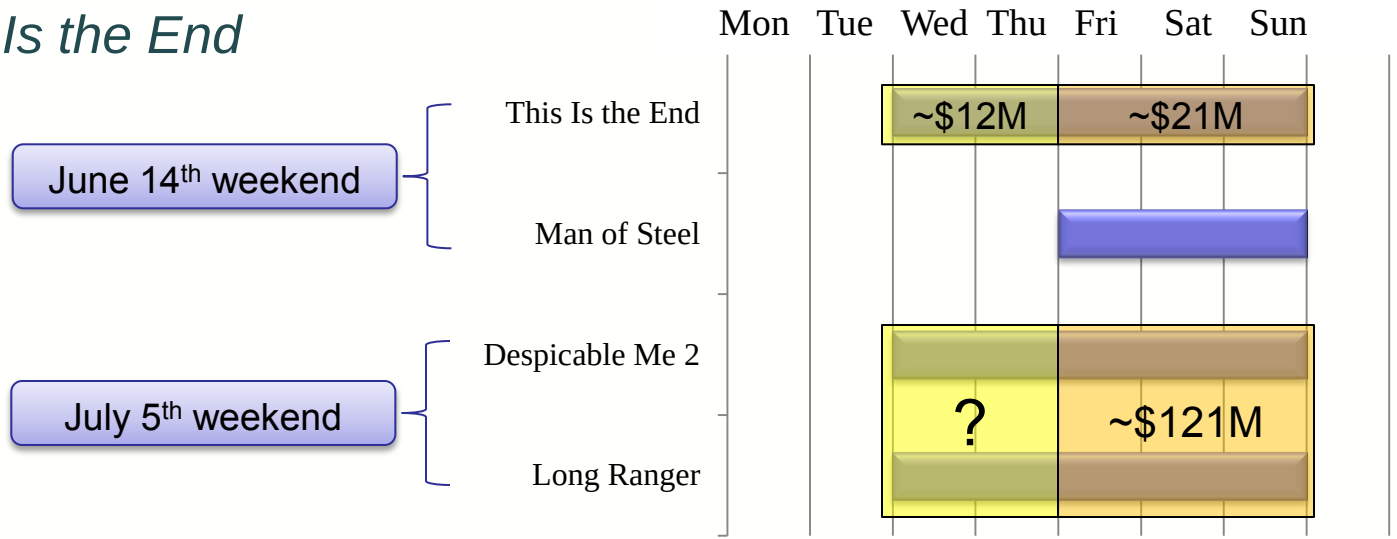


- Similarity

Tweet Number By Day
Tweets Changing Trend
Sentiment River
Movie Sentiment Score
MPAA
Genre
Movie Star Power
Sentiment Wordle
Predicted Gross
Make Your Own Similarity

DM2 and Lone Ranger OW Prediction

- Two movies released on same weekend ---- total gross
  - 5-day long weekend: ~$175M
  - *This Is the End*



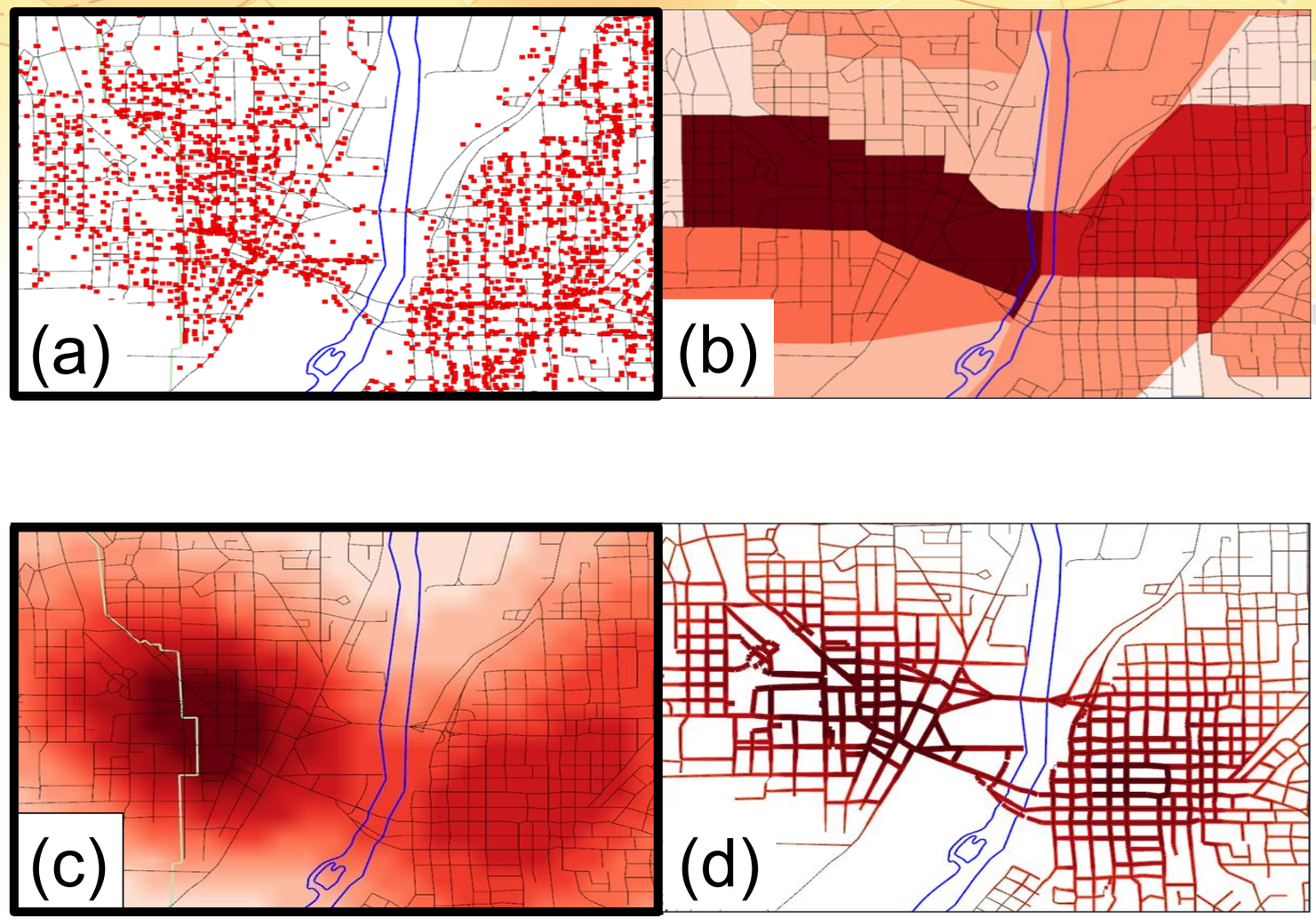


Bristle Maps: A Multivariate Abstraction Technique for Geovisualization

SungYe Kim, Ross Maciejewski Abish Malik, Yun Jang,
David S. Ebert, Tobias Isenberg

Background

- Geovisualization
  a) Point map:
     Cluttering
  b) Choropleth map:
     Modifiable areal unit problem
  c) Heat map:
     One variable per map
  d) Line map:
     Two variables per map
     (i.e., color, thickness)



VIS 2013 VAST • INFOVIS • SCIVIS BIOVIS • LDAV

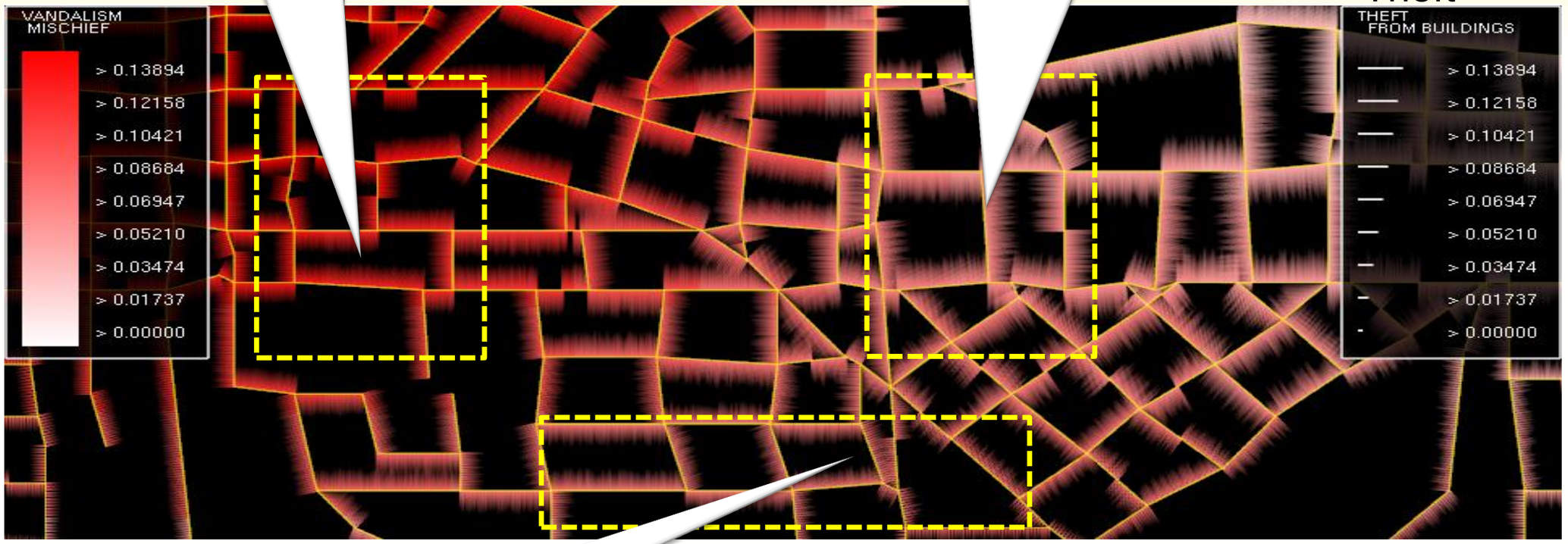
Bristle Map: A multi-attribute visualization

-High vandalism
-Low theft
-Low theft variance

-Low vandalism
-High theft
-High theft variance

Vandalism

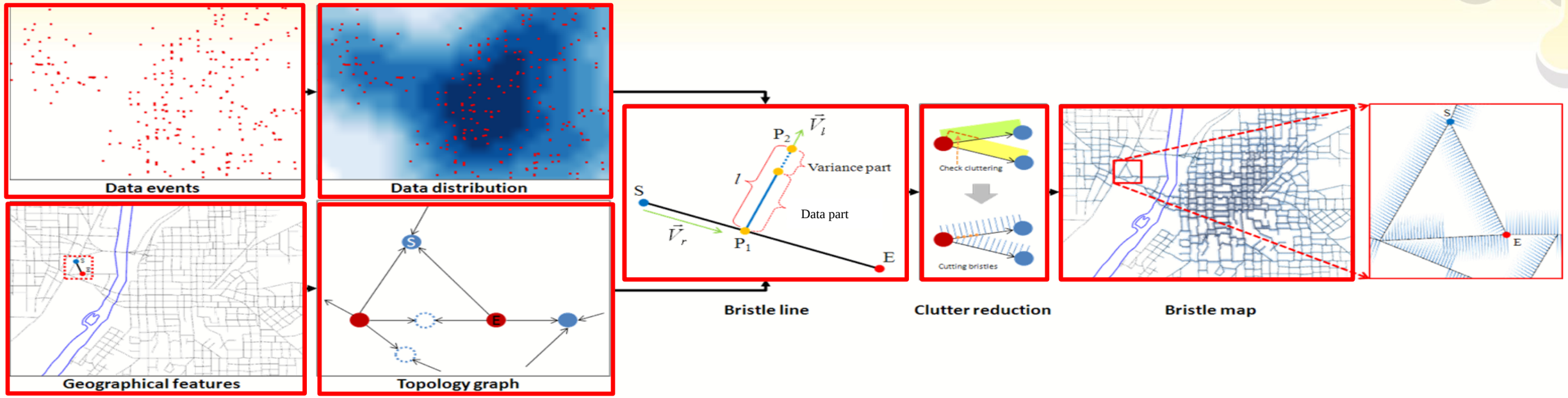
Theft



-Low vandalism
-Low theft
-High theft variance

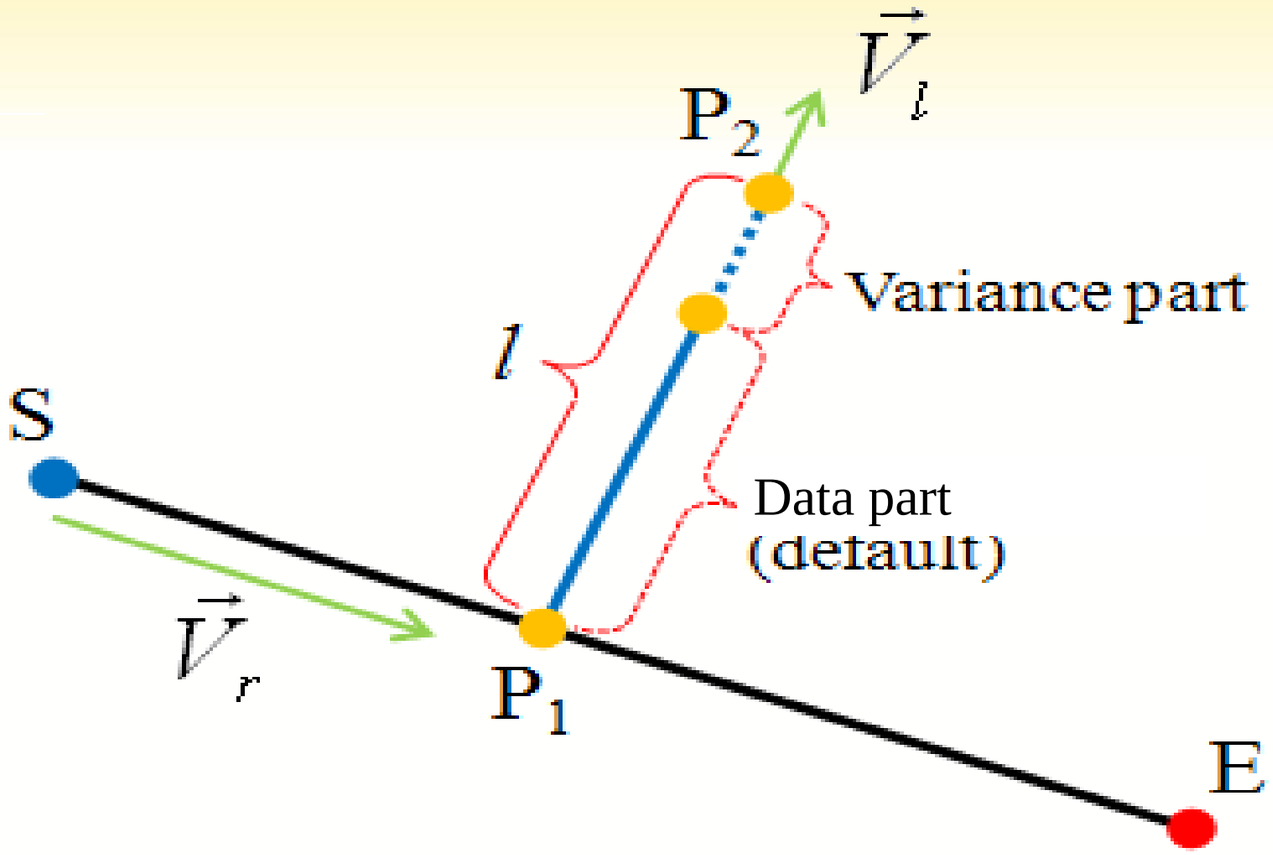
VAST • INFOVIS • SCIVIS
BIOVIS • LDAV
VIS 2013

System Overview



VAST • INFOVIS • SCIVIS
BIOVIS • LDAV
VIS 2013

Bristle Map Generation

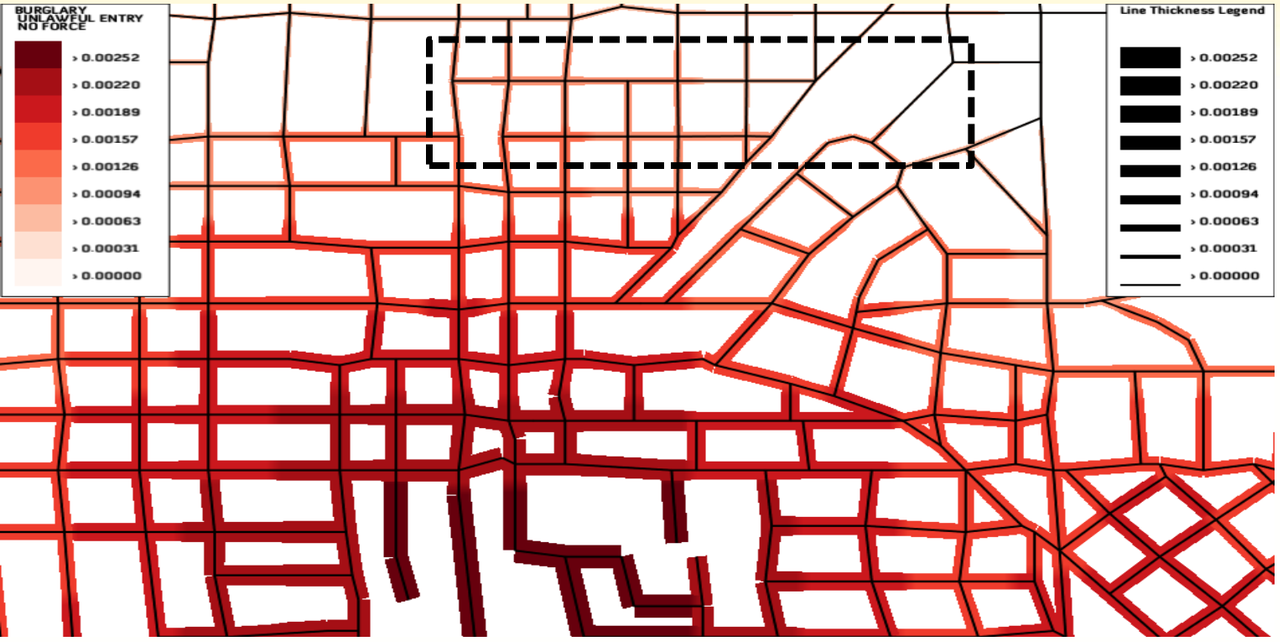


VAST • INFOVIS • SCIVIS
BIOVIS • LDAV
VIS 2013

Encoding Schemes:

Value of Bristle Density

- Line map vs. Bristle map



Line map (color, thickness)



Bristle map (color, density)

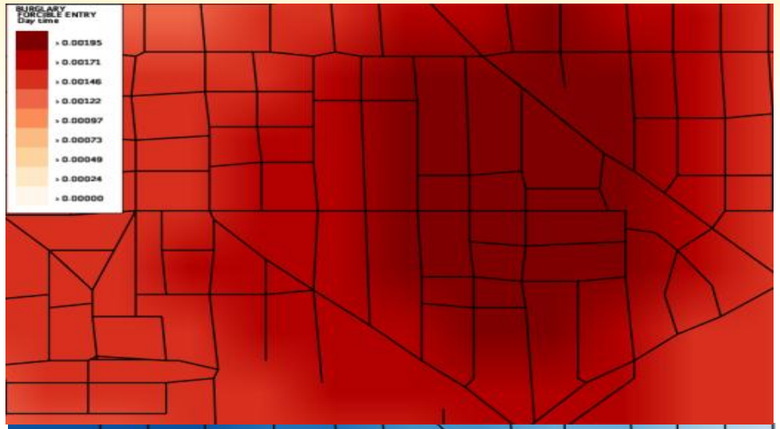
$$\frac{15 \text{ bristle lines}}{\text{unit geo distance}}$$

cf. unit geo distance = 0.0009°

VAST • INFOVIS • SCIVIS
BIOVIS • LDAV
VIS 2013

Encoding Schemes:

Orientation for Bivariate Comparison

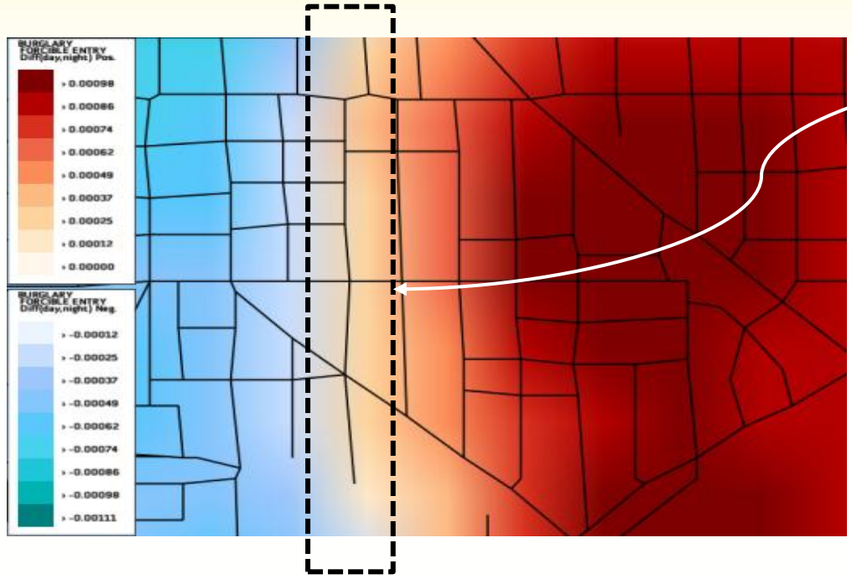


Day



Night

Day ≈ Night



Day - Night

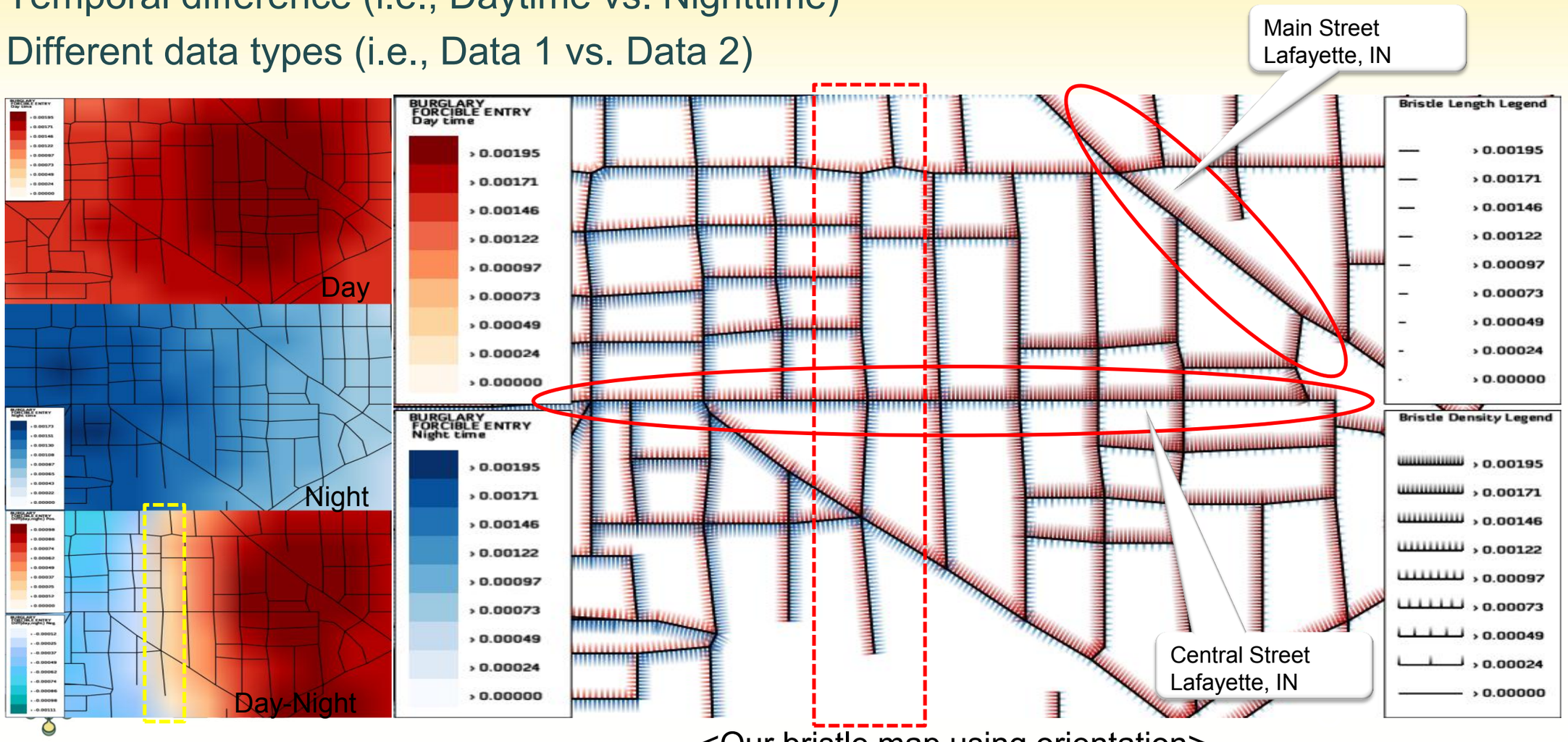
Day < Night

Day > Night

Encoding Schemes:

Orientation for Bivariate Comparison

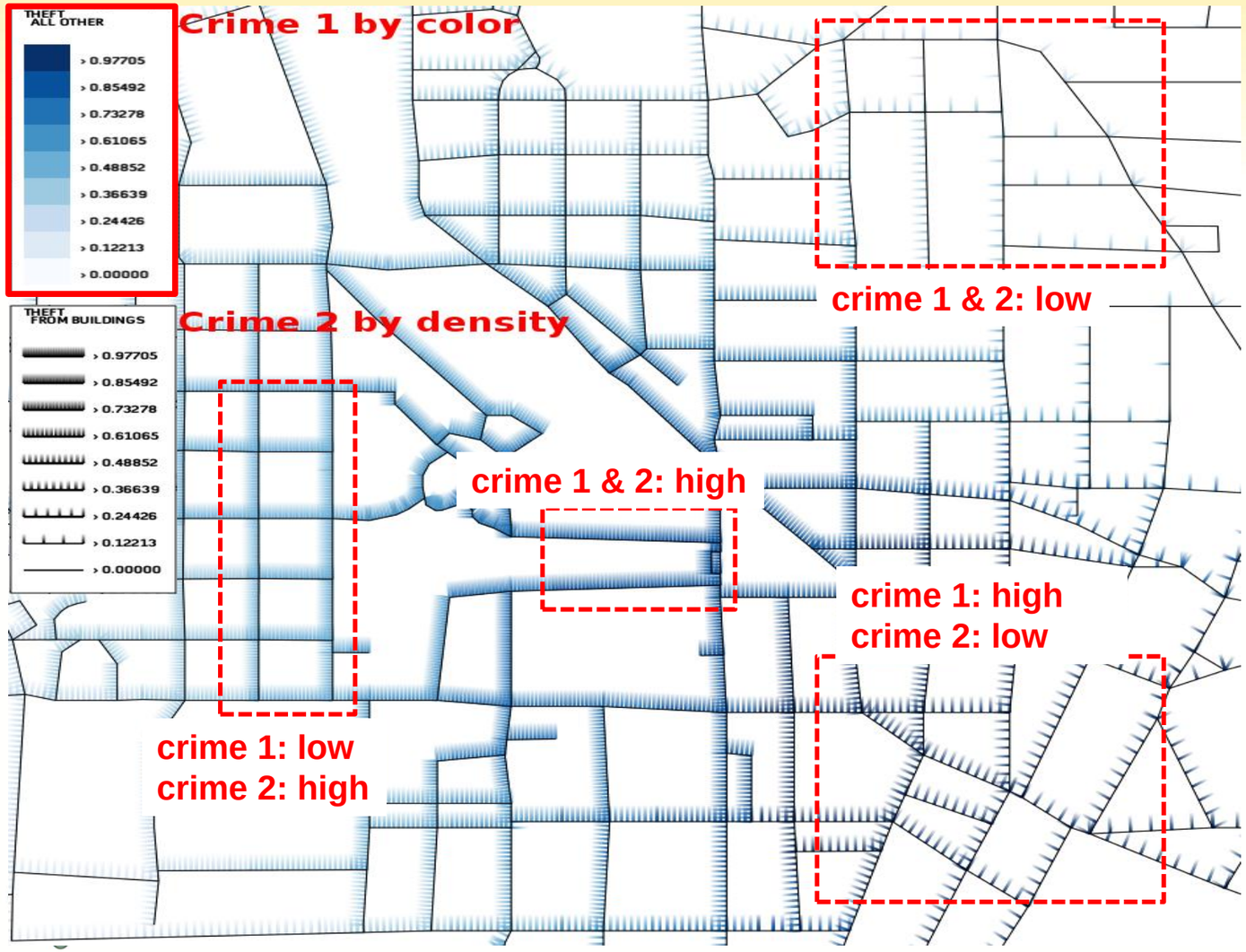
- Temporal difference (i.e., Daytime vs. Nighttime)
- Different data types (i.e., Data 1 vs. Data 2)



<Our bristle map using orientation>

Encoding Schemes:

Multivariate Encoding

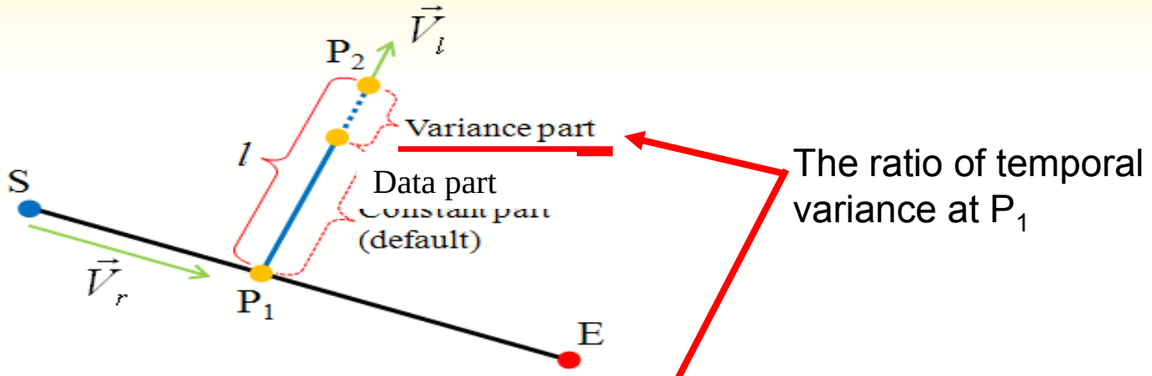


- **Separating** color, length, and density
  - Data 1: Color
  - Data 2: Density
  - Constant length
  - One orientation

Encoding Schemes:

Encoding Data Variance

- The ratio of temporal variance

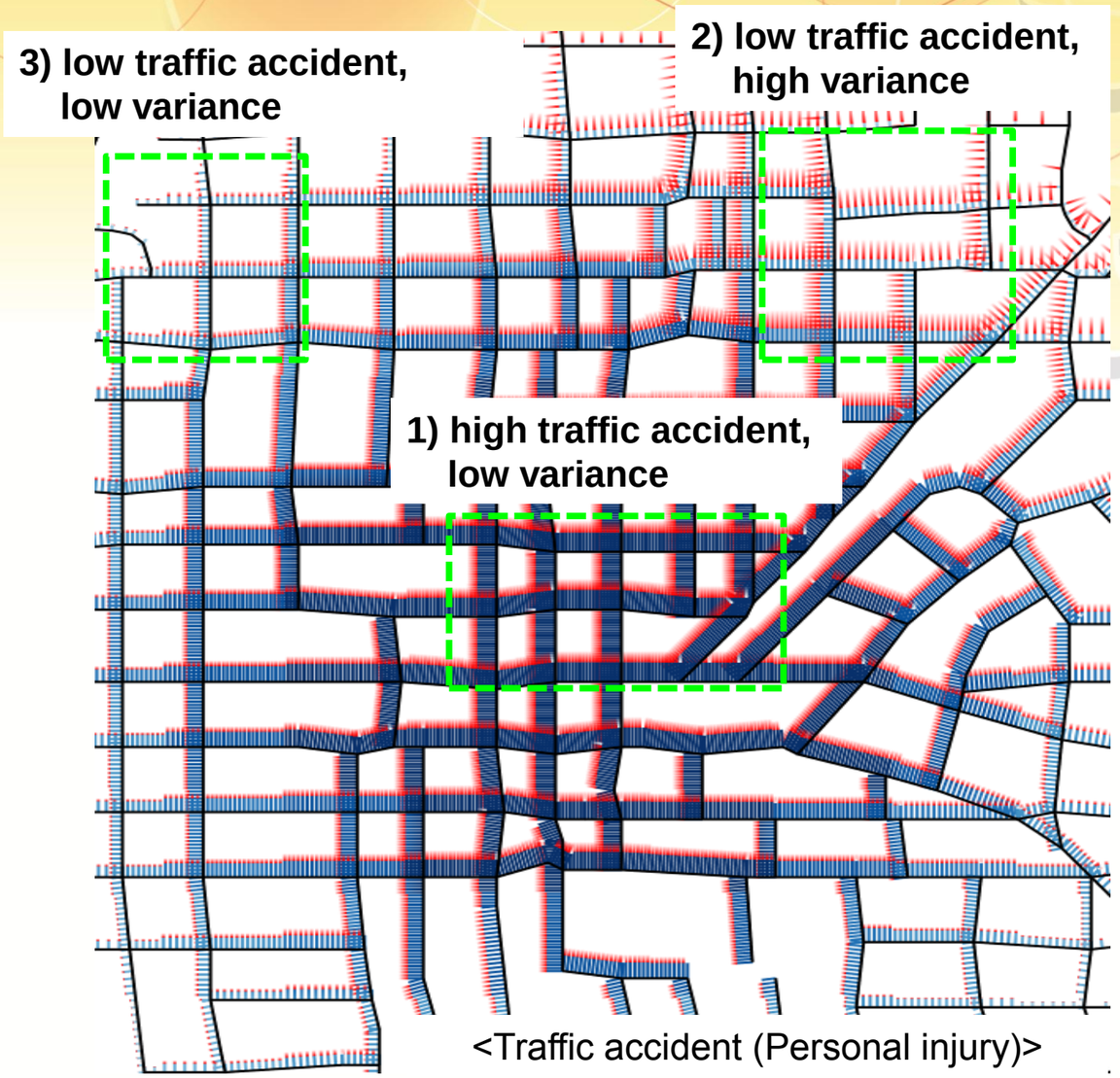


The ratio of temporal variance at P_1

$$L_l = t \times L_{l\,max}$$

$$= (\alpha \times \kappa_{P_1} + \beta \times \upsilon_{P_1}) \times L_{l\,max}$$

$$= \left(\alpha \times \kappa_{P_1} + \beta \times \left(\frac{\tilde{\sigma}_{P_1}}{\tilde{\sigma}_{max}}\right)\right) \times L_{l\,max}$$

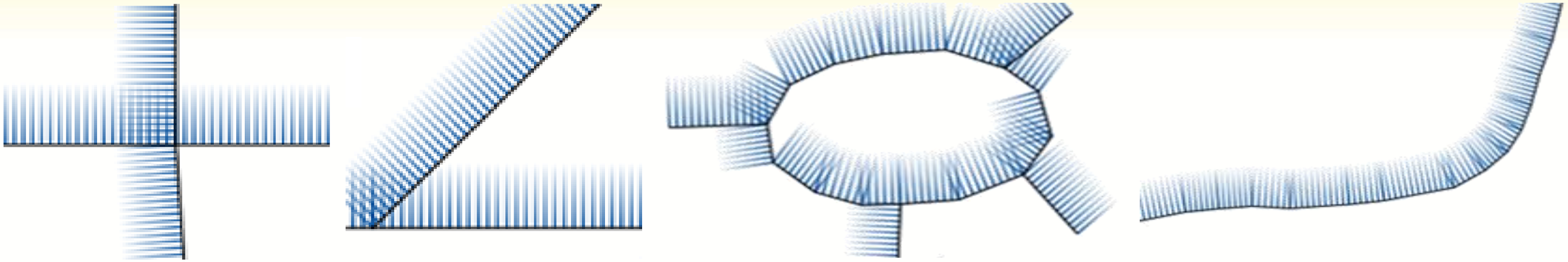


<Traffic accident (Personal injury)>

- Data: color, length (data part), density
- Variance: length (variance part)

Clutter Reduction

- Clutter around **intersections** and **curved lines**



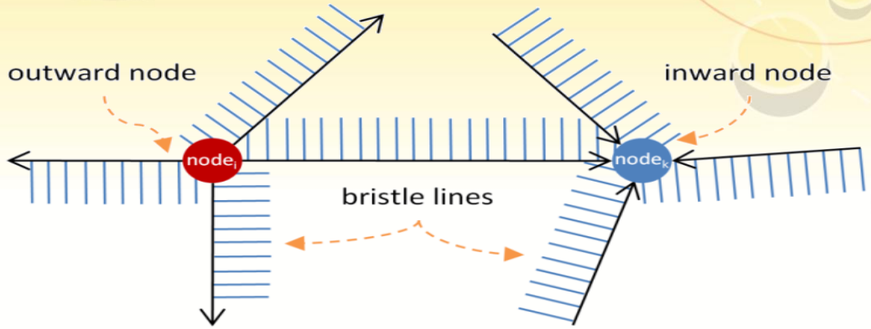
- Two strategies
  - Use a topology graph
  - Cut bristle lines crossing neighbor roads

VIS 2013 VAST • INFOVIS • SCIVIS BIOVIS • LDAV

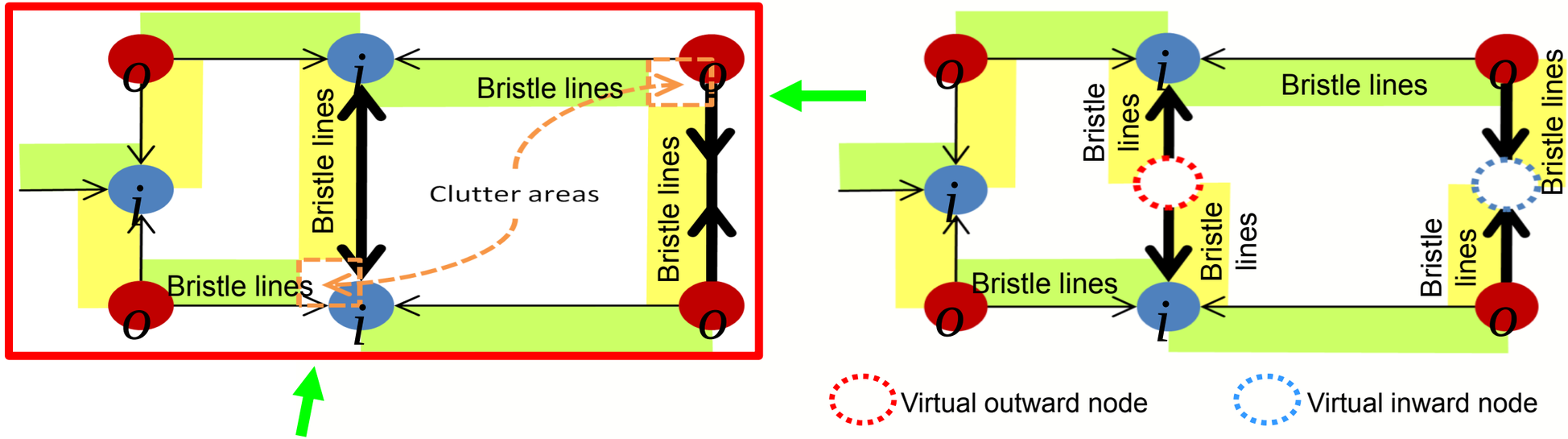
Clutter Reduction:

Using a Topology Graph

- Nodes and directional edges
  - Outward / inward node
  - Edge direction decides bristle orientation

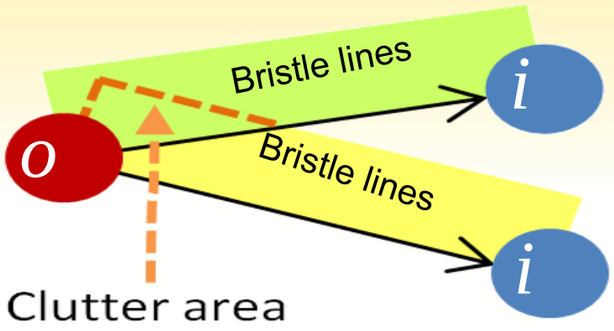


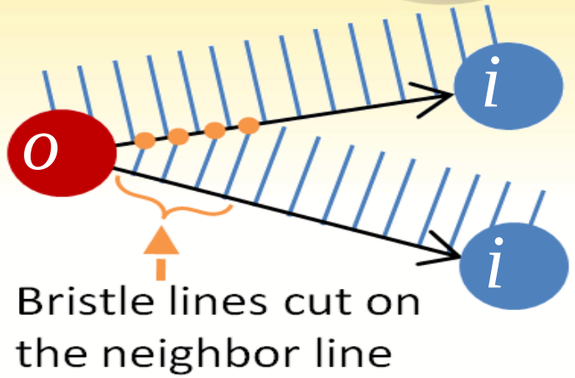
- Virtual nodes
  - Add an outward node in-between two inward nodes
  - Add an inward node in-between two outward nodes

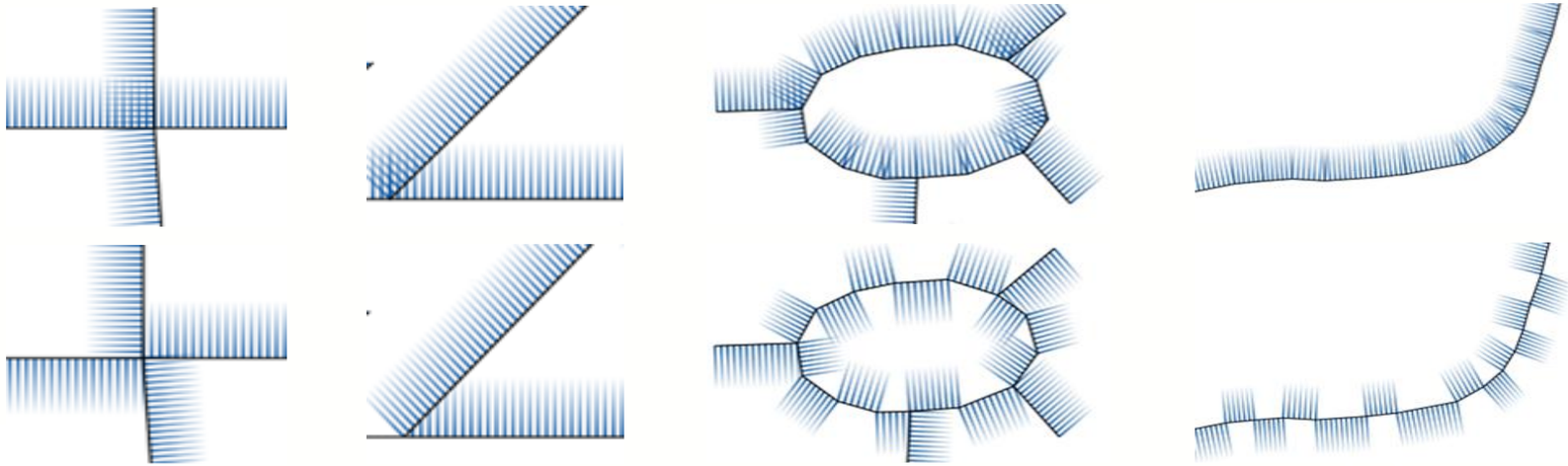


Clutter Reduction:

Avoid Crossing Neighbors

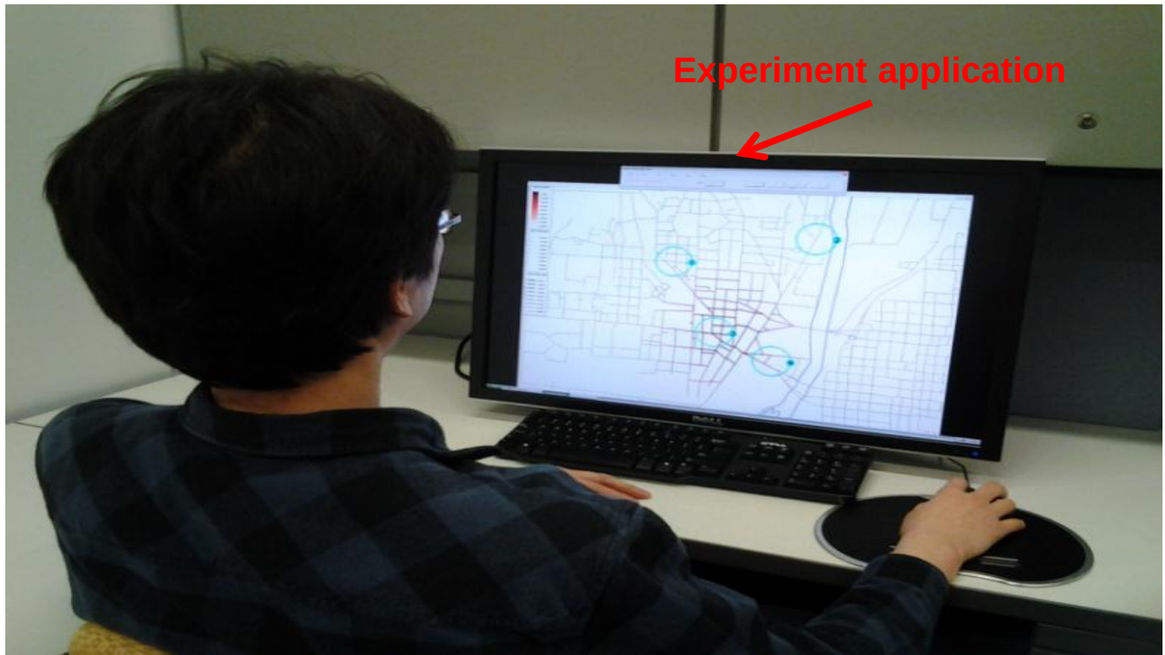






Evaluation

- Effectiveness of bristle maps?
- Participants: 30 students (23 male, 7 female)
- Crime data at Purdue University
- Quantitative measures
  - Task completion time, accuracy



VIS 2013
VAST • INFOVIS • SCIVIS
BIOVIS • LDAV

Evaluation:

Quantitative Evaluation

- Task 1: **Univariate** spatial data encoding techniques
- Task 2: **Bivariate** temporal data encoding techniques
- Task 3: **Multivariate** spatial data encoding techniques

	Data sets	Techniques (or cases)	Trials
Task 1	18	5	90
Task 2	18	5	90
Task 3	15	6	90

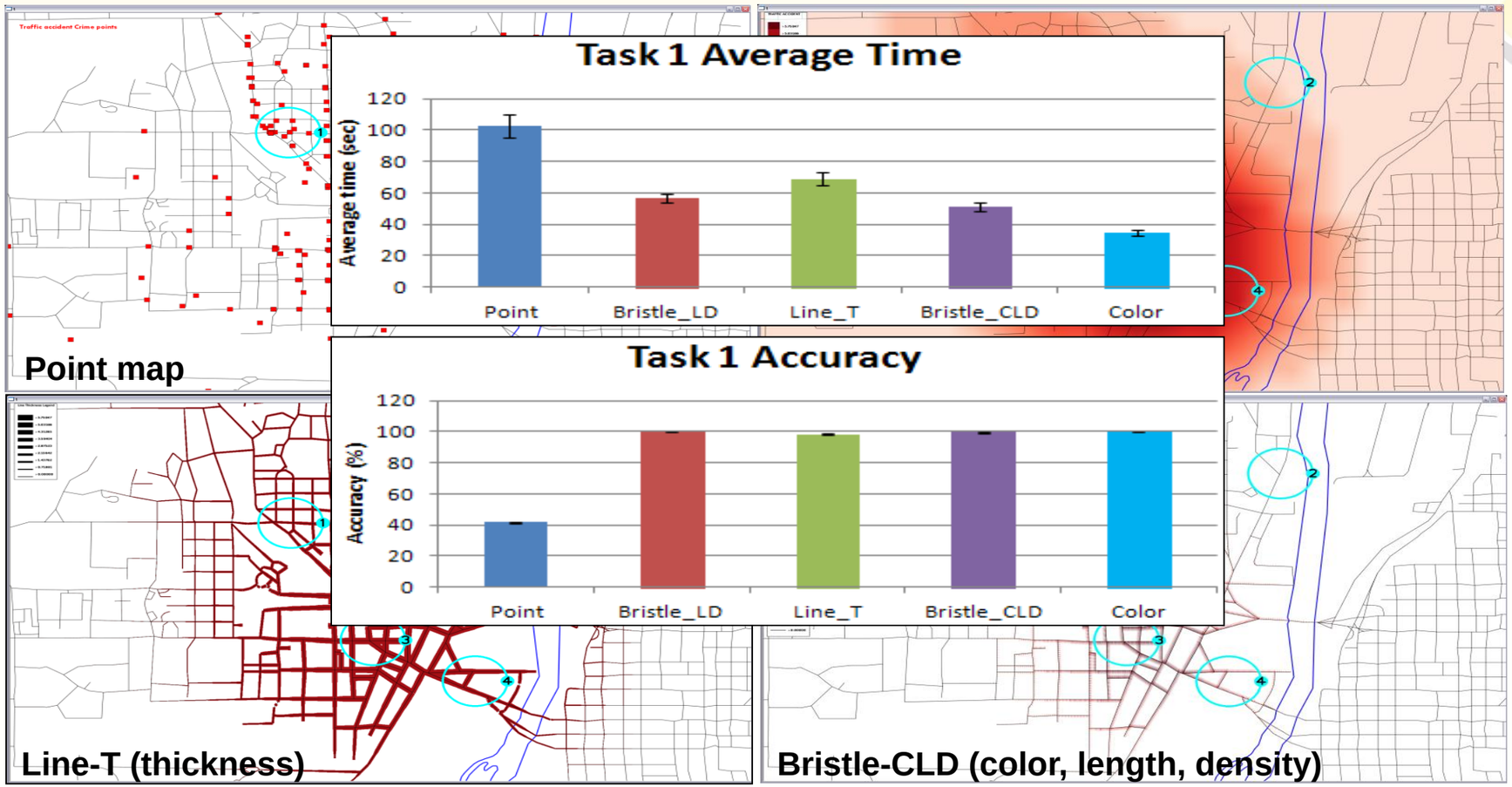
Evaluation:

Task 1

- Results
  - Bristle maps are better than the point and line maps, and equally good as the color maps

5 techniques
: Point map
: Color map
: Line-T
: Bristle-CLD
: Bristle-LD

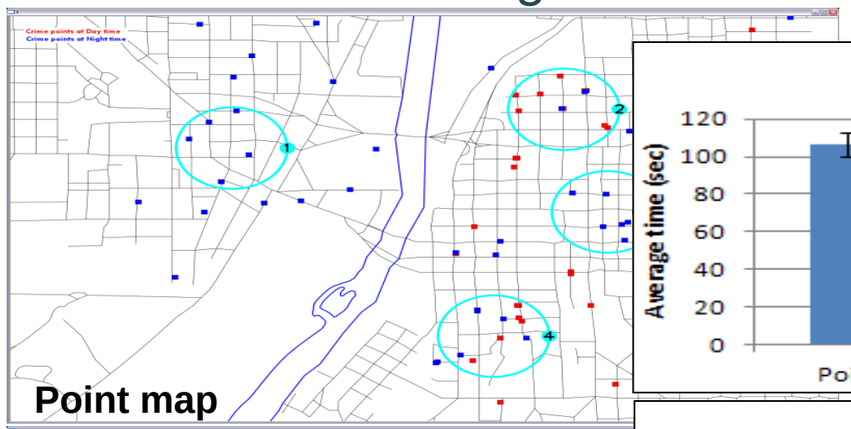
C: color
L: length
D: density
T: thickness



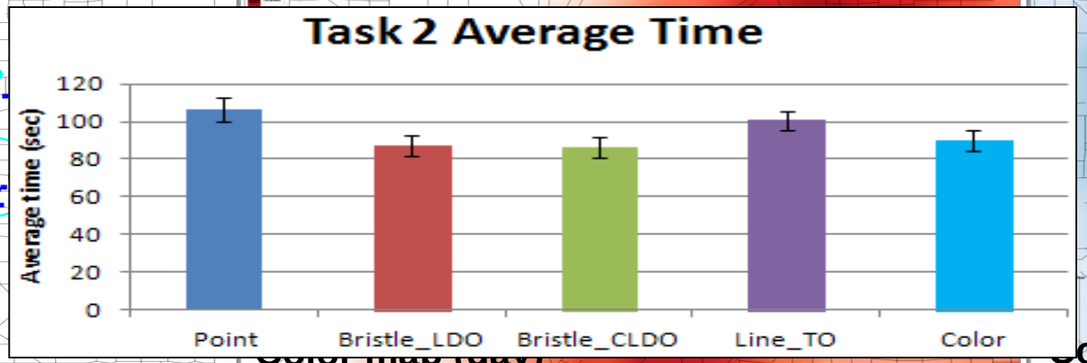
Evaluation:

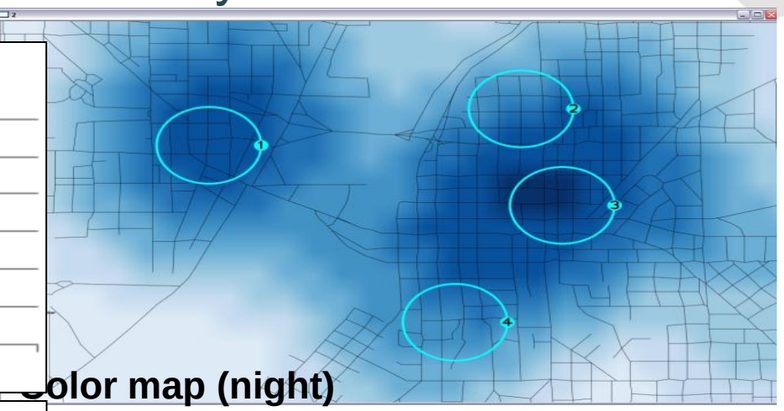
Task 2

- Results
  - Bristle maps are better than the point map, and Bristle-LDO is better than Line-TO
  - Find the region with the highest crime rates including both day and night
  - Find the region with the high crime rates at day, whereas low at night
  - Find the region with the high crime rates at night, whereas low at day

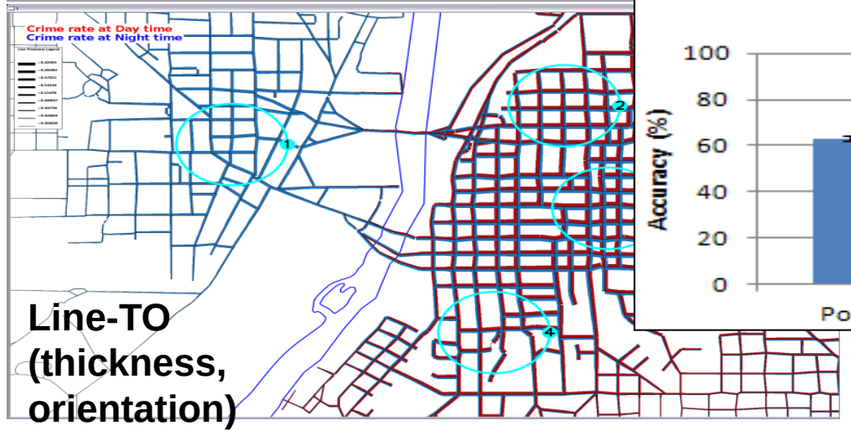


Point map

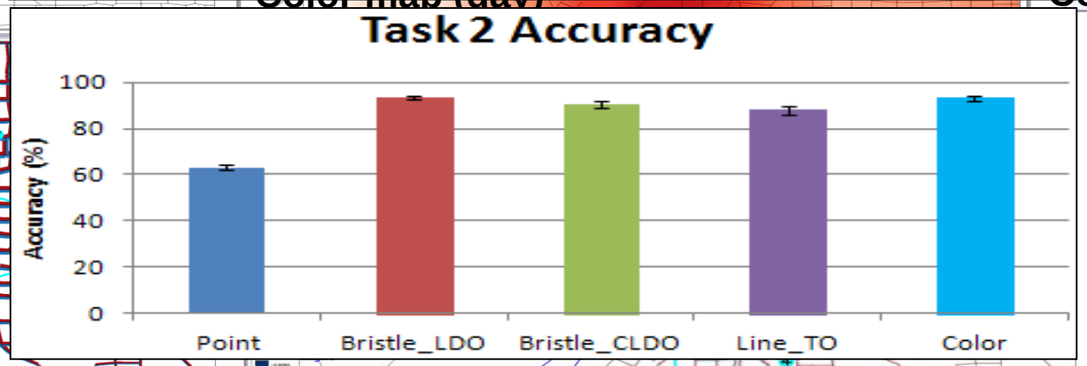




Color map (night)



Line-TO
(thickness, orientation)



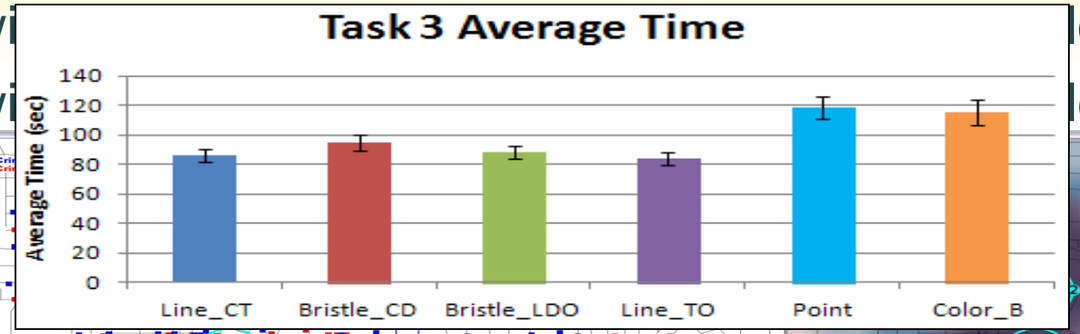
Bristle-CLDO
(color, length, density, orientation)

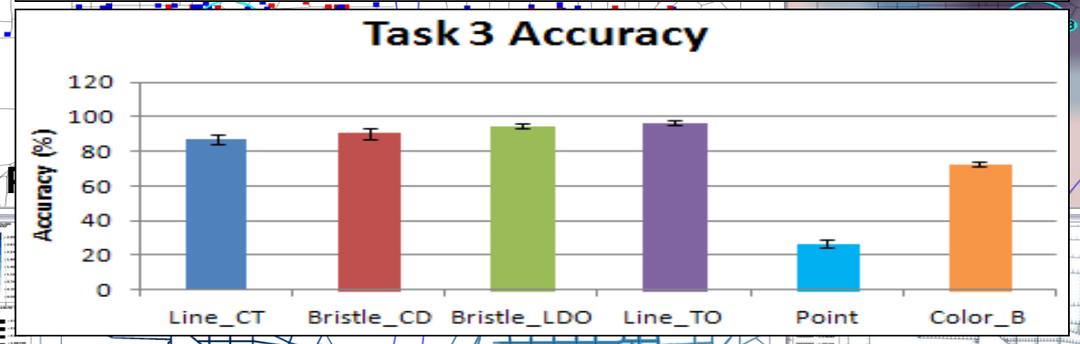
5 techniques
: Point map
: Two color maps
: Line-TO
: Bristle-CLDO
: Bristle-LDO

Evaluation:

Task 3

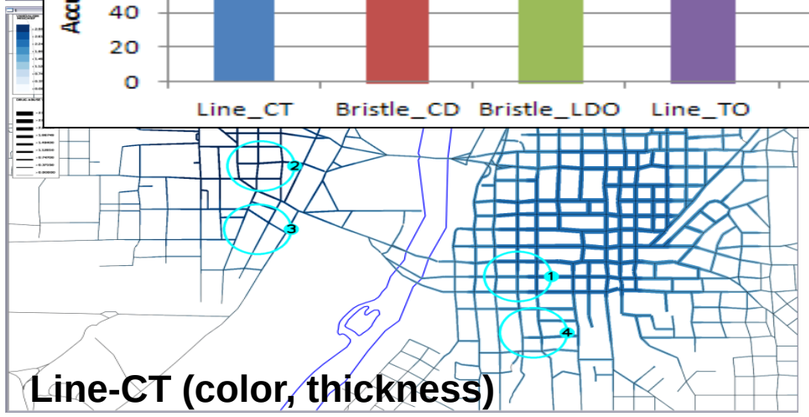
- Results
  - Bristle maps are better than the point and bivariate color maps
  - Find the region with the highest point rates for both crimes
  - Find the region wi… low crime 2
  - Find the region wi… low crime 1

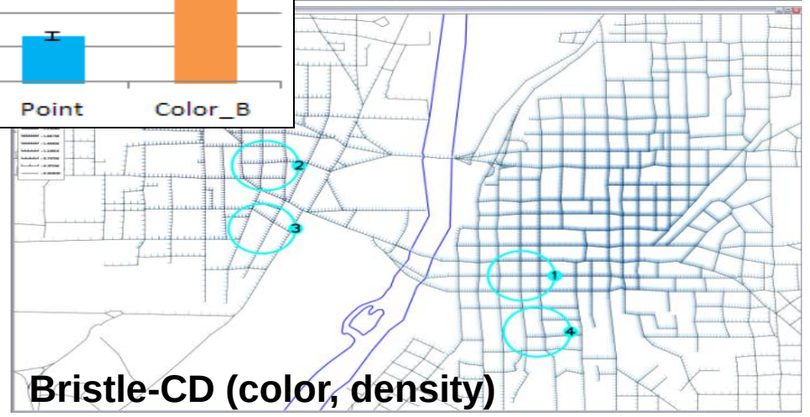




6 techniques
: Point map
: Bivariate color map
: Line-TO
: Line-CT
: Bristle-LDO
: Bristle-CD

C: color
L: length
D: density
O: orientation
B: bivariate
T: thickness



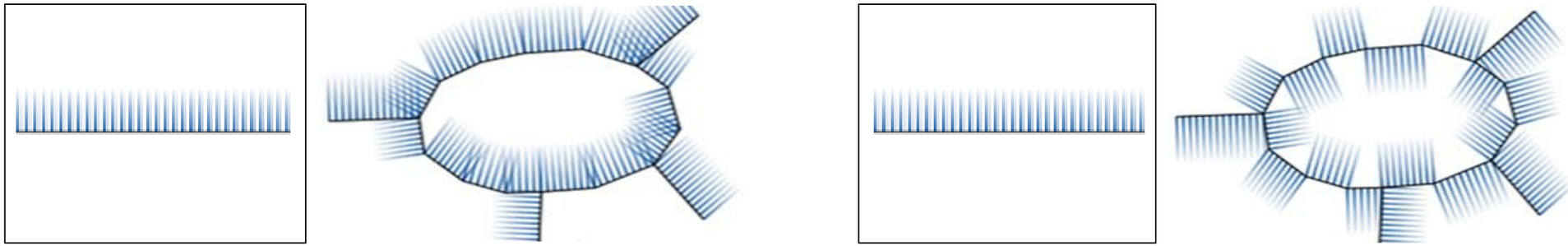


Evaluation:

Clutter Removal

- Question
  - Answer if two regions look either same or different in terms of color, length, and density of bristle lines

Example:



Base vs. before clutter reduction

Base vs. after clutter reduction

- Results
  - Perception of bristle lines before and after applying the clutter reduction is not significantly different in terms of time and accuracy
  - Participants similarly perceived bristle lines before and after clutter reduction

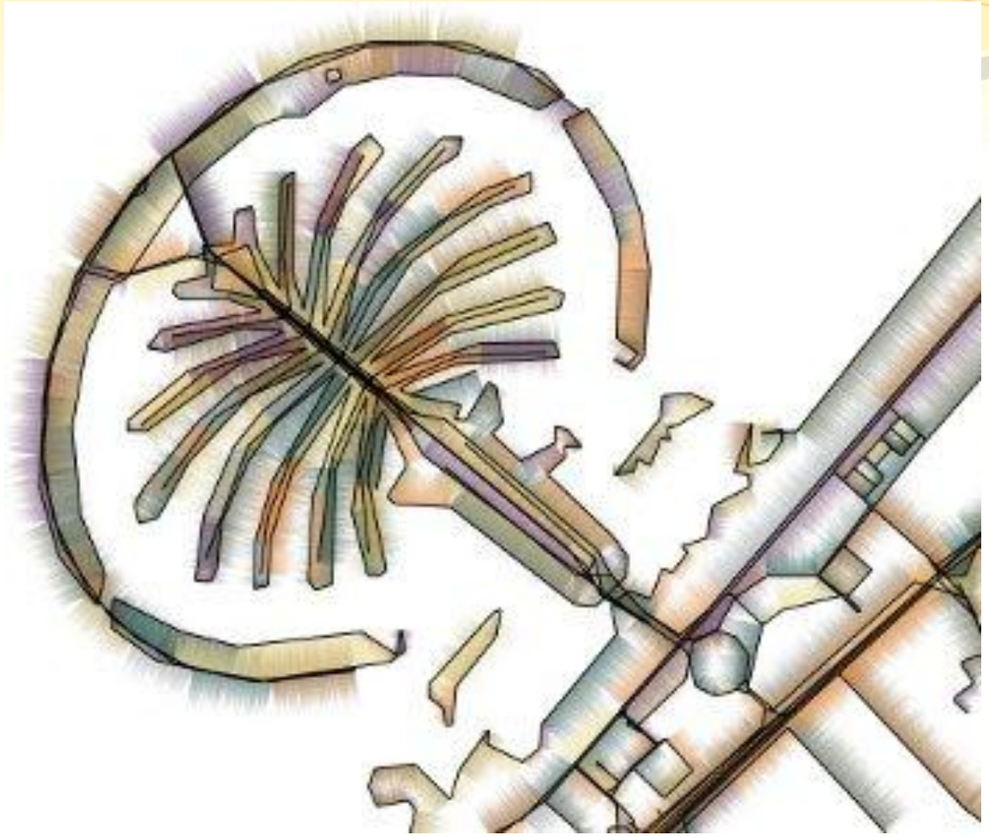
VAST • INFOVIS • SCIVIS
BIOVIS • LDAV
VIS 2013

Evaluation:

Aesthetics

	Point	KDE Map	Bristle	Line
Average	2.26	2.6	2.79	2.34
Std Dev	1.18	.97	1.19	1.10





Tobias Isenberg (2013) *Visual Abstraction and Stylisation of Maps.* The Cartographic Journal, 50(1):8–18, February 2013.

Evaluation:

Summary and Lessons

- As a **univariate** encoding
  - Bristle maps are **better than point, color, and line maps** in terms of time
  - Bristle **density** has a **significant** effect compared to the line map

- As a **bivariate** encoding
  - **Orientation** in bristle maps has
    - **no significant** effect compared to **two color maps**
    - **significant** effect compared to the **bivariate color map**
  - Cognitive load
    - Two color maps < Bristle-(C)LDO < Bivariate color map

Evaluation:

Summary and Lessons

- As a **multivariate** encoding
  - Our bristle maps show **potential in a single view** by providing different parameter combinations
- As a **temporal variance** encoding
  - Bristle map is **significantly better** than other methods
- Bristle maps are an effective multivariate encoding technique
- There are clear cases where the combination is effective in identifying combined areas
- However, for value estimation bristle maps may be cognitively burdensome

Conclusions

- Bristle maps are an effective multivariate encoding technique
- There are clear cases where the combination is effective in identifying combined areas
- However, for value estimation bristle maps may be cognitively burdensome
- We found line maps and heat maps to both be significantly faster than point maps and bristle maps in identifying values (all techniques were equivalent in accuracy)

Arizona State Is Hiring!

- Arizona State University is hiring in Computer Science
  - Big Data (Job #10518)
  - Cyber Security (Job #10519)
  - Next Gen Computing (Job #10523)
  - Software Engineering (Job #10524)
  - Human Centered Computing (Job #10555)
  - Emerging Computational Technologies for Learning (Job #10556)
- http://engineering.asu.edu/hiring
- I have lots of fliers if you want one!



VIS 2013 VAST • INFOVIS • SCIVIS BIOVIS • LDAV

Purdue University is Hiring!

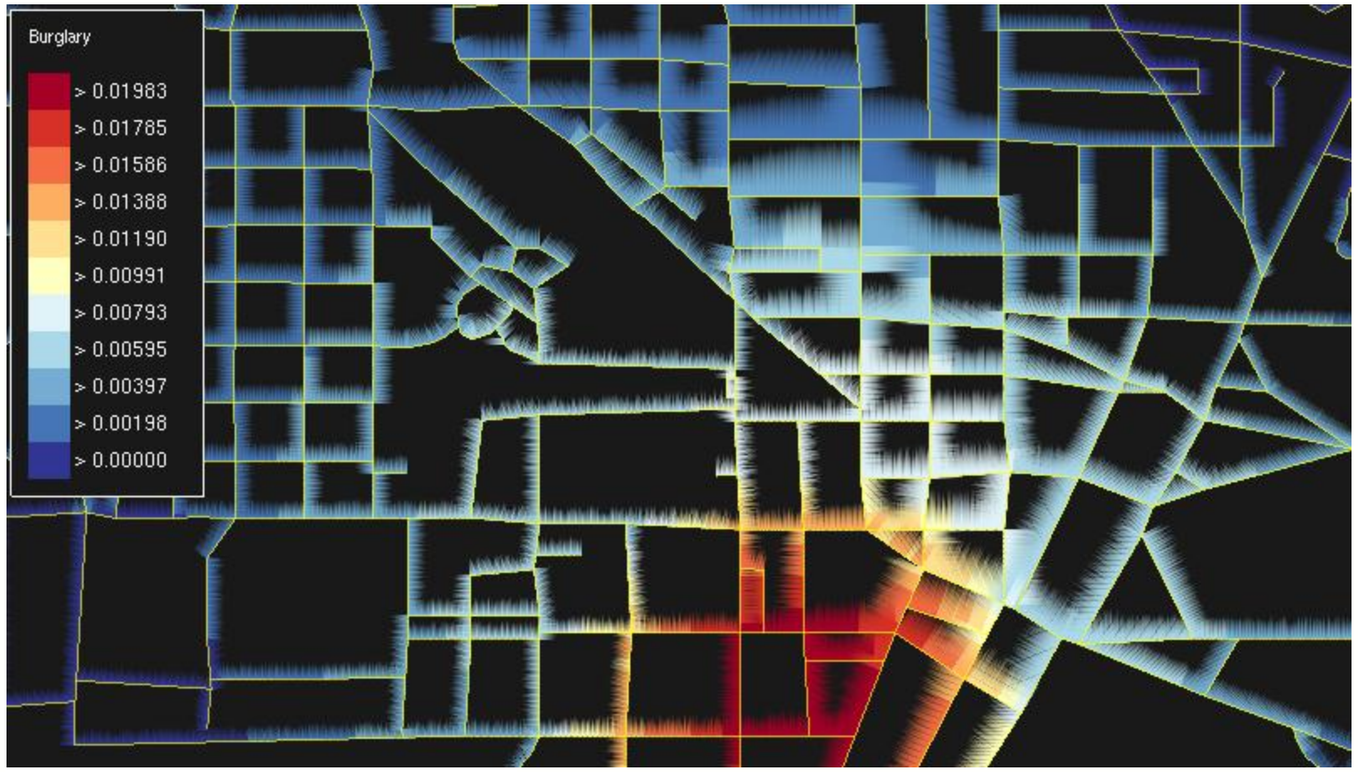
- The School of Electrical and Computer Engineering at Purdue University invites applications for a faculty position at any level in human-centered computing, including but not limited to visualization, visual analytics, human-computer interaction (HCI), imaging, and graphics.

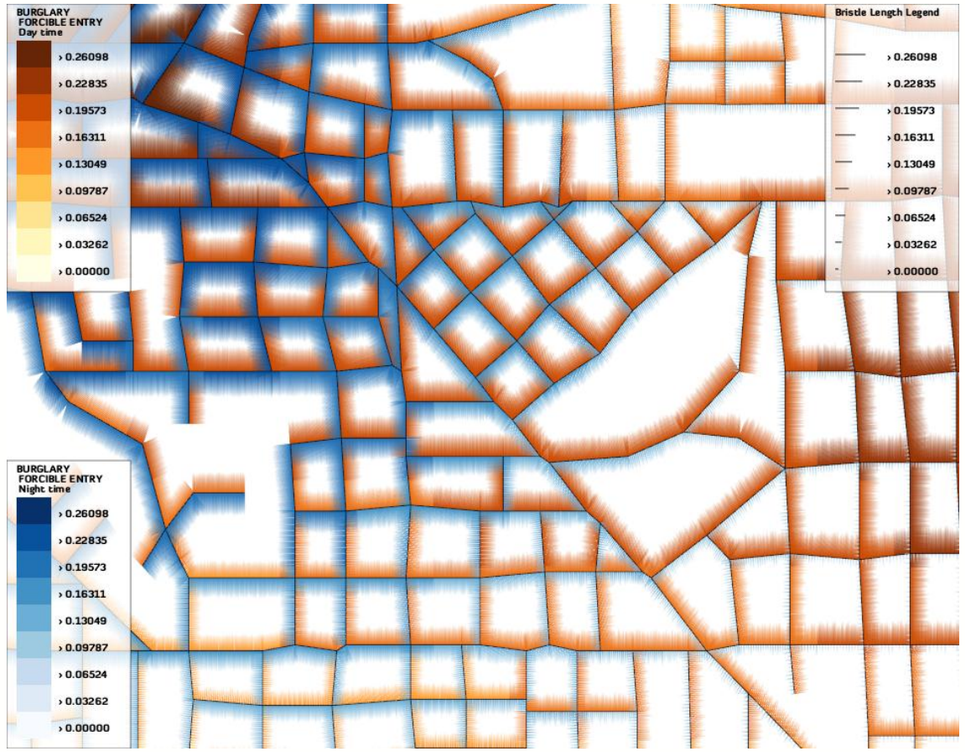


VIS 2013 VAST • INFOVIS • SCIVIS BIOVIS • LDAV

Questions?

- This work was supported by the US Department of Homeland Security's VACCINE Center under Award Number 2009-ST-061-CI0001, the Industrial Strategic technology development program,10041772, funded by the Ministry of Knowledge Economy (MKE, Korea), and the French DIGITEO chair of excellence





More details at rmaciejewski.faculty.asu.edu

Evaluation:

Evaluation Design

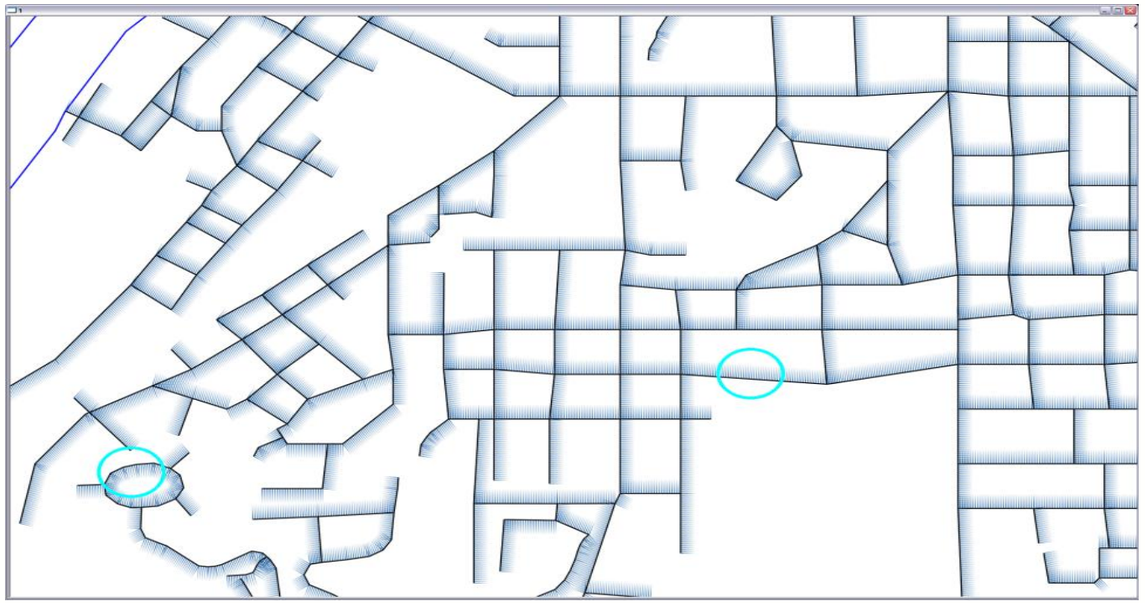
- Hypotheses
  - Bristle maps would be better than other techniques as the complexity level of tasks increased from univariate to multivariate
  - Reason
    - Bivariate color maps and line maps use at most two variables, whereas bristle maps have several encoding parameters
- Participants experience
  - All participants had experience in visualizing data on geographical maps using colors and icons (e.g., paper maps, online map services)
  - Almost daily: 11 participants
  - 1-2 times a week: 17 participants
  - 1-3 time a month: 2 participants
- Apparatus
  - 30” monitor, 2228 × 1478 resolution
  - Windows XP



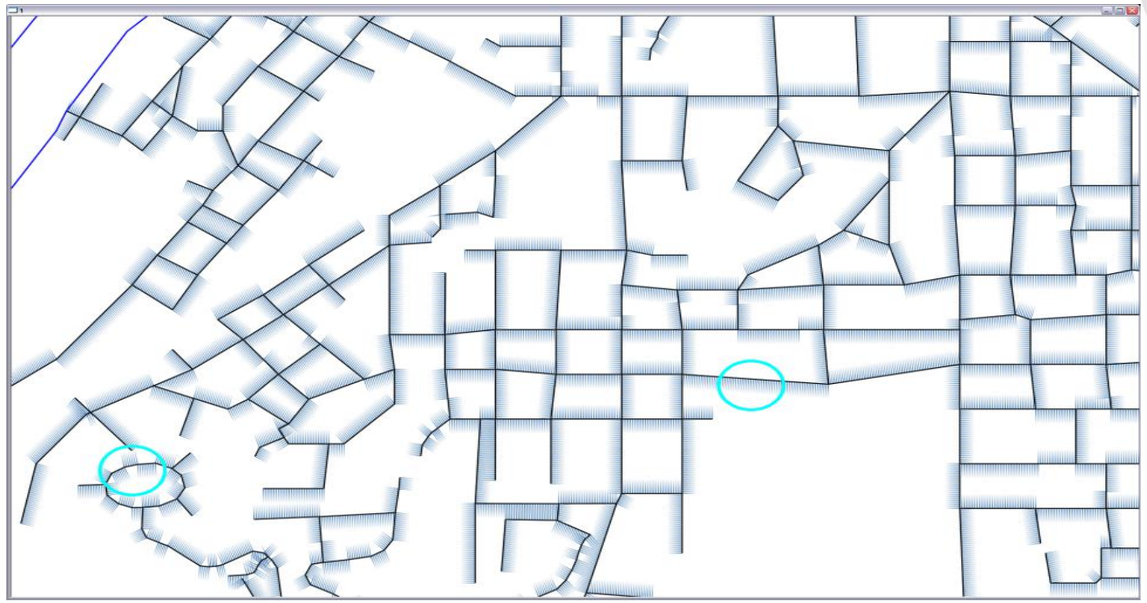
Evaluation Study:

Quantitative Evaluation

- **Task 5** : Do you think that two regions represent same or different amount of crime rates?

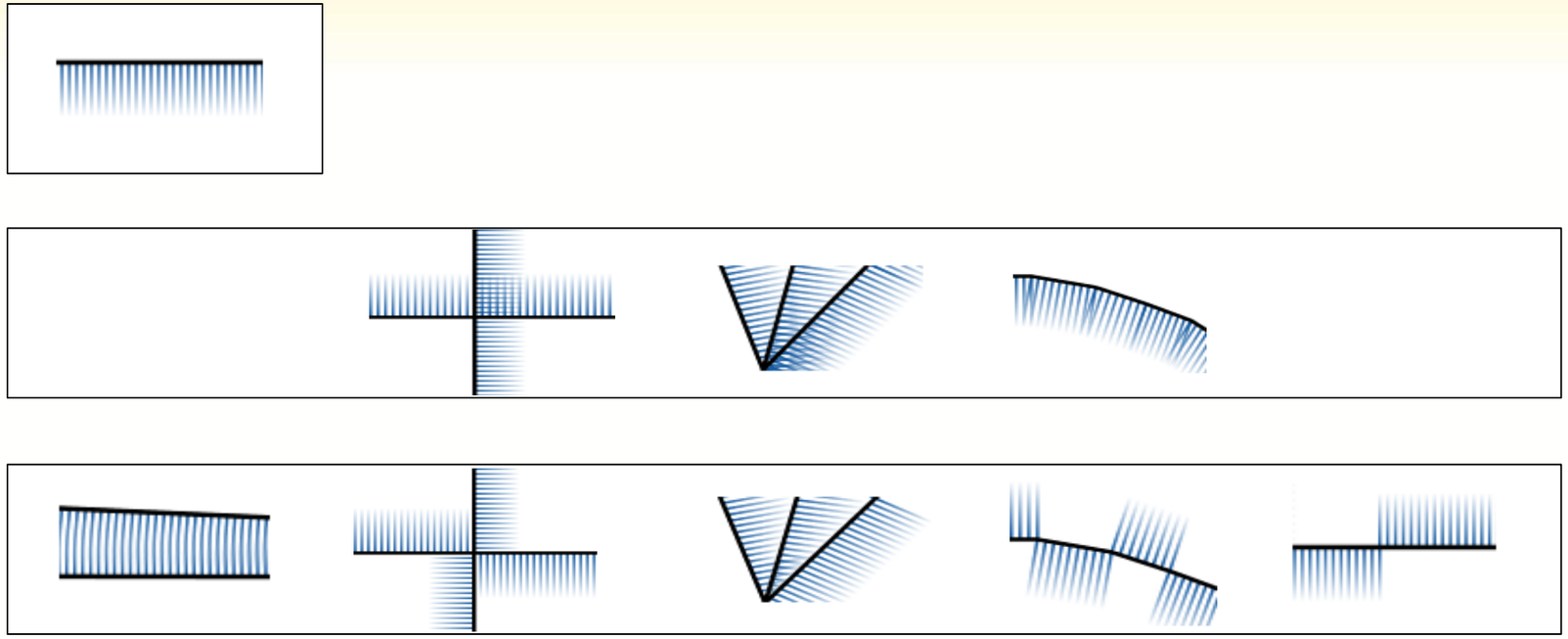


<Base vs. non-clutter reduced bristles>



<Base vs. clutter reduced bristles>

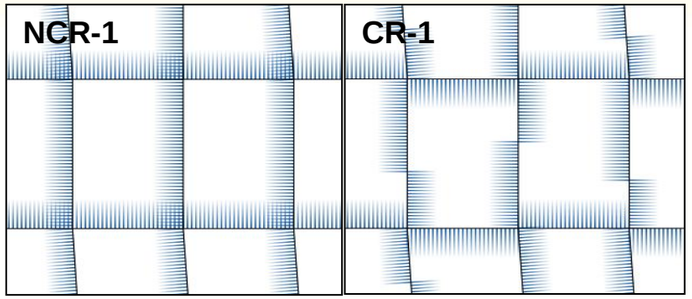
Cases in Clutter Reduction

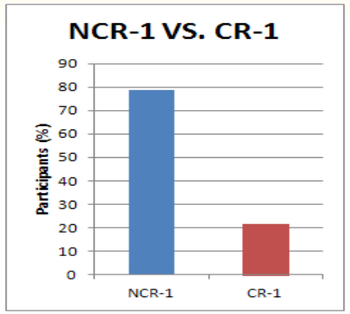


Evaluation Study:

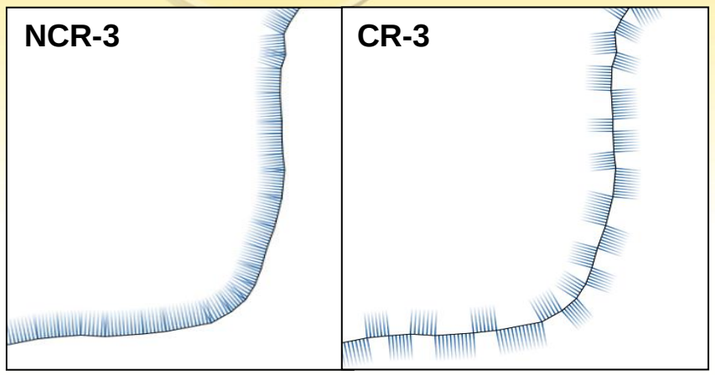
Qualitative Evaluation

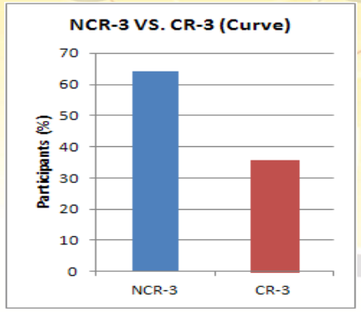
- Question
  - Preference of clutter reduction



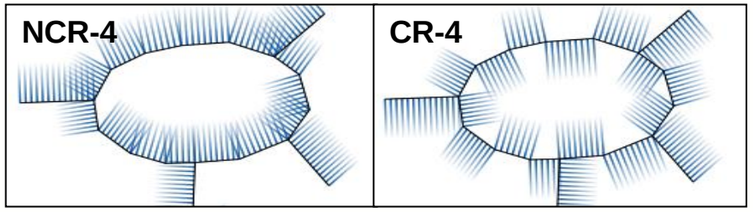


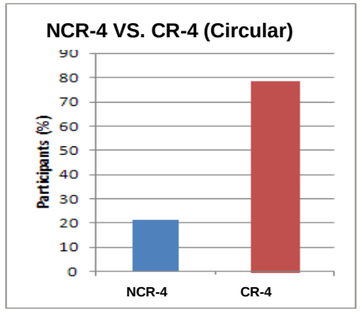
Case 1: Changing orientation using topology



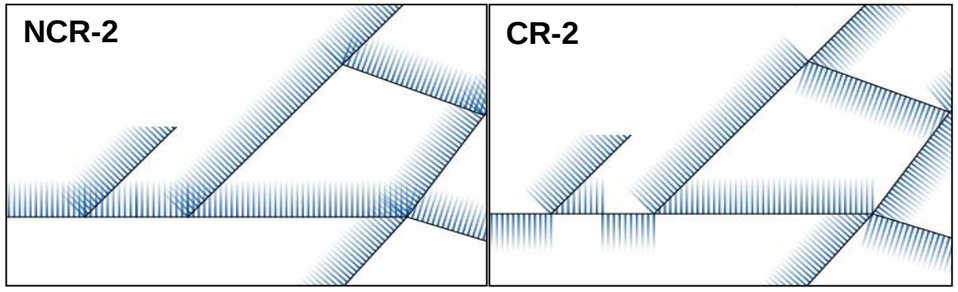


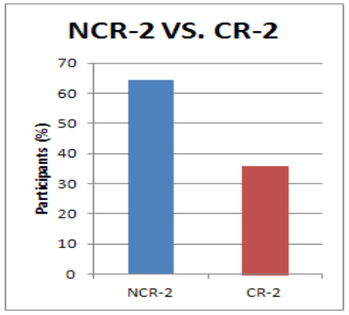
Case 3: Curved roads





Case 4: Circular roads



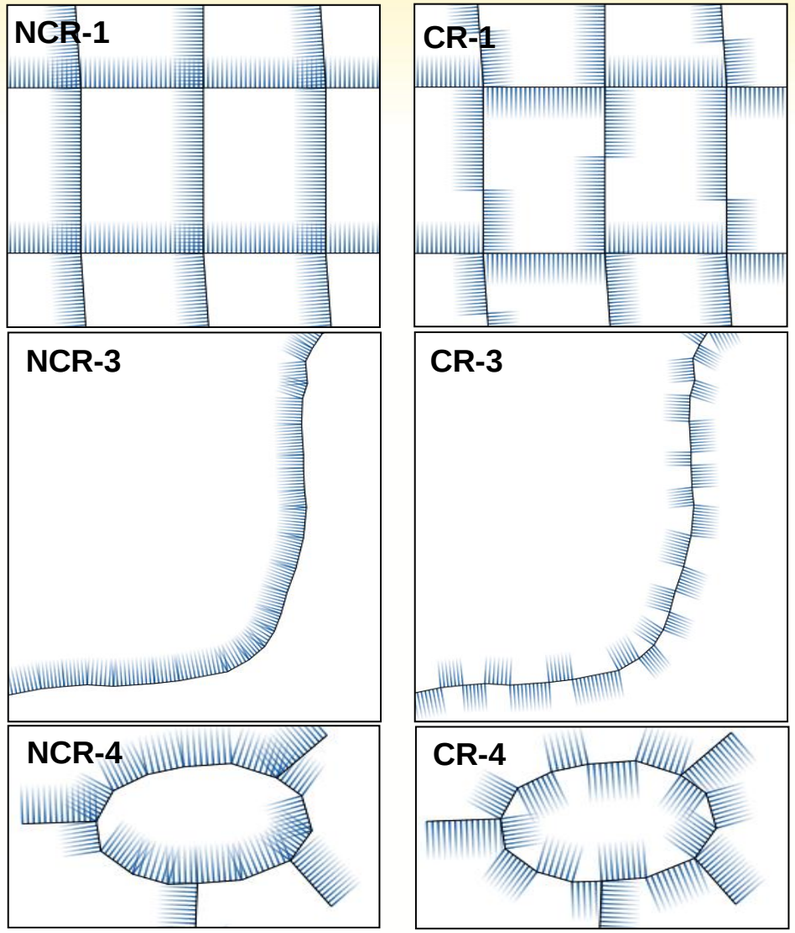


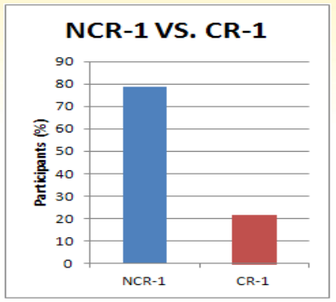
Case 2: Cutting bristles crossing neighbor roads

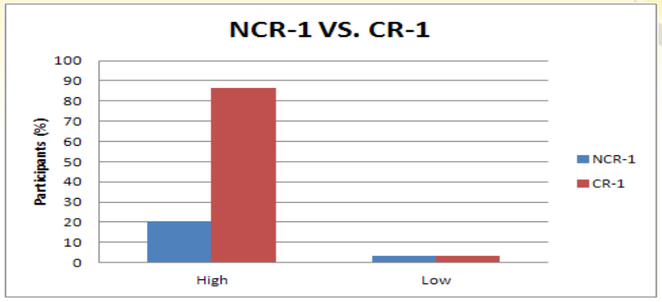
Evaluation Study:

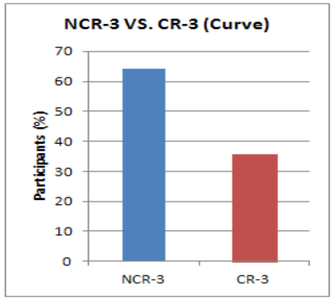
Qualitative Evaluation

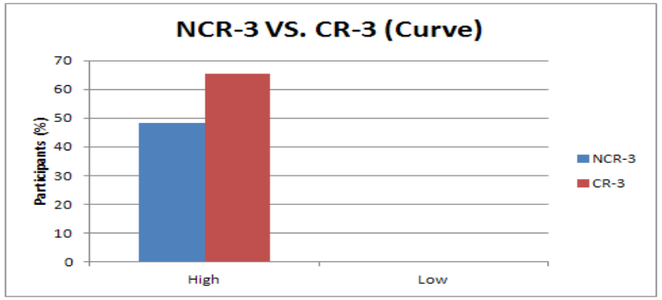
- **Preference** vs. **visual complexity**

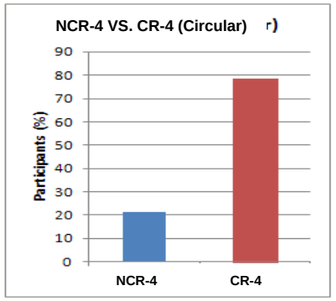


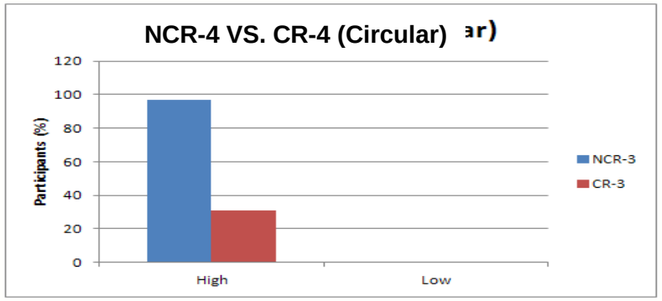












Encoding Schemes:

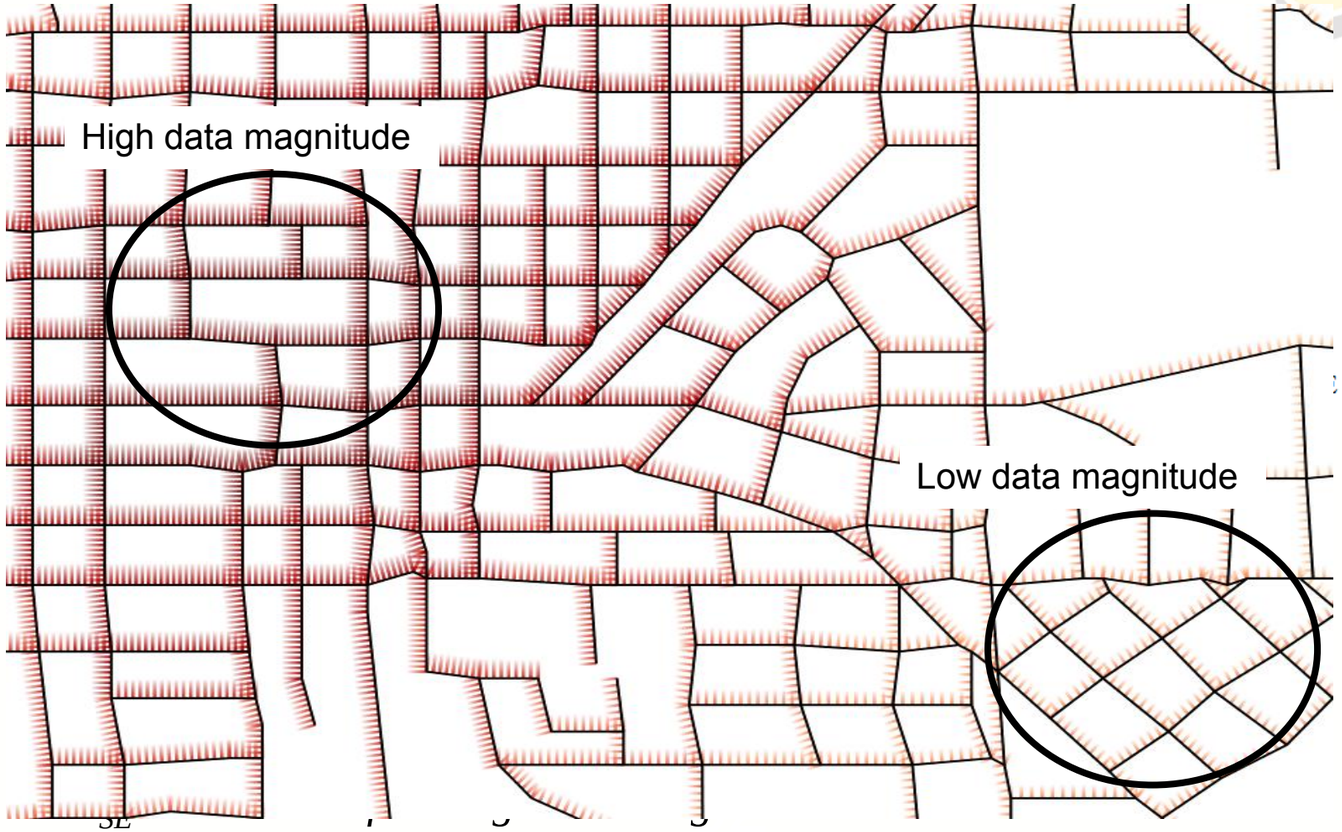
Color, Length and Density as Data Magnitude

- Color
  - Linear mapping from the ratio of data magnitude [0, 1] to color scheme
- Length
- $$L_l = t \times L_{lmax} = (\alpha \times \kappa_{P_1} + \beta \times \upsilon_{P_1}) \times L_{lmax}$$

$$L_{lmax} = \rho \times log_b \left(\frac{1}{N_r} \sum_{i=0}^{N_r - 1} L_{\overline{SE}} \right)$$

- Density

$$N_l = \rho \left(\frac{\zeta}{\lambda} L_{\overline{SE}} \right) \kappa_{\overline{SE}}$$



Evaluation:

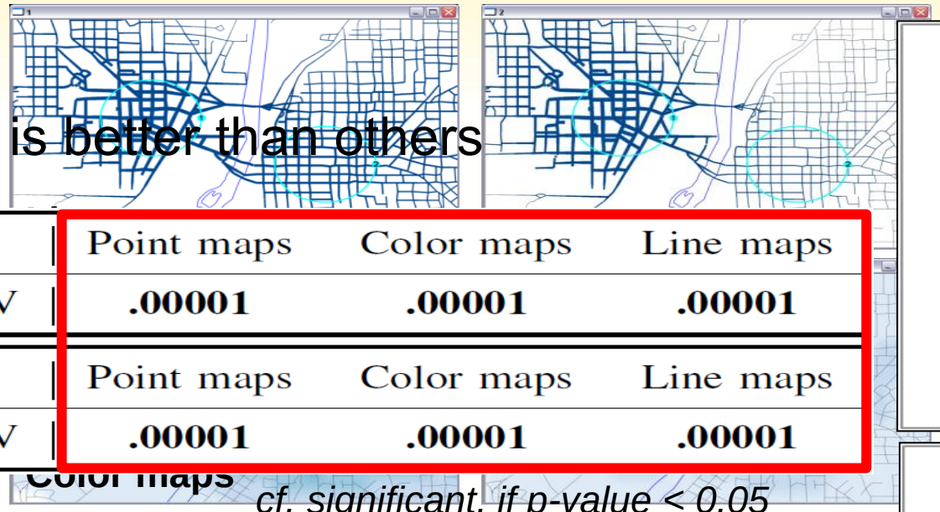
Task 4

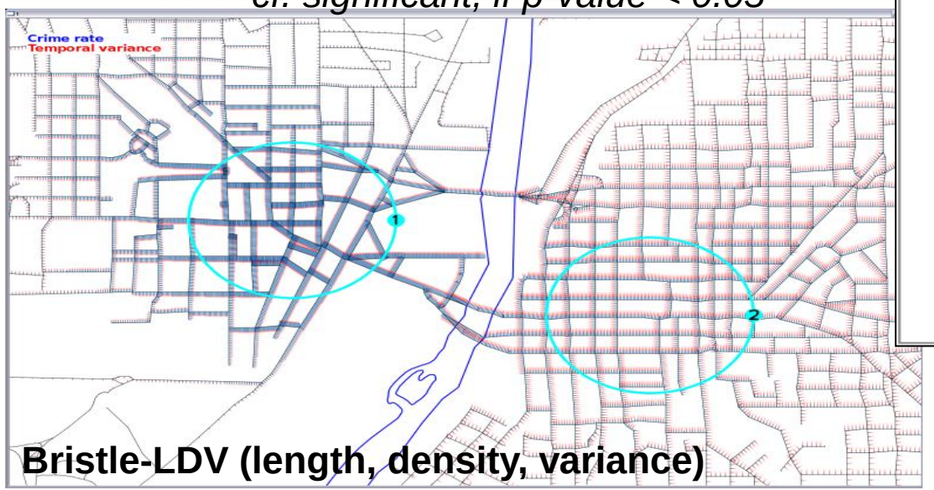
- Question
  - Find the region with the highest temporal variance
- Results
  - Bristle map is better than others

		Point maps	Color maps	Line maps
Time	p-value <			
	Bristle-LDV	**.00001**	**.00001**	**.00001**
Accuracy	p-value <	Point maps	Color maps	Line maps
	Bristle-LDV	**.00001**	**.00001**	**.00001**

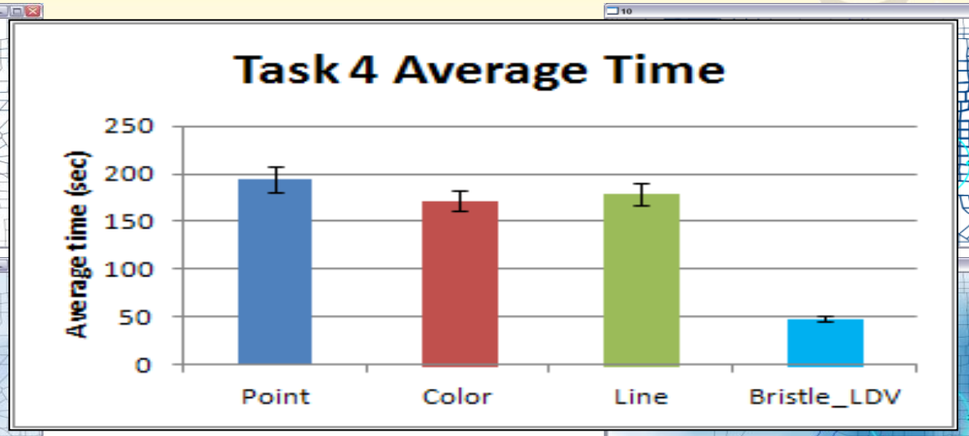
cf. significant, if p-value < 0.05

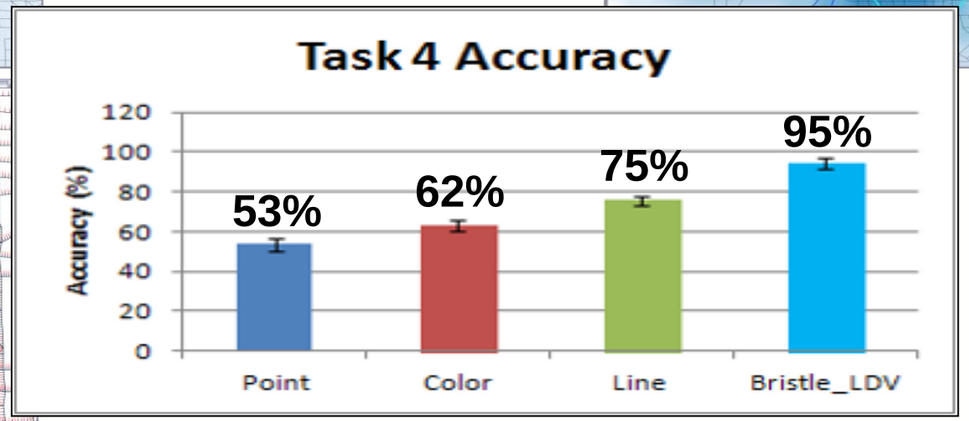
L: length
D: density
V: variance





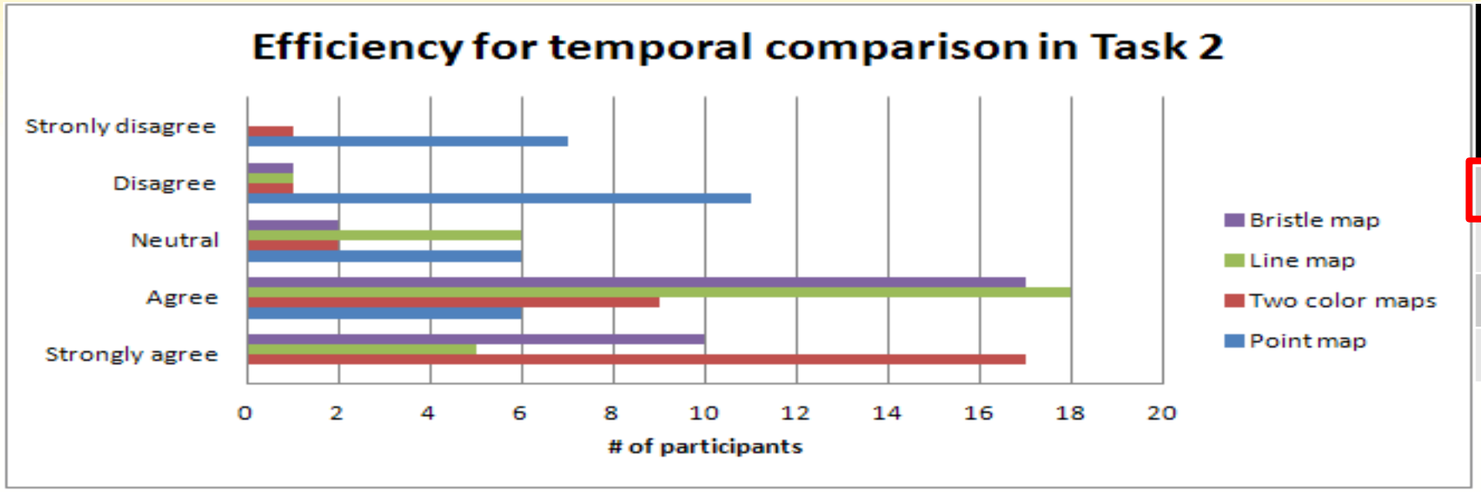
Bristle-LDV (length, density, variance)



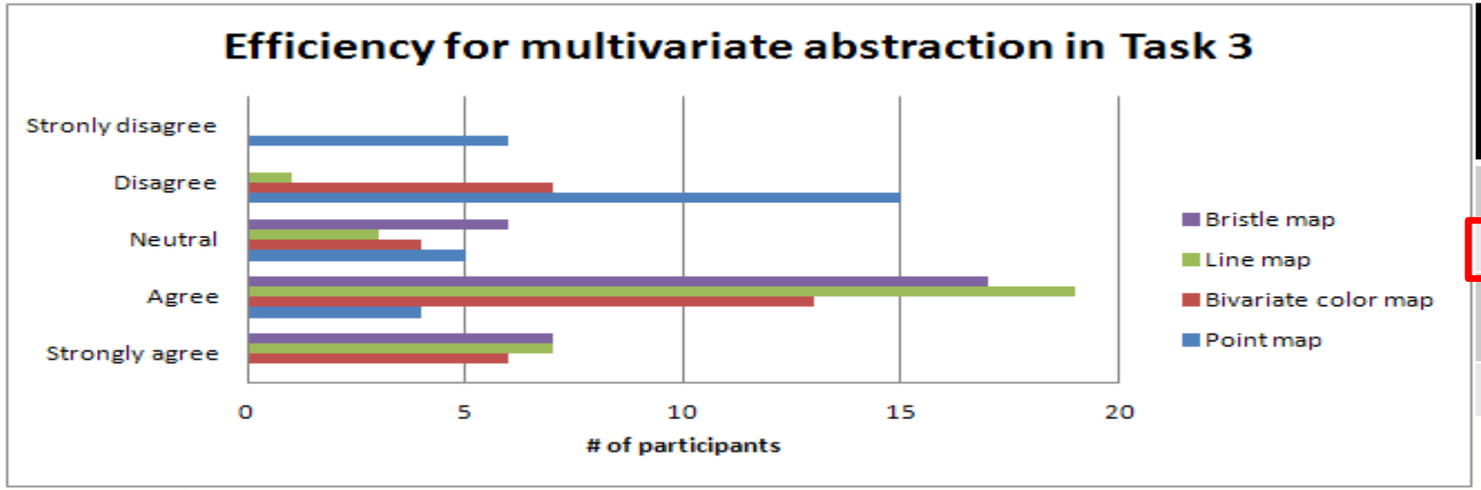


Evaluation:

Qualitative Evaluation (1/2)



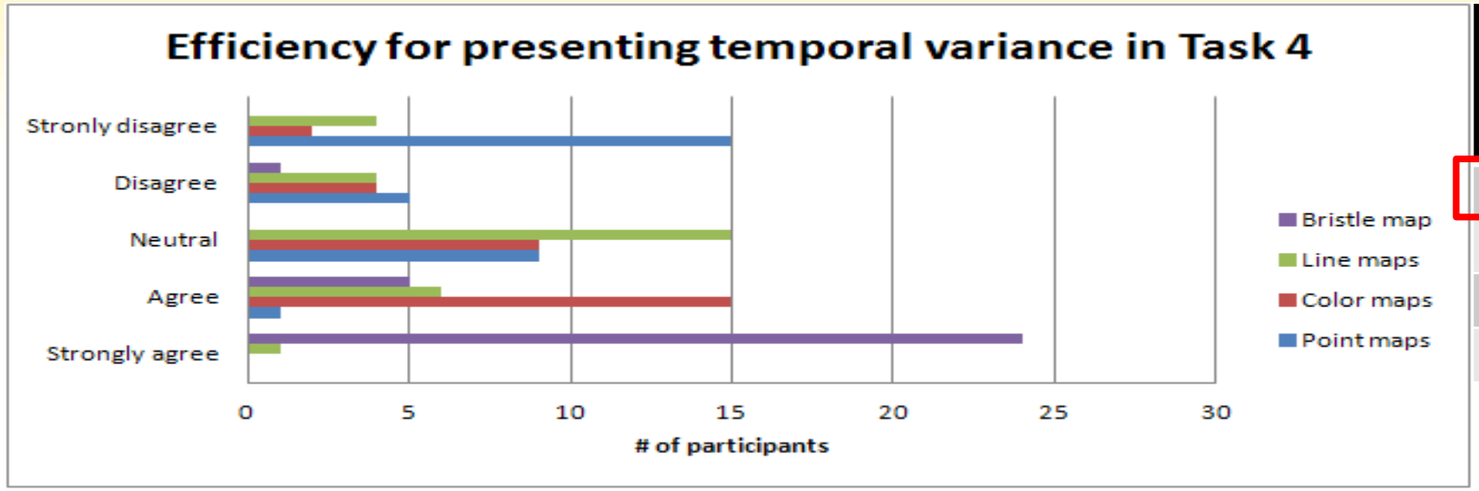
Technique	Agree / strongly agree (total: 30 participants)	%
Bristle map	**27**	**90%**
Two color maps	26	86.7%
Line map	23	76.7%
Point map	6	20%



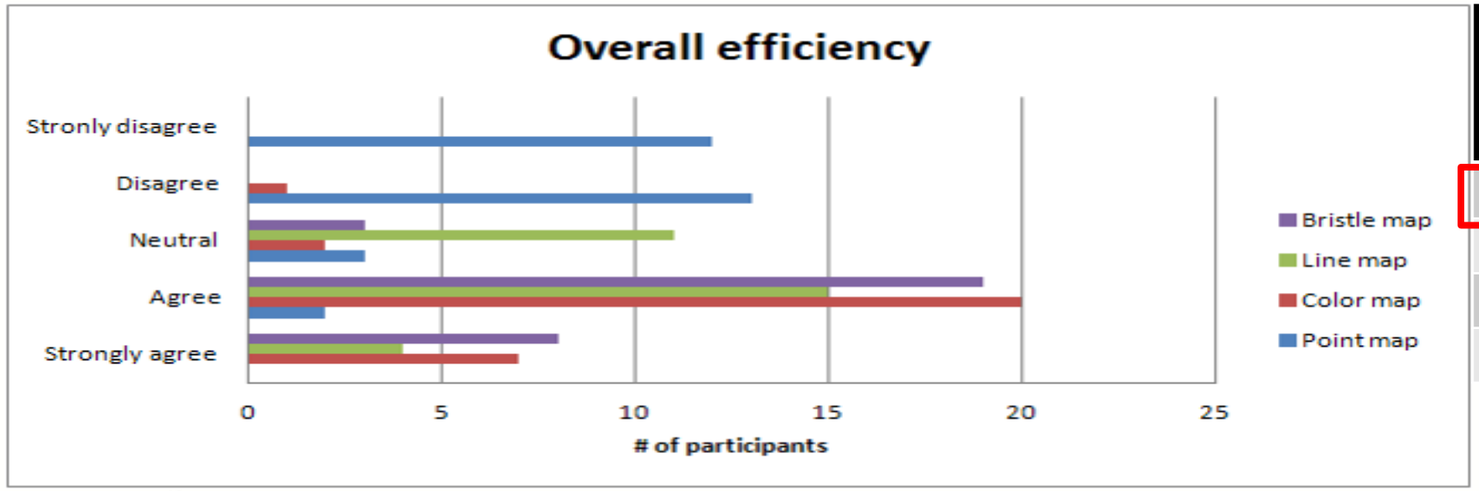
Technique	Agree / strongly agree (total: 30 participants)	%
Line map	26	86.7%
Bristle map	**24**	**80%**
Bivariate color map	19	63.3%
Point map	4	13.3%

Evaluation:

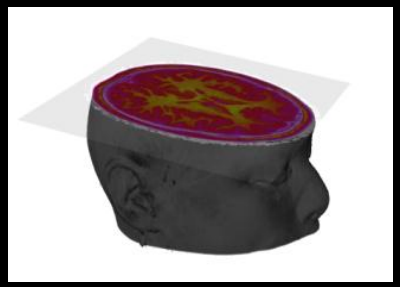
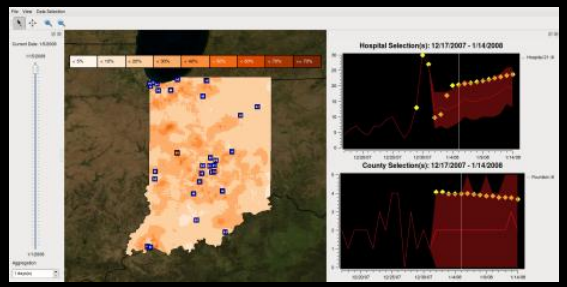
Qualitative Evaluation (2/2)

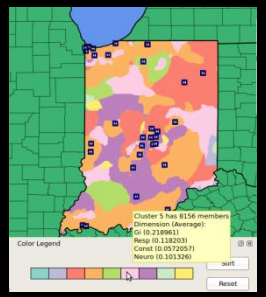


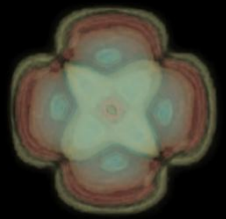
Technique	Agree / strongly agree (total: 30 participants)	%
Bristle map	**29**	**96.7%**
Color maps	15	50%
Line maps	7	23.3%
Point maps	1	3.3%



Technique	Agree / strongly agree (total: 30 participants)	%
Bristle map	**27**	**90%**
Color map	27	90%
Line map	19	63.3%
Point map	2	6.7%



Enabling Predictive Analytics Through Visualization

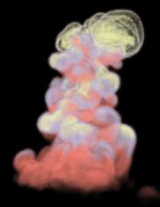
Ross Maciejewski



Ross Maciejewski
E-mail: rmacieje@asu.edu
http://rmaciejewski.faculty.asu.edu/



The Ongoing Data Deluge



- *YouTube users upload 48 hours of new video every minute of the day*
- *In 2009, drones from Iraq and Afghanistan recorded 24 years of video footage*
- *Walmart handles more than 1 million customer transactions every hour*
- *235 Terabytes of data has been collected by the U.S. Library of Congress in April 2011*
- *McKinsey report estimates that by 2018 the U.S. could face a shortage of 140k – 190k data analysts*

Research Challenges in the Data Deluge

Long-term fundamental challenges

- Mobility
- Scale
- Adaptation
- Uncertainty
- Multisource, cross-media analysis
- Numerical to visual analysis

"All this data must be fused with relevant contextual or situational information and visualized in such a way to give our forces a clear picture of threats, options and potential consequences.

The science and technology community has been successful in giving our forces a data advantage; now it's time to give them a decision advantage."

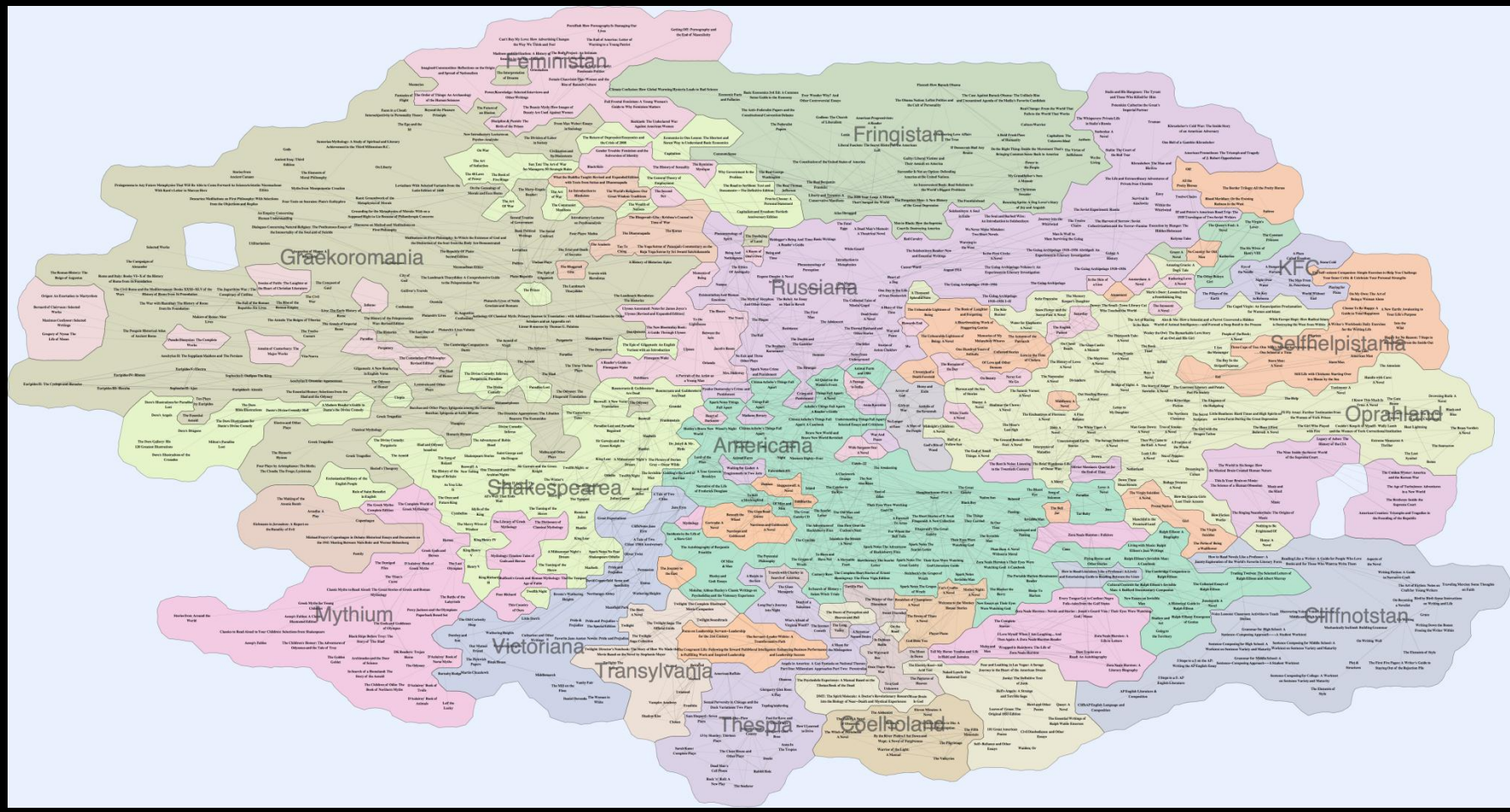
-Zachary J. Lemnios Director, Defense Research and Engineering Department of Defense February 17, 2010

Having Fun In the Data Deluge

- *More people are injured on toilets than by skiing or snowboarding.*
- *More households have dogs as pets than cats, but cat lovers are more likely to have multiple pets.*
- *Americans are eating more peanuts and drinking less coffee.*
- *More Americans belong to a fantasy sports league (10.6 million) than to book clubs (5.7 million).*
- *Book club members are outnumbered by avid bird-watchers (5.8 million).*

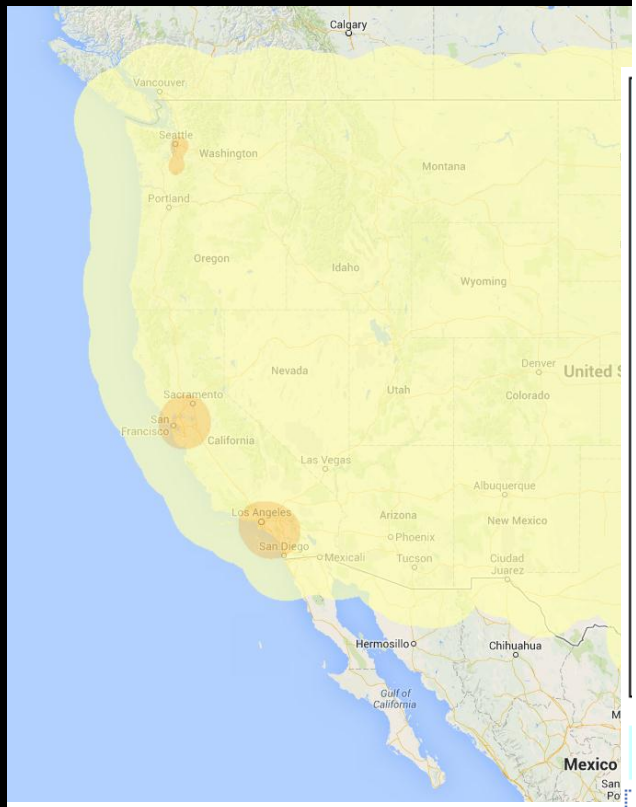
Facts from the Statistical Abstract of the United States, 2013
http://www.nytimes.com/2012/12/20/us/statistical-abstract-paints-america-by-numbers.html?_r=1&

Bookland from University of Arizona



Gansner, E.R.; Yifan Hu; Kobourov, S., "GMap: Visualizing graphs and clusters as maps," *Pacific Visualization Symposium (PacificVis), 2010 IEEE* , vol., no., pp.201,208, 2-5 March 2010

Stephen-land (U



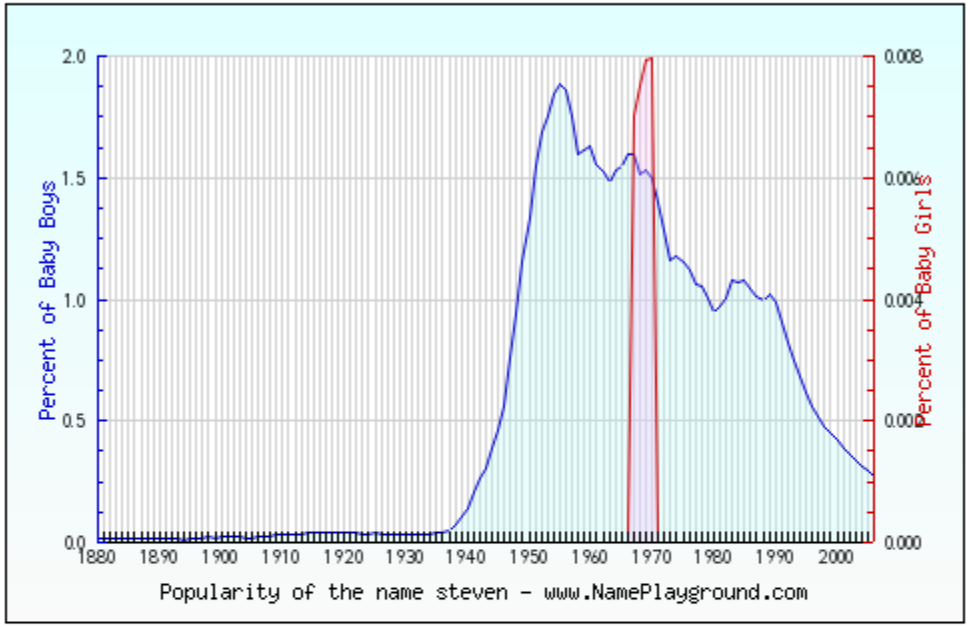
Average Mean Income: $82,708.72

Average Median Income: $65,595.40

$0

Percentage Of Babies Named Steven

Note: Only names in the Top 1000 for each year are represented. [source] Values are normalized for boys and girls to compare trends (vertical scales differ for boys and girls).

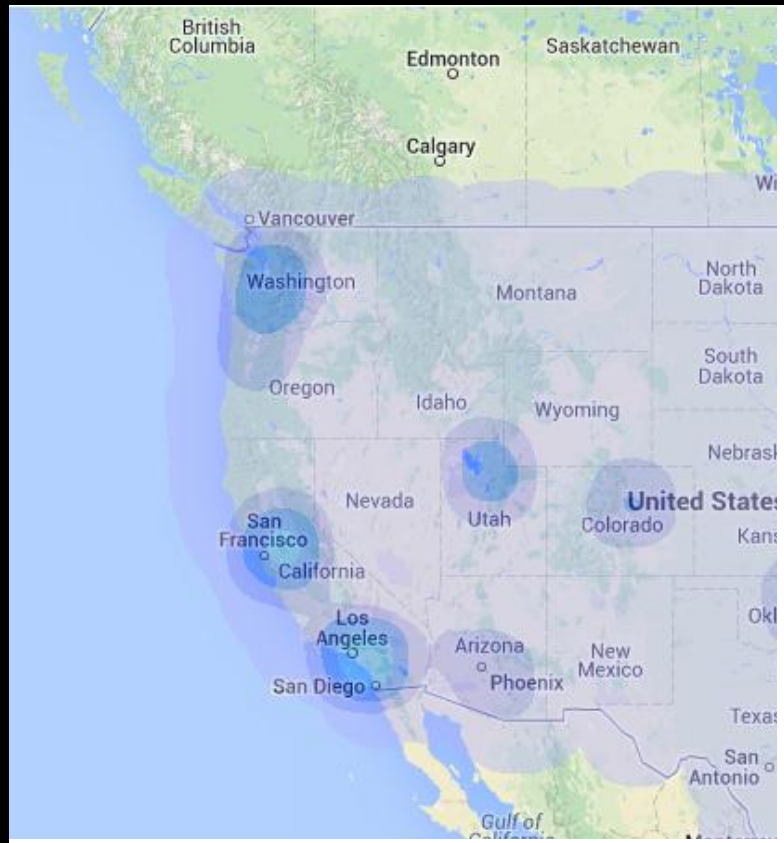


Age Distribution of Steven

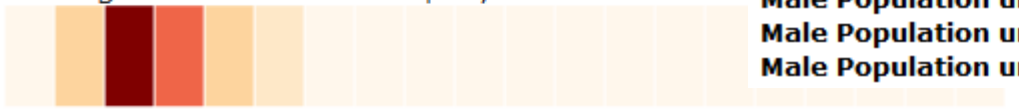
About these totals.

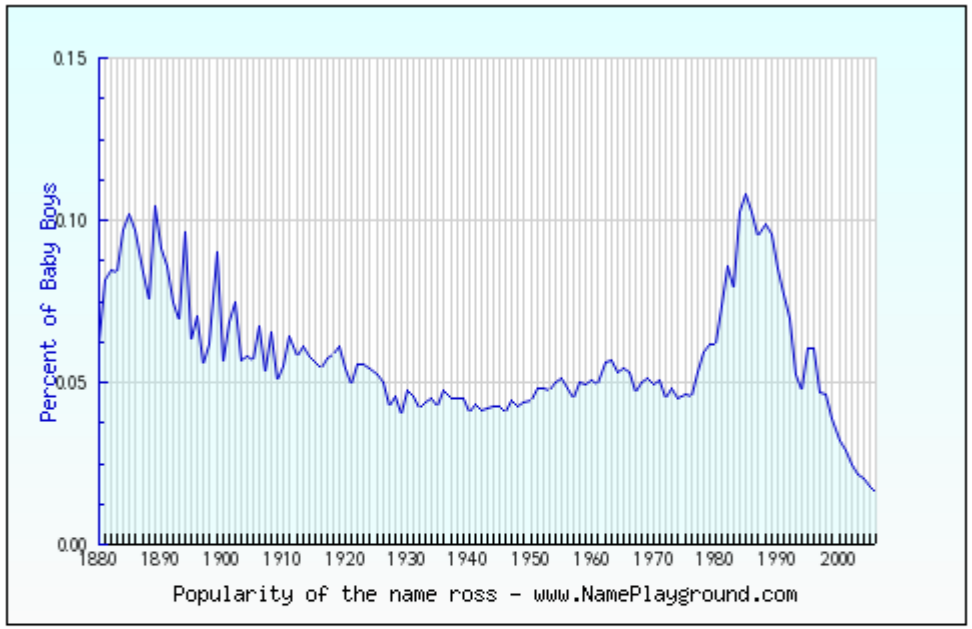
Male Population under 60: 1033434 (83.08%)
Male Population under 50: 691345 (55.58%)
Male Population under 40: 439530 (35.34%)
Male Population under 30: 271310 (21.81%)
Male Population under 20: 97747 (7.86%)
Male Population under 10: 14996 (1.21%)
Female Population under 60: 468 (0.04%)
Female Population under 50: 468 (0.04%)
Female Population under 40: 0 (0.00%)
Female Population under 30: 0 (0.00%)
Female Population under 20: 0 (0.00%)

Ross-Land



Average Mean Income: $75,790.81
Average Median Income: $59,905.57

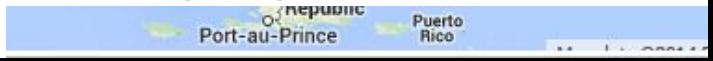




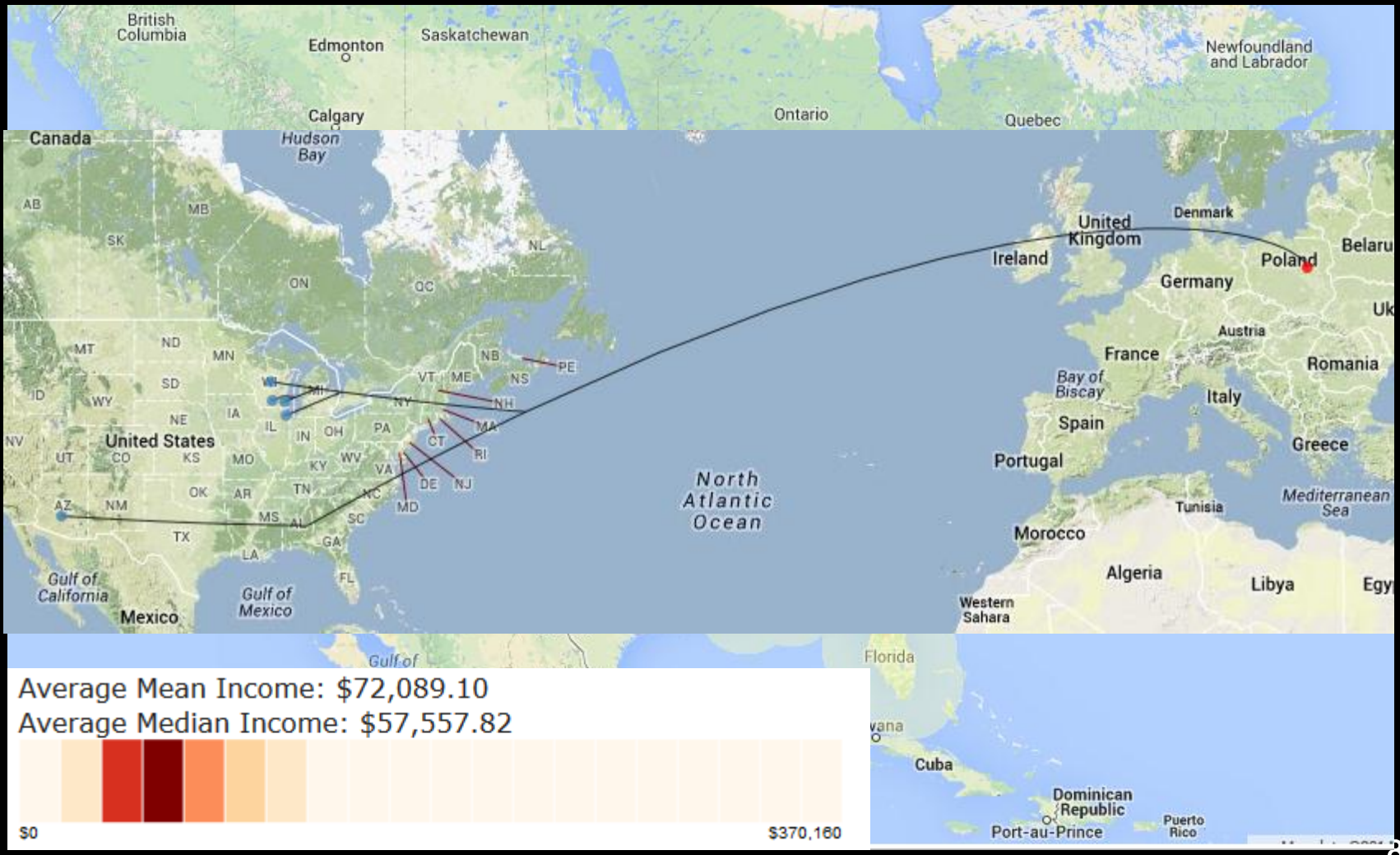
Age Distribution of Ross

About these totals.

Male Population under 60: 52653 (67.19%)
Male Population under 50: 42305 (53.98%)
Male Population under 40: 33703 (43.01%)
Male Population under 30: 23847 (30.43%)
Male Population under 20: 7960 (10.16%)
Male Population under 10: 925 (1.18%)



Maciejewski-Land (USA)



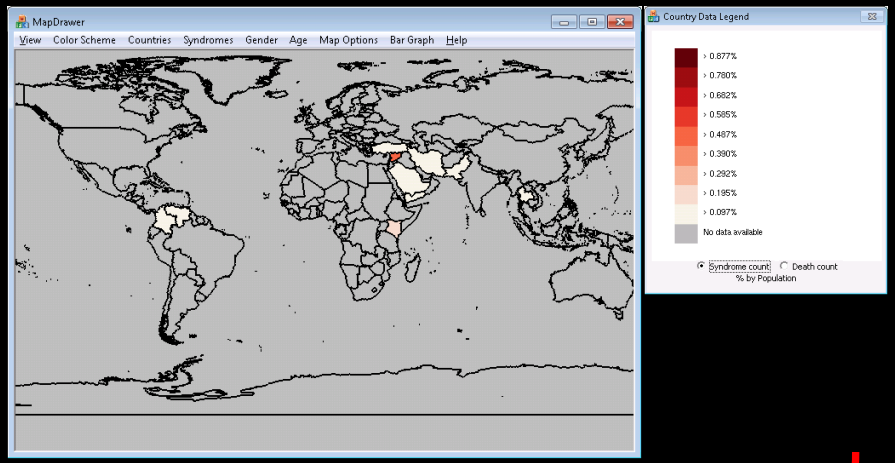
Yifan Zhang, Adnan Muhammad, Paul A. Longley, Ross Maciejewski. Exploring Geo-Genealogy Using Internet Search Histories. *Journal of Maps* (To appear).

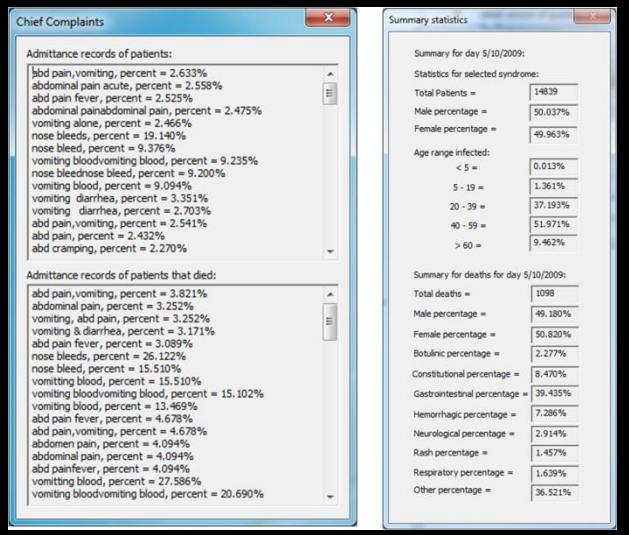
What Was All of That Data?

- *We can mine the white pages (78.4M phone-name-address pairs)*
- *We can link to secondary data sources (census, baby-names)*
- *We can share data to get more data! (Who is likely to search for Ross Maciejewski?)*

Syndromic Surveillance

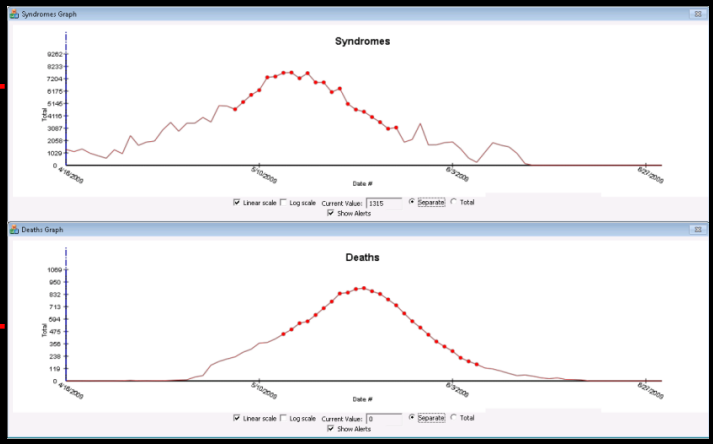
- Syndromic surveillance is the detection of adverse health events focusing on pre-diagnosis information to improve response time
- Pre-diagnosis information can consist of multiple data sources:
  - *Over the counter medicine sales*
  - *News reports on emerging diseases*
  - *Pro-med news feeds*
  - *Emergency department chief complaints*



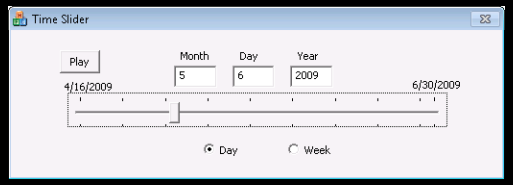


Patient ID	Date	Chief Complaints	Location
9398	4/16/09	Ear pain	Karachi
10816	4/16/09	Stuffy nose	Lebanon
1491	4/16/09	Fever	Allepo
16237	4/16/09	Head bleed	Yemen

CoCo Classifier



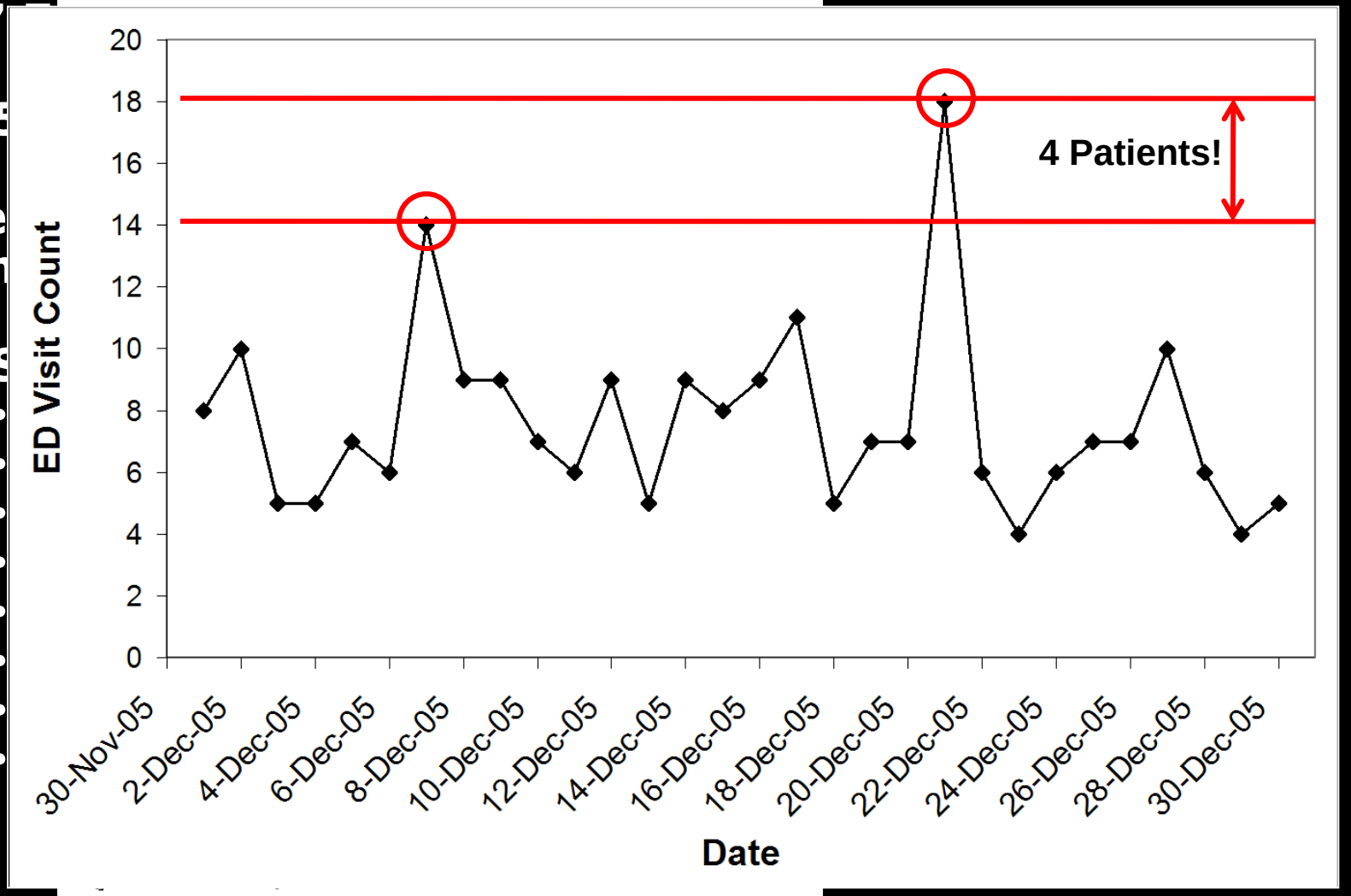
1) Respiratory
2) Gastro-intestinal
3) Neurological
4) Botulinic
5) Constitutional
6) Rash
7) Hemmoraghic
8) Other



VAST 2010 Challenge Award: Support for Future Detection, IEEE Conference on Visual Analytics Science and Technology, October 2010 for "VACCINATED," (Malik, A., Afzal, S., Hodgess, E., Ebert, D., Maciejewski, R.).

Indiana's Public Health Emergency Surveillance System (PHESS)

- 77
- 6,5
- Re
com
- 8 S



Improving Syndromic Surveillance

Interactive visual analytic environment for effective syndromic surveillance and response

- System designed based on collaboration and feedback with state epidemiologists
- Integrated temporal, geospatial, multi-source, multi-scale analytic capability
- Density estimation for data exploration
- Syndromic control charts for temporal alerts
- Demographic filter controls for advanced analysis

Temporal Analysis Methods

For temporal data, we can find anomalies using control chart methods

Control charts consist of a statistic representing some measurement in time (number of patients in a hospital, value of a stock, etc.)

The mean and standard deviation of the statistic is calculated given all the available samples

If the current value is greater than some pre-set number of standard deviations from the mean, then an alert is generated (Shewart suggested three standard deviations)

Control Chart Methods

So, why can't we just find all the alerts for all categories?

For syndromic analysis

- There are hundreds of categorical time signals
- Usually >100 alerts generated per day
- An analyst can process ~10 alerts per day with current tools
- Current solution is simply to set thresholds such that only 10 alerts are generated per day

So, we need a visual analytics pipeline to better incorporate domain expert knowledge. If we take only the 10 worst alerts we may miss the needle in the haystack.

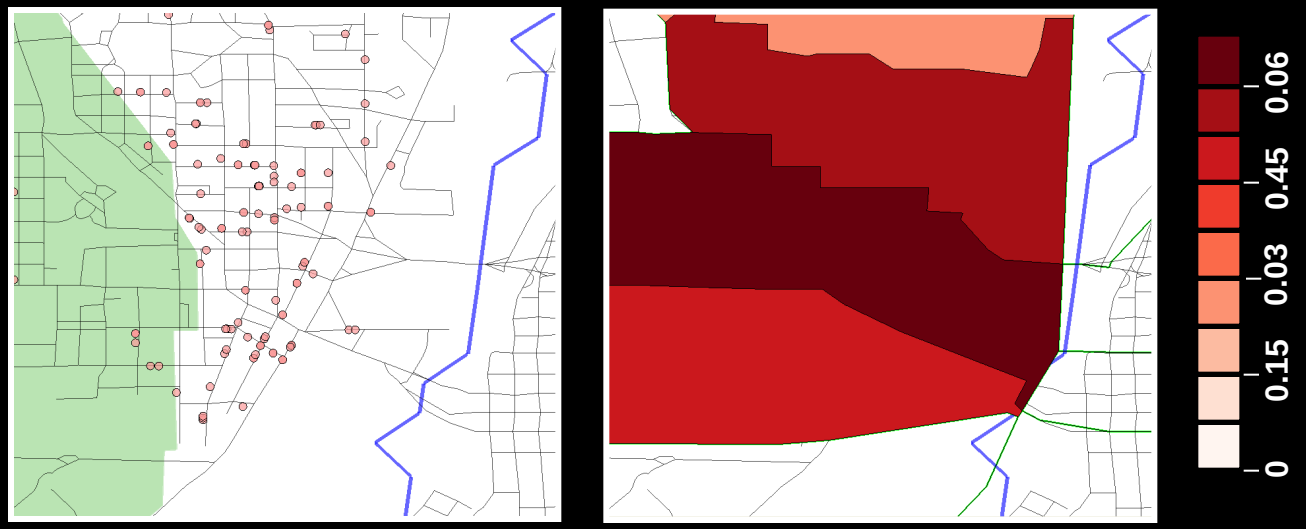
Geospatial Analysis Methods

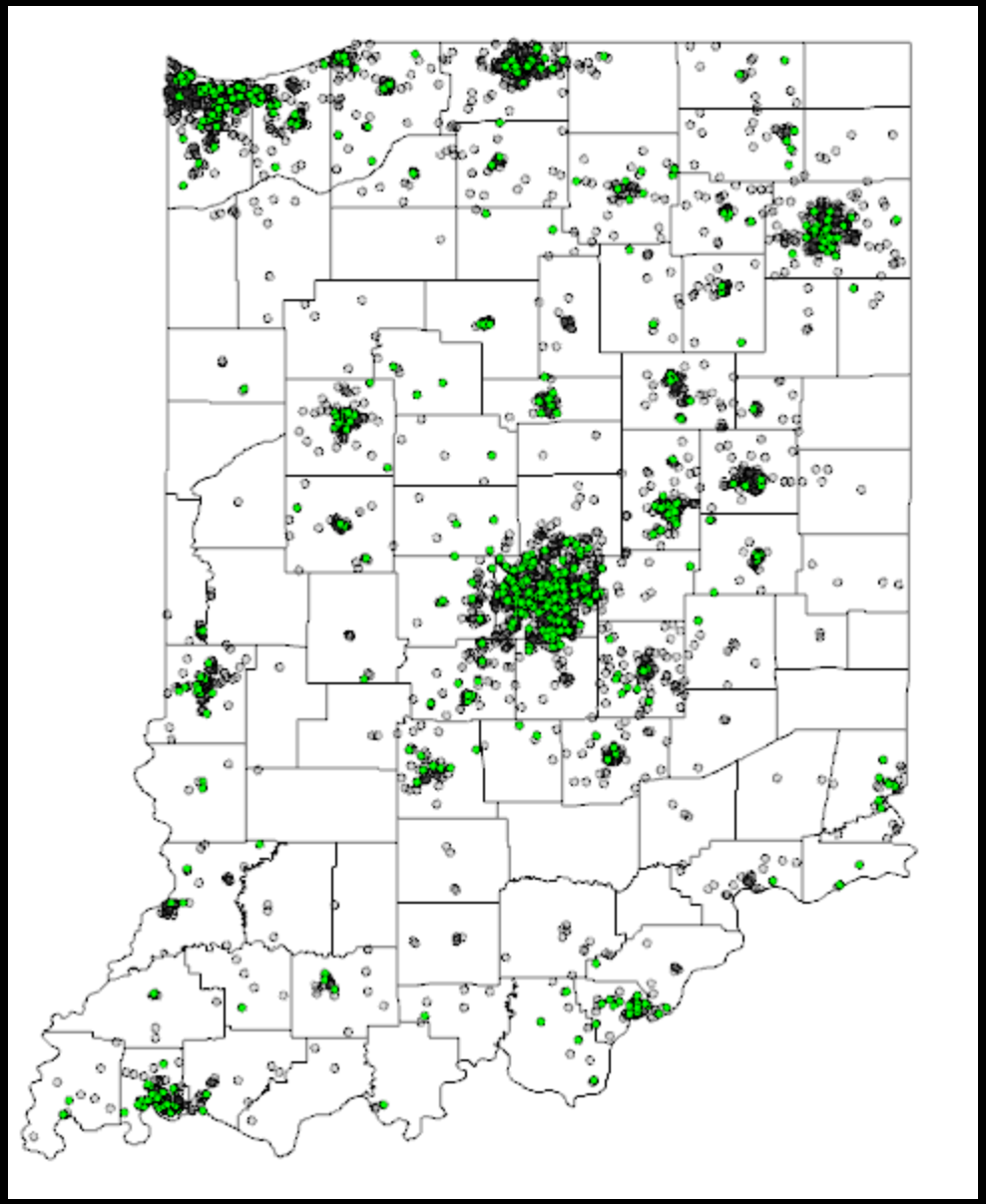
Hotspot

- *An area in which an unexpectedly large amount of syndromes are occurring in space and time*

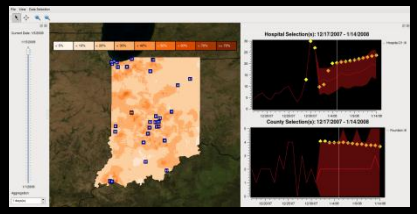
How to display geospatial data?

- *Individual locations give away private information*
- *Aggregation by geographic boundary*





Kernel Density Estimation



$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N} K\left(\frac{x - X_i}{\min(h, d_{i,k})}\right)$$

N = sample size

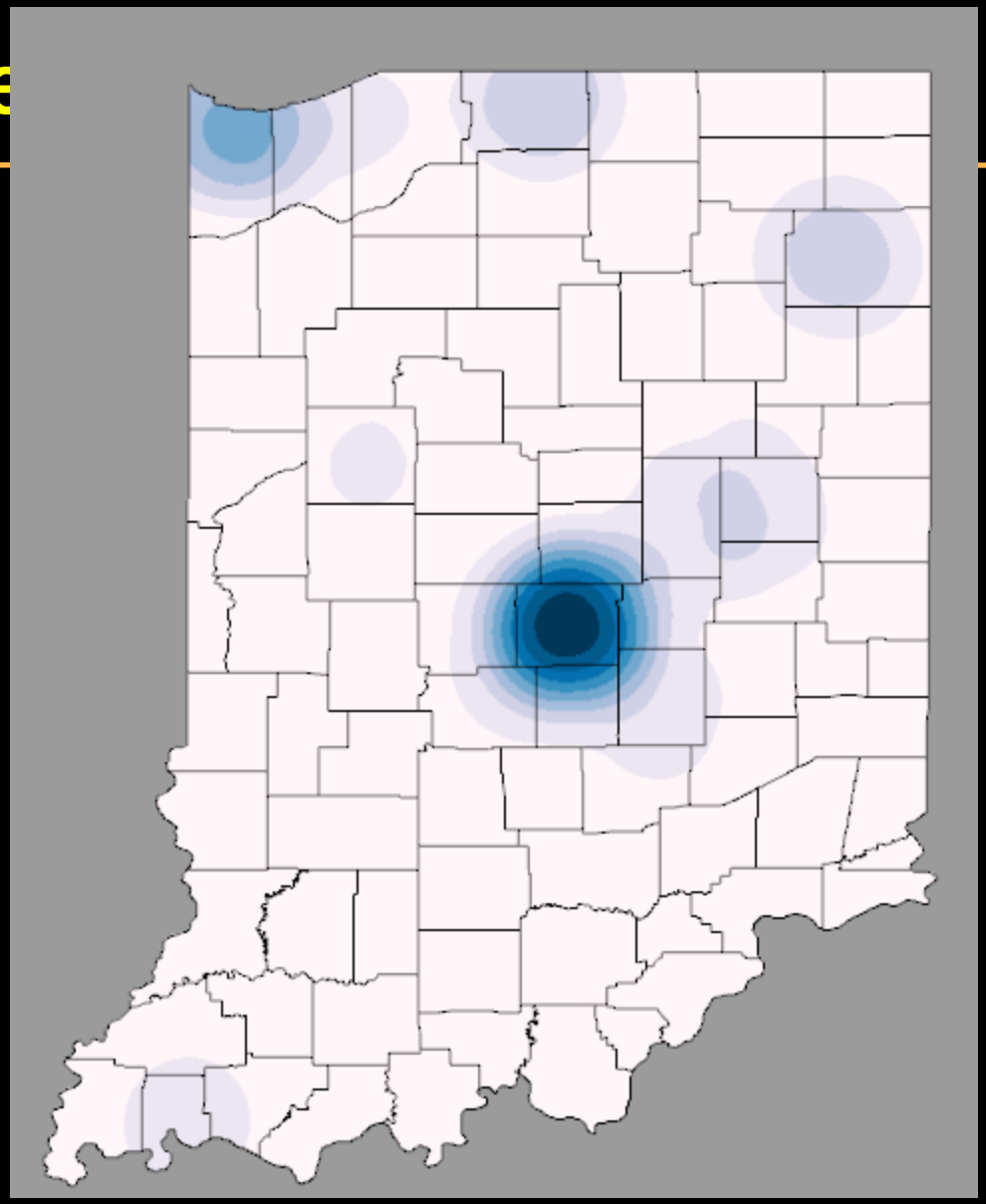
$d_{i,k}$ = distance from i^{th} point to k^{th} nearest neighbor

h = maximum bandwidth

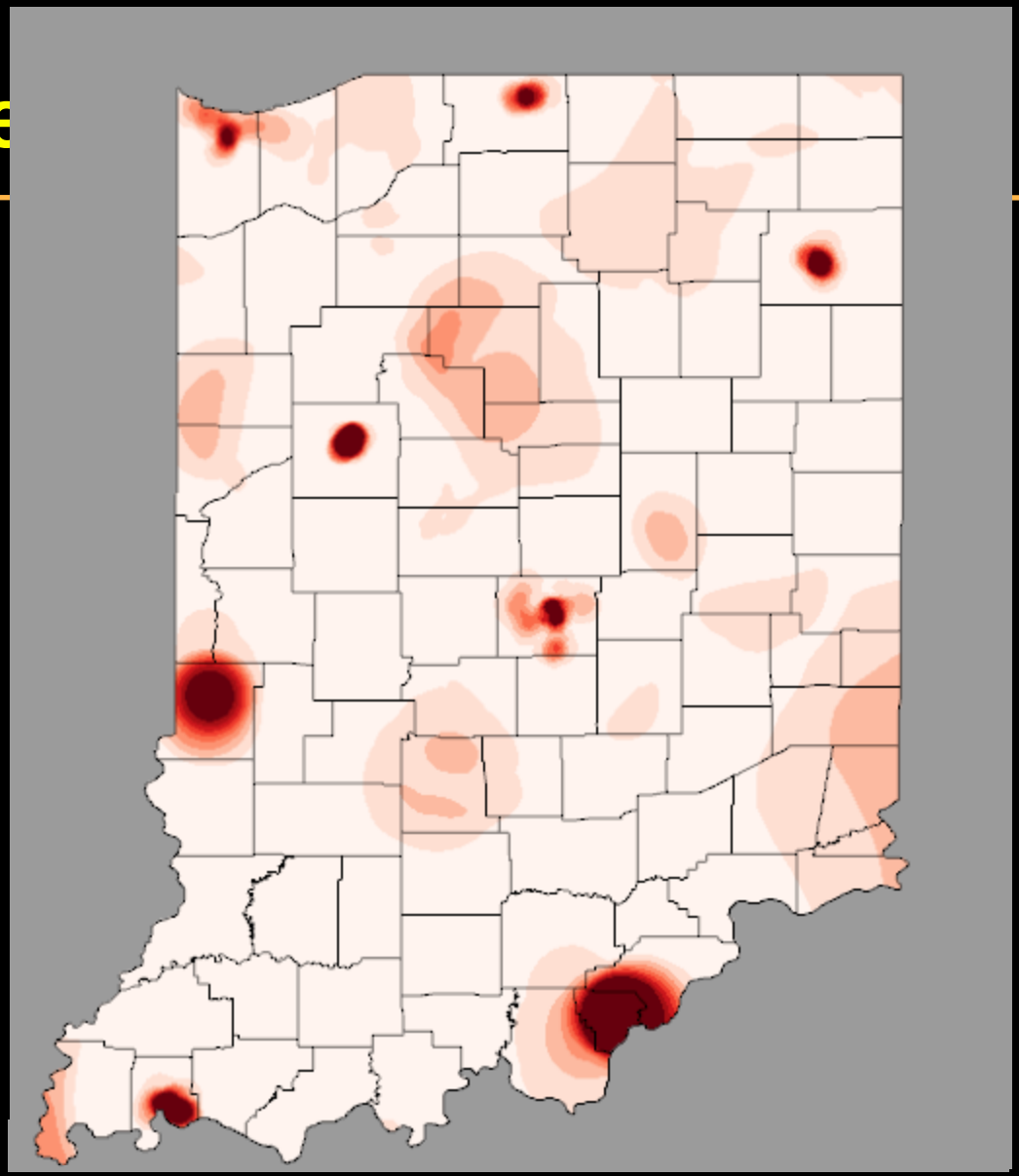
Epanechnikov Kernel

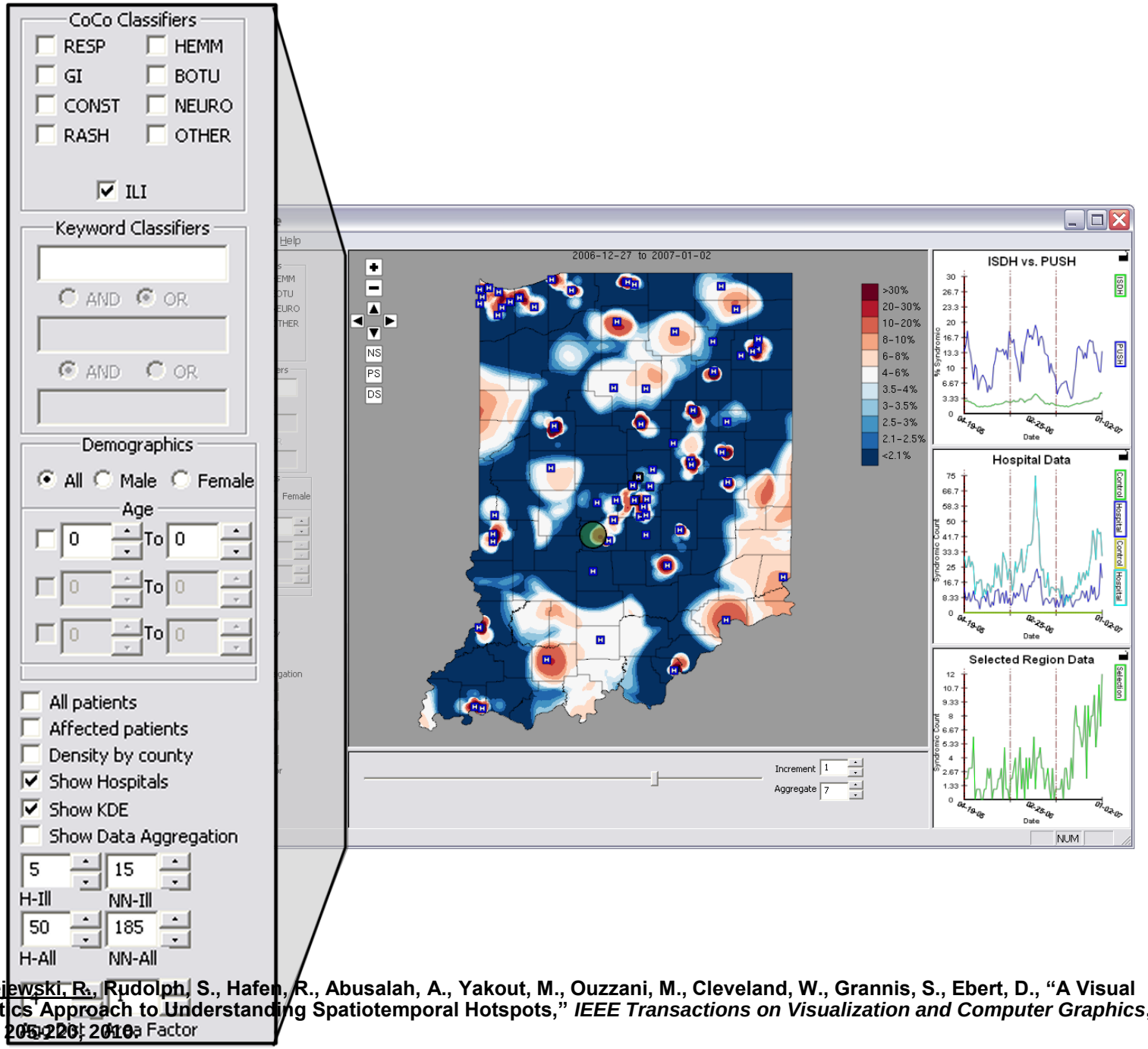
$$K(u) = \frac{3}{4}\left(1 - u^2\right)_{(\|u\| \le 1)}$$

Kernel De



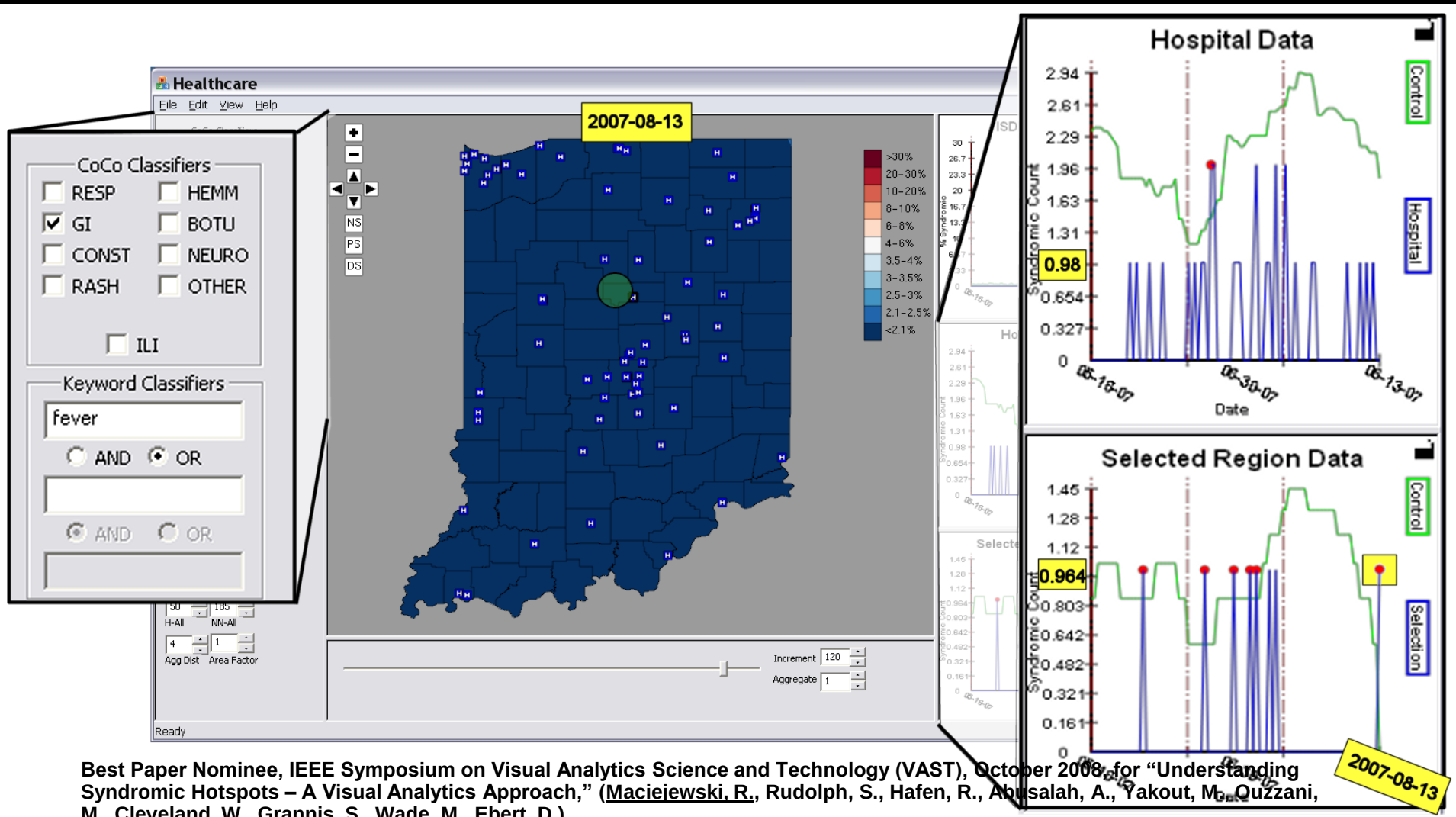
Kernel De



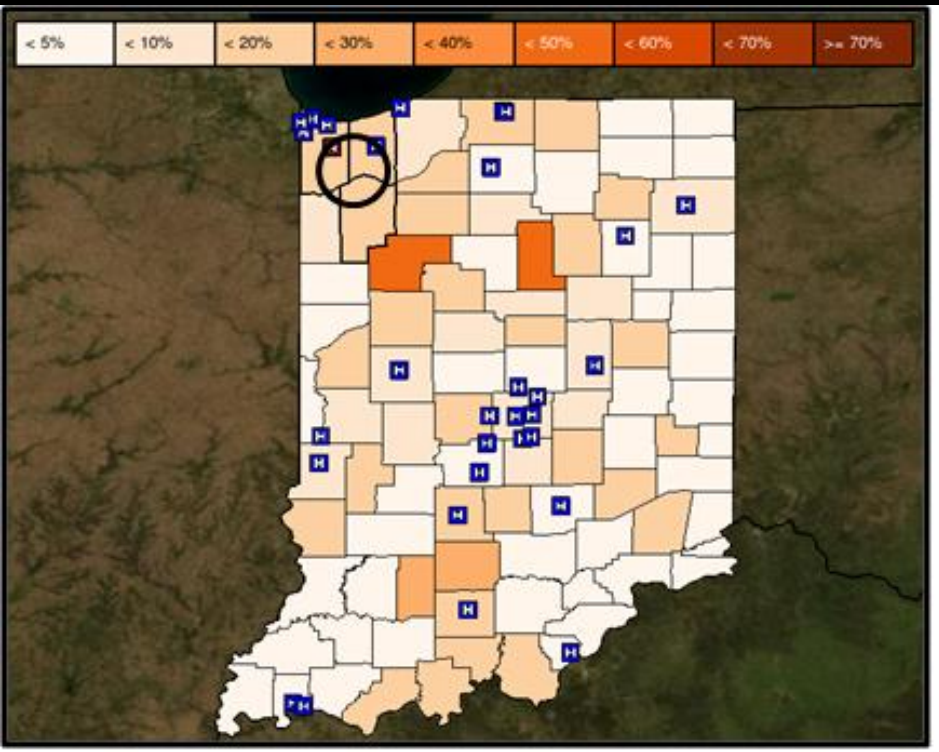


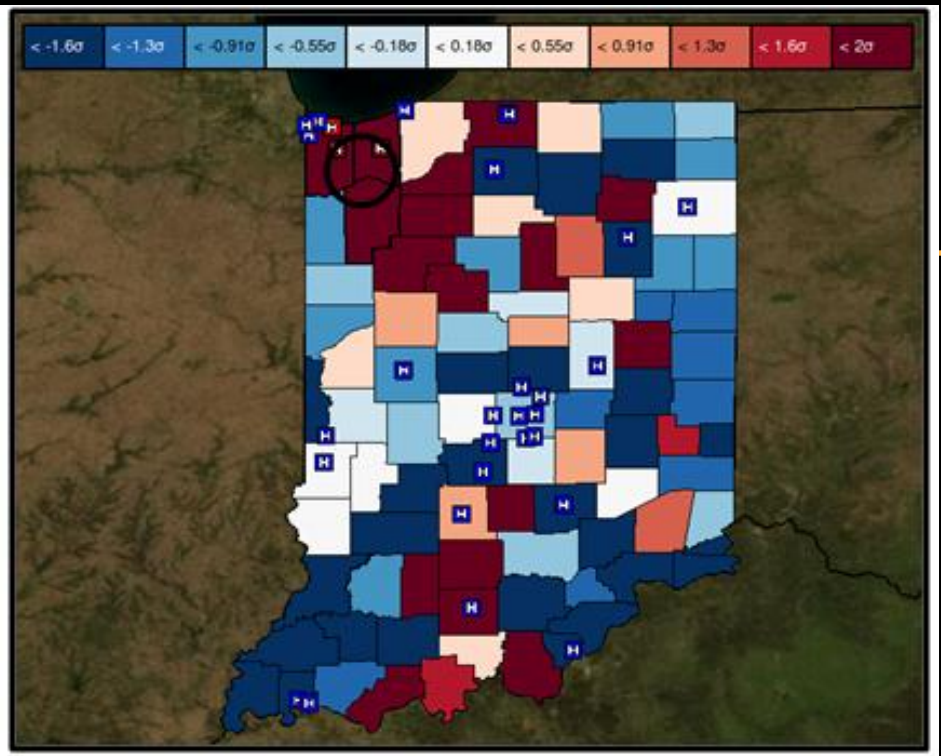
Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W., Grannis, S., Ebert, D., "A Visual Analytics Approach to Understanding Spatiotemporal Hotspots," *IEEE Transactions on Visualization and Computer Graphics*, 16(2): 205-220, 2010.

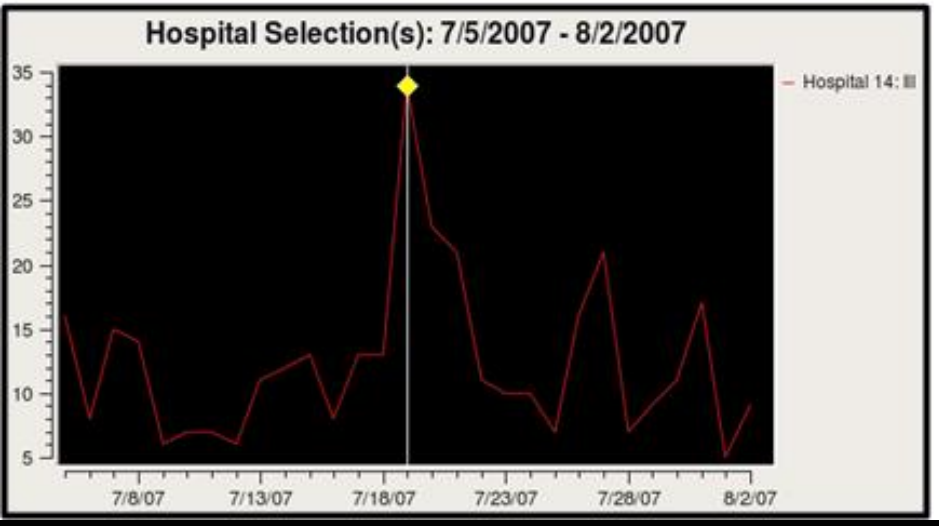
Hypothesis Generation and Exploration

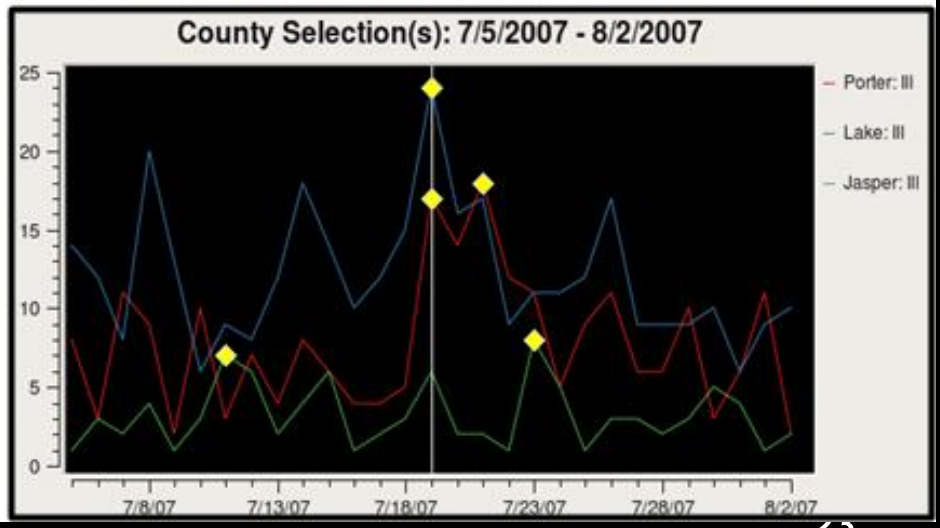


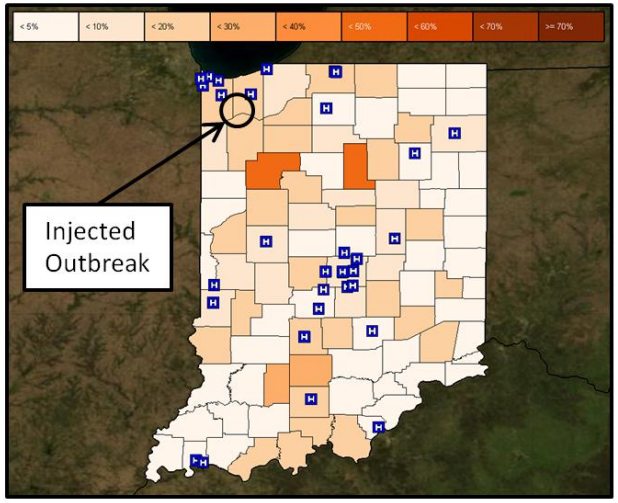
Best Paper Nominee, IEEE Symposium on Visual Analytics Science and Technology (VAST), October 2008, for "Understanding Syndromic Hotspots – A Visual Analytics Approach," (Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W., Grannis, S., Wade, M., Ebert, D.).

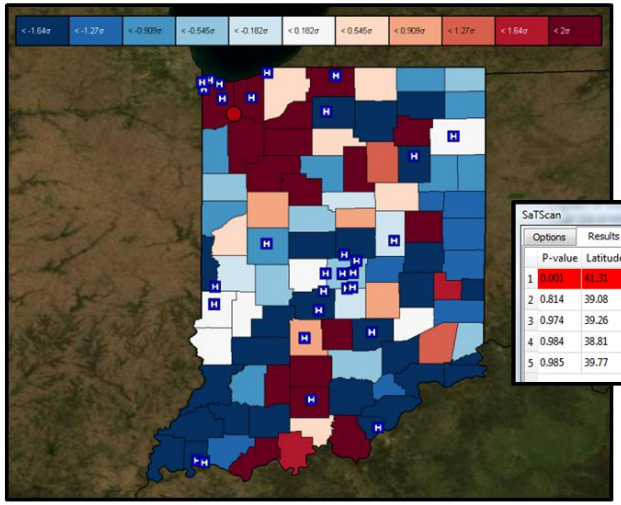

< 5%
< 10%
< 20%
< 30%
< 40%
< 50%
< 60%
< 70%
>= 70%

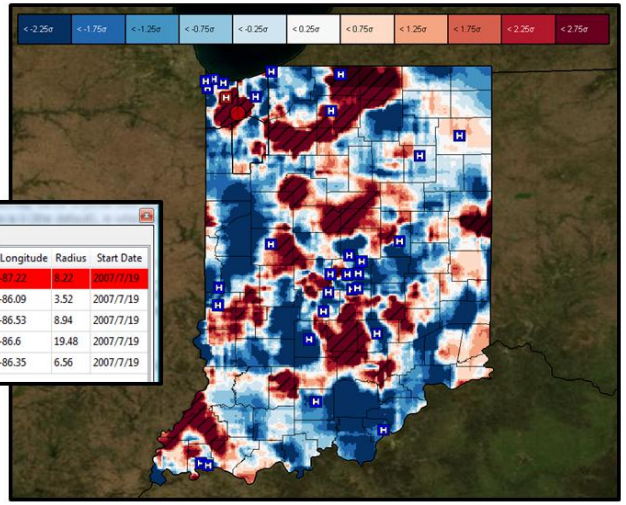

< -1.6σ
< -1.3σ
< -0.91σ
< -0.55σ
< -0.18σ
< 0.18σ
< 0.55σ
< 0.91σ
< 1.3σ
< 1.6σ
< 2σ

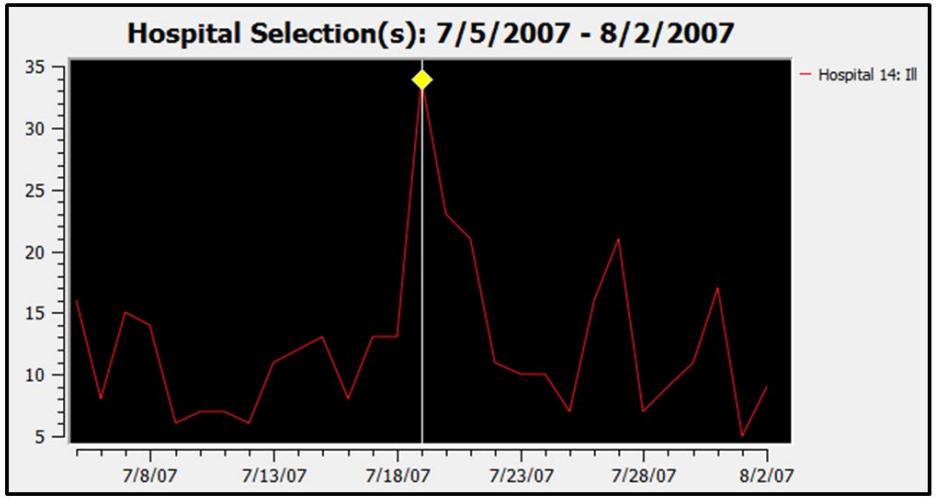

Hospital Selection(s): 7/5/2007 - 8/2/2007
Hospital 14: III
35
30
25
20
15
10
5
7/8/07
7/13/07
7/18/07
7/23/07
7/28/07
8/2/07

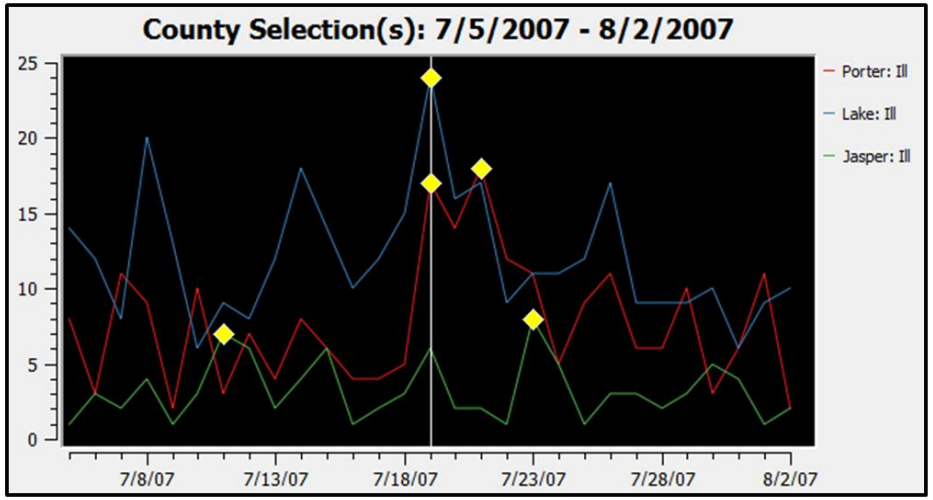

County Selection(s): 7/5/2007 - 8/2/2007
Porter: III
Lake: III
Jasper: III
25
20
15
10
5
0
7/8/07
7/13/07
7/18/07
7/23/07
7/28/07
8/2/07


Injected
Outbreak


SaTScan
Options
Results
P-value Latitude Longitude Radius Start Date
1 0.001 41.31 -87.22 8.22 2007/7/19
2 0.814 39.08 -86.09 3.52 2007/7/19
3 0.974 39.26 -86.53 8.94 2007/7/19
4 0.984 38.81 -86.6 19.48 2007/7/19
5 0.985 39.77 -86.35 6.56 2007/7/19



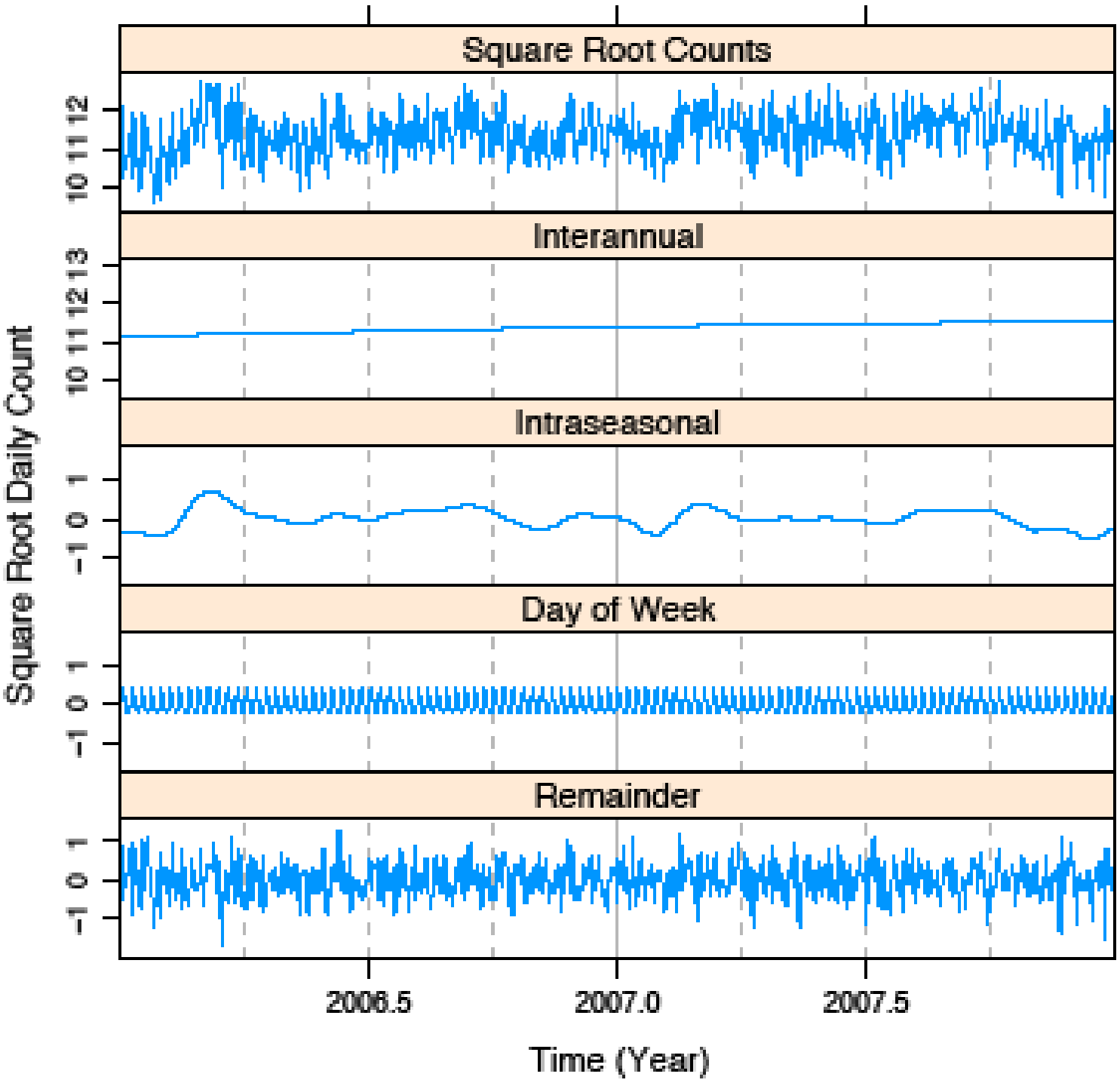

Hospital Selection(s): 7/5/2007 - 8/2/2007
35
30
25
20
15
10
5
7/8/07
7/13/07
7/18/07
7/23/07
7/28/07
8/2/07
Hospital 14: Ill


County Selection(s): 7/5/2007 - 8/2/2007
25
20
15
10
5
0
7/8/07
7/13/07
7/18/07
7/23/07
7/28/07
8/2/07
Porter: Ill
Lake: Ill
Jasper: Ill

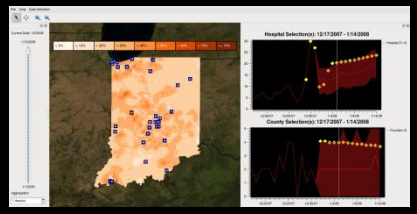
Tim

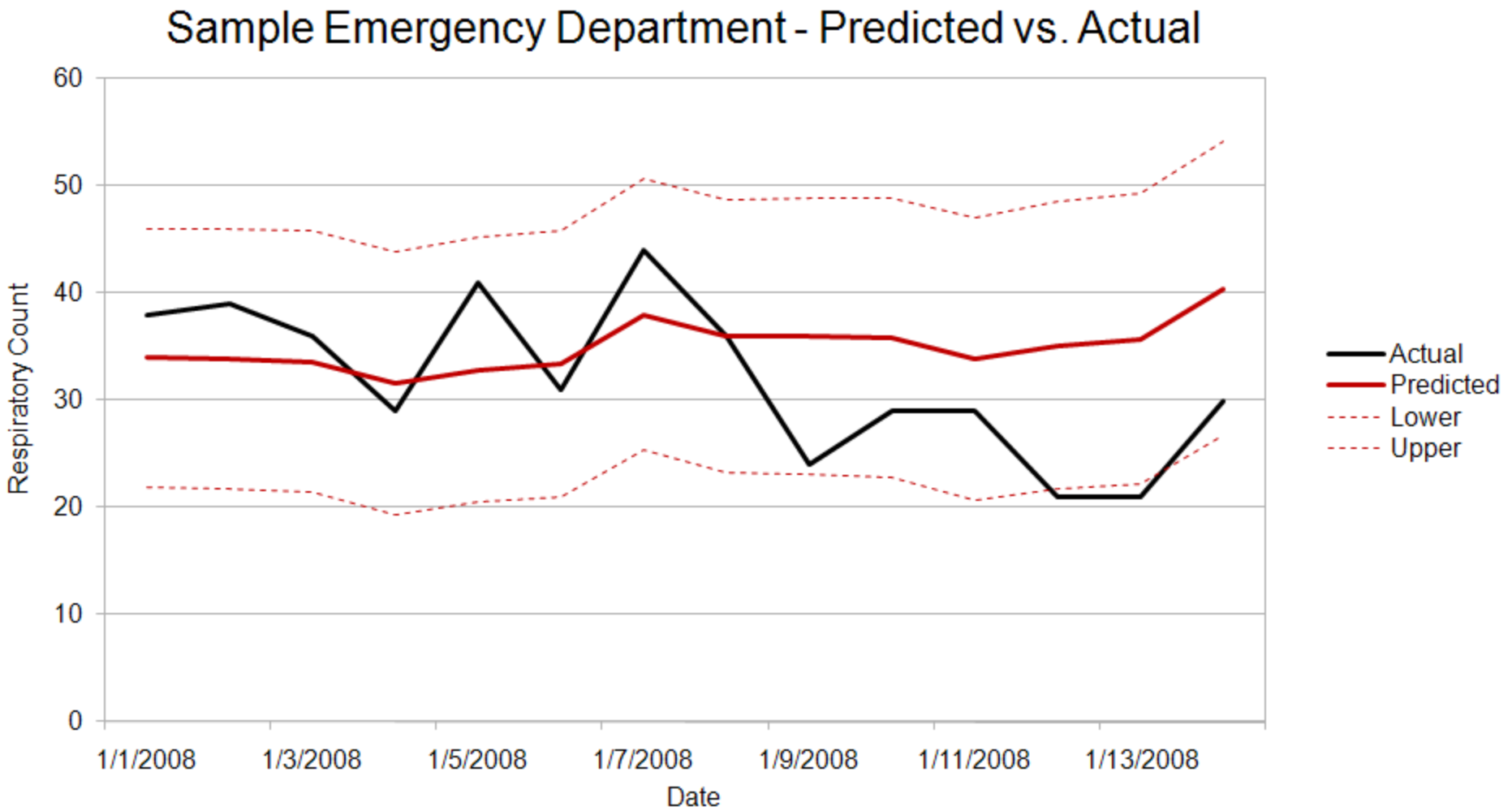
Seas

- T
- F
- O


Square Root Counts
Interannual
Intraseasonal
Day of Week
Remainder
Square Root Daily Count
10 11 12
10 11 12 13
−1 0 1
−1 0 1
−1 0 1
2006.5
2007.0
2007.5
Time (Year)

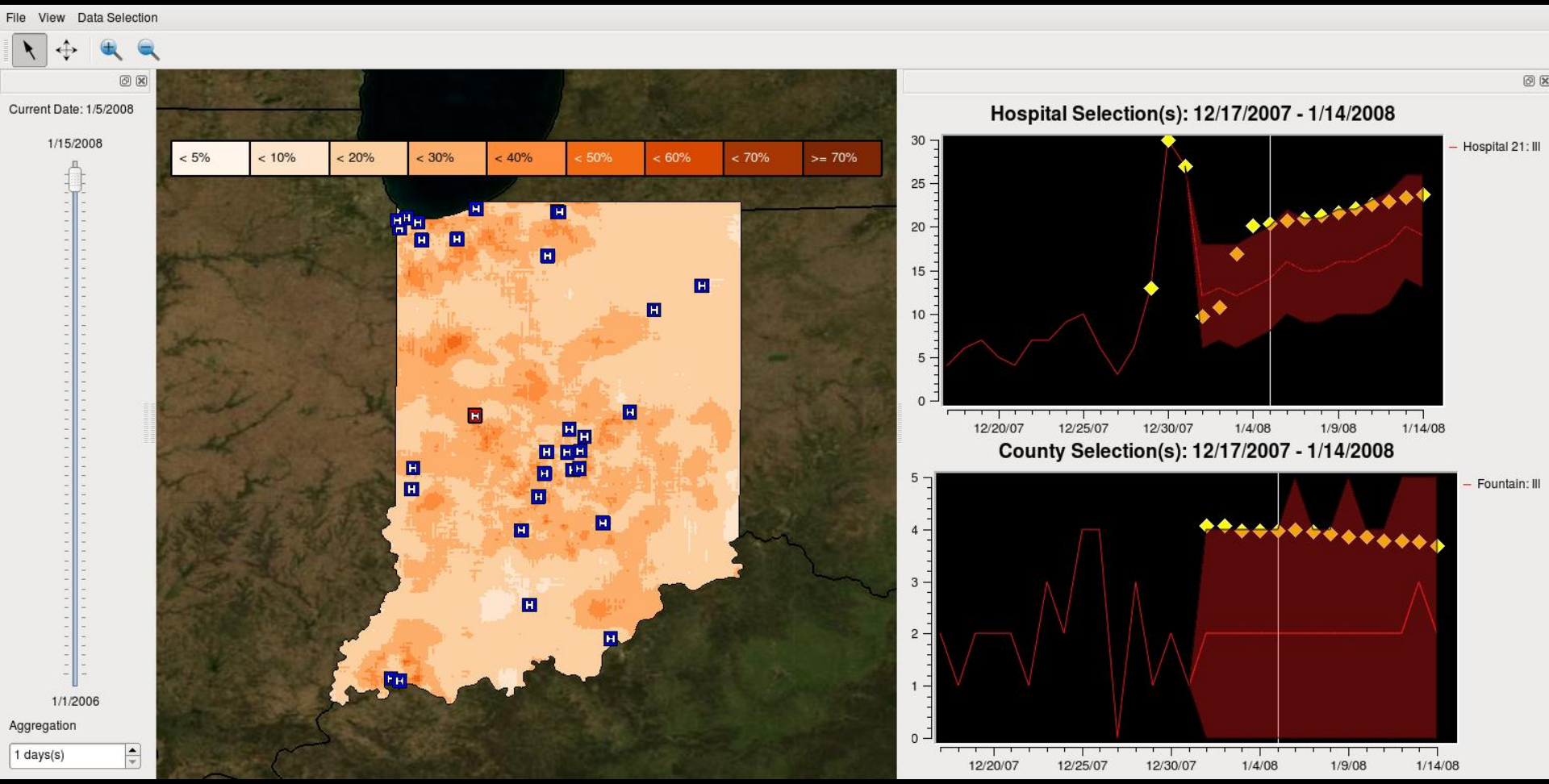
Predictive Visual Analytics





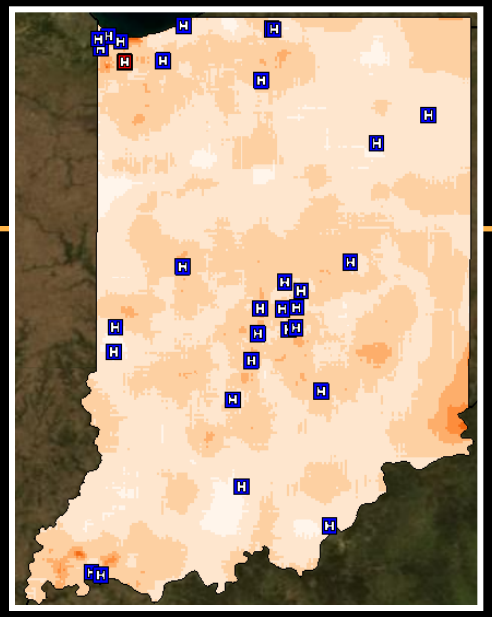
Predictive Visual Analytics

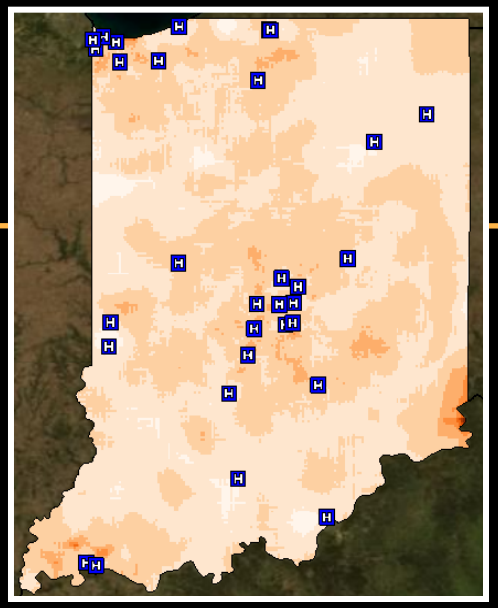


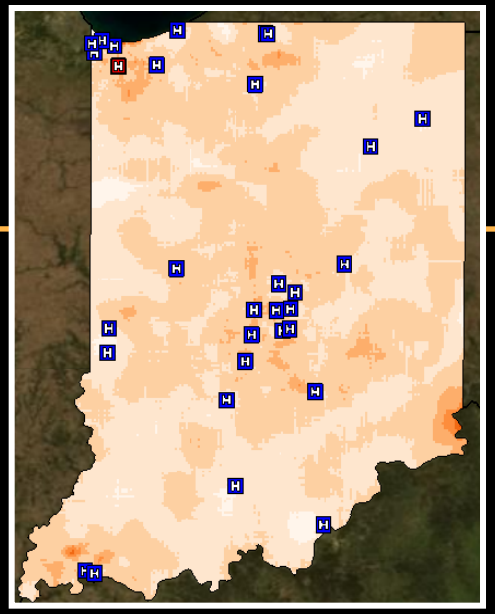


Maciejewski, R., Rudolph, S., Hafen, R., Mitchell, M. A., Larew, S. G., Cleveland, W., Ebert, D., "Forecasting Hotspot – A Predictive Analytics Approach," *IEEE Transactions on Visualization and Computer Graphics*, 17(4): 440-453, 2011.

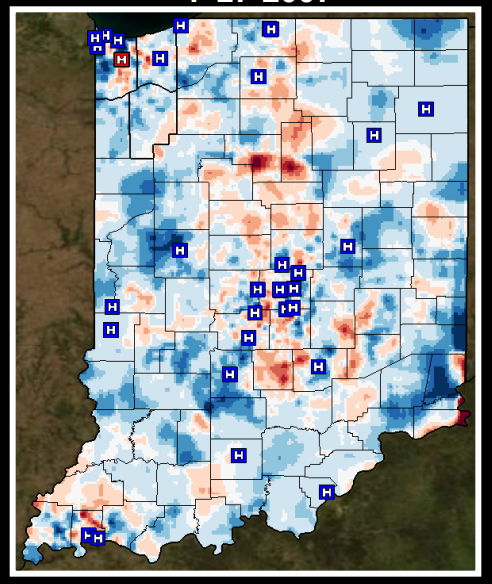
< 5%	< 10%	< 20%	< 30%	< 40%	< 50%	< 60%	< 70%	>= 70%



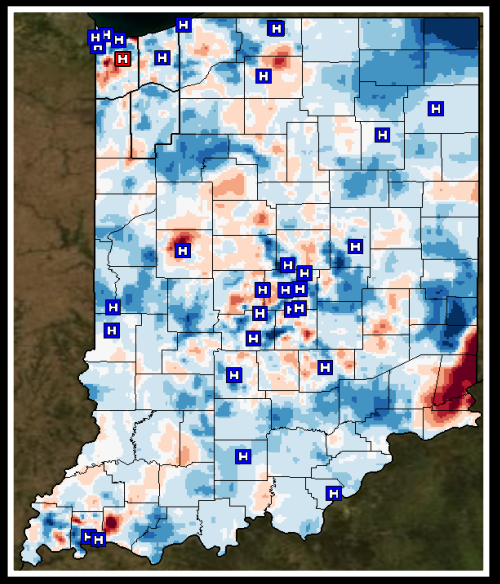




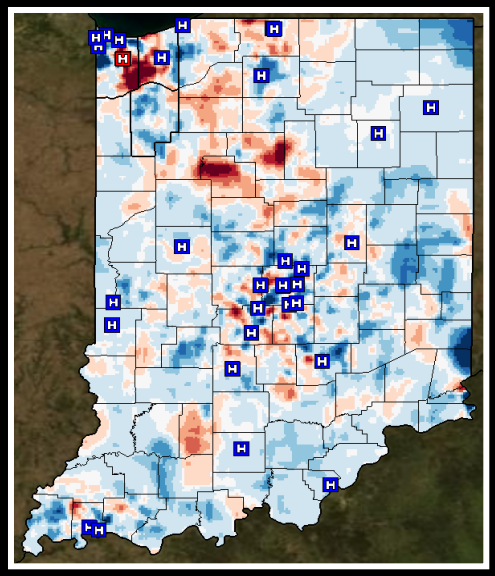
7-17-2007



7-18-2007



7-19-2007



< -8σ	< -6σ	< -4σ	< -2σ	< -1σ	< 1σ	< 2σ	< 4σ	< 6σ	< 8σ	< 10σ

Visual Analytics Law Enforcement Technology (VALET)

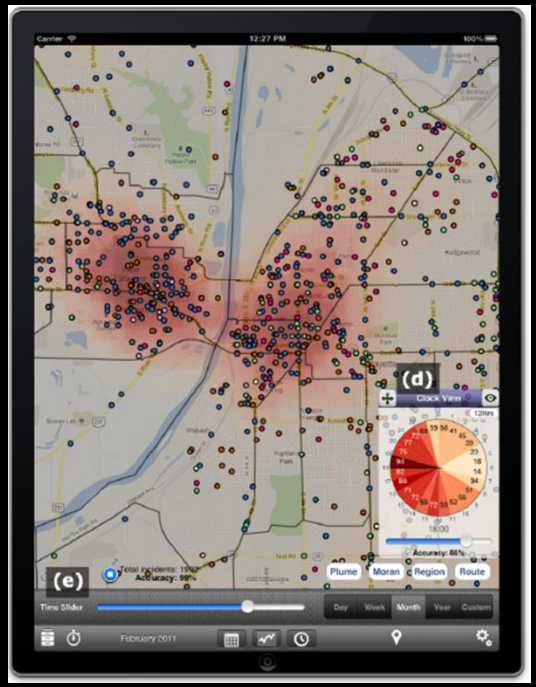
- *VALET was developed in partnership with Purdue University through sponsorship by the Department of Homeland Security*

VALET is designed to ingest any geolocated criminal incident reports

Data Types:
- Shapefiles
- SQL connectivity
- FBI UCR codes

Data Size:
- ~1000 records/day

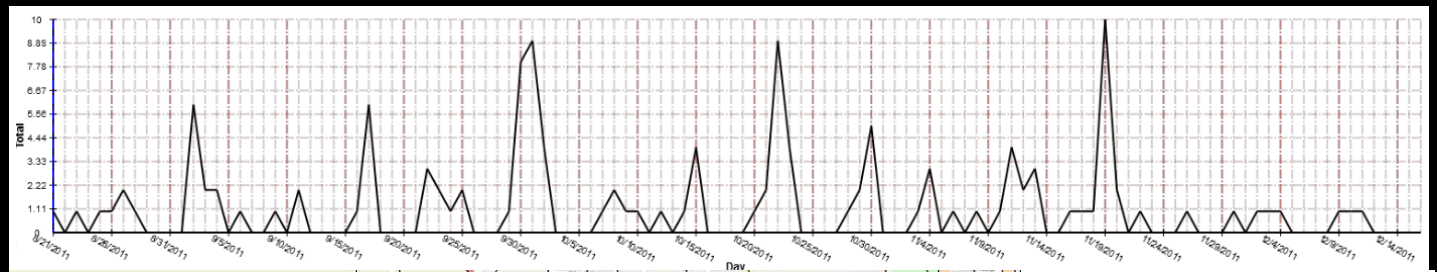
Partners:
- Tippecanoe County Sherriff, IN
- Ohio State Highway Patrol
- New York City Police Department

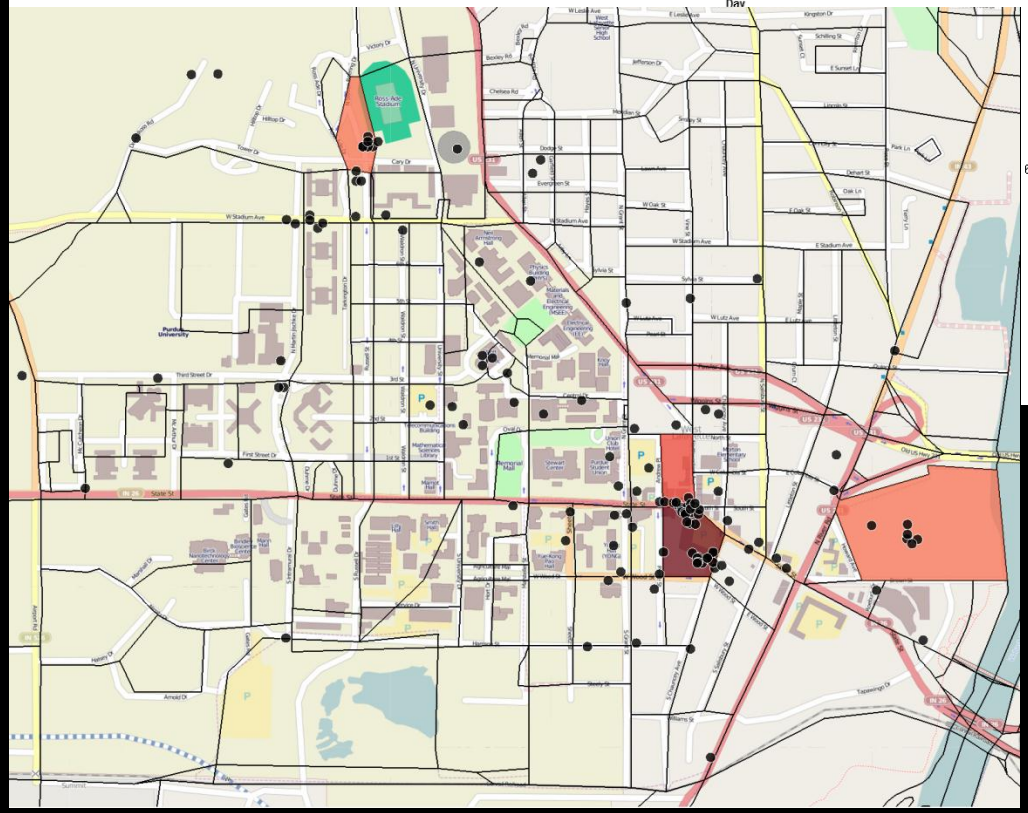


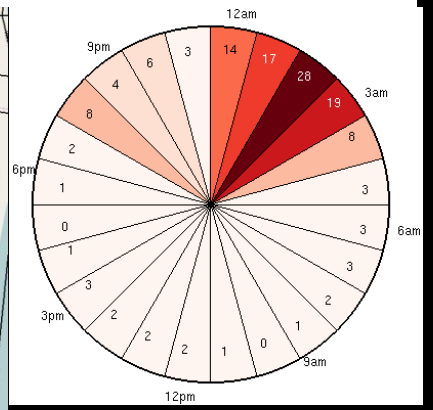
VACCINE
Visual Analytics for Command, Control, and Interoperability Environments
A U.S. Department of Homeland Security Center of Excellence

For More Information Contact:
Dr. Ross Maciejewski
rmacieje@asu.edu
Vader.lab.asu.edu

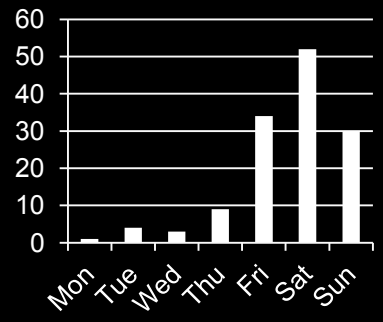
Retrospective Analytics: Drunkenness/Public-Intoxication



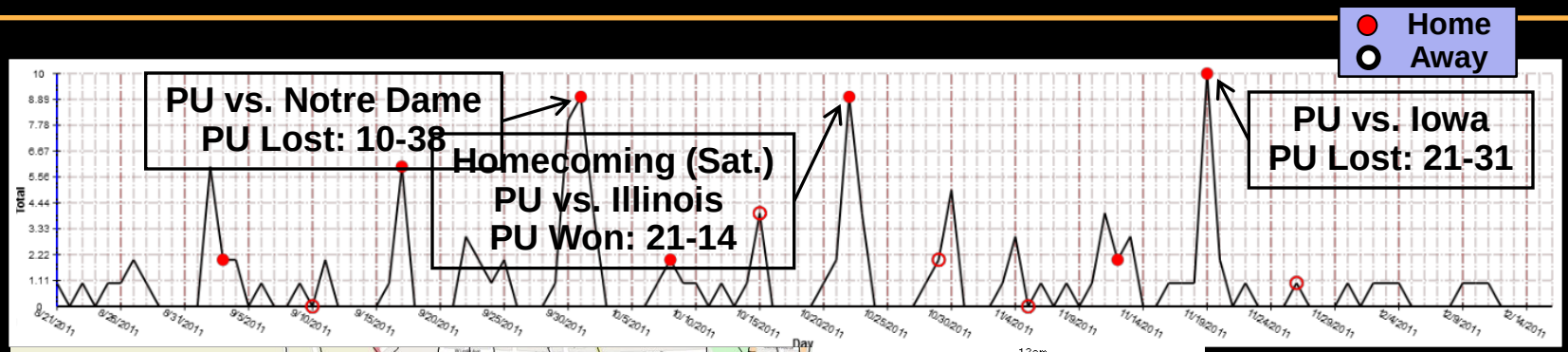


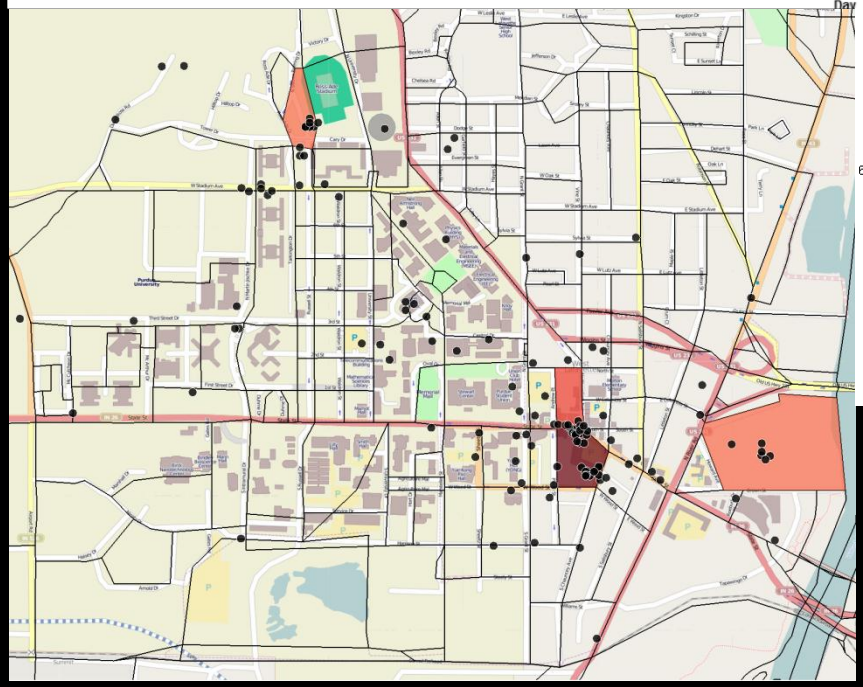


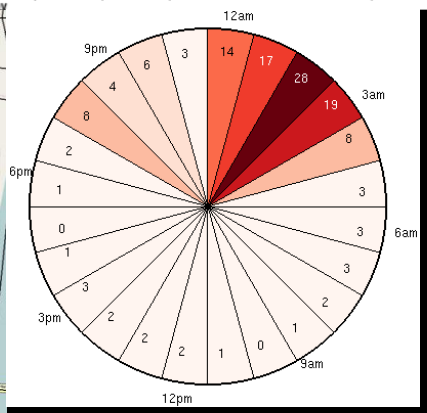
Day-of-the-Week

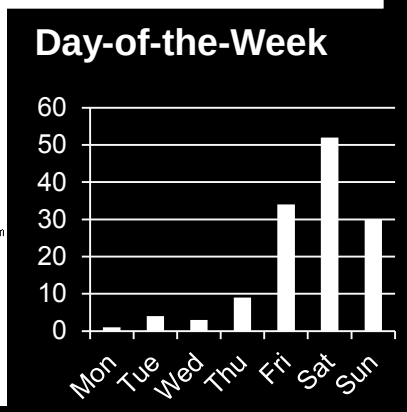


Retrospective Analytics: Drunkenness/Public-Intoxication

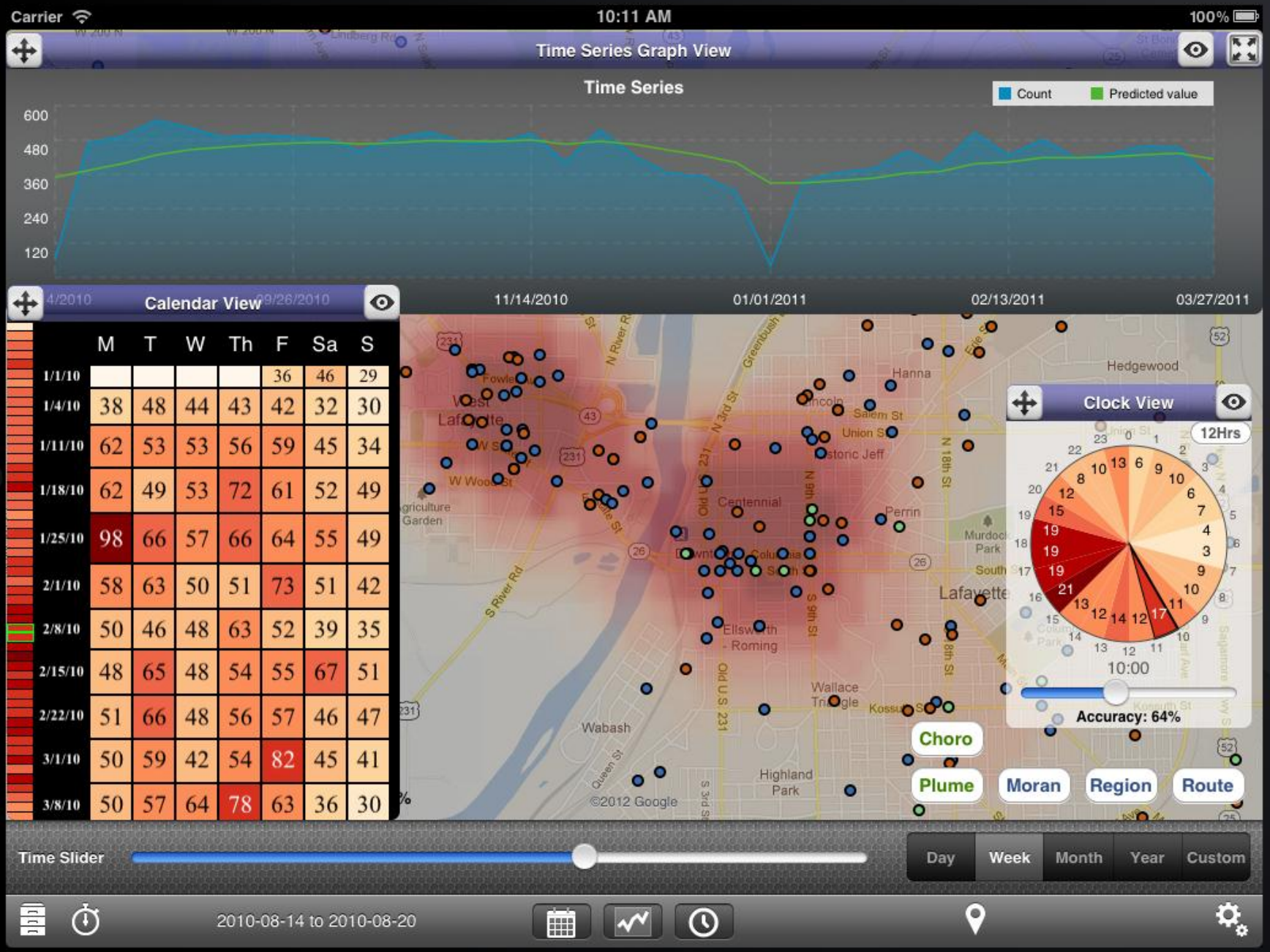








Abish Malik, Ross Maciejewski, Timothy F. Collins and David S. Ebert. Visual Analytics Law Enforcement Toolkit. *IEEE International Conference on Technologies for Homeland Security*, 2010.


Carrier
10:11 AM
100%
Time Series Graph View
Time Series
Count
Predicted value
600
480
360
240
120
Calendar View
11/14/2010
01/01/2011
02/13/2011
03/27/2011
M T W Th F Sa S
1/1/10 36 46 29
1/4/10 38 48 44 43 42 32 30
1/11/10 62 53 53 56 59 45 34
1/18/10 62 49 53 72 61 52 49
1/25/10 98 66 57 66 64 55 49
2/1/10 58 63 50 51 73 51 42
2/8/10 50 46 48 63 52 39 35
2/15/10 48 65 48 54 55 67 51
2/22/10 51 66 48 56 57 46 47
3/1/10 50 59 42 54 82 45 41
3/8/10 50 57 64 78 63 36 30
Clock View
12Hrs
10:00
Accuracy: 64%
Choro
Plume
Moran
Region
Route
©2012 Google
Time Slider
Day
Week
Month
Year
Custom
2010-08-14 to 2010-08-20

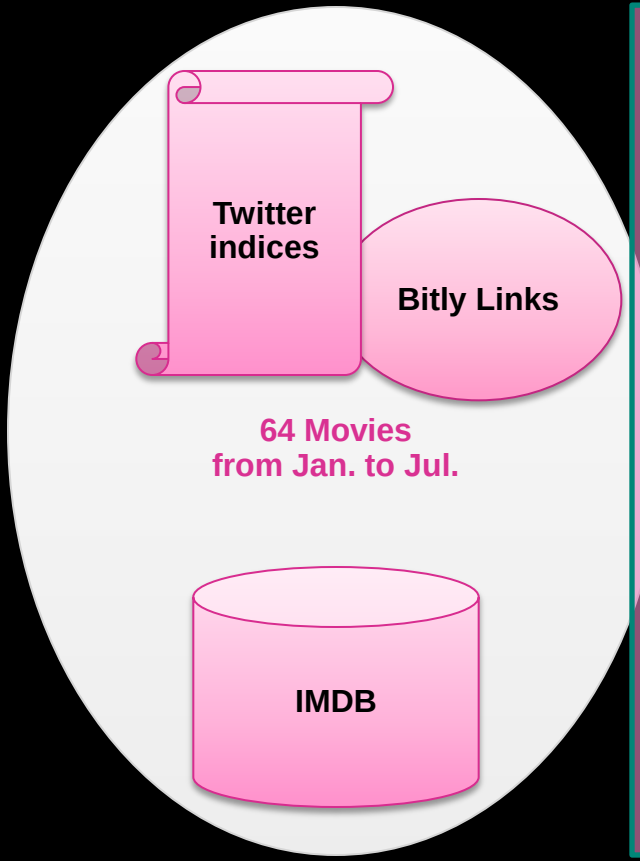
Social Media:Real-time Twitter Monitoring and Integration into Tools

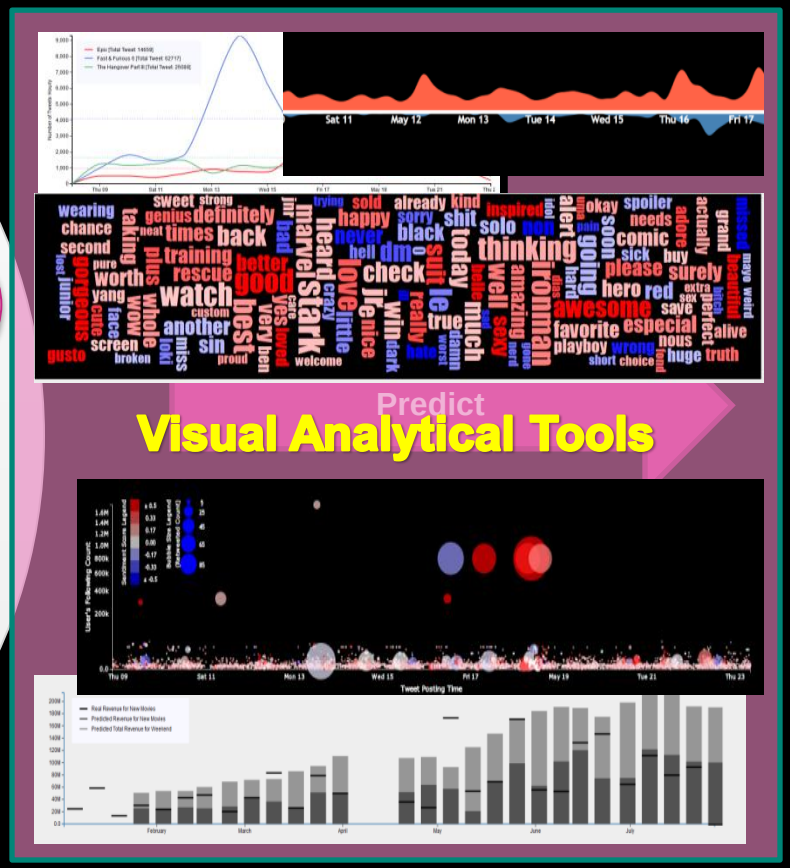
- *Topic extraction using novel STL based remainder estimation technique*
- *Dynamically linked views providing options to monitor emerging / emergent twitter feeds*
- *Topics extracted shown as a dynamic word cloud*



Junghoon Chae, Denis Thom, Harold Bosch, Yun Jang, Ross Maciejewski, David S. Ebert, Thomas Ertl. Spatiotemporal Social Media Analytics for Abnormal Event Detection Using Seasonal-Trend Decomposition. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012

VAST Challenge Description

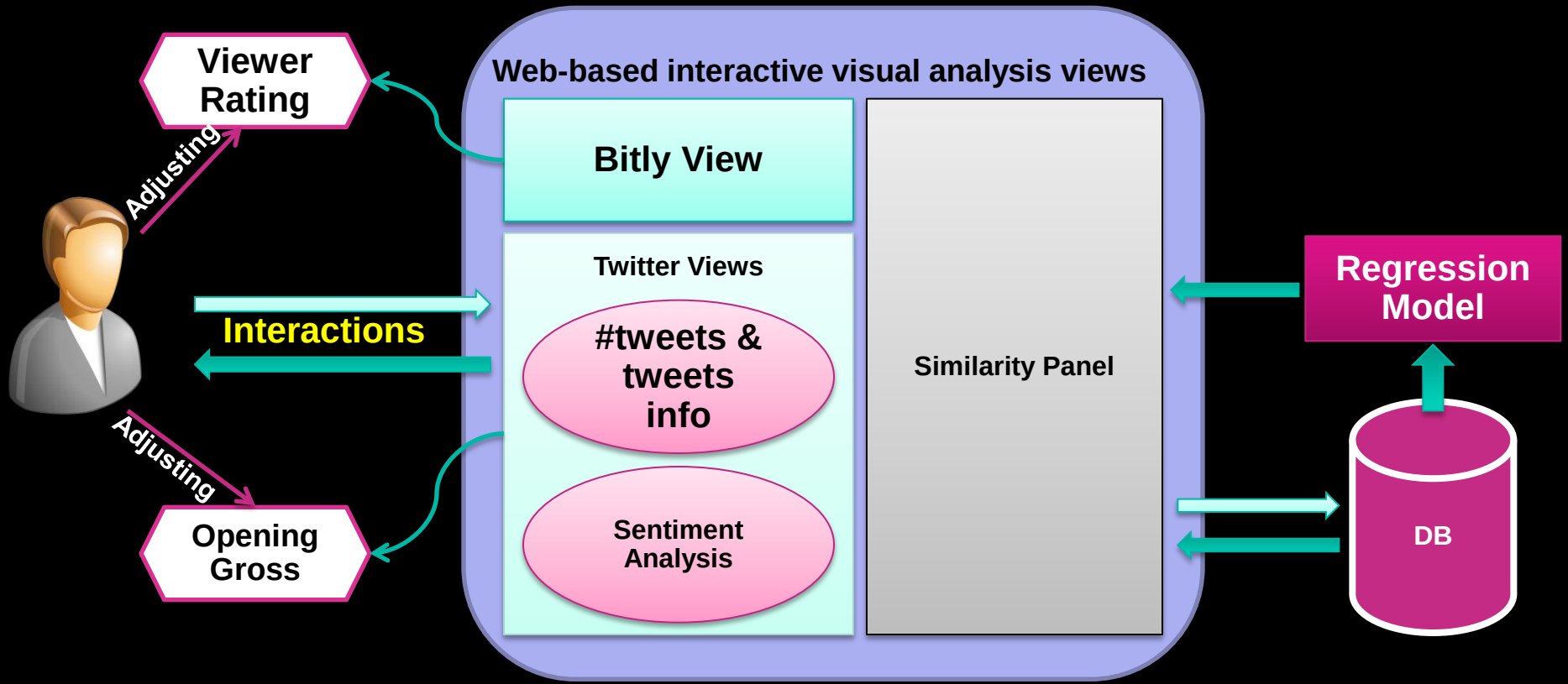




Opening Weekend Gross

Viewer Rating

Overview of Our System



DM2 and Lone Ranger OW Prediction

Total weekend revenue

Individual box-office take

- Multiple linear regression model
- Adjustment
  - *Exclude non-related tweets*
  - *Correct miscalculated sentiment scores*
- Similarity





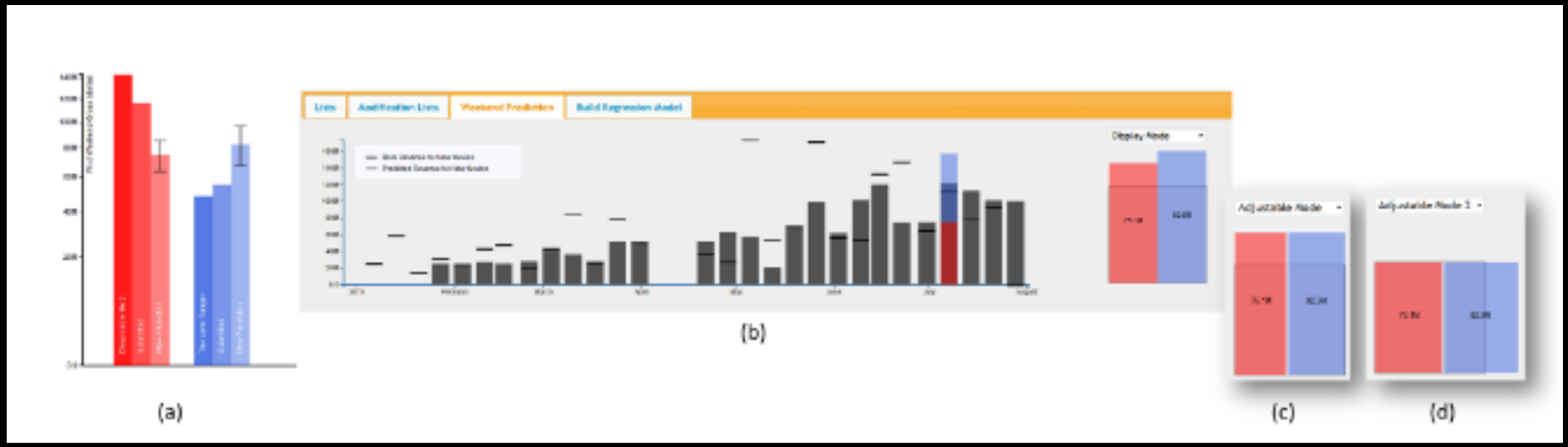
Viewer rating prediction

Prediction finalization

DM2 and Lone Ranger OW Prediction

Two movies released on same weekend ---- total gross

- 3-day weekend total gross prediction: $121M
- 5-day long weekend: ~$175M



DM2 and Lone Ranger OW Prediction

Individual movie OW box-office prediction --- - Multiple Linear Regression Model

Response

Regressors

Variable	Description
MPAA rating	G, PG, PG-13, R, NC-17
Genre	Animation, Action, Thriller, Comedy, etc.
OW	3-day Opening Weekend Gross
TBD	Average daily #tweets
Budget	Approximate movie budget from IMDB
TSS	Tweet sentiment score
MSS	Movie sentiment score
MSP	Movie Star Power

DM2 and Lone Ranger OW Prediction

Individual movie OW box-office prediction --- - Multiple Linear Regression Model

$$OW = \beta_0 + \beta_1 TBD + \beta_2 Budget$$

$$\beta_0 \approx 4.9 \times 10^3$$
$$\beta_1 \approx 4462$$
$$\beta_2 \approx 0.23$$
$$adj - R^2 \approx 0.60$$
$$P - value < 0.05$$

- Lone Ranger $85M +/- $13M
- DM2 $76M +/- $13M

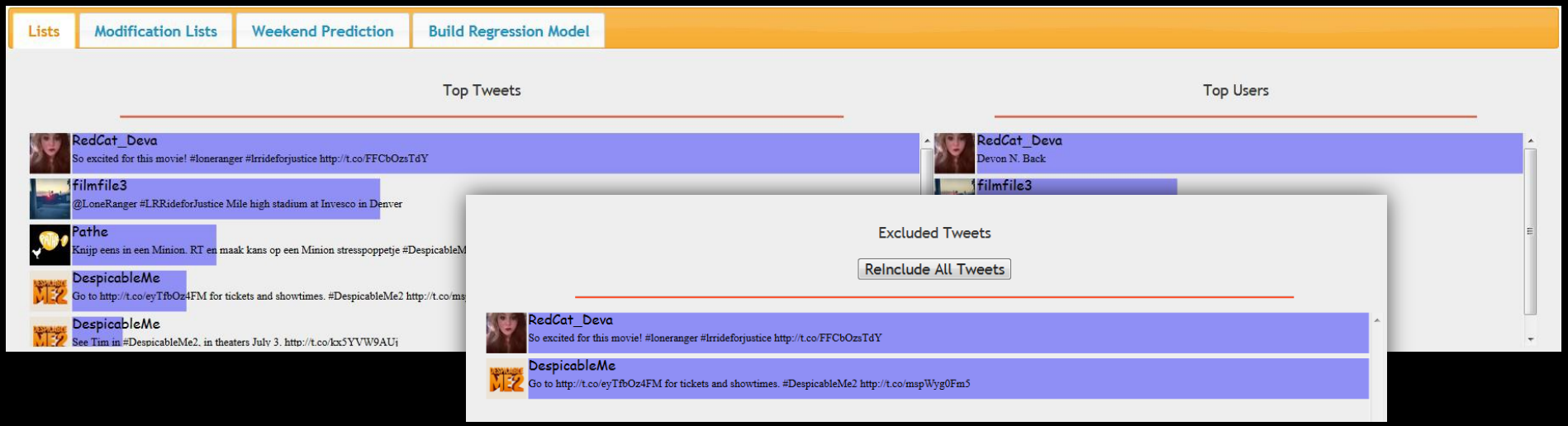
From January to July, each week model parameters were updated with newly collected data.

- Exclude non-related tweets
- Correct miscalculated sentiment scores

DM2 and Lone Ranger OW Prediction

Individual movie OW box-office prediction --- - Model Adjustment

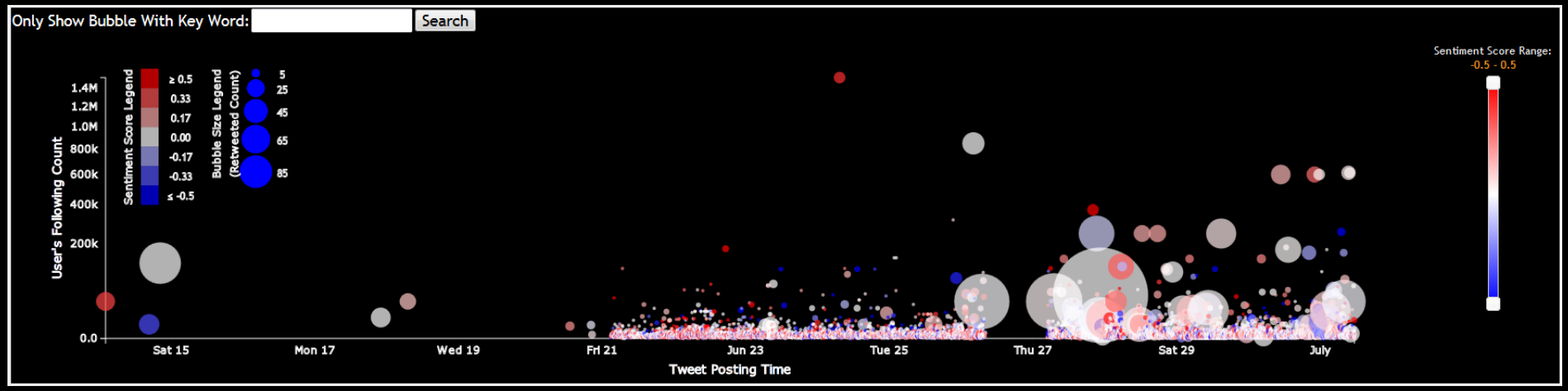
- *Exclude non-related tweets*
  - From tweet/user list or From sentiment analysis view



DM2 and Lone Ranger OW Prediction

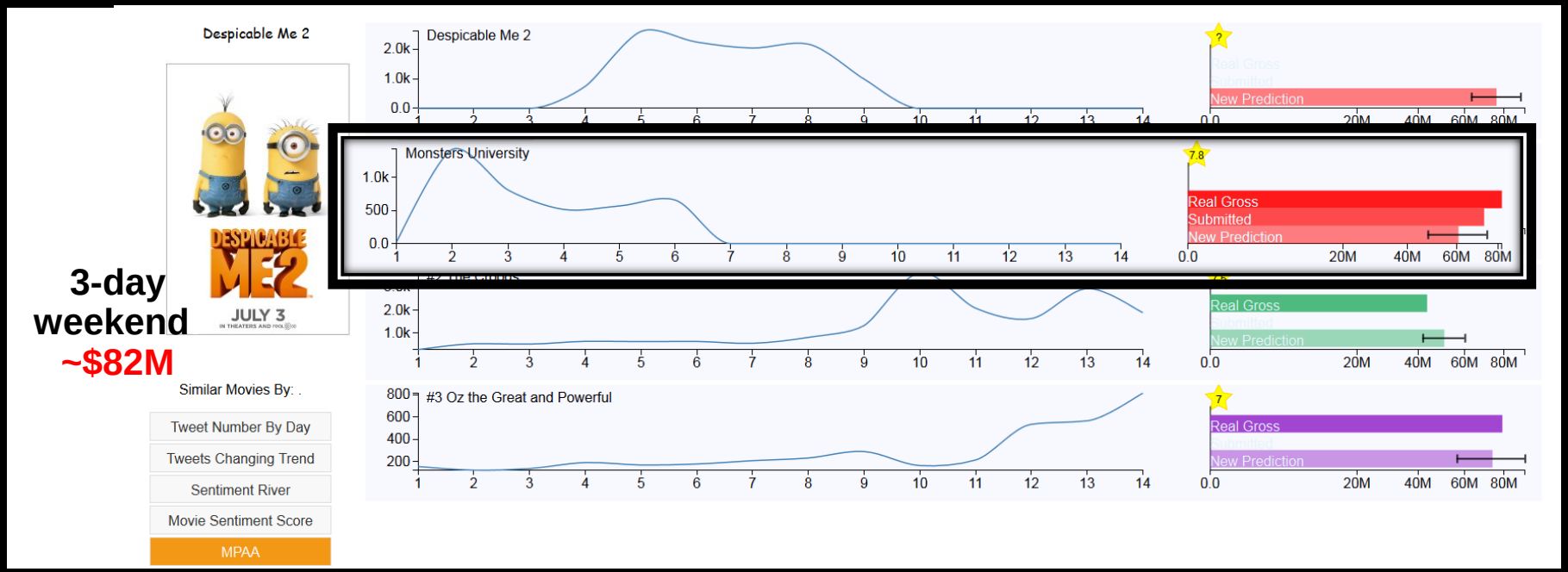
Individual movie OW box-office prediction --- - Model Adjustment

- *Exclude non-related tweets*
  - From tweet/user list or From sentiment analysis view



DM2 and Lone Ranger OW Prediction

Individual movie OW box-office prediction --- - Comparison with similar movies



Individual movie OW box-office prediction --- Comparison with similar movies



Compared with Oz and MU:

3-day weekend ~$82M



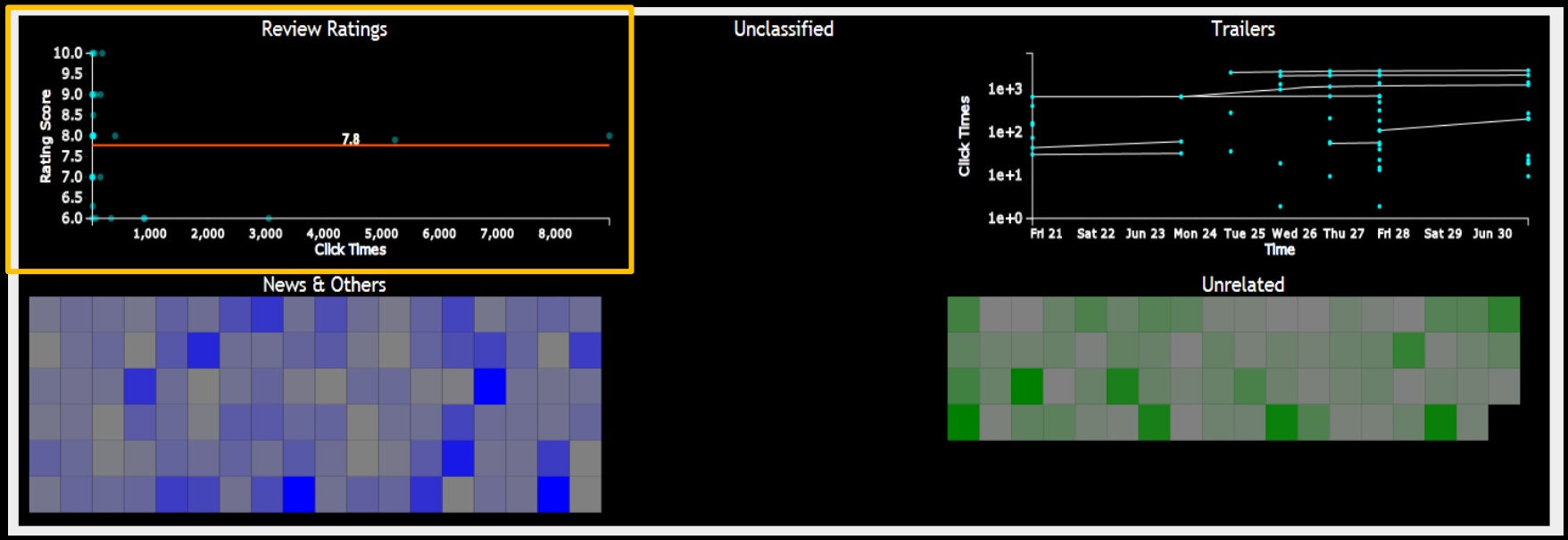
Compared with WWZ:

3-day weekend < $66M

DM2 and Lone Ranger OW Prediction

Viewer Rating Prediction

- DM2: 7.8/10.0 Lone Ranger: 6.4/10.0

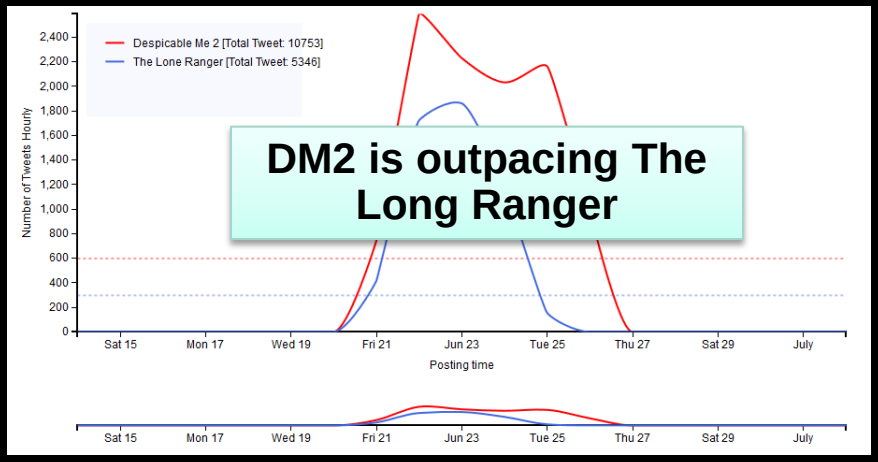


$$Viewer\ Rating = \frac{1}{k}\sum_{i=1}^{k} score(i)$$

DM2 and Lone Ranger OW Prediction

Finalize box-office prediction

- Overall view of two movies by line chart







3-day OW	Viewer rating	5-day OW
~$82M	7.8/10.0	?
~$39M	6.4/10.0	
Total ~$121M		Total ~$175M

DM2 and Lone Ranger OW Prediction

Predictions vs. Real OW

	Opening weekend gross		Viewer Rating	
	prediction	real	prediction	real
DM2 (3-day)	$82M	$83.5M	7.8	7.9
DM2 (5-day)	$117M	$143M		
The Lone Ranger (3-day)	$39M	$29M	6.4	6.8
The Lone Ranger (5-day)	$55M	$48.7M		

Performance Evaluation

Comparison With Peer Teams

Team	Gross Prediction				Viewer Rating			
	Entry	Average Error	STD	MRAE	Entry	Average Error	STD	MRAE
VADER(Interactive)	23	11.213	9.416	0.467	23	0.487	0.460	0.075
Team Prolix	23	16.466	15.195	0.424	20	0.82	0.640	0.129
Uni Konstanz Boxoffice	14	17.056	15.743	3.929	21	0.905	1.519	0.095
CinemAviz	21	17.219	17.677	1.970	21	0.738	0.559	0.114
Team Turboknopf	8	21.9	15.606	0.685	18	0.514	0.426	0.079
elvertoncf - UFMG	3	12.677	9.806	3.009	3	1.323	0.328	0.259
Philipp Omentisch	5	30.657	38.028	0.678	5	0.5	0.324	0.071
CDE IIIT	2	60.6	62.084	0.537	2	0	0	0

Performance Evaluation

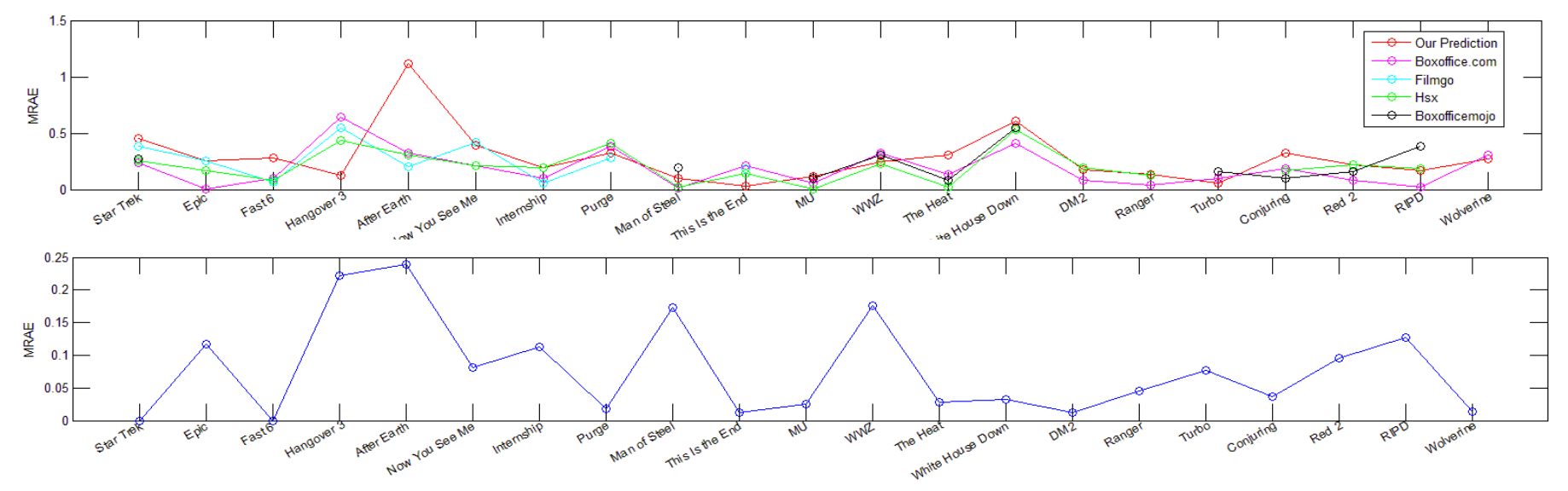
Comparison With Experts

Prediction Source	Entry	Average Error	STD	Average MRAE
VADER (interactive)	21	12.729	9.425	0.285
VADER (No interaction)	21	23.051	22.011	0.501
boxoffice.com	21	8.538	7.466	0.191
filmgo.net	6	12.75	7.409	0.297
hsx	20	9.06	7.397	0.205
boxofficemojo	14	9.864	7.527	0.224

Performance Evaluation

Mean Relative Absolute Error (MRAE)

- Compared with 4 professional website predictions



Yafeng Lu, Feng Wang, Ross Maciejewski. VAST 2013 Mini-Challenge 1: Box Office VAST - Team VADER. *IEEE Conference on Visual Analytics Science and Technology*, October, 2013.

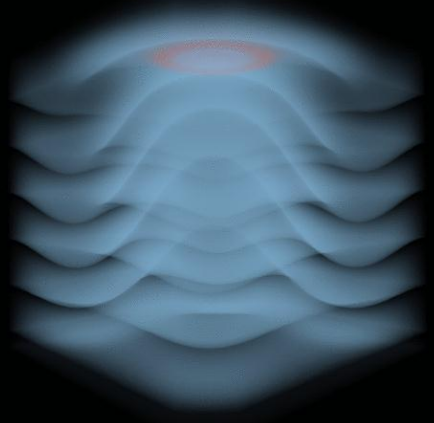
In Summary

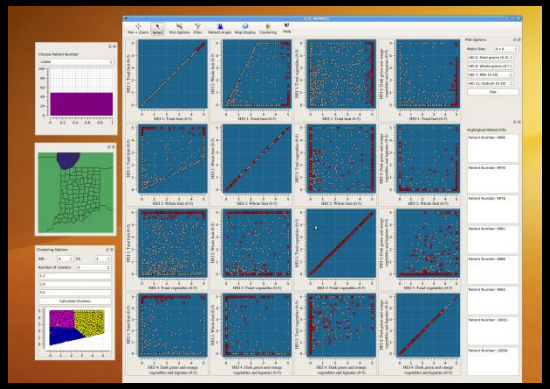
- *Big Data is everywhere (and it's fun!)*
- *We can use it to understand ourselves and the world around us*
- *We need to enable people to gain insight into data to make predictions!*
- *Visualizations help with veracity (ex. Tweet cleaning)*
- *Visualizations can give us variety (lots of ways to look at the data)!*
- *We need tools that can scale (volume) and that are fast (velocity)!*

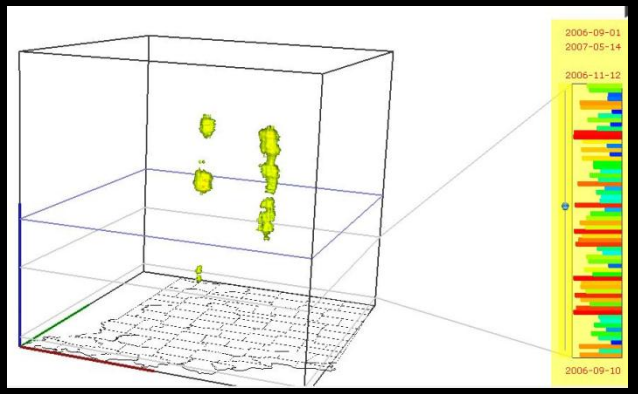
Acknowledgements

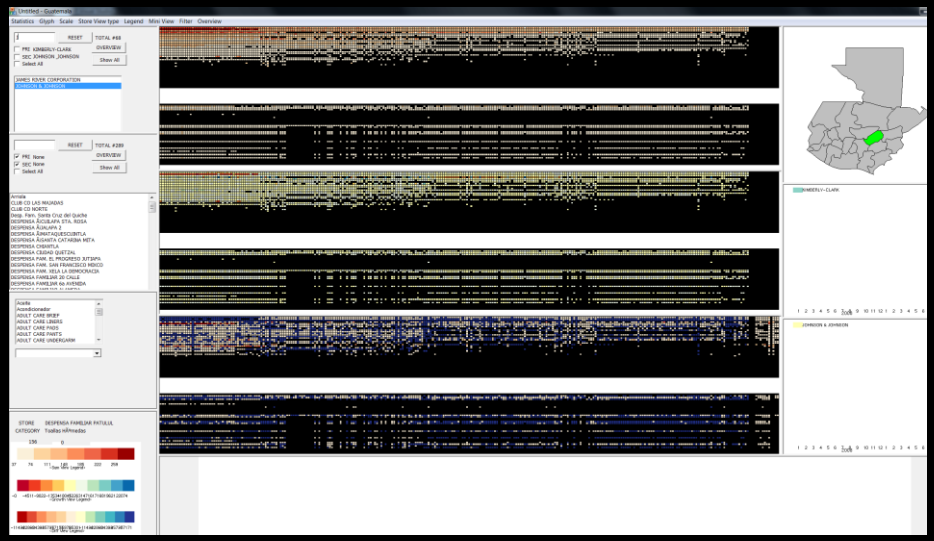
- *This work has been supported by the US Department of Homeland Security's VACCINE Center under grant number 2009-ST-061-CI0001*
- *David S. Ebert, Stephen Rudolph, Timothy Collins, and all my colleagues from Purdue University Visualization and Analytics Center*
- *Ryan Hafen, Benjamin Tyner, Cheng Zheng, David Anderson and William Cleveland from Purdue University Department of Statistics*
- *Larry Glickman, Nita Glickman and George Moore from Purdue University School of Veterinary Medicine*
- *Shaun Grannis from IU School of Medicine*
- *Michael Wade from the Indiana State Department of Health*
- *Lafayette, West Lafayette, Tippecanoe County and Purdue Police Departments*
- *And many others, with their help, this work has been truly interdisciplinary!*

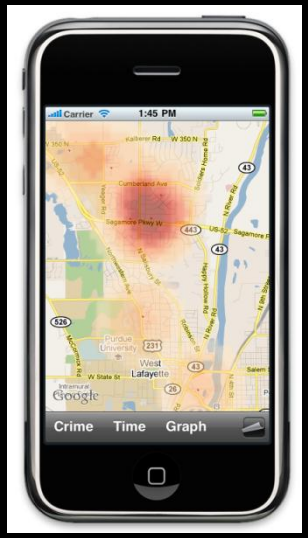
Questions?

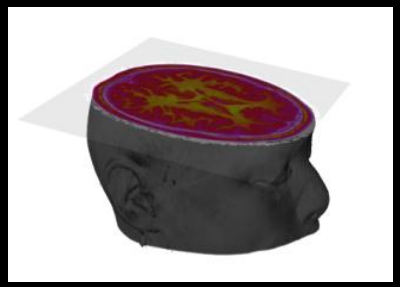
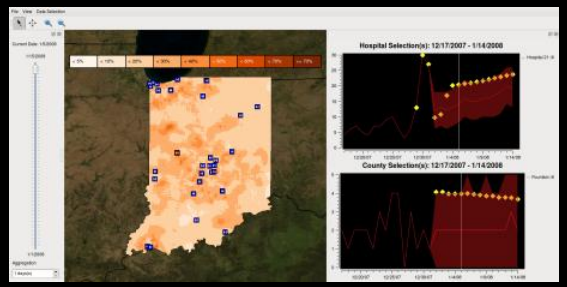


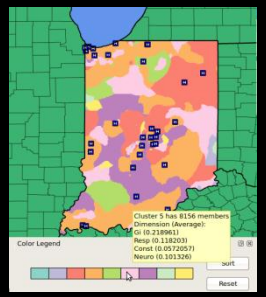


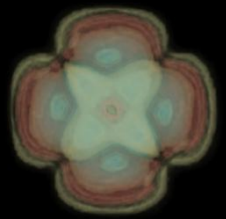

2006-09-01
2007-05-14
2006-11-12
2006-09-10




Carrier
1:45 PM
Crime
Time
Graph



Integrating Visual Analytics Curriculum Into Your Own Classroom

Ross Maciejewski



Ross Maciejewski
E-mail: rmacieje@asu.edu
http://rmaciejewski.faculty.asu.edu/
http://lab.vader.asu.edu



The Ongoing Data Deluge

- Since 2003, digital information has accounted for 90% of all information produced[1]
- In 2009, drones from Iraq and Afghanistan recorded 24 years of video footage
- In 2010, the amount of information added annually to the digital universe was estimated to be nearly 1 ZB
- Wal-mart process > 1 million transactions per hour
- By 2013 Cisco estimates the annual internet traffic will be 667 EBs

1 – C. R. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, and T. S. Yoo, Eds., *NIH-NSF Visualization Research Challenges Report*, IEEE Press, ISBN 0-7695-2733-7, 2006

The Ongoing Data Deluge

- FedEx's ships more than 8.5 million packages per day (http://www.fedex.com/us/about/today/companies/corporation/facts.html)
- Consumers carry more than 1 billion Visa cards worldwide. More than 450 million of those cards are in the United States (http://www.creditcards.com/credit-card-news/credit-card-industry-facts-personal-debt-statistics-1276.php)

Data Overload

- Opportunity: Huge amounts of data available in digital form and ready for analysis!
  - But, how can we make sense of this data?
  - How can we harness this data in the decision making process?
  - How do we avoid being overwhelmed by all of this data?



Why Is This Important to Educators?

- In 2011, a report by the McKinsey Global Institute projected that the US needs 140,000-190,000 works with "deep analytical" expertise[1]
- These are not just computer science jobs!
- Harvard Business Review calls data science "the sexiest job in the 21st century"[2]
- We need to train the next generation of scientists for data analytics!

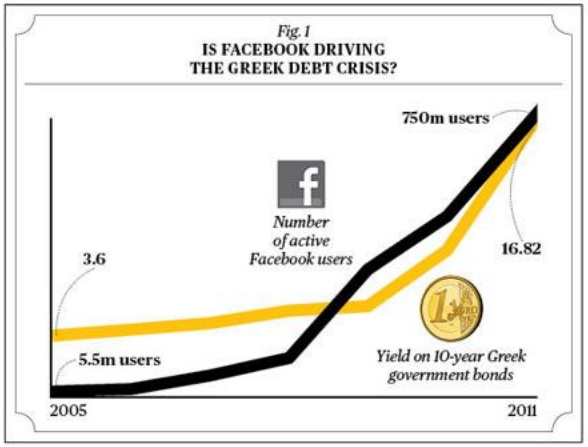
1 - http://www.mckinsey.com/Features/Big_Data
2 - http://www.nytimes.com/2013/04/14/education/edlife/universities-offer-courses-in-a-hot-new-field-data-science.html?pagewanted=all
3 - http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0

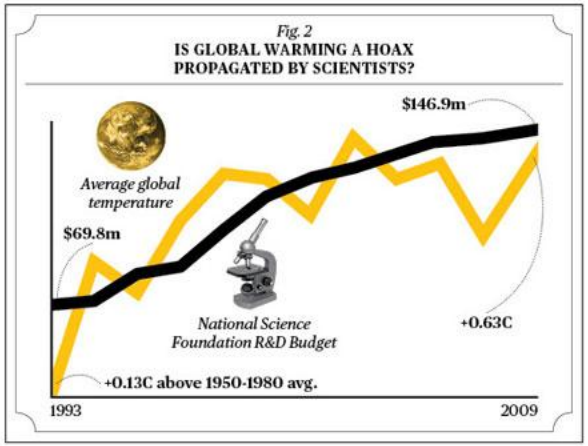
What Goes Into an Degree in “Data Science”?

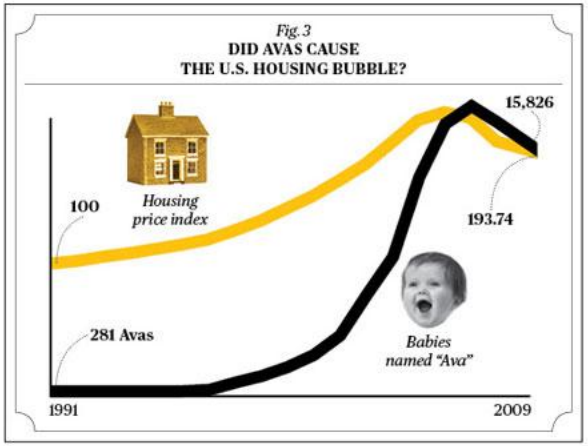
- Typically Data Science is a Masters Program
- Basic programming
- Data cleaning and scraping
- Predictive statistics (regression modeling)
- Ethics classes on data privacy implications
- What about communicating findings, what other skills do we need?

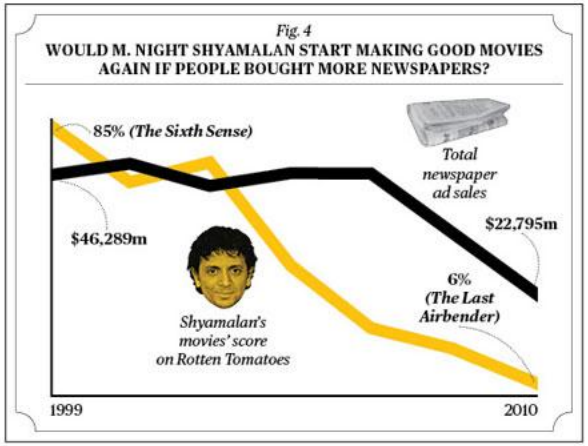
The Power of Visualization

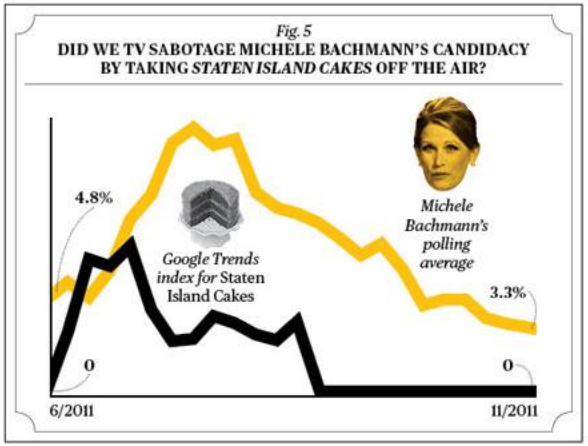
- The goal is to take all of this *data* and transform it into *information*
- How many terabytes of data we have collected doesn't matter, it's how many petaflops of *insights* we can generate from this data
- "*The purpose of computing is insight, not numbers.*" Dedication to his book: RW Hamming (1971). *Introduction to Applied Numerical Analysis*. McGraw Hill.
- We need to make the data understandable to people and a key way of doing this is through *visualization*

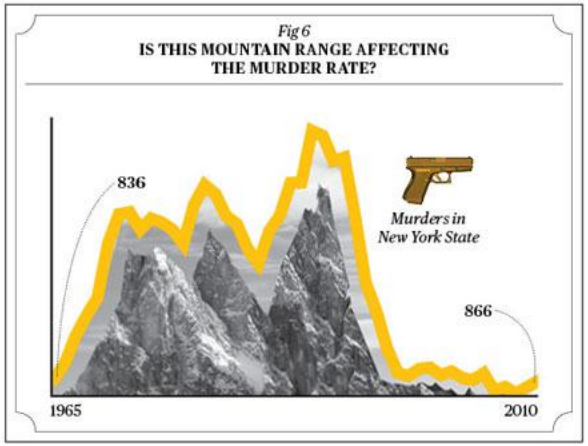












ALAMY(3); BLOOMBERG(1); GETTY IMAGES(7). DATA: FIG 1: FACEBOOK, BLOOMBERG; FIG 2: NASA, NATIONAL SCIENCE FOUNDATION; FIG 3: U.S. SOCIAL SECURITY ADMINISTRATION, FEDERAL HOUSING FINANCE AGENCY; FIG 4: ROTTEN TOMATOES, NEWSPAPER ASSOCIATION OF AMERICA; FIG 5: GOOGLE, REAL CLEAR POLITICS; FIG 6: NEW YORK LAW ENFORCEMENT AGENCY

What Skills Are Students Missing?

Non-Data Components of Graphs

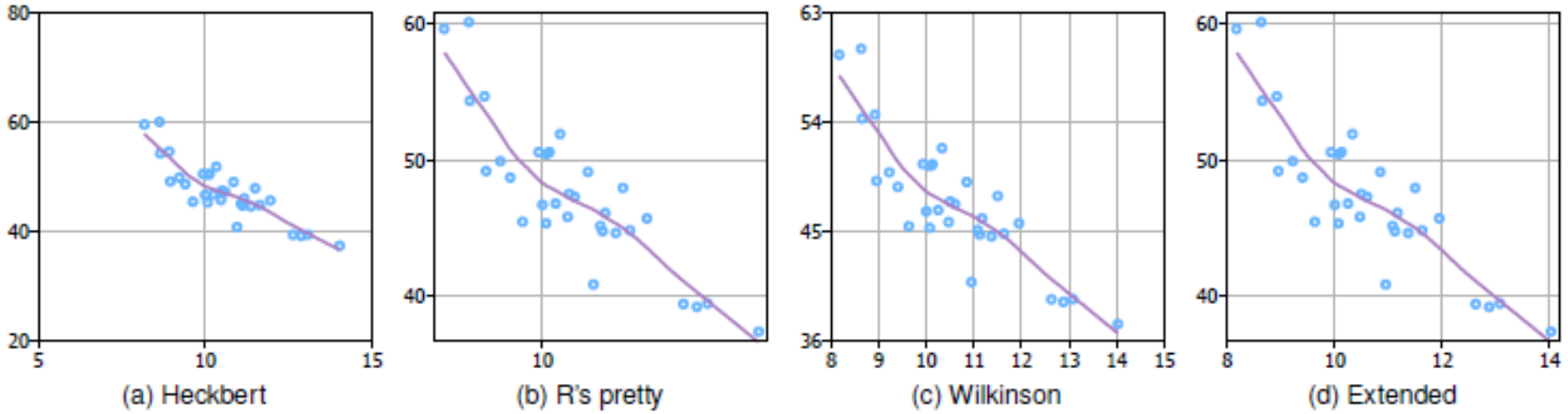
- Axes and legends can often be as important as the data themselves
- Poor axis choices and label choices can lead to confusing visualizations
- Axis tick labels provide cognitive context for most basic plot types
- They support estimation and contribute to the overall appearance of the graphic
- Cleveland suggests choosing the scales so that the data rectangle fills up as much of the scale-line rectangle as possible

W. S. Cleveland. The Elements of Graphing Data. Wadsworth Publ. Co., Belmont, CA, USA, 1985.

Extension of Wilkinson's Algorithm

- Utilizes four components for optimizing tick placement
  - Simplicity $= 1 - \frac{i-1}{|Q|-1}$-j+v
  - Density $= 2 - max\left(\frac{\rho}{\rho_t}, \frac{\rho_t}{\rho}\right)$

	Name	Description
Input	Q	preference-ordered list of nice step sizes
	(d_{min}, d_{max})	data range
	ρ_t	target label density
	fs_t	target font size
	Formats	set of label formats
		(defaults shown in Table 3)
Other	q	element of Q
	j	skip amount
	i	index of q in Q
	(l_{min}, l_{max})	labeling sequence range
	v	equals 1 if labeling includes 0, 0 otherwise



1 – J. Talbot, S. Lin, P. Hanrahan, "**An Extension of Wilkinson's Algorithm for Positioning Tick Labels on Axes**," IEEE Transactions on Visualization and Computer Graphics 16(6): 1036-1043, 2010
2 - L. Wilkinson. The Grammar of Graphics (Statistics and Computing). Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

Graph Aspect Ratios

- Our ability to perceive trends and patterns in a given display is heavily influenced by the aspect ratio
- Aspect ratio affects densities, relative distances and orientations
- Several methods have been proposed for automatically selecting the aspect ratio
- Aspect ratio: a = width/height

1 – J. Talbot, J. Gerth, P. Hanrahan, “Arc Length-based Aspect Ratio Selection,” IEEE Transactions on Visualization and Computer Graphics 17(12): 2276-2282, 2011

How Can We Integrate Data Analytics Into Our Classes?

National Health and Nutrition Examination Survey (NHANES)

Designed to assess health and nutritional status of adults and children in the US

- *Combines interviews and physical examinations*
- *Exams a nationally representative sample of about 5,000 persons per year*
- *Includes demographic, socioeconomic, dietary and health-related questions*

Healthy Eating Index (HEI)

Measure of diet quality that assesses conformance to federal dietary guidance

USDA uses HEI to monitor diet quality of the US population

HEI-2005 uses recommendations of the MyPyramid guidance as the baseline

1 - Guenther, P. M., Reedy, J., Krebs-Smith, S. M., Reeve, B. B., "Evaluation of the Healthy Eating Index-2005," *Journal of the American Dietetic Association*, 108:1854-1896, 2008.

2 – Freedman, L. S., Guenther, P. M., Krebs-Smith, S. M., Kott, P. S., "A Population's Mean Healthy Eating Index-2005 Scores are Best Estimated by the Score of the Population Ratio when One 24-Hour Recall is Available," Journal of Nutrition, 138:1725-1729, 2008.

Healthy Eating Index–2005 components and standards for scoring[1]

Component		Maximum points	Standard for maximum score	Standard for minimum score of zero
Total Fruit (includes 100% juice)	(HEI-1)	5	≥0.8 cup equiv. per 1,000 kcal	No Fruit
Whole Fruit (not juice)	(HEI-2)	5	≥0.4 cup equiv. per 1,000 kcal	No Whole Fruit
Total Vegetables	(HEI-3)	5	≥1.1 cup equiv. per 1,000 kcal	No Vegetables
Dark Green and Orange Vegetables and Legumes[2]	(HEI-4)	5	≥0.4 cup equiv. per 1,000 kcal	No Dark Green or Orange Vegetables or Legumes
Total Grains	(HEI-5)	5	≥3.0 oz equiv. per 1,000 kcal	No Grains
Whole Grains	(HEI-6)	5	≥1.5 oz equiv. per 1,000 kcal	No Whole Grains
Milk[3]	(HEI-7)	10	≥1.3 cup equiv. per 1,000 kcal	No Milk
Meat and Beans	(HEI-8)	10	≥2.5 oz equiv. per 1,000 kcal	No Meat or Beans
Oils[4]	(HEI-9)	10	≥12 grams per 1,000 kcal	No Oil
Saturated Fat	(HEI-10)	10	≤7% of energy[5]	≥15% of energy
Sodium	(HEI-11)	10	≤0.7 gram per 1,000 kcal[5]	≥2.0 grams per 1,000 kcal
Calories from Solid Fats, Alcoholic beverages, and Added Sugars (SoFAAS)	(HEI-12)	20	≤20% of energy	≥50% of energy

[1]Intakes between the minimum and maximum levels are scored proportionately, except for Saturated Fat and Sodium (see note 5).
[2]Legumes counted as vegetables only after Meat and Beans standard is met.
[3]Includes all milk products, such as fluid milk, yogurt, and cheese, and soy beverages.
[4]Includes nonhydrogenated vegetable oils and oils in fish, nuts, and seeds.
[5]Saturated Fat and Sodium get a score of 8 for the intake levels that reflect the 2005 Dietary Guidelines, <10% of calories from saturated fat and 1.1 grams of sodium/1,000 kcal, respectively.

http://www.cnpp.usda.gov/Publications/HEI/healthyeatingindex2005factsheet.pdf

Playing Data Detective

So, we've collected all of these surveys, now what?

Well, how about Exploratory Data Analysis where we visually inspect the data and ask questions about it to help form hypotheses.

What are some easy questions to ask?

- *What is the average … (total fruit consumption)?*
- *What is the average … (milk consumption)?*

1 – J. W. Tukey. Exploratory Data Analysis

NHANES

Utilizing Survey Cycle 2-3

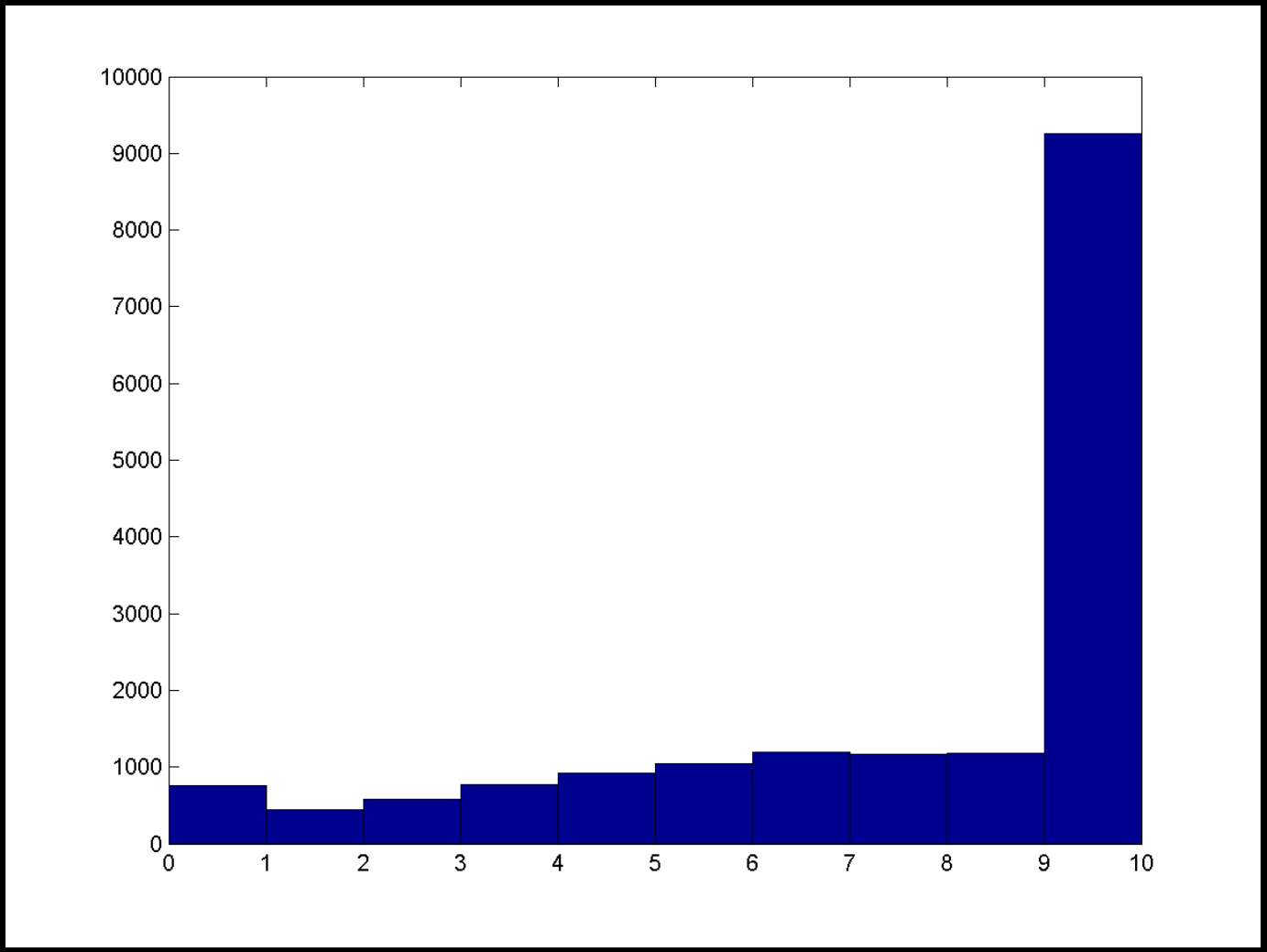
- ***Cycle 2: 2001-2002***
- ***Cycle 3: 2003-2004***
- ***Sample size of ~15000***

Other variables:

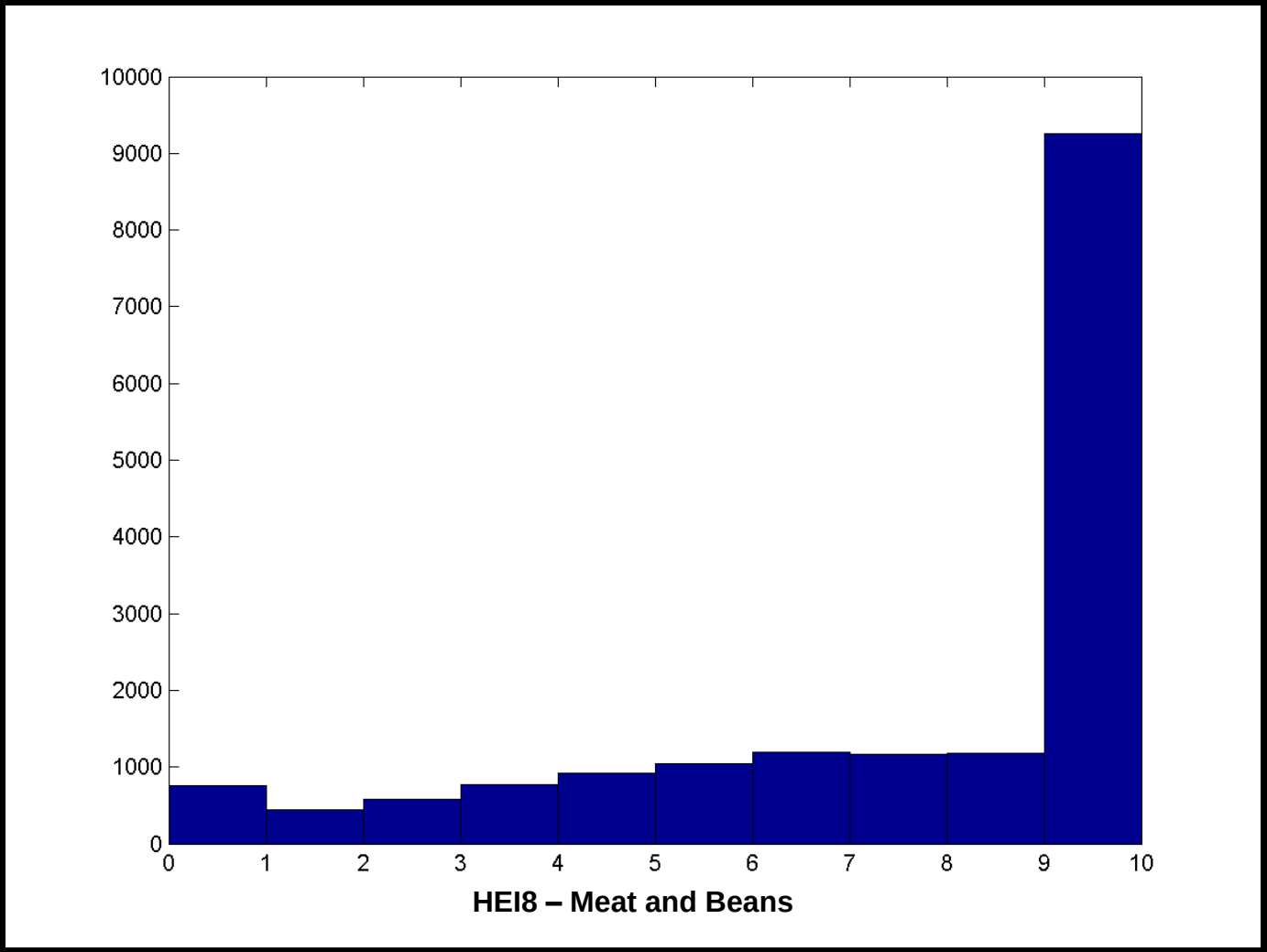
- **Gender**
- **Age**
- **Race**
- **Total energy for the day (kcal)**
- **Total saturated fatty acid (gm)**
- **BMI**

So with more data, maybe we can think of more questions!

What is the average … ?



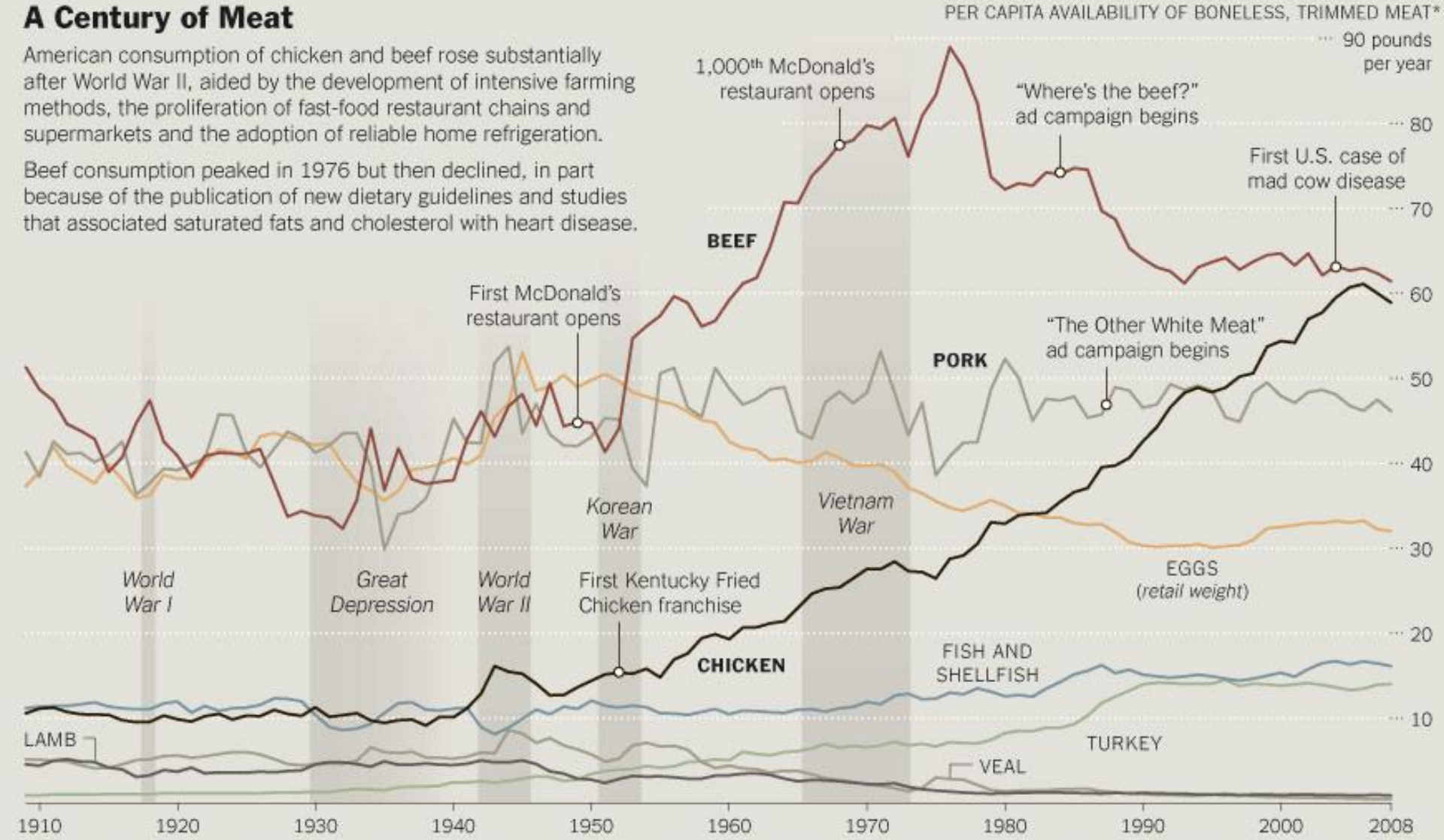
What is the average … ? Meat Consumption



A Century of Meat

American consumption of chicken and beef rose substantially after World War II, aided by the development of intensive farming methods, the proliferation of fast-food restaurant chains and supermarkets and the adoption of reliable home refrigeration.

Beef consumption peaked in 1976 but then declined, in part because of the publication of new dietary guidelines and studies that associated saturated fats and cholesterol with heart disease.



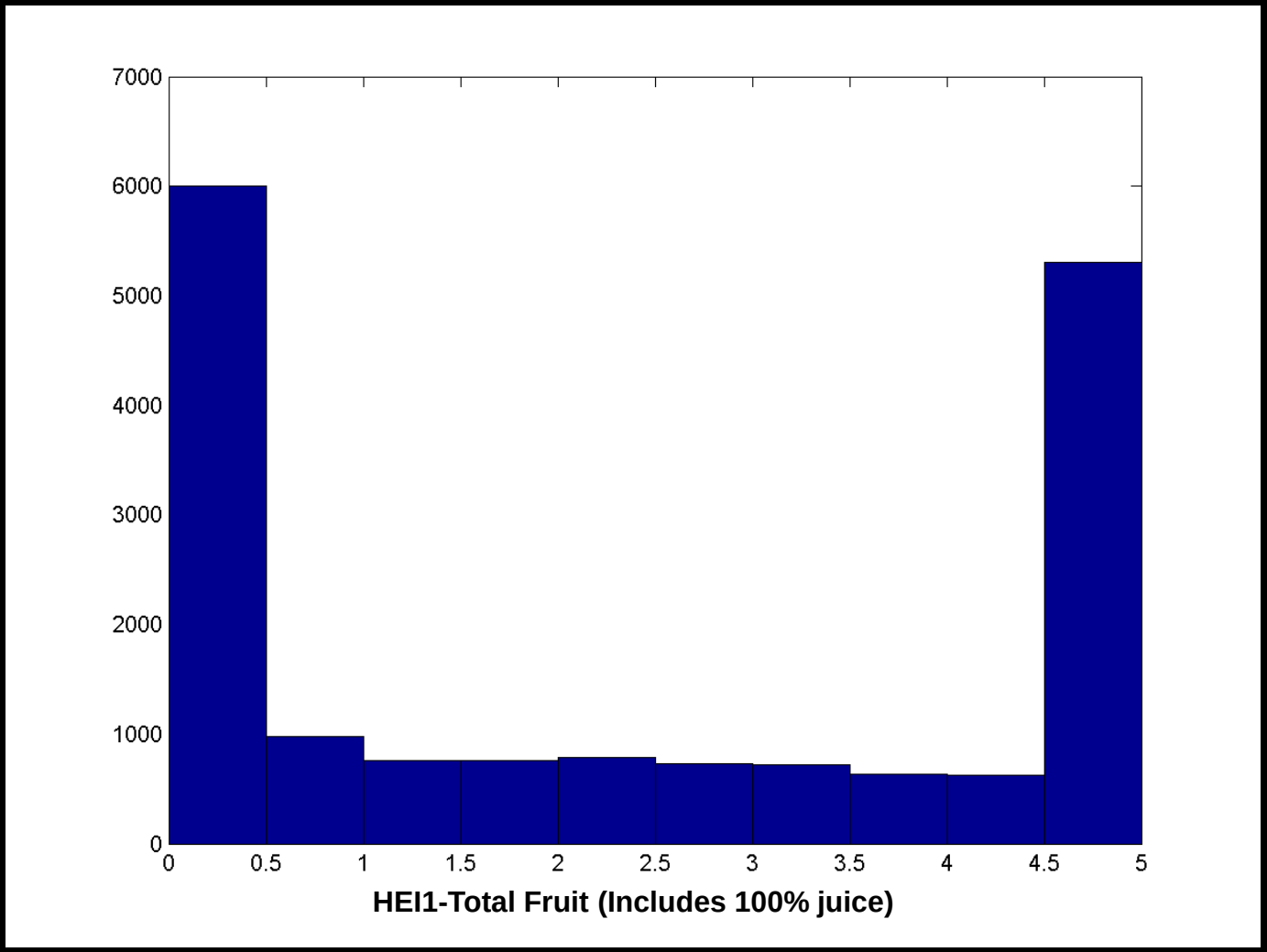
*Note: per capita availability of boneless meat is a proxy for human consumption, and is lower than retail weight or carcass weight. Bones, offal and game are excluded.

Sources: U.S. Department of Agriculture (data); news and company reports; "Putting Meat on the American Table," by Roger Horowitz

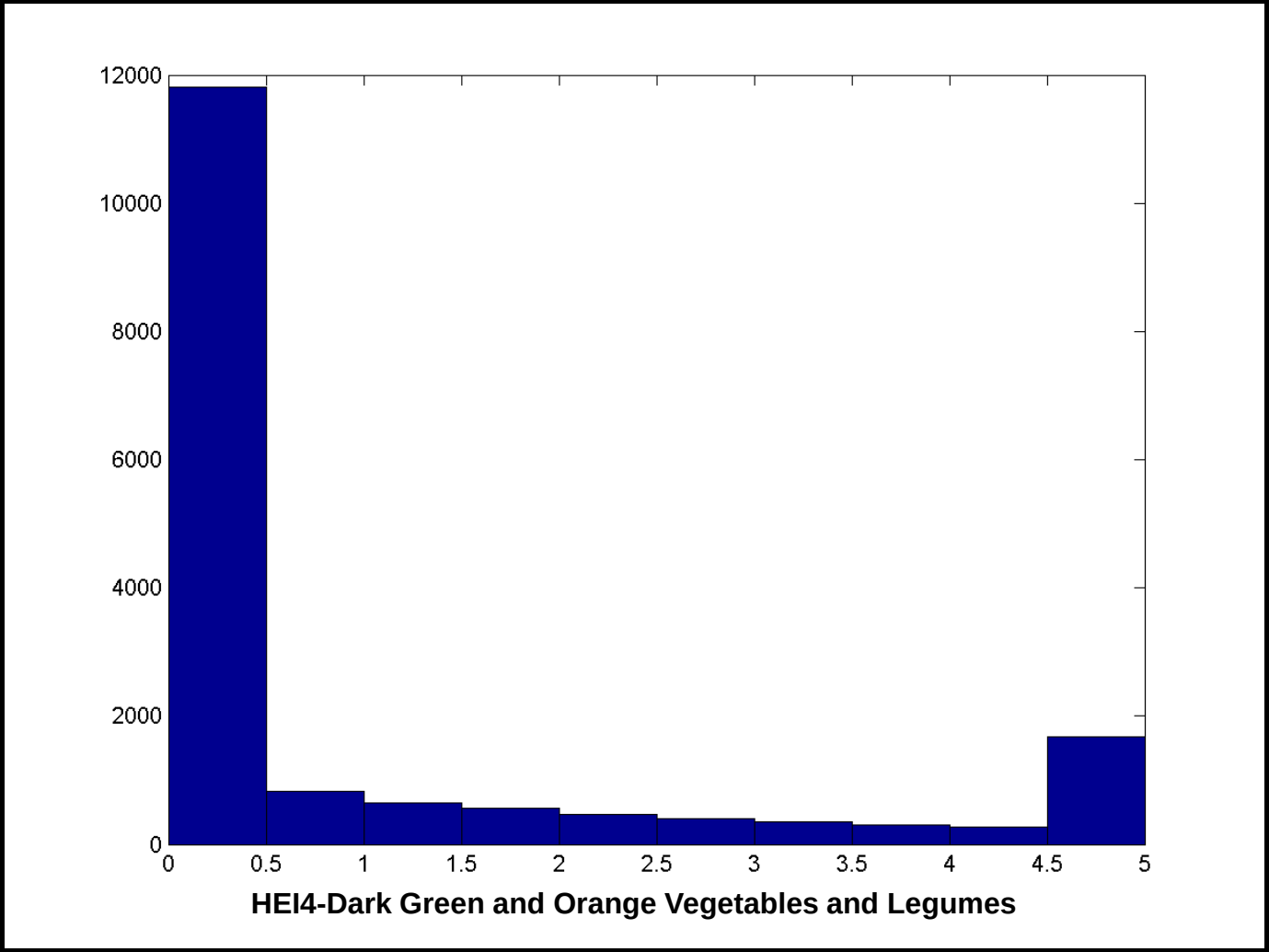
JONATHAN CORUM/THE NEW YORK TIMES

http://www.nytimes.com/imagepages/2011/03/15/science/15food_graphic.html

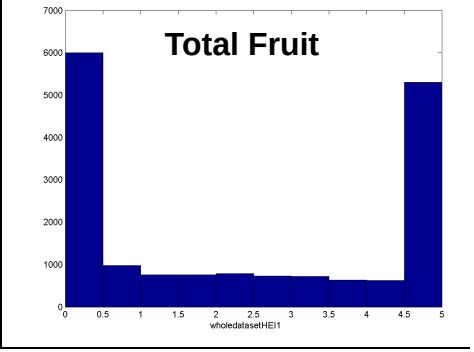
What is the average … ? Total Fruit Consumption

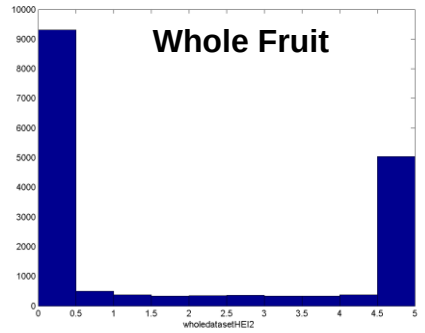


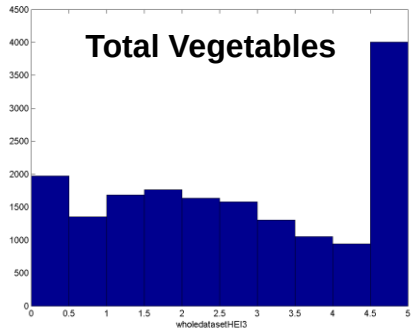
What is the average … ? Dark Green and Orange Vegetable Consumption

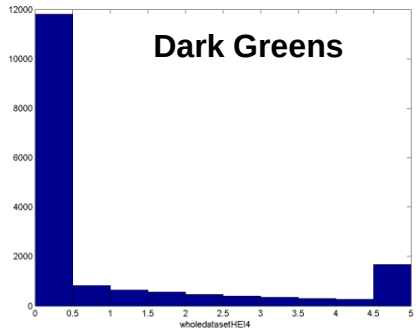


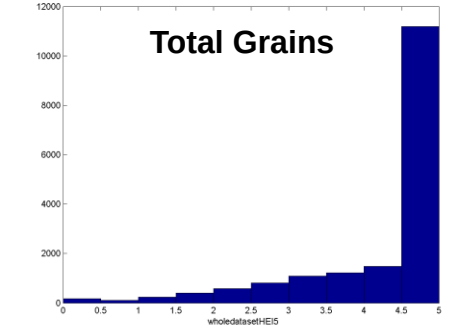
How Do These Compare?

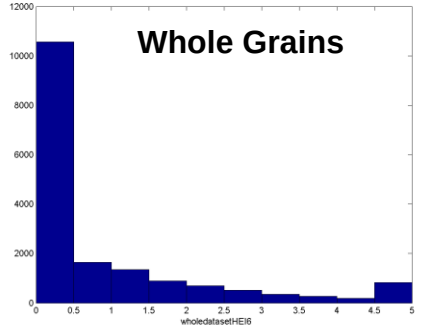


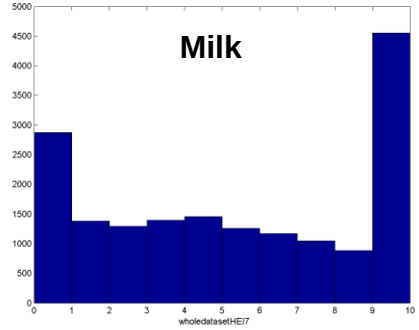


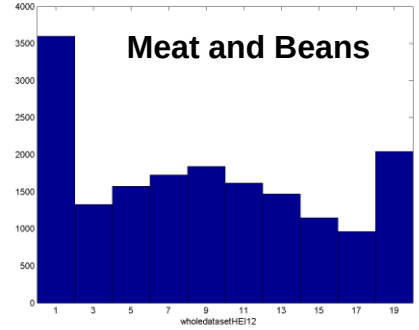


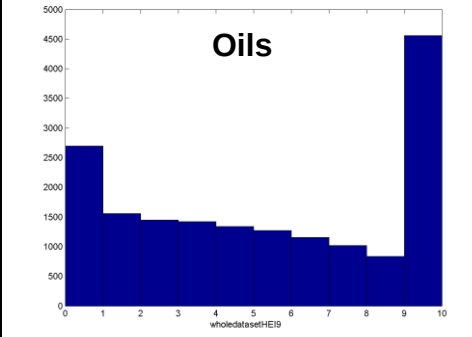


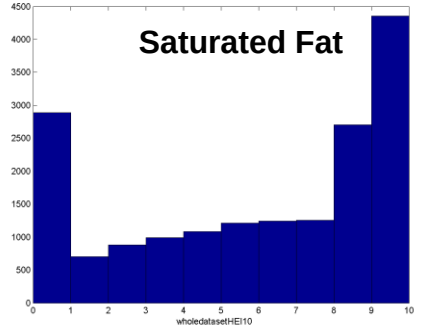


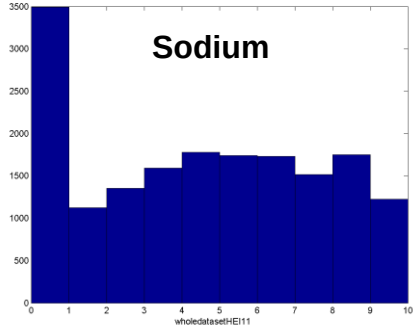


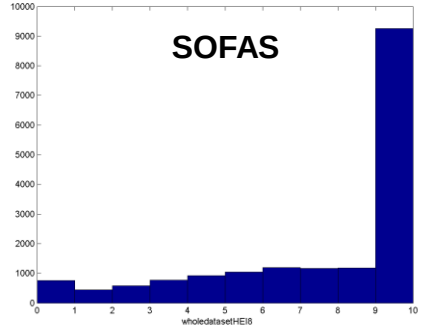












So, What Can I say About the Population as a Whole?

- *We consume lots of SOFAs*
- *We are not good at eating our vegetables*
- *We eat lots of grains …*
- *But not lots of whole grains*

- *But what does this say about someone like you?*

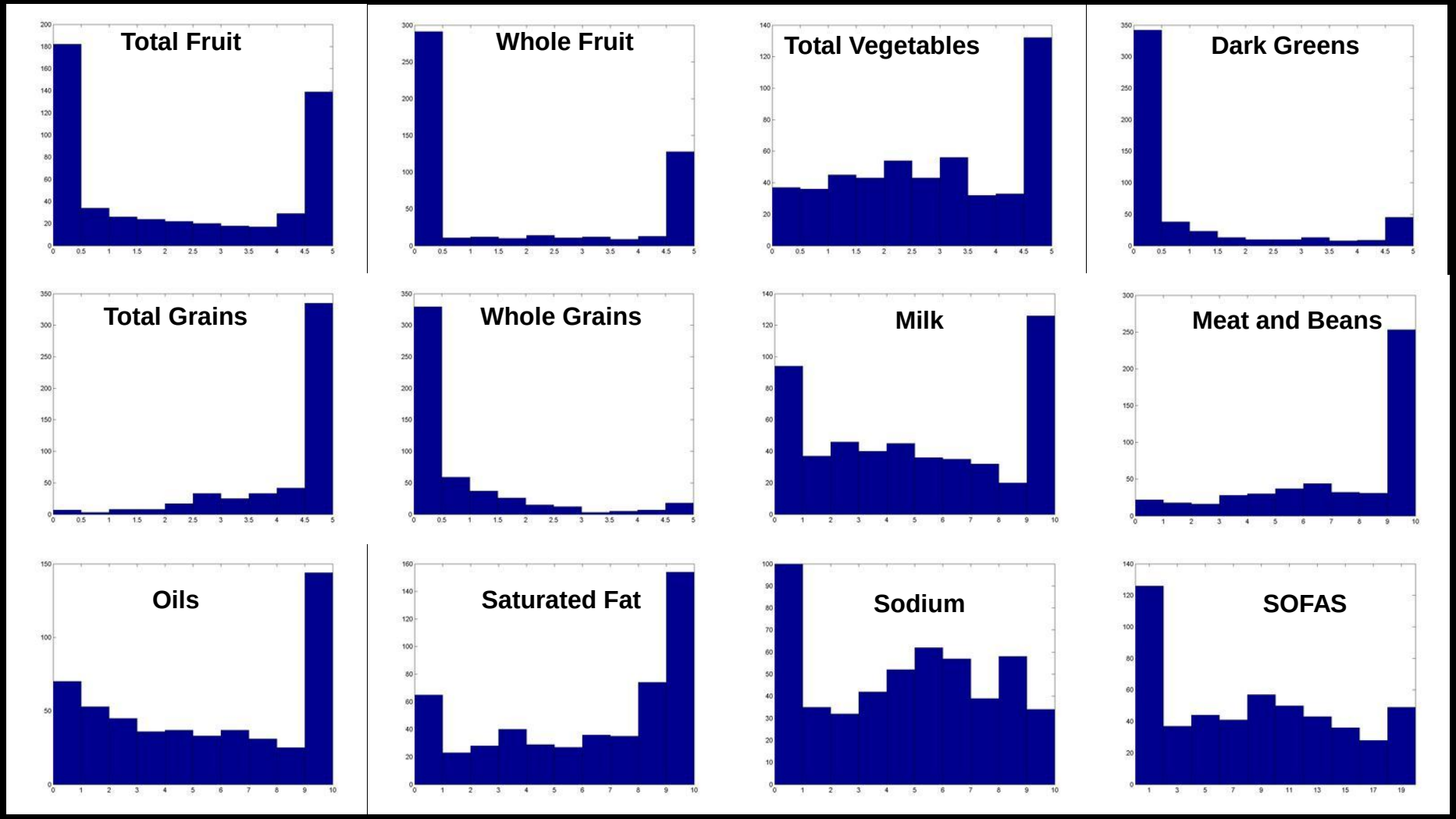
What if I further segment the data? Filter by Age, Gender, Etc.?

Let's look at Females, Age 19-30

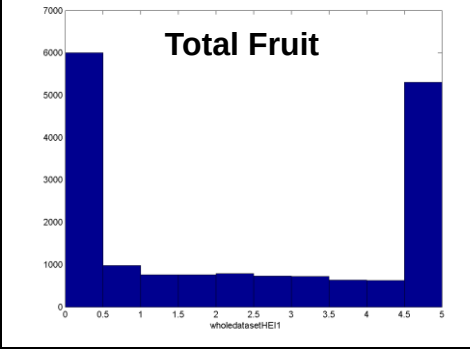
18.5 < BMI < 25.0

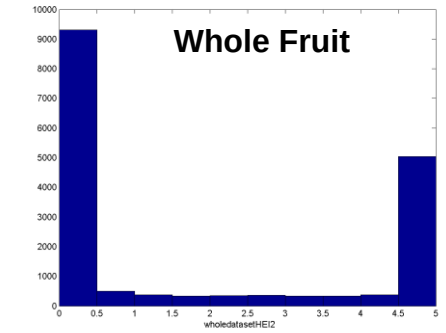
Total population in our sample = 511

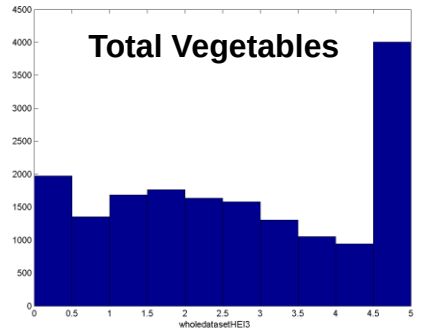
NHANES in Females Age 19-30, Average BMI

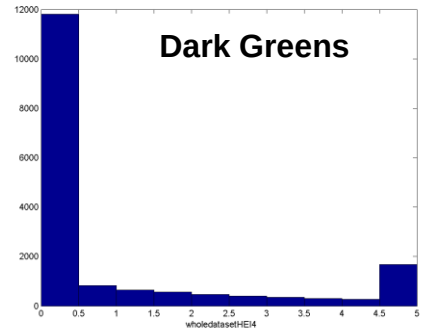


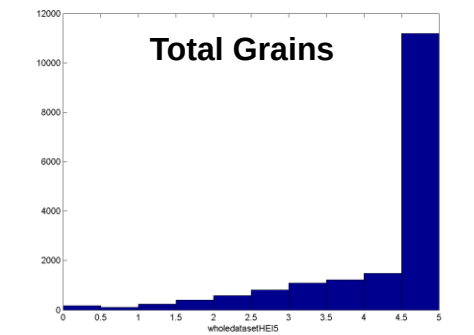
Back to the Entire Population!

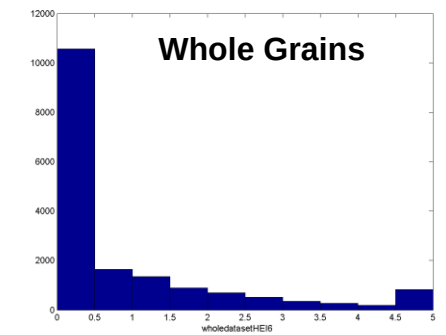


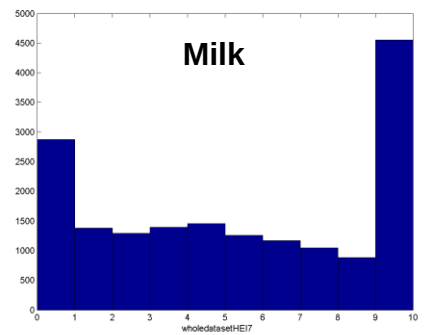


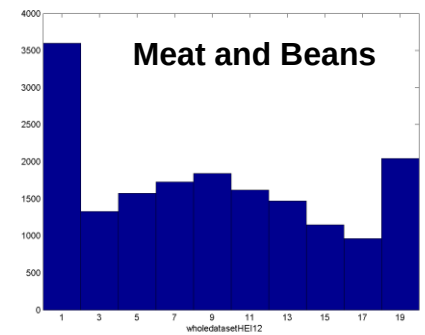


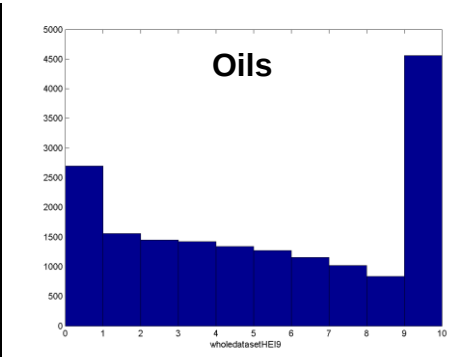


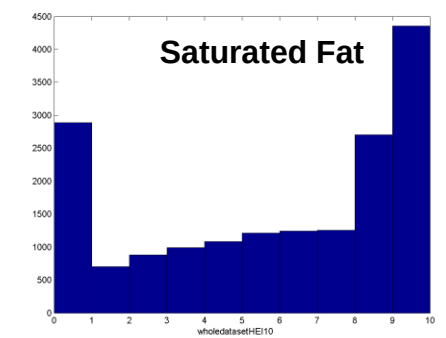


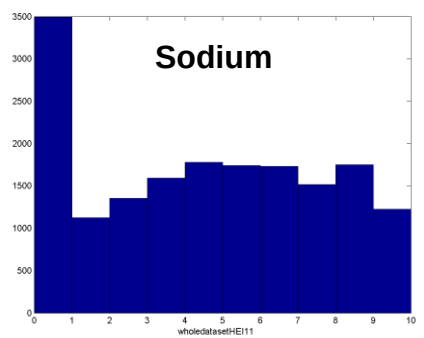


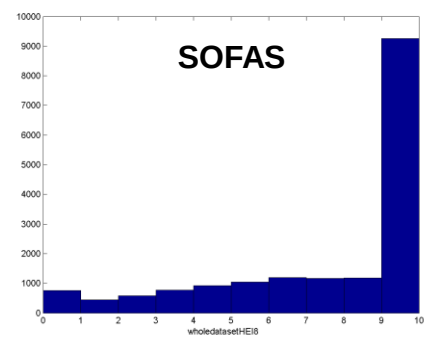












Forming Hypotheses

What do we think we see when comparing this group of females to the entire population?

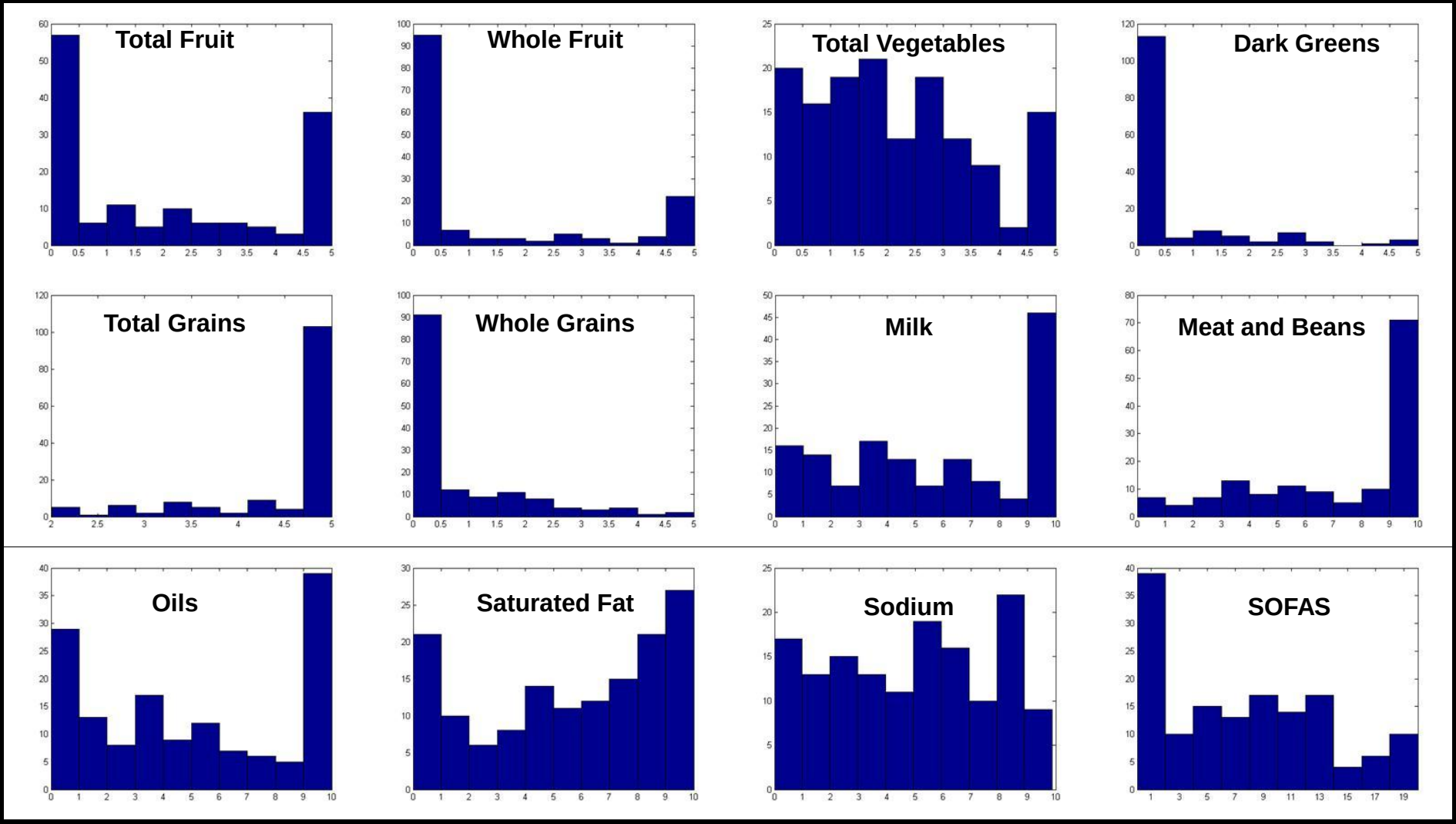
- ***Their SOFA consumption is more evenly distributed***
- ***Their meat and bean consumption is more skewed***

What if I further segment the data? Filter by Age, Gender, Etc.?

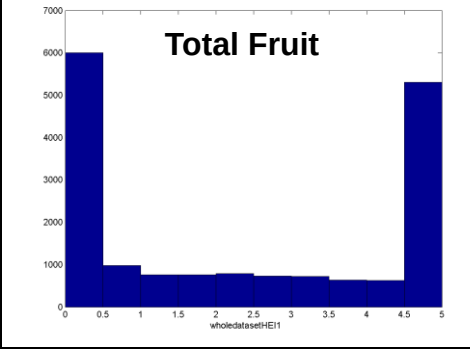
Let's look at Males, Age 14-18, BMI < 18.5

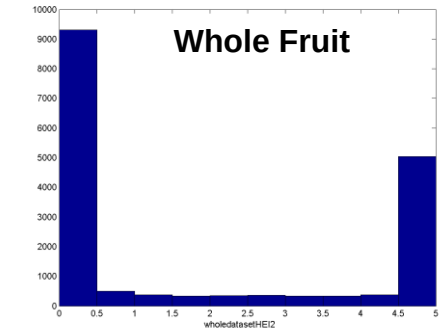
Total population in our sample = 145

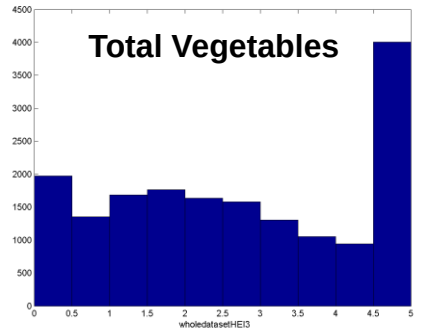
NHANES in Adolescent Underweight Males

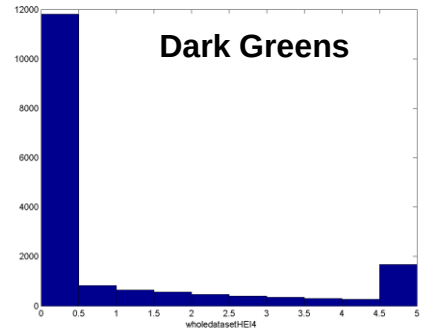


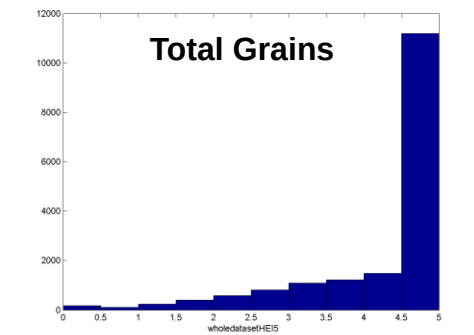
Back to the Entire Population!

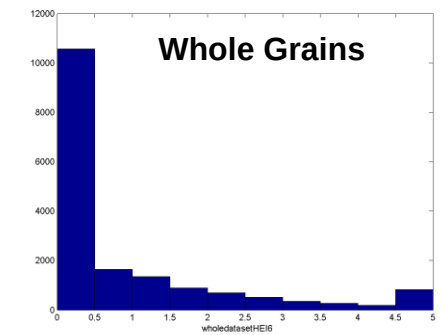


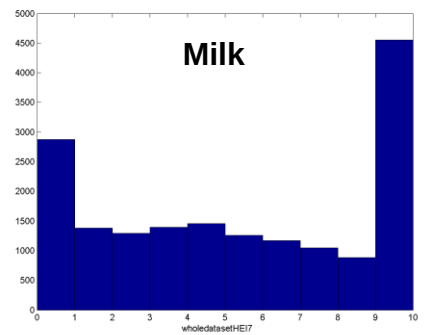


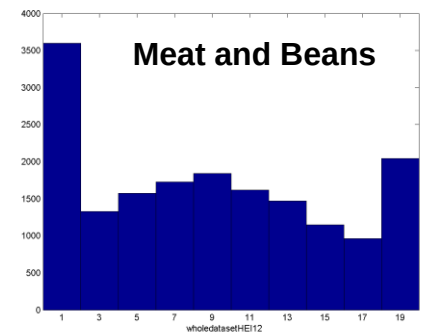


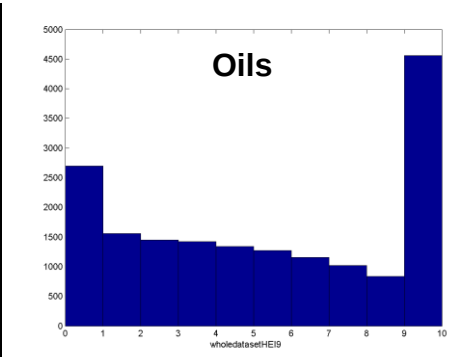


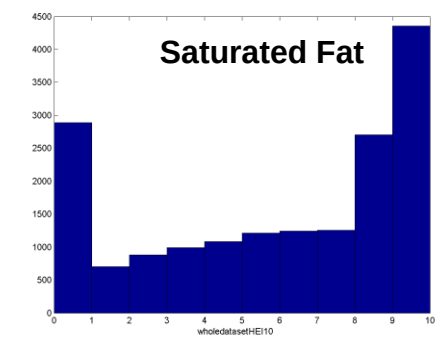


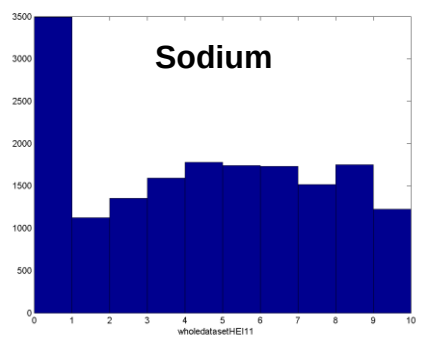


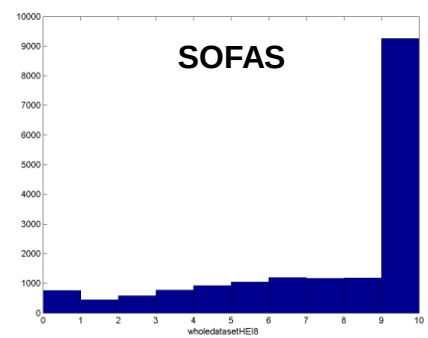












Forming Hypotheses

What do we think we see when comparing this group of females to the entire population?

- ***Their vegetable consumption is more evenly distributed***
- ***Their meat and bean consumption is more skewed***
- ***Their SOFAs are more evenly distributed***

What if I want to ask: How is … related to … ?

From our histograms, we came up with interesting insights into our data

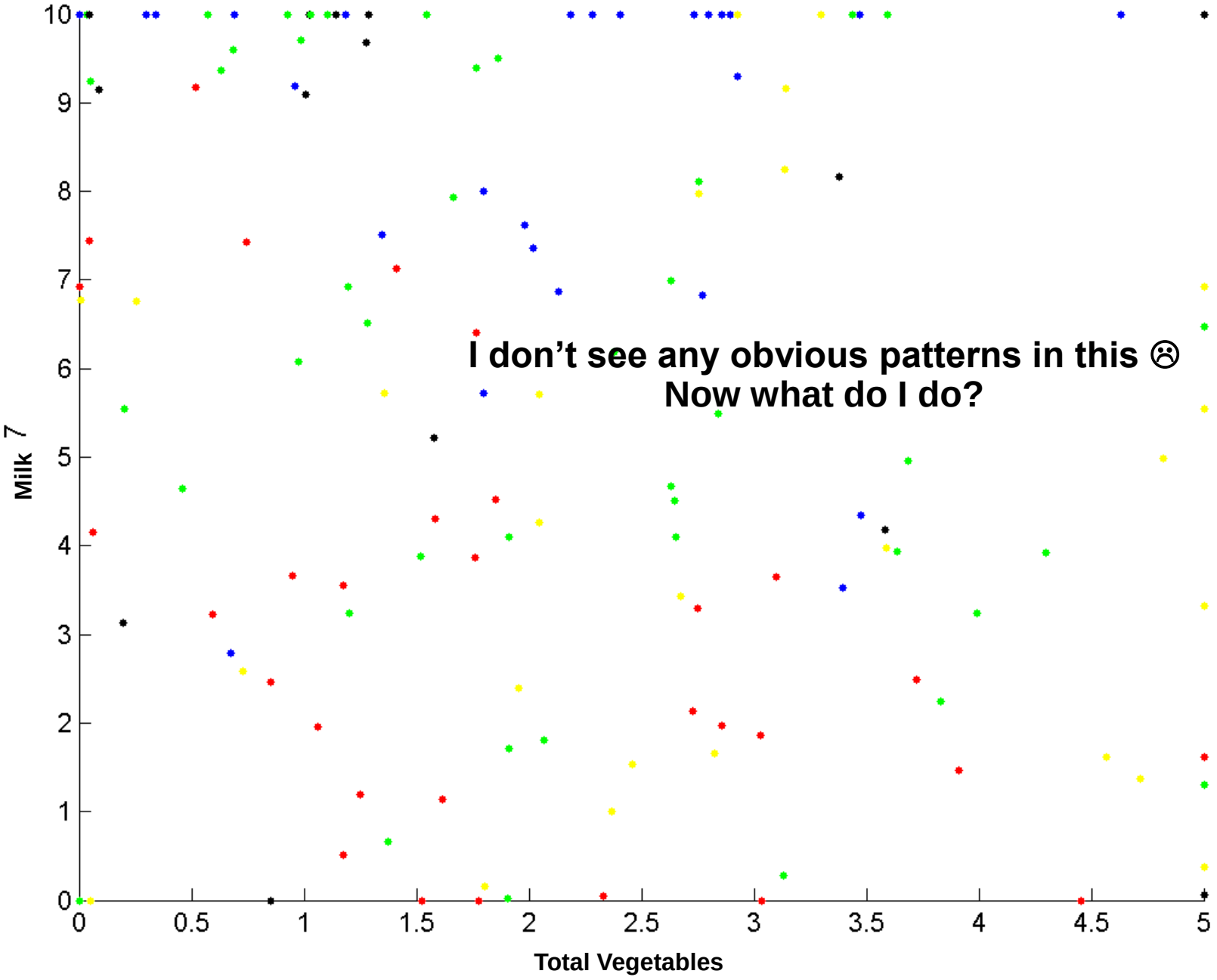
But what if I want to know about relationships between the data?



More like:
Do people that consume more milk consume more vegetables?

We can try a scatterplot!

Group: Underweight Adolescent Males



What if I further question the data? All relationships by Age, Gender?

How many of these did you really expect me to do?

These histograms were made in Excel

One can also use SAS, JMP, MatLab

In all of those programs we have to write scripts, modify variables, etc.

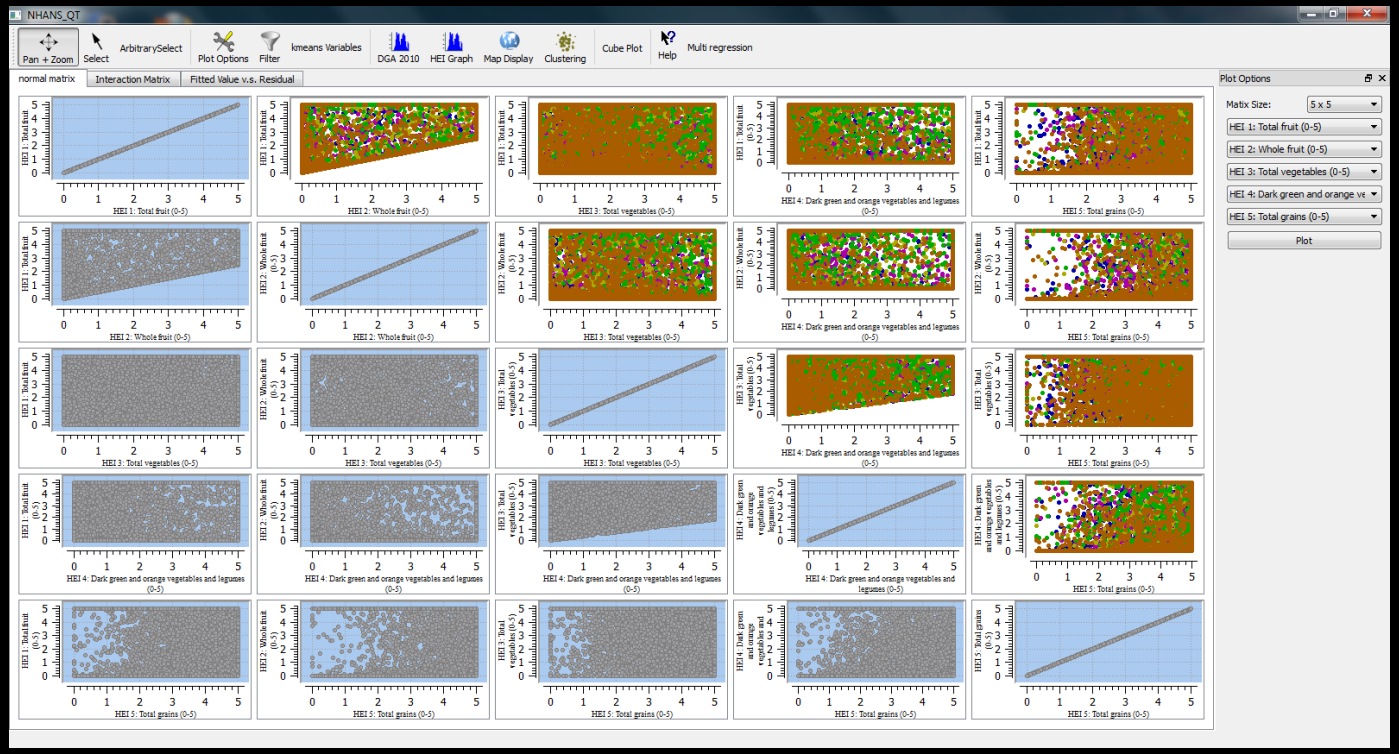
We can't directly manipulate our variables

So, being a data detective can be very tedious

And these were just simple comparisons!

So what do we do now?

We can utilize interactive graphics combined with underlying analytical processes to help us be better detectives!



The Curse of Dimensionality

- *A term coined by Bellman in 1961*
- *Refers to the problems associated with multivariate data analysis as the dimensionality increases the available data becomes sparse*
- *Sparsity is a problem for any method that requires statistical significance*
- *Sometimes data dimensions are redundant and can be reduced with minimal information loss*
- *In visualization we are also limited with screen space and the number of available visual variables, so choosing the most appropriate dimensions is key*

Richard Ernest Bellman; Rand Corporation (1957). *Dynamic programming*. Princeton University Press.

The Curse of Dimensionality

So, what can we do?

- We can incorporate prior knowledge of the data
- We can smooth the target function
- We can reduce the dimensionality

The Practicality of the Curse of Dimensionality

For a given sample size, there is a maximum number of features above which the performance of classifying samples will degrade rather than improve

In most cases, the additional information that is lost when discarding some features is compensated by a more accurate mapping in the lower-dimensional space

So, how do we know what features we can throw away?

For visualization this implies that there are some features of a dataset that will be better to visualize (contain more information) than others!

Curses! I'm Still Overwhelmed!

Probably so, we are creating a lot of data in the world

Overall, NHANES is actually a relatively modest set of data

As our datasets get more variables we need ways to question the data

What I really want to know is are there relationships/groups/clusters in the data, how can I see these when we have so many dimensions?

One way is to just cluster the data and draw these clusters in an arbitrary space (we saw this in the demo)

Another is to try dimensional reduction!

Dimensional Reduction

Two approaches are available to reduce dimensionality

- **Feature extraction:** creating a subset of new features by combinations of the existing features
- **Feature Selection:** choosing a subset of all the features
- Given a feature space $x_i \in R^N$ find a mapping $y = f(x): R^N \rightarrow R^M$ with M<N such that the transformed feature vector $y \in R^M$ preserves (most of) the information or structure in R^N
- An optimal mapping is one that does not increase error

What Does Dimensional Reduction Mean For You?

Feature Selection: choosing a subset of all the features

- Well, that's what we were already doing
- We use our own expert knowledge to select features of interest
- Note we could also use machine learning and statistics here too if we were so inclined

What Does Dimensional Reduction Mean For You?

Feature extraction: creating a subset of new features by combinations of the existing features

- You can think of this in terms of the NHANES data
- Let's say I do an analysis and I find a correlation between SOFA consumption and sodium consumption (the more SOFAs, the more sodium)
- I could combine SOFAs and sodium into a new variable – SSOFAS
- I make up some formula to do this:
  - $SSOFAS = \alpha * SOFAs + \beta * sodium$
- So basically we just combine variables that are correlated

Principle Components Analysis

One of the most commonly applied dimension reduction techniques

PCA is a deterministic analytical procedure that utilizes an orthogonal transformation to reduce a set of sample observations with potentially correlated variables into a set of uncorrelated variables called principal components

The number of principal components will always be less than or equal to the original number of variables in the sample set

. T. Jolliffe, “Principal Component Analysis," *Springer Series in Statistics*, Springer, NY, 2002.

Principle Component Analysis

The main limitation of PCA is that it does not consider class separability since it does not take into account the class label of the feature vector

- PCA simple performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance
- There is no guarantee that the directions of maximum variance will contain the most interesting/important features

That sounded really hard!

- *Yes, but we have software to help us!*
- *In JMP Principal components can be accessed through the Principal Components command on the Analyze > Multivariate Methods menu.*

- *But, how does principal component analysis help me?*
- *What can principal component analysis tell me?*

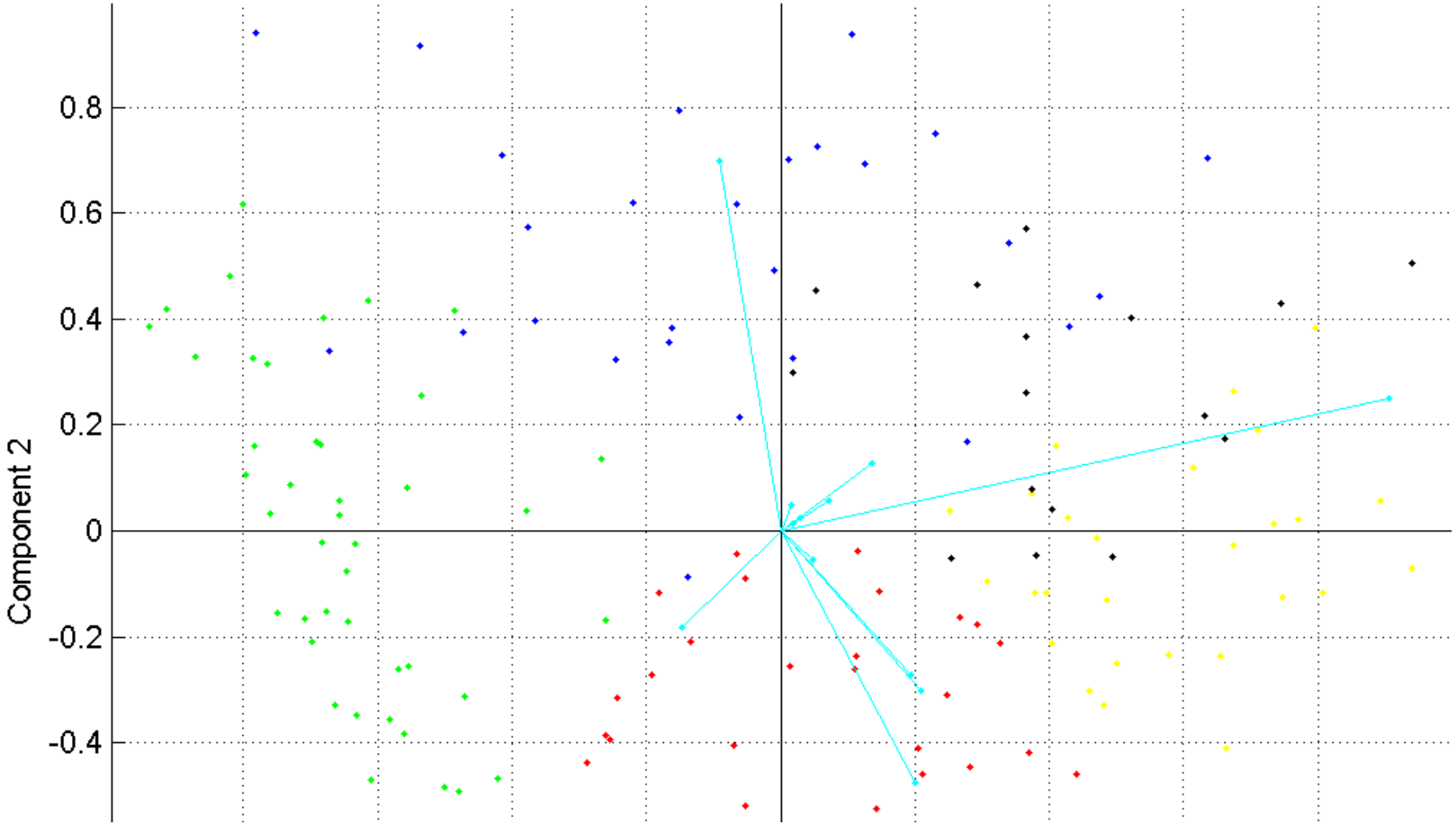
Principal Component Analysis

- *Let's go back to NHANES for an example and focus on the HEI values.*
- *Each person in my data set has 12 HEI values, so you can think of each person as a point in 12-dimensional space*
- *It is hard to understand all the relationships in this space, PCA tries to figure out what dimensions to combine for you automatically*
- *After you run PCA you will get a set of what you can think of as new variables*
- *Since there were 12 variables in our NHANES data, we will get at most 12 principal components (but hopefully less!)*

Principal Component Analysis

- *Ok, you've convinced me to run PCA on NHANES and I get 8 components, but now I'm confused!*
- *The output is such that your first component accounts for as much of the variability in the data set as possible*
- *The second component accounts for the second most variability, etc.*
- *Typically we will plot the data using the first two principal components as axes*
- *We can also write out each component to analyze it*

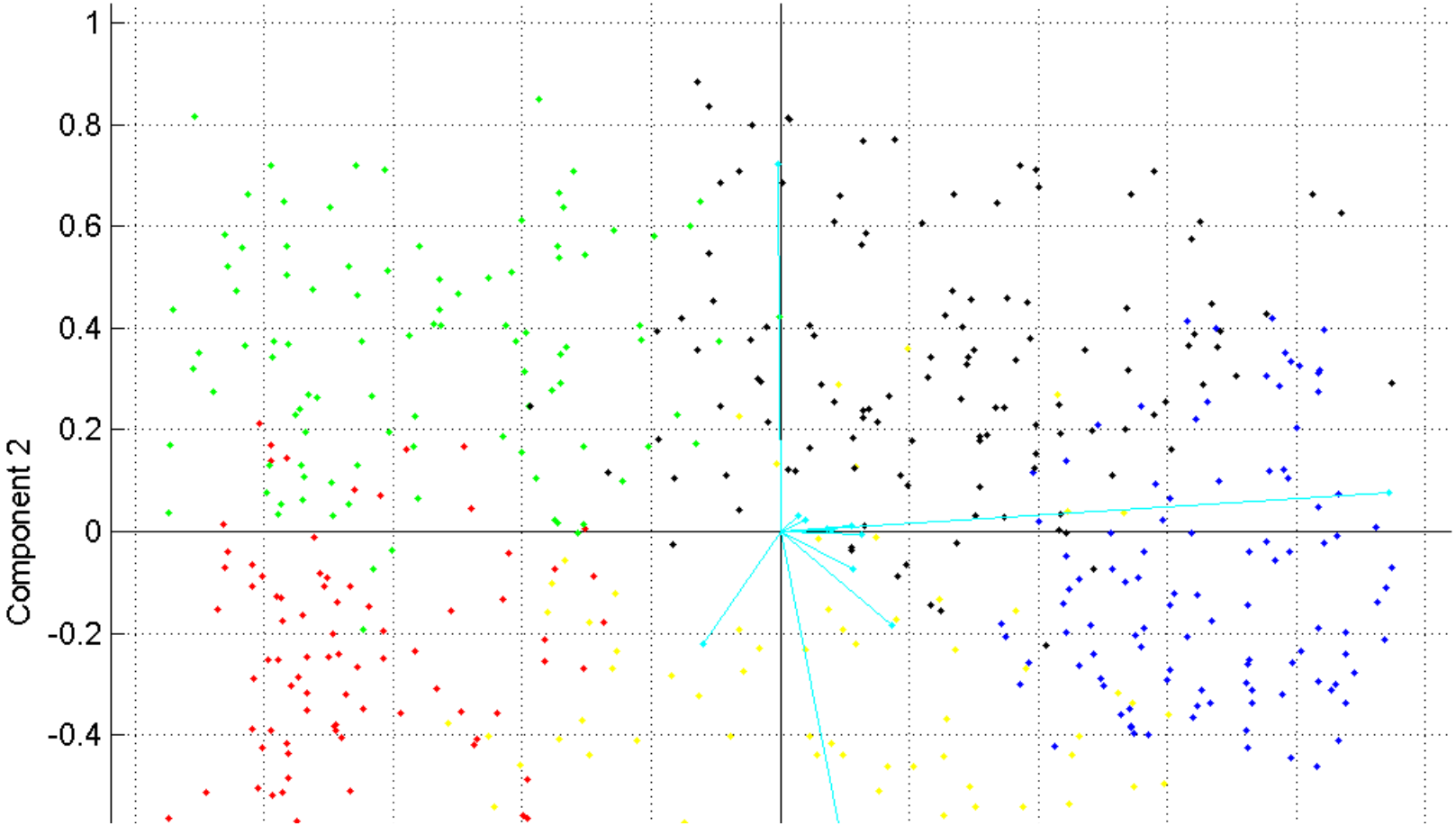
Underweight Adolescent Males



FV1 ≈ 0.14 HEI1 + 0.07 HEI2 +0.02 HEI5 + 0.02 HEI6 + -0.09 HEI7 + 0.19 HEI8 +0.05 HEI3 + 0.03 HEI4 + 0.20 HEI9 + 0.20 HEI10 + -0.15 HEI11 + 0.91 HEI12

FV2 ≈ 0.13 HEI1 + 0.06 HEI2 + 0.05 HEI5 + 0.01 HEI6 + 0.70 HEI7 + -0.27 HEI8 + -0.05 HEI3 + 0.02 HEI4 + -0.48 HEI9 + -0.30 HEI10 + -0.18 HEI11 + 0.25 HEI12

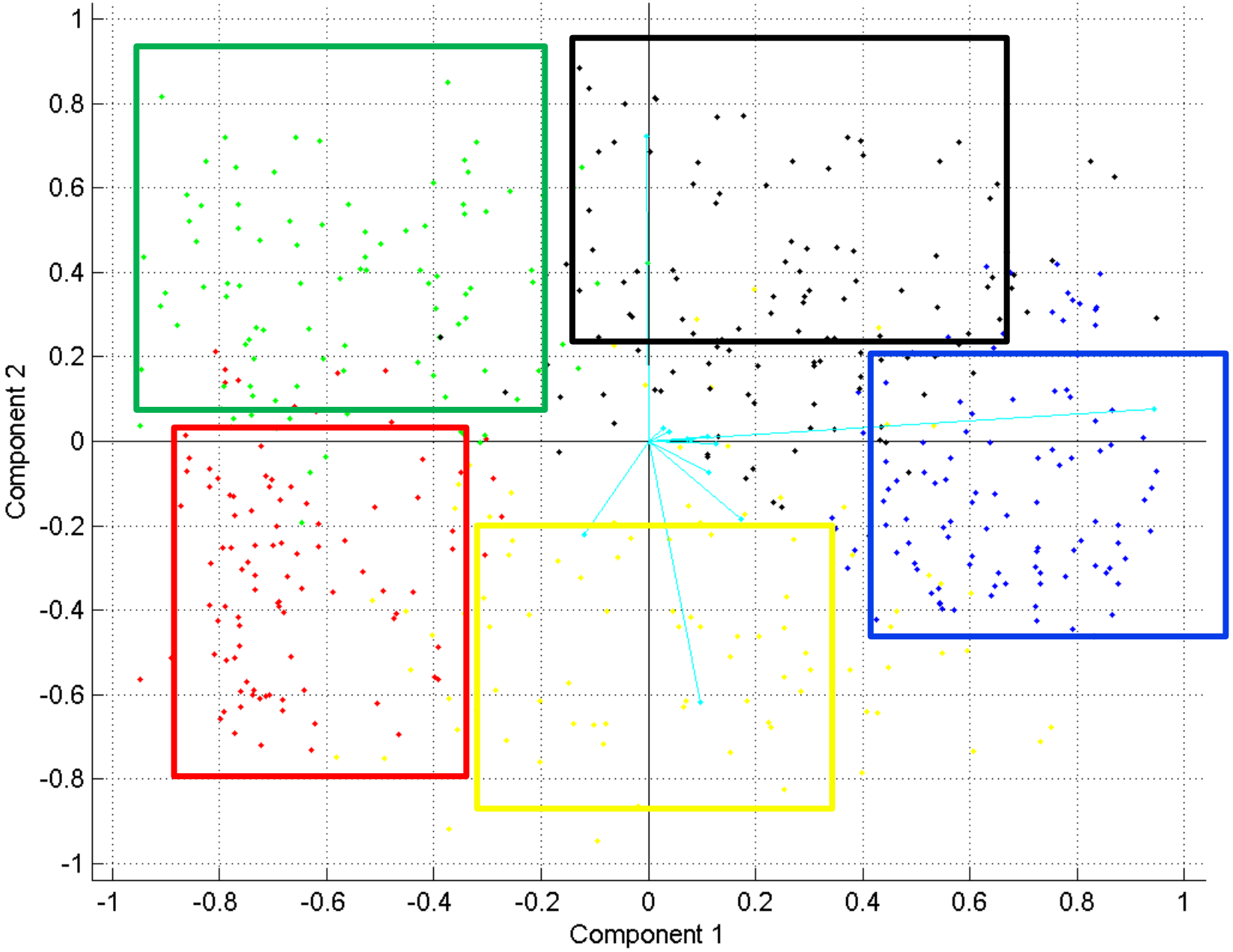
Females Age 19-30, Average BMI



FV1 ≈ 0.13 HEI1 +0.11 HEI2 +0.04 HEI5 + 0.03 HEI6 + -0.00 HEI7 + 0.11 HEI8 +0.08 HEI3 + 0.07 HEI4 + 0.17 HEI9 + 0.10 HEI10 +-0.12 HEI11 + 0.94 HEI12

FV2 ≈-0.01 HEI1 + 0.01 HEI2 +0.02 HEI5 + 0.03 HEI6 + 0.72 HEI7 + -0.07 HEI8 + 0.00 HEI3 + 0.00 HEI4 + -0.18 HEI9 +-0.62 HEI10 +-0.22 HEI11 +0.08 HEI12

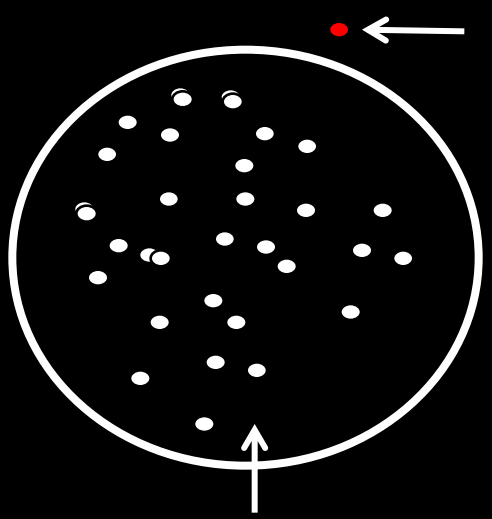
Females Age 19-30, Average BMI



Putting it all together

- ***If I know things about you …***
  - Age, gender, BMI
- ***Can I guess your typical daily diet?***
  - Probably not very accurately
- ***But, if I ask you a few questions, maybe I can get closer!***

Putting it all together



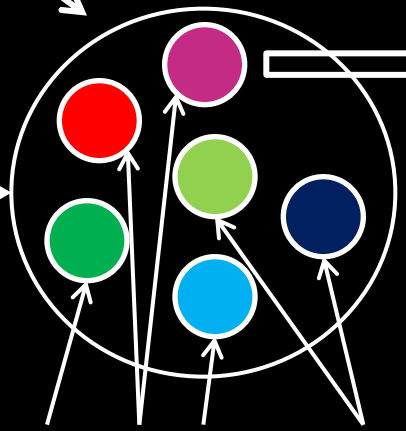
This is you.
Maybe you tell me your age, gender, BMI?
(Inform me maybe?)

This is the NHANES data
I know everyone's diet

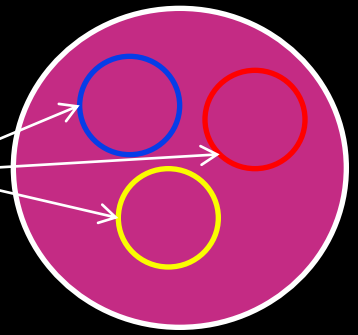
I can first split the data by demographics

I can stick you in one of these groups based on your demographics

Total Population



Split by demographics
(Learned or *a prior*)



Now my demographics groups having groups based on diet!

Now I know your diet!
(Well sort of)

This is still you.
I know you're in this large group
But I want to know your subgroup.

If I know how many sodas you drink, maybe I can narrow it down from three groups to two

Maybe if you ate a salad today tells me more!

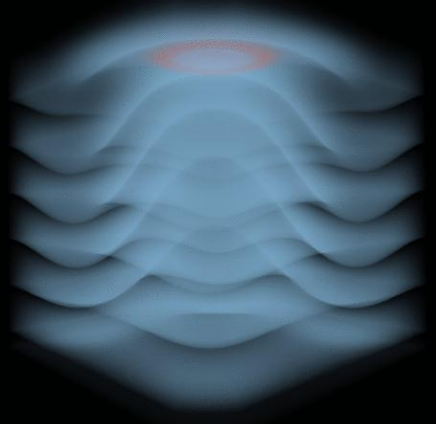
?

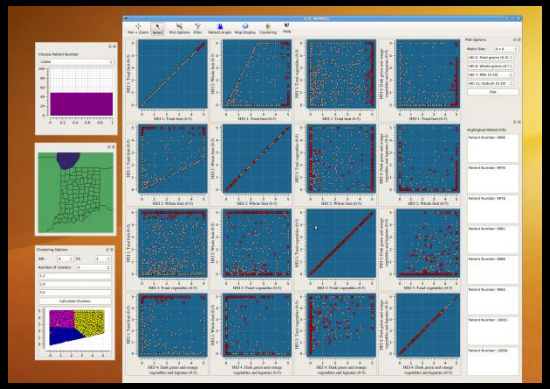
?

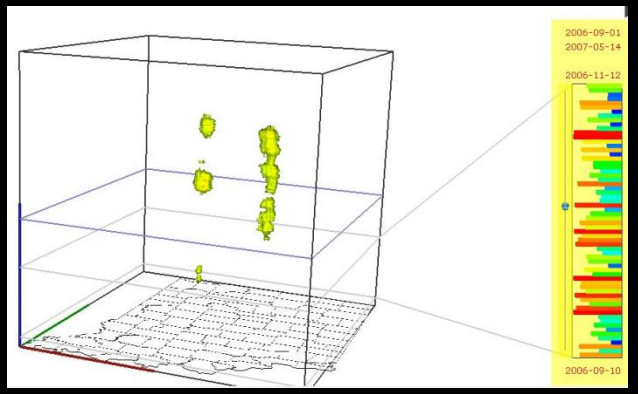
Summary

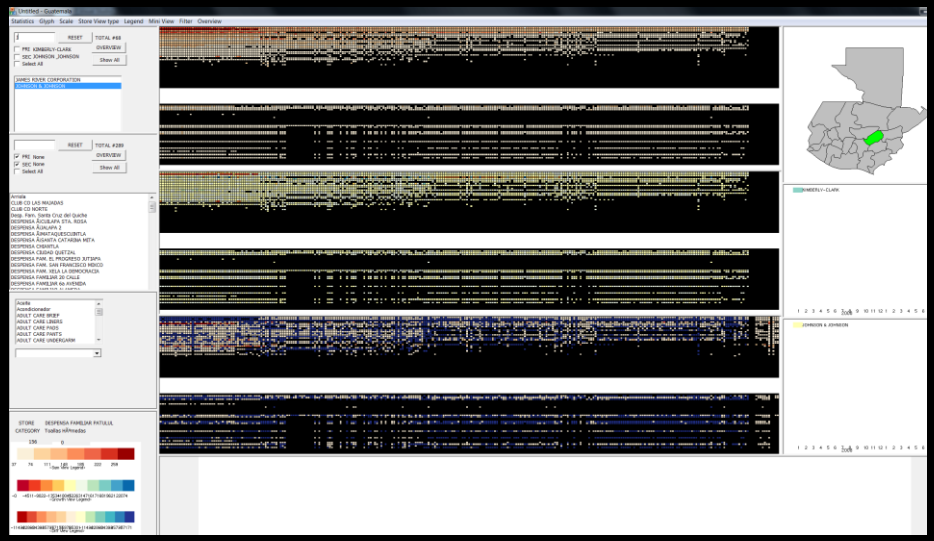
- ***Data analysis is hard!***
- ***We need better tools to be data detectives***
- ***By being data detectives, we can ask questions and maybe create predictions about unknown data based on available data (we make a model of the data)***
- ***Visual analytics helps us in the data detective work!***
- ***But we still need to understand what the analyses do for us so that we can put them into practice.***

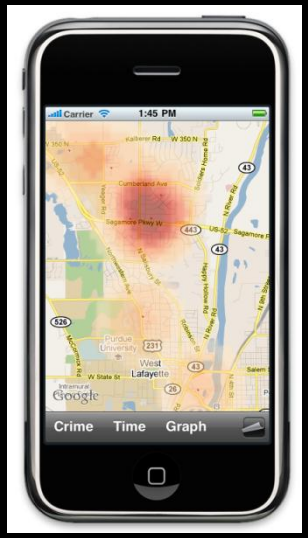
Questions? (http://rmaciejewski.faculty.asu.edu)











VISUAL ANALYTIC TOOLS FOR GEO-GENEALOGY AND GEO-DEMOGRAPHICS

JOSE IBARRA

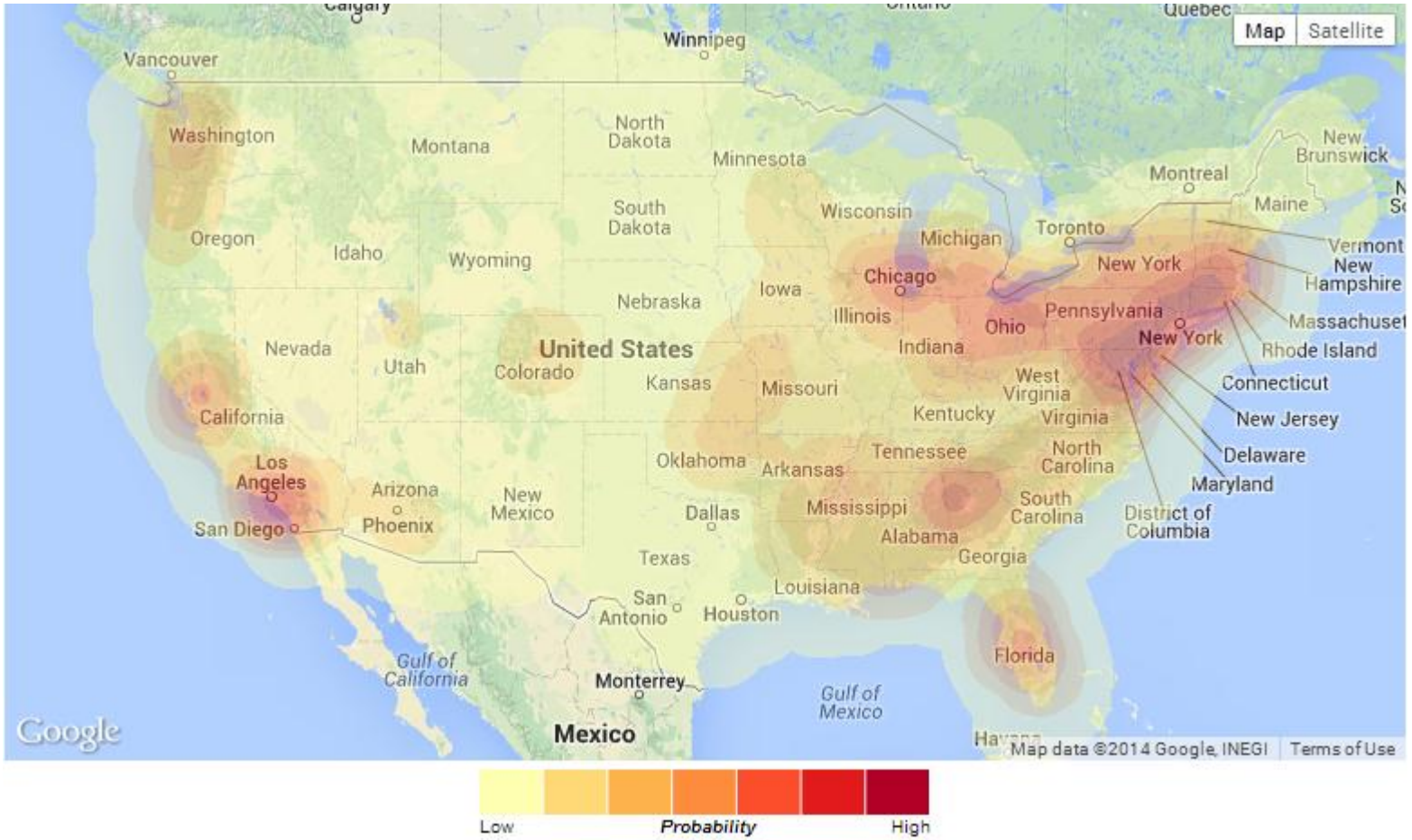
VISUAL ANALYTICS AND DATA EXPLORATION LAB



Visual Analytics &
Data Exploration Research

OVERVIEW

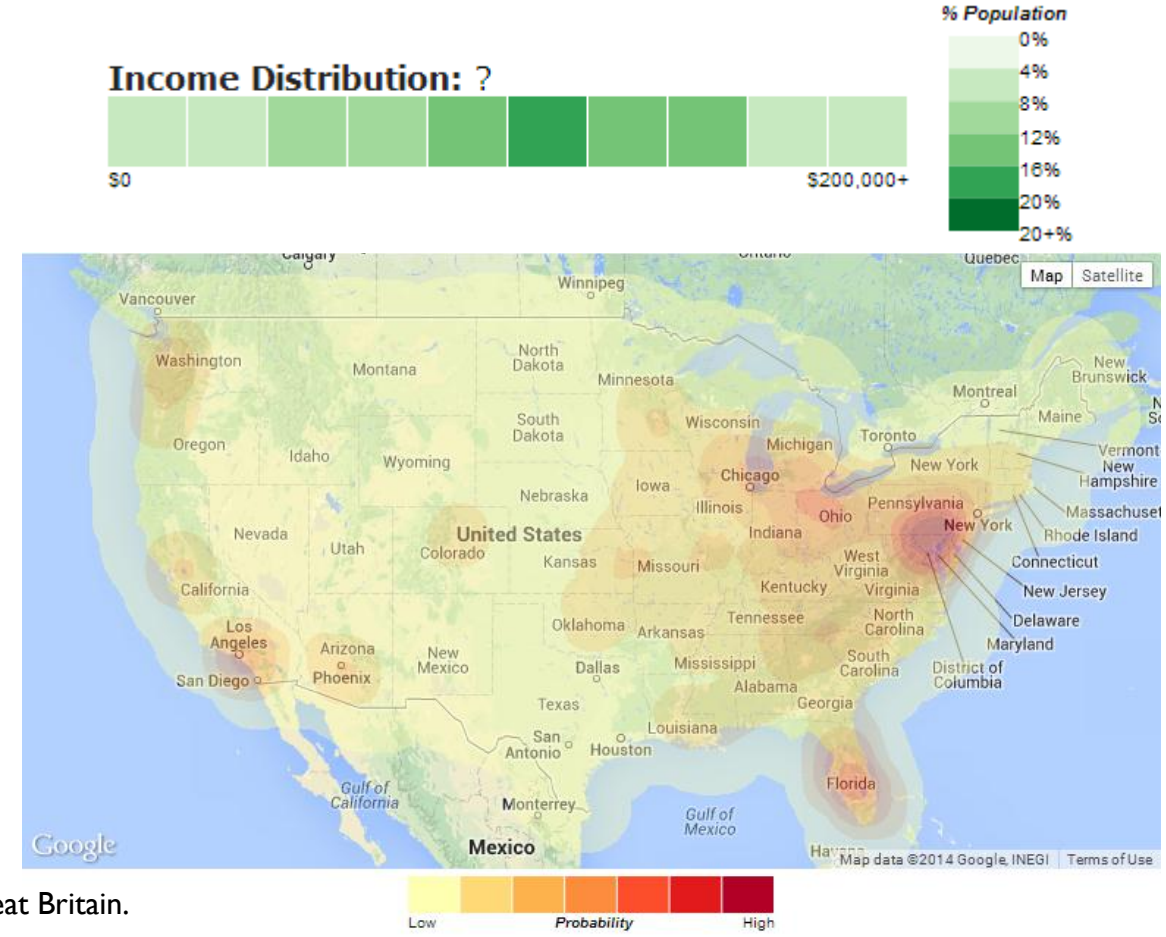
- Introduction
- System
- Density Estimation
- Income Estimation
- Similarity Exploration
- Demo
- Conclusion



Surname: Gates

INTRODUCTION

- Family names
  - Marker for spatial data sets
- Previous spatial work
  - Cheshire and Longley[1]
  - Expansion to other Areas[2]



Surname: Miller

1. CHESHIRE J.A., LONGLEY P.A., SINGLETON A. D.: The surname regions of Great Britain. Journal of Maps 6, 1 (2010), 401–409.
2. CHESHIRE J.A., LONGLEY P.A., YANO K., NAKAYA T.: Japanese surname regions. Papers in Regional Science 92 (2013), In Press.

NAME PROFILER SYSTEM

- Visual analytics mashup[1]
- Utilized United States data
  - Phonebook and Census data stored in MySQL
- Visualize and explore surname spatial distribution
  - Kernel Density Estimation (KDE) heatmaps
- Visualize and explore surname income distributions
  - Color coded histograms
- Explore similar spatial and income distributions between surnames
  - Word clouds of similar names

1. WOOD J., DYKES J., SLINGSBY A., CLARKE K.: Interactive visual exploration of a large spatio-temporal dataset: Reflections on a geovisualization mashup. IEEE Transactions on Visualization and Computer Graphics 13, 6 (2007), 1176–1183.

Surname: ibarra | Regular Distribution ? | Submit

(a)

Income Distribution: ?

$0 — $200,000+

% Population
0%
4%
8%
12%
16%
20%
20+%

Distribution Tool: ? Reset

(c)

$0 — $200,000+

Name Distribution Heatmap: ?

(b)



Low — *Probability* — High

Similar Names Word Cloud Type: ? Core Map Similarity

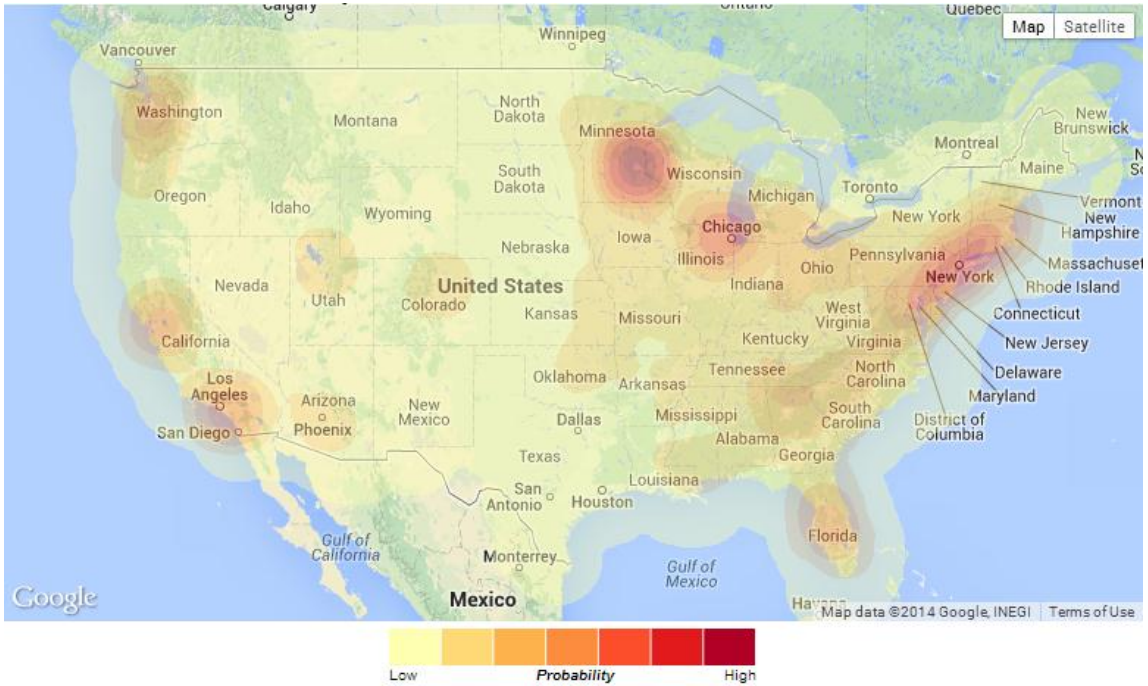
(d)



Low — *Similarity* — High

DENSITY ESTIMATION & AGGREGATION

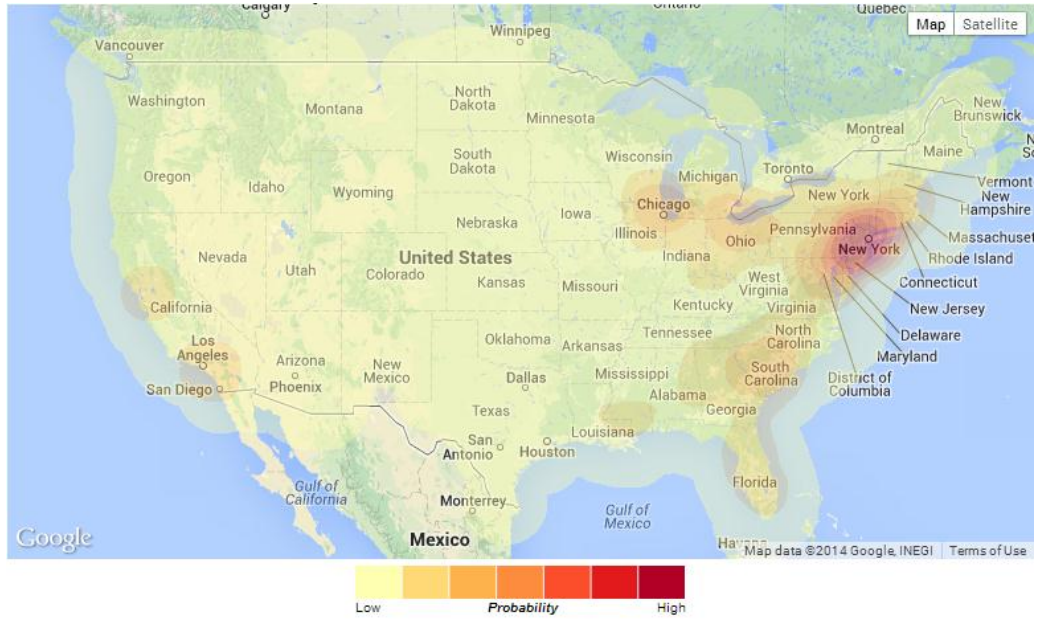
- Heatmap visualization of surnames
  - Probability density function
- Different heatmap types
  - Magnitude
  - Probability distribution
- Only for names with more than 100 records



Surname: Anderson

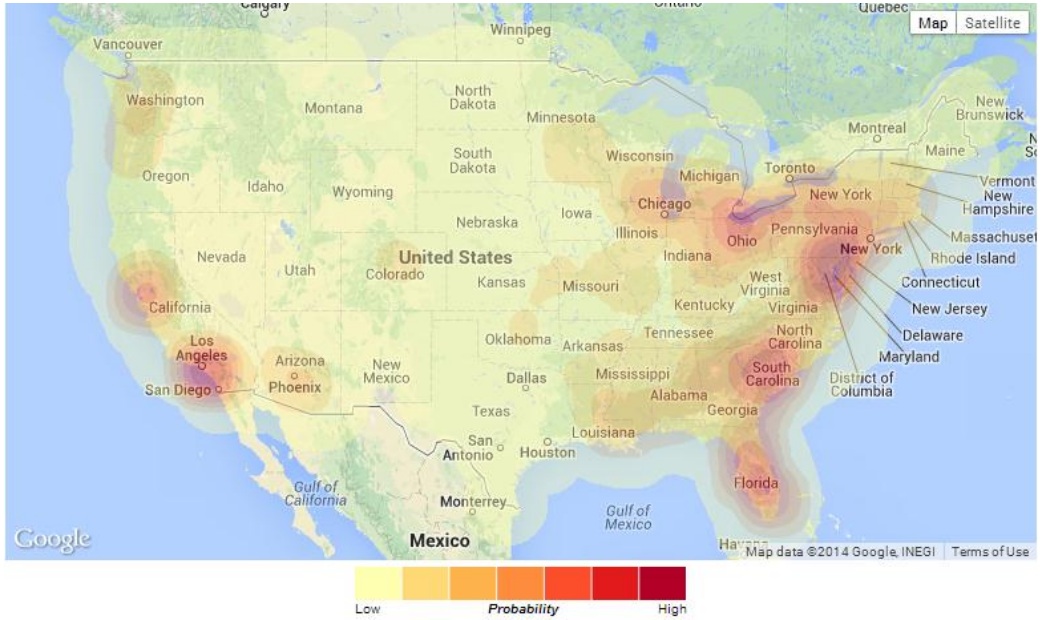
DENSITY ESTIMATION & AGGREGATION

Magnitude



Surname: Mack

Probability Distribution



Surname: Mack

DENSITY ESTIMATION & AGGREGATION

$$\hat{f}_{\mathbf{h}}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N}\left(\prod_{j=1}^{d}\frac{1}{\mathbf{h}_j}K\left(\frac{\mathbf{x}_j - \mathbf{X}_{ij}}{\mathbf{h}_j}\right)\right)$$

- **h** = smoothing pattern
- N = total number of samples
- d = data dimensionality
- K = kernel function (Epanechnikov)

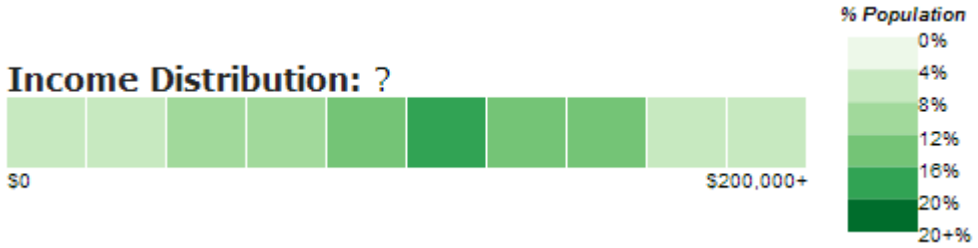
EPANECHNIKOV KERNEL

$$K(u) = \frac{2}{\pi}(1 - u^2)1_{\{u \leq 1\}}$$

- $u^2 \leq 1$
  - Evaluates to one if true
  - Evaluates to zero if false

INCOME ESTIMATION

- 2008-2012 American Community Survey Estimates[1]
- Income distributions for census tract
  - Percentage of population within ten specified income ranges
- Each surname mapped to census tract



Surname: Paterson

1. U.S. CENSUS BUREAU : 2008-2012 American Community Survey 5-Year Estimates, 2013.

INCOME ESTIMATION

$$DX = B$$

- D_{ij} = number of surname records in census tract
  - i = census tract
  - j = surname
- B_{lk} = percentage of census tract population in income bin
  - l = census tract
  - k = income bin
- Solved using non-negative least square solver[1]

1. LAWSON C. L., HANSON R. J.: Solving Least Squares Problems. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1995.

SIMILARITY EXPLORATION: DENSITY ESTIMATION

- Precomputed density estimates
  - Fixed zoom level
  - 170x90 resolution
- L^2 Norm and Core Distance comparison



Surname: Fernandez



Core Distance



Surname: Wong

SIMILARITY EXPLORATION: L^2 NORM

- Sum of squares
- 40 minutes to compute

$$||\mathbf{x}-\mathbf{y}|| = \sum_{k=1}^{n} (x_k - y_k)^2$$

AKERS





LILLARD



SIMILARITY EXPLORATION: CORE DISTANCE

- Previous work by Cheshire and Longley[1]
- Core selection
  - Five largest local maxima from each density estimate
- 30-50 seconds of computation time



Surname: Longley





Surname: Littlefield

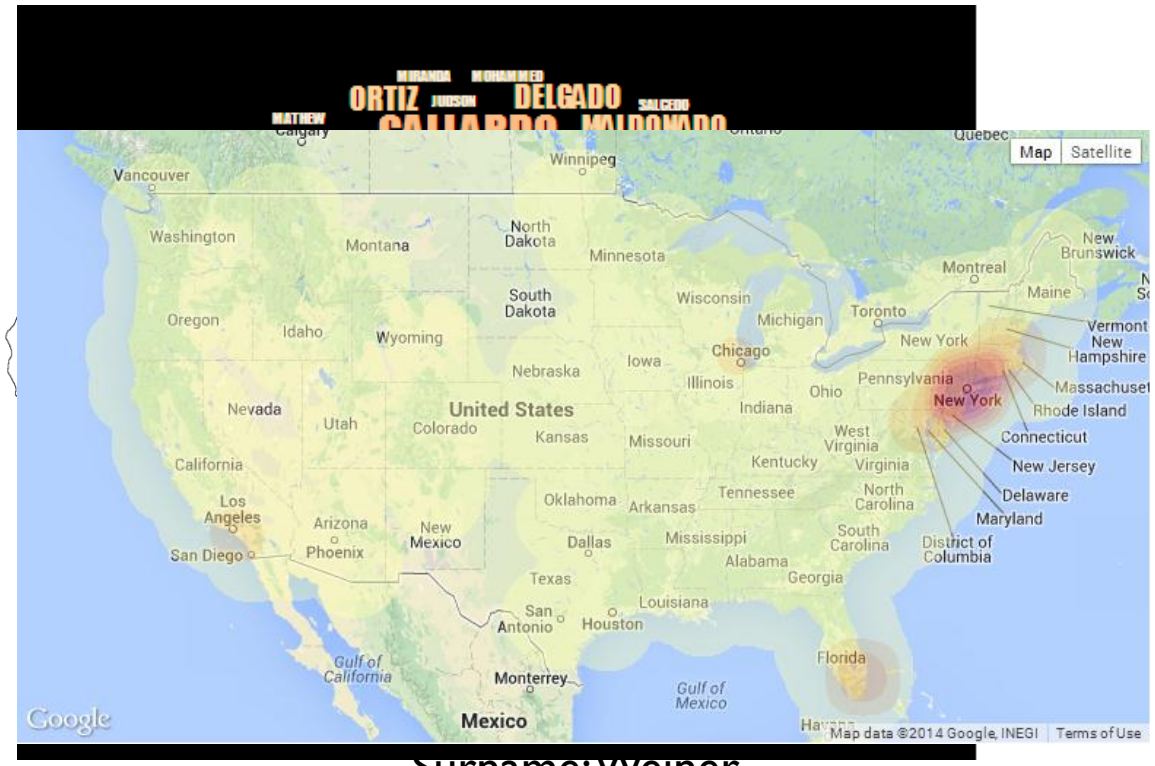
1. CHESHIRE J.A., LONGLEY P.A.: Identifying spatial concentrations of surnames. International Journal of Geographics Information Science 26 (2012), 309–325.

CORE DISTANCE CALCULATION

- Select five highest local maxima
- Store maxima for each surname
- Compare maxima between each surname



Surname: Ibarra

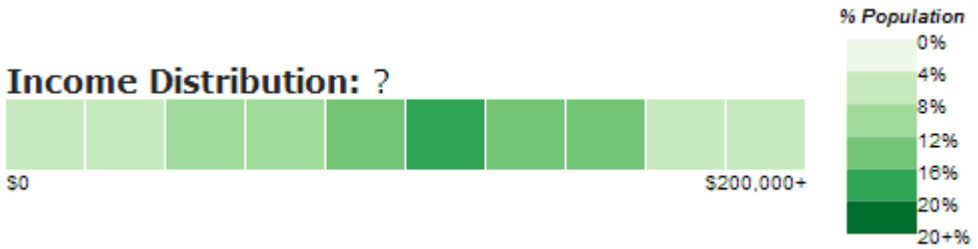


Surname: Weiner

SIMILARITY EXPLORATION: INCOME ESTIMATION

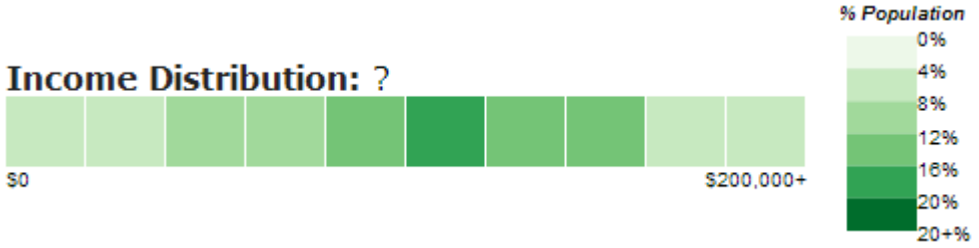
- L^2 norm comparison
- User defined comparison
  - L^2 norm comparison

$$||\mathbf{x} - \mathbf{y}|| = \sum_{k=1}^{n} (x_k - y_k)^2$$

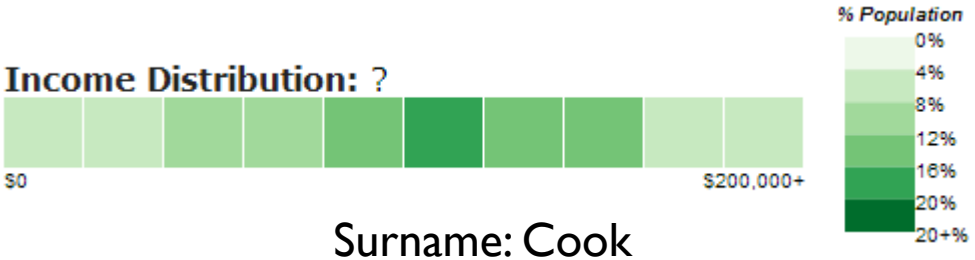


Surname: Maciejewski

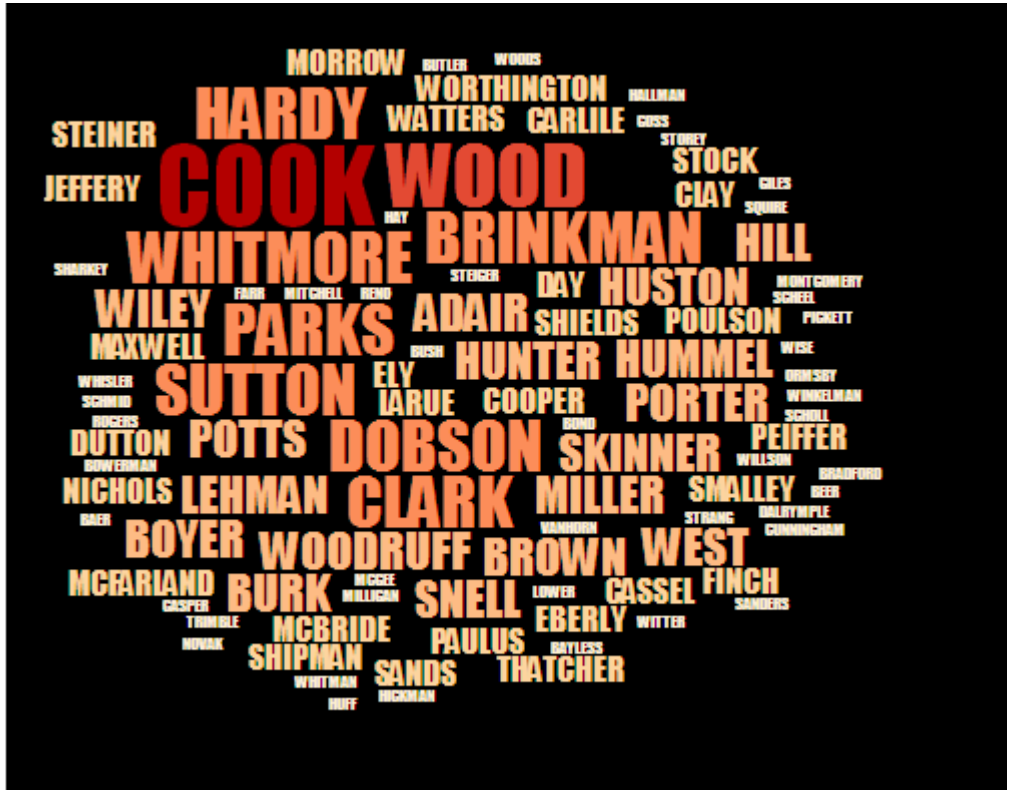
INCOME SIMILARITY EXPLORATION: L^2 NORM



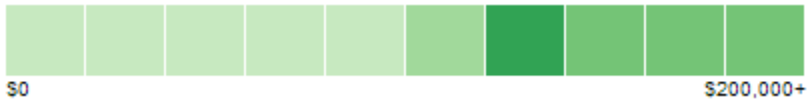
Surname: Smith



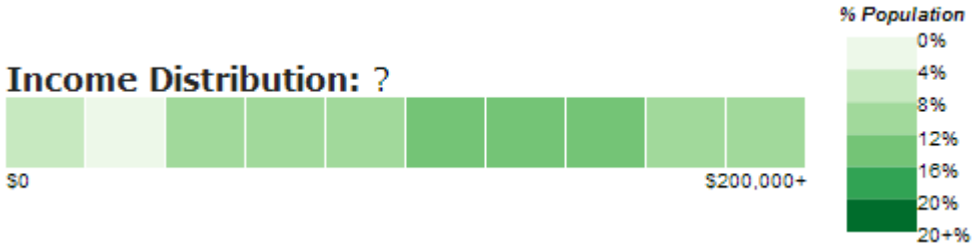
Surname: Cook



INCOME SIMILARITY EXPLORATION: USER DEFINED INCOME DISTRIBUTION



User Defined Distribution



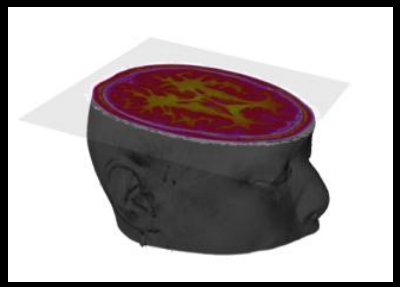
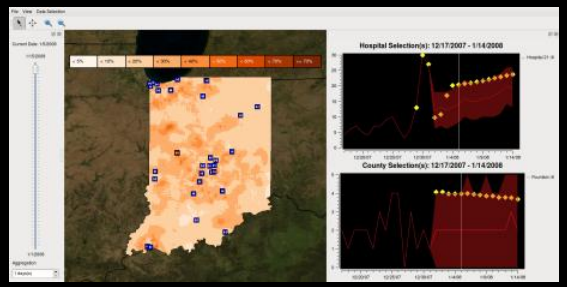
Surname: Mehta

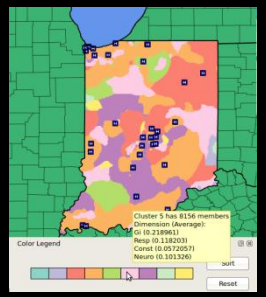


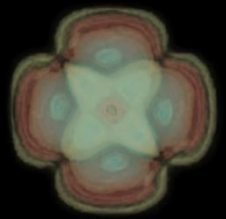
DEMO

http://ec2-54-201-211-102.us-west-2.compute.amazonaws.com:8080/nameweb/html/

Questions?



Visual Analytics Summer School 2013: Fundamentals of Visual Analytics

Ross Maciejewski



Ross Maciejewski
E-mail: rmacieje@asu.edu
http://rmaciejewski.faculty.asu.edu/
http://lab.vader.asu.edu



The Ongoing Data Deluge

- Since 2003, digital information has accounted for 90% of all information produced[1]
- In 2009, drones from Iraq and Afghanistan recorded 24 years of video footage
- In 2010, the amount of information added annually to the digital universe was estimated to be nearly 1 ZB
- Wal-mart process > 1 million transactions per hour
- By 2013 Cisco estimates the annual internet traffic will be 667 EBs

1 – C. R. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, and T. S. Yoo, Eds., *NIH-NSF Visualization Research Challenges Report*, IEEE Press, ISBN 0-7695-2733-7, 2006

The Ongoing Data Deluge

- FedEx's ships more than 8.5 million packages per day (http://www.fedex.com/us/about/today/companies/corporation/facts.html)
- Consumers carry more than 1 billion Visa cards worldwide. More than 450 million of those cards are in the United States (http://www.creditcards.com/credit-card-news/credit-card-industry-facts-personal-debt-statistics-1276.php)

Data Overload

- Opportunity: Huge amounts of data available in digital form and ready for analysis!
  - But, how can we make sense of this data?
  - How can we harness this data in the decision making process?
  - How do we avoid being overwhelmed by all of this data?



The Power of Visualization

- The goal is to take all of this *data* and transform it into *information*
- How many terabytes of data we have collected doesn't matter, it's how many petaflops of *insights* we can generate from this data
- "*The purpose of computing is insight, not numbers.*" Dedication to his book: RW Hamming (1971). *Introduction to Applied Numerical Analysis*. McGraw Hill.
- We need to make the data understandable to people and a key way of doing this is through *visualization*

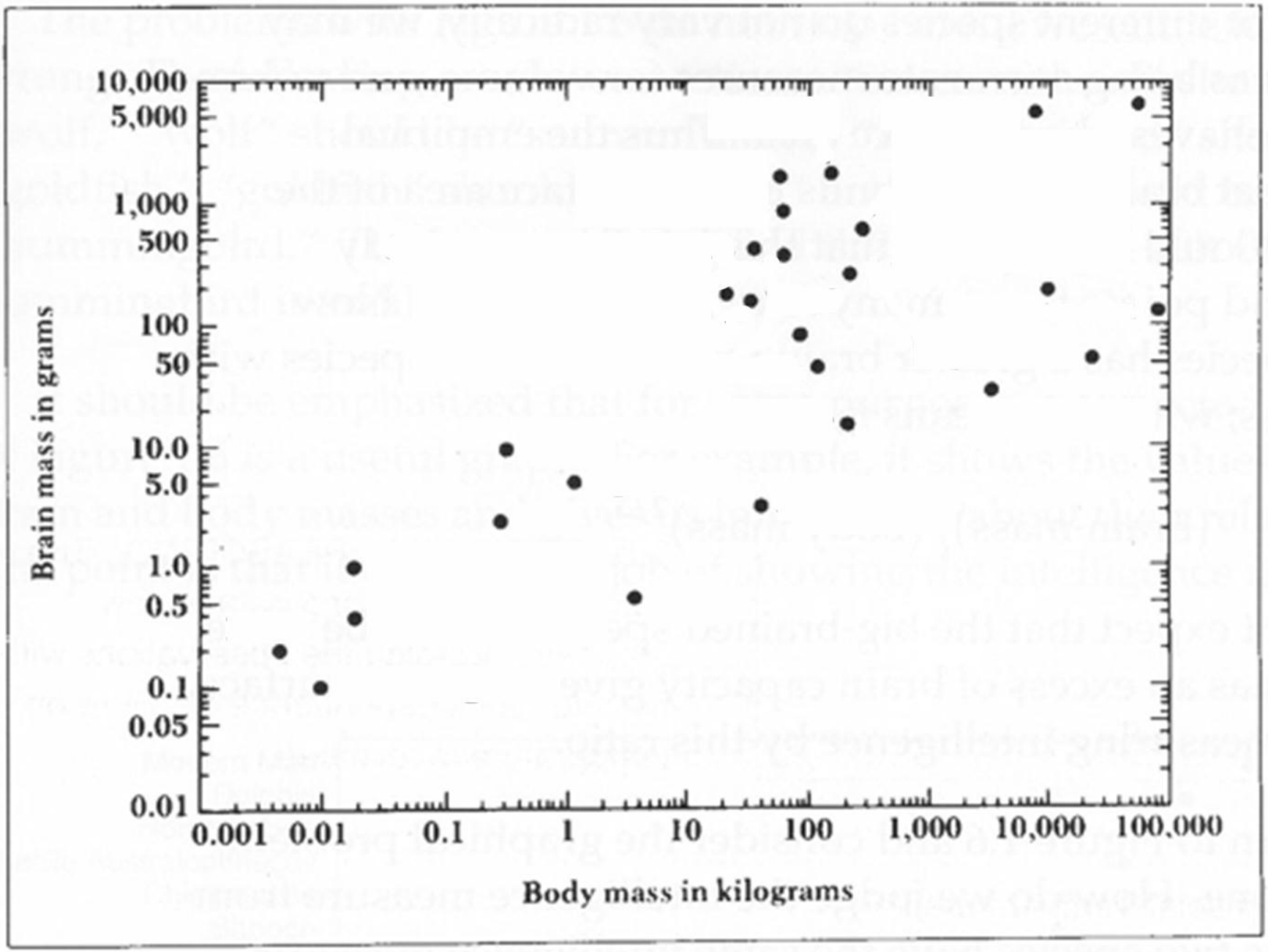
	A	B	C	D
1	Animal	Brain Weight (kg)	Body Weight (kg)	
2	Mountain beav	1.35	465	
3	Cow	465	423	
4	Grey wolf	36.33	119.5	
5	Goat	27.66	115	
6	Guinea pig	1.04	5.5	
7	Dipliodocus	11700	50	
8	Asian elephant	2547	4603	
9	Donkey	187.1	419	
10	Horse	521	655	
11	Potar monkey	10	115	
12	Cat	3.3	25.6	
13	Giraffe	529	680	
14	Gorilla	207	406	
15	Human	62	1320	
16	African elephar	6654	5712	
17	Triceratops	9400	70	
18	Rhesus monke	6.8	179	
19	Kangaroo	35	56	
20	Golden hamste	0.12	1	
21	Mouse	0.023	0.4	

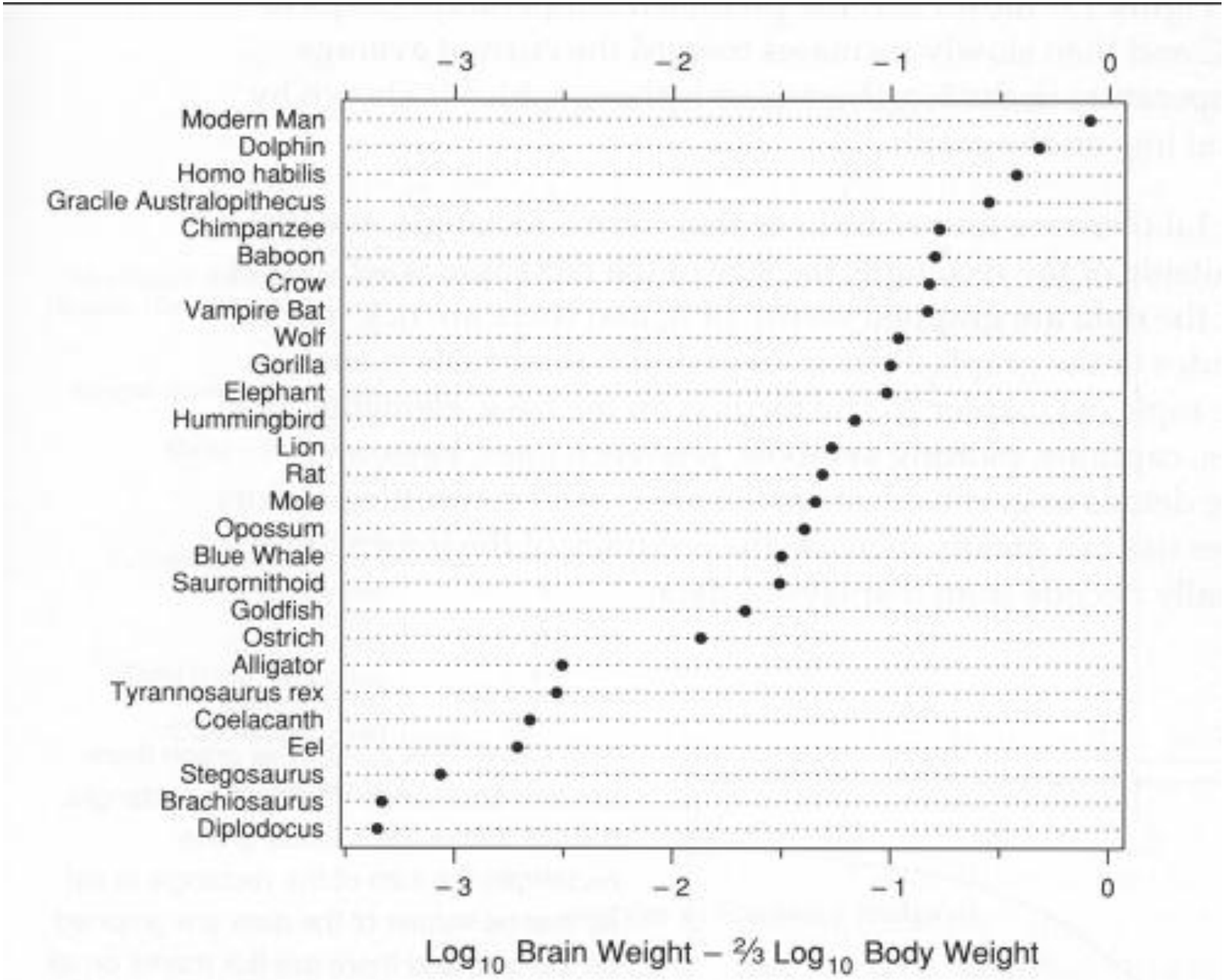
Which animal weights the least/most?

Is there a relationship between brain weight and body weight?

If so, are there any outliers?

Rousseeuw, P.J. & Leroy, A.M. (1987) Robust Regression and Outlier Detection. Wiley, p. 57

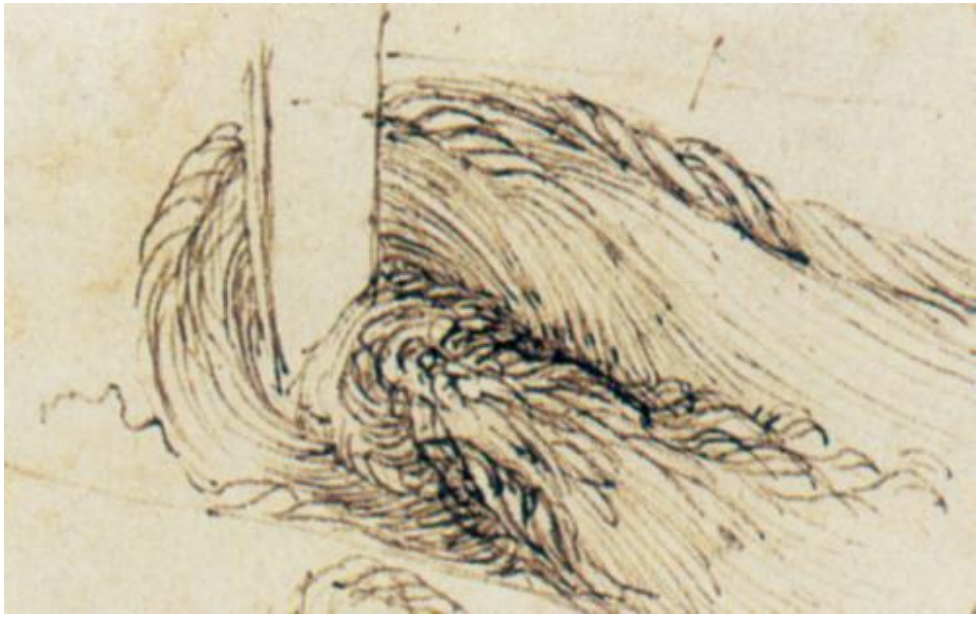

Brain mass in grams
10,000
5,000
1,000
500
100
50
10.0
5.0
1.0
0.5
0.1
0.05
0.01
0.001
0.01
0.1
1
10
100
1,000
10,000
100,000
Body mass in kilograms



WS Cleveland – The Elements of Graphing Data, 1994

Visualization Through the Ages

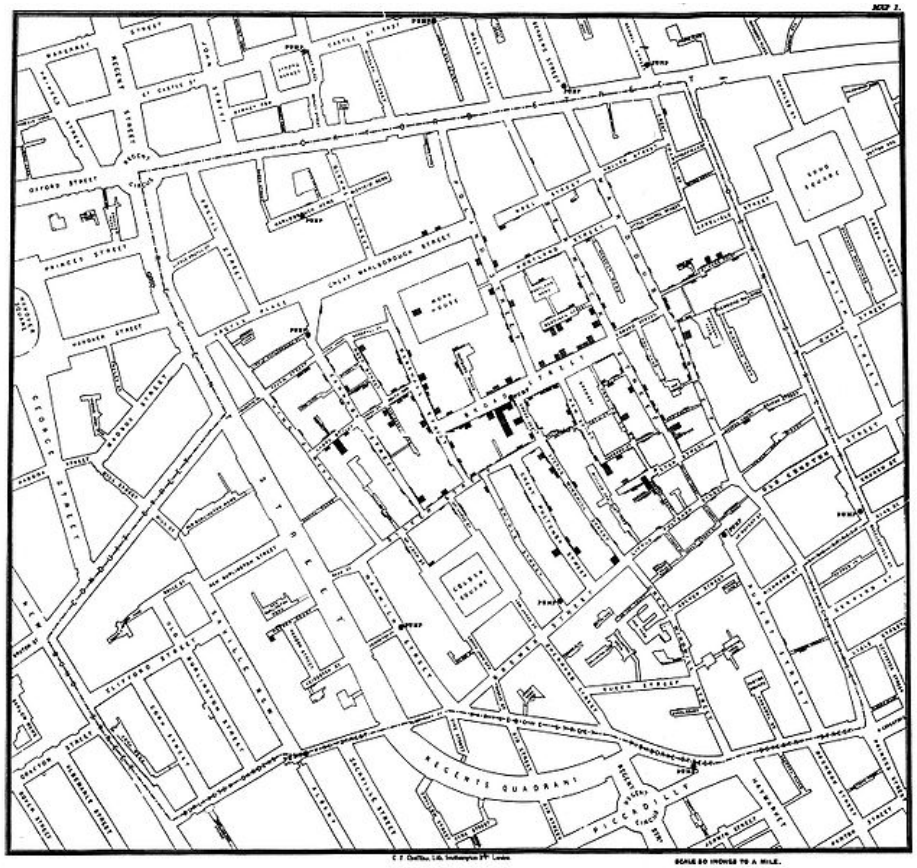
- Hand-drawn illustration[1] of water by Leonardo da Vinci from his studies to determine the processes underlying water flow (1510)



1 – Leonardo da Vinci. *Old Man Seated on Rocky Outcrop, Seen in Profile to the Right, with Water Studies*. Windsor Castle, Royal Liberty, RL 12579r, c. 1510-1513.

Visualization Through the Ages

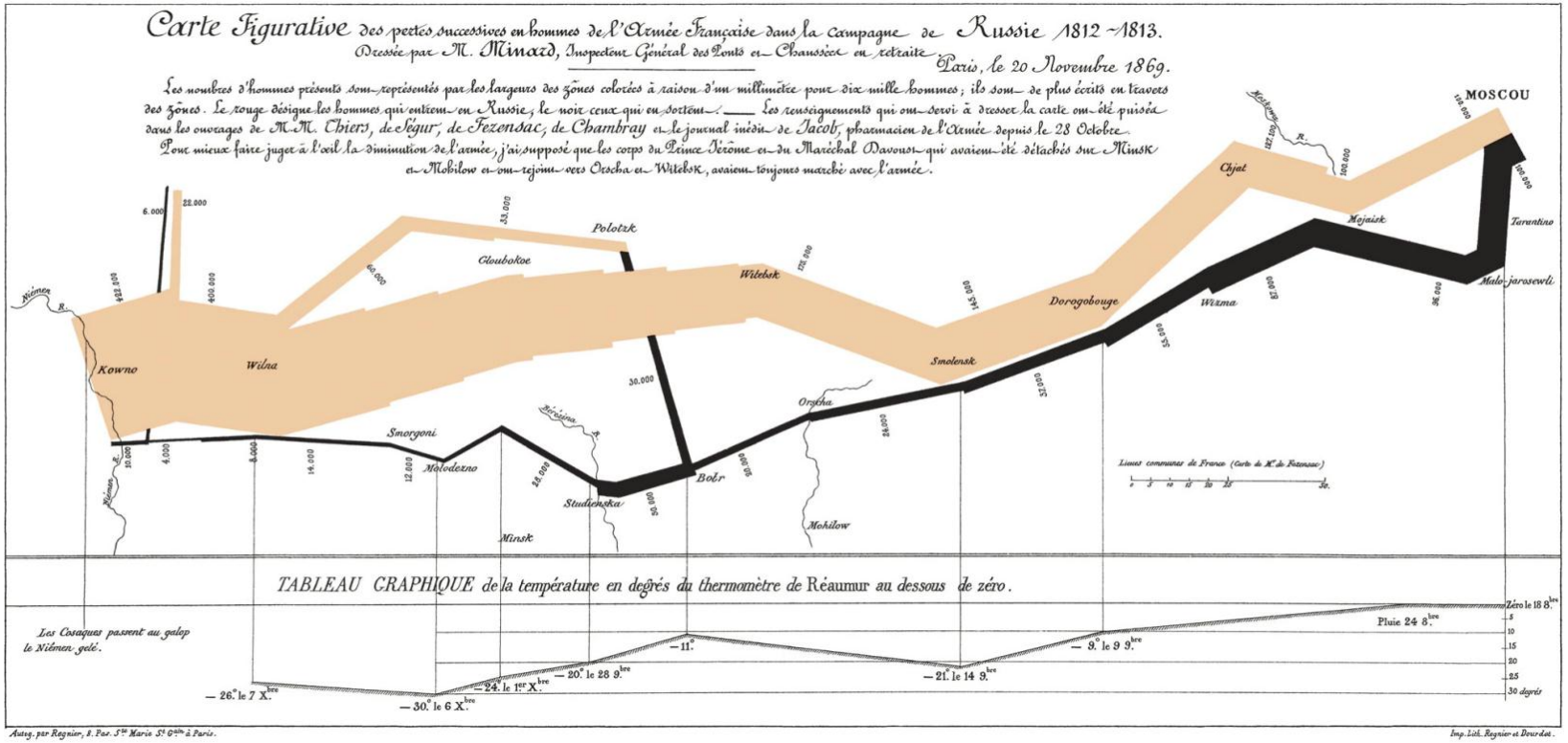
- John Snow's map of Cholera cases in the London epidemic of 1854[1]



1 – The Visual Display of Quantitative Information, E. R. Tufte, 1983

Visualization Through the Ages

- Charles Minard's 1869 flow map[1] showing Napoleon's disastrous Russian campaign of 1812.



1 – Charles Joseph Minard: Mapping Napoleon's March, 1861 by John Corbett, Center for Spatially Integrated Social Science

Visualization Through the Ages

- James Maxwell's[1] thermodynamic surface sculpture (1874)
- Three-dimensional plot of the states of a fictitious water-like substance
- Coordinates are volume (x), entropy (y), and energy (z)

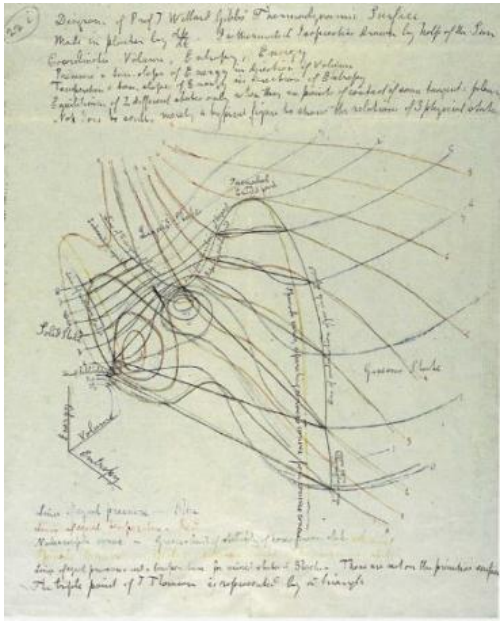


Plate IV. Diagram of the lines on Gibbs' thermodynamic surface, 8 July 1875 (Number 564).



1 – James Clerk Maxwell and P. M. Harman (2002), The Scientific Letters and Papers of James Clerk Maxwell, Volume 3: 1874-1879, Cambridge University Press

Why is Visualization Helpful?

- It utilizes the high bandwidth of the human visual systems
- The human mind is fast and parallel
- Humans are great at visual pattern recognition
- We have pre-attentive visual phenomena
- Visual tools can extend memory and cognitive capacity
- We think visually!
- "A picture is worth a thousand words." – *Printers Ink*, pp. 96-97 December 8, 1921

Why Is Visualization So Helpful?

- Amplifies cognition
  - Expands our working memory by allowing us to offload results
  - Reduces the search time
  - Pattern detection and recognition can be improved through perceptual cues and inference
  - Visualizations can be designed to control attention interaction for improving cognition

Readings in Information Visualization: Using Vision to Think, SK Card, J Mackinlay and B.Shneiderman, 1999

What is Visualization?

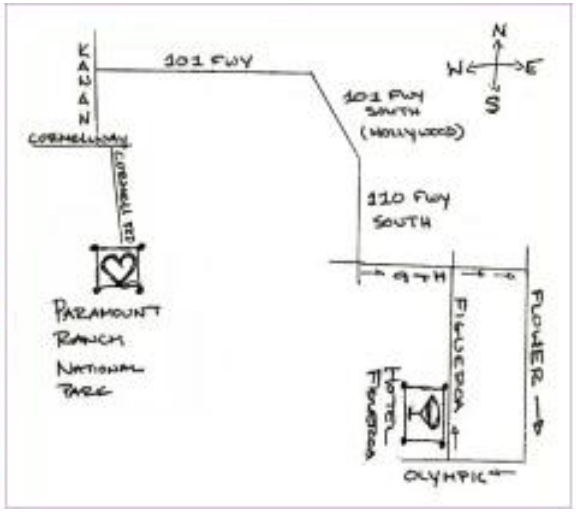
- "The use of computer-supported, interactive visual representations of data to amplify cognition."[1]
- This is not simply the process of making a graphic or an image, the goal is to create insight, not pretty pictures
- We want to help people form a mental image of something and internalize their own understanding
- We want to promote discovery, decision making and explanations

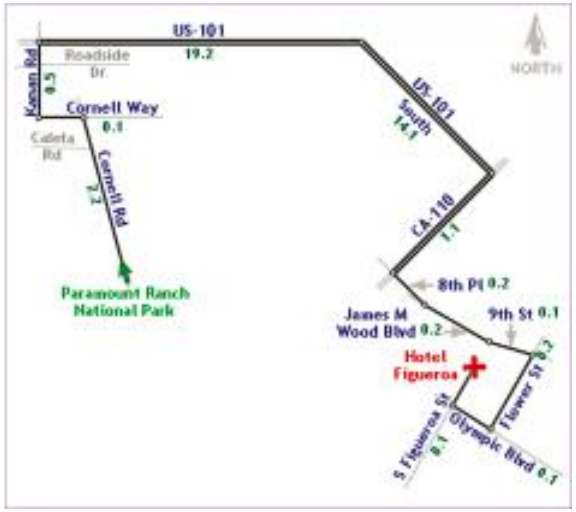
1-Readings in Information Visualization: Using Vision to Think, SK Card, J Mackinlay and B.Shneiderman, 1999

What is Visualization

- We want to find and utilize cognitive and perceptual principles
- We want to optimize our visualizations and our interactions with the visualization according to these principles







Maneesh Agrawala, Chris Stolte: Rendering effective route maps: improving usability through generalization. SIGGRAPH 2001: 241-249

Scientific Visualization

- Primarily relates to and represents something physical or geometrics
- The structure of the data is typically defined or given
- Examples
  - Air flow over a wing
  - Stresses on a girder
  - Organs in the human body
  - Molecular bonding

1 – Descriptions on this slide are borrowed from John Stasko's "InfoVis Overview" lecture: http://www.cc.gatech.edu/~stasko/7450/Notes/overview.pdf

Information Visualization

- Primarily relates to data that does not have a direct physical correspondence
- Notion of the data is abstract
- Examples
  - Baseball statistics
  - Stock trends
  - My social network

1 – Descriptions on this slide are borrowed from John Stasko's "InfoVis Overview" lecture: http://www.cc.gatech.edu/~stasko/7450/Notes/overview.pdf

Purpose of Visualization

- Analysis – Understand your data better and act upon that understanding
  - Given a data set, compare, contrast, assess, evaluate
  - Solve a problem!
- Presentation – Communicate and inform others more effectively
- Visualization is most useful in **exploratory data analysis**[1]

1 – J. W. Tukey. Exploratory Data Analysis

The Information Seeking Mantra

Overview first,

then zoom and filter,

details on demand

1 - B. Shneiderman*. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations*. In the Proceedings of the IEEE Symposium on Visual Languages, pp. 336-343, 1996.

Tasks for Visualization

- **Searching and browsing** - find a specific piece of information, inspect data, seek information
- **Analyze** – do comparisons, differences, look for outliers, extrema and patterns
- **Monitor** – look for changes and trends

"Contained within the data of any investigation is information that can yield conclusions to questions not even originally asked. That is, there can be surprises in the data … To regularly miss surprises by failing to probe thoroughly with visualization tools is terribly inefficient because the cost of intensive data analysis is typically very small compared with the cost of data collection."

W. S. Cleveland

The Elements of Graphing Data

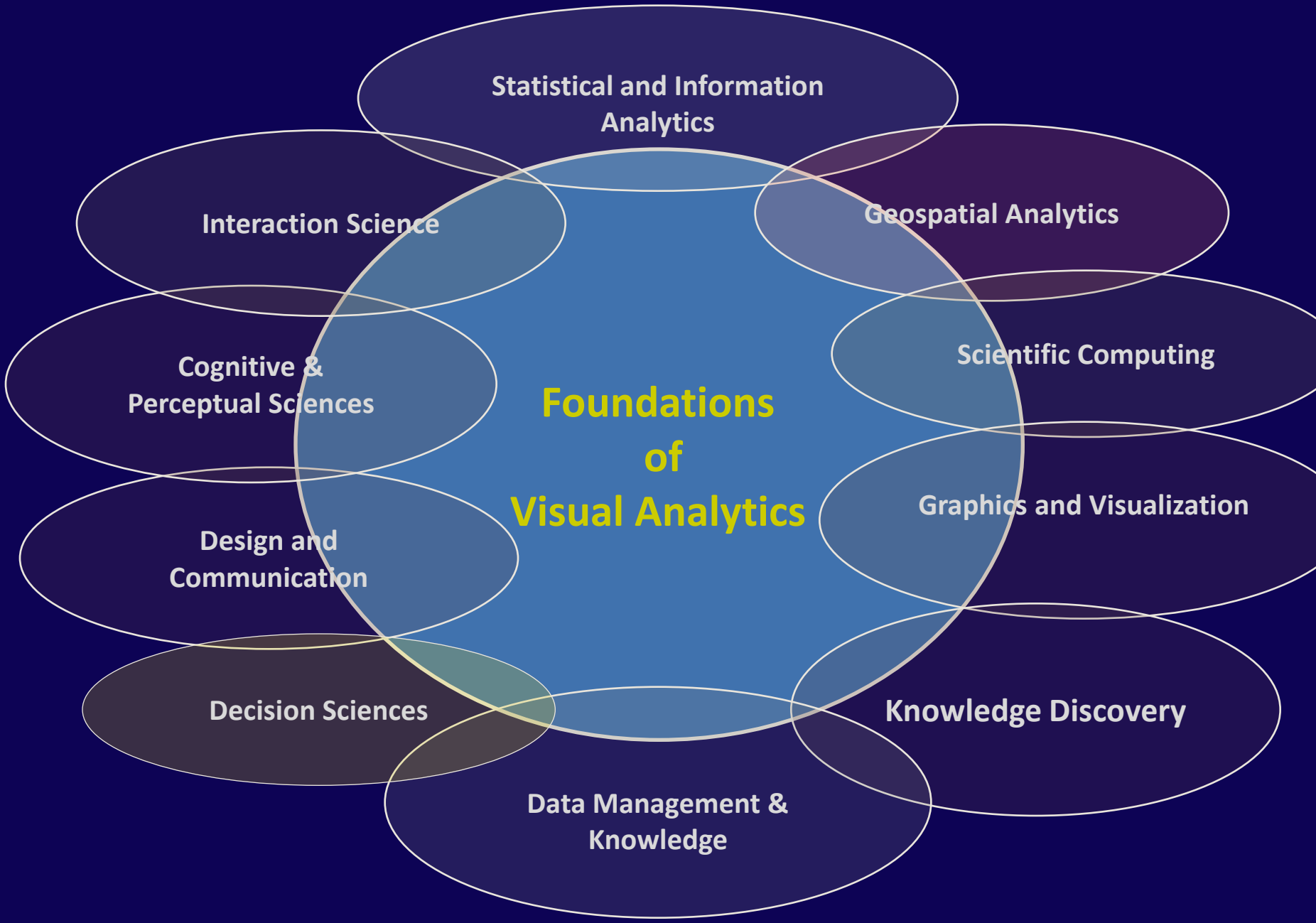
Is This the Total Solution?

- Traditional visualization misses several key factors in how people solve difficult problems
- Visual analytics aims to overcome these shortfalls and create an interactive human-computer exploration and decision making environment

Visual Analytics

- Visualization is good for exploring data, but we can do more than just explore
- "Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces"[1]
- Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets.[2]
- "A graphic display has many purposes but it achieves its highest value when it forces us to see what we were not expecting."[3]

1 - J. J. Thomas and K. A. Cook (Ed.). *Illuminating the Path: The R&D Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.
2 – Keim et al. chapter in *Information Visualization: Human-Centered Issues and Perspectives*, 2008.
3 – H. Wainer. *A Graphic Discovery: A Trout in the Milk and Other Visual Adventures*.


Statistical and Information Analytics
Interaction Science
Geospatial Analytics
Cognitive & Perceptual Sciences
Scientific Computing
Foundations of Visual Analytics
Graphics and Visualization
Design and Communication
Decision Sciences
Knowledge Discovery
Data Management & Knowledge

Visual Analytics

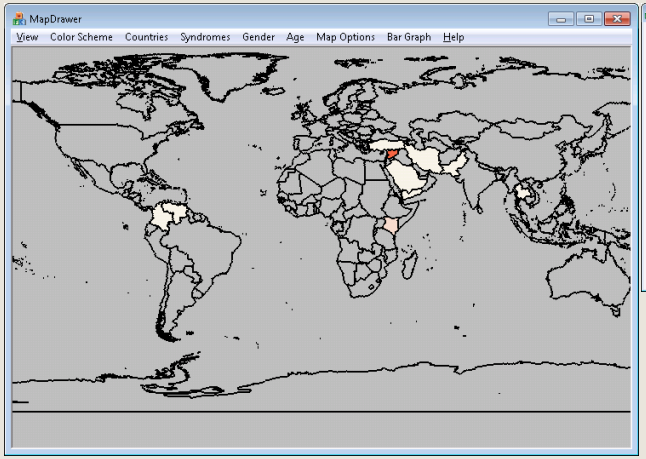
- Visualization is good for exploring data, but we can do more than just explore
- "Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces"[1]
- Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex data sets.[2]
- "A graphic display has many purposes but it achieves its highest value when it forces us to see what we were not expecting."[3]

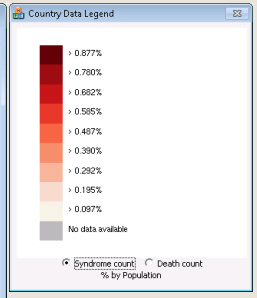
1 - J. J. Thomas and K. A. Cook (Ed.). *Illuminating the Path: The R&D Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.
2 – Keim et al. chapter in *Information Visualization: Human-Centered Issues and Perspectives*, 2008.
3 – H. Wainer. *A Graphic Discovery: A Trout in the Milk and Other Visual Adventures*.

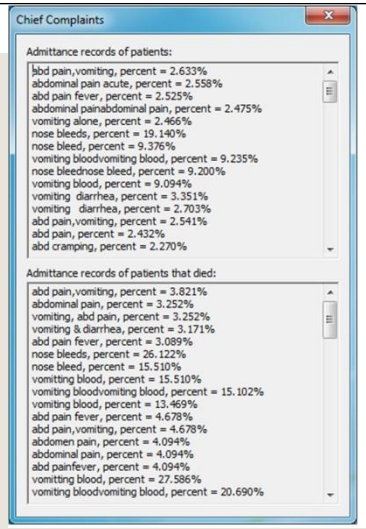
A Quick Example

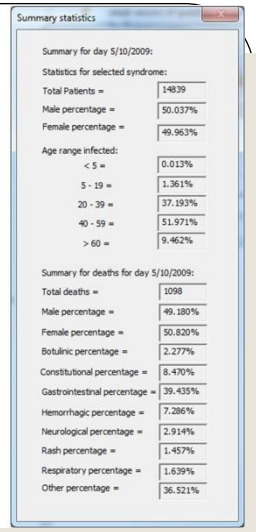
Syndromic Surveillance

- Syndromic surveillance is the detection of adverse health events focusing on pre-diagnosis information to improve response time
- Pre-diagnosis information can consist of multiple data sources:
  - Over the counter medicine sales
  - News reports on emerging diseases
  - Pro-med news feeds
  - Emergency department chief complaints



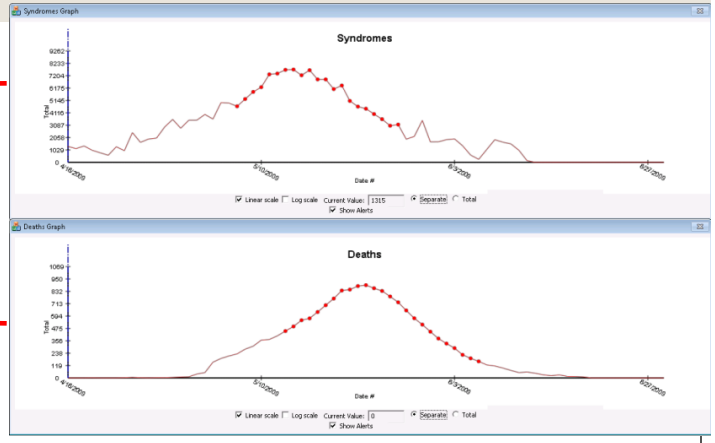


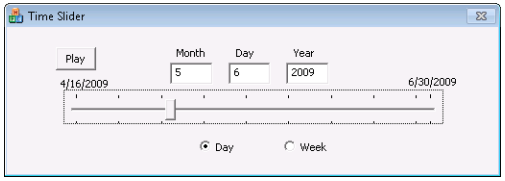




Patient ID	Date	Chief Complaints	Location
9398	4/16/09	Ear pain	Karachi
10816	4/16/09	Stuffy nose	Lebanon
1491	4/16/09	Fever	Allepo
16237	4/16/09	Head bleed	Yemen

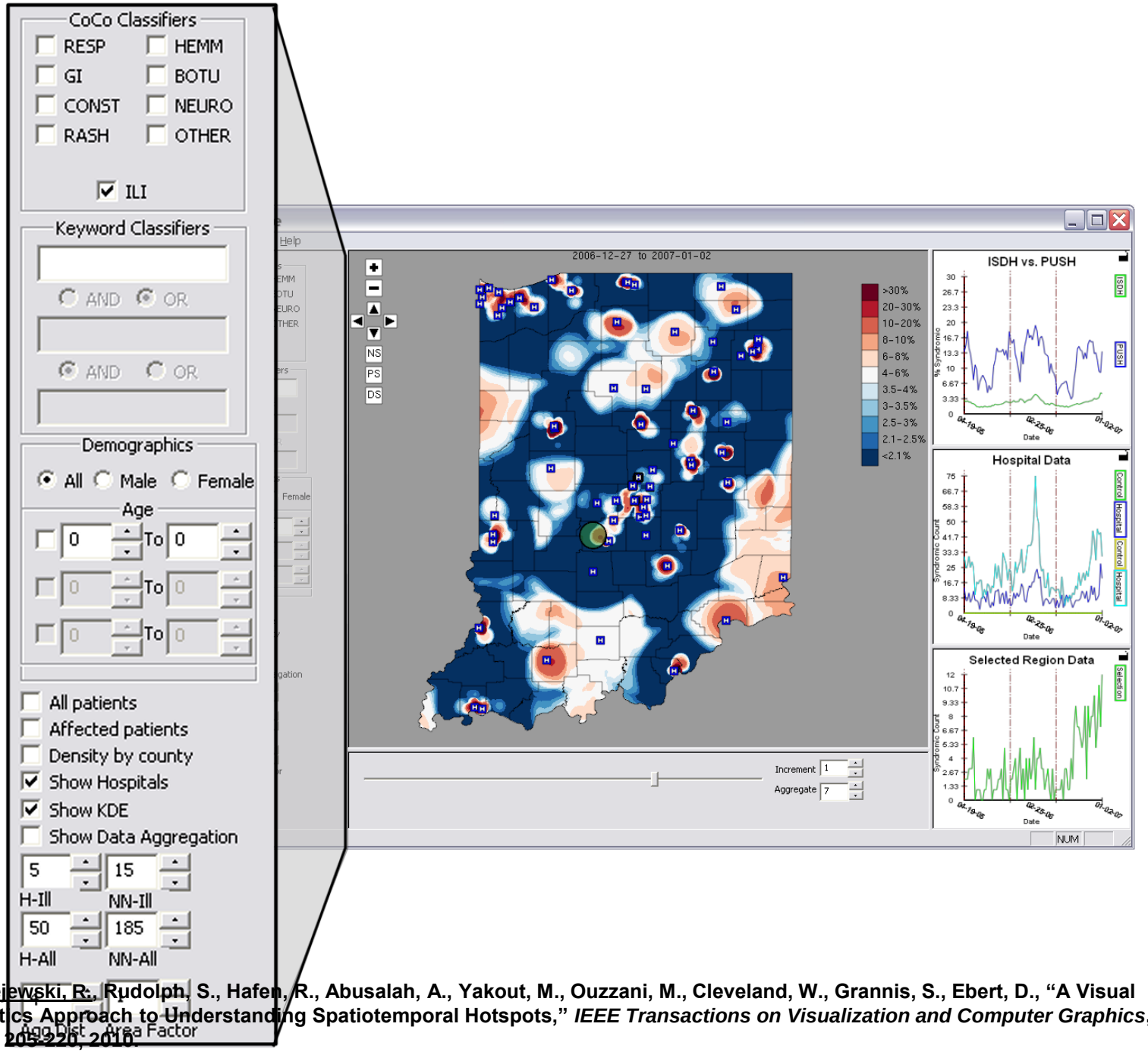
CoCo Classifier





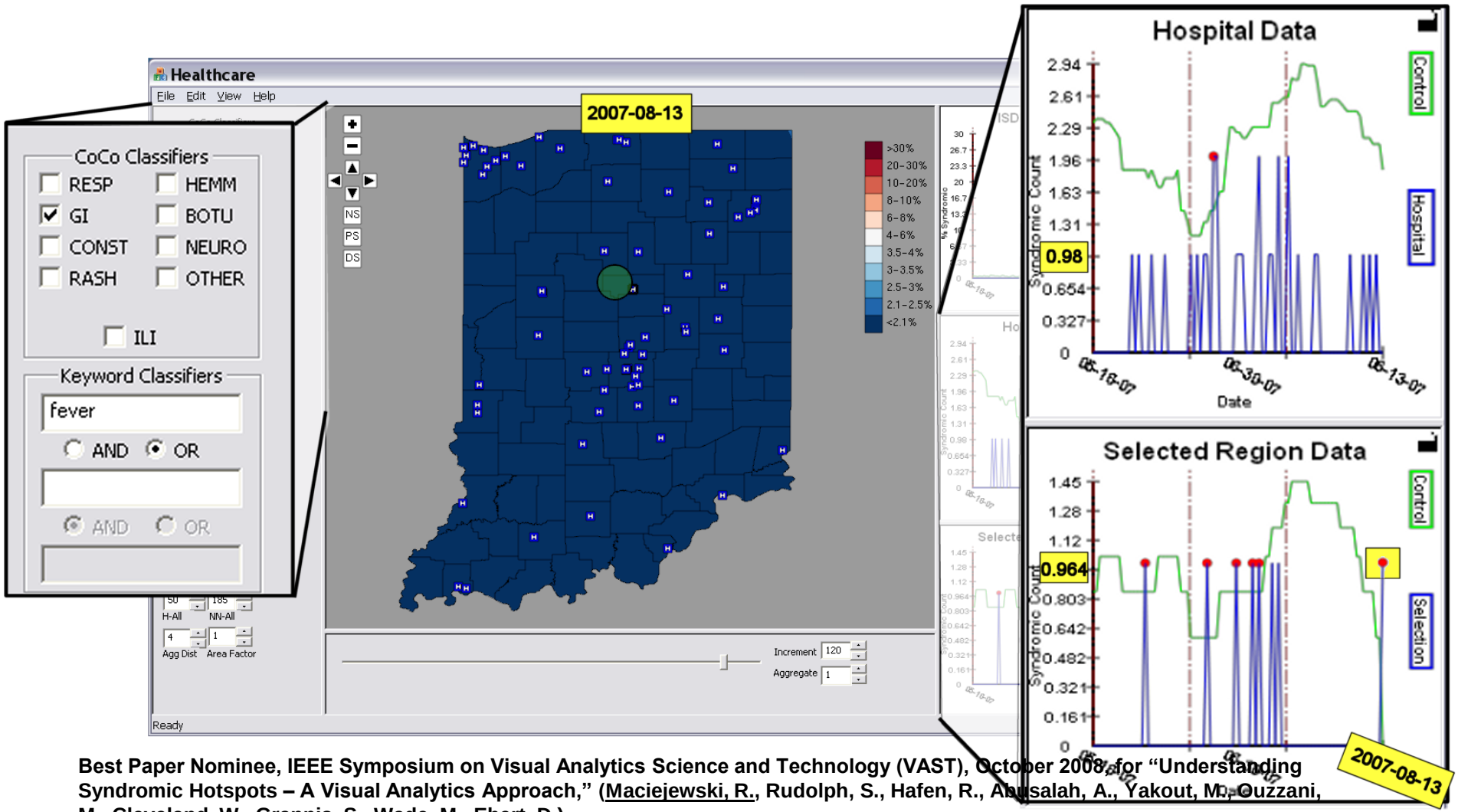
1) **Respiratory**
2) **Gastro-intestinal**
3) **Neurological**
4) **Botulinic**
5) **Constitutional**
6) **Rash**
7) **Hemmoraghic**
8) **Other**

VAST 2010 Challenge Award: Support for Future Detection, IEEE Conference on Visual Analytics Science and Technology, October 2010 for "VACCINATED," (Malik, A., Afzal, S., Hodgess, E., Ebert, D., Maciejewski, R.).



Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W., Grannis, S., Ebert, D., "A Visual Analytics Approach to Understanding Spatiotemporal Hotspots," *IEEE Transactions on Visualization and Computer Graphics*, 16(2): 205-220, 2010.

Hypothesis Generation and Exploration



Best Paper Nominee, IEEE Symposium on Visual Analytics Science and Technology (VAST), October 2008 for "Understanding Syndromic Hotspots – A Visual Analytics Approach," (Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W., Grannis, S., Wade, M., Ebert, D.).

Seems simple enough?

Data Models

- Data models are structured forms suitable for computer-based transformations
- These structures exist in the original data or are derivable from the original data
- Structures retain the information and knowledge content and the related context within the original data
- These structures are transformable into lower-dimensional representations for visualization and analysis

J. J. Thomas and K. A. Cook (Ed.). *Illuminating the Path: The R&D Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.

Data Models vs. Conceptual Models

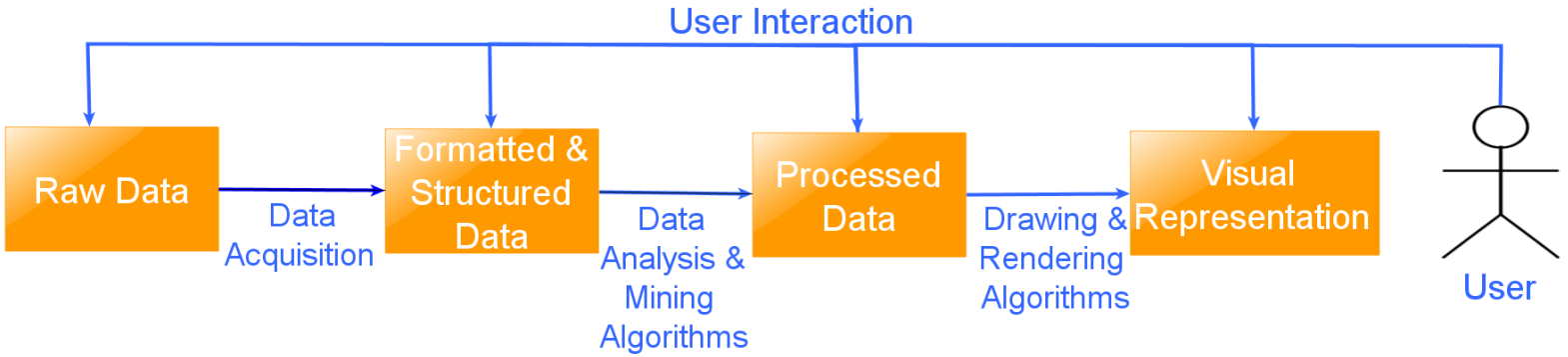
- Data models are mathematical abstractions
  - We can perform numerical operations, addition, subtraction, etc.
- Conceptual models are our mental constructs
  - These contain semantic structure and support reasoning
  - Think of giving directions to someone using landmarks

Data Types

- **Nominal**
  - Data whose categories have no implied ordering
  - Examples include political affiliations of a population
- **Ordinal**
  - Data that has a specified order, but no specified distance metric
  - Examples include beverage sizes at McDonalds (Small, medium, large)
- **Interval**
  - Data that has measurable distances
  - Examples include periods of time (second, minute, etc.) – the zero point is arbitrary
- **Ratio**
  - Same as interval, but include a zero point
  - Example include Celsius scale, height above sea level

1- SS Stevens, "On the Theory of Scales of Measurement," *Science*, 103(2684):677-680, 1946.

The Visual Analytics Pipeline



- We want to take these different data types and map them to an appropriate visual representation
- **Data Analysis** – data are prepared for visualization (smooth, interpolate, transform)
- **Filtering** – A subset of the data (usually user defined) is selected for visualization
- **Mapping** – Data are mapped to geometric primitives and their attributes
- **Rendering** – Geometric data are transformed to image data

Image taken from http://www.pafkiet.edu.pk/Default.aspx?tabid=693

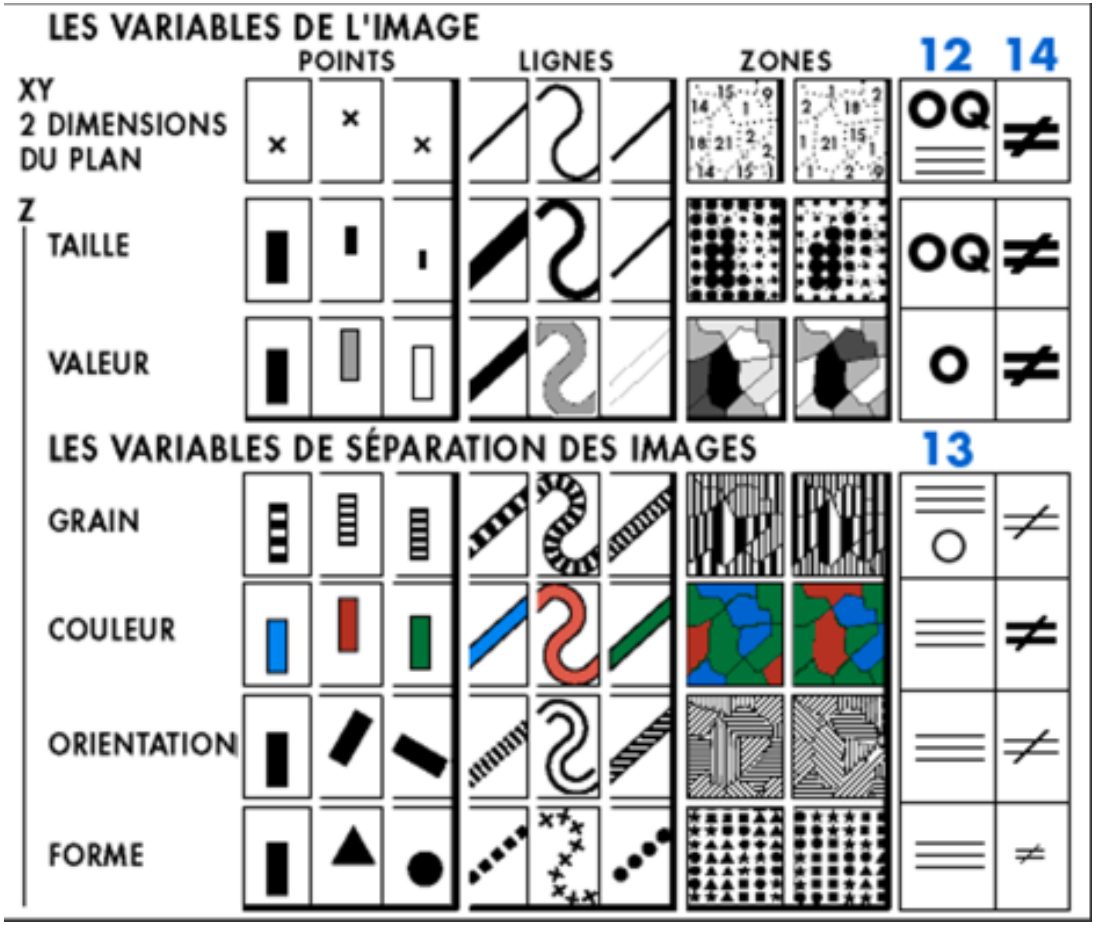
Haber, R. & McNabb, D. A. (1990). Visualization idioms: A conceptual model for scientific visualization systems. In Visualization in Scientific Computing.

Visual Representations

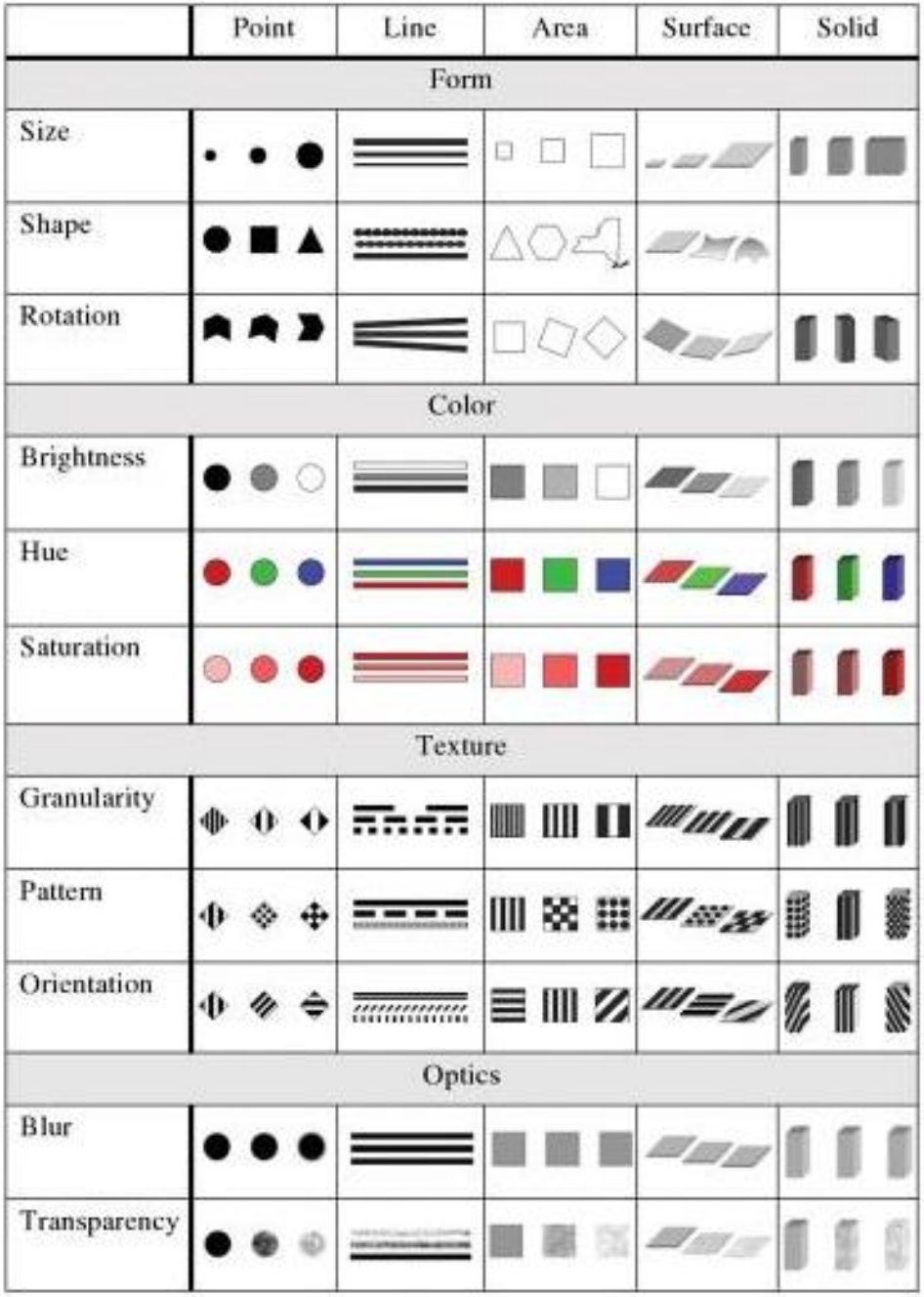
- Visual representations translate data into a visible form, highlighting important features, such as commonalities and anomalies
- Visual representations make it easy for users to perceive salient aspects of their data quickly
- These visual representations augment the cognitive reasoning process with perceptual reasoning which enhances the analytical reasoning process

Bertin's Visual Variables

- Visualization is concerned primarily with a mapping to visual form
- [x,y]
  - Position
- [z]
  - Size (Taille)
  - Value (Valeur)
  - Color (Couleur)
  - Texture (Grain)
  - Orientation
  - Shape (Forme)



J Bertin (1967), *The Semiology of Graphics*



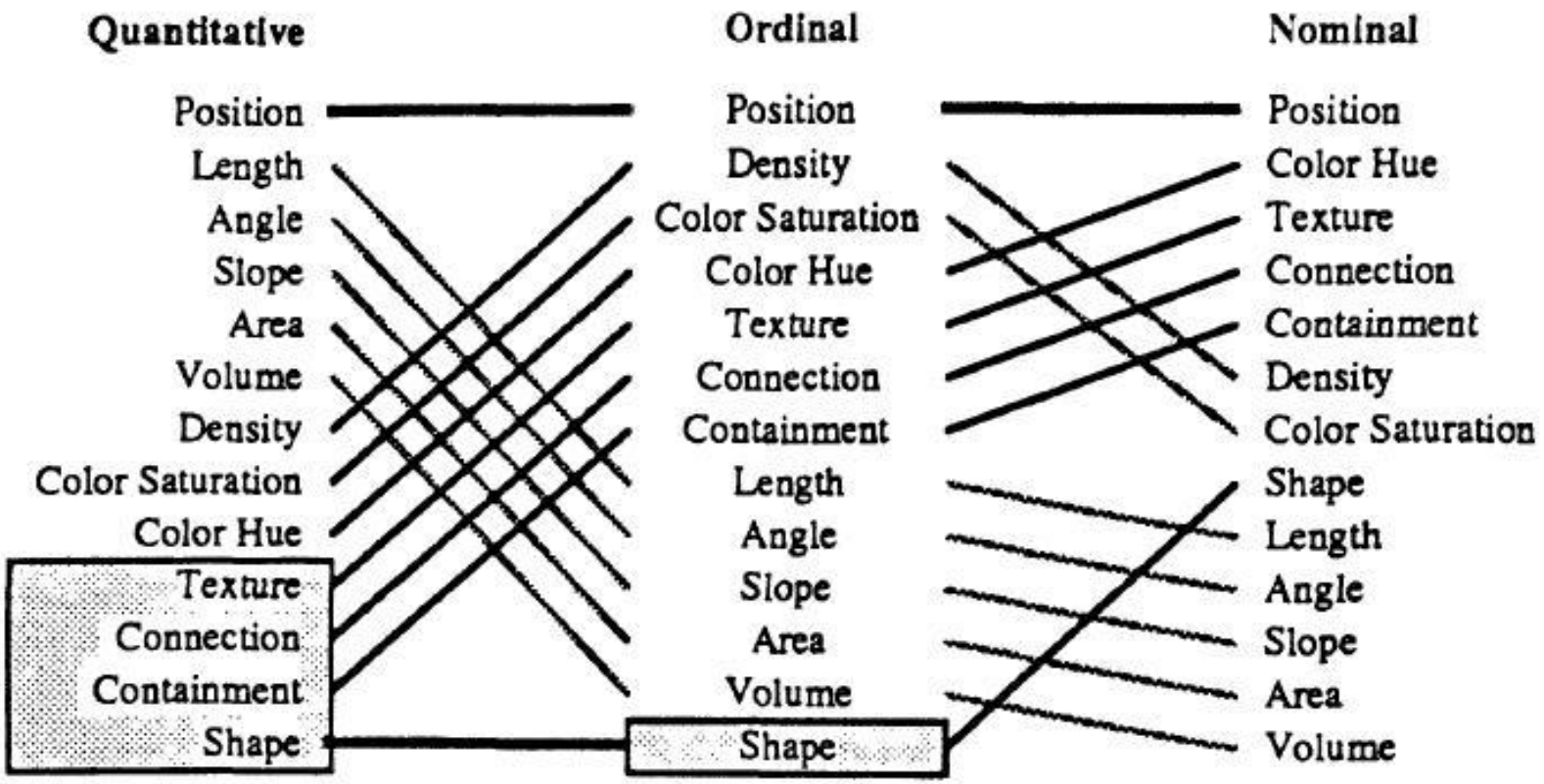
L Wilkinson (2005) *The Grammar of Graphics*, p. 317

How Many Variables to Use?

- The number of visual variables necessary for the representation is at least equal to the number of components in the information
- With three components, the information can be perceived as a single image
- Otherwise, we need several images (this is often termed *small multiples* and will be discussed in a future lecture)
- The number of components is the best basis for a classification of graphic constructions

J Bertin (1967), *The Semiology of Graphics*

How Do I Choose the Encoding?



J Mackinlay. Automating the Design of Graphical Presentations of Relational Information, ACM TOG 5:2, 1986.

Expressing Data With Color

- Results from research on visual attention can be used to assign visual features to data values
- One of the key components of visually representing data is choosing the appropriate color scale
- There is no "best" color scale
- Choice depends on[1,2]
  - data type
  - problem domain
  - visual representation
  - Questions the analyst is asking of the data
- While there is no "best" choice, there are design principles

1 - H. Levkowitz and G. T. Herman, "Color scales for image data," IEEE Computer Graphics and Applications, vol. 12, pp. 72-80, 1992
2 - P. Rheingans, "Task-based color scale design," Proceedings of Applied Image and Pattern Recognition, pp. 35-43, 1999.

Design Principles for Color Schemes

- *Order*[1] – Given a univariate data type, the color scale that is chosen to map the data must represent a perceived ordering
- *Separation*[2] – Important differences between ranges of the variable should be represented by colors that can be perceived as being different
  - Not only should they be perceived as different, but also equal
- *Aesthetics*[3] – color map should be aesthetically pleasing, contain a maximum perceptual resolution, and ordering should be intuitive

1 - B. E. Trumbo, “Theory for Coloring Bivariate Statistical Maps," *The American Statistician*, vol. 35, no. 4, pp. 220-226, 1981.
2 - H. Levkowitz and G. T. Herman, "Color scales for image data," IEEE Computer Graphics and Applications, vol. 12, pp. 72-80, 1992
3 - K. Moreland, “Diverging Color Maps for Scientific Visualization," *Proceedings of the 5th International Symposium on Visual Computing*, December 2009.

Univariate Color Schemes

- **Qualitative scheme**
  - Rainbow color scale is one of the most commonly used, but it is a poor color map in a large variety of domain problems
  - Ordering of the hues is unintuitive
  - Nominal data types can use this scale as no ordering is implied



1 - D. Borland and R. M. Taylor, “Rainbow Color Map (Still) Considered Harmful,“ *Computer Graphics & Applications*, vol. 27, no. 2, pp. 14-17, 2007.
2 -P. Rheingans, “Task-based color scale design," *Proceedings of Applied Image and Pattern Recognition*, pp. 35-43, 1999.
3 - M. A. Harrower and C. A. Brewer, “ColorBrewer.org: An online tool for selecting color schemes for maps," *The Catrtographic Journal*, vol. 40, no. 1, pp. 27-37, 2003.

Univariate Color Schemes

- **Sequential color scheme**
  - Simplest is the gray scale map where variable is mapped to brightness
  - Sequential maps represent ordered data
  - Dark colors typically represent high ranges, bright, low
  - Benefits are that the scale is intuitive
  - Weakness is that limited number of distinguishable colors can be represented



1 - D. Borland and R. M. Taylor, “Rainbow Color Map (Still) Considered Harmful,“ *Computer Graphics & Applications*, vol. 27, no. 2, pp. 14-17, 2007.
2 -P. Rheingans, “Task-based color scale design," *Proceedings of Applied Image and Pattern Recognition*, pp. 35-43, 1999.
3 - M. A. Harrower and C. A. Brewer, “ColorBrewer.org: An online tool for selecting color schemes for maps," *The Catrtographic Journal*, vol. 40, no. 1, pp. 27-37, 2003.

Univariate Color Schemes

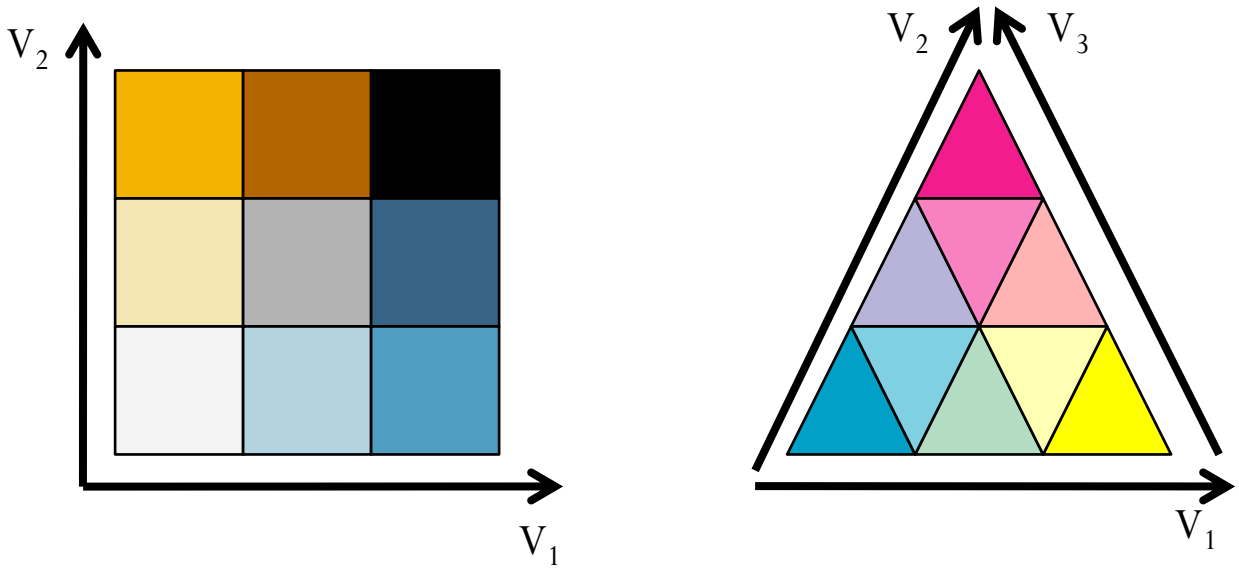
- **Divergent color scheme**



  - Provides means for variable comparisons
  - Best suited for ratio data where there is some meaningful zero point
  - Scale lacks a natural ordering of colors
  - Careful choices must be made in choosing high and low ends
  - Can use concept of cool (blues) and warm (reds and yellow) colors

1 - C. Hardin and L. Maffi, *Color Categories in Thought and Language*, Cambridge University Press, 1997.
2 -P. Rheingans, “Task-based color scale design," *Proceedings of Applied Image and Pattern Recognition*, pp. 35-43, 1999.
3 - M. A. Harrower and C. A. Brewer, “ColorBrewer.org: An online tool for selecting color schemes for maps," *The Catrtographic Journal*, vol. 40, no. 1, pp. 27-37, 2003.

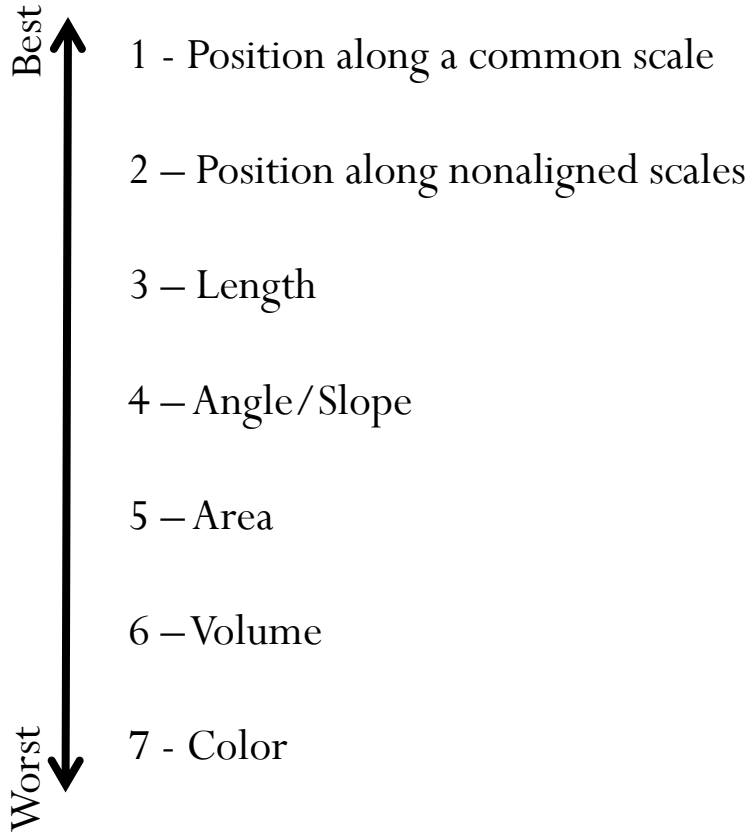
Multivariate Color Schemes



M. A. Harrower and C. A. Brewer, “ColorBrewer.org: An online tool for selecting color schemes for maps," *The Catrtographic Journal*, vol. 40, no. 1, pp. 27-37, 2003.

Cleveland's Heirarchy[1]

- Cleveland evaluated elements when isolated
- Tasks were restricted to magnitude and ratio comparisons
- Research indicates this hierarchy may be best in pre-attentive stages or when focusing only on portions of a graphic[2,3,4]



1 – WS Cleveland, *The Elements of Graphing Data*, 1985

2 – Spence, I., and Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology* 5, 61-77

3 – Simkin, D., and Hastie, R. (1987). An information processing analysis of graph perception. *Journal of the American Statistical Association*, 82, 454-465

4 – Carswell, C. M. (1992). Choosing specifiers: An evaluation of the basic tasks model of graphical perception. *Human Factors*, 4, 535-554

What if I Combine Encodings?

- So, what happens when we combine several scales in a single display?
- Can we represent one quantitative dimension with color and another with orientation and expect a perceiver to respond to both dimensions?
- Do these things make psychological sense?

L Wilkinson (2005) *The Grammar of Graphics*

Integral Versus Separable Dimensions

- "A configuration has properties that have to be expressed as some form of interaction or interrelation between the components, be they features or dimensions."[1]
- *Integral dimensions* are not as easily decomposable by perceivers as *separable dimensions*
- Separating hue from brightness in a color (integral dimensions) are harder to decompose than say size and texture (separable dimensions)

1 – Garner, W. R. (1981). The analysis of unanalyzed perceptions. In M. Kubovy and J. R. Pomerantz (Eds.), *Perceptual Organization* (pp. 119-139). Hillsdale, NJ: Lawrence Erlbaum.

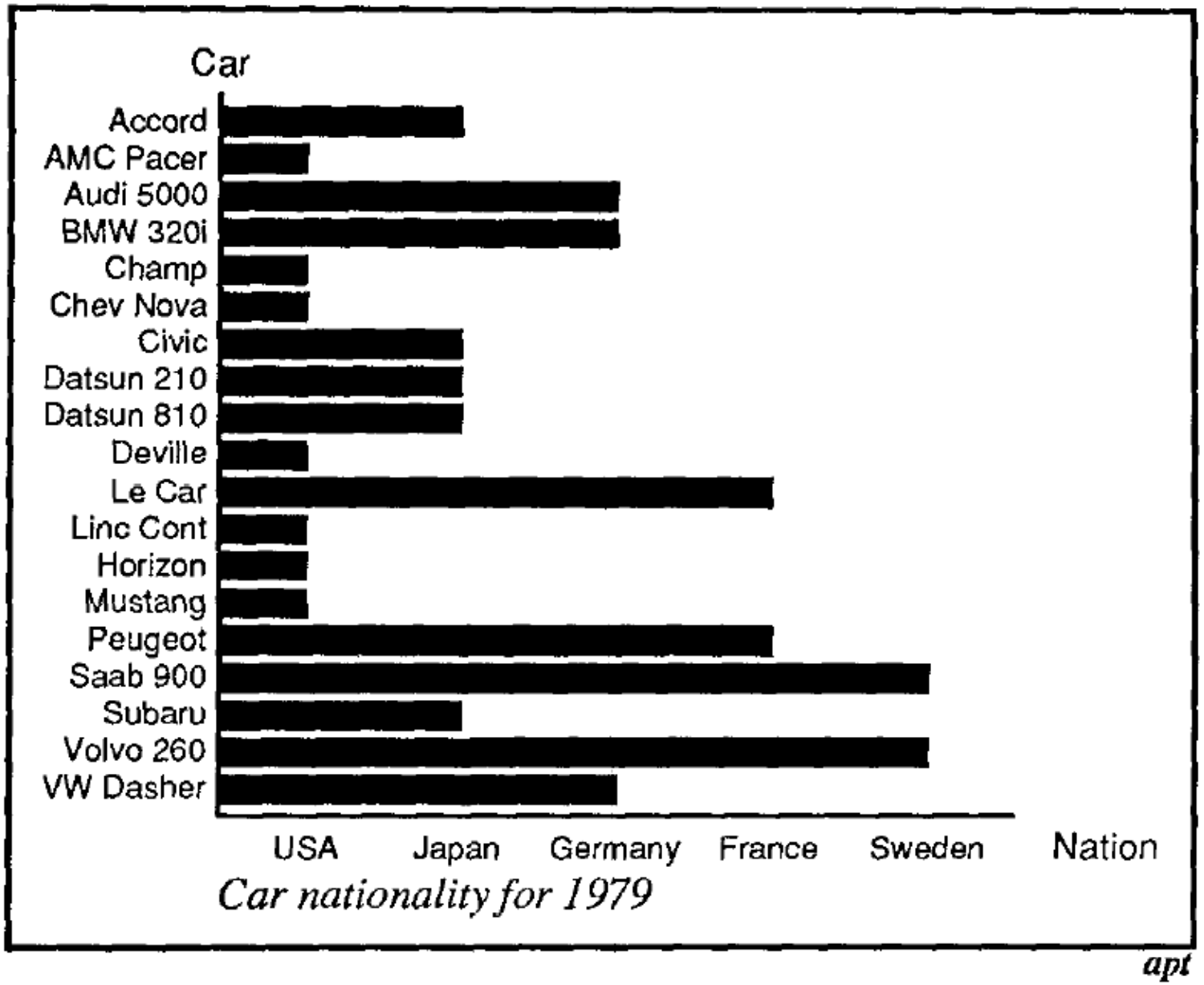
Combinatorics of Encodings

- Challenge:
  - Pick the best encodings from the exponential number of possibilities $(n+1)^8$

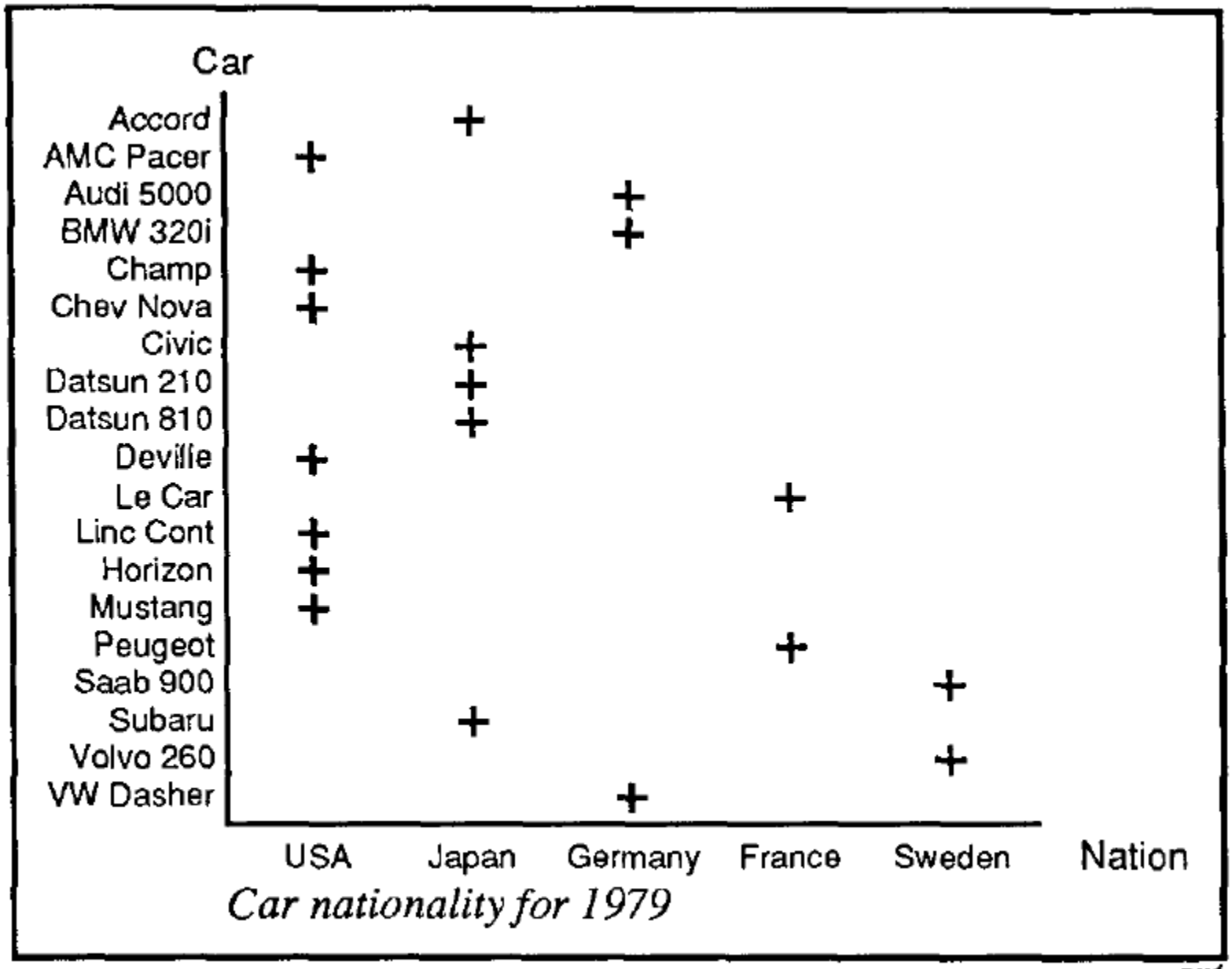
- Principle of Consistency: The properties of the image should match the properties of the data

- Principle of Importance Ordering: Encode the most important information in the most effective way

1 – This slide is borrowed from Pat Hanrahan's "From Data to Image" lecture:
http://graphics.stanford.edu/courses/cs448b-04-winter/lectures/encoding/



1 – J. Mackinlay, Automating the Design of Graphical Presentations of Relational Information, ACM Transactions on Graphics, 5(2): 110-141, 1986.



1 – J. Mackinlay, Automating the Design of Graphical Presentations of Relational Information, ACM Transactions on Graphics, 5(2): 110-141, 1986.

The Curse of Dimensionality

- A term coined by Bellman in 1961
- Refers to the problems associated with multivariate data analysis as the dimensionality increases the available data becomes sparse
- Sparsity is a problem for any method that requires statistical significance
- Sometimes data dimensions are redundant and can be reduced with minimal information loss
- In visualization we are also limited with screen space and the number of available visual variables, so choosing the most appropriate dimensions is key

Richard Ernest Bellman; Rand Corporation (1957). *Dynamic programming*. Princeton University Press.

The Curse of Dimensionality

- So, what can we do?
  - We can incorporate prior knowledge of the data
  - We can smooth the target function
  - We can reduce the dimensionality

The Practicality of the Curse of Dimensionality

- For a given sample size, there is a maximum number of features above which the performance of classifying samples will degrade rather than improve
- In most cases, the additional information that is lost when discarding some features is compensated by a more accurate mapping in the lower-dimensional space
- So, how do we know what features we can throw away?
- For visualization this implies that there are some features of a dataset that will be better to visualize (contain more information) than others!

Dimensional Reduction

- Two approaches are available to reduce dimensionality
  - **Feature extraction:** creating a subset of new features by combinations of the existing features
  - **Feature Selection:** choosing a subset of all the features

$$\begin{bmatrix} x_1 \\ x_2 \\ \\ \\ x_N \end{bmatrix} \rightarrow \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \\ \\ x_{i_M} \end{bmatrix} \qquad \begin{bmatrix} x_1 \\ x_2 \\ \\ \\ x_N \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ y_2 \\ \\ \\ y_M \end{bmatrix} = f\left(\begin{bmatrix} x_1 \\ x_2 \\ \\ \\ x_N \end{bmatrix}\right)$$

  - Given a feature space $x_i \in R^N$ find a mapping $y = f(x): R^N \rightarrow R^M$ with M<N such that the transformed feature vector $y \in R^M$ preserves (most of) the information or structure in R^N
  - An optimal mapping is one that does not increase error

Principle Components Analysis

- One of the most commonly applied dimension reduction techniques
- PCA is a deterministic analytical procedure that utilizes an orthogonal transformation to reduce a set of sample observations with potentially correlated variables into a set of uncorrelated variables called principal components
- The number of principal components will always be less than or equal to the original number of variables in the sample set

. T. Jolliffe, \Principal Component Analysis," *Springer Series in Statistics*, Springer, NY, 2002.

Principle Component Analysis

- The main limitation of PCA is that it does not consider class separability since it does not take into account the class label of the feature vector
  - PCA simple performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance
  - There is no guarantee that the directions of maximum variance will contain the most interesting/important features
  - This is very similar to Multi-Dimensional Scaling (MDS)

Data Mining

- A lot of the previous techniques for dimensional reduction are typically used in communities such as data mining and machine learning
- Can be used to enhance the visualizations, show us what is similar in a multidimensional space
- However, if we reduce the space, we need to think about how we communicate these changes to the end user

Visual Analytics

- We want to know how to best create images of data effectively in order to facilitate advanced data analysis
  - Spotting outliers
  - Discriminating clusters
  - Checking distributional and other assumptions
  - Examining relationships
  - Comparing mean differences
  - Observing time-based processing

Representation

- What are the two main ways of presenting multivariate data sets?
  - Directly (textually) – Tables
  - Symbolically (pictures) – Graphs

- How do we decide which to use, and when?

1 – Descriptions on this slide are borrowed from John Stasko's "Multivariate Data & Tables and Graphs" lecture: http://www.cc.gatech.edu/~stasko/7450/Notes/data.pdf

Tables?

- Use tables when
  - The document will be used to look individual value
  - The document will be used to compare individual values
  - Precise values are required
  - The quantitative info to be communicated involves more than one unit of measure
- Use graphs when
  - The message is contained in the shape of the values
  - The document will be used to reveal relationships among values

1 – Descriptions on this slide are borrowed from John Stasko's "Multivariate Data & Tables and Graphs" lecture: http://www.cc.gatech.edu/~stasko/7450/Notes/data.pdf

2 - S. Few, Show Me the Numbers: Designing Tables and Graphs to Enlighten,

Graphs?

- Graph
  - Visual display that illustrates one or more relationships among entities
  - Shorthand way to present information
  - Allows a trend, pattern or comparison to be easily comprehended
- Critical to remain task-centric
  - Why do you need a graph?
  - What questions are being answered?
  - What data is needed to answer those questions?
  - Who is the audience?

1 – Descriptions on this slide are borrowed from John Stasko's "Multivariate Data & Tables and Graphs" lecture: http://www.cc.gatech.edu/~stasko/7450/Notes/data.pdf

Histograms

- Typical first look visualization method
  - Shows the shape of the data distribution
- The choice of the histogram bin width greatly impacts the resultant visualization
- There is no "best" number of bins, instead, different bin sizes can reveal different features of the data
- Some methods work to determine an optimal number of bins, but these methods make assumptions on the underlying data distribution

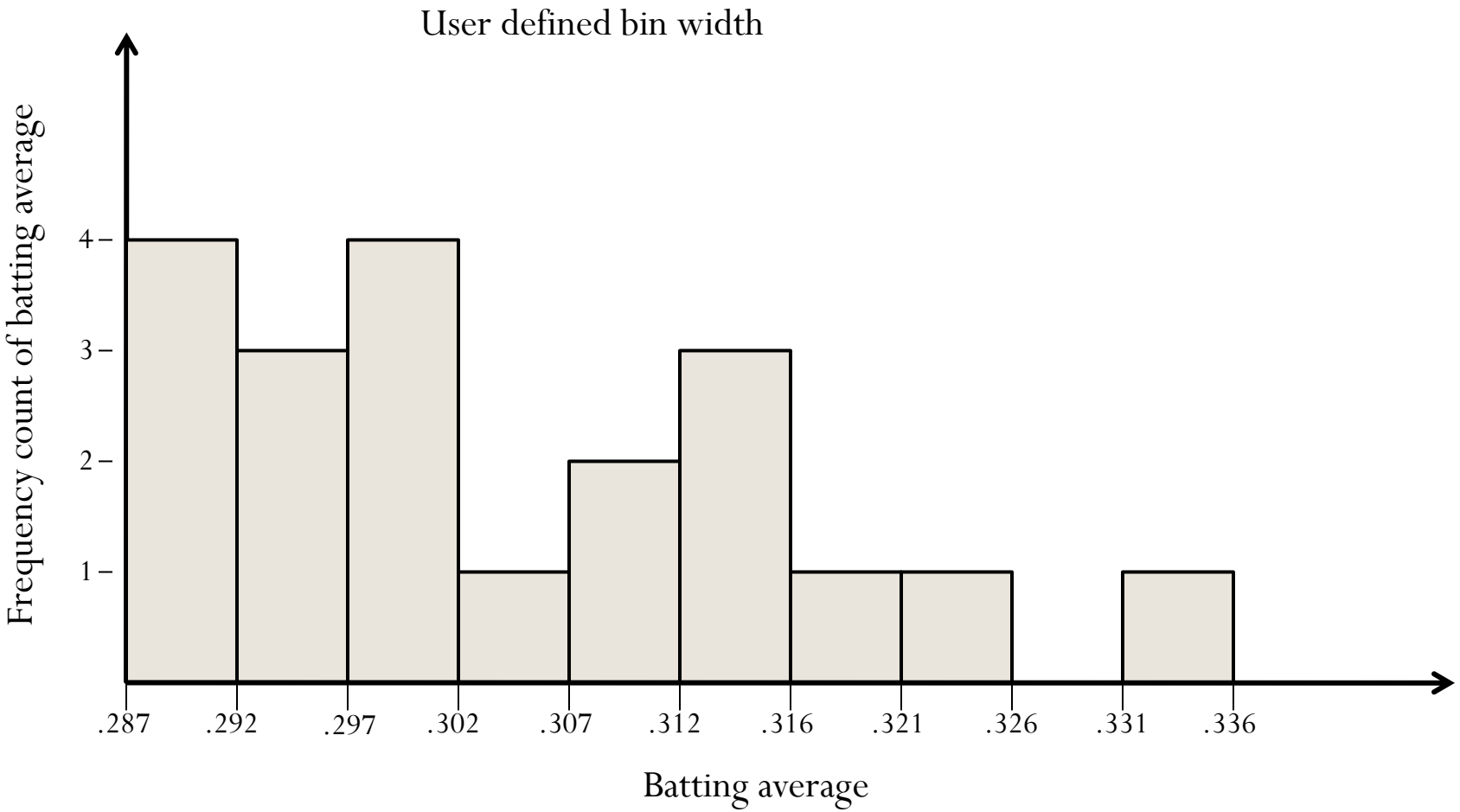
Histogram Binning

- Number of bins (k) can be user specified or chosen from a suggested bin width (h) such that
- $k = \left\lceil \frac{\max x - \min x}{h} \right\rceil$
- Common choices for k include the square-root choice where $k = \sqrt{N}$
- Sturge's formula[1] where $k = \lceil \log N + 1 \rceil$
- Scott's choice[2] where $h = \frac{3.5\sigma}{N^{\frac{1}{3}}}$
- Freedman-Diaconis rule[3] $h = 2IQR(x)N^{-\frac{1}{3}}$

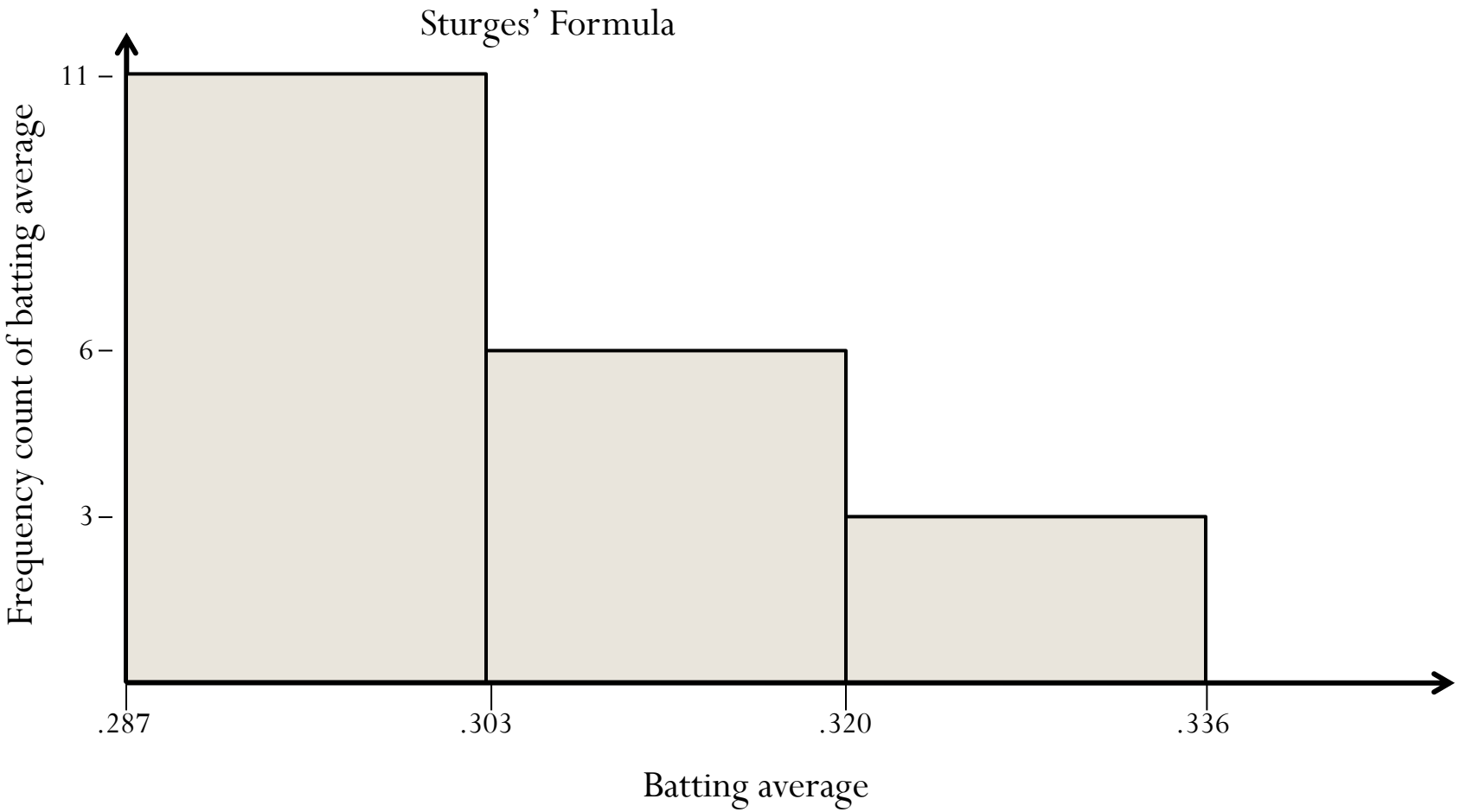
1 – H. A. Sturges. The choice of a class interval. *Journal of the American Statistical Association, p. 65-66, 1926*

2 - D. W. Scott. On optimal and data-based histograms. *Biometrika, 66(3):605-610, 1979.*

3 - Freedman, David; Diaconis, Persi, "On the histogram as a density estimator: L_2 theory". *Probability Theory and Related Fields* (Heidelberg: Springer Berlin) **57** (4): 453–476, 1981.


User defined bin width
Frequency count of batting average
4
3
2
1
.287
.292
.297
.302
.307
.312
.316
.321
.326
.331
.336
Batting average

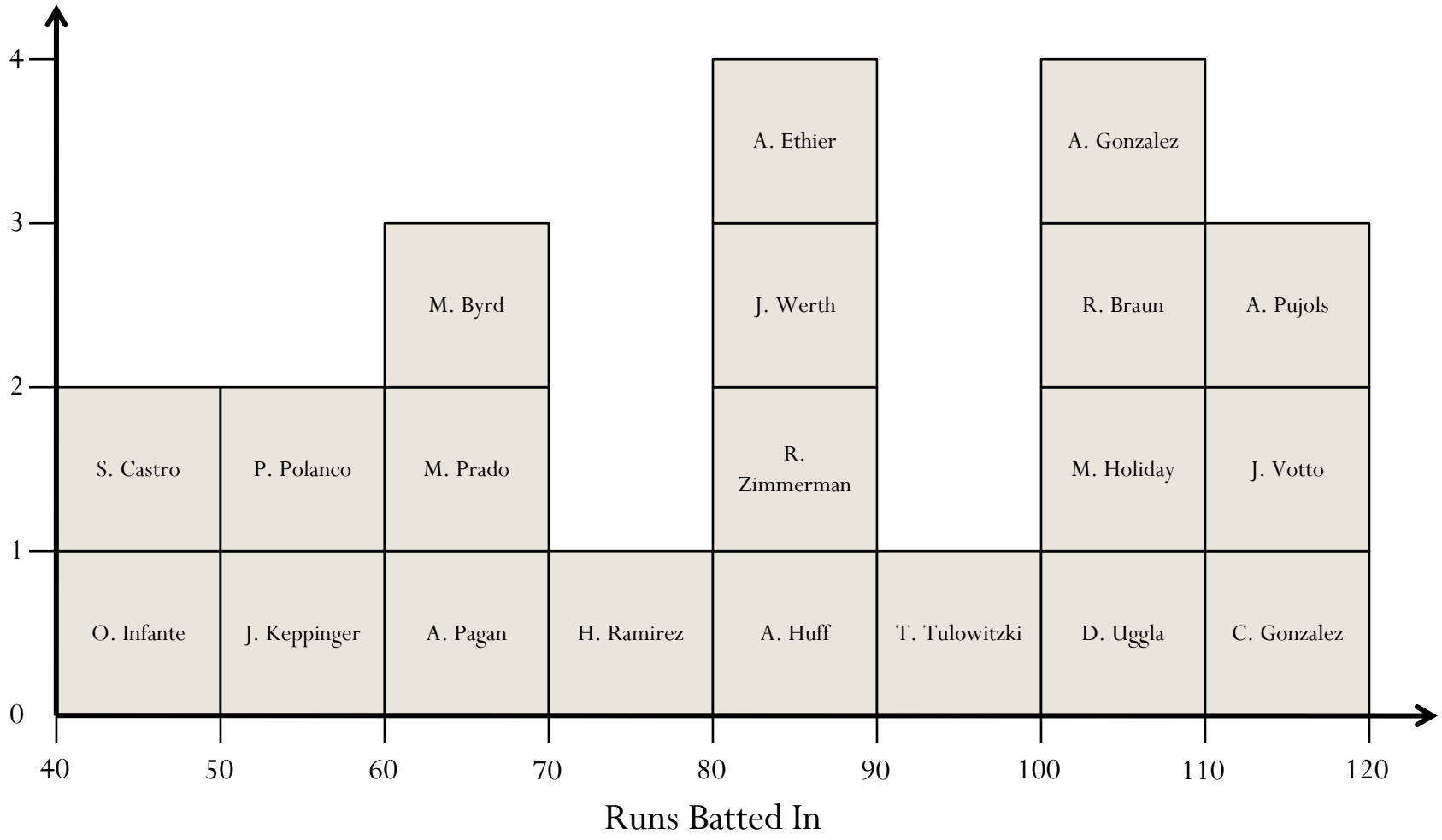

Square root choice
Frequency count of batting average
7
6
5
4
3
2
1
.287
.297
.307
.316
.326
.336
Batting average


Sturges' Formula
Frequency count of batting average
11
6
3
.287
.303
.320
.336
Batting average

Histograms

- Drawbacks
  - Density estimate depends on starting position of the bins
  - Discontinuities of the estimate are not due to the underlying density, only artifacts of the chosen bin location
  - As the number of dimensions grow, need many samples or else most bins would be empty
- Histograms are best suited for quick visualizations in 1 or 2 dimensions

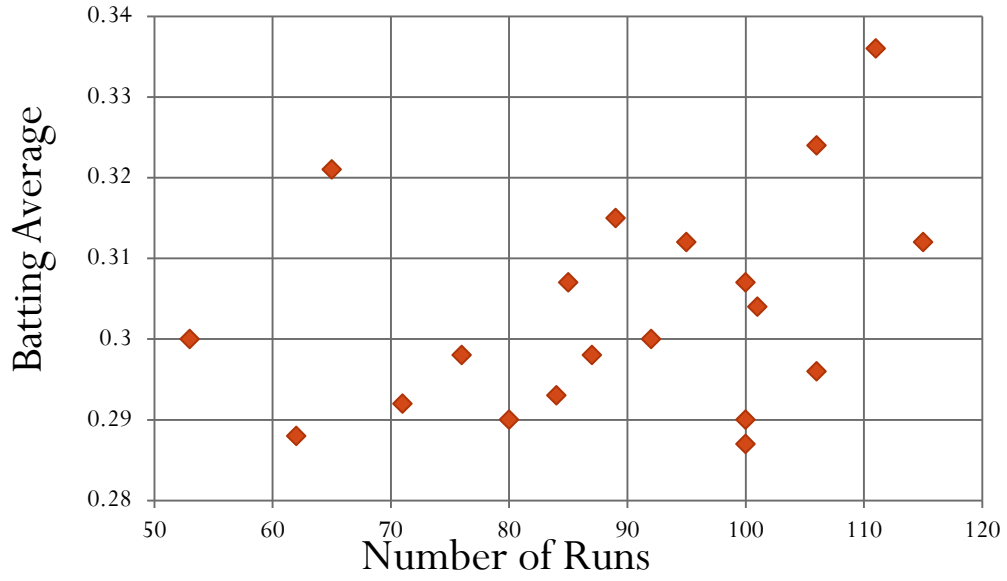
Bivariate Case – Stacked Bar Graph



T. N. Dang, L. Wilkinson and A. Anand, "Stacking Graphic Elements to Avoid Over-Plotting," *IEEE Transactions on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1044-1052, 2010.

Bivariate Case - Scatterplot

- Visualizes discrete data values along two axes
- Used as a means of analyzing bivariate relationships
- Quick means of assessing outliers, clusters and distributions
- Putting a line through the data can help assess trends, but can also mislead viewer



Non-Data Components of Graphs

- Axes and legends can often be as important as the data themselves
- Poor axis choices and label choices can lead to confusing visualizations
- Axis tick labels provide cognitive context for most basic plot types
- They support estimation and contribute to the overall appearance of the graphic
- Cleveland suggests choosing the scales so that the data rectangle fills up as much of the scale-line rectangle as possible

W. S. Cleveland. The Elements of Graphing Data. Wadsworth Publ. Co., Belmont, CA, USA, 1985.

J. Talbot, S. Lin, P. Hanrahan, "**An Extension of Wilkinson's Algorithm for Positioning Tick Labels on Axes**," IEEE Transactions on Visualization and Computer Graphics 16(6): 1036-1043, 2010

Graph Aspect Ratios

- Our ability to perceive trends and patterns in a given display is heavily influenced by the aspect ratio
- Aspect ratio affects densities, relative distances and orientations
- Several methods have been proposed for automatically selecting the aspect ratio
- Aspect ratio: a = width/height

J. Talbot, J. Gerth, P. Hanrahan, “Arc Length-based Aspect Ratio Selection,” IEEE Transactions on Visualization and Computer Graphics 17(12): 2276-2282, 2011

Data Distributions

- The type of data distribution affects not only the way the data should be analyzed, but also the way it should be visualized!

Skewness

- Measure of the asymmetry of the probability distribution
- Can be positive, negative or undefined
- Negative skew indicates that the tail on the left side is longer than on the right
- Positive skew indicates that the tail on the right side is longer than on the left
- A zero value indicates the tails are relatively evenly distributed on both sides of the mean, but does not necessarily imply symmetry

Visualizing Skewed Data

- Plotting skewed data compresses data values into smaller regions, resulting in lower visual fidelity[1]
- How to overcome skewness?
  - Remove data outliers
  - Interactive techniques (zoom, pan, brush)
- This can help for the visual analysis, but not always!
  - What if I want to compare data sets with different skews?
  - What if I want to perform statistical analyses?
- Pre-conditioning data through transformations is a key step for appropriate analysis and visualization.

1 - W. S. Cleveland. *Visualizing Data*. Hobart Press, 1993

The Power Transformation

- We can transform data to a Normal approximation by a power transformation[1]
- The Normal (Gaussian) distribution has many features which make it popular

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  - It can be fully characterized with two parameters
  - The probability of any value can be obtained by knowing how many standard deviations it is away from the mean
  - Many statistical measures and tests are well defined for this distribution
- Must be careful not to characterize non-Normal data as Normal

- Such transformations reduce the effects of:
  - Skewness (transformations changes the range of data values and helps fit data onto the display)
  - Random noise (transformations can help show global trends)
  - Monotone spread

1 – J. W. Tukey. On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, 28: 602-632, 1955.

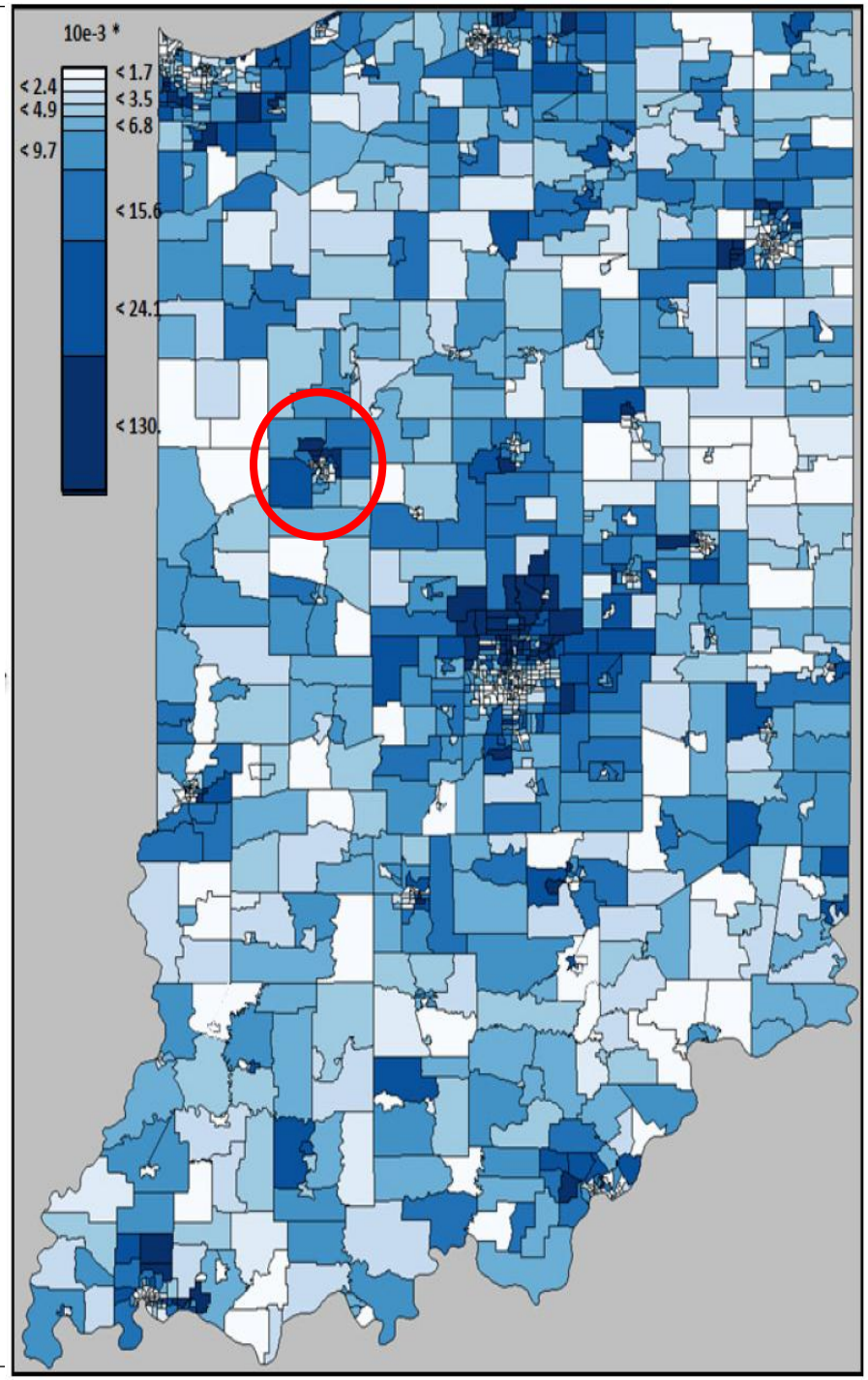
The Power Transformation

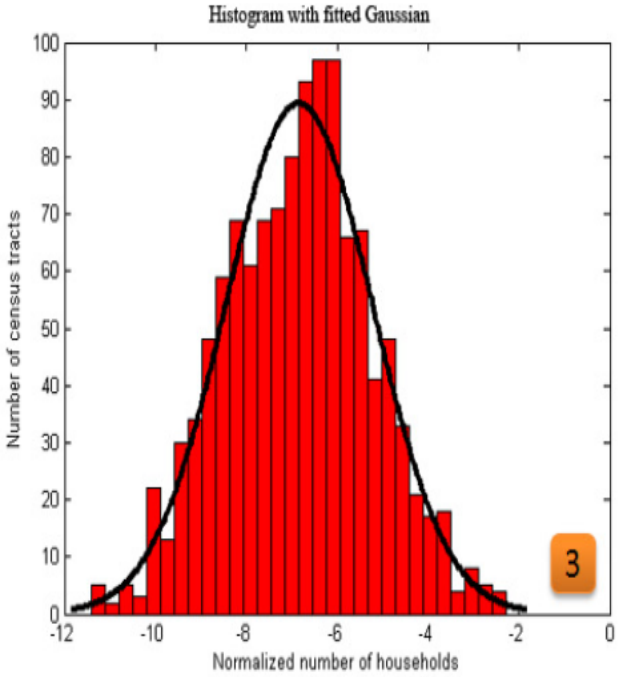
- The Box-Cox[1] Power Transformation
  - x is the observed (recorded) data and λ is the power

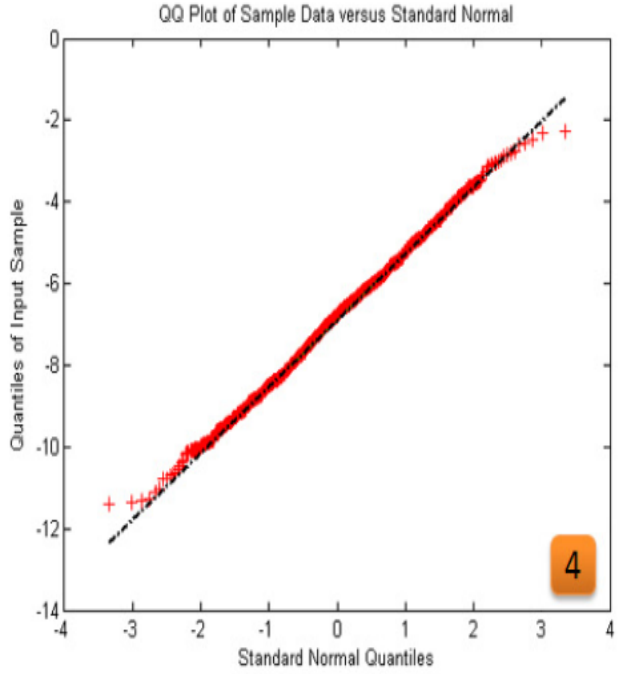
$$x(\lambda) = \frac{x^{\lambda} - 1}{\lambda}, (\lambda \neq 0)$$
$$= ln(x), (\lambda = 0)$$

- The power (λ) is chosen by maximizing the logarithm of the likelihood function

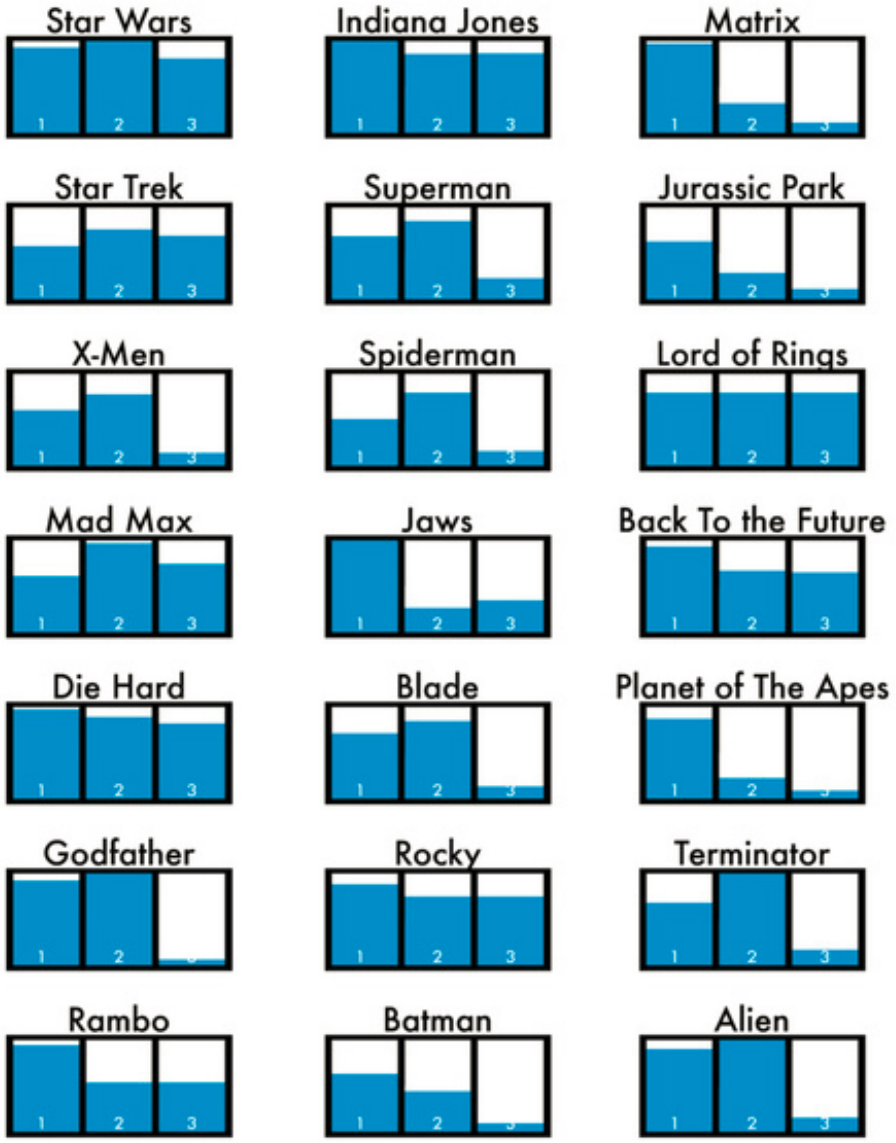
1 – G. Box and D. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological), 26(2):211-252, 1964.*


10e-3 *
< 2.4
< 4.9
< 9.7
< 1.7
< 3.5
< 6.8
< 15.6
< 24.1
< 130.


Histogram with fitted Gaussian
Number of census tracts
Normalized number of households
3


QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
4

THE TRILOGY METER

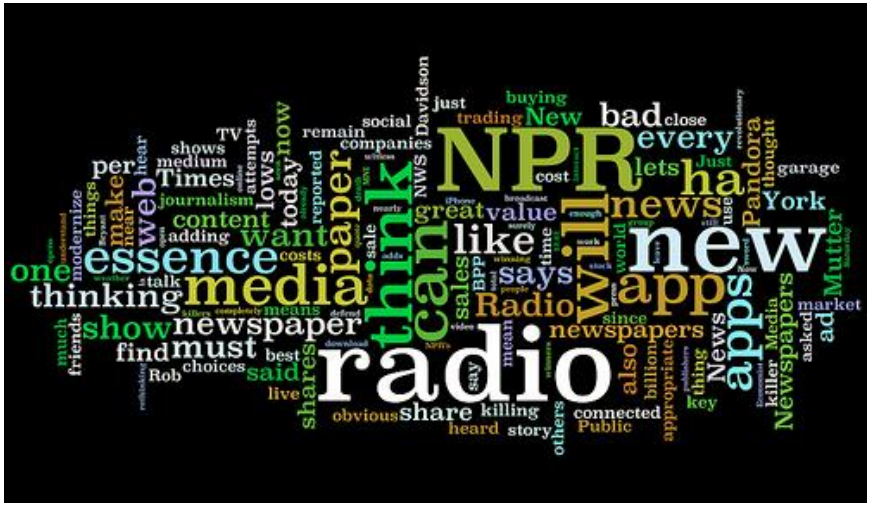


http://www.juiceanalytics.com/writing/better-know-visualization-small-multiples/

Tag Clouds and Wordles

- Visual representation for text data where words are placed and scaled based on some statistical measures
- Font size is typically determined by the number of instances a word is used

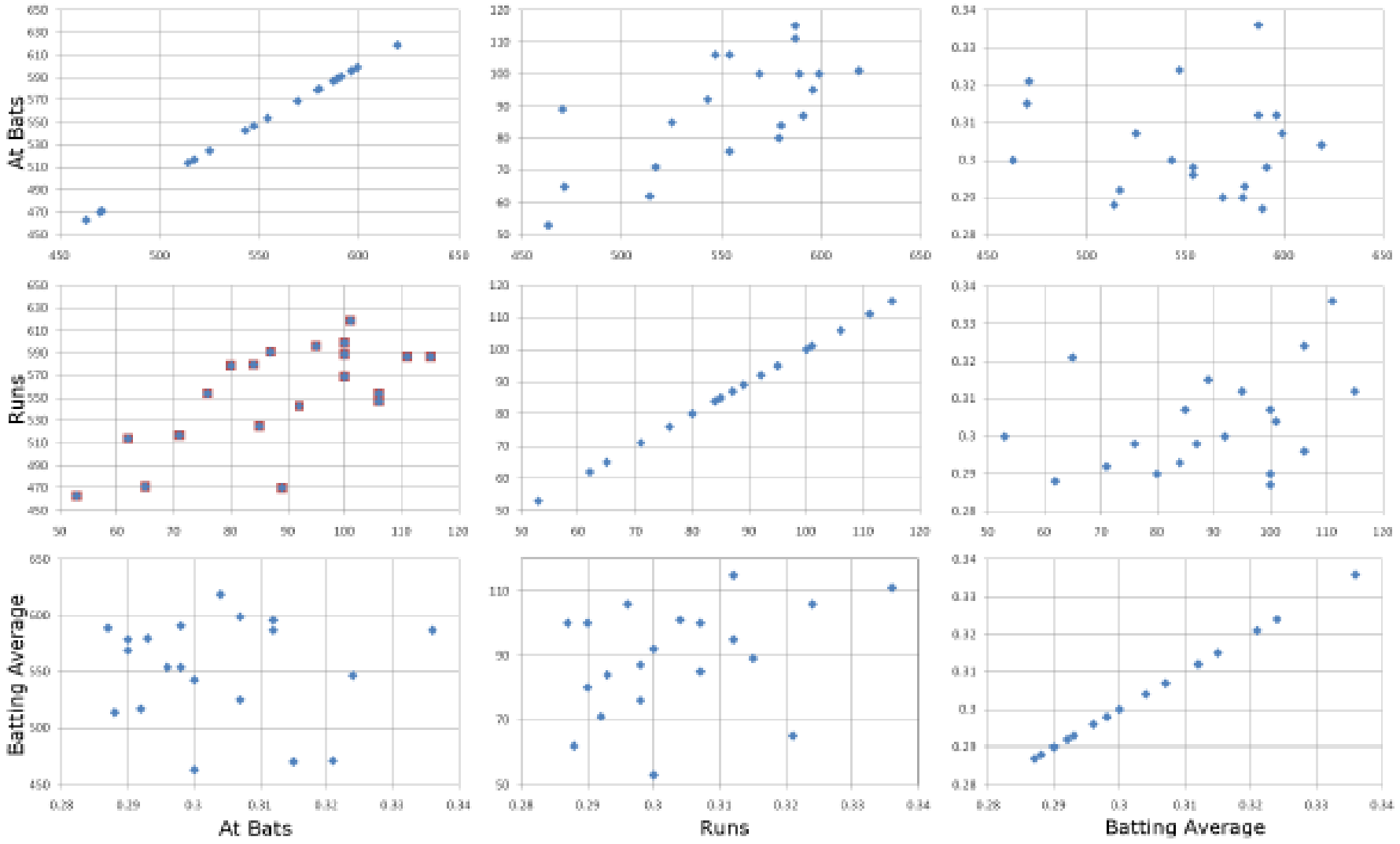




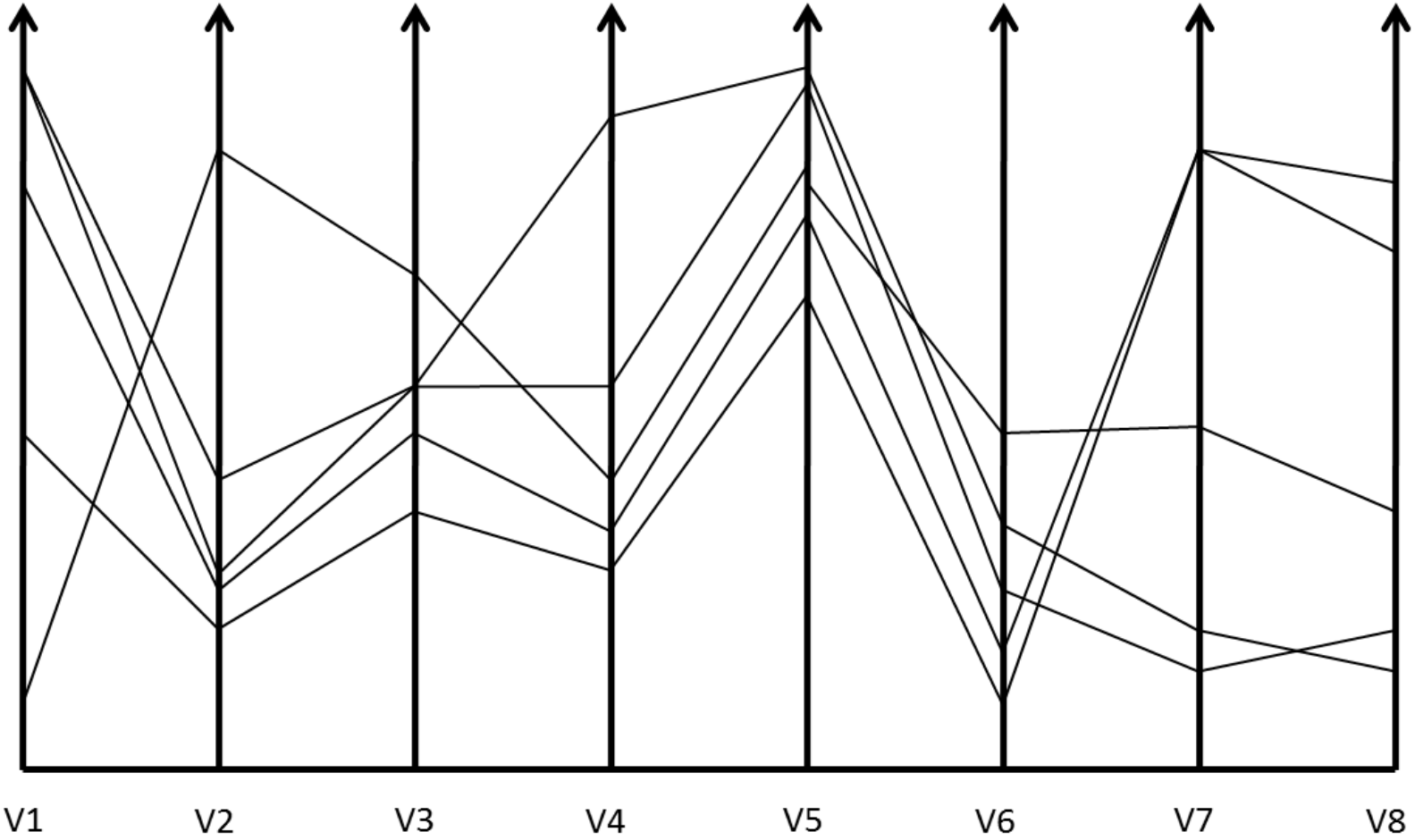
Text cloud comparing 2002 State of the Union Address by U.S. President Bush and 2011 State of the Union Address by President Obama. Created at TagCrowd.

Viegas, FB, Wattenberg, M, Feinberg, J, "Participatory Visualization with Wordle, IEEE Transactions on Visualization and Computer Graphics 15(6)1137-1144

Multivariate Case - Scatterplot Matrix



Parallel Coordinate Plots



1 - Inselberg A (1985) The Plane with Parallel Coordinates. The Visual Computer1(4):69-91
2 - Ankerst M, Berchtold B, Keim DA (1998) Similarity clustering of dimensions for an enhanced visualization of multidimensional data. IEEE Symposium on Information Visualization pp 52-62

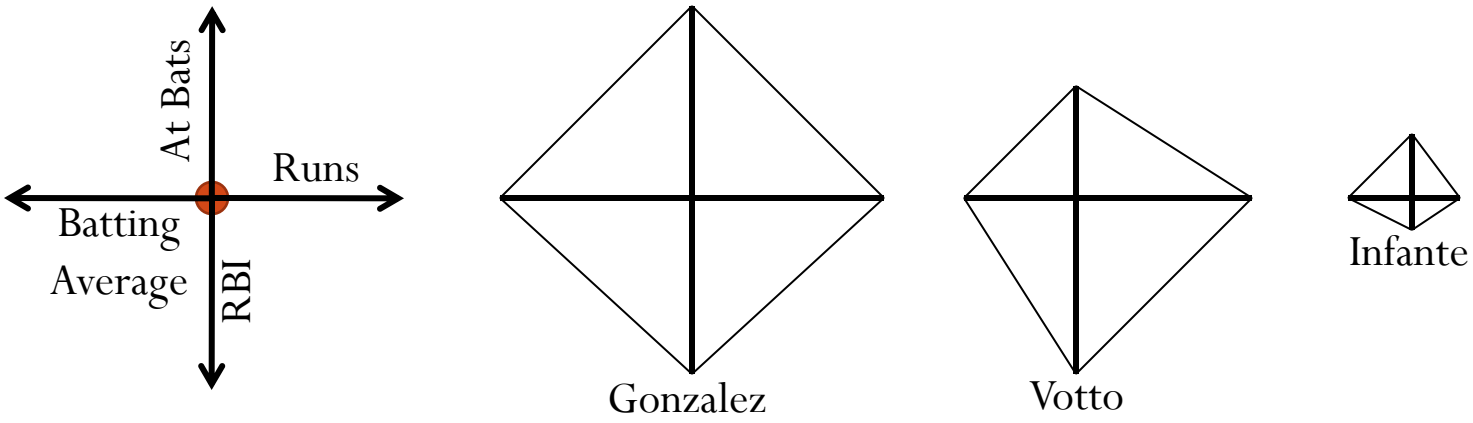
Issues With Parallel Coordinate Plots

- Different variables can take different values with very different ranges
  - Need to normalize data ranges (maybe do a power transformation and then a normalization?)
- Order of the parallel coordinate plots has a major impact on the resultant visualization
- The more variables we plot, the more lines we get and the more clutter that we get

1 -Yang, J., Peng, W., Ward, M.O., Rundensteiner, E.A., Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In *Proc. of IEEE Symposium on Information Visualization* (2003), pp. 105–112.
2 – Zhou, H., Yuan, X., Qu, Huamin, Cui, W., Chen, B., "Visual Clustering in Parallel Coordinates," Computer Graphics Forum 27(3) 1047-1054, 2008.

Star Plot

- Lay out axes in a radial layout, length of a ray emanates from a central point
- Rays are then joined together by a polyline drawn around the outside



S. E. Fienberg,"Graphical Methods in Statistics," *The American Statistician*, vol. 33 pp. 156-178, 1979.

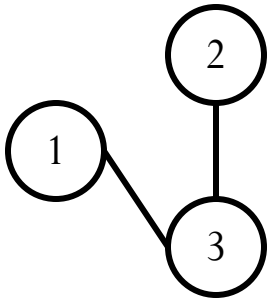
Connections

- So, we've talked about multivariate data
- In these data, there can also be connections (e.g., person-person, person-location)
- One of the most common ways to represent and explore these connections is through graph drawing and visualization
- In fact, there is a whole research community that focuses on the creation of graph layout techniques: http://www.gd2012.org/

What is a Graph?

- **Vertices** (nodes) connected by **Edges** (links)
- We have several ways to represent a graph
  - Adjacency list

Adjacency List
1:3
2:3
3:1,2



  - Adjacency matrix (this is analogous to the weights matrix in last lecture)

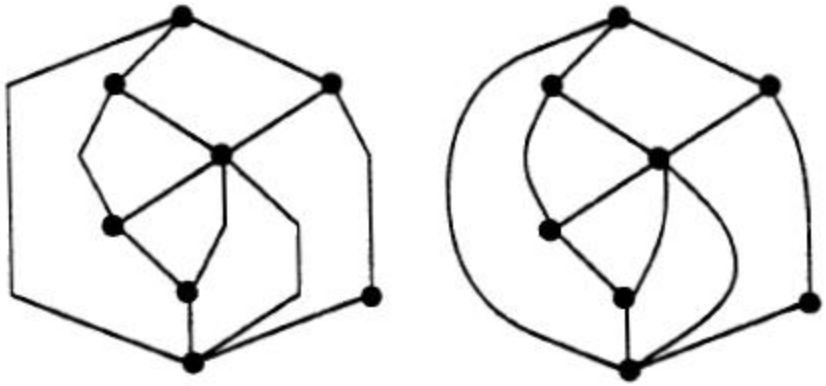
	1	2	3
1	0	0	1
2	0	0	1
3	1	1	0

Graph Representations

- Adjacency list representation is usually preferred because it provides the most compact way to represent *sparse* graphs
  - Sparse graph: $|E| << |V|^2$
- Adjacency matrix representation may be preferred for a highly connected graph

How to Draw a Graph?

- Various graphic standards have been proposed for the representations
  - Vertices are typically circles or boxes and edges are a simple curve
- So, a graph drawing algorithm takes in the edges and vertices of a graph and outputs a pictorial representation based on some algorithmic graphic standard



How to Draw a Graph?

- A graph has infinitely many ways to be laid out
- What we are concerned with is the usefulness of the drawing or the capability of conveying the meaning and relationships
- These are often expressed by means of aesthetics
  - A fundamental and classic aesthetic is the minimization of edge crossings
  - Might want to minimize the area on the screen that the graph takes up
  - May want to optimize displays of symmetry

P Eades and R Tamassia, "Algorithms for Drawing Graphs: An Annotated Bibliography," *Technical Report No. CS-89-09*, Brown University, October 1989.

Force-Directed Methods

- Use a physical analogy to draw graphs
- View a graph as a system of objects with forces acting between them
- Assumption is that a balanced system gives a good layout
- Specifically, a system configuration with locally minimum energy
  - The sum of forces on each object is zero

Force-Directed Methods

- Consider each node/vertex to be the object of the system
- The nodes interact with each other based on some forces
- Edges do not interact with each other, instead, they add new forces to each node
- Equilibrium configuration is when the sum of forces on each node is zero
- Methods typically have two parts
  - Model – a force system defined by vertices and edges
  - Algorithm – a technique for finding equilibrium

Graph Drawing Shortcomings

- Difficult to encode more variables of data cases
  - Shape
  - Size
  - Color
  - All of these can clash with the basic node-link structure

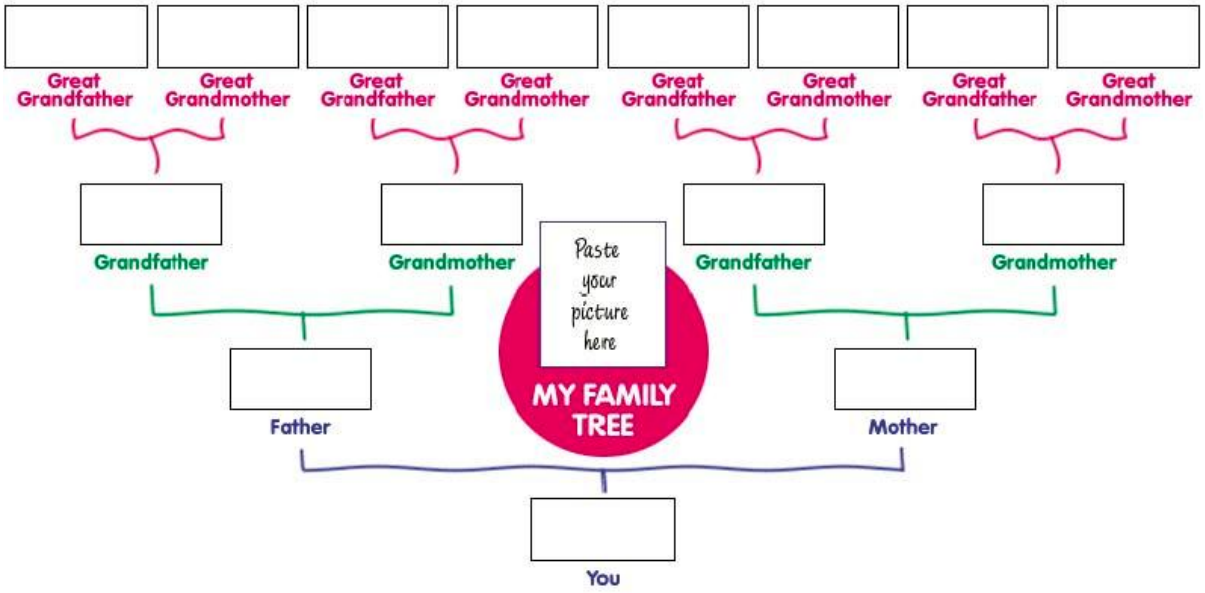
Parts of this slide borrowed from John Stasko's Information Visualization Course Lecture: Hierarchies and Trees 2(Space Filling): http://www.cc.gatech.edu/~stasko/7450/Notes/tree2.pdf

Hierarchies

- So we've been discussing Graphs
- Hierarchies are often represented as a particular type of graph called a tree which are:
  - Directed
  - Acyclic
- Two main representation schemes
  - Node-link (this is similar to the graphical representations we've discussed earlier)
  - Space-Filling

Hierarchies

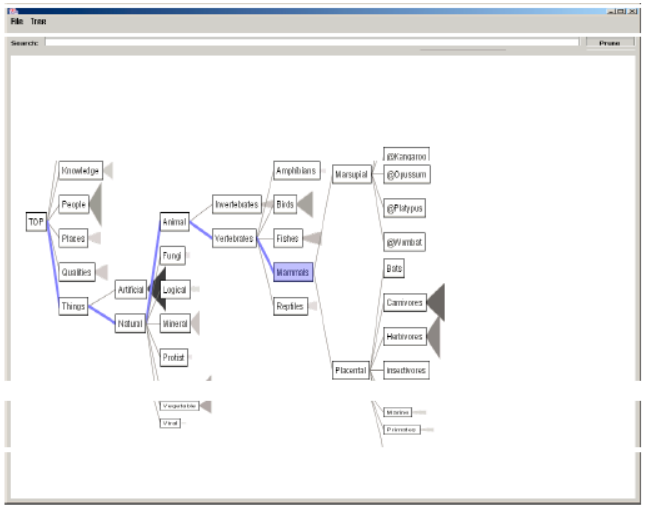
- Definition
  - Data repository in which cases are related to subcases
  - Can be thought of as imposing an ordering in which cases are parents or ancestors of other cases

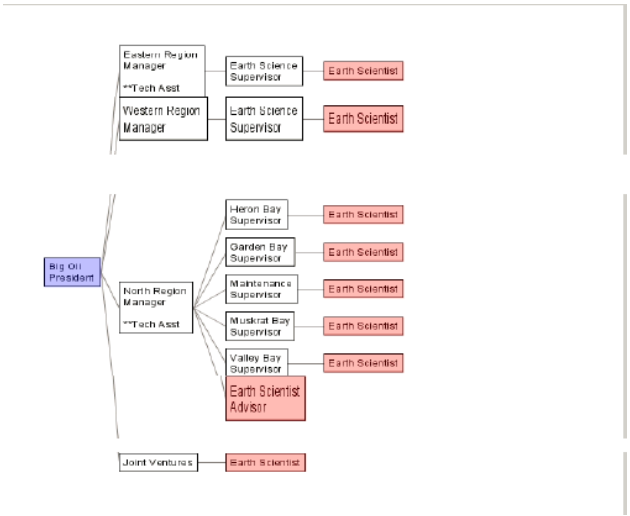


Parts of this slide borrowed from John Stasko's Information Visualization Course Lecture: Hierarchies and Trees 1(Node-Link): http://www.cc.gatech.edu/~stasko/7450/11s/Talks/tree1.pdf

Space Tree

- Visualization techniques try to overcome some of these issues in node link tree diagrams
- *Space Tree* by Plaisant et al.
  - Dynamic rescaling of branches to best fit available screen space
  - Utilized preview icons to summarize branch topology

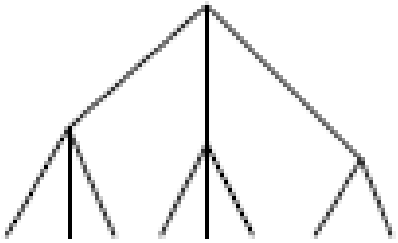


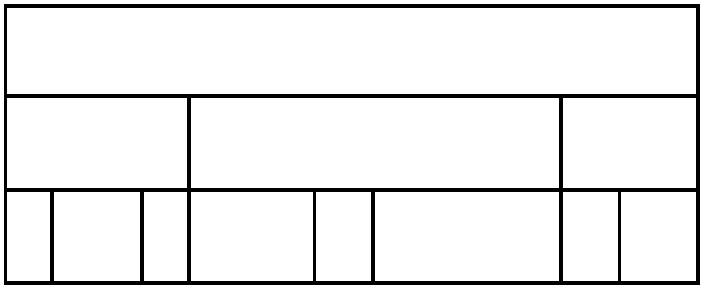


Plaisant, C.; Grosjean, J.; Bederson, B.B.; , "SpaceTree: supporting exploration in large node link tree, design evolution and empirical evaluation," *IEEE Symposium on Information Visualization*, pp. 57- 64, 2002

Space-Filling Representation

- Each item now occupies an area
- Children are contained under the parent





One example: "Icicle plot"

Parts of this slide borrowed from John Stasko's Information Visualization Course Lecture: Hierarchies and Trees 2(Space Filling): http://www.cc.gatech.edu/~stasko/7450/Notes/tree2.pdf

Treemap

- Space filling representation developed by Johnson and Shneiderman
- Children are drawn inside their parent
- Alternate horizontal and vertical slicing at each successive level
- Use area to encode other variable of data items

Johnson, B. and Shneiderman, B. Treemaps: A Space-Filling Approach to the Visualization of Hierarchical Information Structures. In Proceedings of the IEEE Information Visualization '91, pages 275–282, IEEE, 1991.

Treemap Algorithm

Draw()

{

1) Change orientation from parent (horiz/vert)

2) Read all files and directories at this level

3) Make rectangles for each, scaled to size

4) Draw rectangles using appropriate size and color

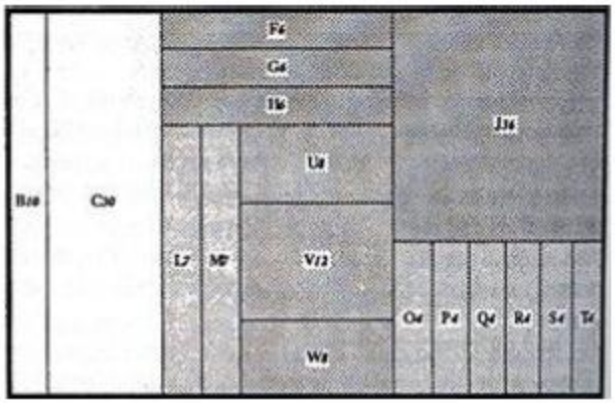
5) For each directory

a) Make recursive call using its rectangle as focus

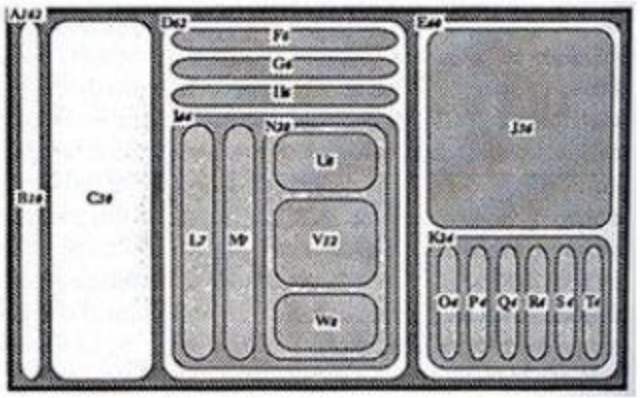
}

Parts of this slide borrowed from John Stasko's Information Visualization Course Lecture: Hierarchies and Trees 2(Space Filling): http://www.cc.gatech.edu/~stasko/7450/Notes/tree2.pdf

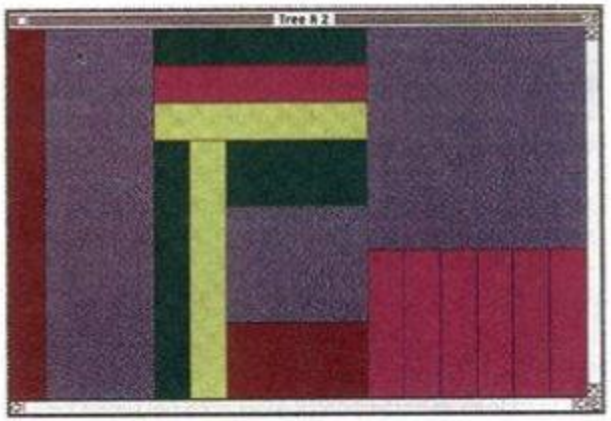
Nested vs. Non-Nested

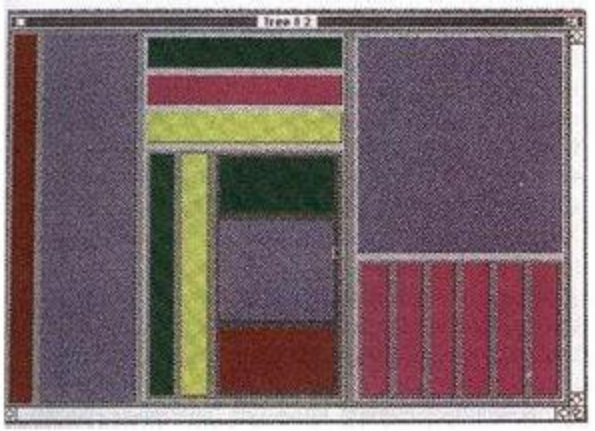


Non-nested Tree-Map



Nested Tree-Map

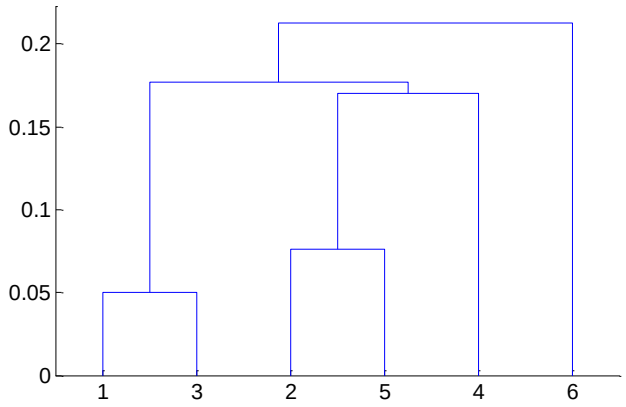


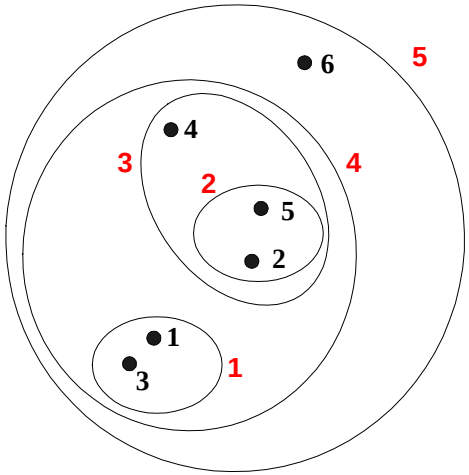


Parts of this slide borrowed from John Stasko's Information Visualization Course Lecture: Hierarchies and Trees 2(Space Filling): http://www.cc.gatech.edu/~stasko/7450/Notes/tree2.pdf

Hierarchical Clustering

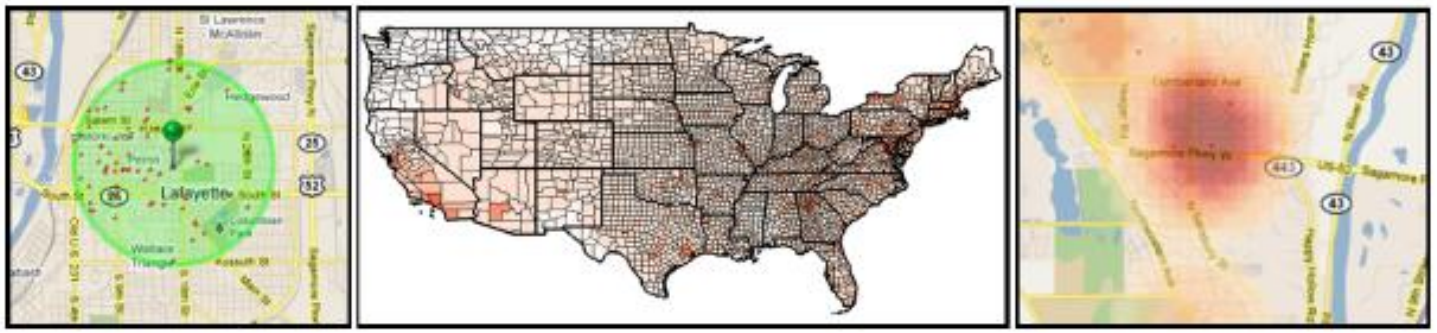
- Produces a set of ***nested clusters*** organized as a hierarchical tree
- Can be visualized as a **dendrogram** (along with other options)
  - A tree-like diagram that records the sequences of merges or splits



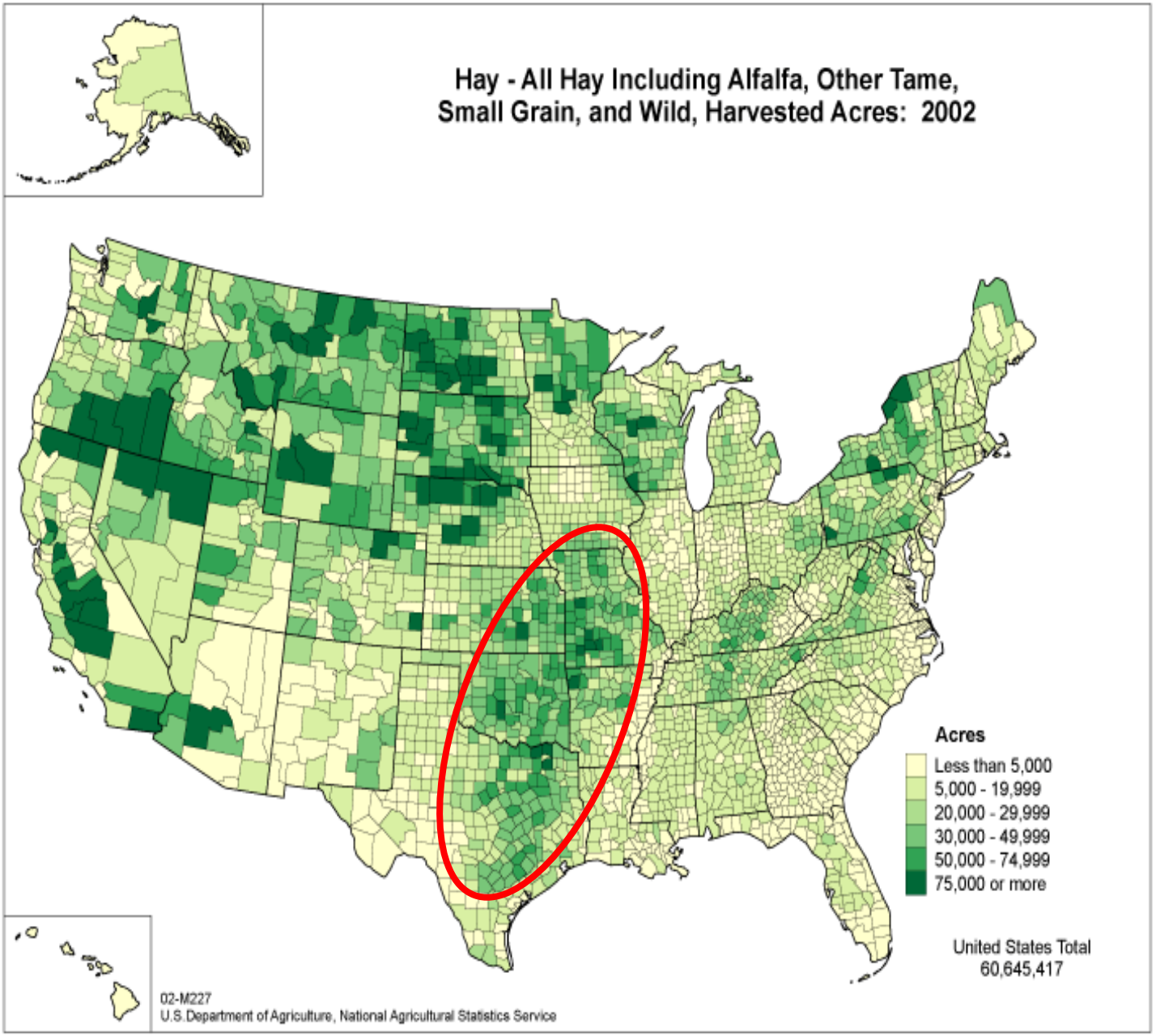


Geovisualization

- Geovisualization primarily denotes tools and techniques that are designed to support analyses that focus on datasets with a geographic component
- Visual representations are designed and built utilizing cartographic principles
- Looking for trends over geographic regions



http://www.nass.usda.gov



riable

he

tion Phenomena.

Coloring Choropleth Maps[1]

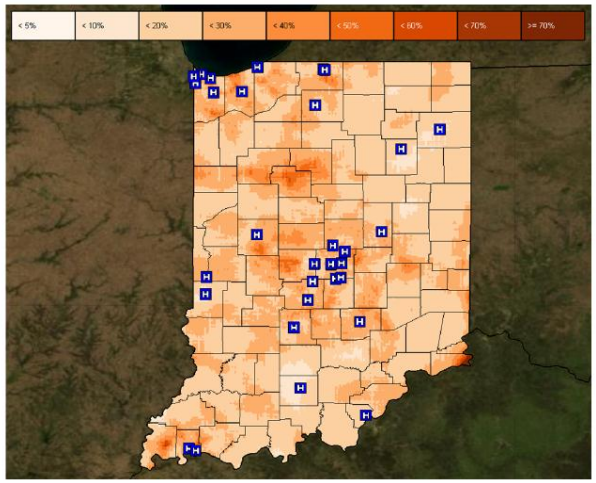
- The number of colors depends on the number of classes (bins)
- Too many classes can overwhelm the user and distract them from seeing trends
- Too many classes can compromise legibility as colors become difficult to distinguish
- Typical cartographic rule of thumb is 5-7 classes
- Typical coloring schemes[8] include sequential, divergent and qualitative
- For more details on mapping as a visual representation, see [2]

1 – C. A. Brewer. http://colorbrewer.org

2 – A. MacEachren. *How MapsWork*. Guilford Press, 1995.

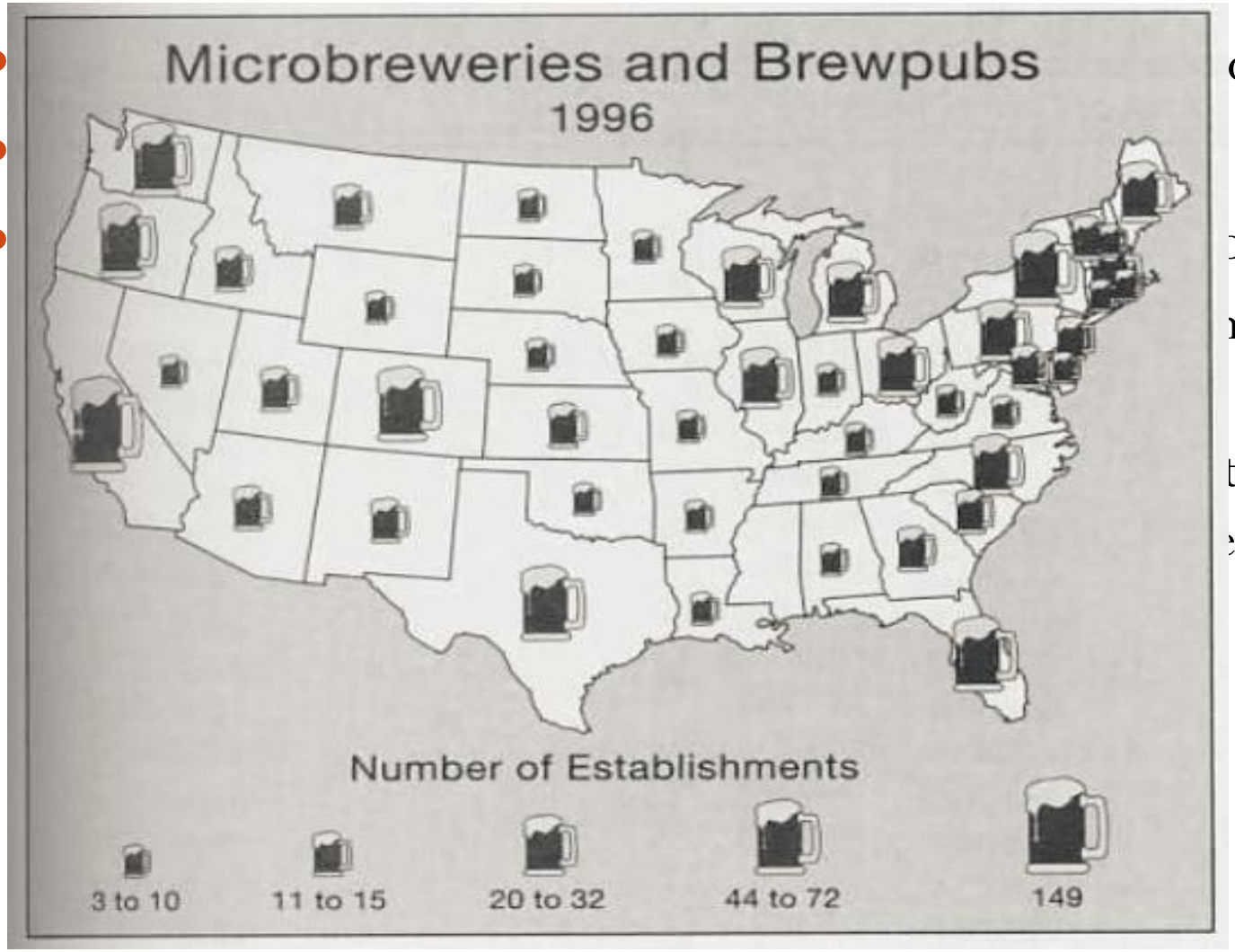
Isopleth Map

- Isarithmic map or contour map is created by interpolating a set of isolines between sample pints of known values
- Created from collection of point data and then interpolating unknown values between them
- Interpolation methods:
  - Kernel density estimation (see earlier lectures)
  - Inverse distance
  - Kriging



TA Slocum, RB McMaster, FC Kessler and HH Howard, "Thematic Cartography and Geovisualization," Prentice Hall Series, 2009.

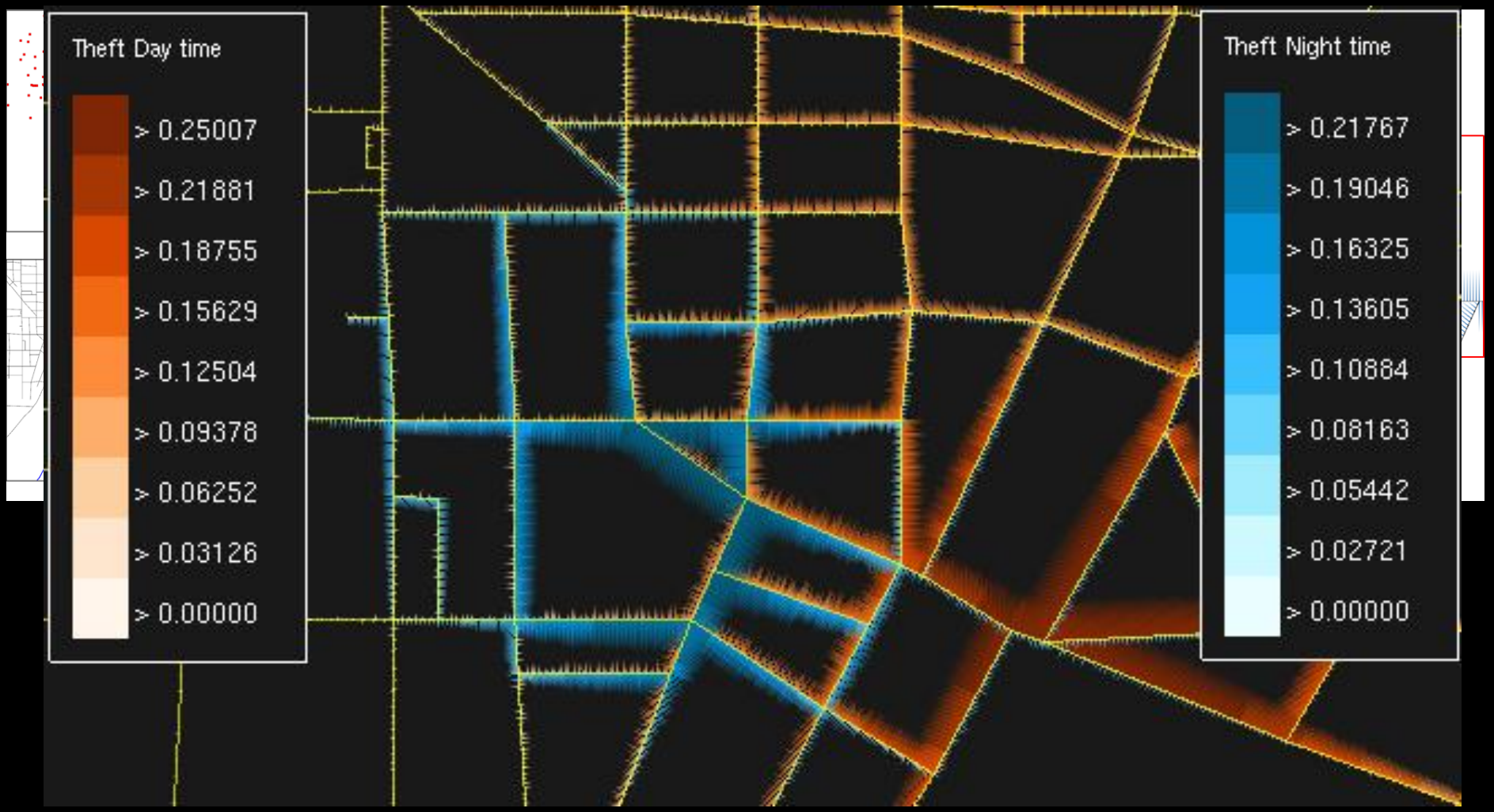
Proportional Symbol Mapping



- ons
-
- oint data
  - ıs
  - t data is
  - entroid)

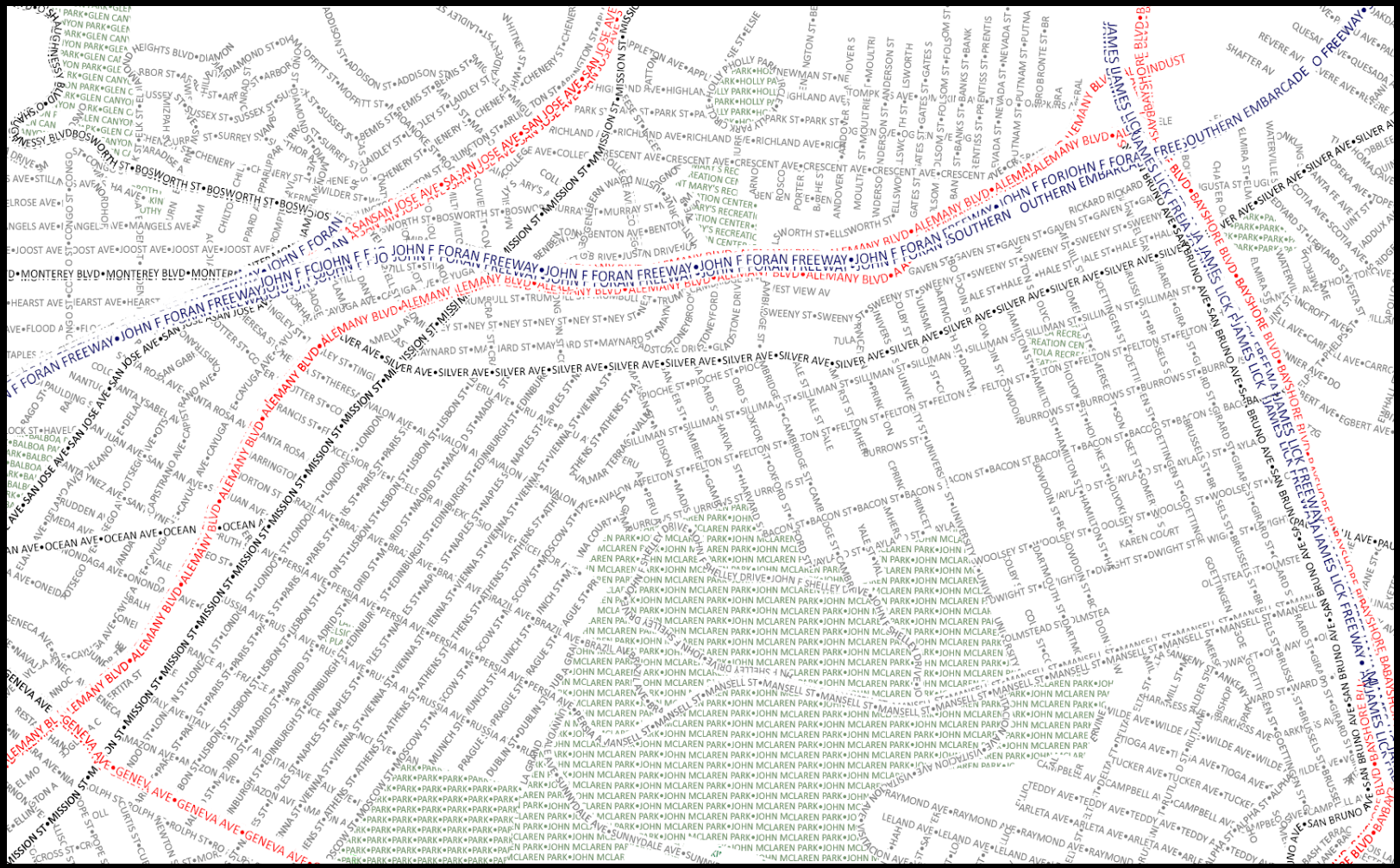
TA Slocum, RB McMaster, FC Kessler and HH Howard, "Thematic Cartography and Geovisualization," Prentice Hall Series, 2009.

Bristle Maps: A Multivariate Abstraction Technique for Geovisualization



Kim, S., Maciejewski, R., Malik, A., Jang, Y., Ebert, D. S., Isenberg, T., "Bristle Maps: A Multivariate Abstraction Technique for Geovisualization," *IEEE Transactions on Visualization and Computer Graphics*, (To appear)

Typographic Maps: An Illustrative Abstraction Technique



1 – Afzal, S., Maciejewski, R., Jang, Y., Elmqvist, N., Ebert. D. S., "Spatial Text Visualization Using Automatic Typographic Maps," *IEEE Transactions on Visualization and Computer Graphics*, 18(12): 2556-2564, 2012.

Time-Oriented Data

- Time is an outstanding dimension reflected by Shneiderman's Task by Data Type Taxonomy[1]
- Time oriented data is ubiquitous, stock markets, movie trends, business, medicine
- Each data case is likely an event of some kind, with one variable being the date and time

W. Aigner, S. Miksch, W. Muller, H. Schumann, C. Tominski, "Visual Methods for Analyzing Time-Oriented Data", *IEEE Trans. On Visualization and Computer Graphics*, Vol. 14, No. 1, Jan.-Feb. 2008, pp. 47-60.

Time Series

- A random selection of 4000 graphs from 15 newspapers and magazines worldwide showed that between 1974 and 1980, 75% of these graphs were time series[1]
- What questions can we ask of these visuals?[2]
  - Does a data object exist at a certain time?
  - When does a certain data object exist?
  - How long does a data object exist?
  - How fast and how much does the data object change?
  - What order to objects appear/disappear?
  - Is there a cyclical pattern to appearances?
  - Which objects exist simultaneously?

1 – E. Tufte, “The Visual Display of Quantitative Information,” 1983.
2 – A. MacEachren, “How Maps Work,: Representation, Visualization and Design,” The Guilford Press, 1995.

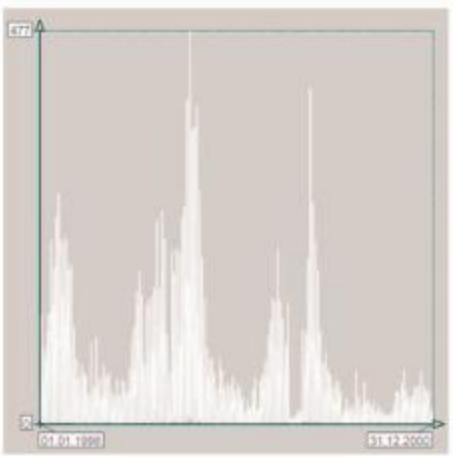
Time

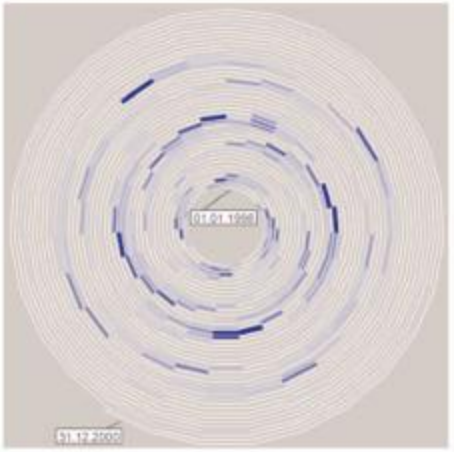
- Time is ordered
  - We assume one time point precedes another
  - The notion of time being ordered is closely bound to the notion of causality (things that happen close together in time are often related)
- Time is continuous
  - We often break time into discrete components for analysis
  - Most people think of time as a flowing stream
- Time is cyclical
- Time is independent of location

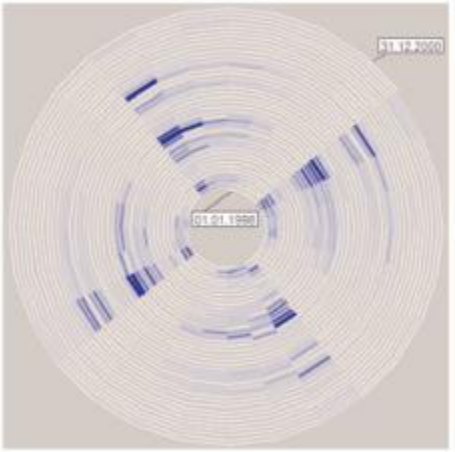
L. Wilkinson, The Grammar of Graphics, *Springer*, 2005

Linear vs. Cyclical Time

- Linear time assumes a starting point and defines the time domain with data elements from the past and future
- Natural processes may be cyclic (e.g., seasons)
  - The ordering of points in a cyclic time domain would be meaningless
  - Winter comes before summer, but also after summer



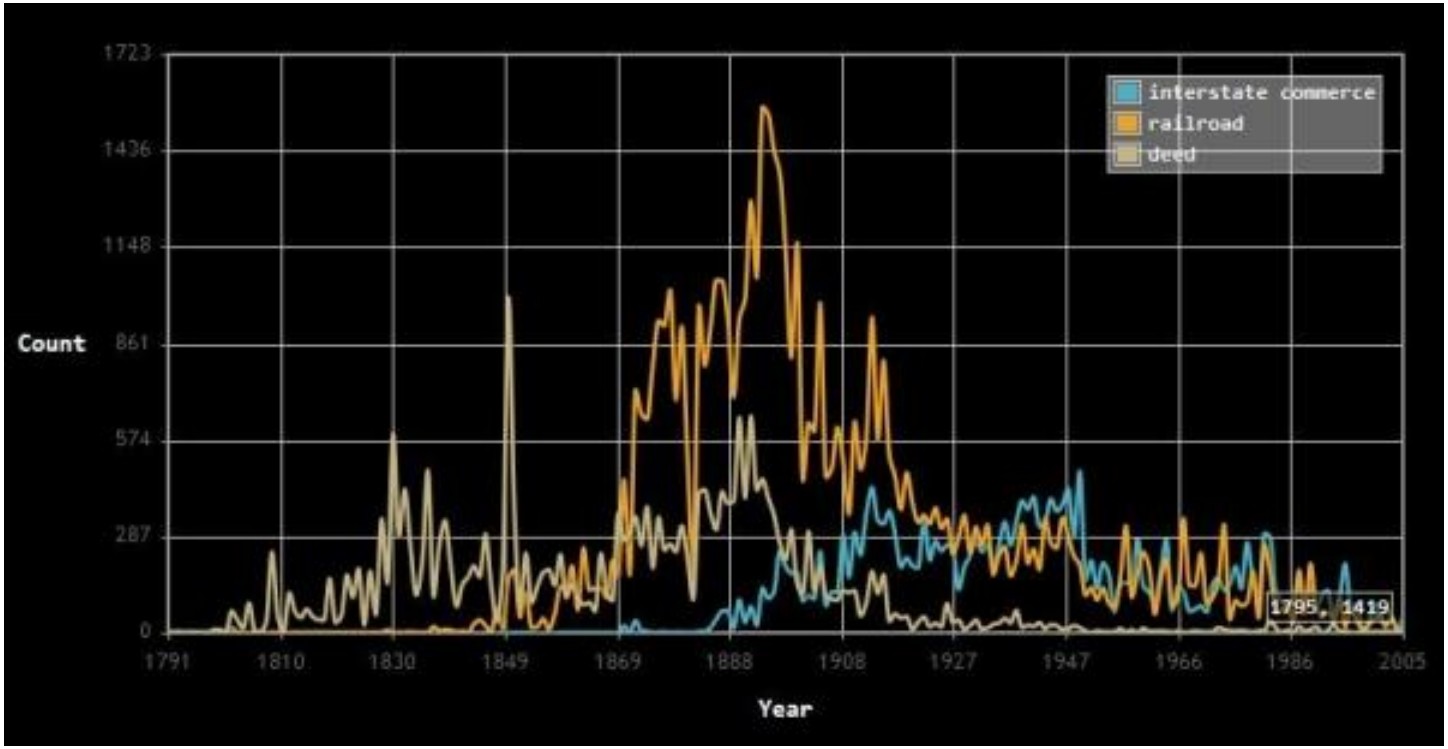




W. Aigner, S. Miksch, W. Muller, H. Schumann, C. Tominski, "Visual Methods for Analyzing Time-Oriented Data", *IEEE Trans. On Visualization and Computer Graphics*, Vol. 14, No. 1, Jan.-Feb. 2008, pp. 47-60.

Linear Time

- Standard presentation
  - 2D line graph with time on x-axis and value on y-axis



New Database Creates Time-Series Plots of Phrases in U.S. Supreme Court Opinions December 16, 2011 | Posted by: Scott C. Idleman: http://law.marquette.edu/facultyblog/2011/12/16/new-database-creates-time-series-plots-of-phrases-in-u-s-supreme-court-opinions/

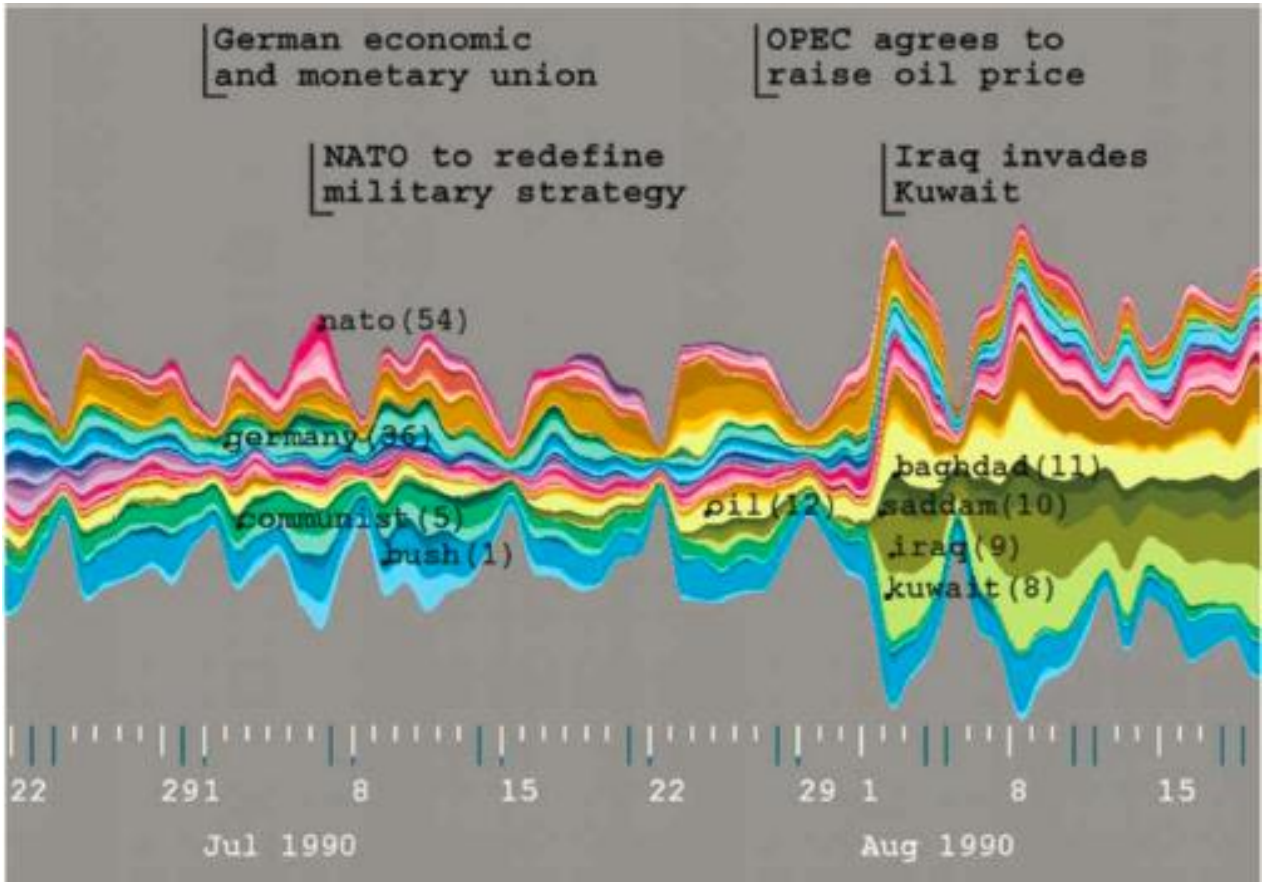
Sparkline

- "Intense, simple, wordlike graphics"



http://sparkline.org/

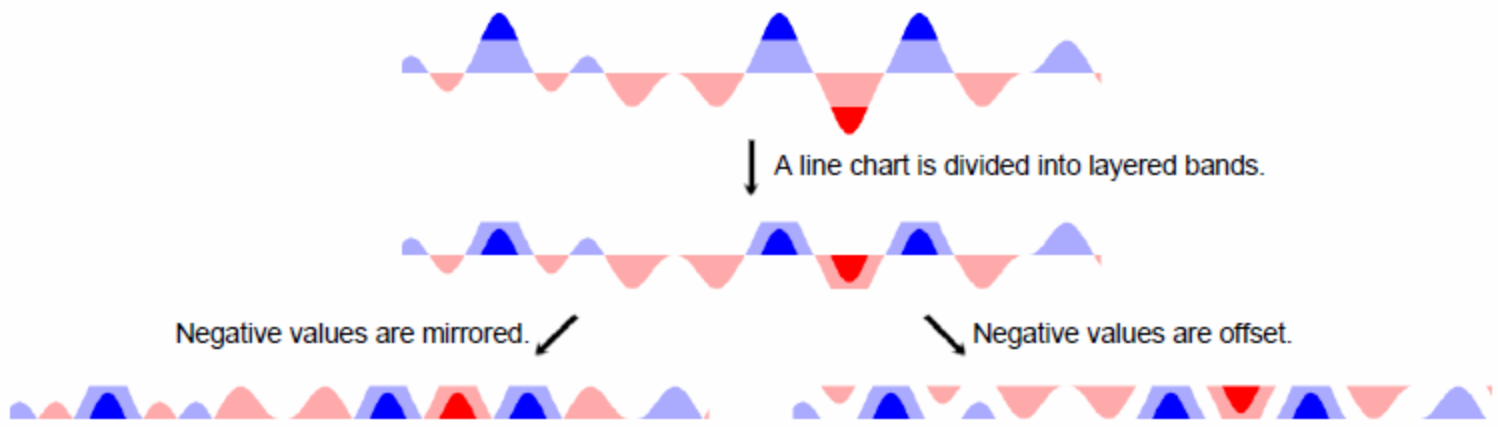
Theme River



1– S. Havre, E. Hetzler, P. Whitney, and L. Nowell. *Themeriver:Visualizing thematic changes in large document collections*. IEEE Transactions on Visualization and Computer Graphics 8(1):9-20, 2002.

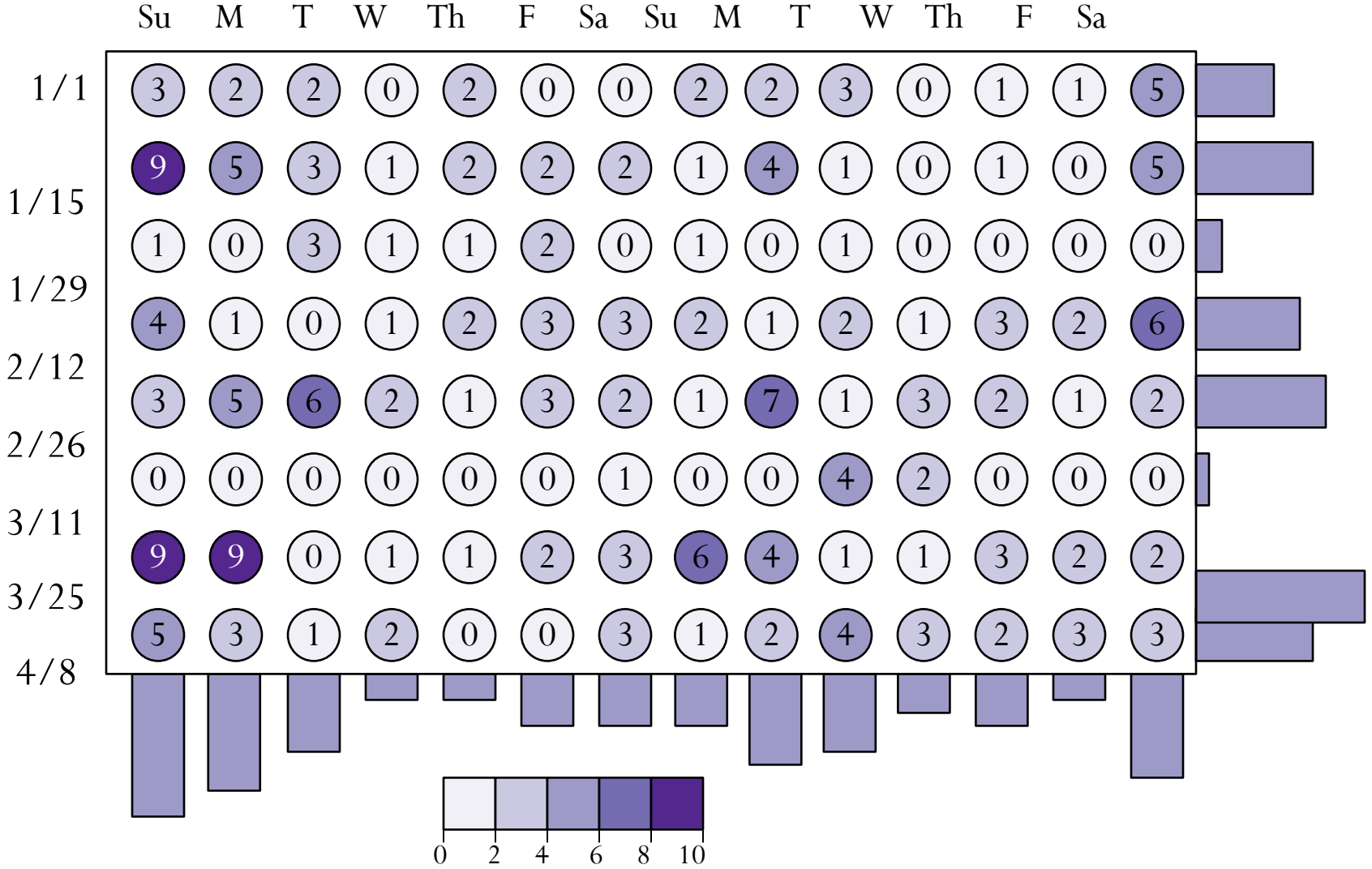
* - Theme River Image from: http://www.nytimes.com/interactive/2008/02/23/movies/20080223_REVENUE_GRAPHIC.html#

Sizing the Horizon



Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. 2009. Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the 27th international conference on Human factors in computing systems* (CHI '09). ACM, New York, NY, USA, 1303-1312.

Calendar Based Visualization



1 – J. J. Van Wijk and E. R. Van Selow. *Cluster and calendar based visualization of time series data*. Proceedings of the 1999 IEEE Symposium on Information Visualization, 1999.

Coordinated Multiple Views

- Instead of trying to make the "best" visualization of all of our data, perhaps we can use multiple views
- Data can be expressed in a variety of ways
- Given the multivariate nature of data, a single statistical graphic may not be enough
- Interactive graphics systems can provide multiple representations of the data
- These representations can be *linked* or *coordinated*

North C, Shneiderman B (2000) Snap-together visualization Evaluating coordination usage and construction. International Journal of Human-Computer Studies 51:715-739

Interaction Types

- Highlighting and focus
- Drill-down and hyperlinks
- Overview and context
- Changing parameters
- Changing representations
- Temporal fusion

A. Dix and G. Ellis. **Starting Simple - adding value to static visualisation through simple interaction.** *Proceedings of Advanced Visual Interfaces* , 1998

Interaction Types

- How do we think of interacting with data?
  - Comparing and sorting
    - Provide a selection of data graphics to compare values and sort these values based on information
  - Adding/filtering variables
  - Highlighting
  - Aggregating/combining variables
  - Zoom and Pan
  - Rescale
  - Annotate
  - Bookmark

S. Few. *Now You See It*.

Interaction

- Really, the key is user intent
  - Focusing on what a user wants to achieve through a specific interaction technique
- Interaction is done by a person for a purpose
  - Seeking more information
  - Solving a problem
  - Exploratory analysis
  - Analytic discourse

* - Material borrowed from John Stasko's lecture on Interaction: http://www.cc.gatech.edu/~stasko/7450/Notes/interaction.pdf

Interaction in Visualization

- Broken down into 7 categories
  - Select
  - Explore
  - Reconfigure
  - Encode
  - Abstract/Elaborate
  - Filter
  - Connect

Ji Soo Yi; Youn ah Kang; Stasko, J.T.; Jacko, J.A.; , "Toward a Deeper Understanding of the Role of Interaction in Information Visualization," *IEEE Transactions on Visualization and Computer Graphics,* vol.13, no.6, pp.1224-1231, Nov.-Dec. 2007

Visual Analysis vs. Visual Analytics

- The choice of visual representation will influence the user's analysis
- Massaging the data can help improve the visual representation
- It also makes some data analysis methods easier
- Adding interaction allows the user to form and explore their own hypotheses
- But, how do we find "important" features in the data?

Data Mining

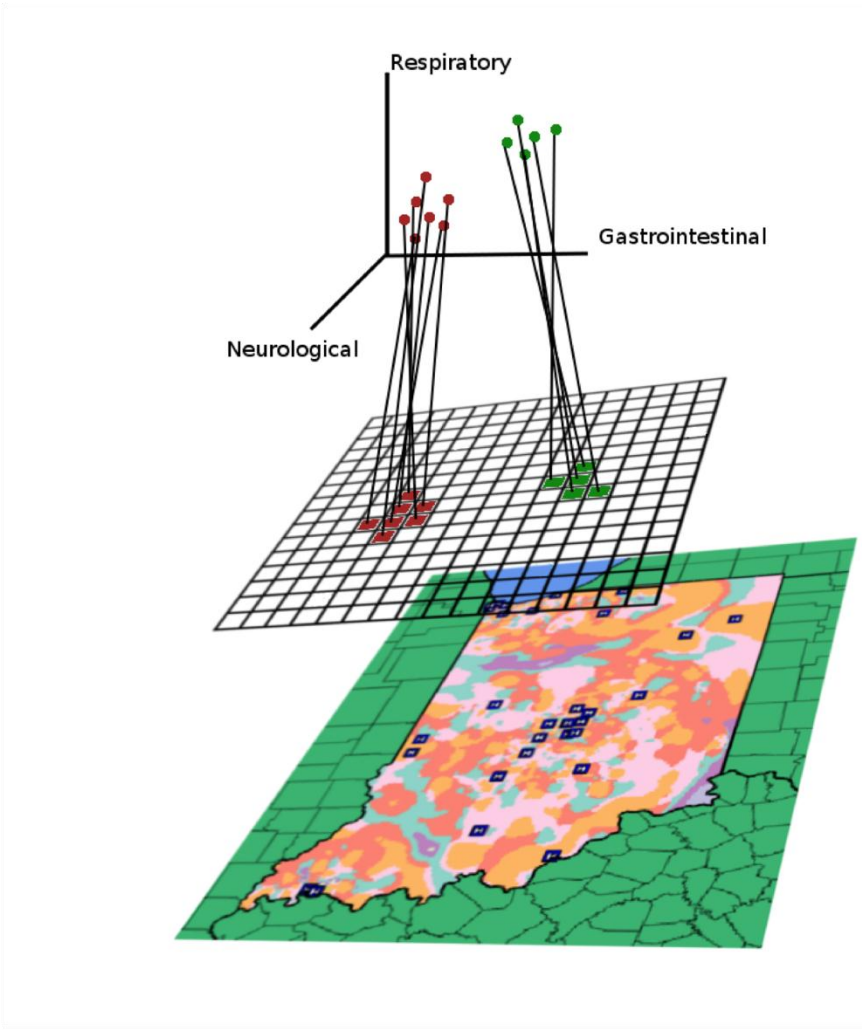
- Data mining domain has techniques for examining data, looking for patterns, looking for anomalies
- Can be used to enhance the visualizations, show us what is important
- Can also be used in exploratory analysis, “I think this looks interesting, show me similar trends.”

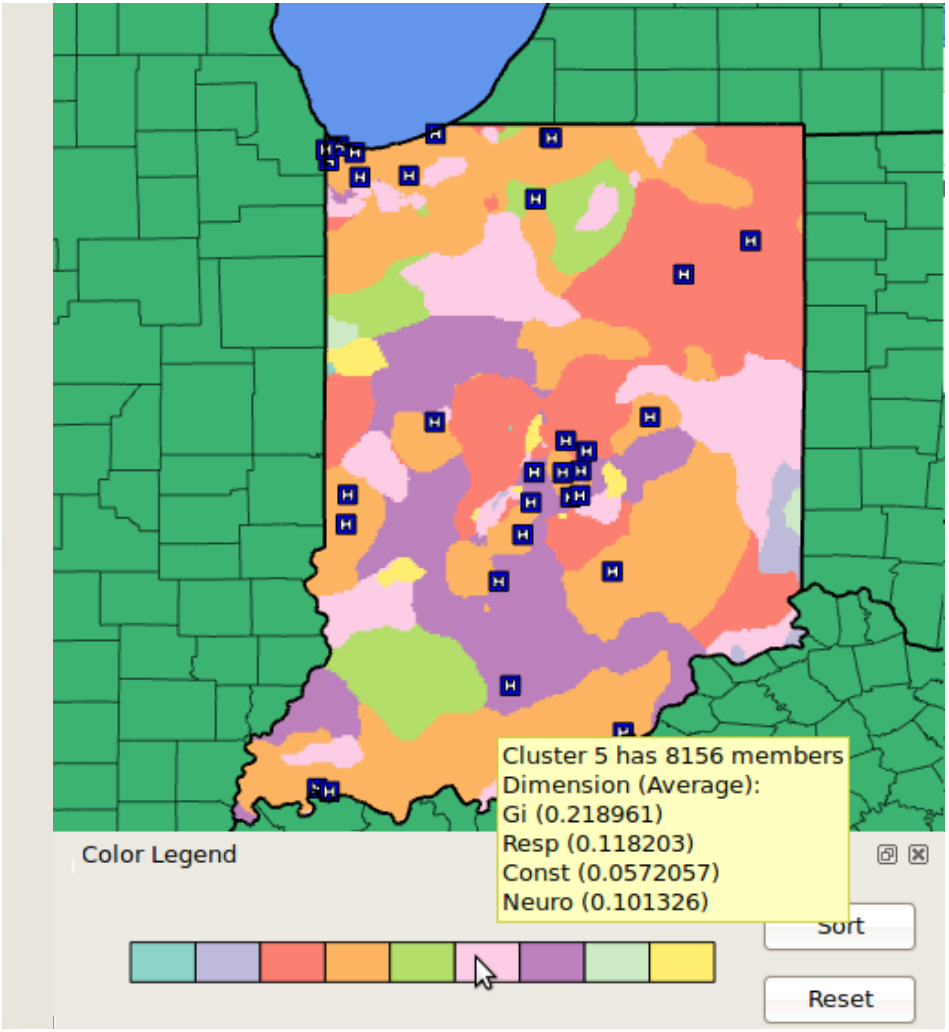
What questions do we tend to ask in an analysis?

- What else is "close" in the data set? What points are close?
- What is the "trend"?
- If we have geographic data:
  - Do neighboring spatial regions have similar values?
  - Do neighboring spatial regions have similar temporal trajectories?
  - Are disparate spatial regions showing similar trajectories or values?
  - Are temporal trends lagging or leading in neighboring spatial regions?

1 - S. Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

Multivariate Clustering

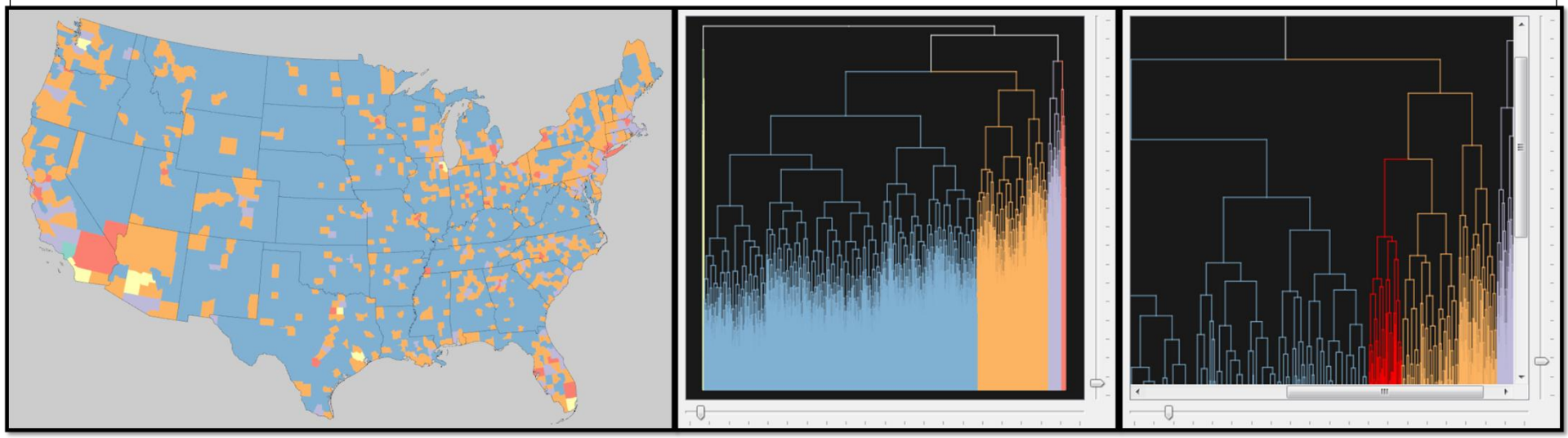




Multivariate Clustering

- Hierarchical Clustering
  - Produces a set of nested clusters organized as a hierarchical tree
  - Can be visualized as a dendrogram

- No assumptions on the number of clusters
  - Any desired number of clusters can be obtained by 'cutting' the dendogram at the proper level



Multivariate Distance Function

- In these high-dimensional clusters, we don't necessarily have a concept of how far away two points are
- We want to know how spatial units are related in high-dimensional space and observe these patterns in some sort of projection

Metric Names	Formula
Euclidean	$dist = \sqrt{\sum_i (v_i^1 - v_i^2)^2}$
Manhattan	$dist = \sum_i \lvert v_i^1 - v_i^2 \rvert$
Canberra	$dist = \sum_i \frac{\lvert v_i^1 - v_i^2 \rvert}{\lvert v_i^1 \rvert + \lvert v_i^2 \rvert}$
Binary	$dist = \sum_i v_i^1 = v_i^2$

Temporal Analysis

- Many of these questions depend on temporal trajectories
- We can apply time series similarity work as a means of determining which regions have similar trajectories
- One common metric is the sequence normalized metric which will focus on matching trends as opposed to exact values

$$dist_{SNE}(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}\left(\frac{q_i-\mu_q}{\sigma_q}-\frac{p_i-\mu_p}{\sigma_p}\right)^2}.$$

What questions do we tend to ask in an analysis?

- Key Temporal questions tend to focus on change[1]
  - What is the magnitude of the change?
  - What is the shape of the changes?
  - How fast is the change occurring (what is the velocity of the change)?
  - What is the direction of the change?

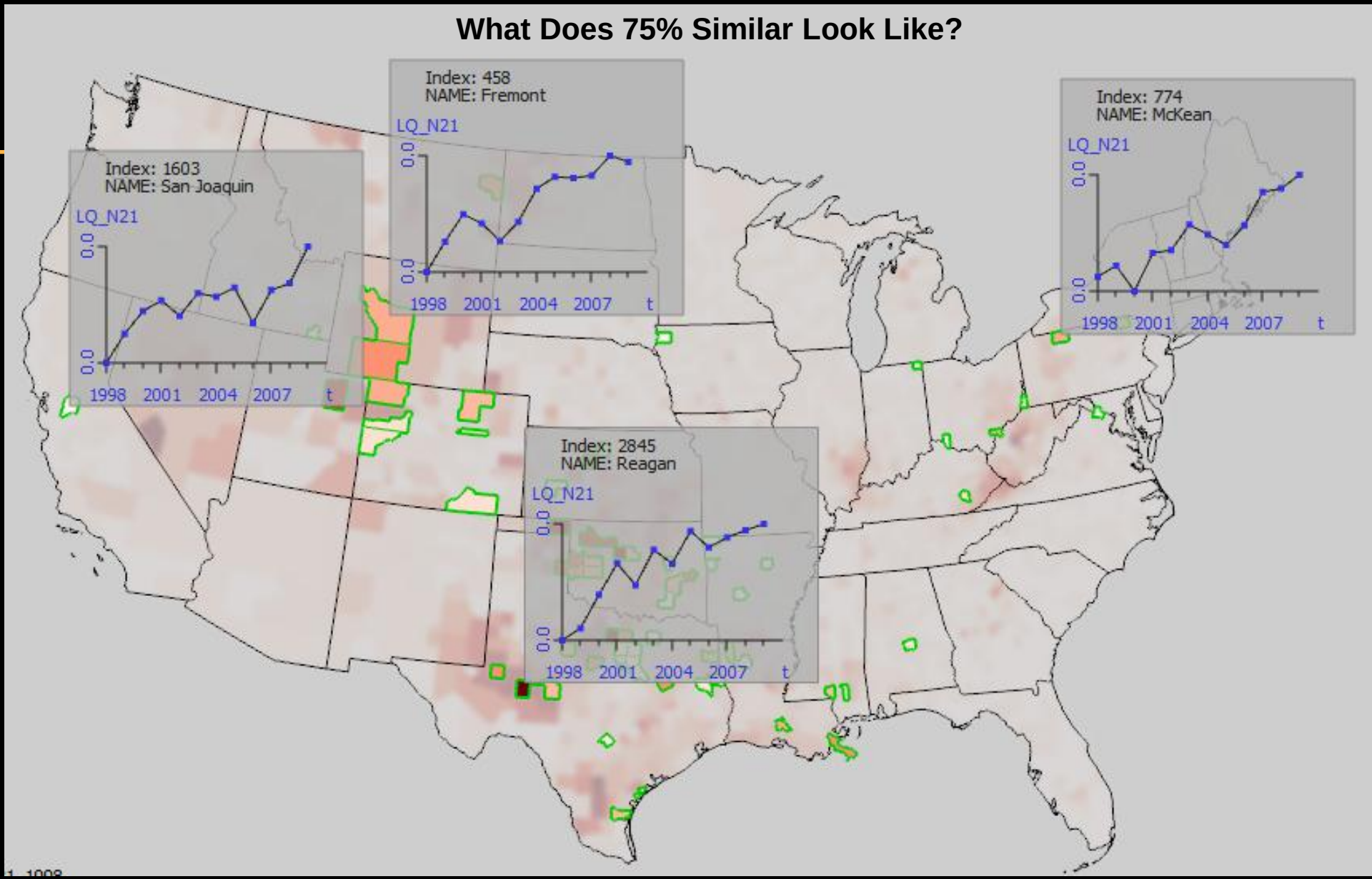
1 - S. Few. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

Similarity

- Many of these questions depend on temporal trajectories
- We can apply time series similarity work as a means of determining which regions have similar trajectories
- One common metric is the sequence normalized metric which will focus on matching trends as opposed to exact values

$$dist_{SNE}(\mathbf{p},\mathbf{q}) = \sqrt{\sum_{i=1}^{n}\left(\frac{q_i-\mu_q}{\sigma_q}-\frac{p_i-\mu_p}{\sigma_p}\right)^2}.$$

What Does 75% Similar Look Like?



Finding Similarities in Time Series

- Indexing problem
  - Find all stocks whose fluctuations are similar to stock X
- Subsequence Similarity
  - Find out other days in which stock x had similar movements as today
- Clustering
  - Group regions that have similar temporal patterns
- Rule Discovery
  - Find rules like if x goes up, y goes up, etc.

Temporal Analysis Methods

- For temporal data, we can find anomalies using control chart methods
- Control charts consist of a statistic representing some measurement in time (number of patients in a hospital, value of a stock, etc.)
- The mean and standard deviation of the statistic is calculated given all the available samples
- If the current value is greater than some pre-set number of standard deviations from the mean, then an alert is generated (Shewart suggested three standard deviations)

Summary

- Visual Analytics combines a wide breadth of topics
  - Visual representations
  - Perception
  - Statistics
  - Human Computer Interaction
  - Machine Learning and Data Mining
  - Cognition and Many More
- This tutorial only glossed over some of the more commonly used methods and functions as an introduction

A Walkthrough Example

National Health and Nutrition Examination Survey (NHANES)

Designed to assess health and nutritional status of adults and children in the US

- *Combines interviews and physical examinations*
- *Exams a nationally representative sample of about 5,000 persons per year*
- *Includes demographic, socioeconomic, dietary and health-related questions*

Healthy Eating Index (HEI)

Measure of diet quality that assesses conformance to federal dietary guidance

USDA uses HEI to monitor diet quality of the US population

HEI-2005 uses recommendations of the MyPyramid guidance as the baseline

1 - Guenther, P. M., Reedy, J., Krebs-Smith, S. M., Reeve, B. B., "Evaluation of the Healthy Eating Index-2005," *Journal of the American Dietetic Association*, 108:1854-1896, 2008.

2 – Freedman, L. S., Guenther, P. M., Krebs-Smith, S. M., Kott, P. S., "A Population's Mean Healthy Eating Index-2005 Scores are Best Estimated by the Score of the Population Ratio when One 24-Hour Recall is Available," Journal of Nutrition, 138:1725-1729, 2008.

Healthy Eating Index–2005 components and standards for scoring[1]

Component		Maximum points	Standard for maximum score	Standard for minimum score of zero
Total Fruit (includes 100% juice)	(HEI-1)	5	≥0.8 cup equiv. per 1,000 kcal	No Fruit
Whole Fruit (not juice)	(HEI-2)	5	≥0.4 cup equiv. per 1,000 kcal	No Whole Fruit
Total Vegetables	(HEI-3)	5	≥1.1 cup equiv. per 1,000 kcal	No Vegetables
Dark Green and Orange Vegetables and Legumes[2]	(HEI-4)	5	≥0.4 cup equiv. per 1,000 kcal	No Dark Green or Orange Vegetables or Legumes
Total Grains	(HEI-5)	5	≥3.0 oz equiv. per 1,000 kcal	No Grains
Whole Grains	(HEI-6)	5	≥1.5 oz equiv. per 1,000 kcal	No Whole Grains
Milk[3]	(HEI-7)	10	≥1.3 cup equiv. per 1,000 kcal	No Milk
Meat and Beans	(HEI-8)	10	≥2.5 oz equiv. per 1,000 kcal	No Meat or Beans
Oils[4]	(HEI-9)	10	≥12 grams per 1,000 kcal	No Oil
Saturated Fat	(HEI-10)	10	≤7% of energy[5]	≥15% of energy
Sodium	(HEI-11)	10	≤0.7 gram per 1,000 kcal[5]	≥2.0 grams per 1,000 kcal
Calories from Solid Fats, Alcoholic beverages, and Added Sugars (SoFAAS)	(HEI-12)	20	≤20% of energy	≥50% of energy

[1]Intakes between the minimum and maximum levels are scored proportionately, except for Saturated Fat and Sodium (see note 5).
[2]Legumes counted as vegetables only after Meat and Beans standard is met.
[3]Includes all milk products, such as fluid milk, yogurt, and cheese, and soy beverages.
[4]Includes nonhydrogenated vegetable oils and oils in fish, nuts, and seeds.
[5]Saturated Fat and Sodium get a score of 8 for the intake levels that reflect the 2005 Dietary Guidelines, <10% of calories from saturated fat and 1.1 grams of sodium/1,000 kcal, respectively.

http://www.cnpp.usda.gov/Publications/HEI/healthyeatingindex2005factsheet.pdf

Playing Data Detective

So, we've collected all of these surveys, now what?

Well, how about Exploratory Data Analysis where we visually inspect the data and ask questions about it to help form hypotheses.

What are some easy questions to ask?

- *What is the average … (total fruit consumption)?*
- *What is the average … (milk consumption)?*

1 – J. W. Tukey. Exploratory Data Analysis

NHANES

Utilizing Survey Cycle 2-3

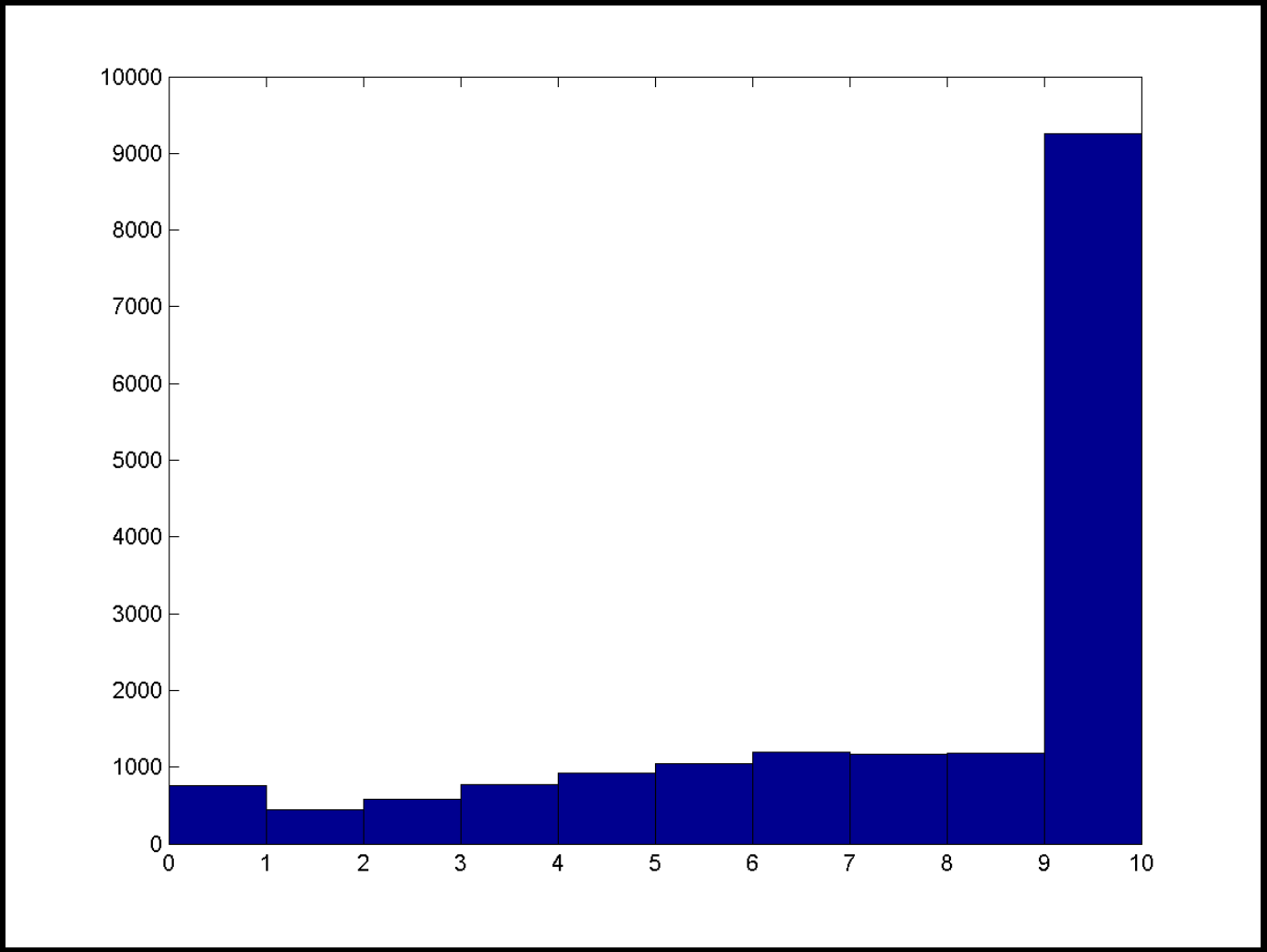
- ***Cycle 2: 2001-2002***
- ***Cycle 3: 2003-2004***
- ***Sample size of ~15000***

Other variables:

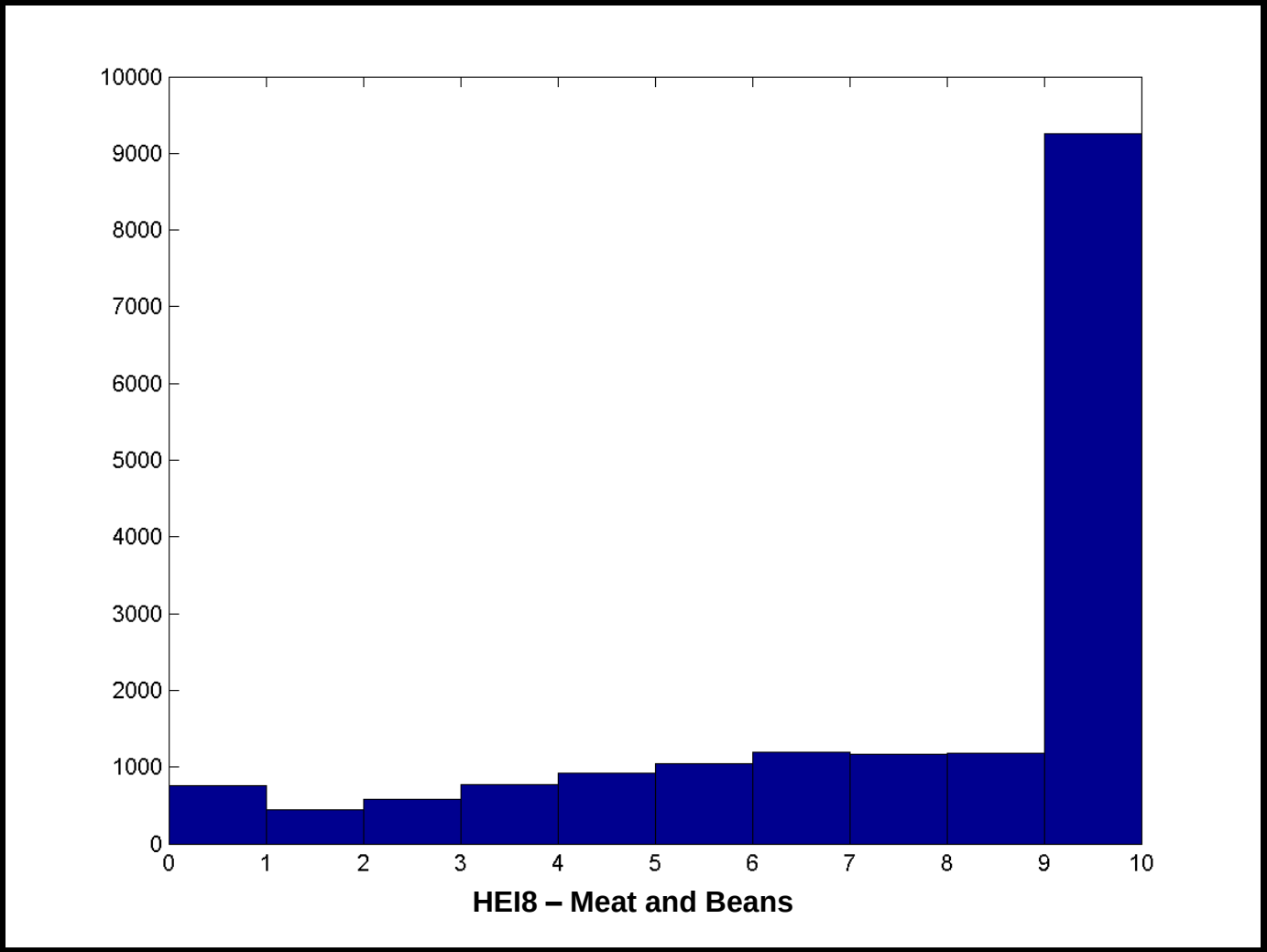
- **Gender**
- **Age**
- **Race**
- **Total energy for the day (kcal)**
- **Total saturated fatty acid (gm)**
- **BMI**

So with more data, maybe we can think of more questions!

What is the average … ?



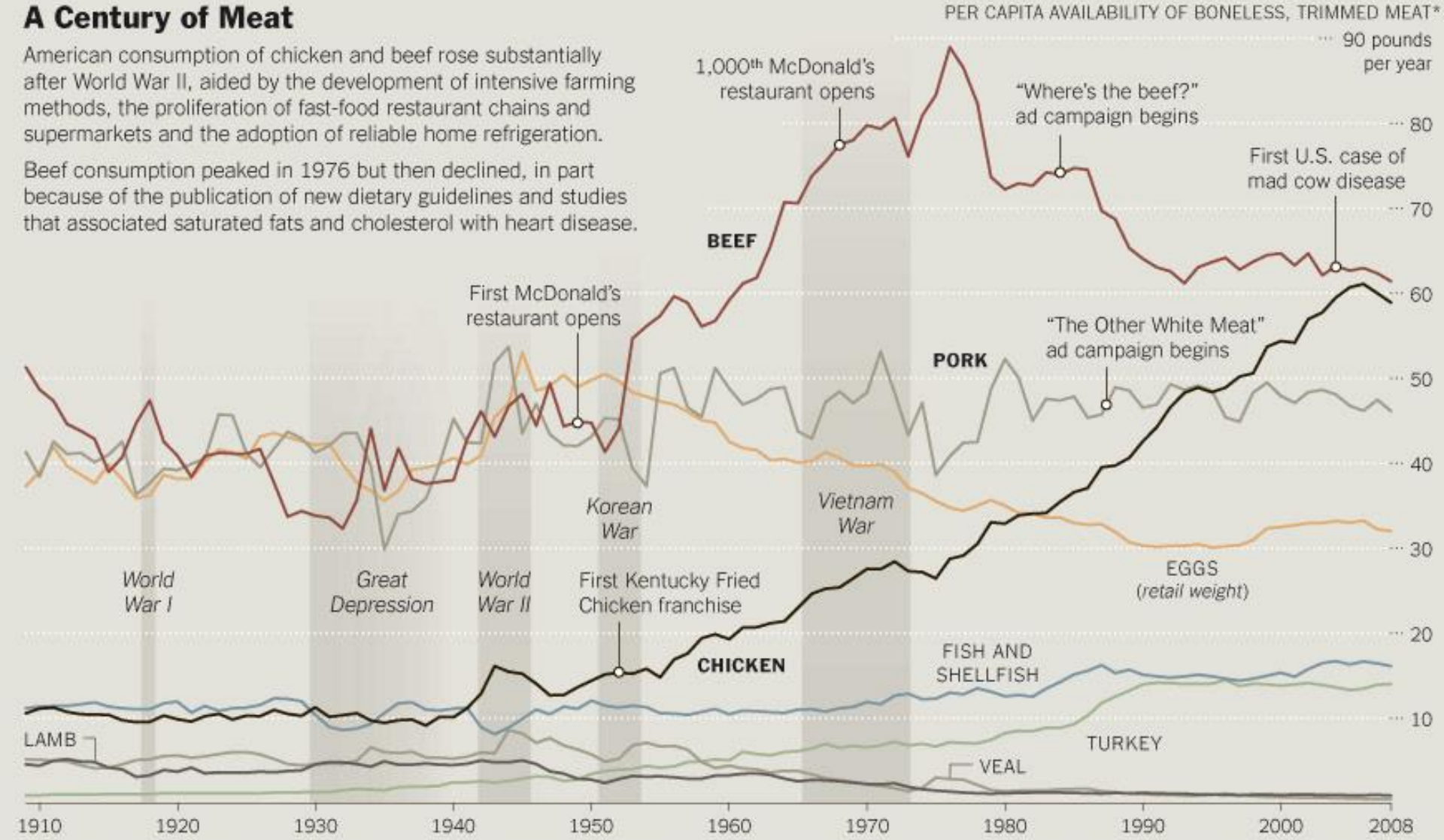
What is the average … ? Meat Consumption



A Century of Meat

American consumption of chicken and beef rose substantially after World War II, aided by the development of intensive farming methods, the proliferation of fast-food restaurant chains and supermarkets and the adoption of reliable home refrigeration.

Beef consumption peaked in 1976 but then declined, in part because of the publication of new dietary guidelines and studies that associated saturated fats and cholesterol with heart disease.



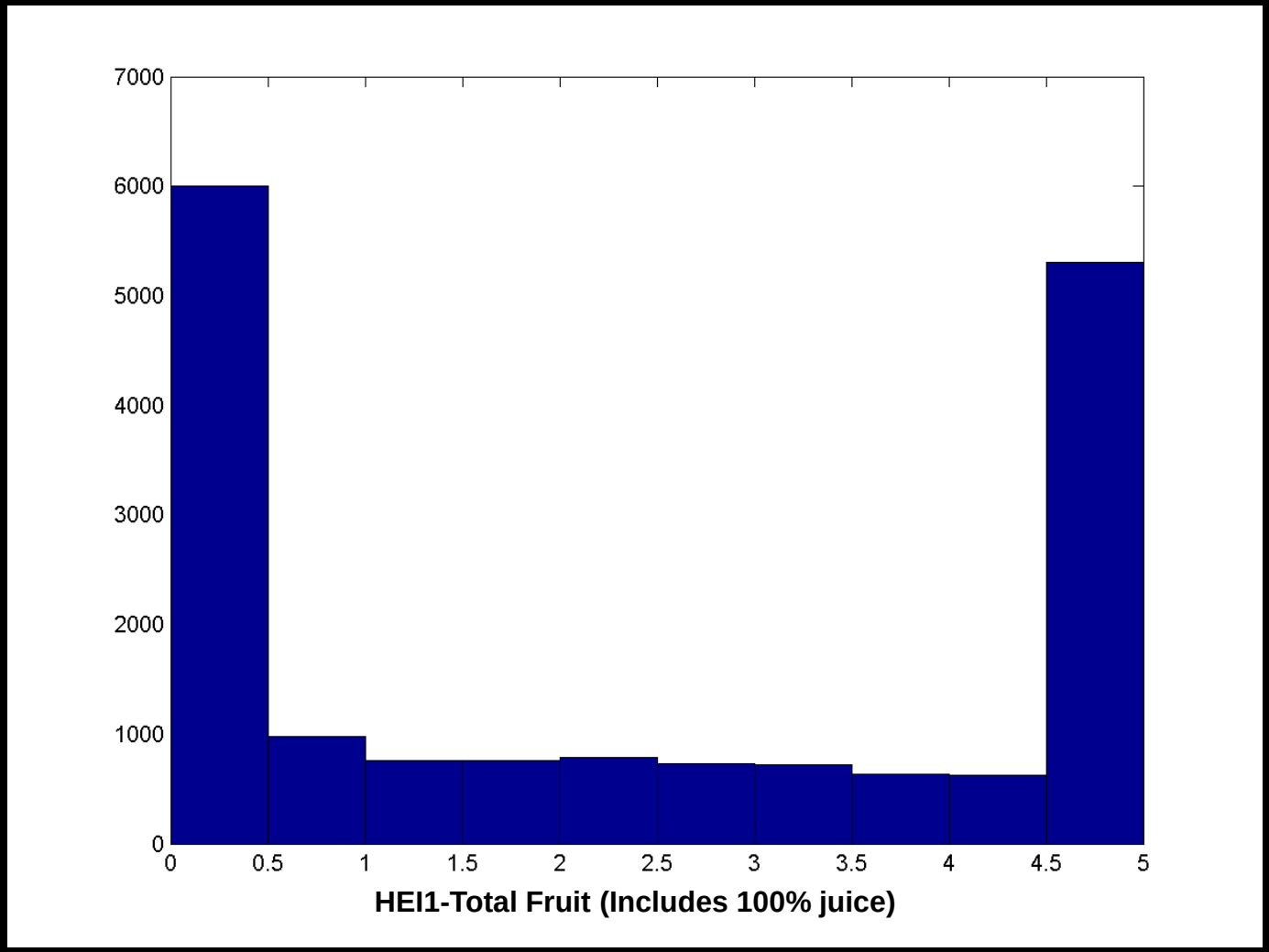
*Note: per capita availability of boneless meat is a proxy for human consumption, and is lower than retail weight or carcass weight. Bones, offal and game are excluded.

Sources: U.S. Department of Agriculture (data); news and company reports; "Putting Meat on the American Table," by Roger Horowitz

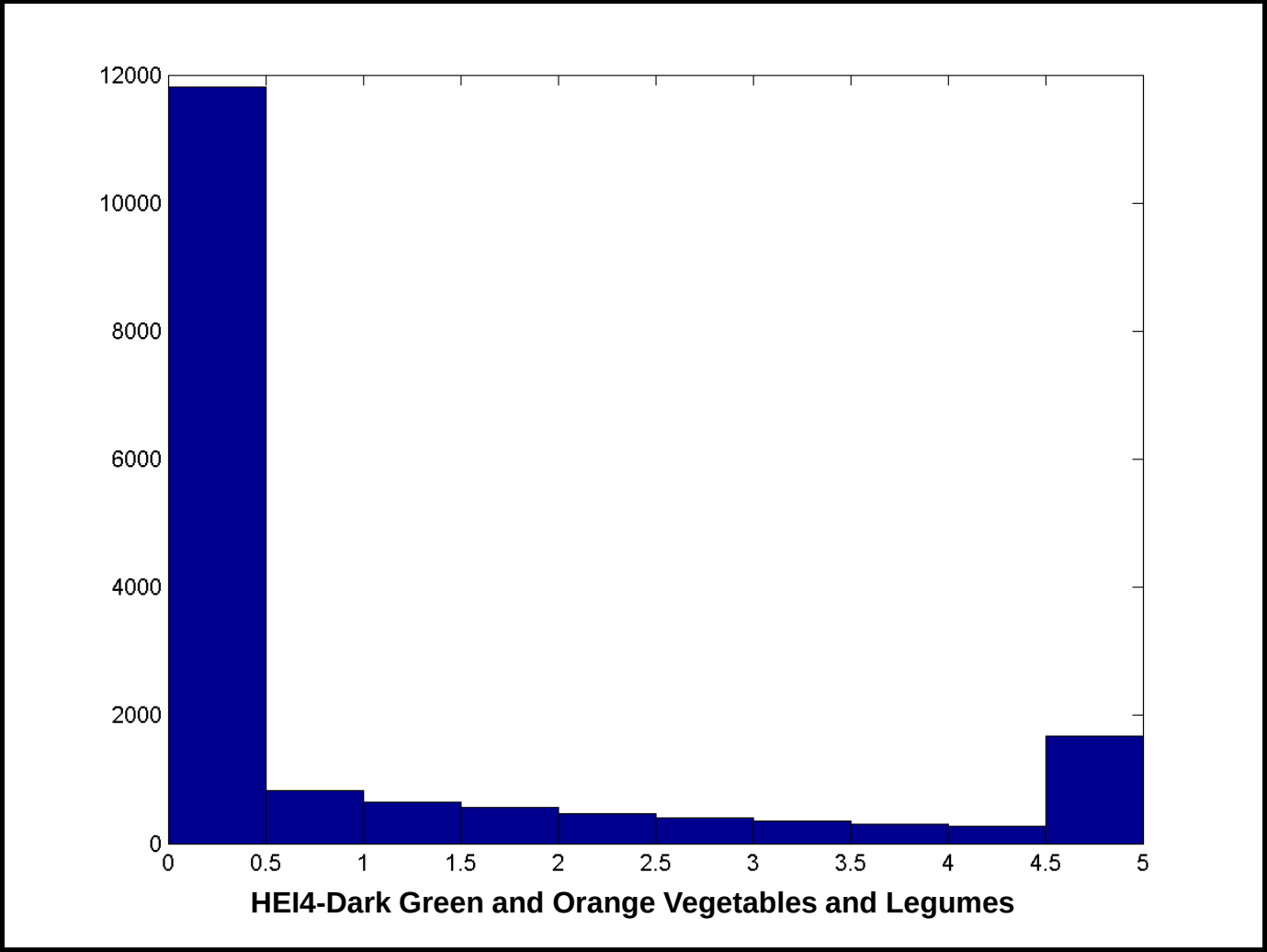
JONATHAN CORUM/THE NEW YORK TIMES

http://www.nytimes.com/imagepages/2011/03/15/science/15food_graphic.html

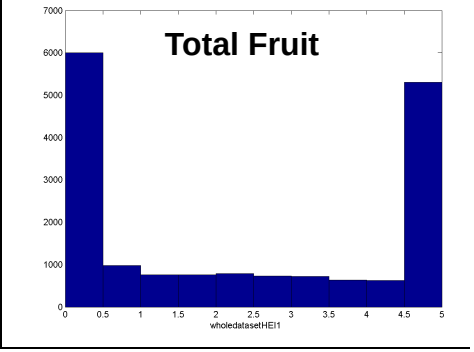
What is the average … ? Total Fruit Consumption

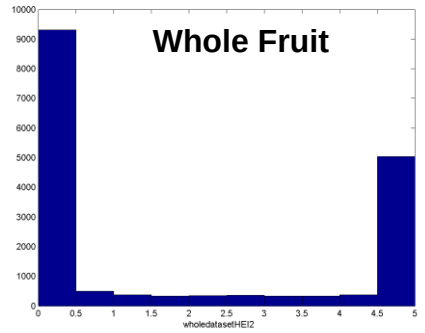


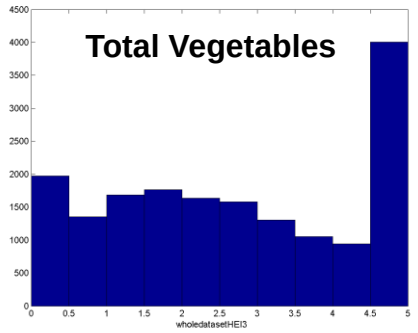
What is the average … ? Dark Green and Orange Vegetable Consumption

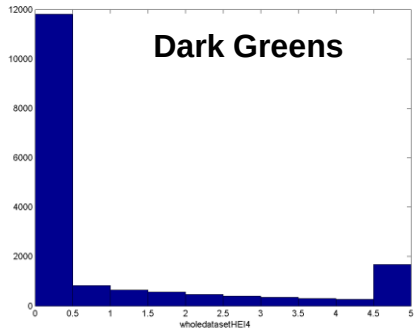


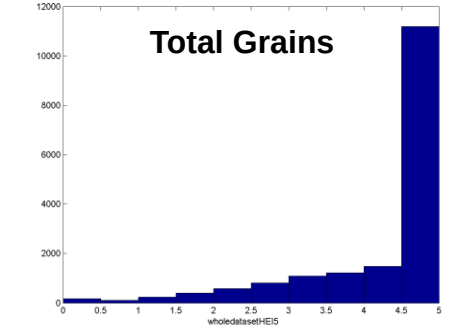
How Do These Compare?

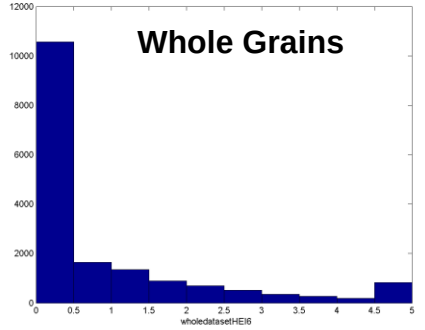


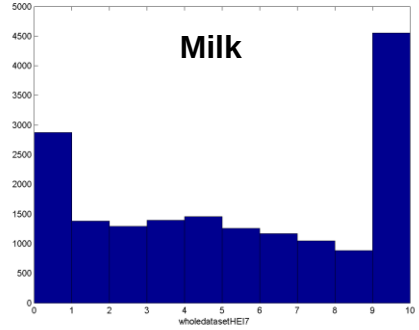


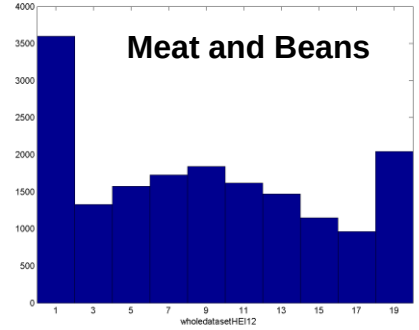


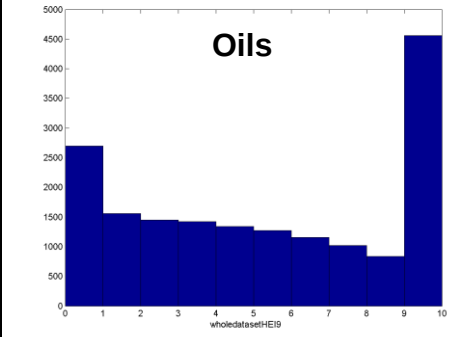


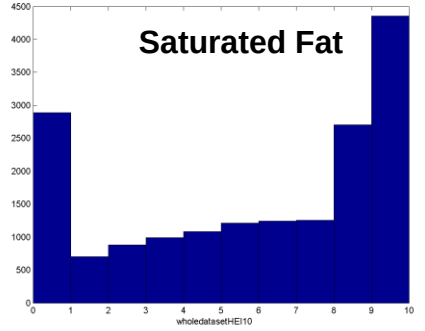


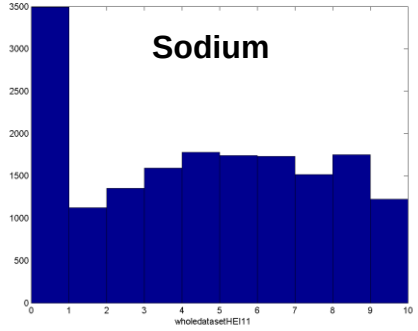


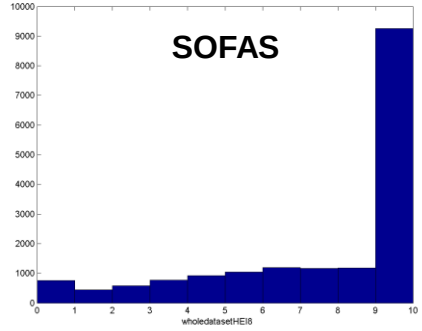












So, What Can I say About the Population as a Whole?

- *We consume lots of SOFAs*
- *We are not good at eating our vegetables*
- *We eat lots of grains …*
- *But not lots of whole grains*

- *But what does this say about someone like you?*

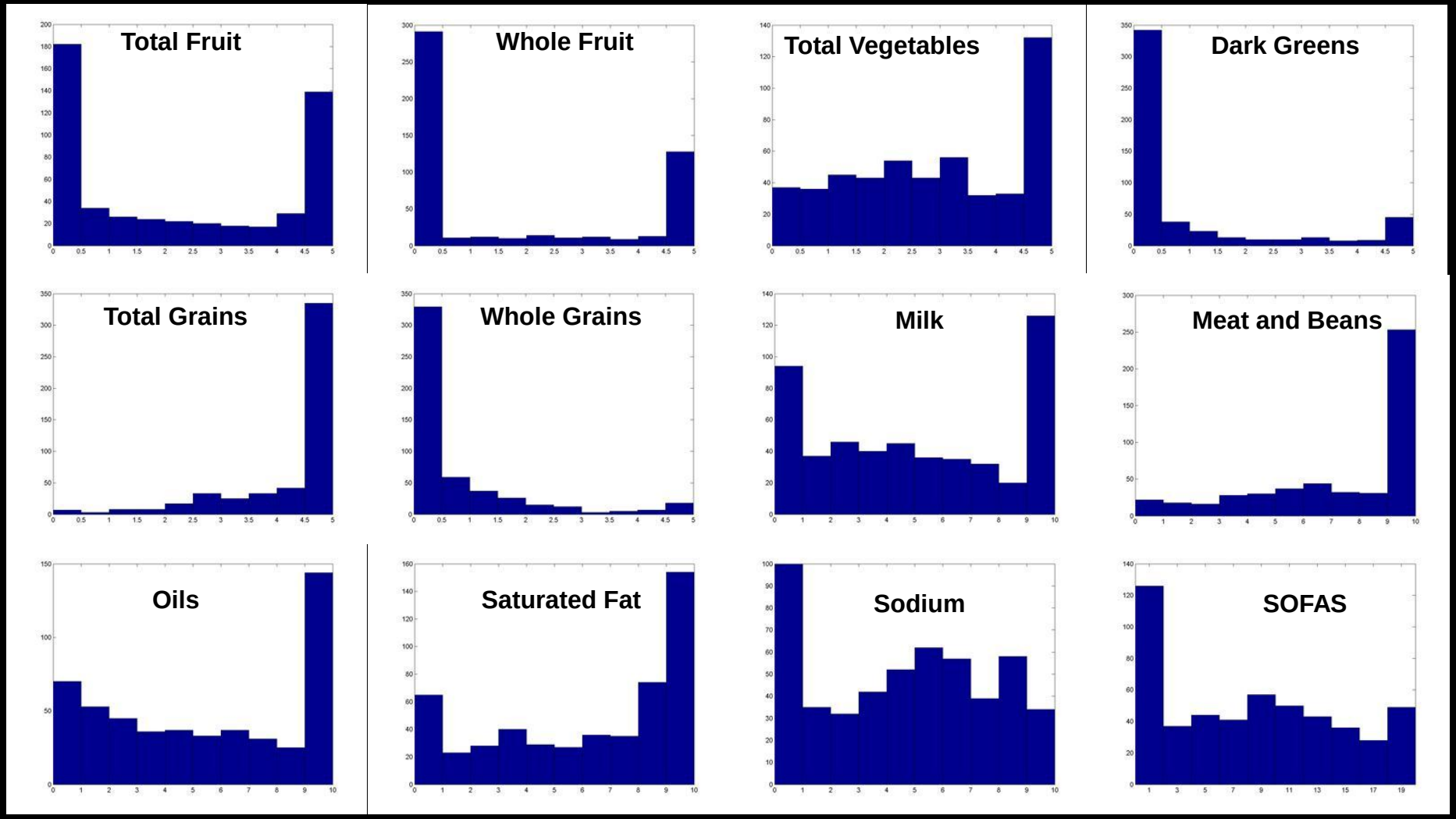
What if I further segment the data? Filter by Age, Gender, Etc.?

Let's look at Females, Age 19-30

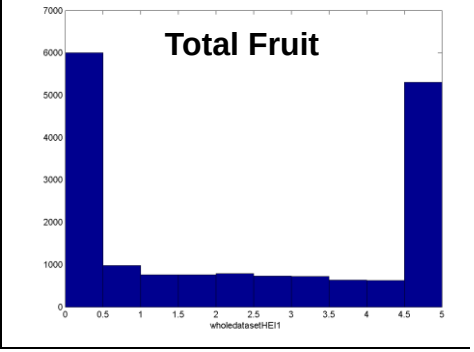
18.5 < BMI < 25.0

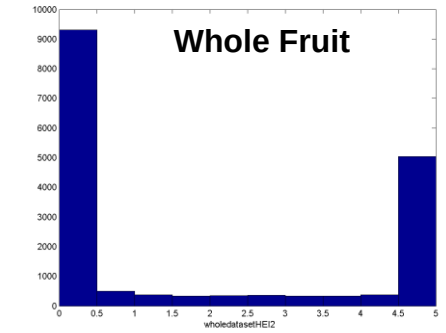
Total population in our sample = 511

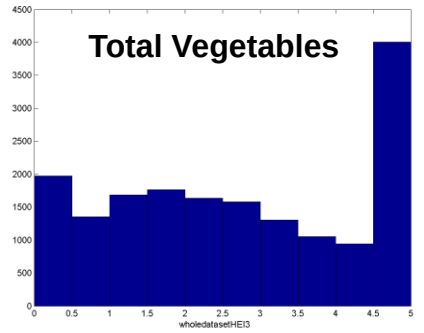
NHANES in Females Age 19-30, Average BMI

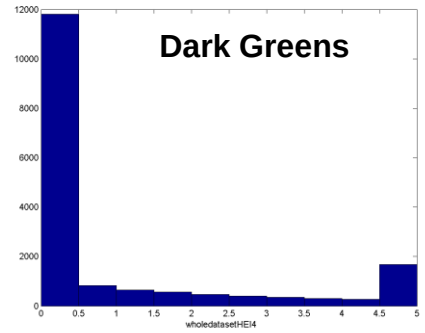


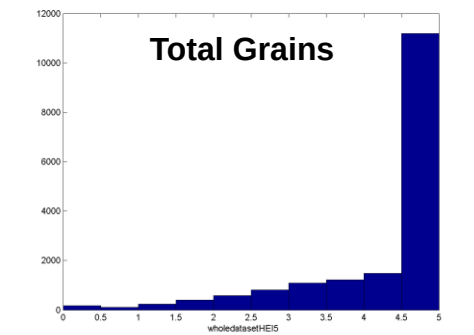
Back to the Entire Population!

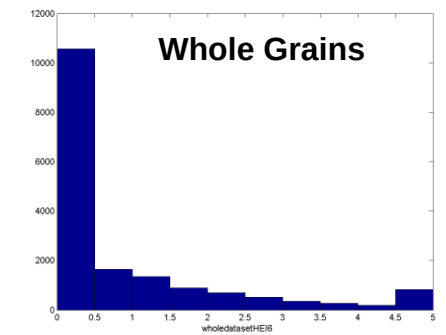


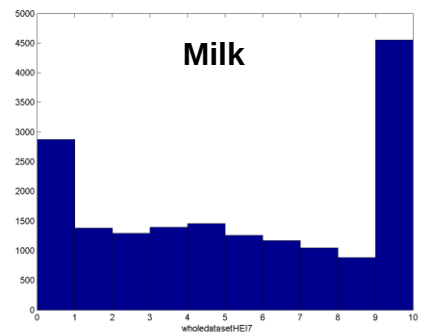


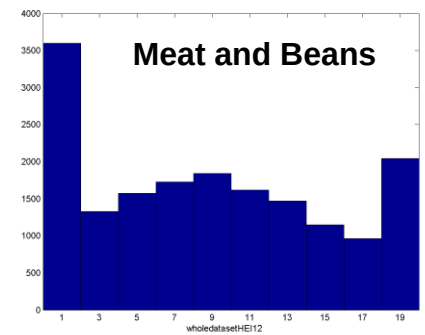


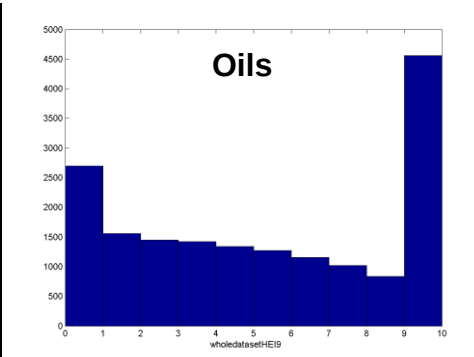


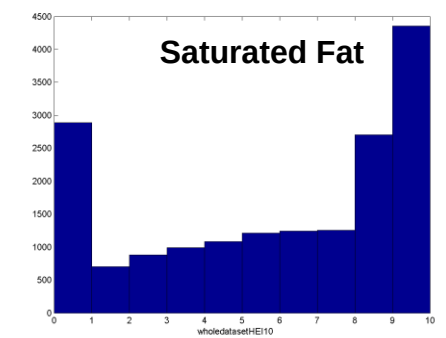


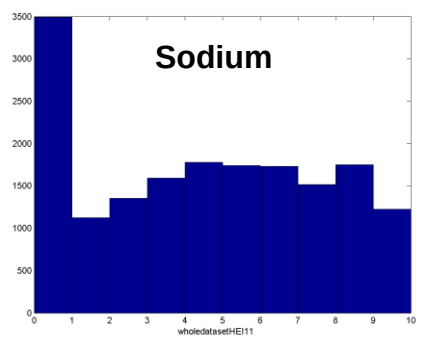


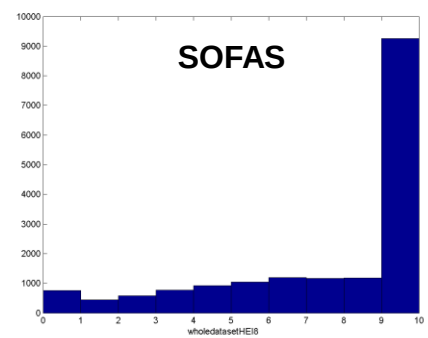












Forming Hypotheses

What do we think we see when comparing this group of females to the entire population?

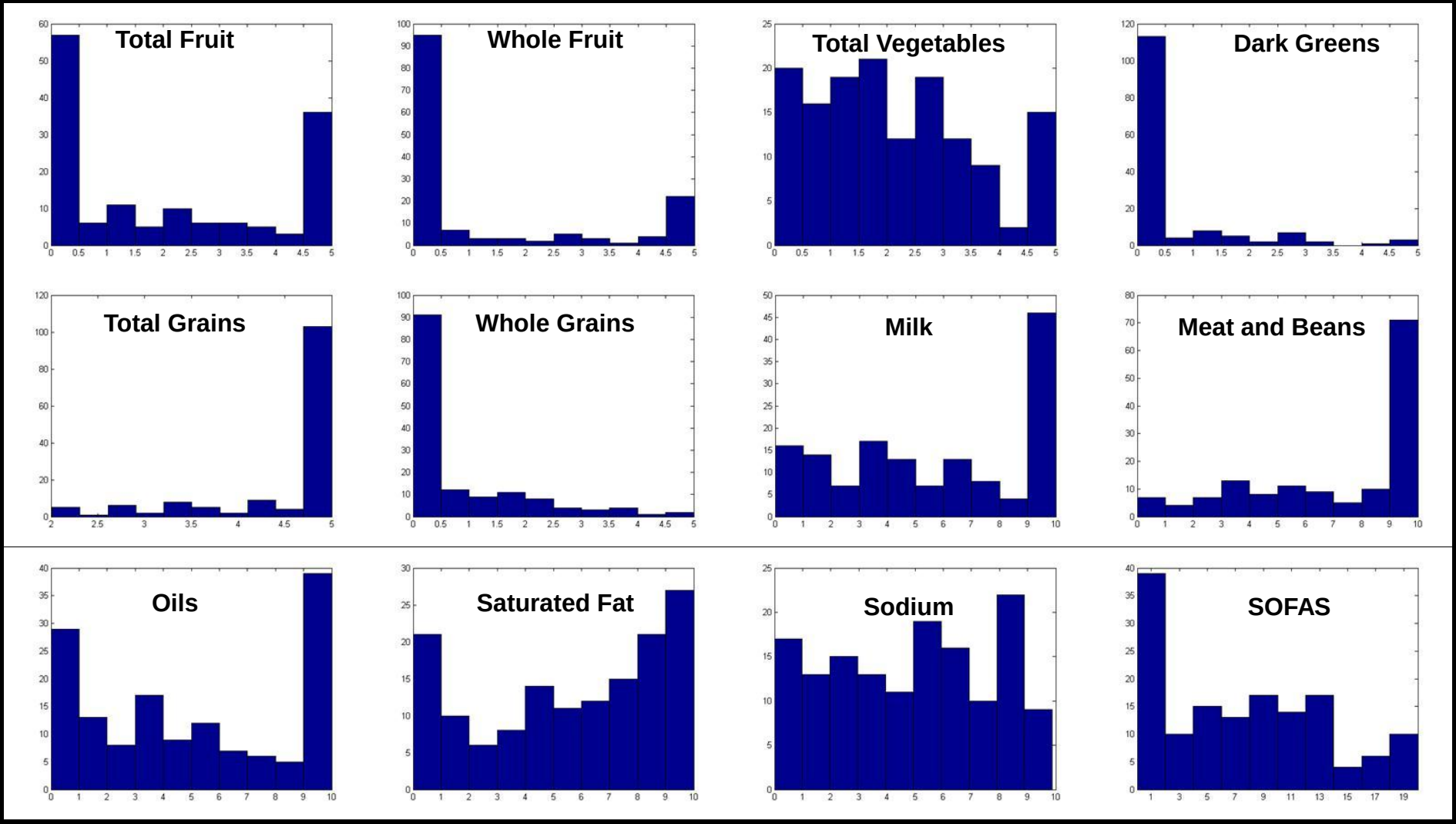
- ***Their SOFA consumption is more evenly distributed***
- ***Their meat and bean consumption is more skewed***

What if I further segment the data? Filter by Age, Gender, Etc.?

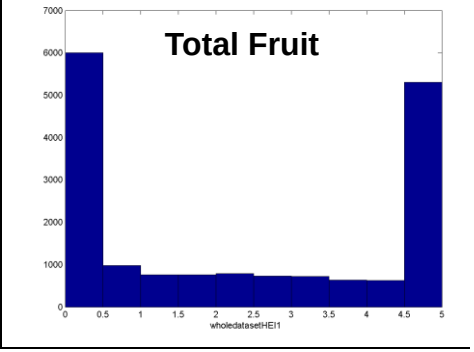
Let's look at Males, Age 14-18, BMI < 18.5

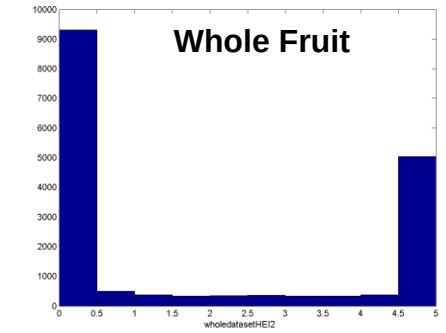
Total population in our sample = 145

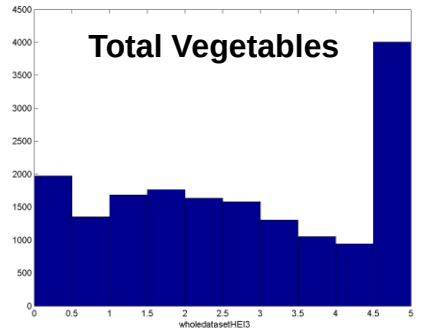
NHANES in Adolescent Underweight Males

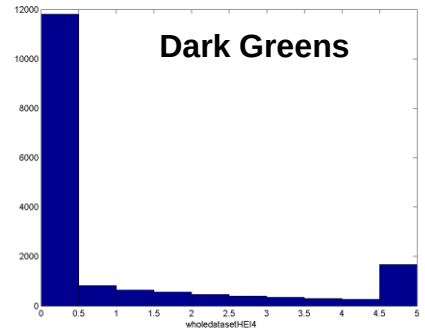


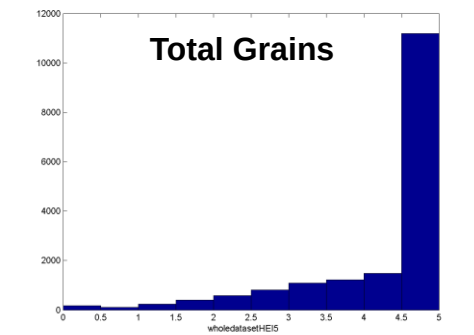
Back to the Entire Population!

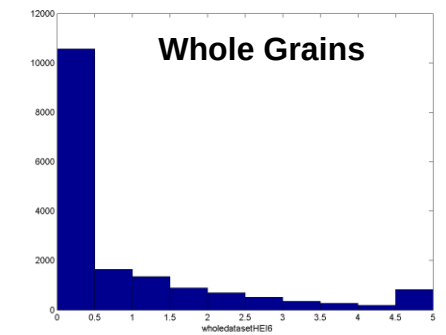


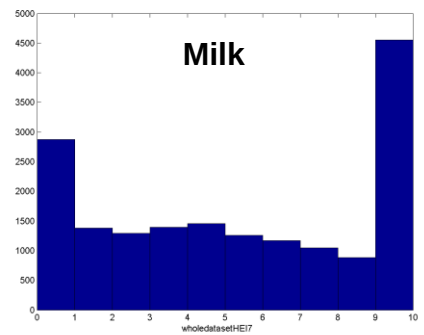


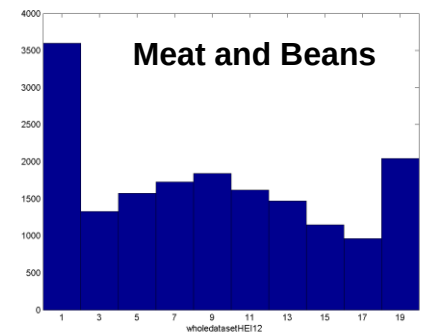


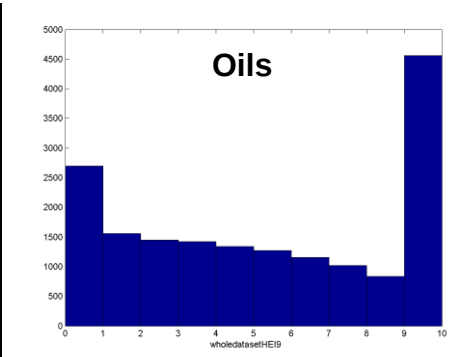


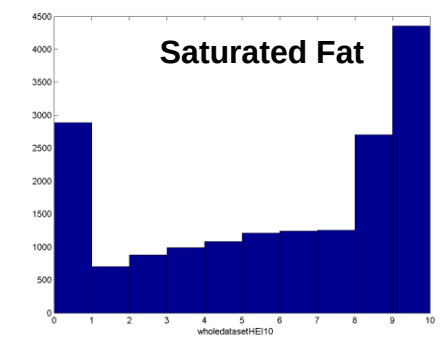


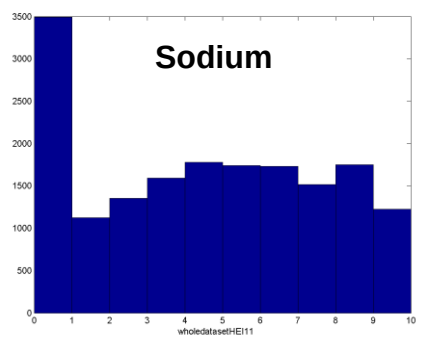


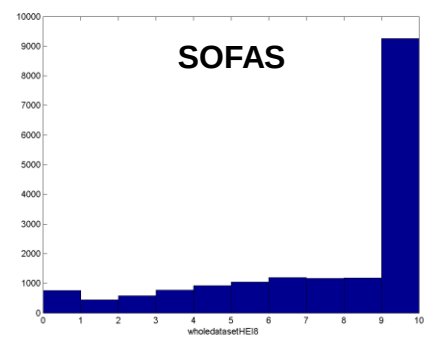












Forming Hypotheses

What do we think we see when comparing this group of females to the entire population?

- ***Their vegetable consumption is more evenly distributed***
- ***Their meat and bean consumption is more skewed***
- ***Their SOFAs are more evenly distributed***

What if I want to ask: How is … related to … ?

From our histograms, we came up with interesting insights into our data

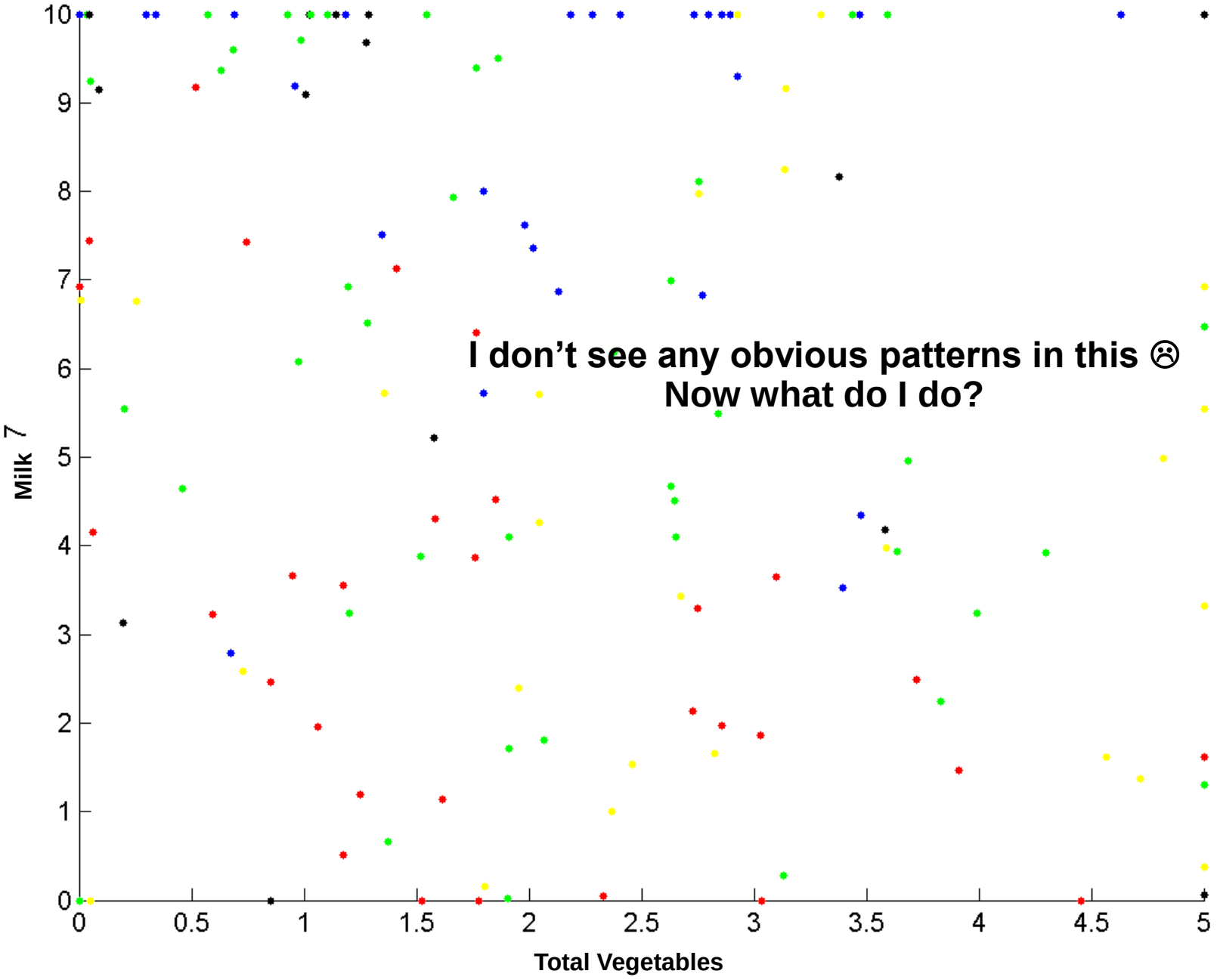
But what if I want to know about relationships between the data?



More like:
Do people that consume more milk consume more vegetables?

We can try a scatterplot!

Group: Underweight Adolescent Males



What if I further question the data? All relationships by Age, Gender?

How many of these did you really expect me to do?

These histograms were made in Excel

One can also use SAS, JMP, MatLab

In all of those programs we have to write scripts, modify variables, etc.

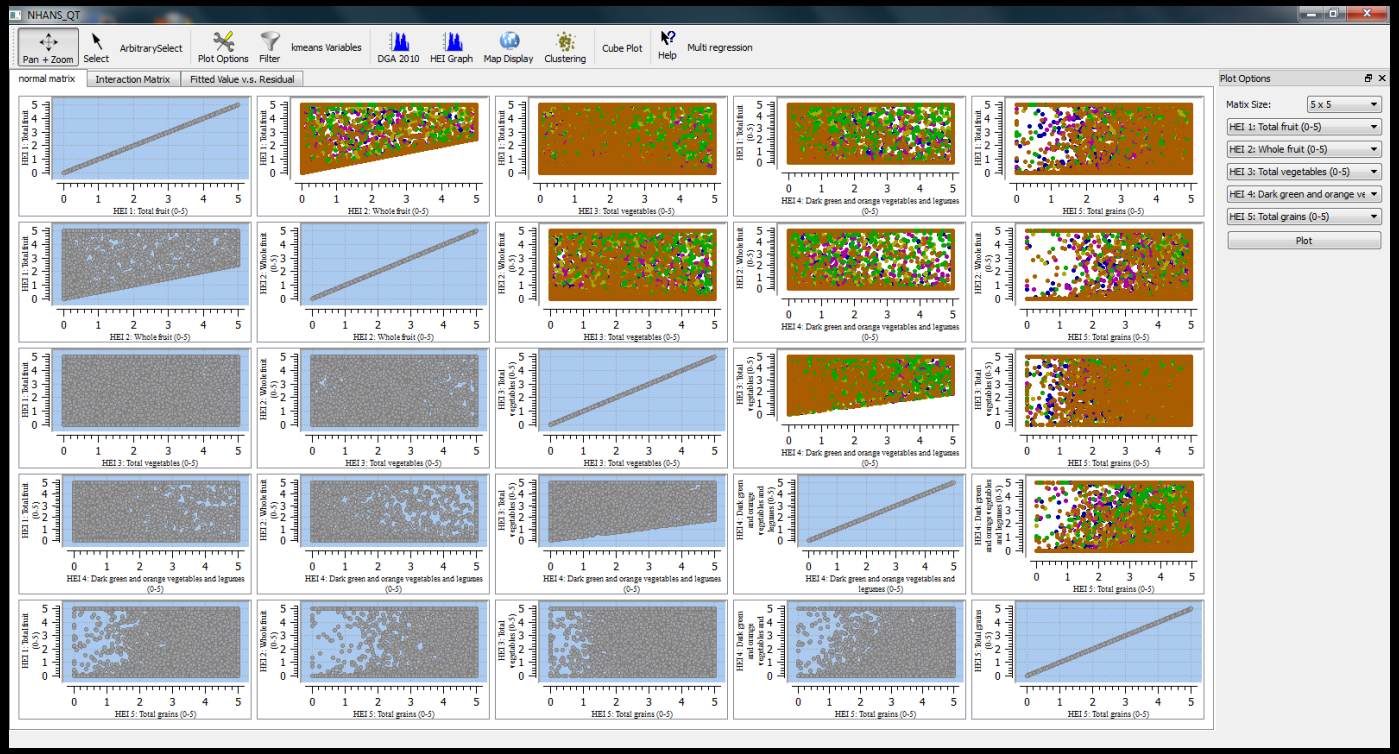
We can't directly manipulate our variables

So, being a data detective can be very tedious

And these were just simple comparisons!

So what do we do now?

We can utilize interactive graphics combined with underlying analytical processes to help us be better detectives!



The Curse of Dimensionality

A term coined by Bellman in 1961

Refers to the problems associated with multivariate data analysis as the dimensionality increases the available data becomes sparse

Sparsity is a problem for any method that requires statistical significance

Sometimes data dimensions are redundant and can be reduced with minimal information loss

In visualization we are also limited with screen space and the number of available visual variables, so choosing the most appropriate dimensions is key

Richard Ernest Bellman; Rand Corporation (1957). *Dynamic programming*. Princeton University Press.

The Curse of Dimensionality

So, what can we do?

- We can incorporate prior knowledge of the data
- We can smooth the target function
- We can reduce the dimensionality

The Practicality of the Curse of Dimensionality

For a given sample size, there is a maximum number of features above which the performance of classifying samples will degrade rather than improve

In most cases, the additional information that is lost when discarding some features is compensated by a more accurate mapping in the lower-dimensional space

So, how do we know what features we can throw away?

For visualization this implies that there are some features of a dataset that will be better to visualize (contain more information) than others!

Curses! I'm Still Overwhelmed!

Probably so, we are creating a lot of data in the world

Overall, NHANES is actually a relatively modest set of data

As our datasets get more variables we need ways to question the data

What I really want to know is are there relationships/groups/clusters in the data, how can I see these when we have so many dimensions?

One way is to just cluster the data and draw these clusters in an arbitrary space (we saw this in the demo)

Another is to try dimensional reduction!

Dimensional Reduction

Two approaches are available to reduce dimensionality

- **Feature extraction:** creating a subset of new features by combinations of the existing features
- **Feature Selection:** choosing a subset of all the features
- Given a feature space $x_i \in R^N$ find a mapping $y = f(x): R^N \rightarrow R^M$ with M<N such that the transformed feature vector $y \in R^M$ preserves (most of) the information or structure in R^N
- An optimal mapping is one that does not increase error

What Does Dimensional Reduction Mean For You?

Feature Selection: choosing a subset of all the features

- Well, that's what we were already doing
- We use our own expert knowledge to select features of interest
- Note we could also use machine learning and statistics here too if we were so inclined

What Does Dimensional Reduction Mean For You?

Feature extraction: creating a subset of new features by combinations of the existing features

- You can think of this in terms of the NHANES data
- Let's say I do an analysis and I find a correlation between SOFA consumption and sodium consumption (the more SOFAs, the more sodium)
- I could combine SOFAs and sodium into a new variable – SSOFAS
- I make up some formula to do this:
  - $SSOFAS = \alpha * SOFAs + \beta * sodium$
- So basically we just combine variables that are correlated

Principle Components Analysis

One of the most commonly applied dimension reduction techniques

PCA is a deterministic analytical procedure that utilizes an orthogonal transformation to reduce a set of sample observations with potentially correlated variables into a set of uncorrelated variables called principal components

The number of principal components will always be less than or equal to the original number of variables in the sample set

. T. Jolliffe, “Principal Component Analysis," *Springer Series in Statistics*, Springer, NY, 2002.

Principle Component Analysis

The main limitation of PCA is that it does not consider class separability since it does not take into account the class label of the feature vector

- PCA simple performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance
- There is no guarantee that the directions of maximum variance will contain the most interesting/important features

That sounded really hard!

- *Yes, but we have software to help us!*
- *In JMP Principal components can be accessed through the Principal Components command on the Analyze > Multivariate Methods menu.*

- *But, how does principal component analysis help me?*
- *What can principal component analysis tell me?*

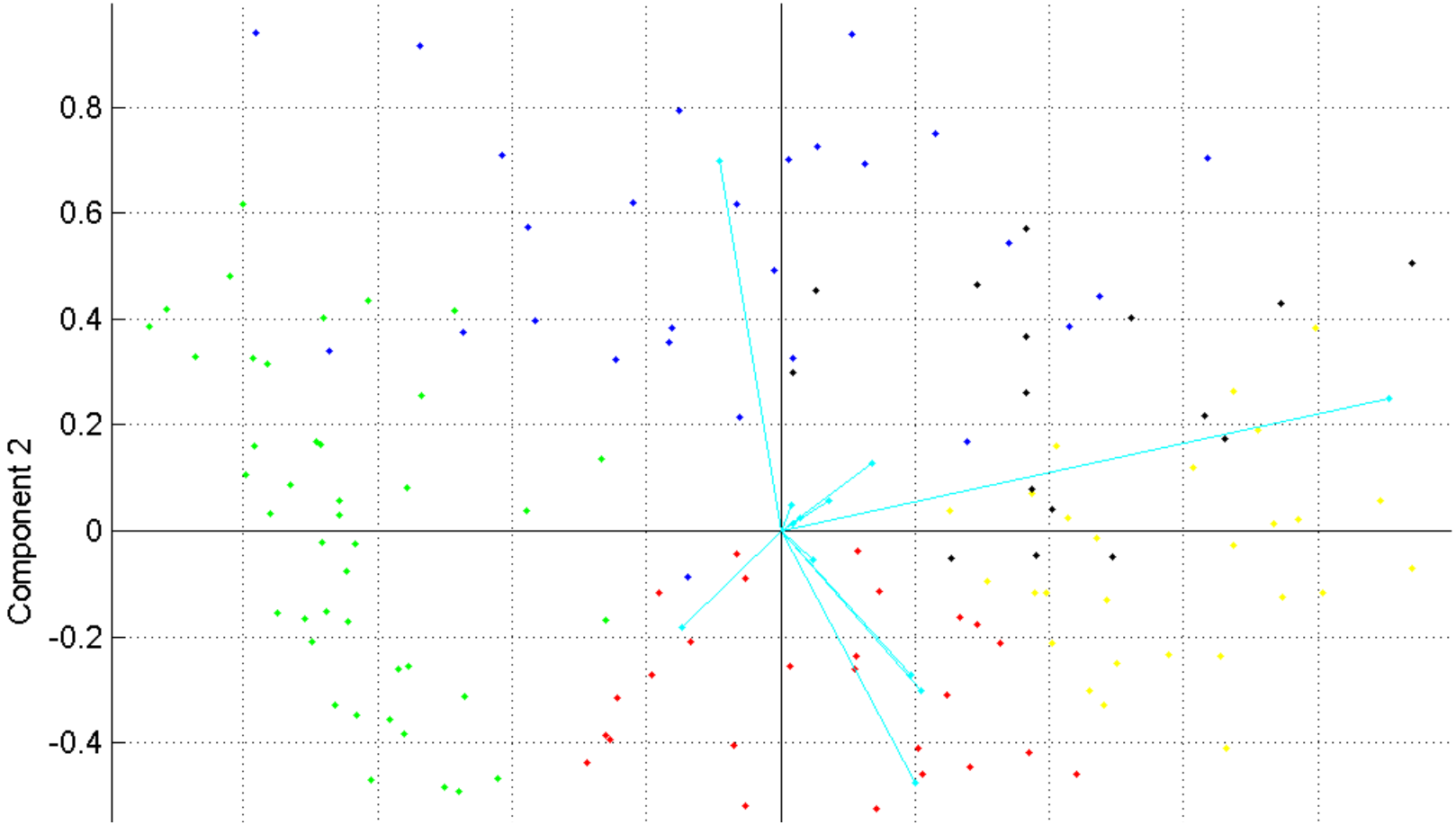
Principal Component Analysis

- *Let's go back to NHANES for an example and focus on the HEI values.*
- *Each person in my data set has 12 HEI values, so you can think of each person as a point in 12-dimensional space*
- *It is hard to understand all the relationships in this space, PCA tries to figure out what dimensions to combine for you automatically*
- *After you run PCA you will get a set of what you can think of as new variables*
- *Since there were 12 variables in our NHANES data, we will get at most 12 principal components (but hopefully less!)*

Principal Component Analysis

- *Ok, you've convinced me to run PCA on NHANES and I get 8 components, but now I'm confused!*
- *The output is such that your first component accounts for as much of the variability in the data set as possible*
- *The second component accounts for the second most variability, etc.*
- *Typically we will plot the data using the first two principal components as axes*
- *We can also write out each component to analyze it*

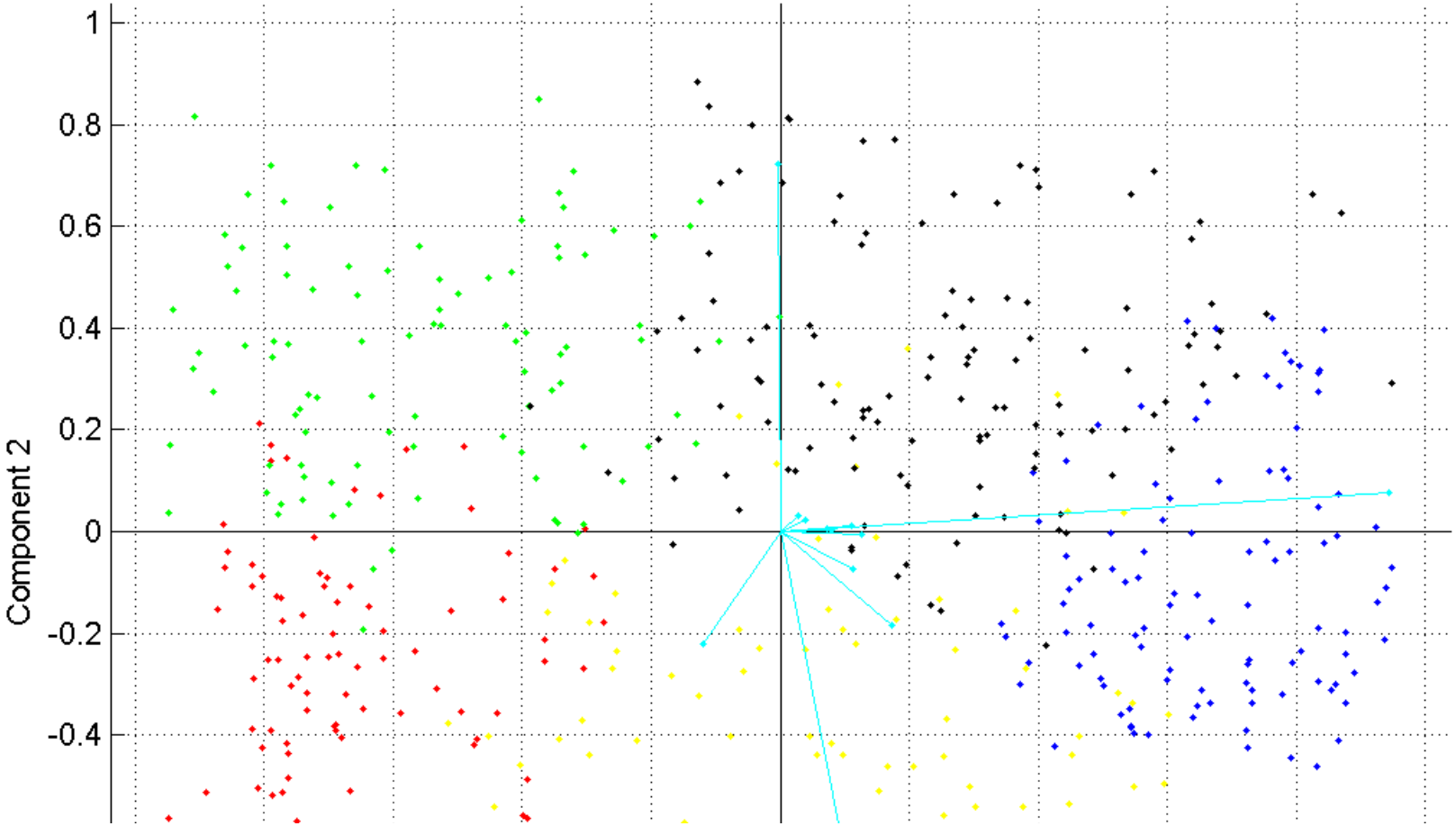
Underweight Adolescent Males



FV1 ≈ 0.14 HEI1 + 0.07 HEI2 +0.02 HEI5 + 0.02 HEI6 + -0.09 HEI7 + 0.19 HEI8 +0.05 HEI3 + 0.03 HEI4 + 0.20 HEI9 + 0.20 HEI10 + -0.15 HEI11 + 0.91 HEI12

FV2 ≈ 0.13 HEI1 + 0.06 HEI2 + 0.05 HEI5 + 0.01 HEI6 + 0.70 HEI7 + -0.27 HEI8 + -0.05 HEI3 + 0.02 HEI4 + -0.48 HEI9 + -0.30 HEI10 + -0.18 HEI11 + 0.25 HEI12

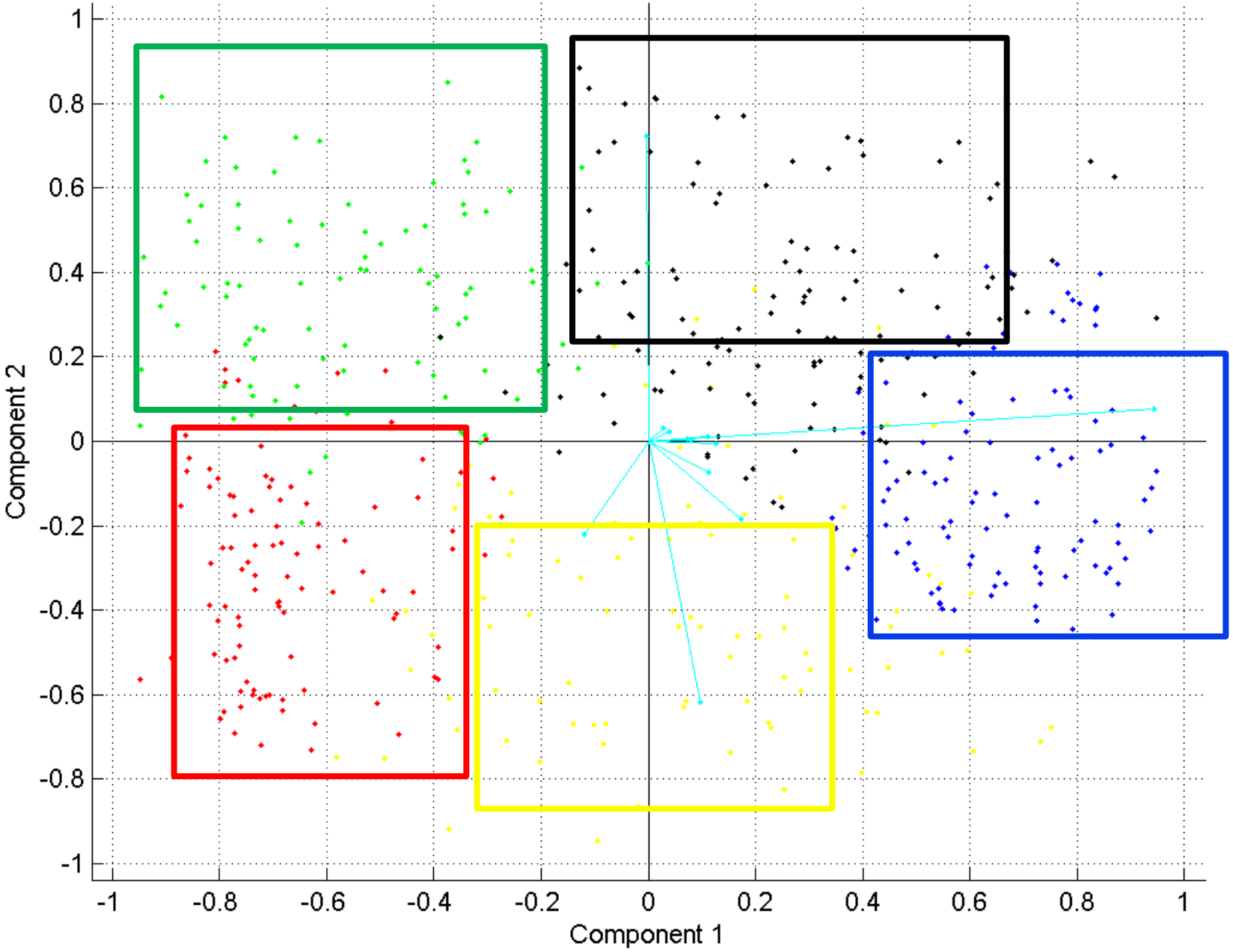
Females Age 19-30, Average BMI



FV1 ≈ 0.13 HEI1 +0.11 HEI2 +0.04 HEI5 + 0.03 HEI6 + -0.00 HEI7 + 0.11 HEI8 +0.08 HEI3 + 0.07 HEI4 + 0.17 HEI9 + 0.10 HEI10 +-0.12 HEI11 + 0.94 HEI12

FV2 ≈-0.01 HEI1 + 0.01 HEI2 +0.02 HEI5 + 0.03 HEI6 + 0.72 HEI7 + -0.07 HEI8 + 0.00 HEI3 + 0.00 HEI4 + -0.18 HEI9 +-0.62 HEI10 +-0.22 HEI11 +0.08 HEI12

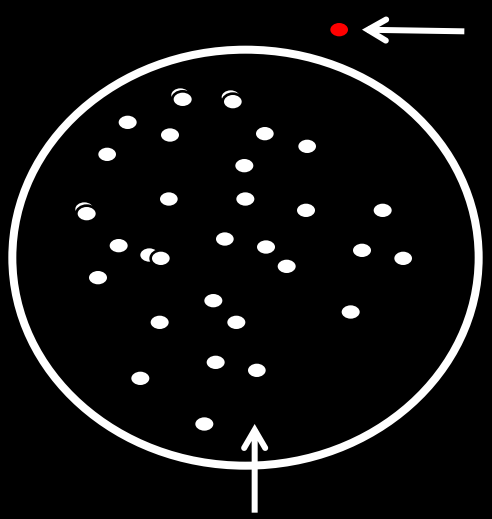
Females Age 19-30, Average BMI



Putting it all together

- ***If I know things about you …***
  - Age, gender, BMI
- ***Can I guess your typical daily diet?***
  - Probably not very accurately
- ***But, if I ask you a few questions, maybe I can get closer!***

Putting it all together



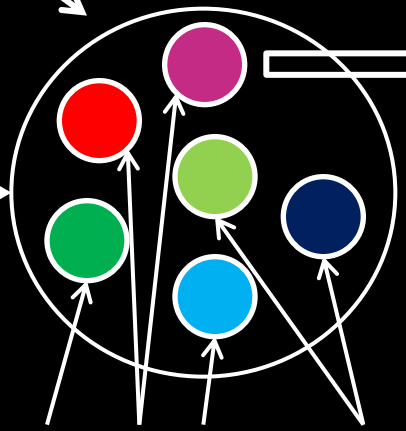
This is you.
Maybe you tell me your age, gender, BMI?
(Inform me maybe?)

This is the NHANES data
I know everyone's diet

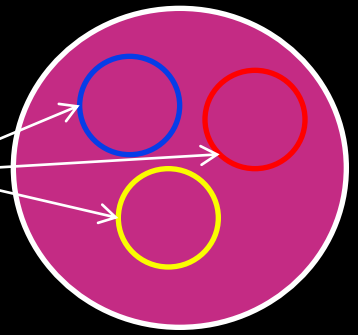
I can first split the data by demographics

I can stick you in one of these groups based on your demographics

Total Population



Split by demographics
(Learned or *a prior*)



Now my demographics groups having groups based on diet!

Now I know your diet!
(Well sort of)

This is still you.
I know you're in this large group
But I want to know your subgroup.

If I know how many sodas you drink, maybe I can narrow it down from three groups to two

Maybe if you ate a salad today tells me more!

?

?

Summary

- ***Data analysis is hard!***
- ***We need better tools to be data detectives***
- ***By being data detectives, we can ask questions and maybe create predictions about unknown data based on available data (we make a model of the data)***
- ***Visual analytics helps us in the data detective work!***
- ***But we still need to understand what the analyses do for us so that we can put them into practice.***

Dr. Ross's Shameless Plug

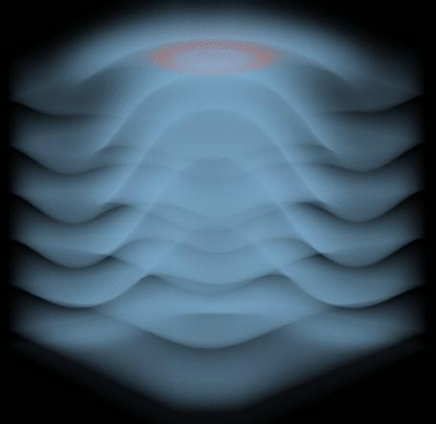
- ***You can play some of my non-VA apps***
  - Search for hedcutr on Facebook and you can make this:

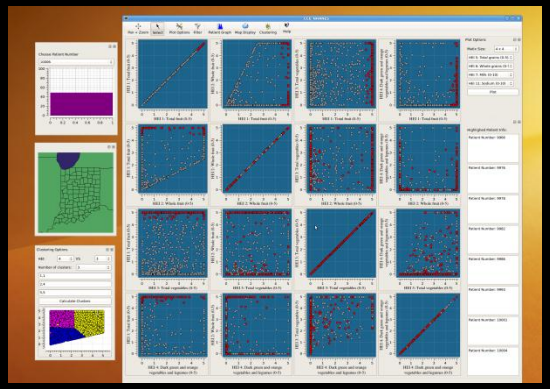


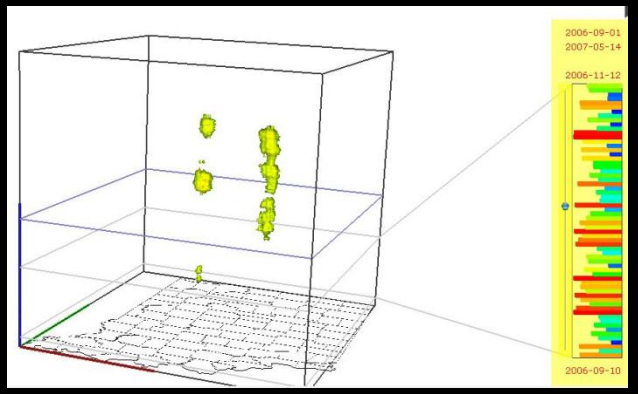


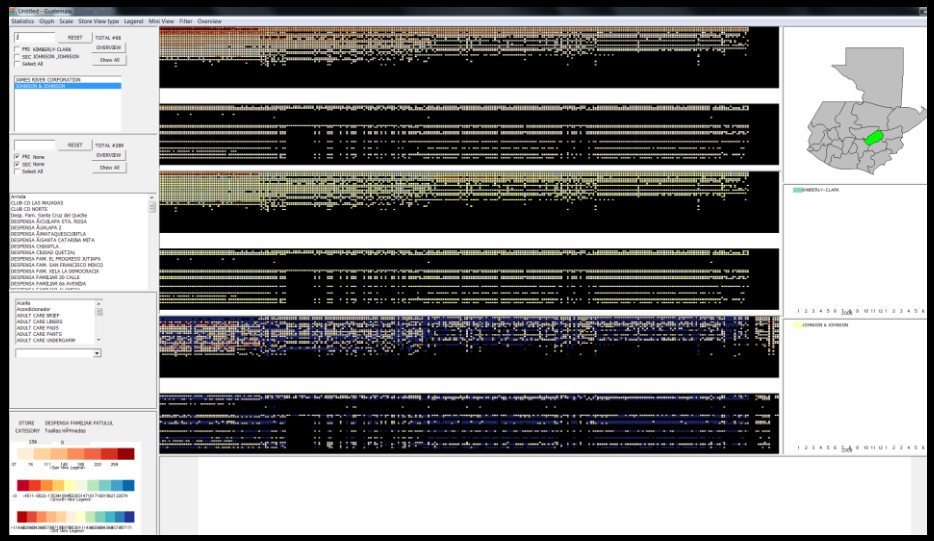
  - I'm collecting data on the rendering parameters to see what people like!

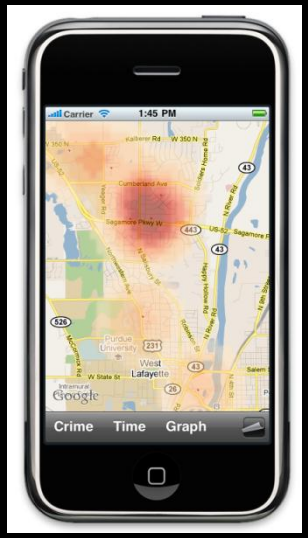
Questions? (http://rmaciejewski.faculty.asu.edu)













VAST · INFOVIS · SCIVIS
BIOVIS · LDAV

VIS 2013

Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw

Carsten Görg, Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo,


University of Colorado


Adobe


Cornell University


Georgia Tech

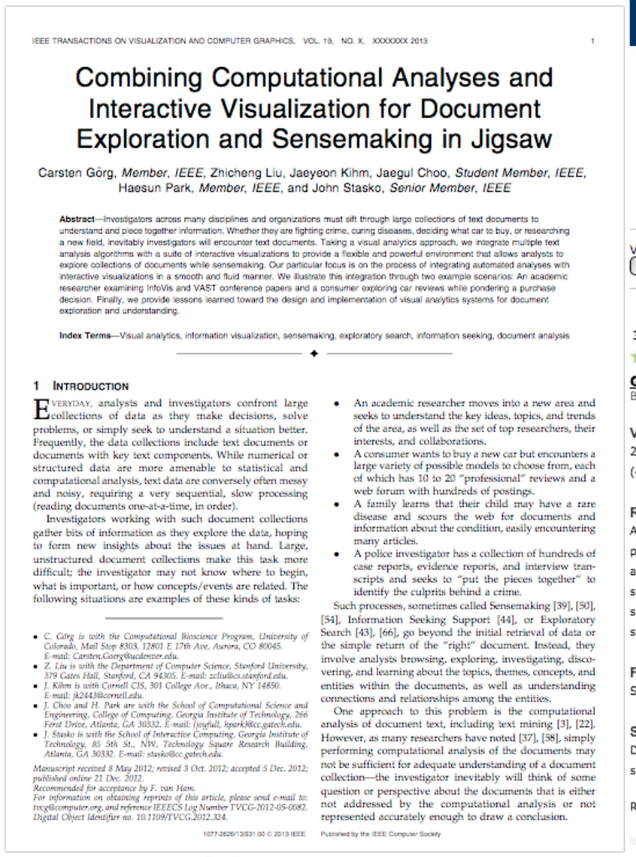
Haesun Park, and John Stasko


Georgia Tech


Georgia Tech

Text documents are everywhere ...

Academic Papers



IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 19, NO. X, XXXXXXX 2013 1

Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw

Carsten Görg, *Member, IEEE*, Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo, *Student Member, IEEE*, Haesun Park, *Member, IEEE*, and John Stasko, *Senior Member, IEEE*

Abstract—Investigators across many disciplines and organizations must sift through large collections of text documents to understand and piece together information. Whether they are fighting crime, curing diseases, deciding what car to buy, or researching a new field, inevitably investigators will encounter text documents. Taking a visual analytics approach, we integrate multiple text analysis algorithms with a suite of interactive visualizations to provide a flexible and powerful environment that allows analysts to explore collections of documents while sensemaking. Our particular focus is on the process of integrating automated analyses with interactive visualizations in a smooth and fluid manner. We illustrate this integration through two example scenarios: An academic researcher examining InfoVis and VAST conference papers and a consumer exploring car reviews while pondering a purchase decision. Finally, we provide lessons learned toward the design and implementation of visual analytics systems for document exploration and understanding.

Index Terms—Visual analytics, information visualization, sensemaking, exploratory search, information seeking, document analysis

1 Introduction

Everyday, analysts and investigators confront large collections of data as they make decisions, solve problems, or simply seek to understand a situation better. Frequently, the data collections include text documents or documents with key text components. While numerical or structured data are more amenable to statistical and computational analysis, text data are conversely often messy and noisy, requiring a very sequential, slow processing (reading documents one-at-a-time, in order).

Investigators working with such document collections gather bits of information as they explore the data, hoping to form new insights about the issues at hand. Large, unstructured document collections make this task more difficult; the investigator may not know where to begin, what is important, or how concepts/events are related. The following situations are examples of these kinds of tasks:

- An academic researcher moves into a new area and seeks to understand the key ideas, topics, and trends of the area, as well as the set of top researchers, their interests, and collaborations.
- A consumer wants to buy a new car but encounters a large variety of possible models to choose from, each of which has 10 to 20 "professional" reviews and a web forum with hundreds of postings.
- A family learns that their child may have a rare disease and scours the web for documents and information about the condition, easily encountering many articles.
- A police investigator has a collection of hundreds of case reports, evidence reports, and interview transcripts and seeks to "put the pieces together" to identify the culprits behind a crime.

Such processes, sometimes called Sensemaking [39], [50], [54], Information Seeking Support [44], or Exploratory Search [43], [66], go beyond the initial retrieval of data or the simple return of the "right" document. Instead, they involve analysts browsing, exploring, investigating, discovering, and learning about the topics, themes, concepts, and entities within the documents, as well as understanding connections and relationships among the entities.

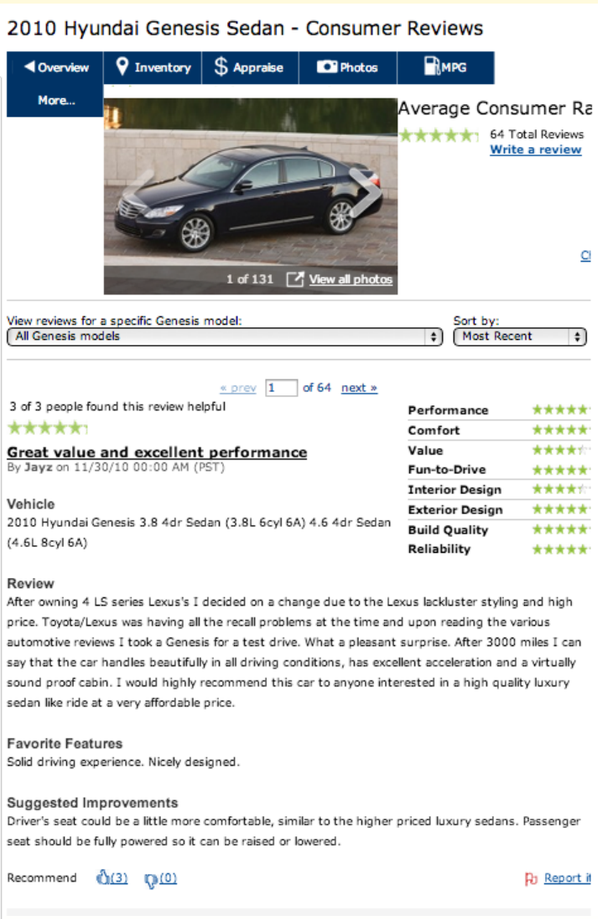
One approach to this problem is the computational analysis of document text, including text mining [3], [22]. However, as many researchers have noted [37], [58], simply performing computational analysis of the documents may not be sufficient for adequate understanding of a document collection—the investigator inevitably will think of some question or perspective about the documents that is either not addressed by the computational analysis or not represented accurately enough to draw a conclusion.

• C. Görg is with the Computational Bioscience Program, University of Colorado, Mail Stop 8303, 12801 E 17th Ave, Aurora, CO 80045. E-mail: Carsten.Goerg@ucdenver.edu.
• Z. Liu is with the Department of Computer Science, Stanford University, 379 Gates Hall, Stanford, CA 94305. E-mail: zcliu@cs.stanford.edu.
• J. Kihm is with Cornell CIS, 301 College Ave., Ithaca, NY 14850. E-mail: jk244@cornell.edu.
• J. Choo and H. Park are with the School of Computational Science and Engineering, College of Computing, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30332. E-mail: {joegfull, hpark}@cc.gatech.edu.
• J. Stasko is with the School of Interactive Computing, Georgia Institute of Technology, 85 5th St., NW, Technology Square Research Building, Atlanta, GA 30332. E-mail: stasko@cc.gatech.edu.

Manuscript received 8 May 2012; revised 3 Oct. 2012; accepted 5 Dec. 2012; published online 21 Dec. 2012.
Recommended for acceptance by F. van Ham.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-2012-05-0082.
Digital Object Identifier no. 10.1109/TVCG.2012.324.

1077-2626/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

Product Reviews



2010 Hyundai Genesis Sedan - Consumer Reviews

Overview | Inventory | Appraise | Photos | MPG | More...

Average Consumer Ra

64 Total Reviews
Write a review

1 of 131 View all photos

View reviews for a specific Genesis model: All Genesis models

Sort by: Most Recent

« prev 1 of 64 next »

3 of 3 people found this review helpful

Great value and excellent performance

By Jayz on 11/30/10 00:00 AM (PST)

Performance	★★★★★
Comfort	★★★★★
Value	★★★★★
Fun-to-Drive	★★★★★
Interior Design	★★★★★
Exterior Design	★★★★★
Build Quality	★★★★★
Reliability	★★★★★

Vehicle

2010 Hyundai Genesis 3.8 4dr Sedan (3.8L 6cyl 6A) 4.6 4dr Sedan (4.6L 8cyl 6A)

Review

After owning 4 LS series Lexus's I decided on a change due to the Lexus lackluster styling and high price. Toyota/Lexus was having all the recall problems at the time and upon reading the various automotive reviews I took a Genesis for a test drive. What a pleasant surprise. After 3000 miles I can say that the car handles beautifully in all driving conditions, has excellent acceleration and a virtually sound proof cabin. I would highly recommend this car to anyone interested in a high quality luxury sedan like ride at a very affordable price.

Favorite Features

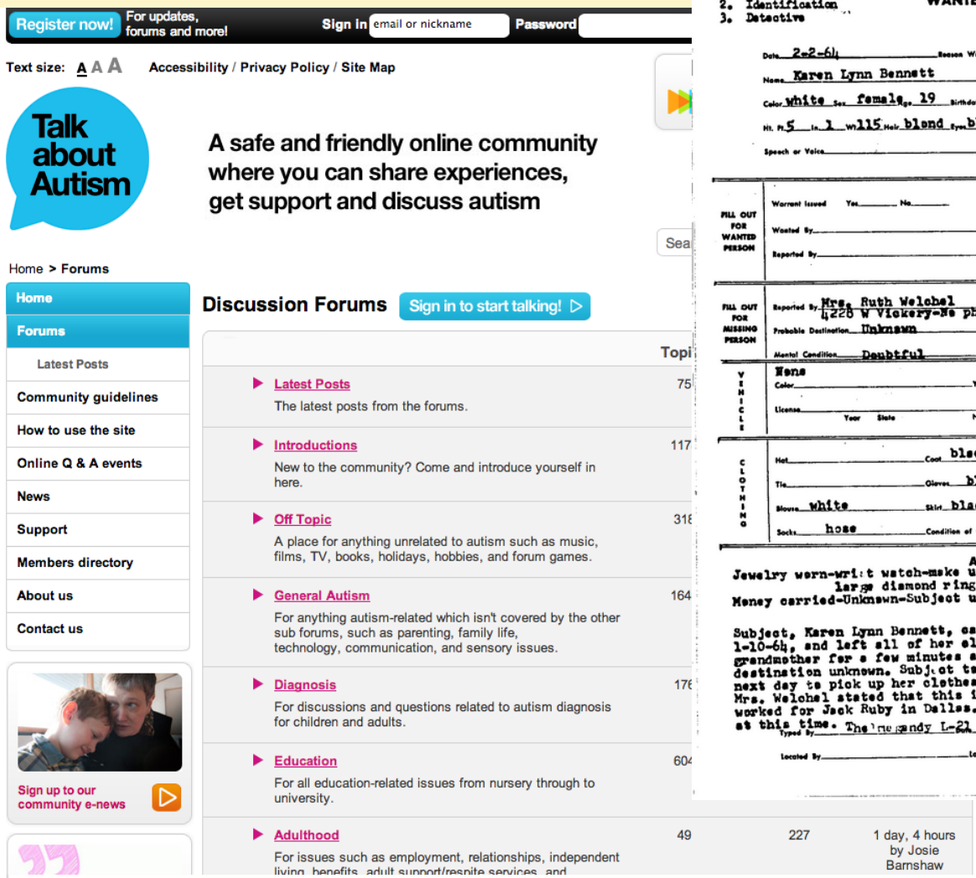
Solid driving experience. Nicely designed.

Suggested Improvements

Driver's seat could be a little more comfortable, similar to the higher priced luxury sedans. Passenger seat should be fully powered so it can be raised or lowered.

Recommend (3) (0) Report it

Health Forums



Register now! For updates, forums and more! Sign in email or nickname Password

Text size: A A A Accessibility / Privacy Policy / Site Map

Talk about Autism

A safe and friendly online community where you can share experiences, get support and discuss autism

Home > Forums

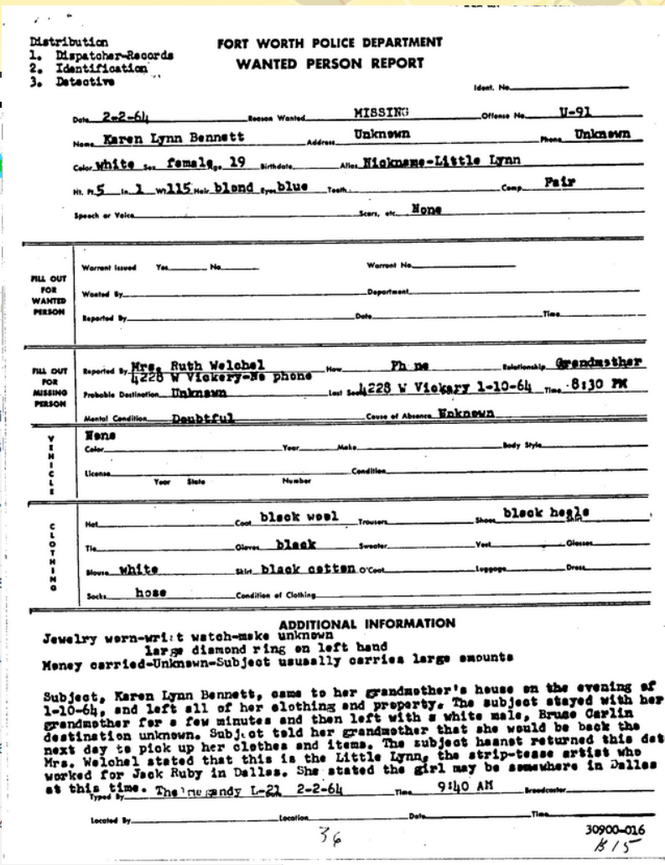
- Home
- Forums
- Latest Posts
- Community guidelines
- How to use the site
- Online Q & A events
- News
- Support
- Members directory
- About us
- Contact us

Sign up to our community e-news

Discussion Forums Sign in to start talking!

Latest Posts The latest posts from the forums.	75		
Introductions New to the community? Come and introduce yourself in here.	117		
Off Topic A place for anything unrelated to autism such as music, films, TV, books, holidays, hobbies, and forum games.	318		
General Autism For anything autism-related which isn't covered by the other sub forums, such as parenting, family life, technology, communication, and sensory issues.	164		
Diagnosis For discussions and questions related to autism diagnosis for children and adults.	176		
Education For all education-related issues from nursery through to university.	604		
Adulthood For issues such as employment, relationships, independent	49	227	1 day, 4 hours by Josie Barnshaw

Police Reports



Distribution
1. Dispatcher-Records
2. Identification
3. Detective

FORT WORTH POLICE DEPARTMENT
WANTED PERSON REPORT

Ident. No.

Date 2-2-64 Reason Wanted MISSING Offense No. U-91
Name Karen Lynn Bennett Address Unknown Phone Unknown
Color White Sex female Age 19 Birthdate Alias Nickname-Little Lynn
Ht. ft. 5 in. 1 Wt. 115 Hair blond Eyes blue Teeth Comp. Fair
Speech or Voice Scars, etc. None

FILL OUT FOR WANTED PERSON
Warrant Issued Yes No Warrant No.
Wanted By Department
Reported By Date Time

FILL OUT FOR MISSING PERSON
Reported By Mrs. Ruth Welchel How Ph ne Relationship Grandmother
4228 W Vickery-Ne phone
Probable Destination Unknown Last Seen 4228 W Vickery 1-10-64 Time 8:30 PM
Mental Condition Doubtful Cause of Absence Unknown

VEHICLE
None
Color Year Make Body Style
License Year State Number Condition

CLOTHING
Hat Coat black wool Trousers Shoes black heels
Tie Gloves black Sweater Vest Glasses
Blouse white Skirt black cotton Coat Luggage Dress
Socks hose Condition of Clothing

ADDITIONAL INFORMATION
Jewelry worn-wrist watch-make unknown
large diamond ring on left hand
Money carried-Unknown-Subject usually carries large amounts

Subject, Karen Lynn Bennett, came to her grandmother's house on the evening of 1-10-64, and left all of her clothing and property. The subject stayed with her grandmother for a few minutes and then left with a white male, Bruce Carlin destination unknown. Subject told her grandmother that she would be back the next day to pick up her clothes and items. The subject hasnot returned this dat Mrs. Welchel stated that this is the Little Lynn, the strip-tease artist who worked for Jack Ruby in Dallas. She stated the girl may be somewhere in Dallas at this time. The friendly L-21 2-2-64 Time 9:40 AM
Typed By
Located By Location Date Time

30900-016
8/15

Investigative Analysis

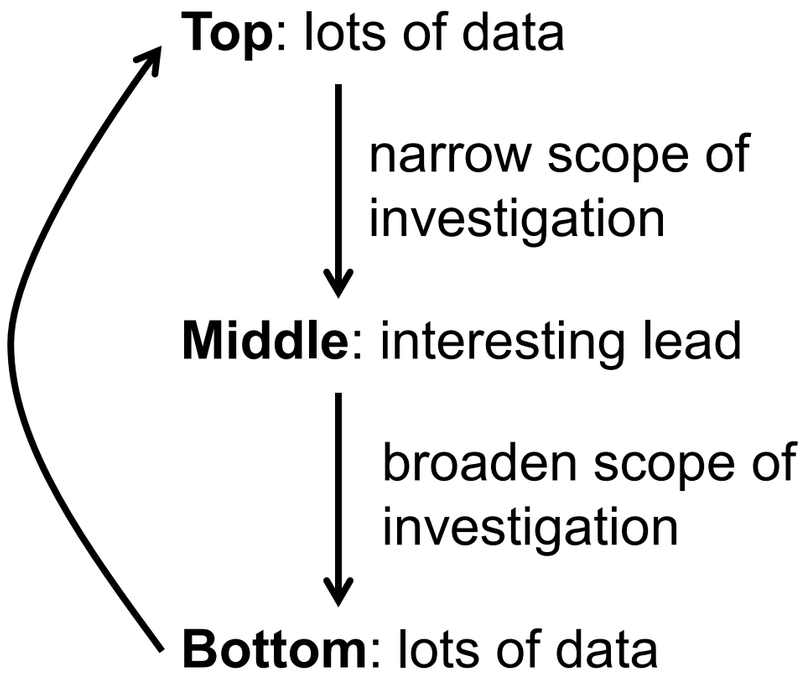
- Assist analysts (police/intelligence, lawyers, reporters, scientists) confronting “information overload” problems while doing exploratory data analysis

- Given a collection of textual documents, assist analysts in
  1. Information foraging
  2. Sense-making

Two Types of Scenarios

1. Targeted analysis scenarios
2. Open ended, strategic analysis scenarios

Hourglass-shaped investigations





VIS 2013 VAST • INFOVIS • SCIVIS BIOVIS • LDAV

Text Analytic Approaches

Text Mining &
Computational Analysis

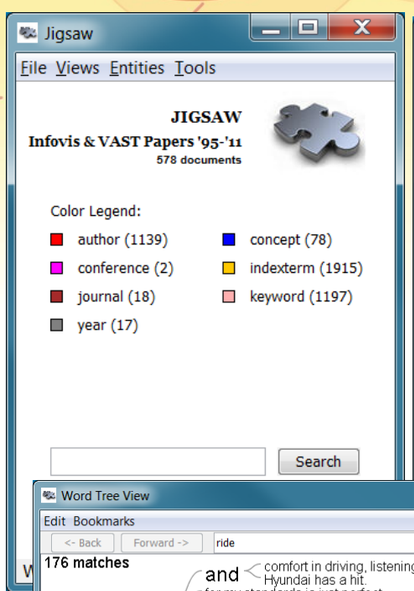
Interactive Text Visualization

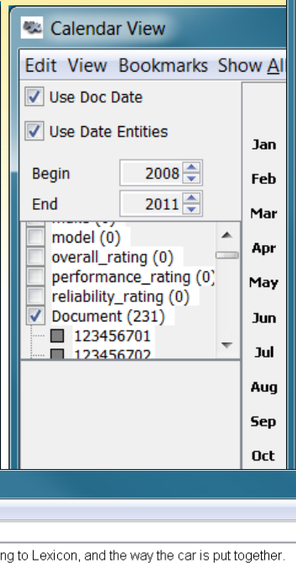
contribution
of this paper

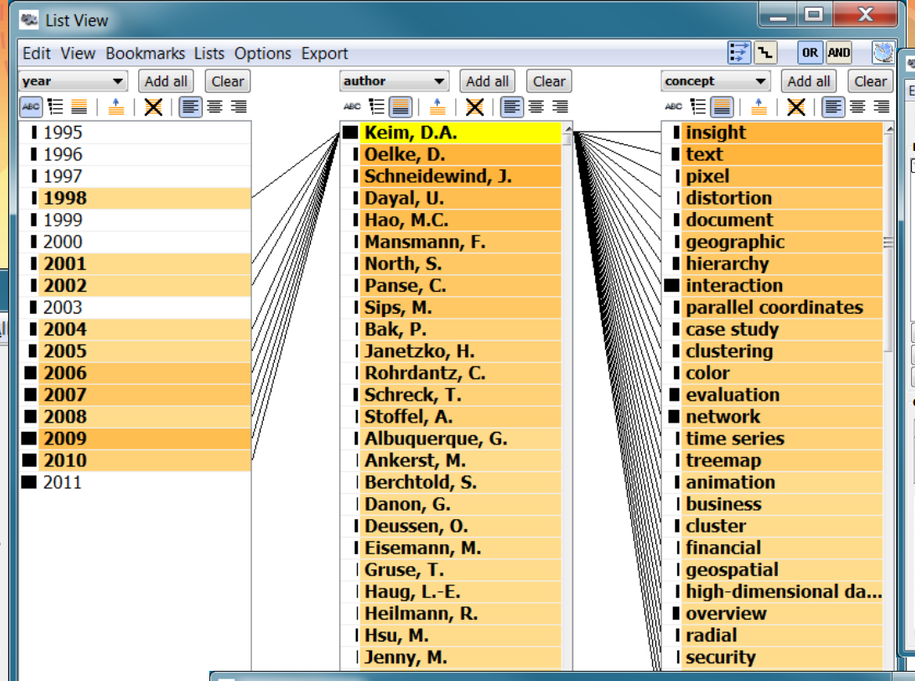
Integration

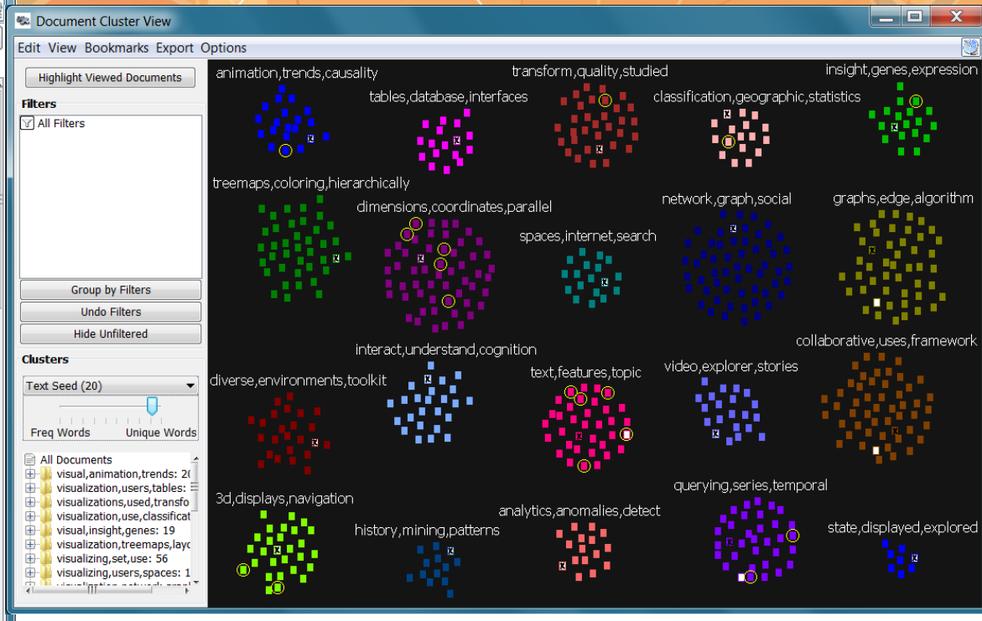
Visual Text Analytics

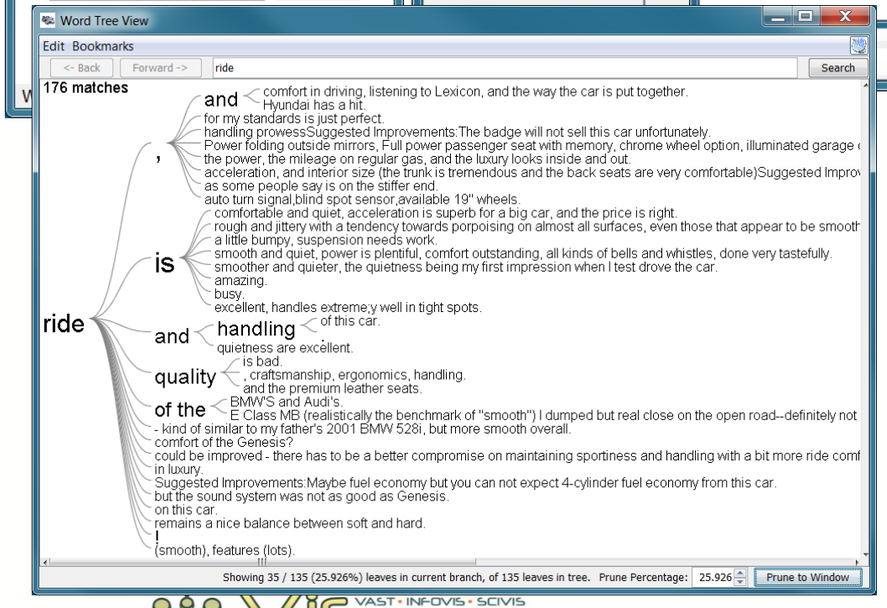
Jigsaw

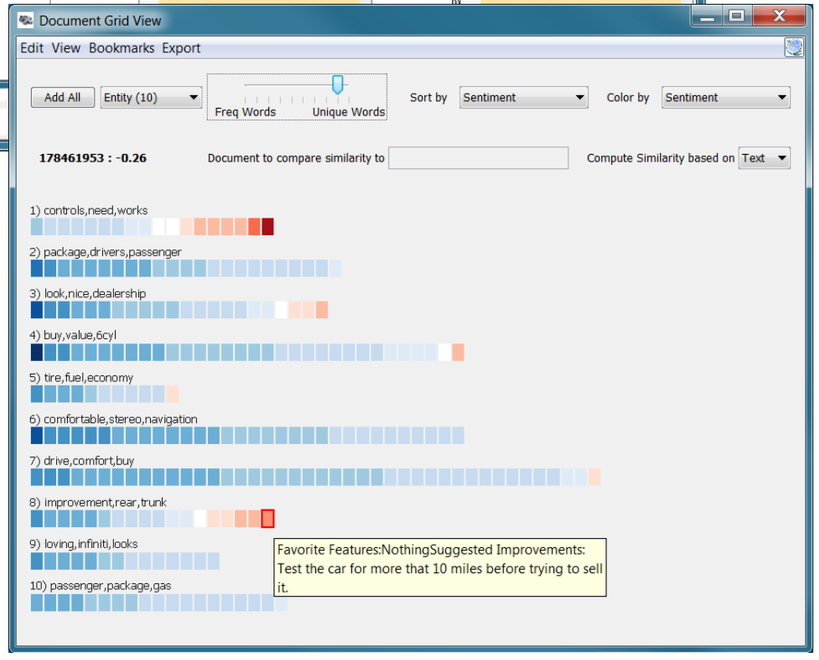


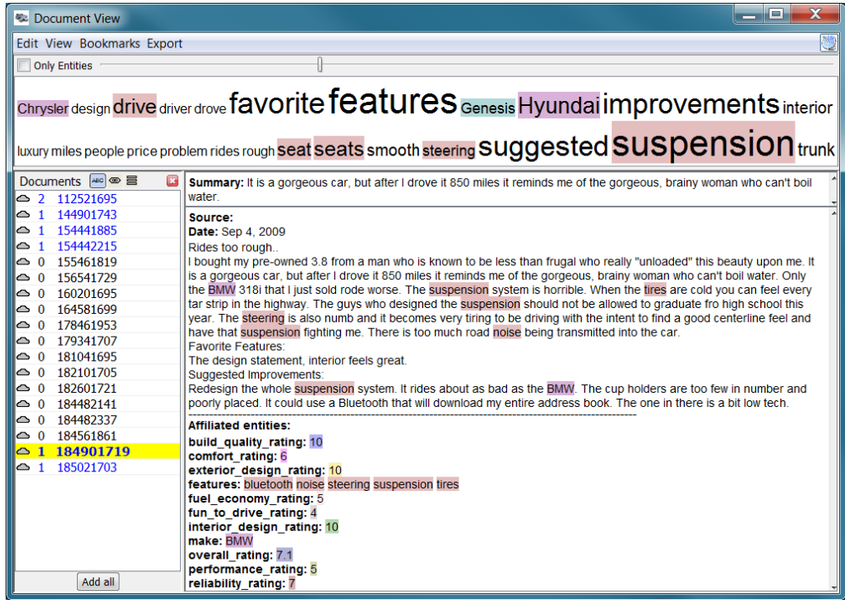












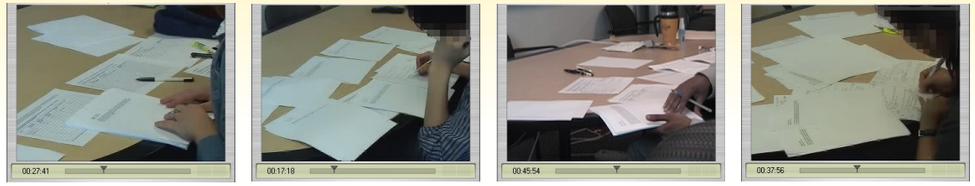


Stasko, Görg, and Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, 2008.

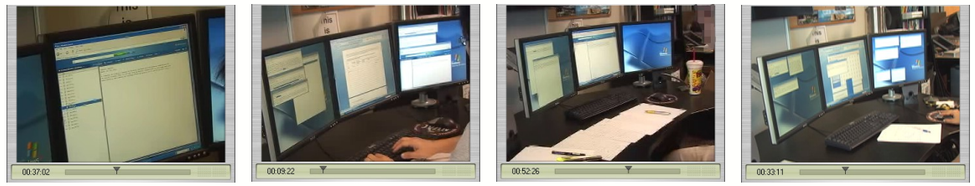
Evaluation Study

4 Settings

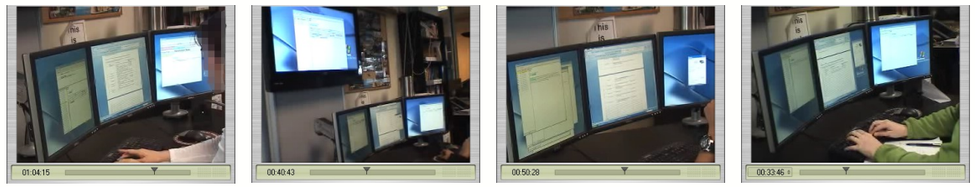
Paper



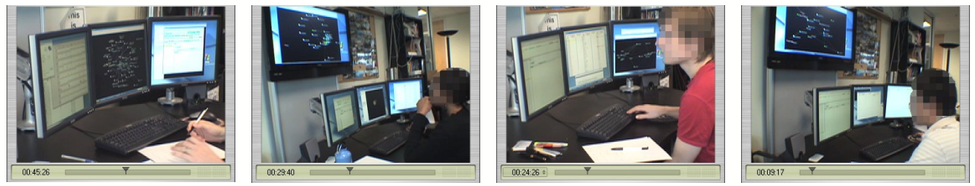
Desktop



Entity



Jigsaw



4 Strategies

1. Overview, filter & detail
   - **Summarization**
   - **Doc Metrics**
2. Build from detail
   - **Doc Similarity**
   - **Recommend related docs**
3. Hit the keyword
4. Find a clue and follow the trail
   - **Doc Clustering**

Kang, Görg, and Stasko. How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation. *IEEE TCVG*, 17(5):570–583, May 2011.

Computational support for hourglass-shaped investigation

Top: lots of data

narrow scope of investigation

Middle: interesting lead

broaden scope of investigation

Bottom: lots of data



Document clustering
Sentiment analysis

Document similarity
Recommending related documents

Computational Text Analyses

- Document Summarization

  (one sentence summaries, word clouds, and keyword summaries)

- Document Similarity

  (based on text or entities, cosine similarity)

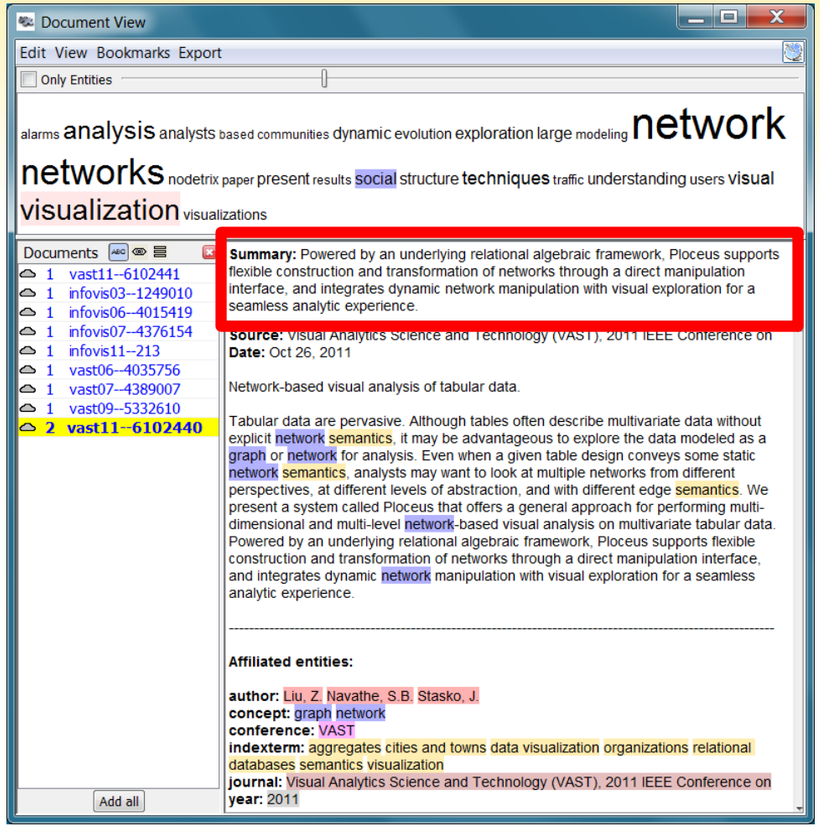
- Document Clustering

  (based on text or entities, k-means clustering)

- Sentiment Analysis

  (hierarchical classification using Lingpipe and dictionary-based approach)

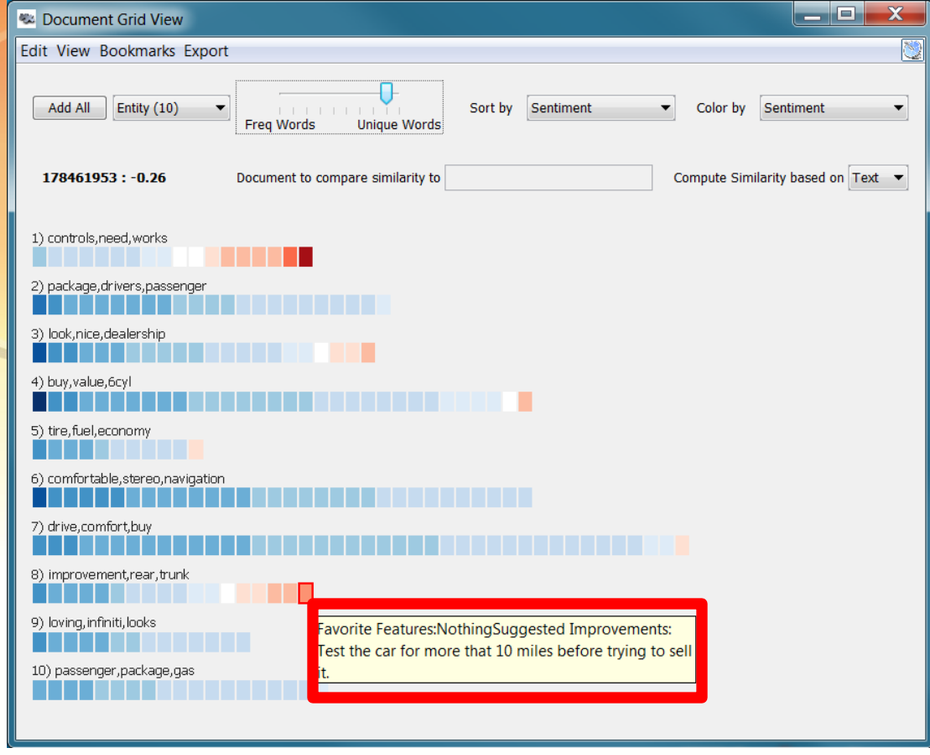
Document Summarization

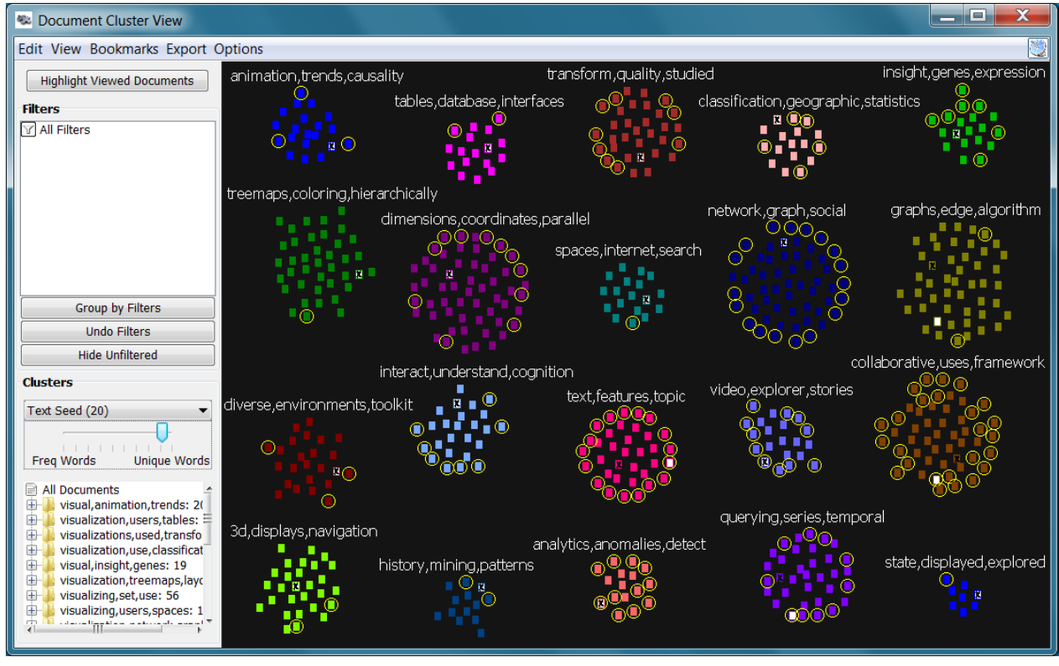


word clouds (multiple docs)

one sentence summaries (single doc)

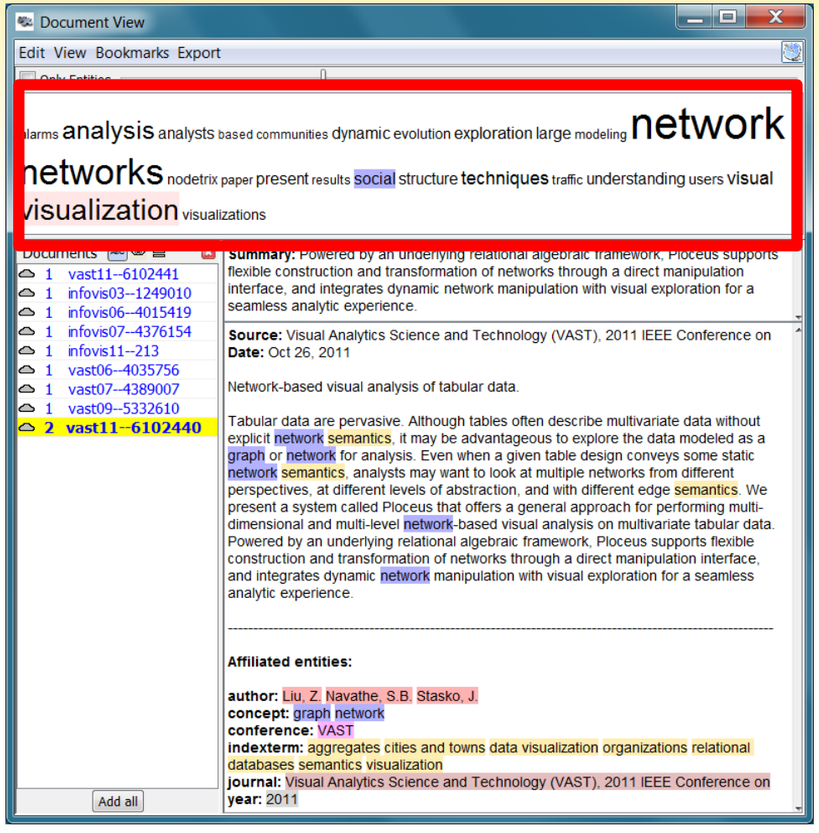
keyword summaries (multiple docs)





VIS 2013 VAST · INFOVIS · SCIVIS BIOVIS · LDAV

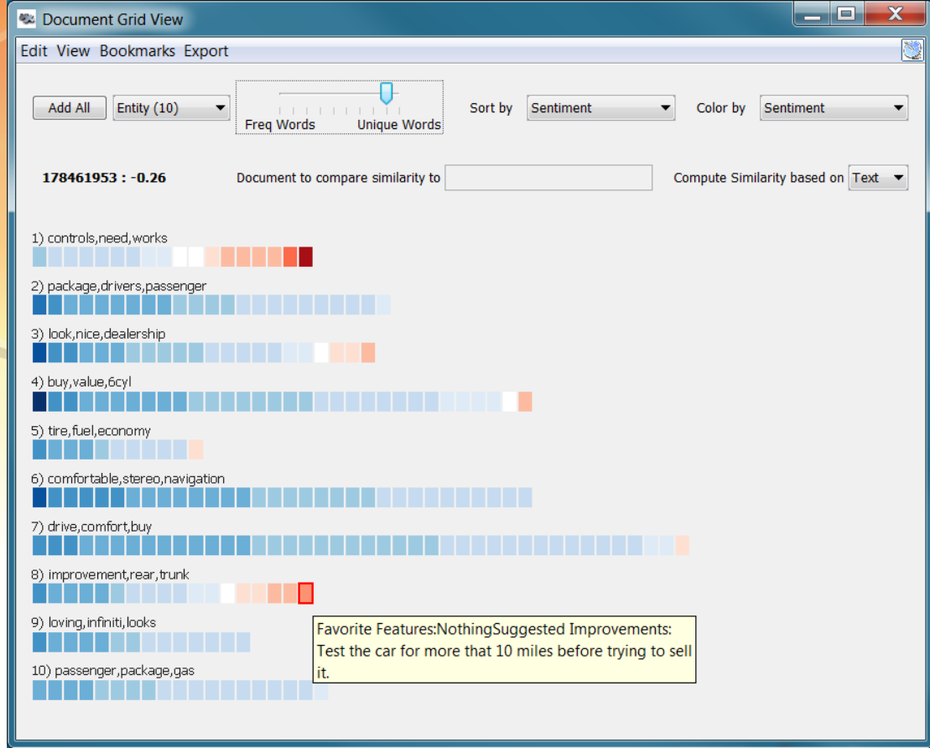
Document Summarization

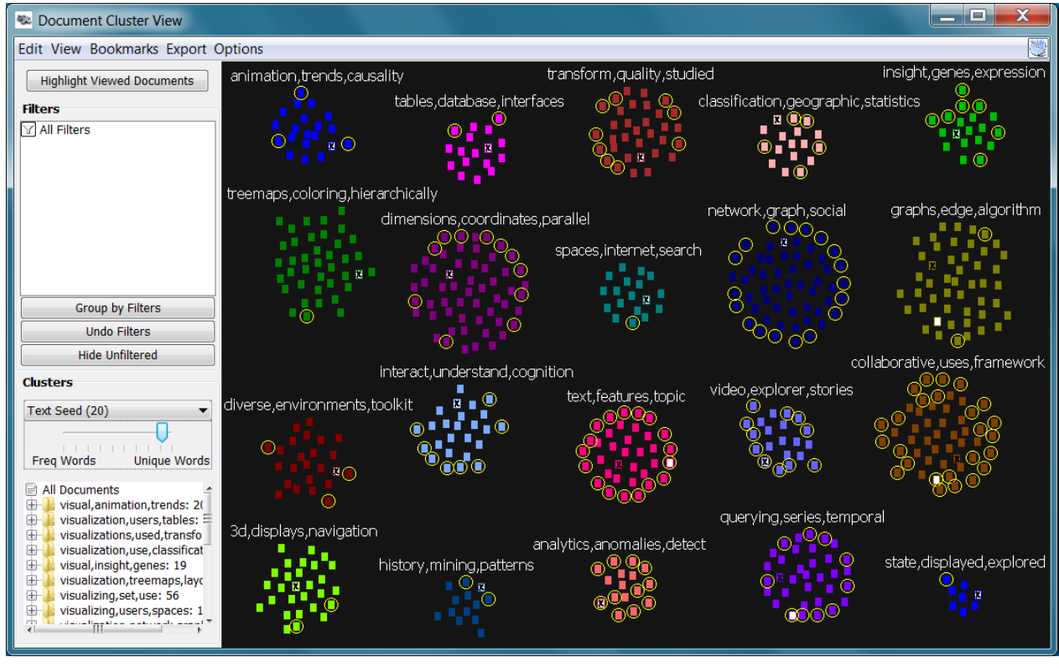


word clouds (multiple docs)

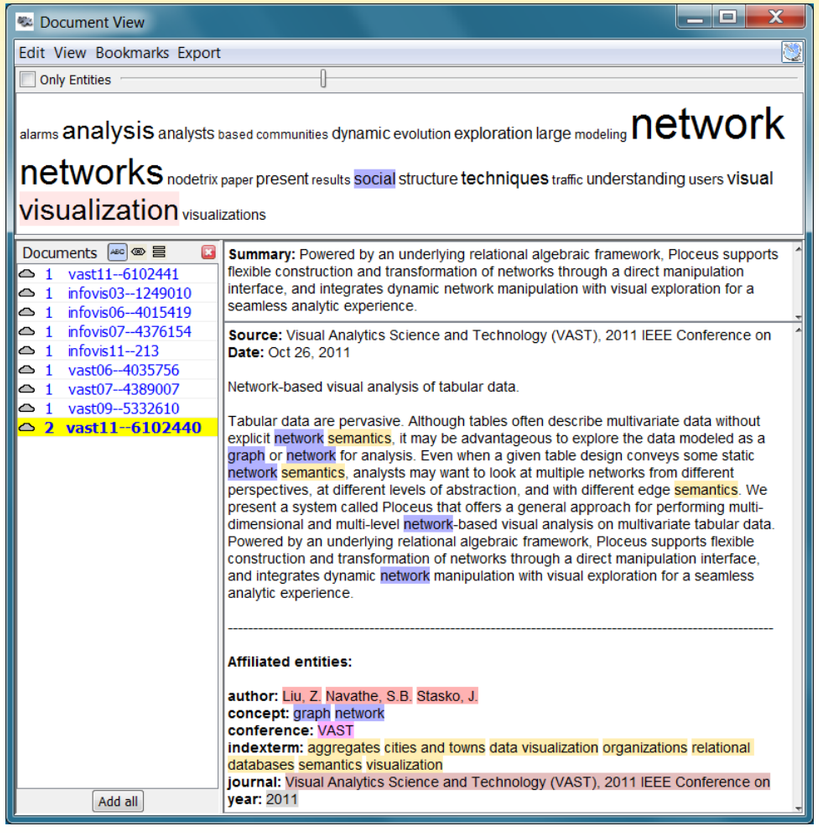
one sentence summaries (single doc)

keyword summaries (multiple docs)





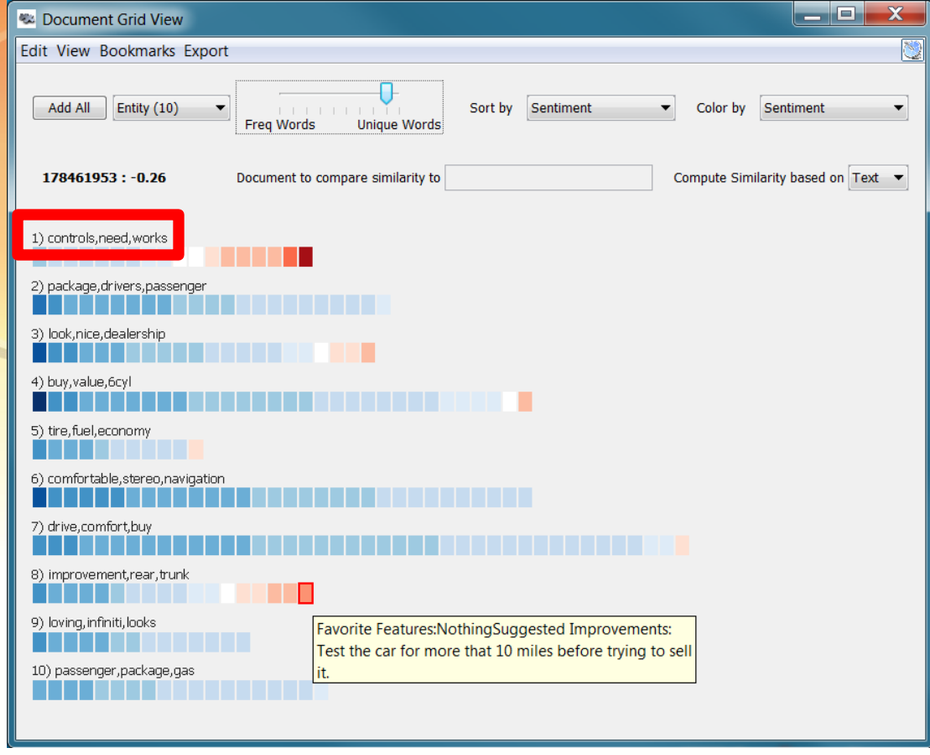
Document Summarization

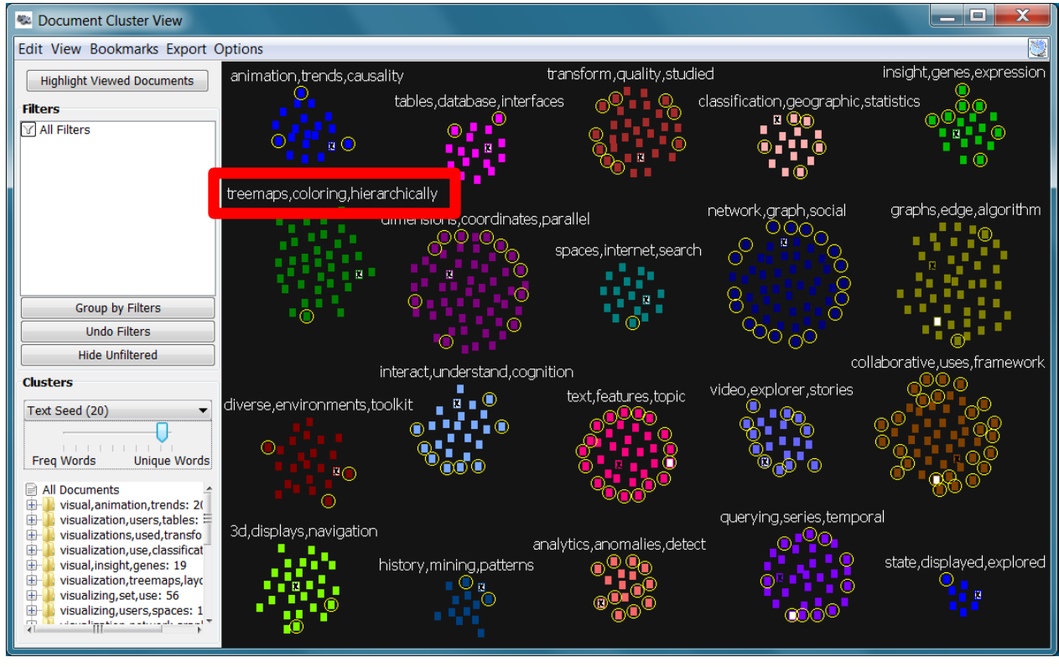


word clouds (multiple docs)

one sentence summaries (single doc)

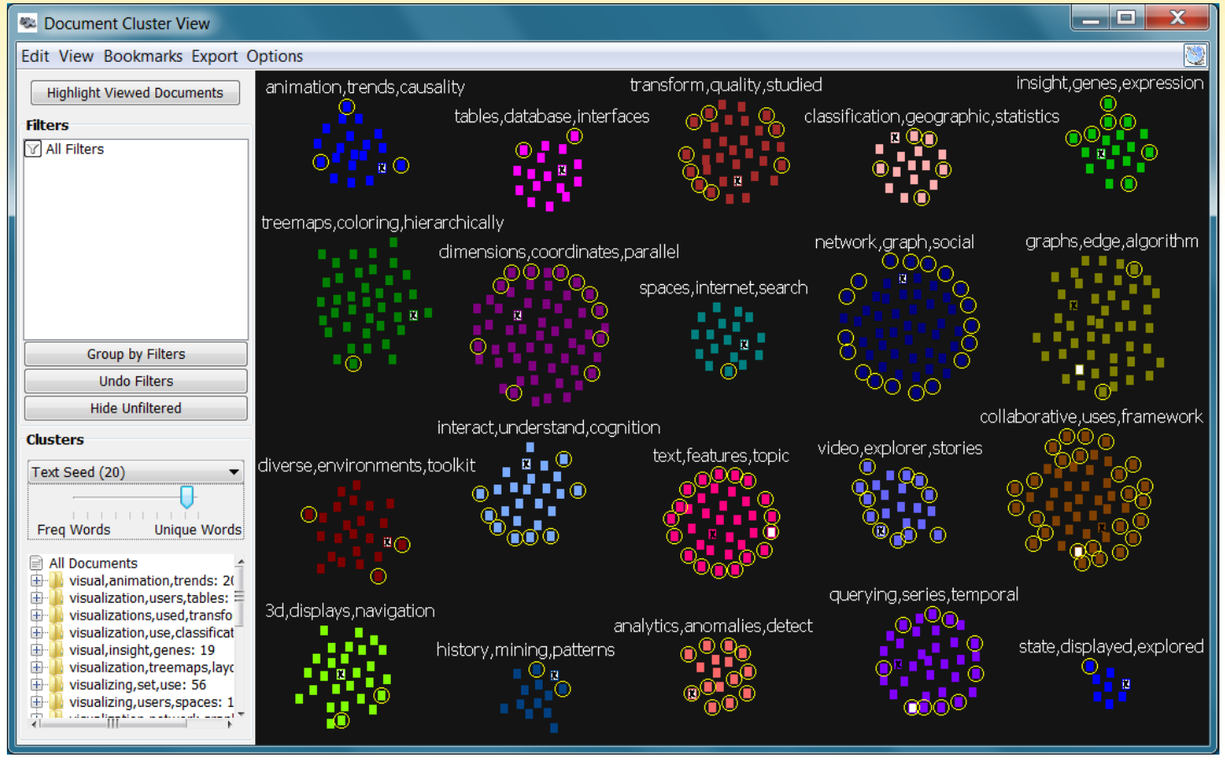
keyword summaries (multiple docs)

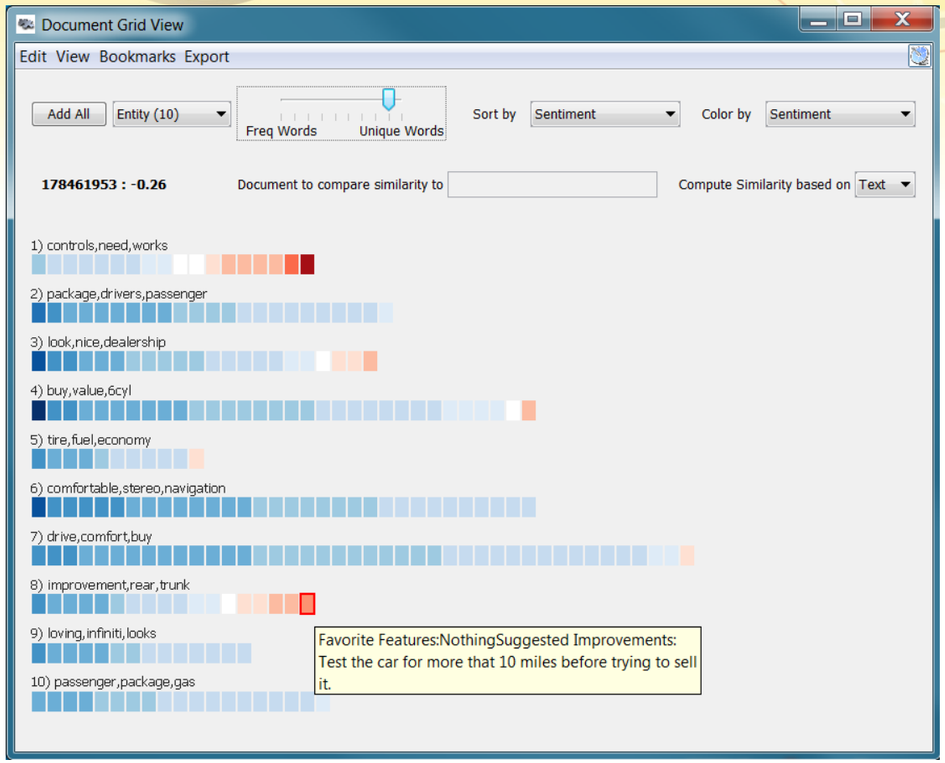






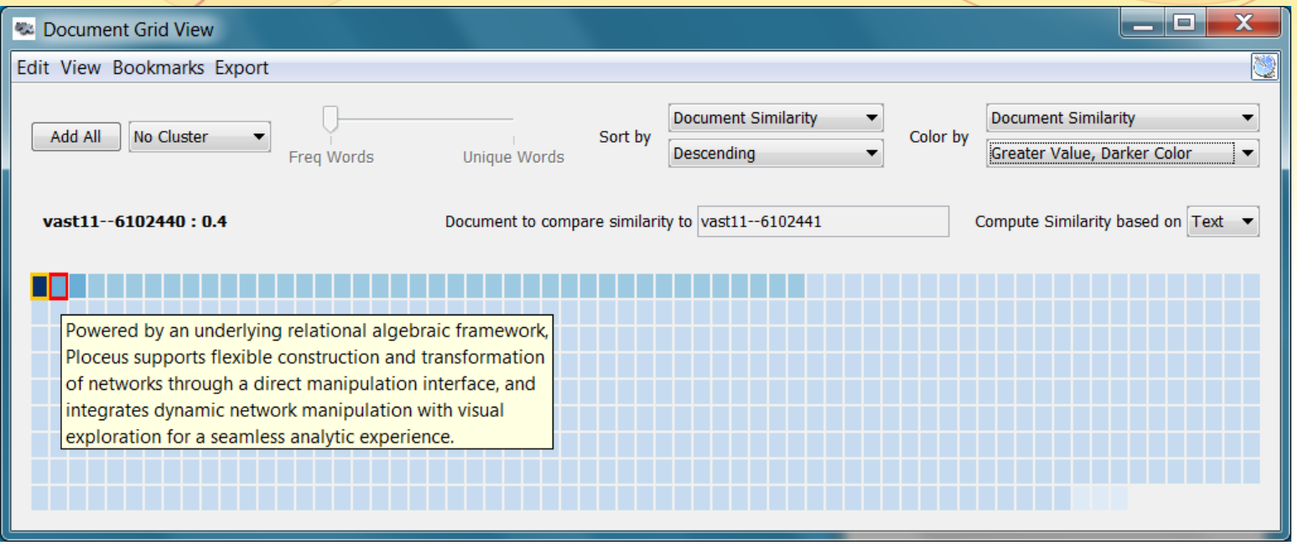
Document Clustering





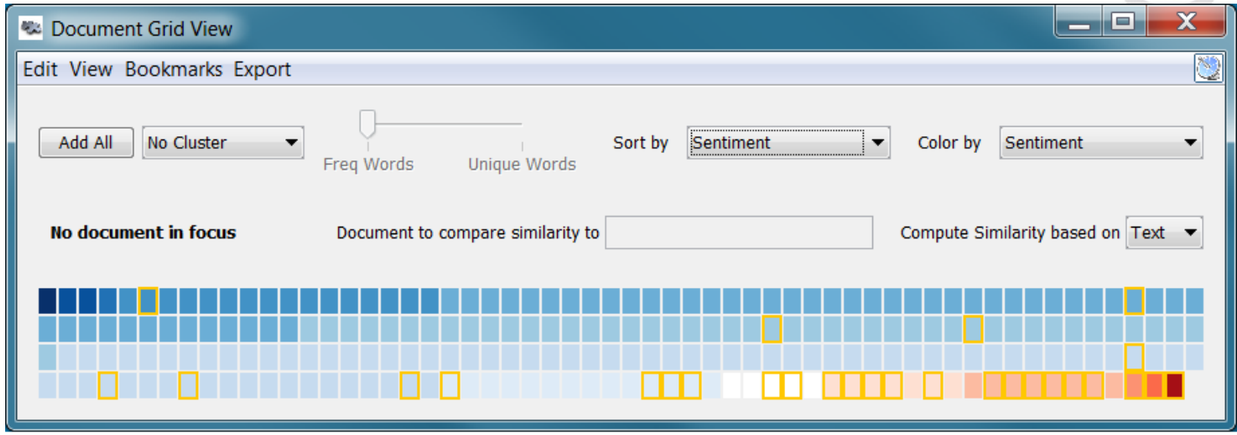
VIS 2013
VAST • INFOVIS • SCIVIS
BIOVIS • LDAV

Document Similarity & Document Metrics

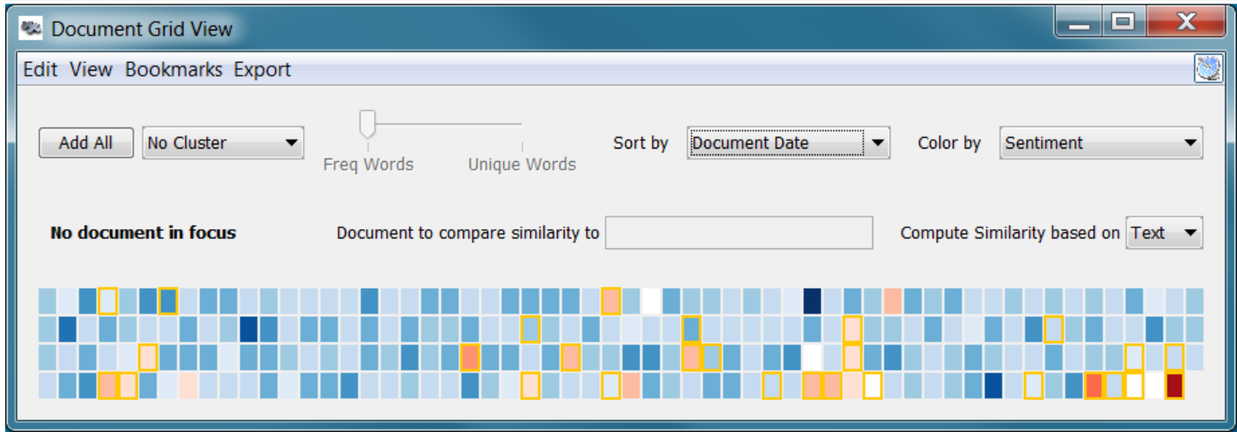


Documents ordered and colored by similarity to selected reference document

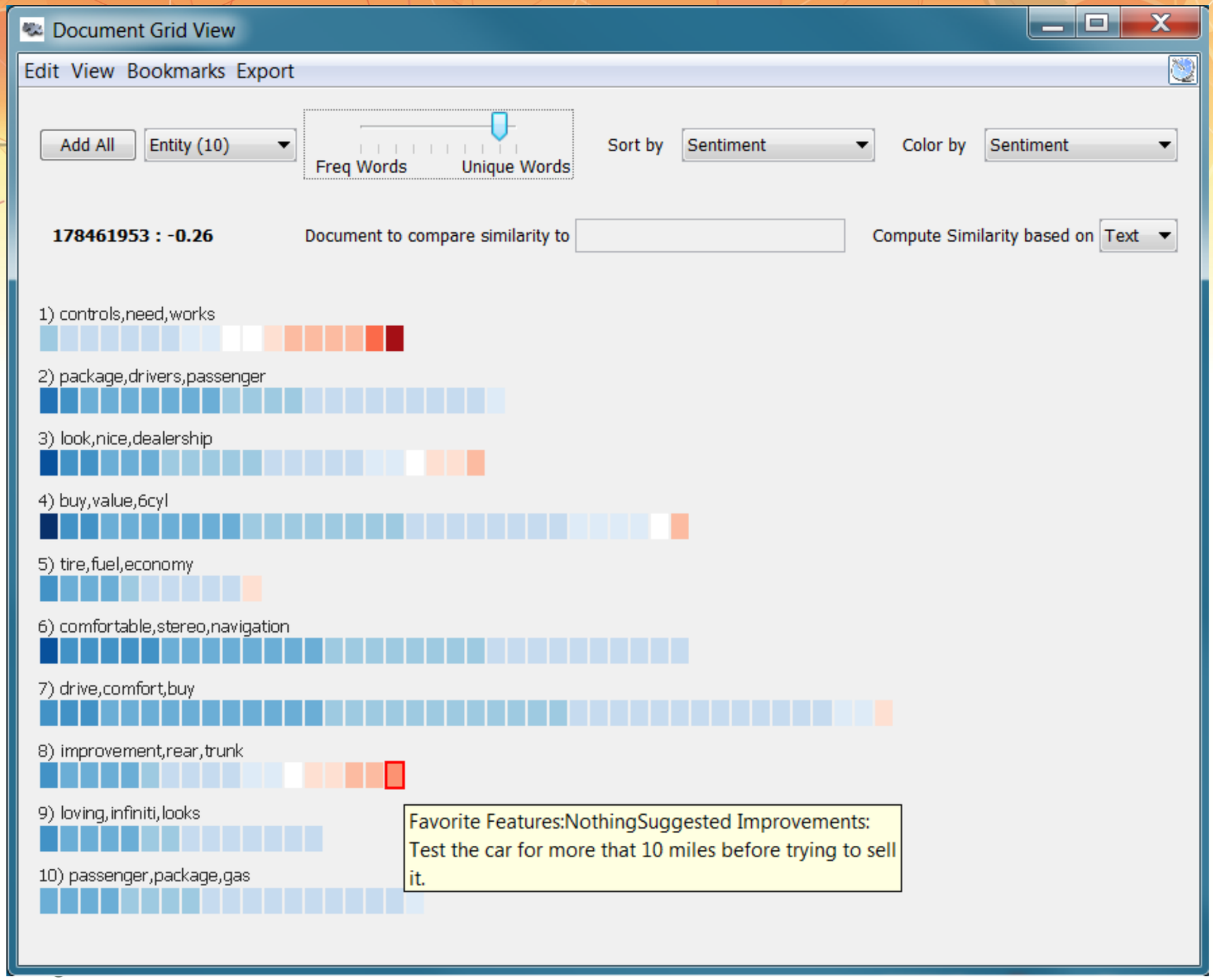
Documents ordered and colored by sentiment



Documents ordered by date and colored by sentiment



VAST · INFOVIS · SCIVIS · BIOVIS · LDAV
VIS 2013



Videos available in IEEE DL

Lessons learned

1. Make different computational analysis results available throughout the system in a variety of different contexts and views
2. Flexibly allow analysis output also to be used as input
3. Integrate different, independent computational analysis measures through interactive visualization to extend functionality and power
4. Provide computational support for both analysis directions: narrowing down as well as widening the scope of an investigation
5. Expose algorithm parameters in an interactive user accessible way

Contributions

1. Methods for fluidly integrating computational text analysis and visualization approaches

2. Illustrating of the utility through example usage scenarios

3. Lessons learned toward the design and construction of visual analytics systems for document exploration

Acknowledgments



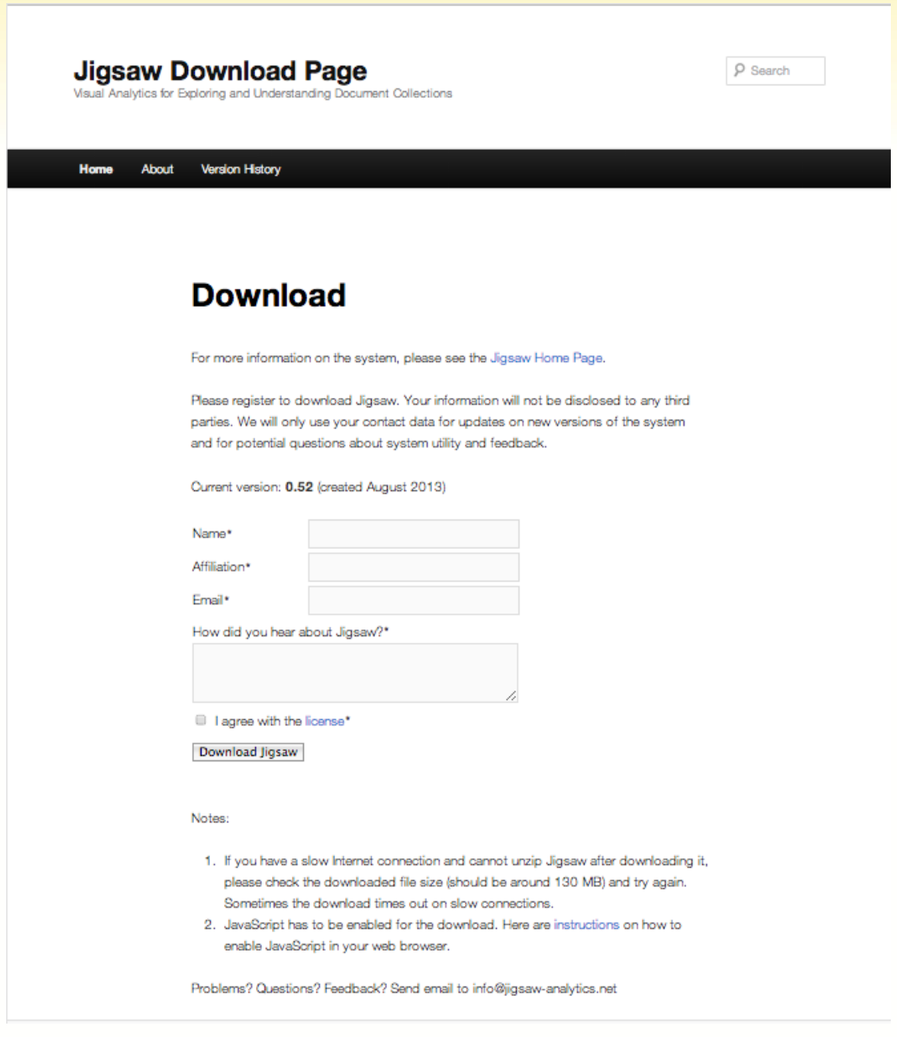
IIS-091788
CCF-0808863



2009-ST-061-CI0001

Get your own copy (incl. InfoVis/VAST paper dataset)

www.jigsaw-analytics.net





Plug-in architecture for your visualizations and algorithms

EuroVis ‘14 Capstone
June 13, 2014



The Value of Visualization… and Why Interaction Matters

John Stasko

School of Interactive Computing
Georgia Institute of Technology

stasko@cc.gatech.edu









Pennard Golf Club



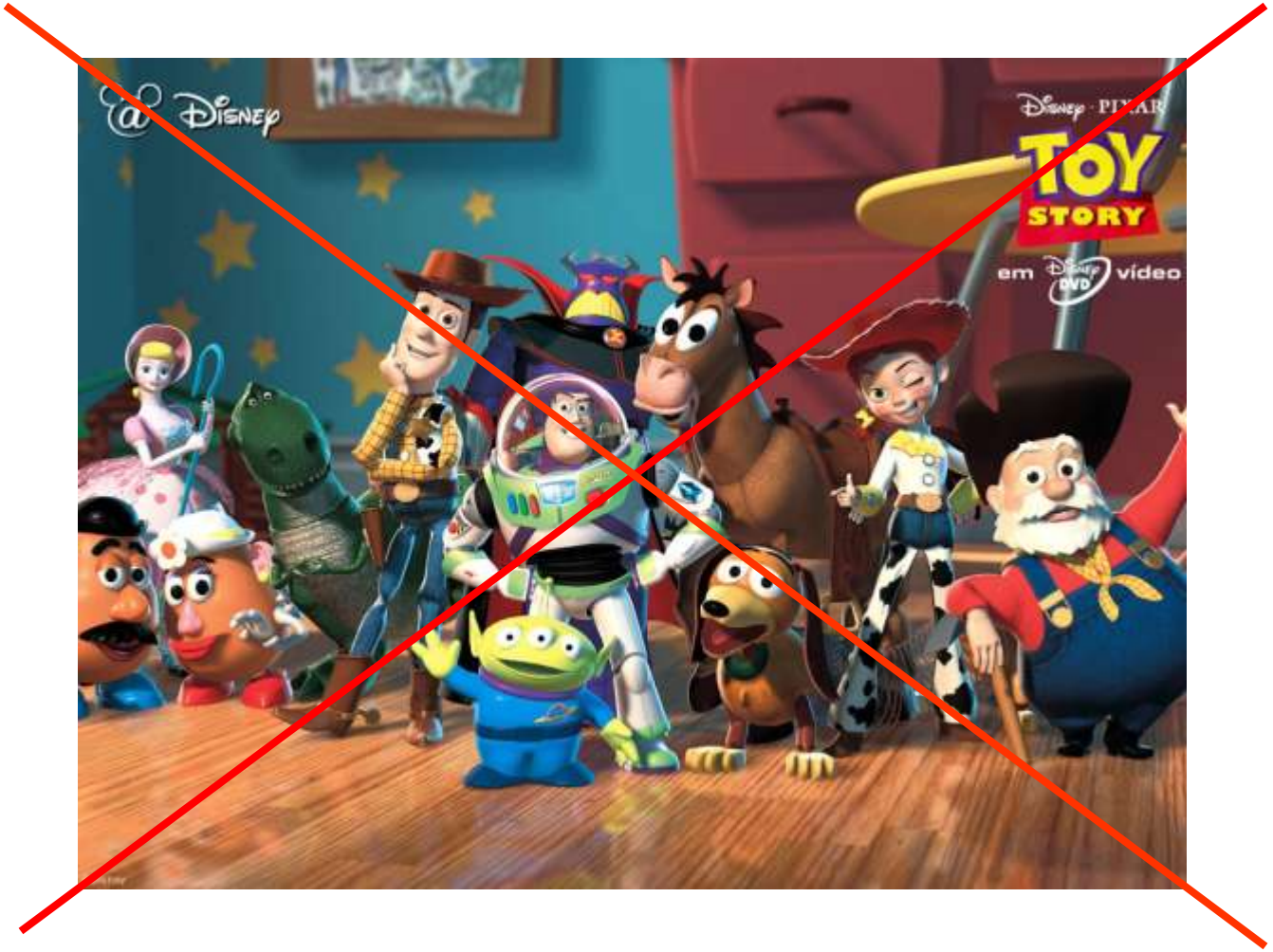
Data Visualization

Making pretty pictures?



Data Visualization

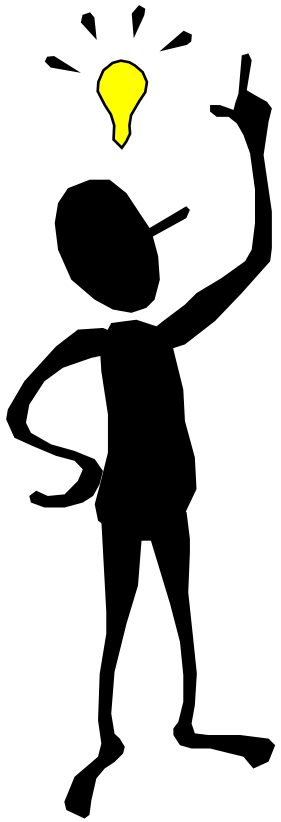
~~Making pretty pictures~~



Data Visualization

A cognitive process

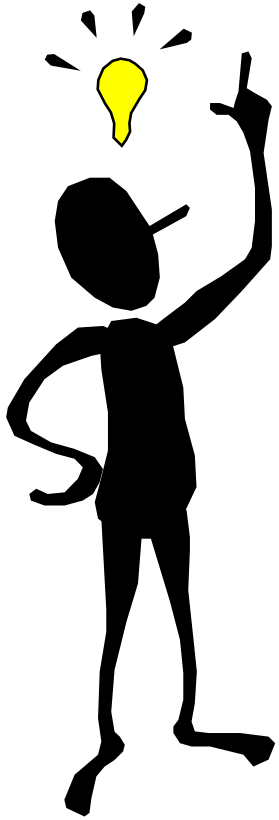
Gain an understanding



Data Visualization

A cognitive process

Gain an **understanding**

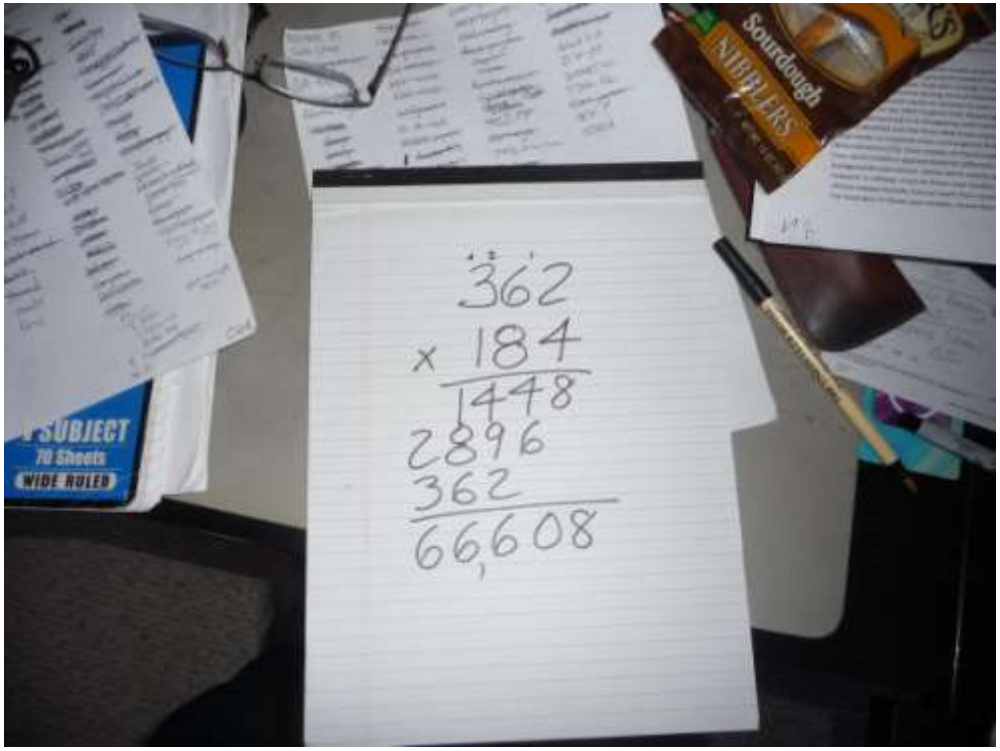


Visuals help us think

Provide a frame of reference, temporary storage area

Cognition $\rightarrow$ Perception

Pattern matching



Visualization's Value

Need to make the case better externally



The Value of Information Visualization

Jean-Daniel Fekete[1], Jarke J. van Wijk[2], John T. Stasko[3], and Chris North[4]

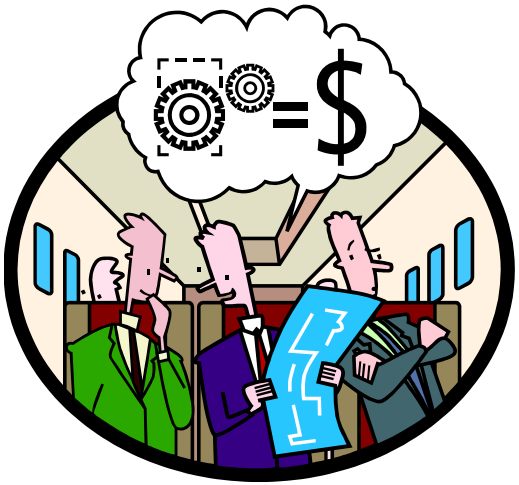
[1] Université Paris-Sud, INRIA, Bât 490,
F-91405 Orsay Cedex, France,
Jean-Daniel.Fekete@inria.fr,
http://www.aviz.fr/~fekete/
[2] Department of Mathematics and Computing Science,
Eindhoven University of Technology, P.O. Box 513,
5600 MB EINDHOVEN, The Netherlands,
vanwijk@win.tue.nl,
http://www.win.tue.nl/~vanwijk/
[3] School of Interactive Computing, College of Computing & GVU Center,
Georgia Institute of Technology, 85 5th St., NW,
Atlanta, GA 30332-0760, USA,
stasko@cc.gatech.edu,
http://www.cc.gatech.edu/~john.stasko
[4] Dept of Computer Science, 2202 Kraft Drive,
Virginia Tech, Blacksburg, VA 24061-0106, USA,
north@vt.edu,
http://people.cs.vt.edu/~north/

Abstract. Researchers and users of Information Visualization are convinced that it has value. This value can easily be communicated to others in a face-to-face setting, such that this value is experienced in practice. To convince broader audiences, and also, to understand the intrinsic

Information Visualization: Human-Centered Issues and Perspectives
Springer, 2008

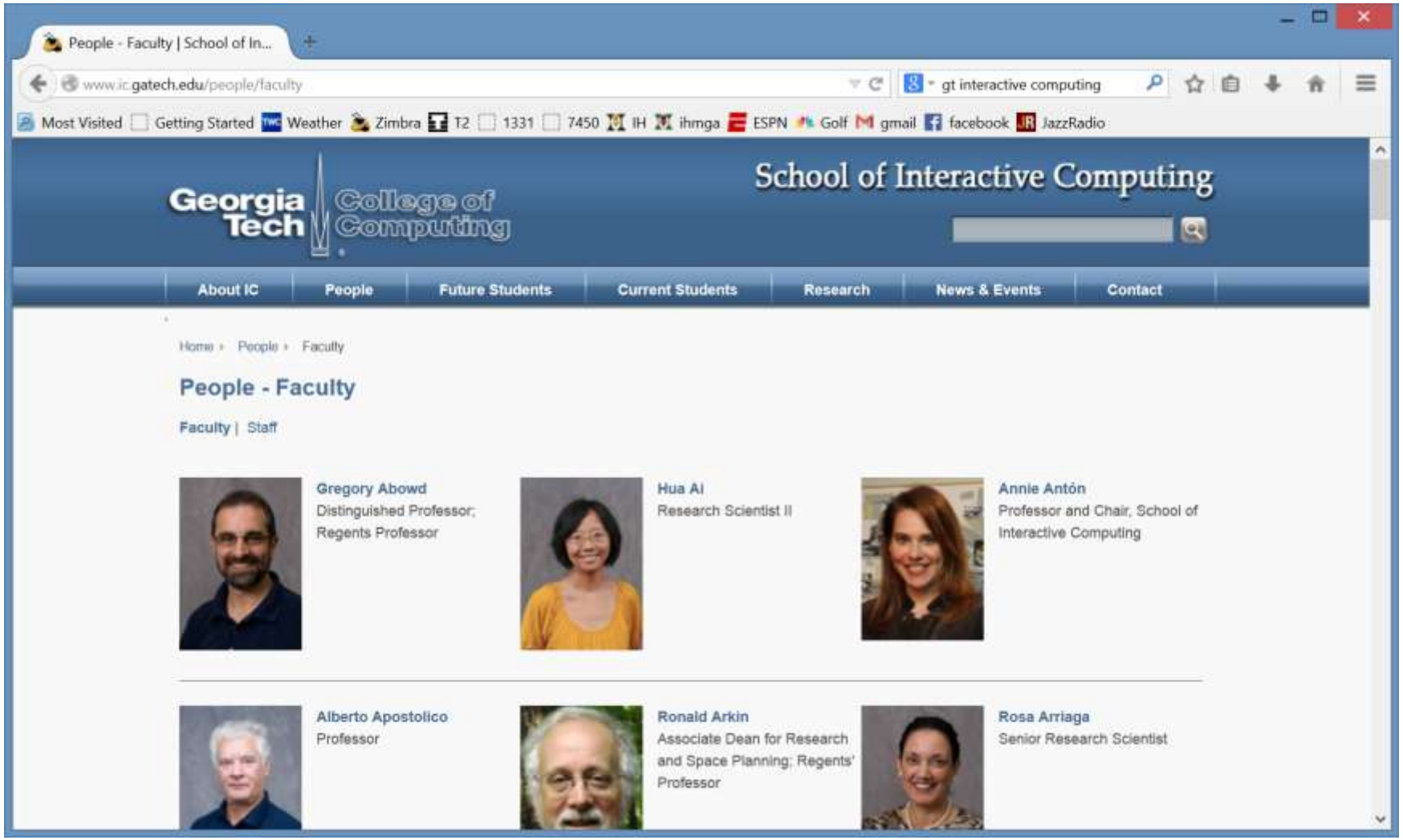
Need for Visualization

Articulating value requires identifying purpose



Three stories…




People - Faculty | School of In...
www.ic.gatech.edu/people/faculty
gt interactive computing
Most Visited
Getting Started
Weather
Zimbra
T2
1331
7450
IH
ihmga
ESPN
Golf
gmail
facebook
JazzRadio
Georgia Tech College of Computing
School of Interactive Computing
About IC
People
Future Students
Current Students
Research
News & Events
Contact
Home › People › Faculty
People - Faculty
Faculty | Staff
Gregory Abowd
Distinguished Professor; Regents Professor
Hua Ai
Research Scientist II
Annie Antón
Professor and Chair, School of Interactive Computing
Alberto Apostolico
Professor
Ronald Arkin
Associate Dean for Research and Space Planning; Regents' Professor
Rosa Arriaga
Senior Research Scientist

Instagram Visualization

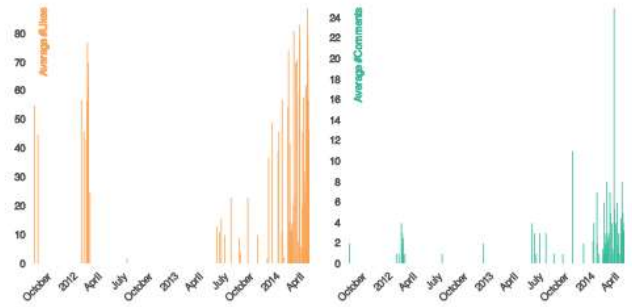
Result for the hashtag # cat with # GO

Options

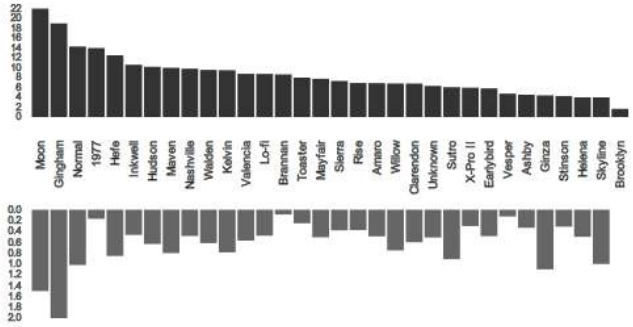
Number of photos per day (photos get from API, not the actual photo counts on Instagram)

Enable Tooltip Enable Brush





Average numbers of likes per photos per filter



Average numbers of comments per photos per filter

Popular photos

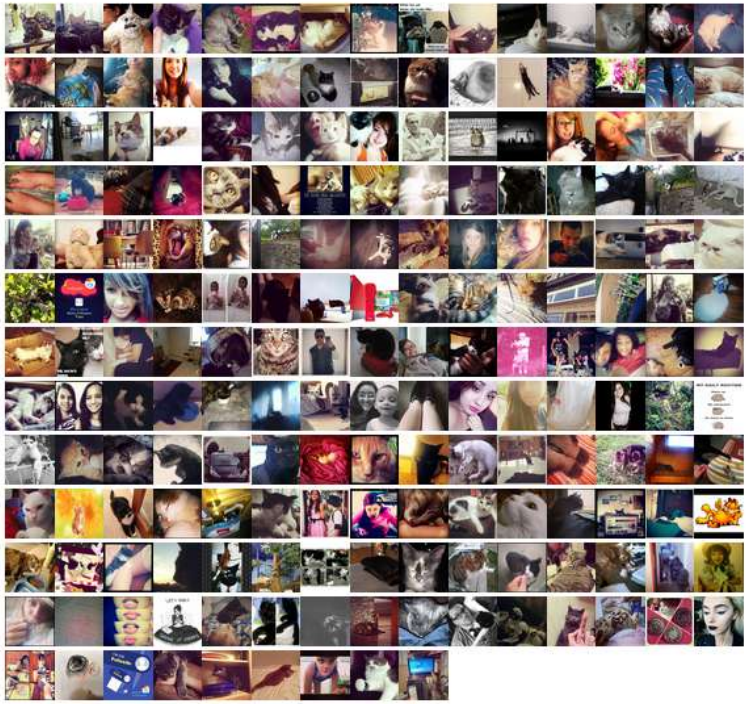


Photo Navigator On Off

Color schemes of the photo



Chloe Xie



Harry Callahan (Clint Eastwood)

“A man’s got to know his limitations.”
Magnum Force, 1973

Thought

If you can articulate very precisely what you're seeking, visualization likely isn't your best approach

OK, so what is visualization good for?

Applications of Visualization

Presentation

Analysis

Explanatory

Exploratory

1. Presentation

Communicate data and ideas

Explain and inform

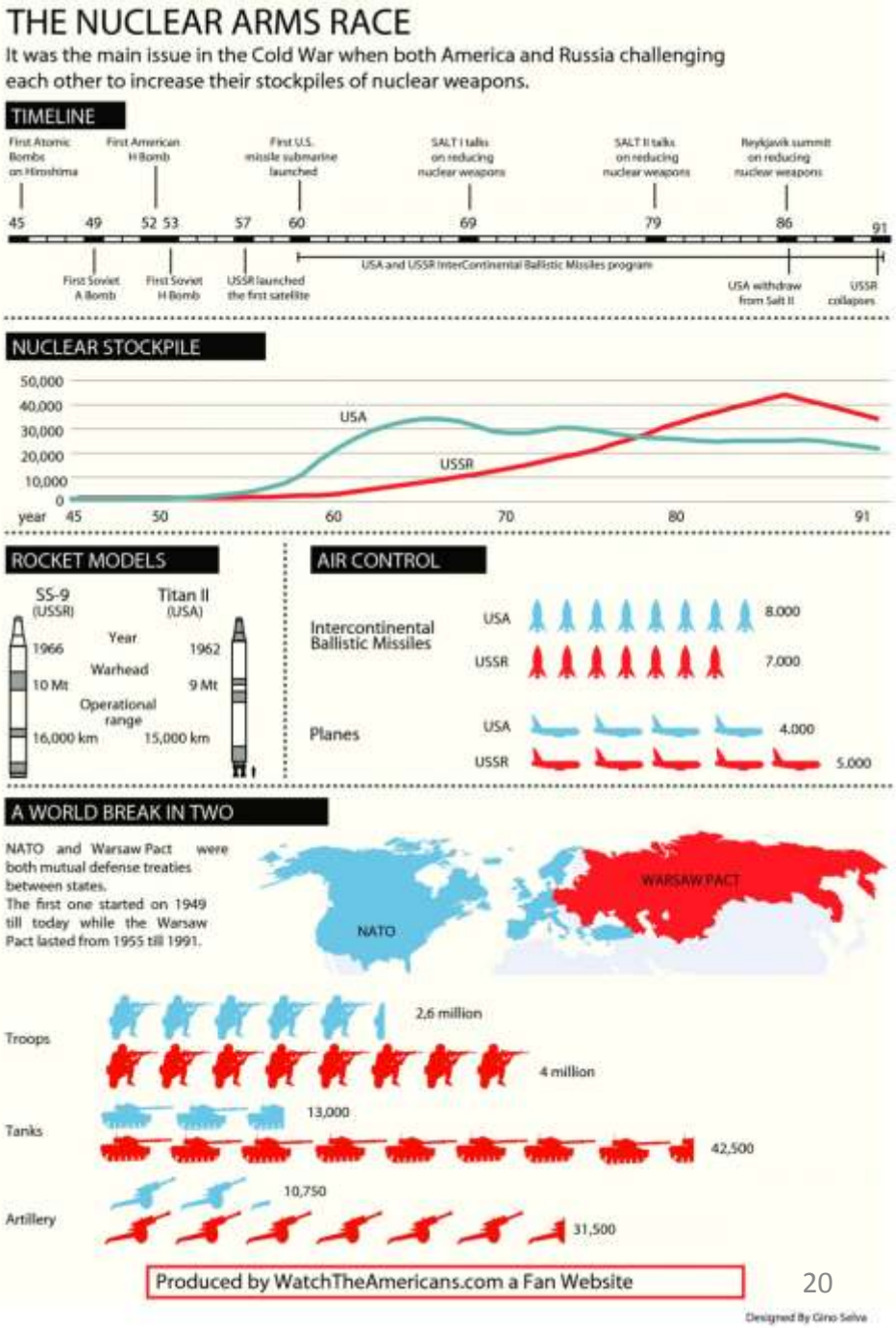
Provide evidence and support

Influence and persuade



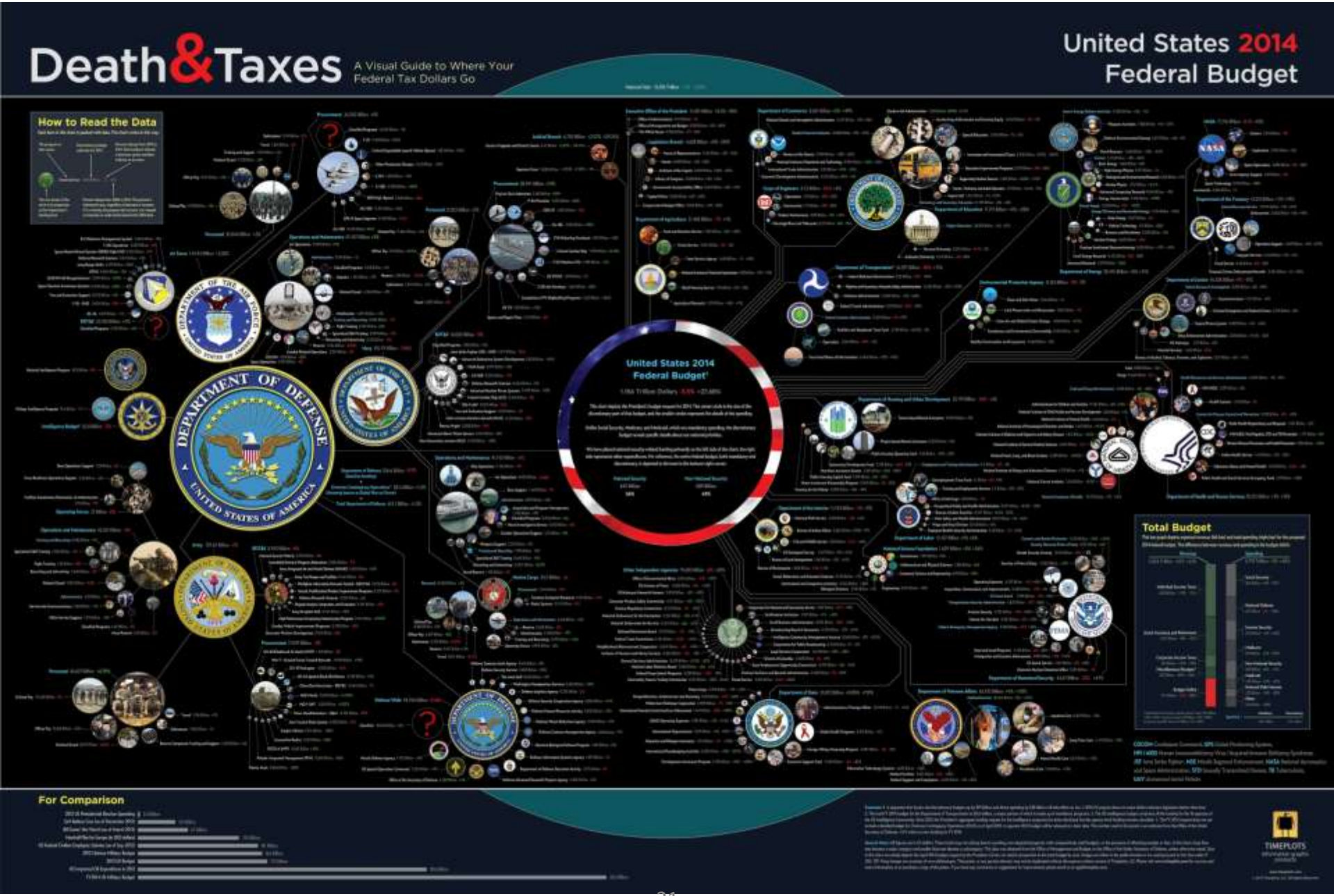
Infographics

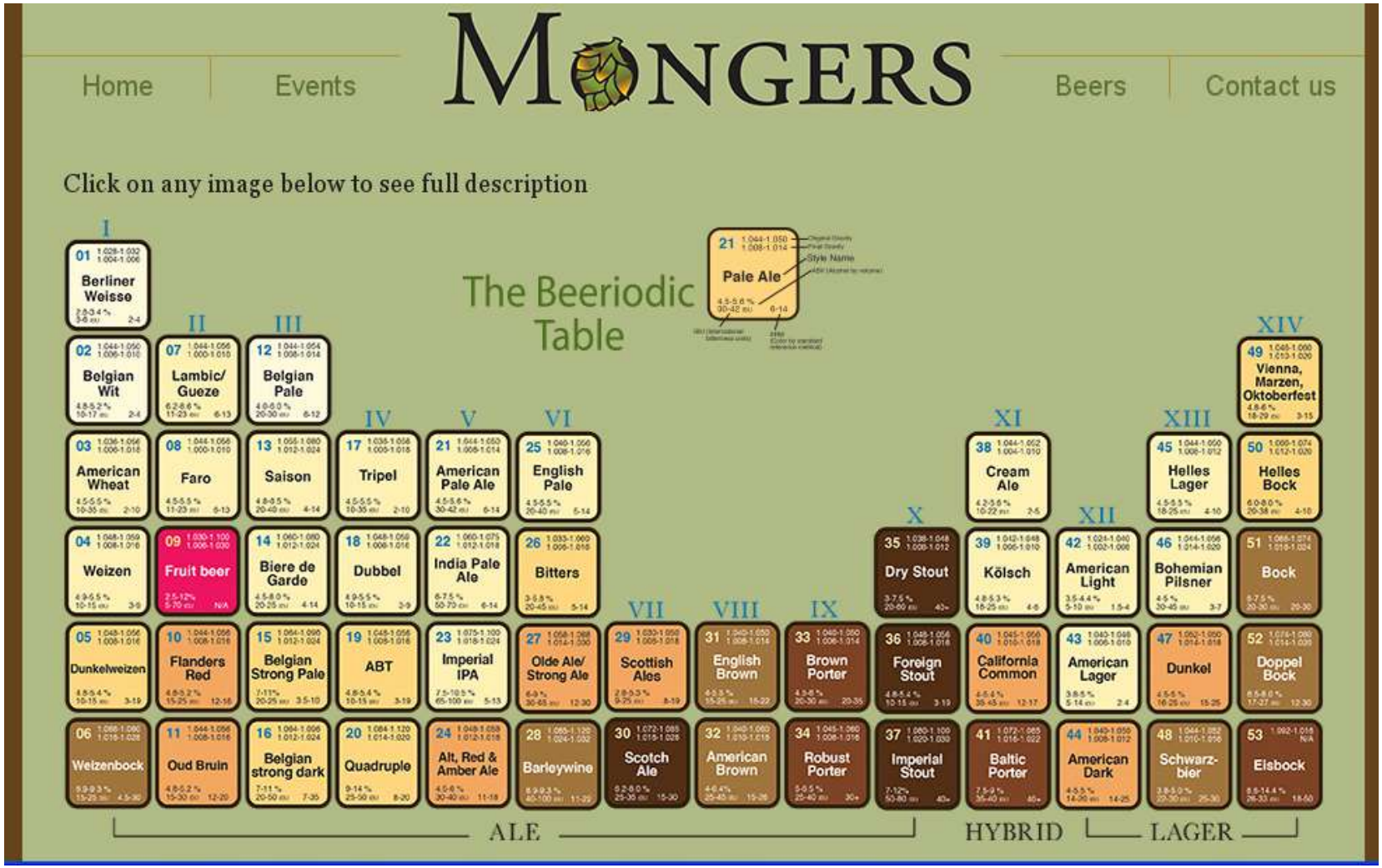
All the rage…



http://visual.ly/nuclear-arms-race

http://visual.ly/death-and-taxes-2014-us-federal-budget





http://thebeermongers.com/beers/

NY Times 2006

755

Steroids or Not, the Pursuit Is On

Barry Bonds is taking aim at the career home run record. He needs only six more to tie Babe Ruth and 47 to equal Hank Aaron.



Lines are cumulative home runs.

Hank Aaron
755 homers
23 seasons



Babe Ruth
714 homers
22 seasons



Barry Bonds
708 homers
20 seasons

Bonds takes lead
Home runs after 16 seasons

Bonds	567
Aaron	554
Ruth	516

755
714
23 seasons
22 seasons
20 seasons
Bonds was injured last season. He played 14 games and hit 6 homers.

600
400
200
0
14th season

According to allegations in a book about Bonds, he began taking steroids before the 1999 season, his 14th in the league. Two seasons later, he hit 73 home runs, surpassing Aaron's career pace.

5 seasons
10
15

Homer Pace After Age 34

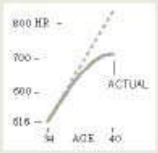
If the accusations are correct, Bonds was 34 in his first season on steroids. Here are projected home run paces for each player after age 34.

PROJECTED PACE BASED ON AVERAGE OF PREVIOUS FIVE SEASONS

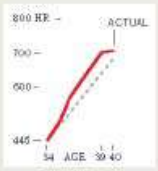
Aaron
Actual homers slightly outpace projected homers for five seasons.



Ruth
Averaged 46.4 homers a season from age 30 to 34. Averaged 42.5 for next four seasons.



Bonds
From age 35 to 39, he averaged 14 more homers a season than projected.



Note: Ages as of July 1 of each season.

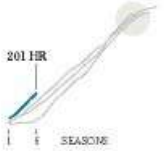


Others Taking Aim

Alex Rodriguez
Is ahead of the pace set by all three home run leaders.



Albert Pujols
Averaging 40 homers a season, he has started stronger than the three leaders did.

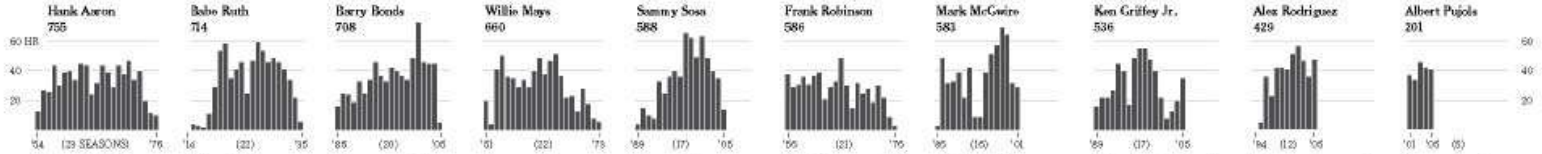


Ken Griffey Jr.
Many thought he would be the first to catch Ruth and Aaron until injuries limited his output.

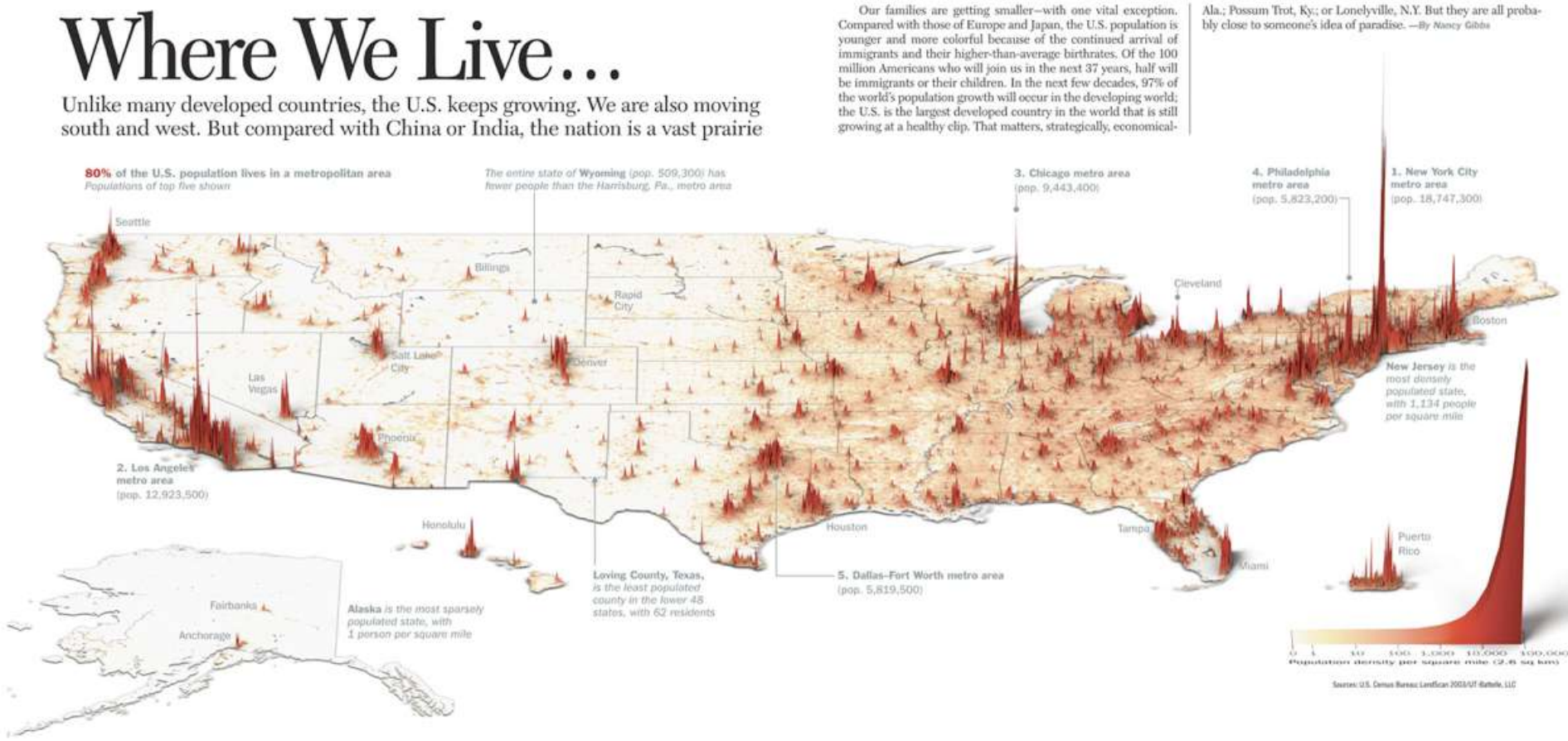


Differing Paths to the Top of the Charts

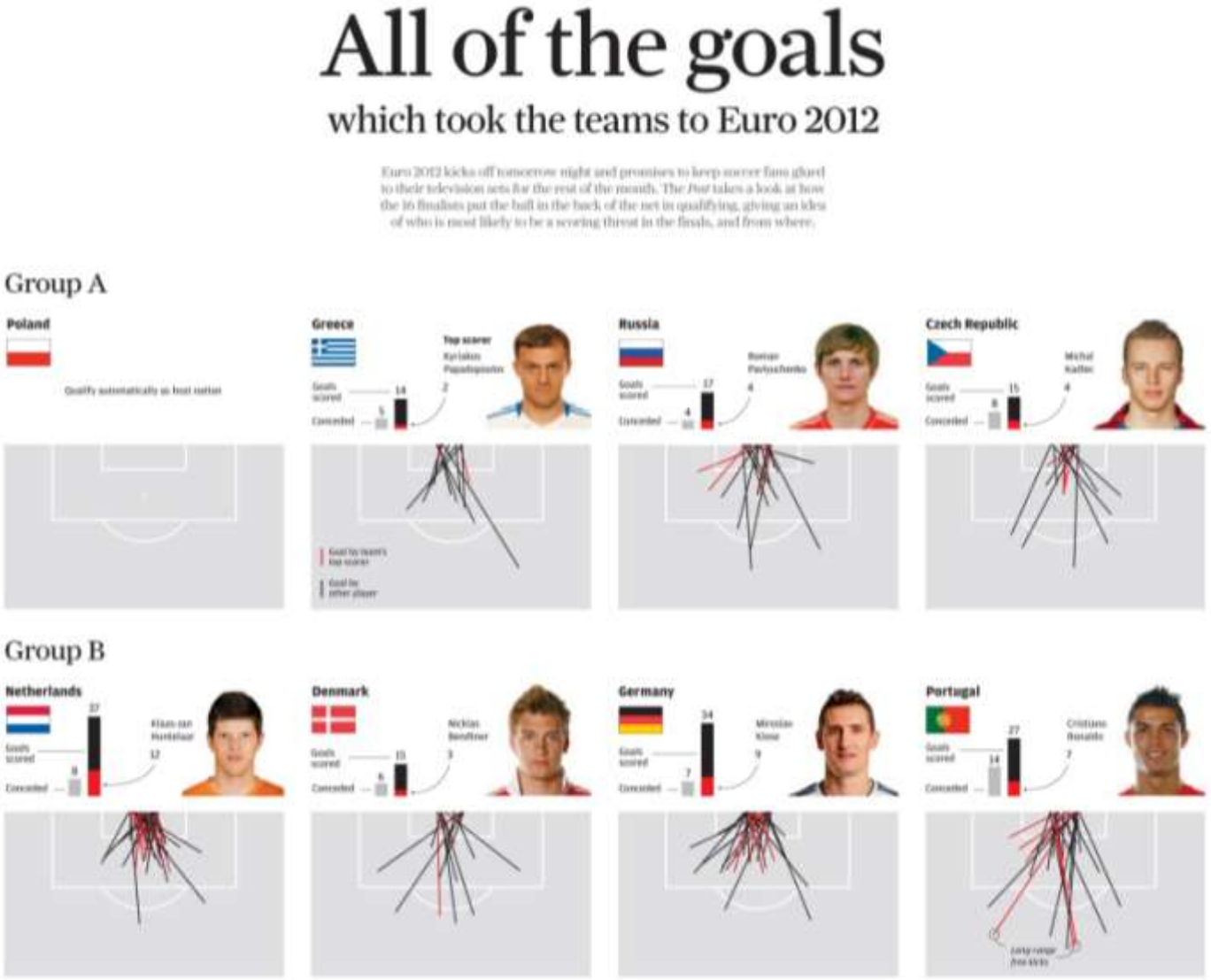
The top seven players on the career home run list, along with a look at Griffey (12th), Rodriguez (37th) and Pujols (tied 257th).



Amanda Cox and Joe Ward/The New York Times



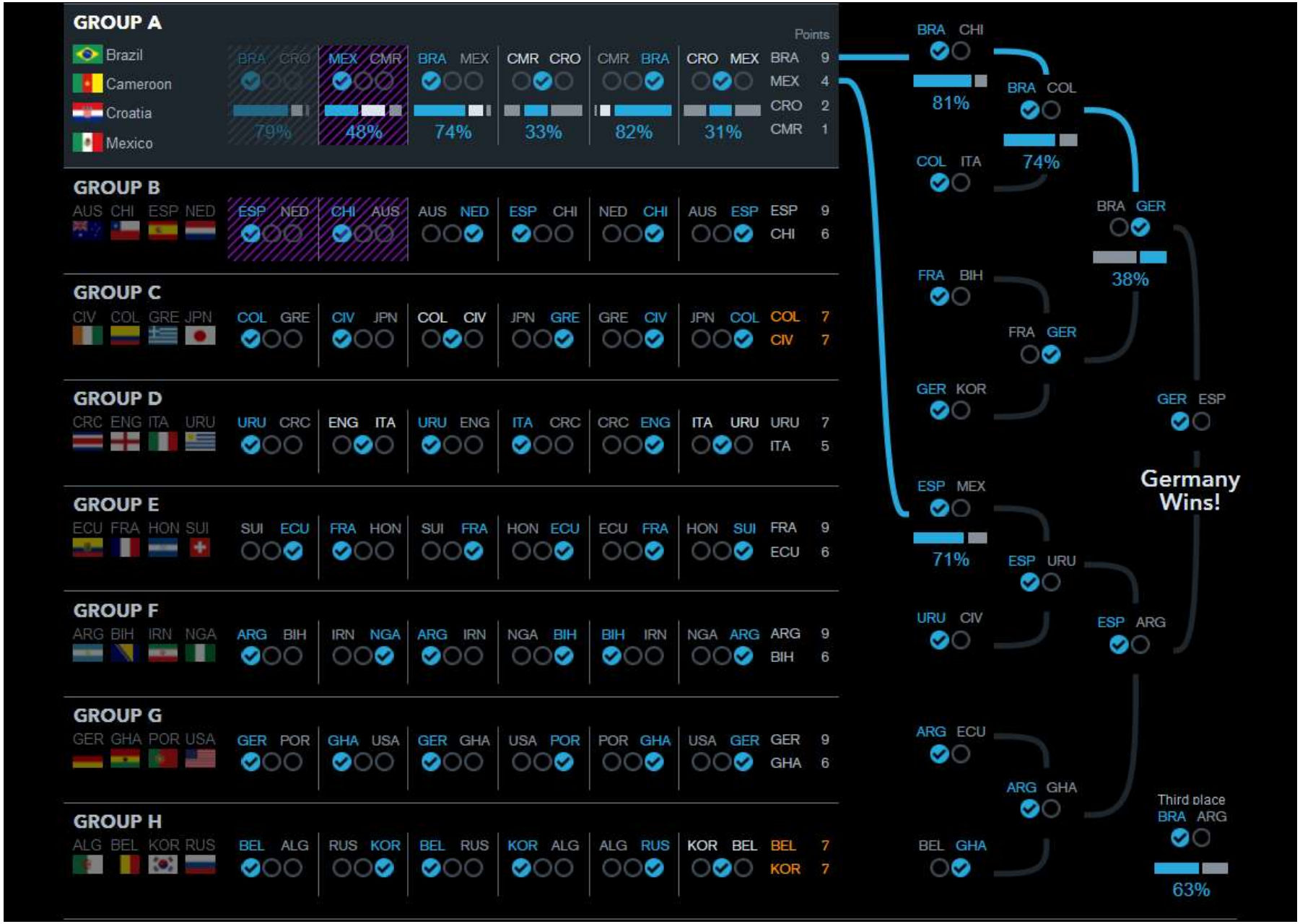
http://infographicsnews.blogspot.com/2009/04/mantras-joe-lertolas-maps.html



Simon Scarr
South China Morning Post
June 7, 2012

http://www.visualisingdata.com/blog/wp-content/uploads/2013/04/GOALS.jpg

http://www.bloomberg.com/visual-data/world-cup



The Times
June 12, 2014

the game • 2014 World Cup

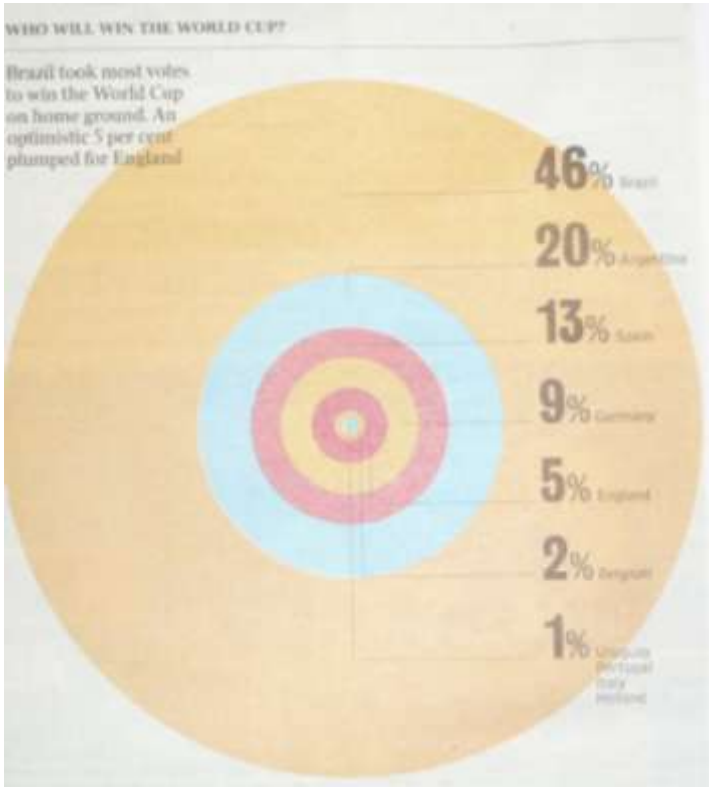
TIMES READER PREDICTIONS

More than 3,000 *Times* readers have gone online to give their predictions about how the World Cup finals will develop. Starting with the groups, readers have worked their way through the competition and their thoughts are collated below. Participants were asked which teams would qualify from the group stages and who would win the tournament overall. There was little disagreement, with only group E throwing up diverse viewpoints.

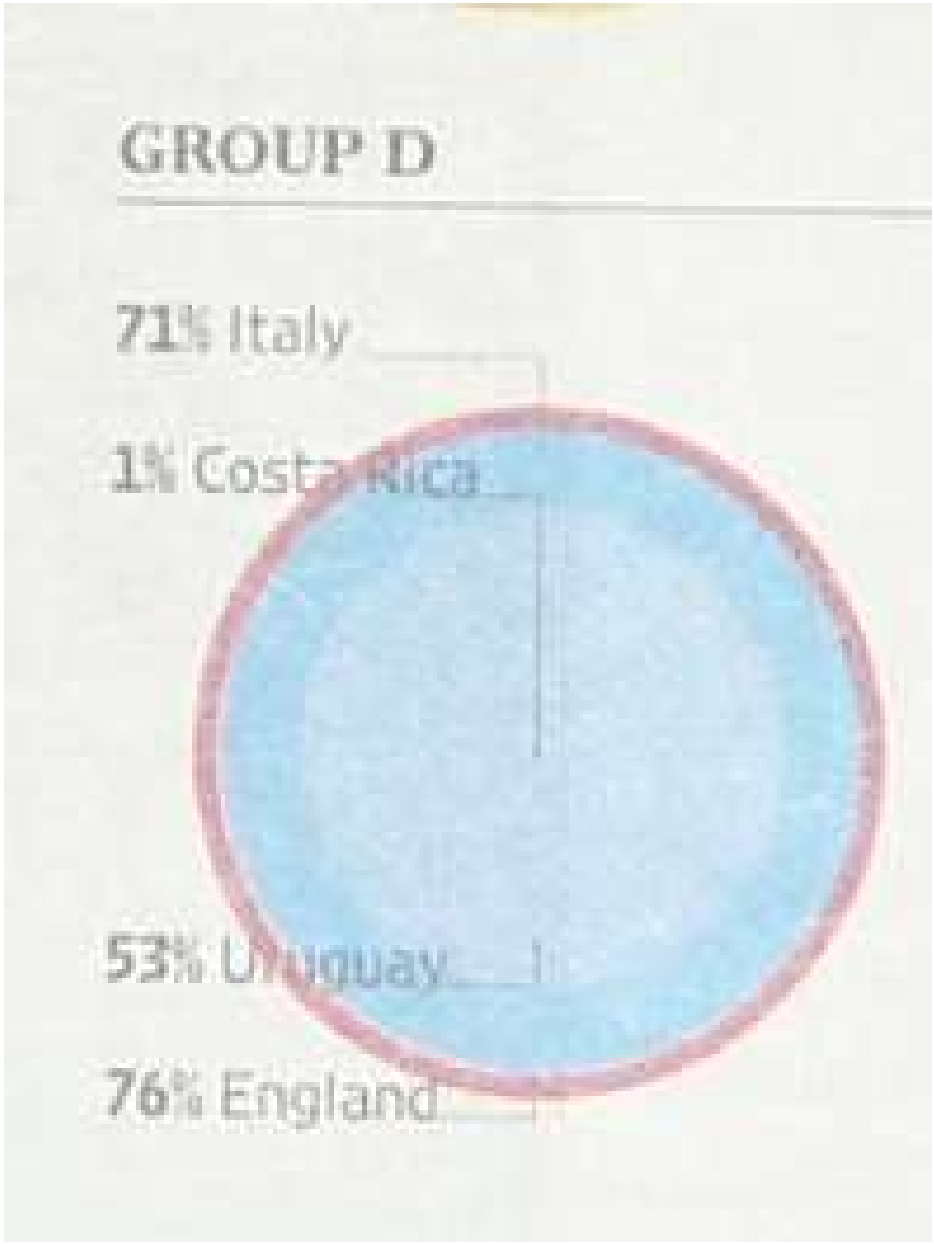
THE GROUP STAGES

Everyone thought Brazil would qualify from group A but were split over who join the hosts in the round of 16 by finishing second. Australia were the least fancied team to advance, a mere 1 per cent of people think they would get out of group B

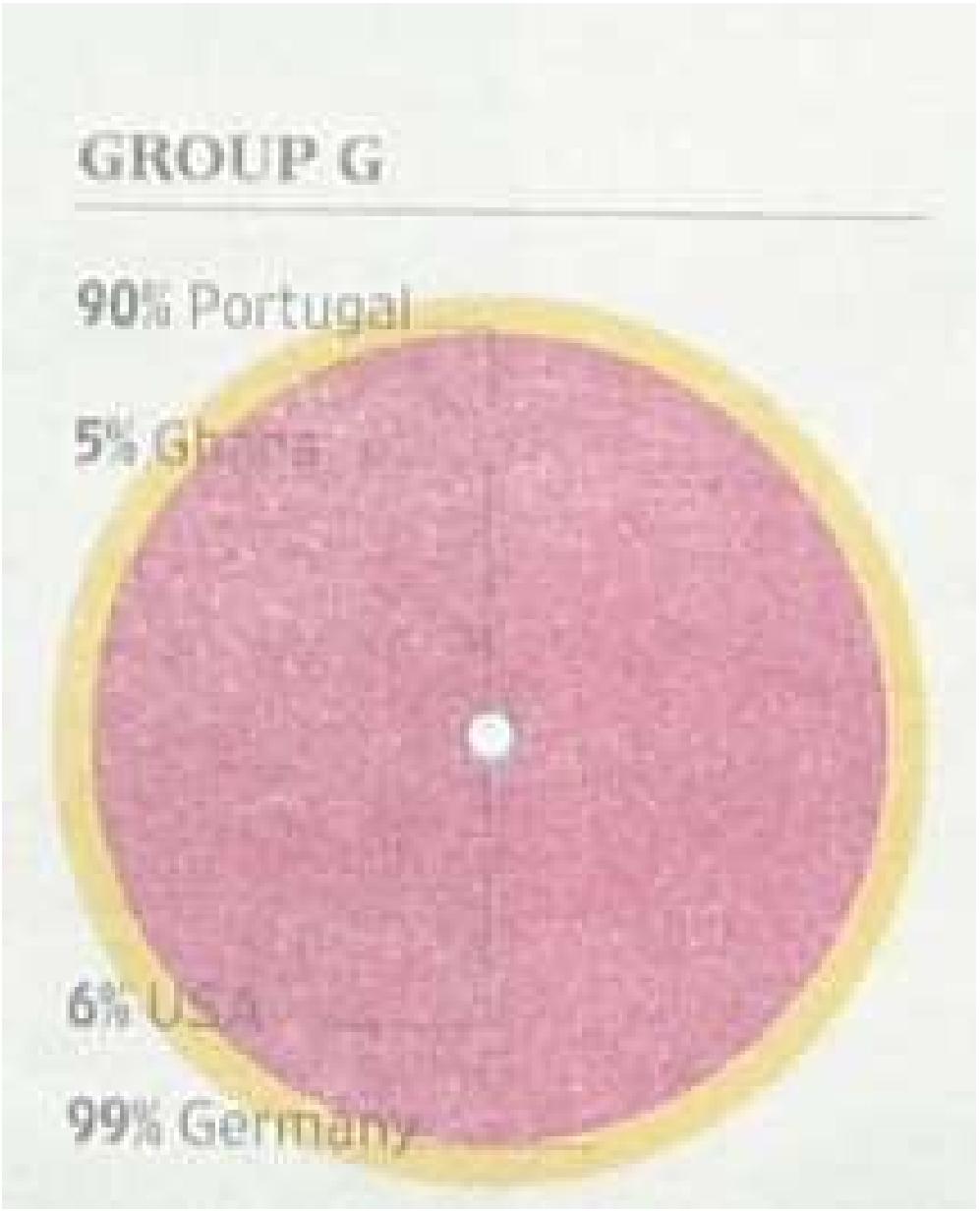




The Times
June 12, 2014



The Times
June 12, 2014



Frequent presentation goals

Clarify

Focus

Highlight

Simplify

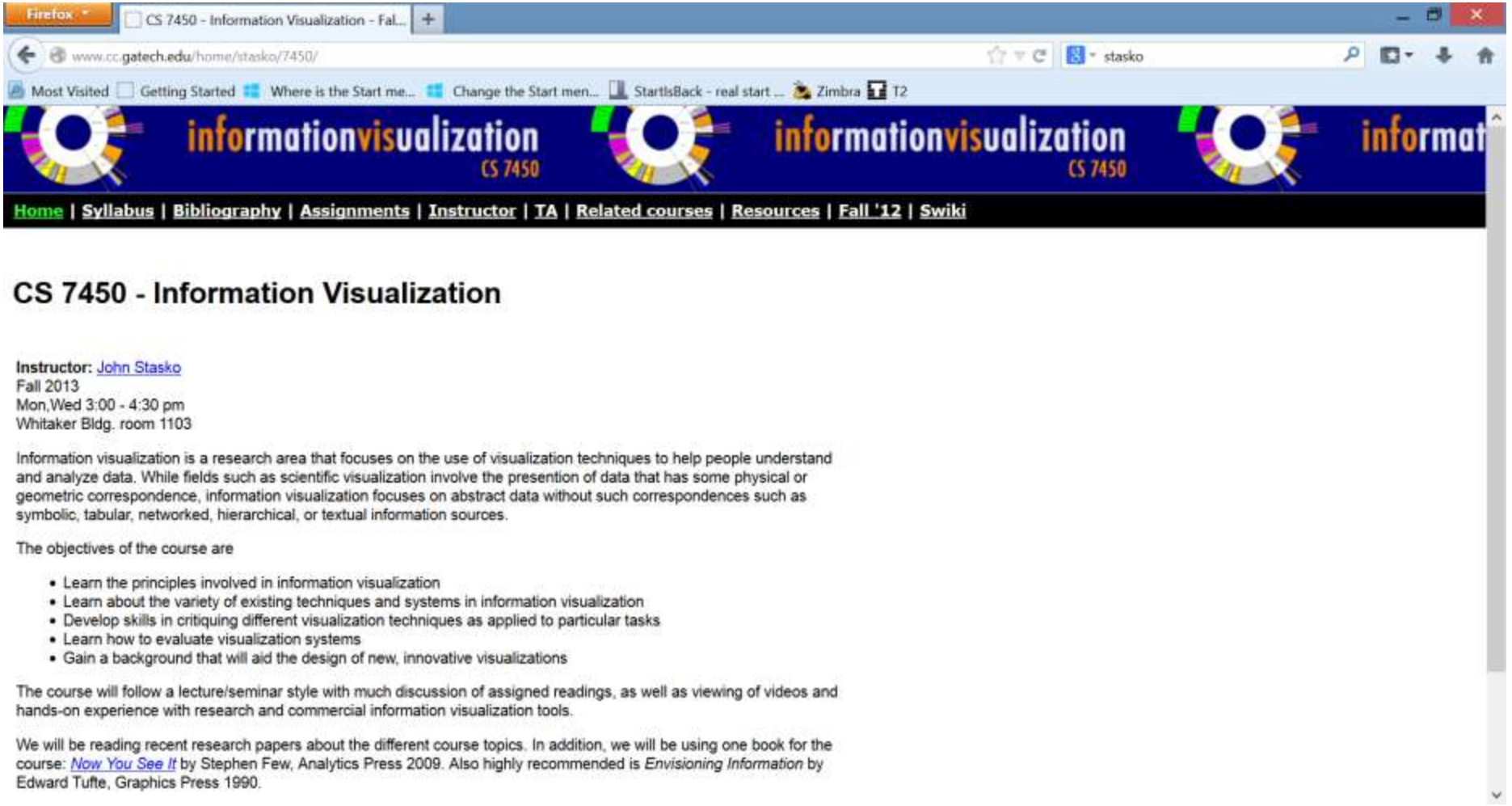
Persuade

May just show a few variables
and/or a subset of the data cases

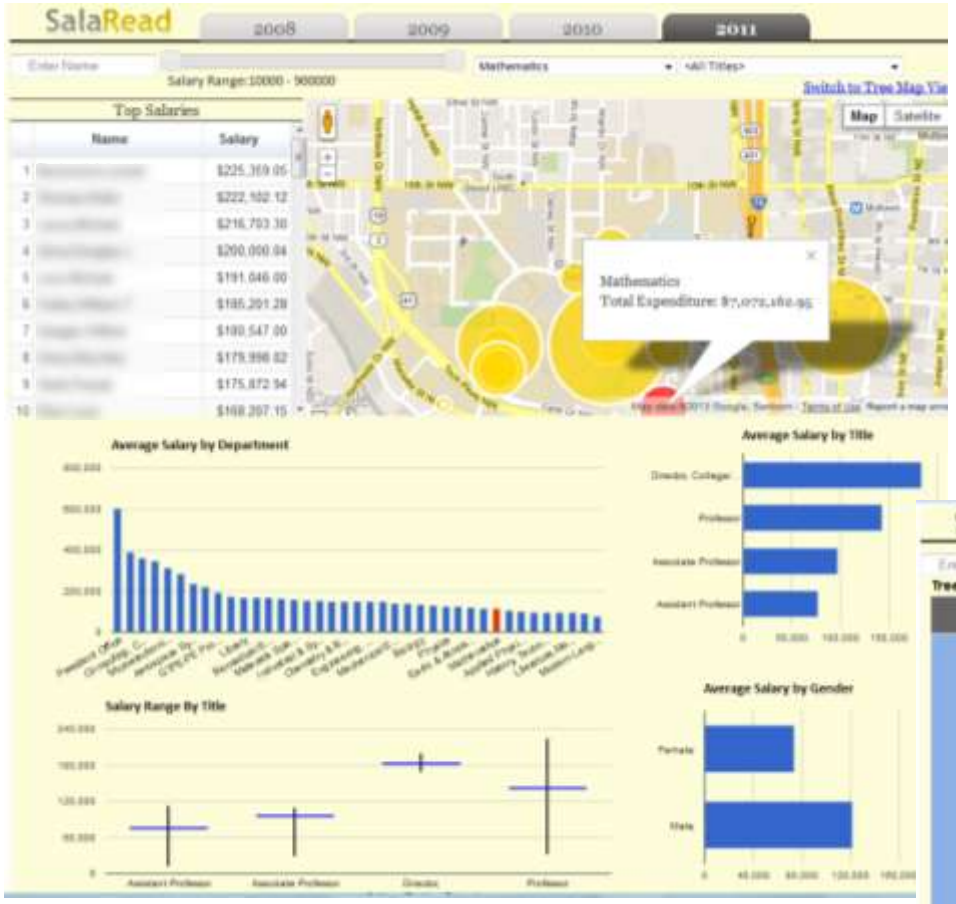
Simply presenting data ***visually*** can have a profound impact

My Class

http://www.cc.gatech.edu/~stasko/7450



CS 7450
Fall ‘12
Design project

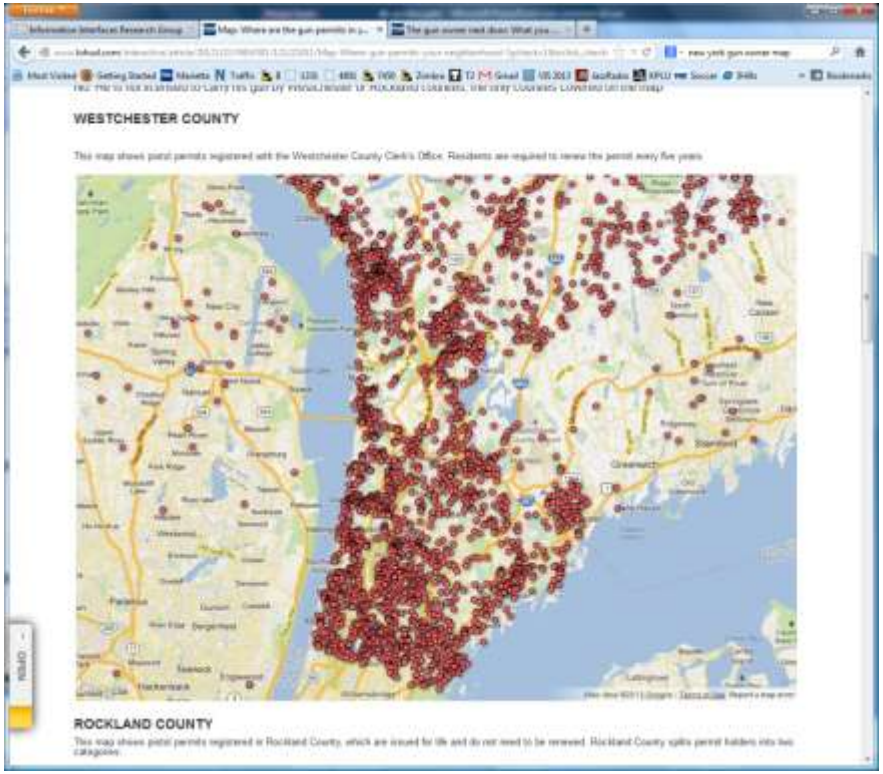




Nate Osborne
Nitya Noronha
Ameya Zambre
Pratik Zaveri

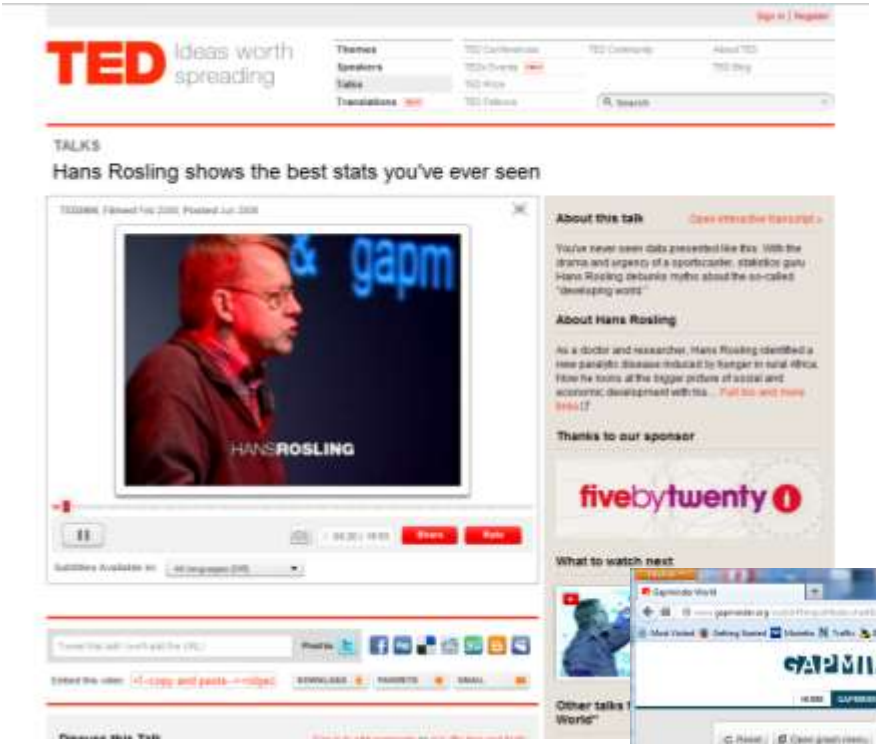
Gun ownership in New York counties





http://www.lohud.com/apps/pbcs.dll/article?AID=2012312230056&nclick_check=1

http://www.lohud.com/interactive/article/20121223/NEWS01/121221011/
Map-Where-gun-permits-your-neighborhood-?gcheck=1&nclick_check=1



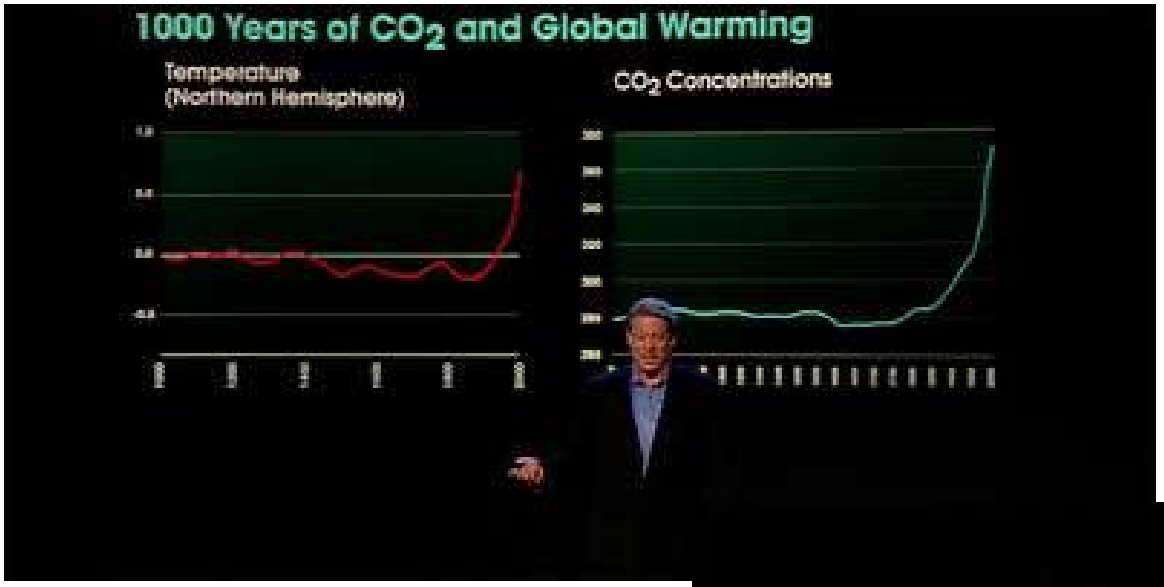
Hans Rosling
Gapminder

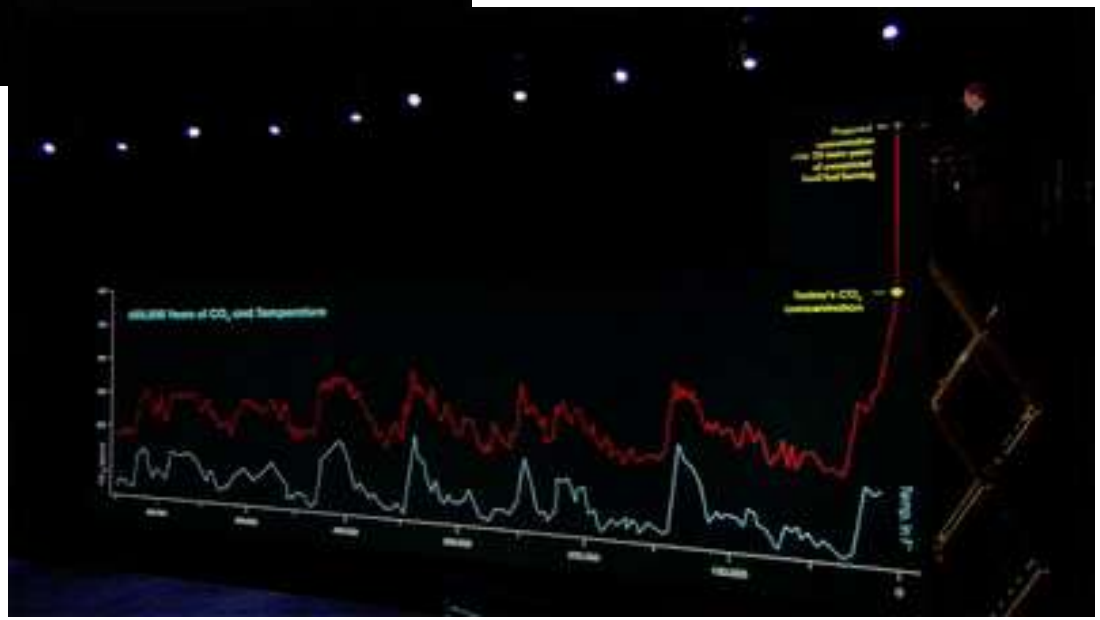


2006

http://www.ted.com/index.php/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen.html

An Inconvenient Truth





Gore made extensive use of data graphics



2. Analysis

Explore the data

Assess a situation

Determine how to proceed

Decide what to do

Many Data Analysis Approaches

Statistics

Database & information retrieval

Data mining

Machine learning





“Contained within the data of any investigation is information that can yield conclusions to questions not even originally asked. That is, there can be surprises in the data…To regularly miss surprises by failing to probe thoroughly with visualization tools is terribly inefficient because the cost of intensive data analysis is typically very small compared with the cost of data collection.”

W. Cleveland
The Elements of Graphing Data

Frequent analysis goals

Show many variables

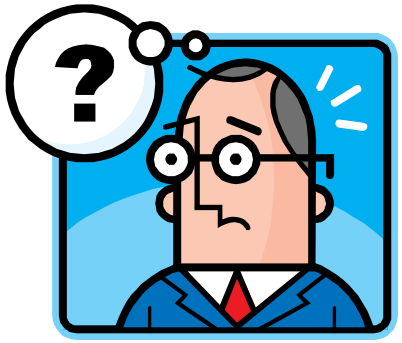
Illustrate overview and detail

Facilitate comparison

Display may not be easy to interpret at first

Visualization most useful in **exploratory data analysis**

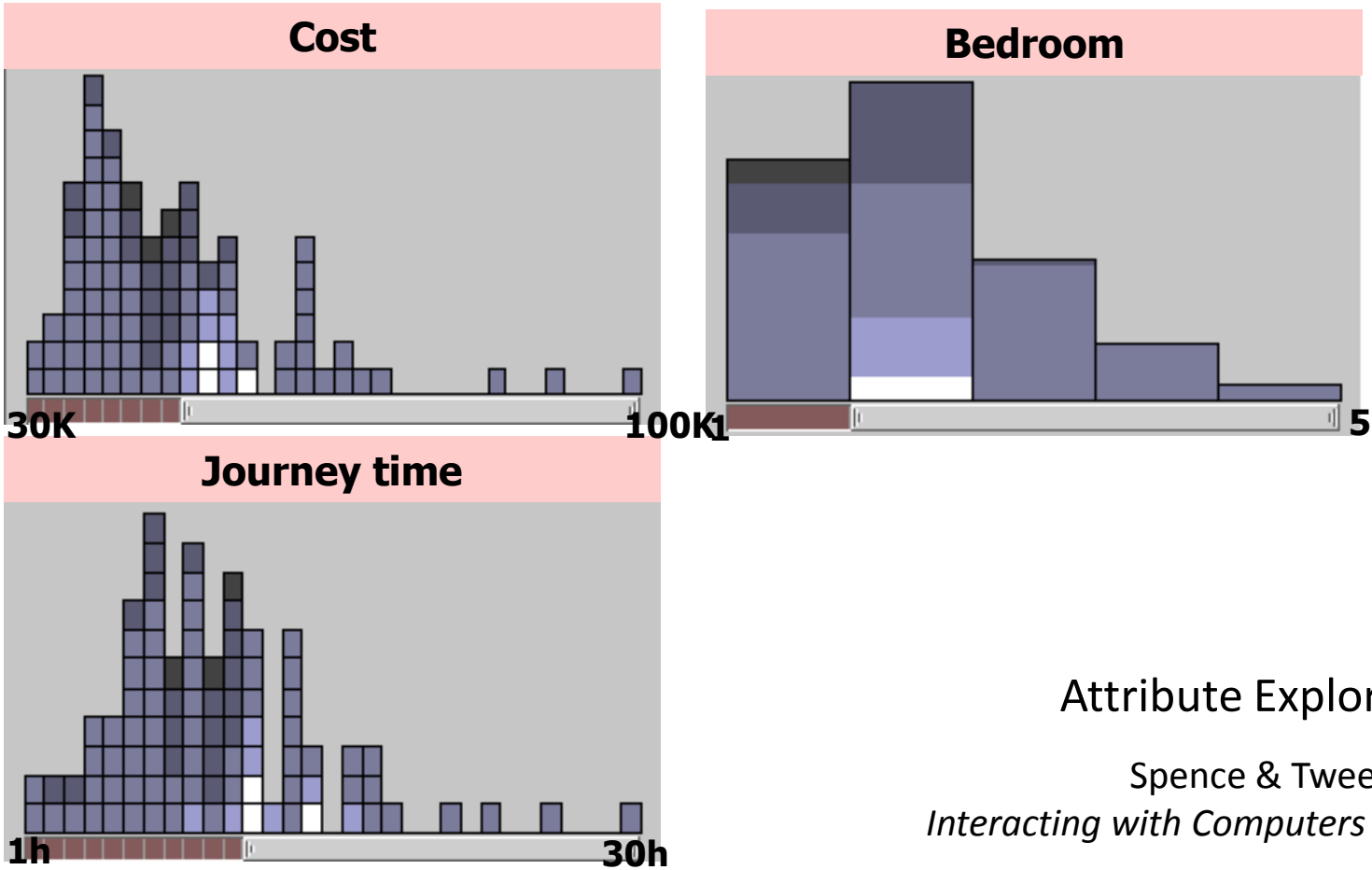
- Don’t know what you’re looking for
- Don’t have a priori questions
- Want to know what questions to ask



At Its Heart: Uncertainty

Nothing is perfect

Humans trade off choices, compromise



Attribute Explorer

Spence & Tweedie

Interacting with Computers '98

Thought

Even analysis is about effective communication

Visualization's Value?

Visualization's Value?

$$V_{alue} = T + I + E + C$$

Visualization's Value?

$$V_{alue} = T + I + E + C$$

Ability to minimize the total **time** needed to answer a wide variety of questions about the data

(Without formal queries, Interaction really helps)

What kinds of questions?

"Low-level" tasks

Retrieve value
Filter
Compute derived value
Find extremum
Sort
Determine range
Characterize distribution
Find anomalies
Cluster
Correlate

Amar, Eagan & Stasko
InfoVis '95

Visualization's Value?

$$V_{alue} = T + \mathbf{I} + E + C$$

Ability to spur and discover **insights** or insightful questions about the data

(Would be very difficult with only the data)

What is Insight?

An individual observation about the data by the participant, a unit of discovery

Complex
Deep
Qualitative
Relevant
Unexpected

Saraiya, North, & Duca
TVCG '05

North
IEEE CG&A '06

What is Insight?

Sudden grasp of new relationships that are necessary to solve a problem and that were not learned in the past

Bernstein, Penner, Clarke-Stewart & Roy
Psychology, 6th edition

What is Insight?

Is not spontaneous “aha!” moments (eg, in cognitive science)
Is knowledge-building and model-confirmation
    Like a substance that people acquire with the aid of systems

Chang, Ziemkiewicz, Green, & Ribarsky
IEEE CG&A ‘09

Visualization's Value?

$$V_{alue} = T + I + E + C$$

Ability to convey an overall **essence** or take-away sense of the data

(The big picture: Whole is greater than the sum of the parts)

Overview and detail

Focus + context

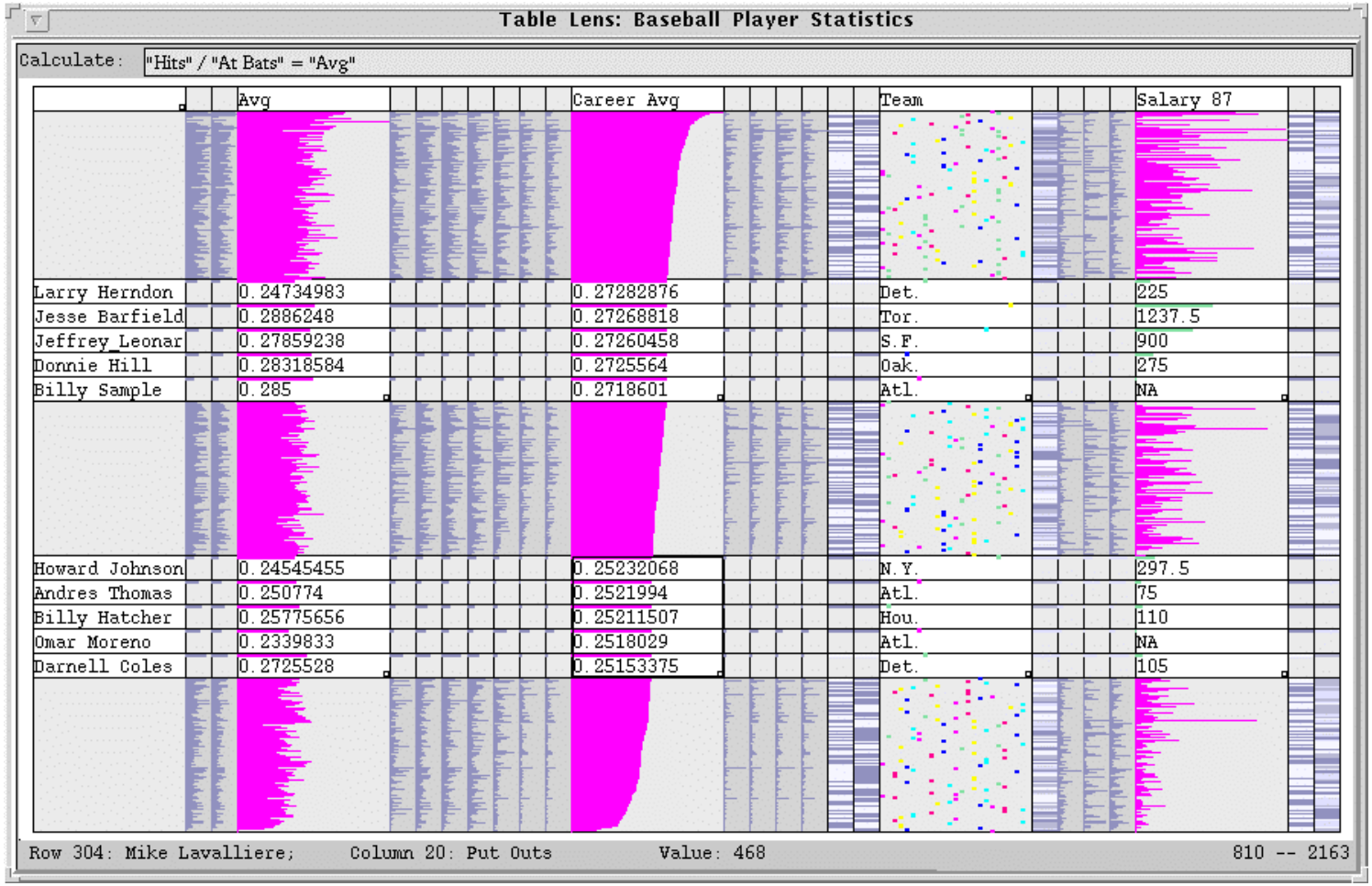


Table Lens

Rao & Card
CHI ‘94

Visualization's Value?

$$V_{alue} = T + I + E + C$$

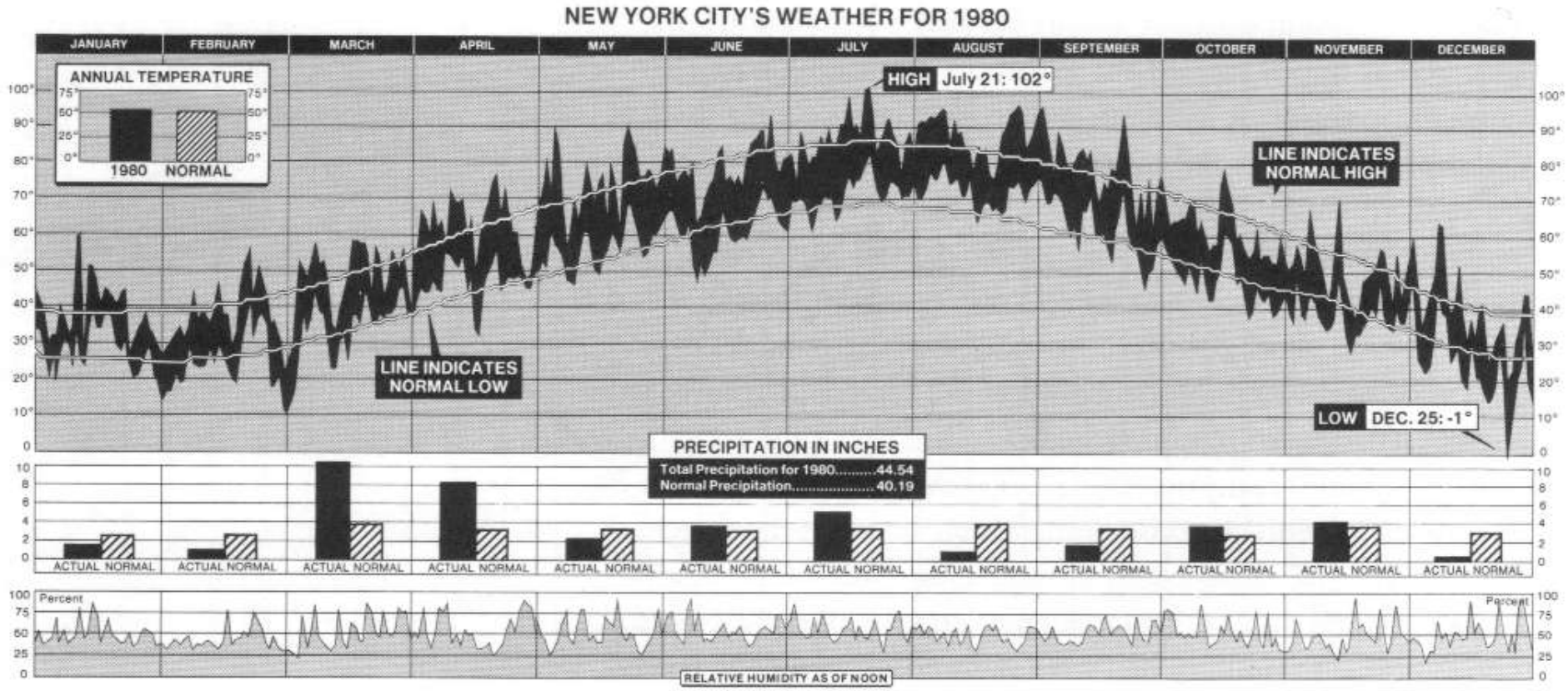
Ability to generate **confidence** and trust about the data, its domain and context

(Beneficial data analysis process side effects)

Some examples

1.

2220 numbers



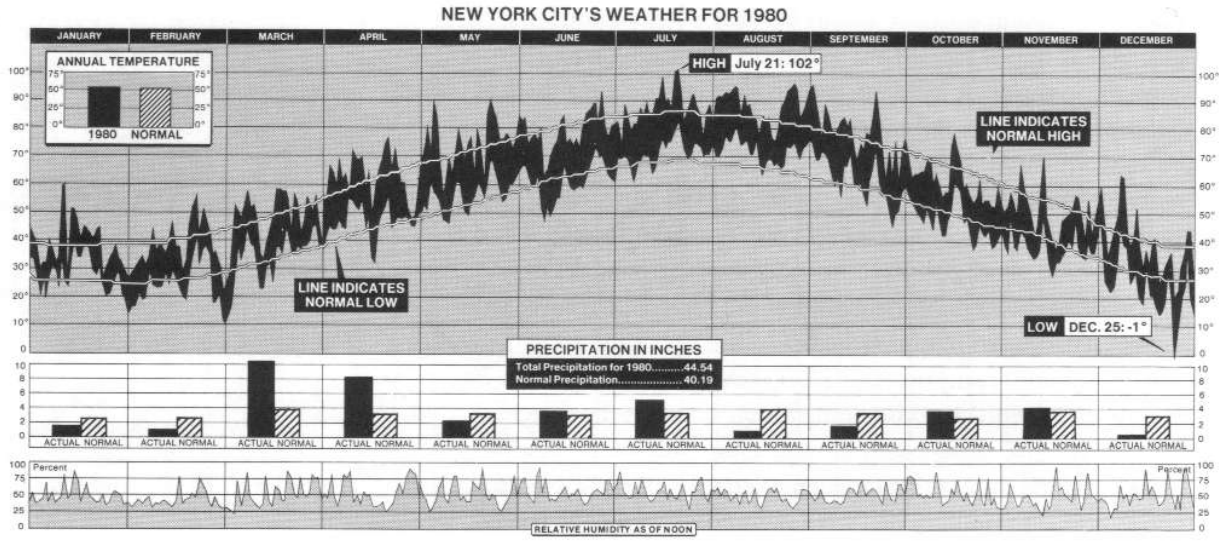
New York Times, January 11, 1981, p. 32.

E. Tufte
The Visual Display of Quantitative Information
1983

Data Values

365	High temp for each day
365	Low temp for each day
365	Avg high temp for each day
365	Avg low temp for each day
365	Precipitation for each day
365	Humidity for each day
12	Precipitation for each month
12	Avg precipitation for each month
1	Precipitation for the year
1	Avg precipitation per year
1	Highest temp (& day) for the year
1	Lowest temp (&day) for the year
1	Avg daily temp for the year
1	Avg daily temp per year

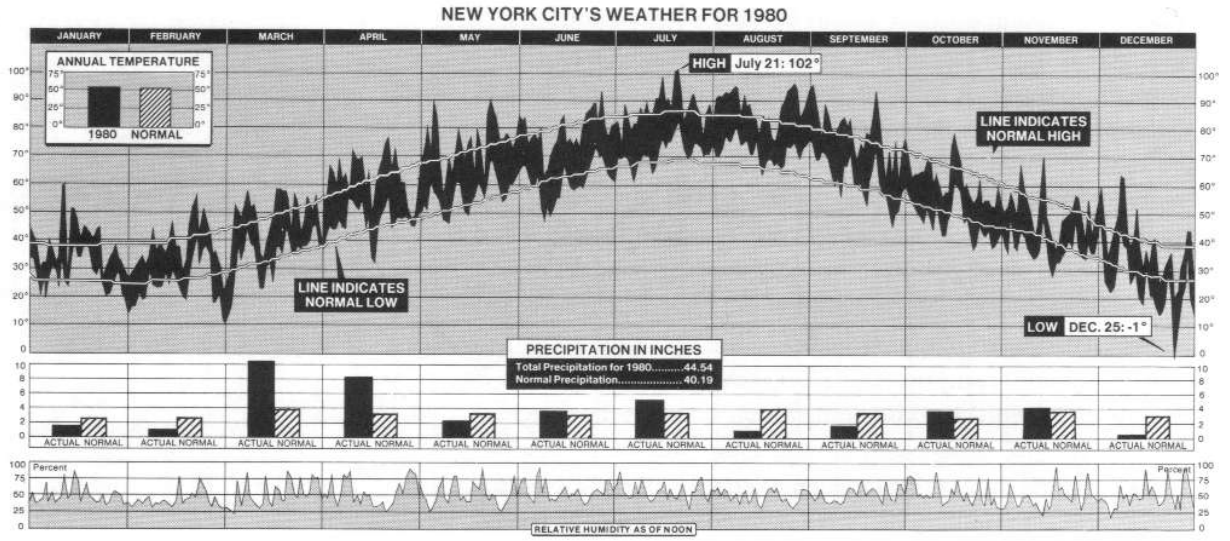
Visualization's Value



New York Times, January 11, 1981, p. 32.

$$V_{alue} = T + I + E + C$$

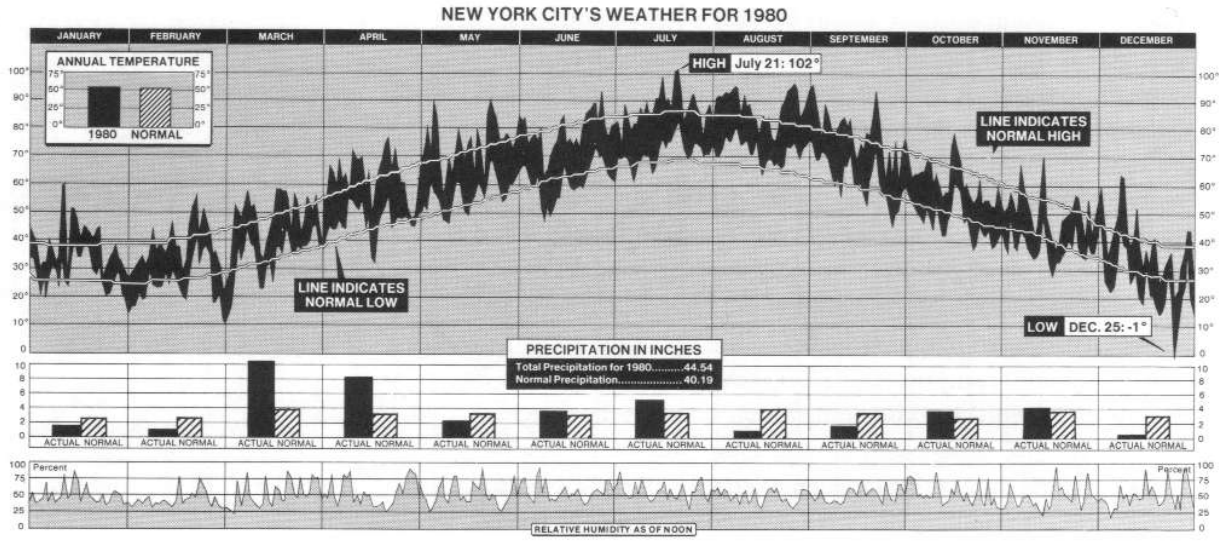
Visualization's Value



New York Times, January 11, 1981, p. 32.

$$V_{alue} = T + I + E + C$$

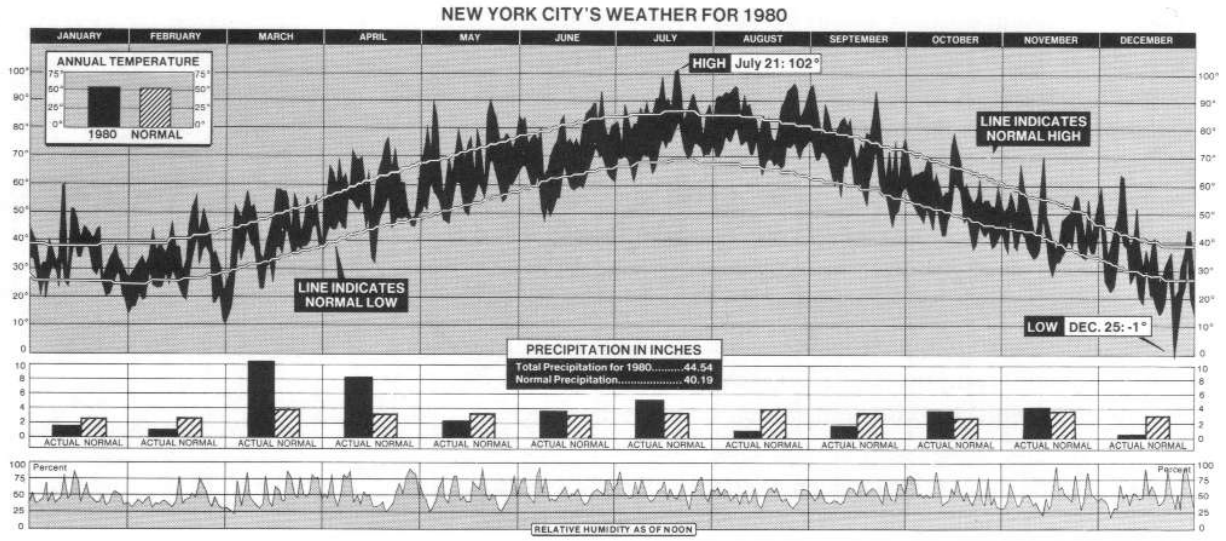
Visualization's Value



New York Times, January 11, 1981, p. 32.

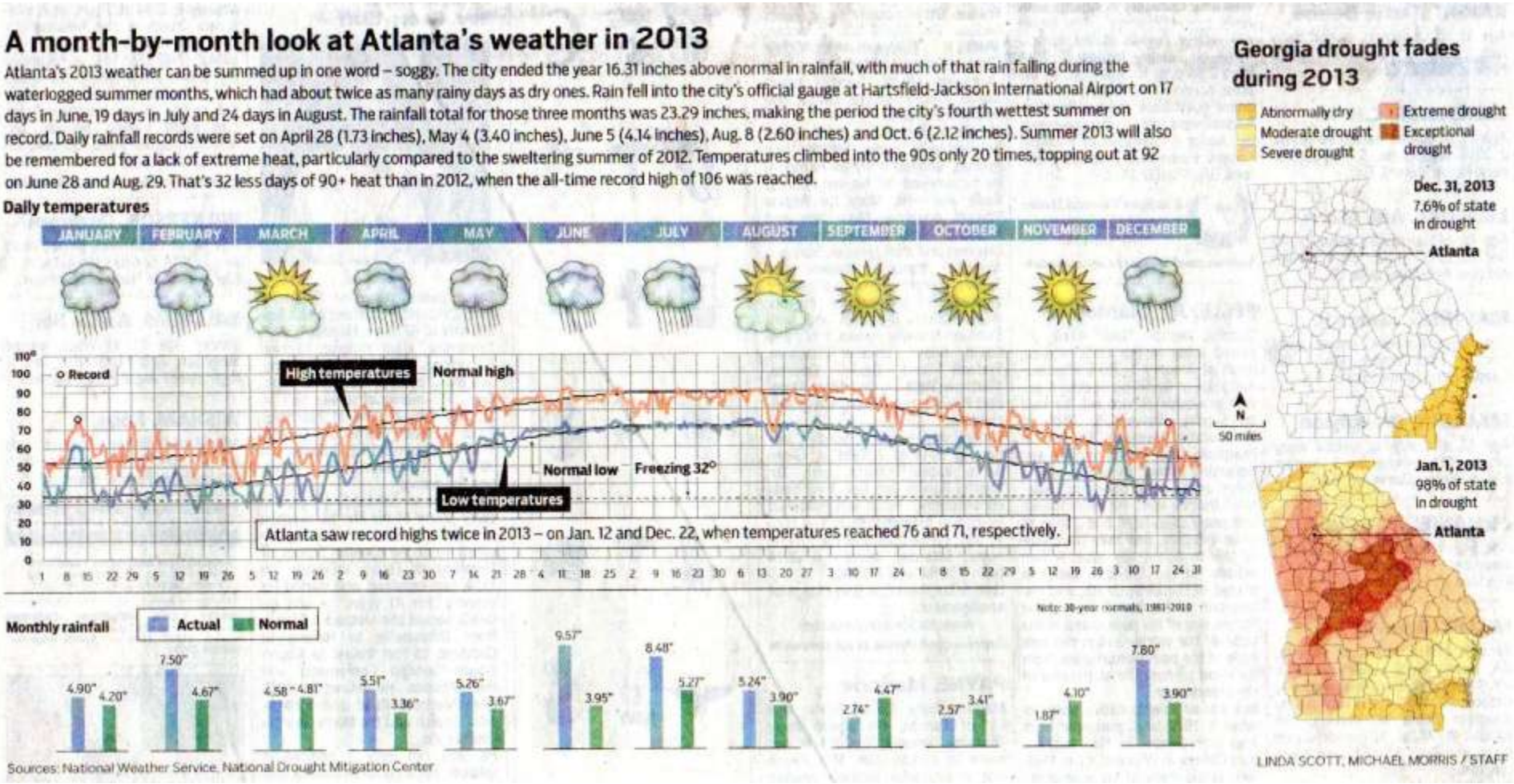
$$V_{alue} = T + I + E + C$$

Visualization's Value



New York Times, January 11, 1981, p. 32.

$$V_{alue} = T + I + E + C$$

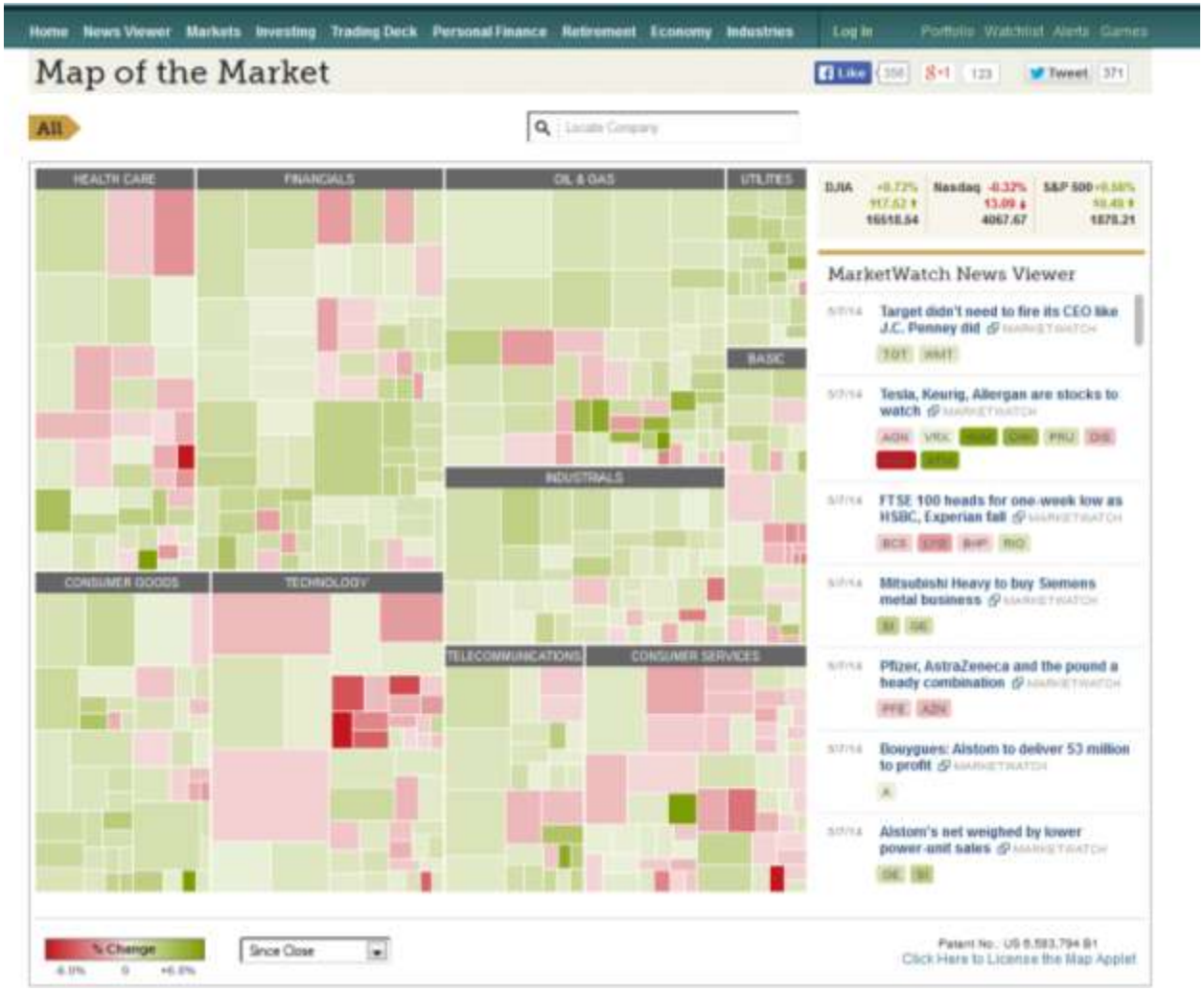


Atlanta Journal Constitution
Jan. 3, 2014

2.

http://www.marketwatch.com/tools/stockresearch/marketmap

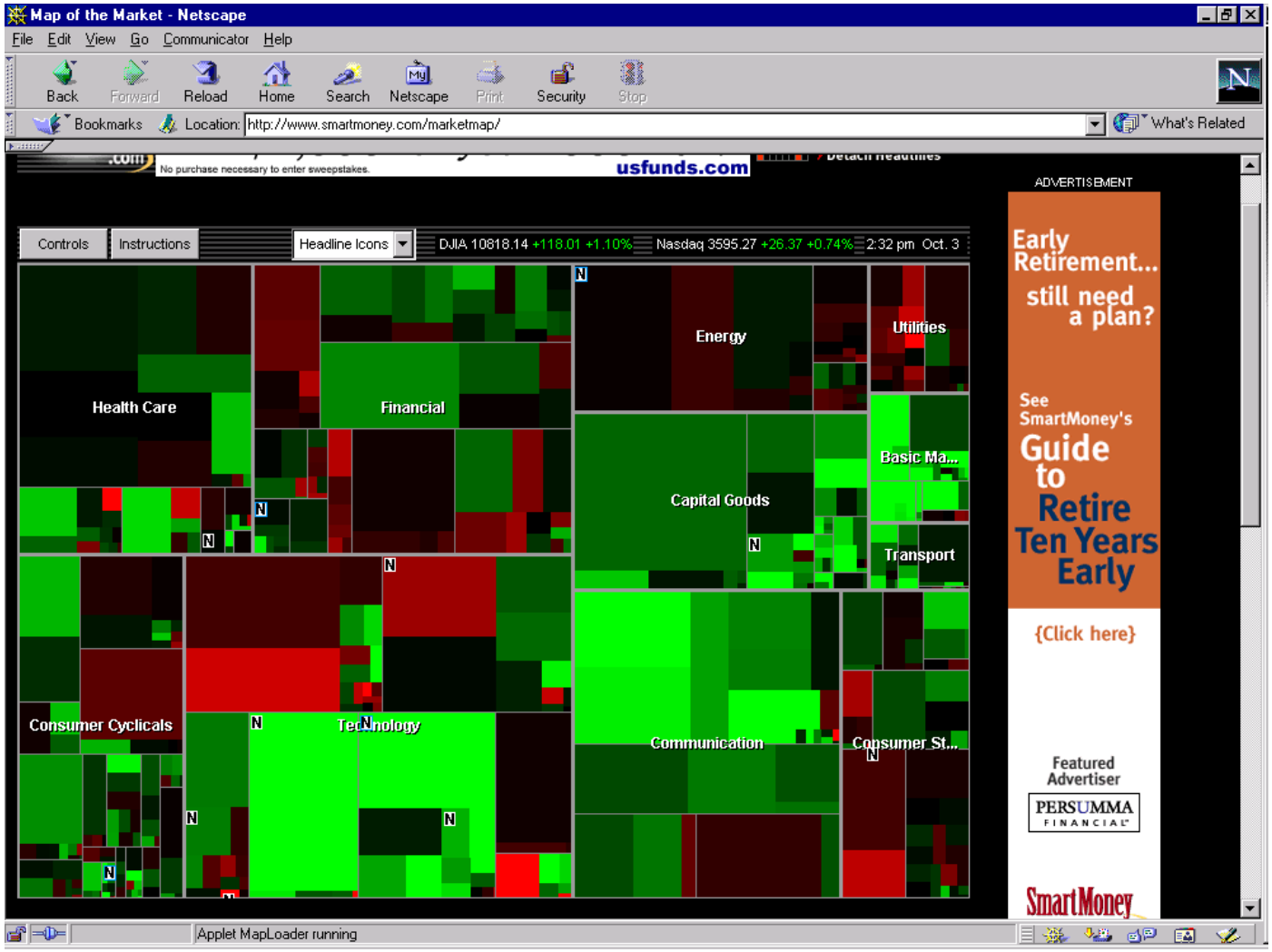
Map of the Market



http://www.smartmoney.com/marketmap

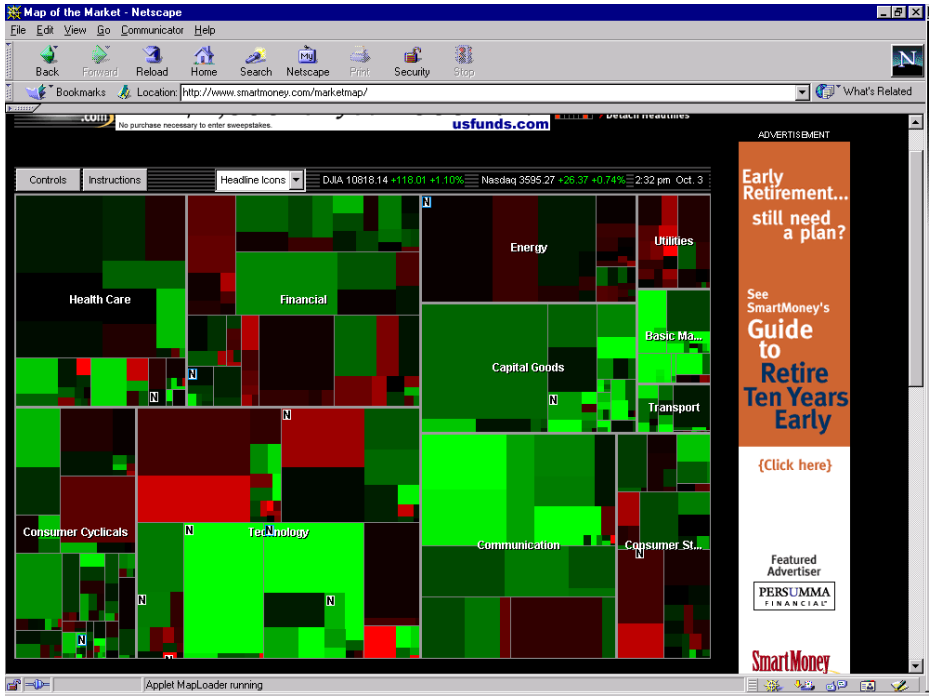
extinct

Map of the Market (old school)



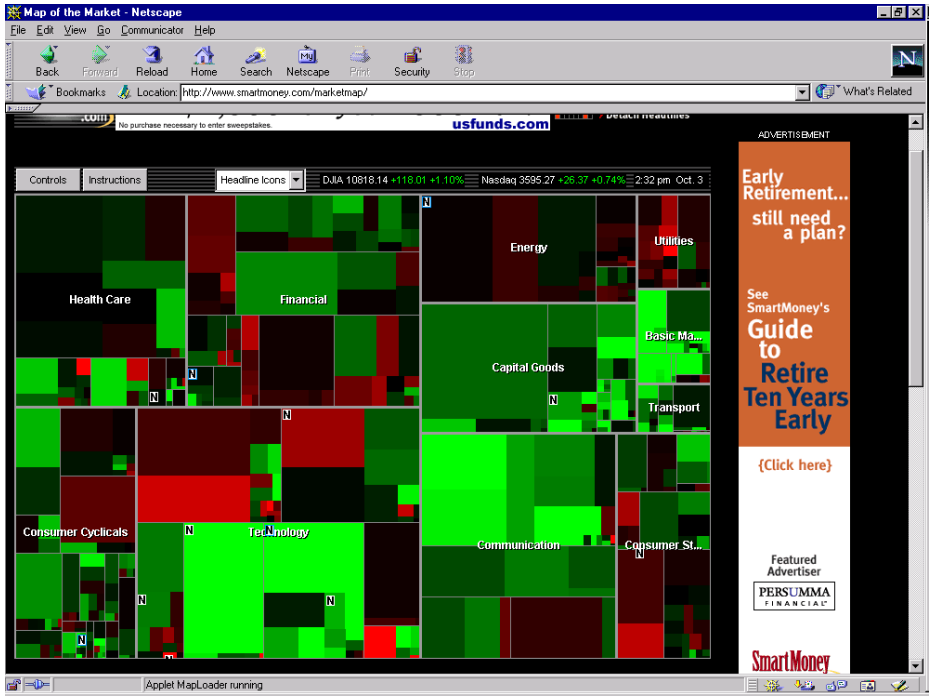
Wattenberg
CHI '99 EA

Visualization's Value



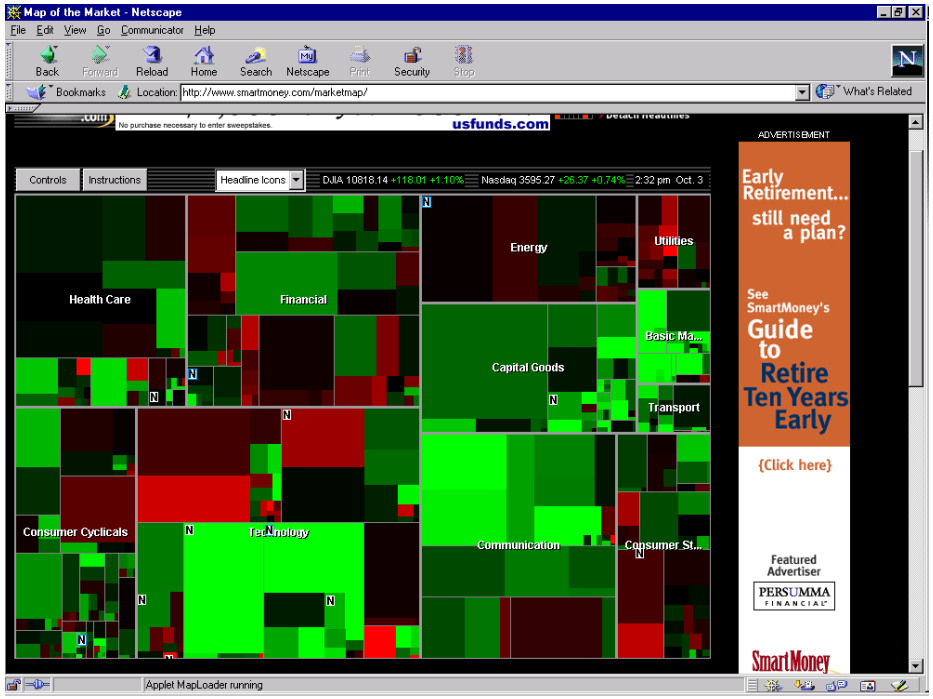
$$V_{alue} = T + I + E + C$$

Visualization's Value

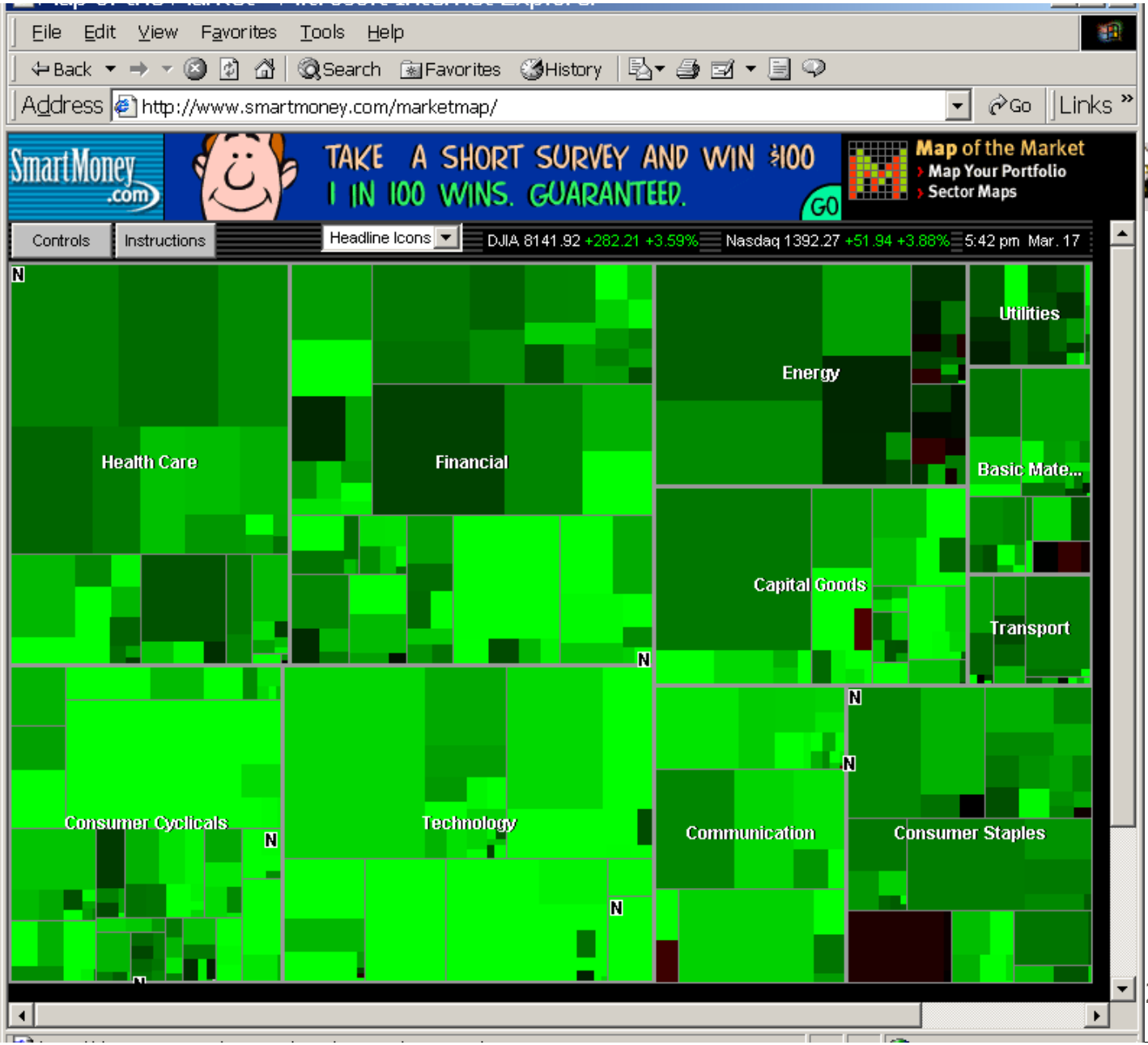


$$V_{alue} = T + I + E + C$$

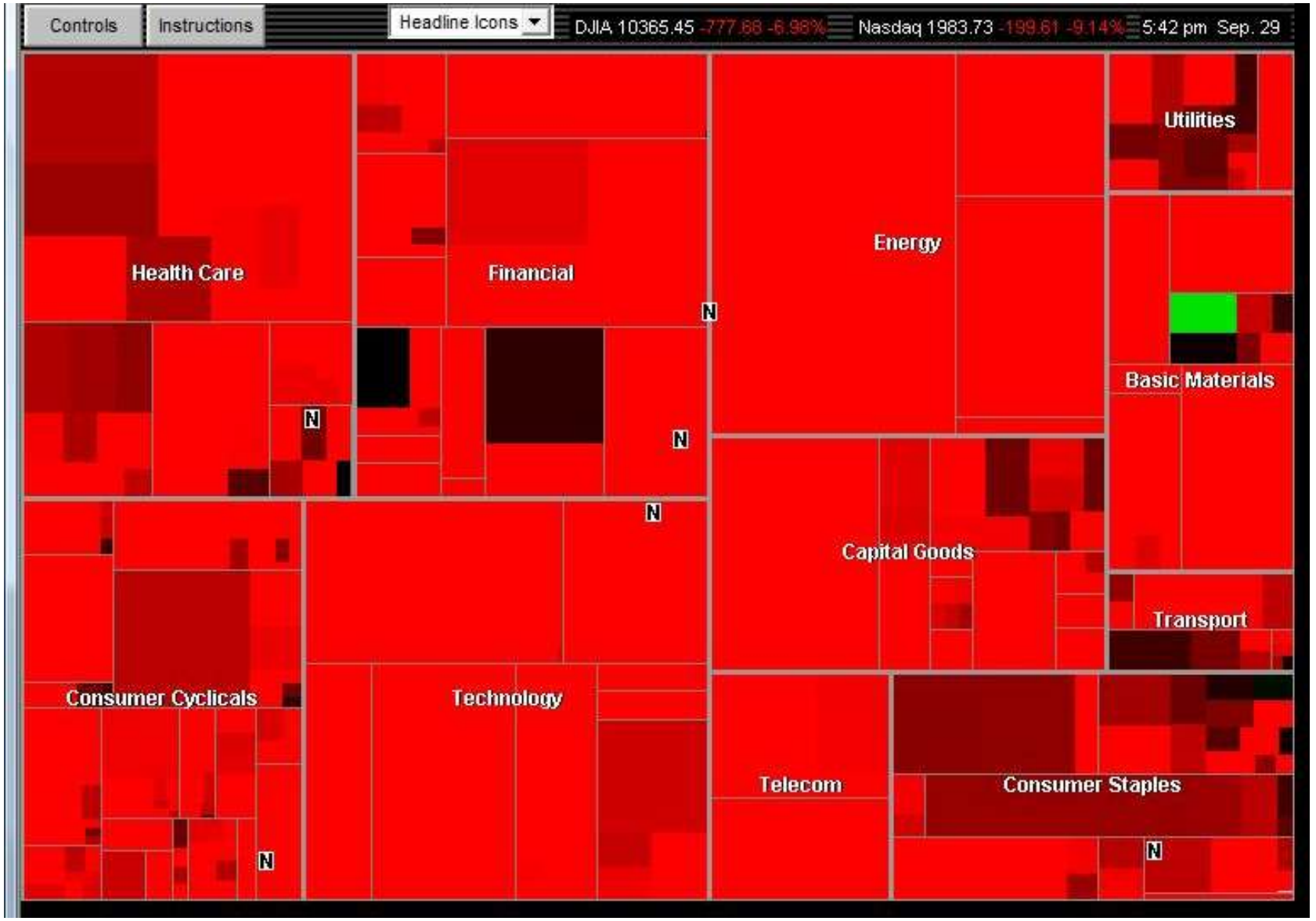
Visualization's Value



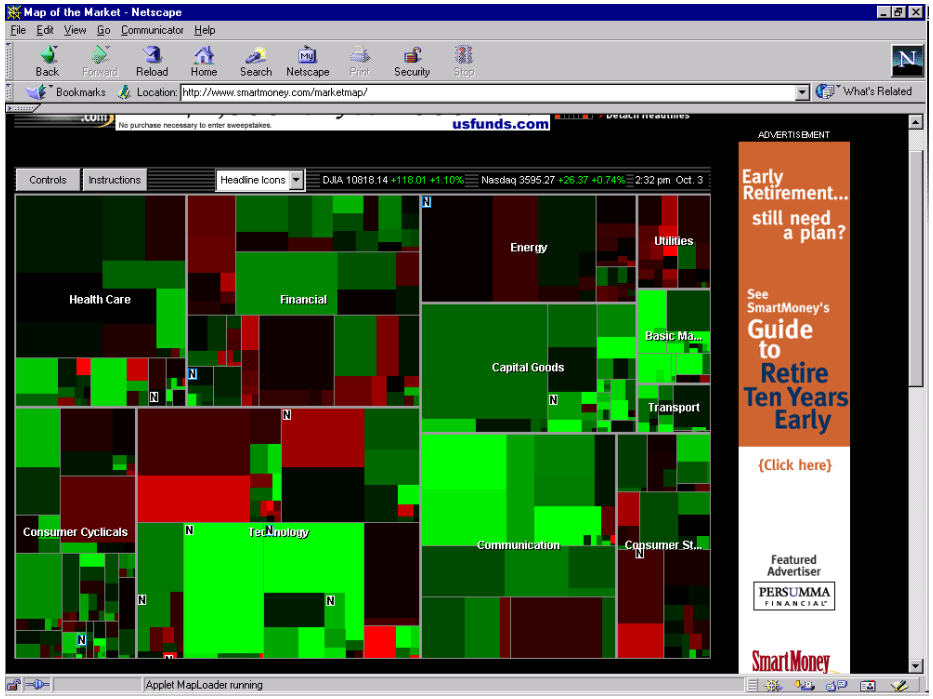
$$V_{alue} = T + I + E + C$$


File Edit View Favorites Tools Help
Back Search Favorites History
Address http://www.smartmoney.com/marketmap/
Go Links
SmartMoney.com
TAKE A SHORT SURVEY AND WIN $100
1 IN 100 WINS. GUARANTEED.
GO
Map of the Market
Map Your Portfolio
Sector Maps
Controls
Instructions
Headline Icons
DJIA 8141.92 +282.21 +3.59%
Nasdaq 1392.27 +51.94 +3.88%
5:42 pm Mar. 17
Health Care
Financial
Energy
Utilities
Basic Mate...
Capital Goods
Transport
Consumer Cyclicals
Technology
Communication
Consumer Staples

Spr 24


Controls
Instructions
Headline Icons
DJIA 10365.45 -777.68 -6.98%
Nasdaq 1983.73 -199.61 -9.14%
5:42 pm Sep. 29
Health Care
Financial
Energy
Utilities
Basic Materials
Capital Goods
Transport
Consumer Cyclicals
Technology
Telecom
Consumer Staples
N

Visualization's Value

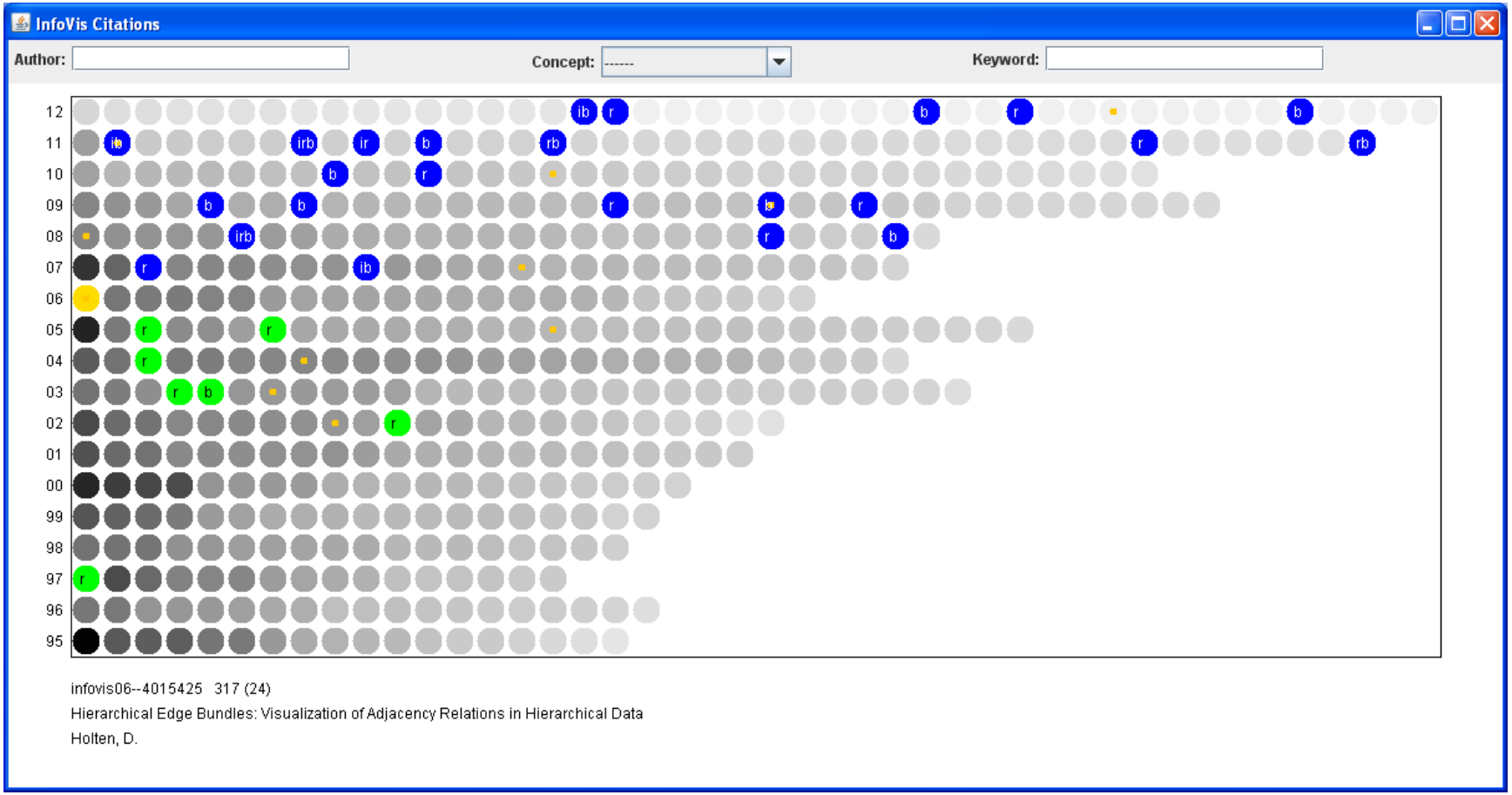


$$V_{alue} = T + I + E + C$$

3.

http://www.cc.gatech.edu/gvu/ii/citevis

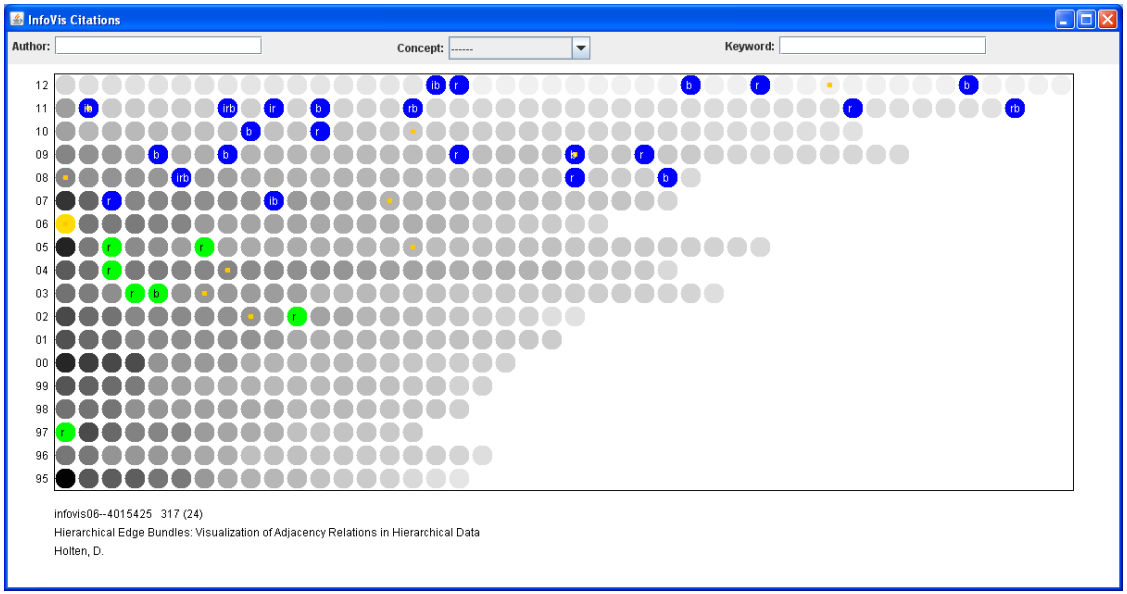
CiteVis



Demo

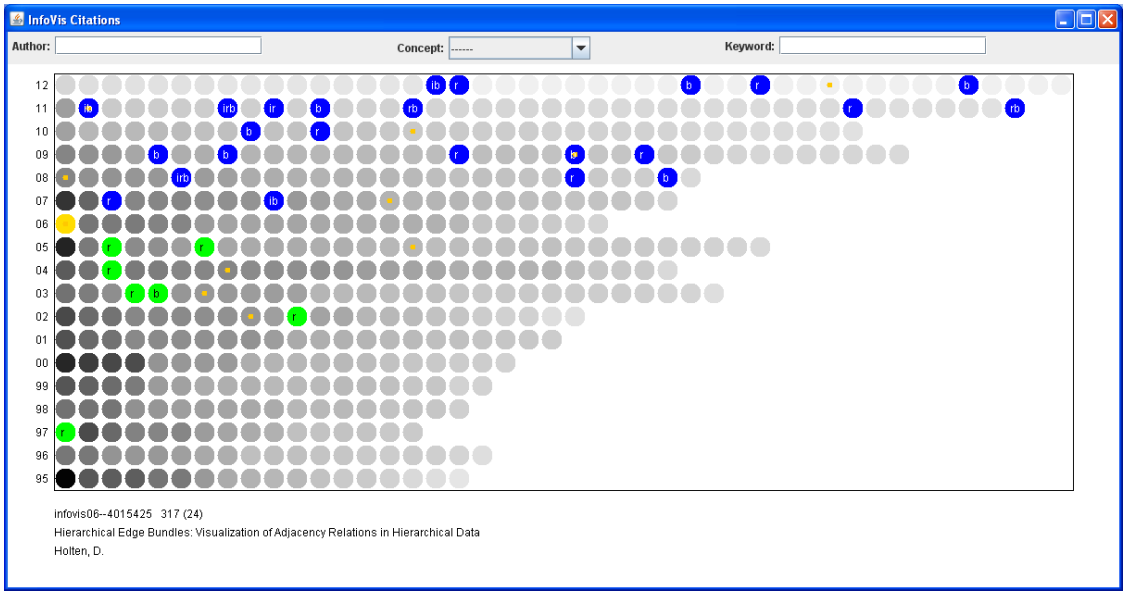
Stasko, Choo, Han, Hu, Pileggi, Sadana & Stolper
InfoVis poster '13

Visualization's Value



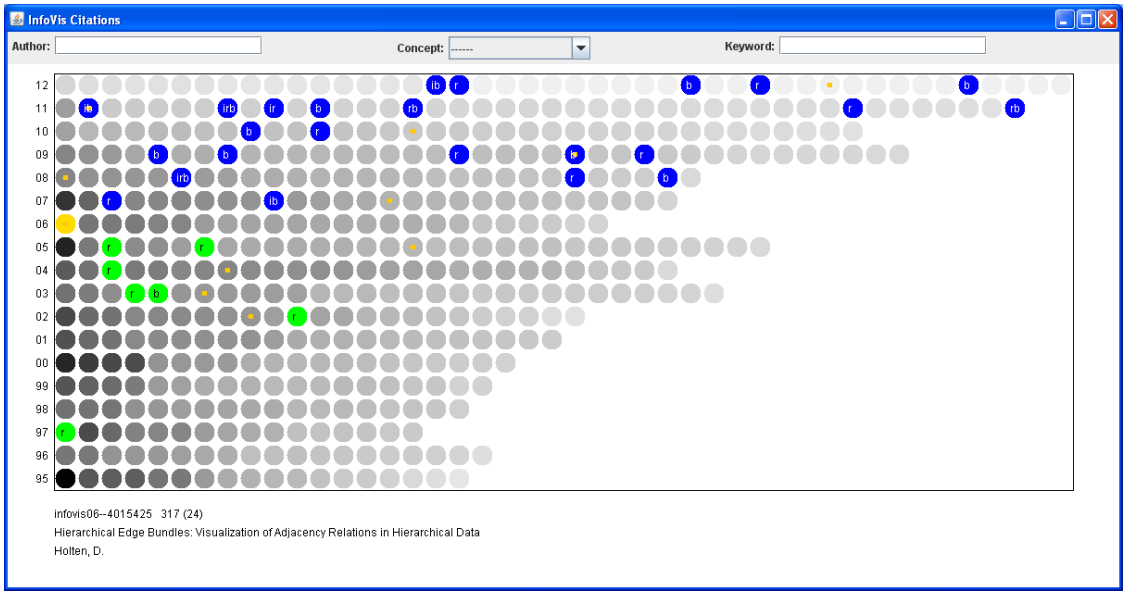
$$V_{alue} = T + I + E + C$$

Visualization's Value



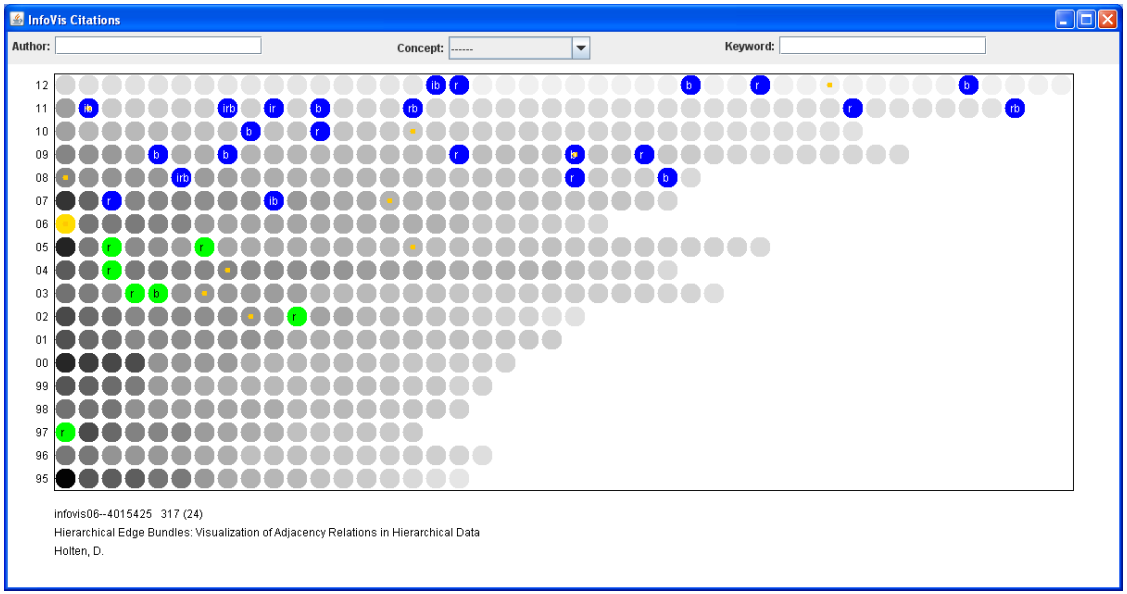
$$V_{alue} = T + I + E + C$$

Visualization's Value



$$V_{alue} = T + I + E + C$$

Visualization's Value



$$V_{alue} = T + I + E + C$$

Value

Not evaluation in traditional sense

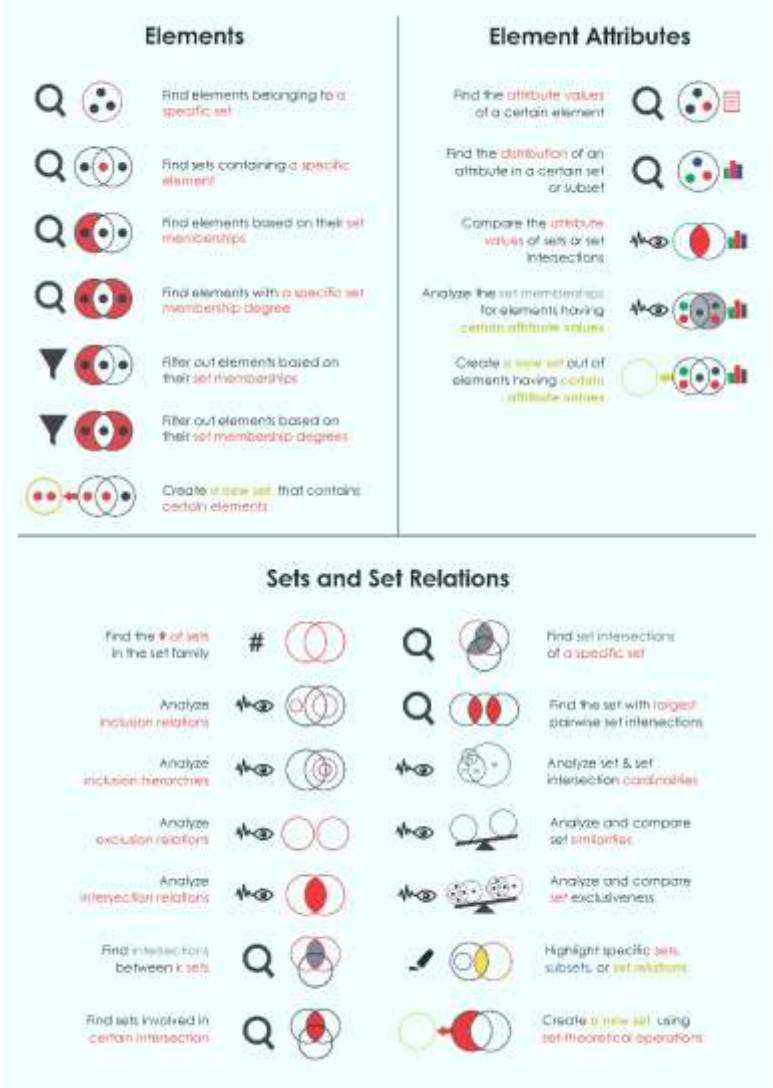
HCI: Benchmark tasks

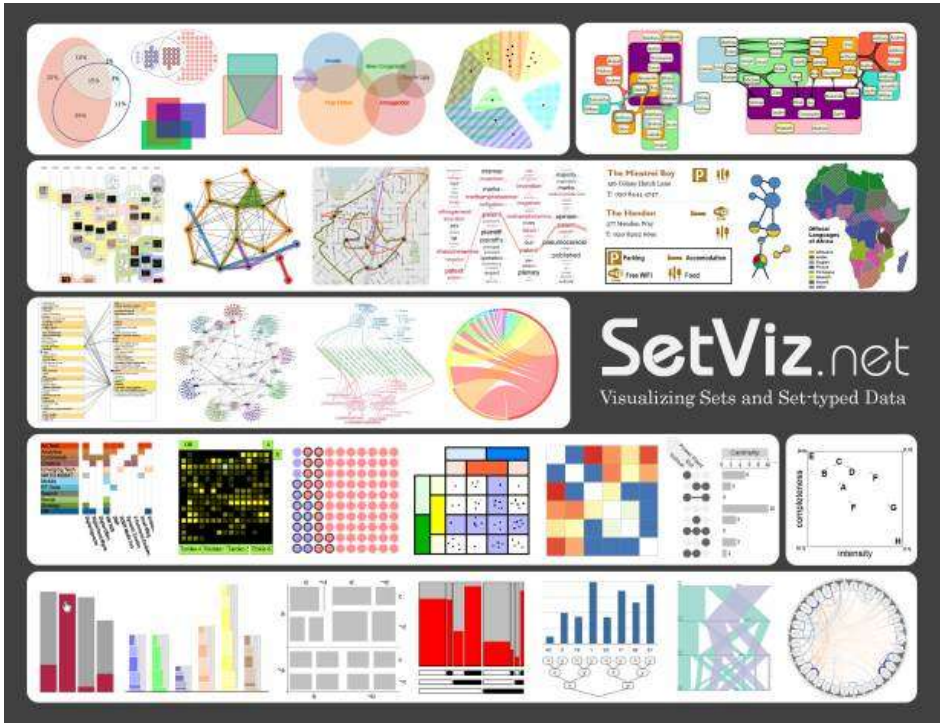
Visualization: Vagueness or absence of specific tasks

More about exploration and understanding



STAR Report: Set Visualization





Alsallakh, Micallef, Aigner, Hauser, Miksch & Rodgers
EuroVis ‘14

Data Visualization 101



Problem:
You have a lot of data (& attributes) to understand

Do you?

- Pack all the data into one complex representation
- Spread the data into multiple coordinated views
- Use interaction to reveal different subsets of the data

Constituents

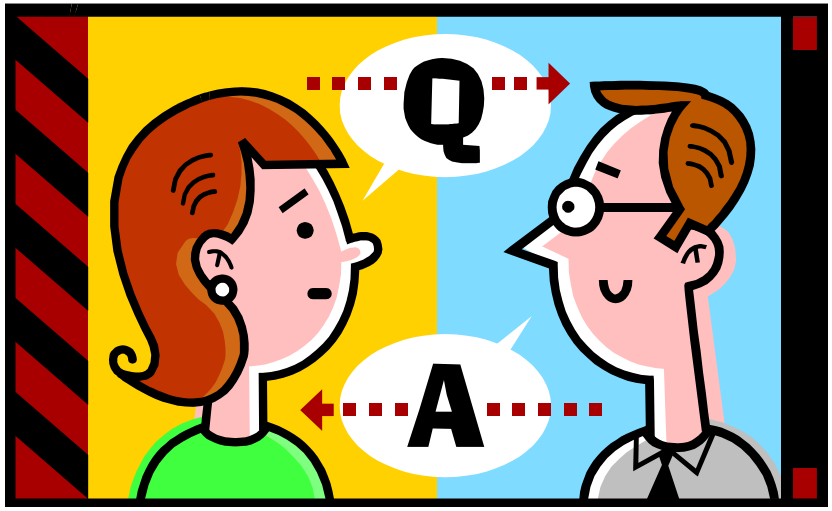
Two key aspects of data visualization

- Representation
- Interaction

> "The effectiveness of information visualization hinges on two things: its ability to clearly and accurately represent information and our ability to interact with it to figure out what the information means."
>
> S. Few, *Now you see it*

Interaction is Vital

Engage in a dialog with your data



Fundamental nature: Equal sibling with representation or subordinate facilitator?

Interaction

Why interact?

Intent

1. Select
2. Explore
3. Reconfigure
4. Encode
5. Abstract/Elaborate
6. Filter
7. Connect

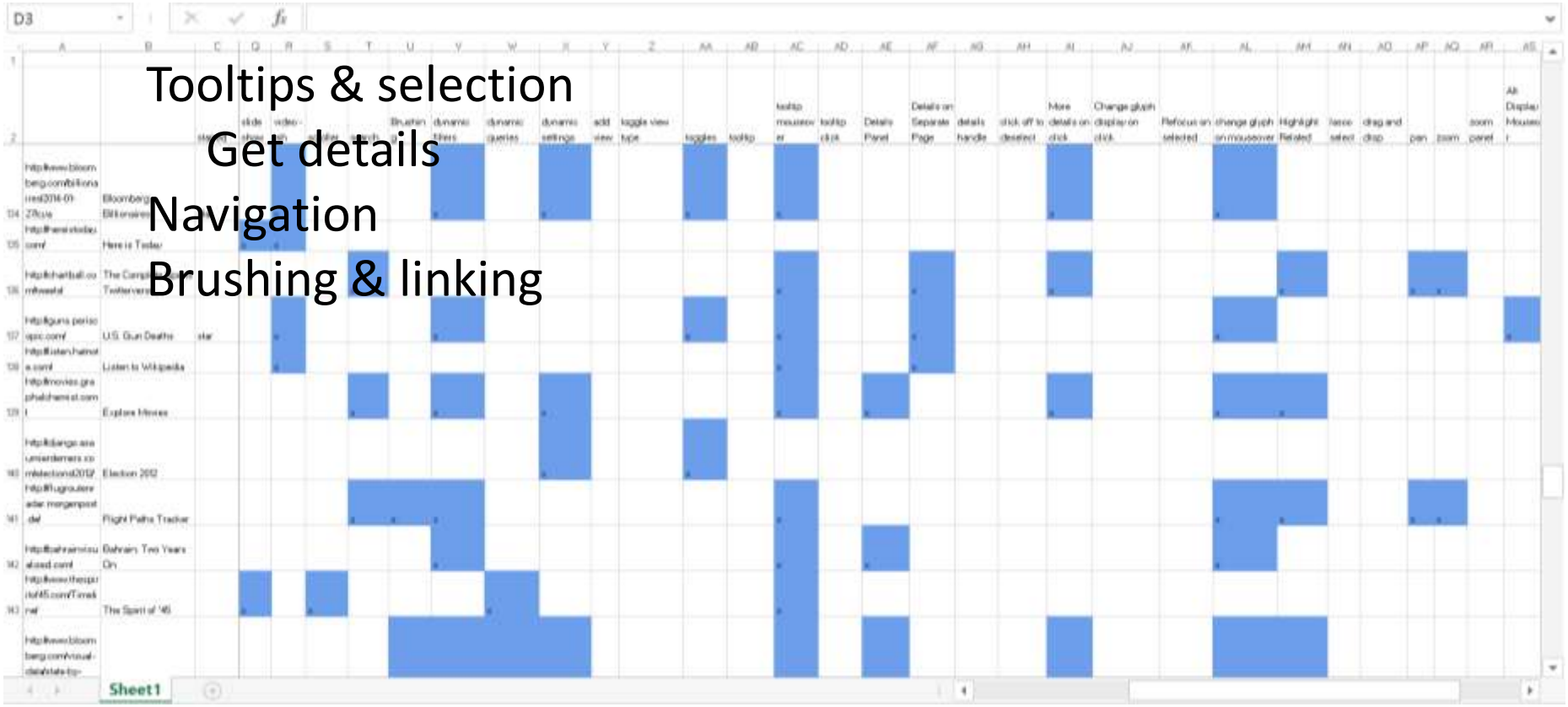
Yi, Kang, Stasko & Jacko
TVCG (InfoVis) '07

Chad Stolper



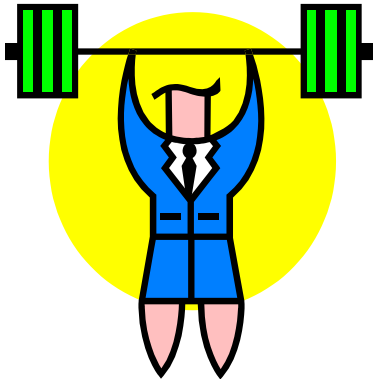
Interaction

How manifested today?



Interaction

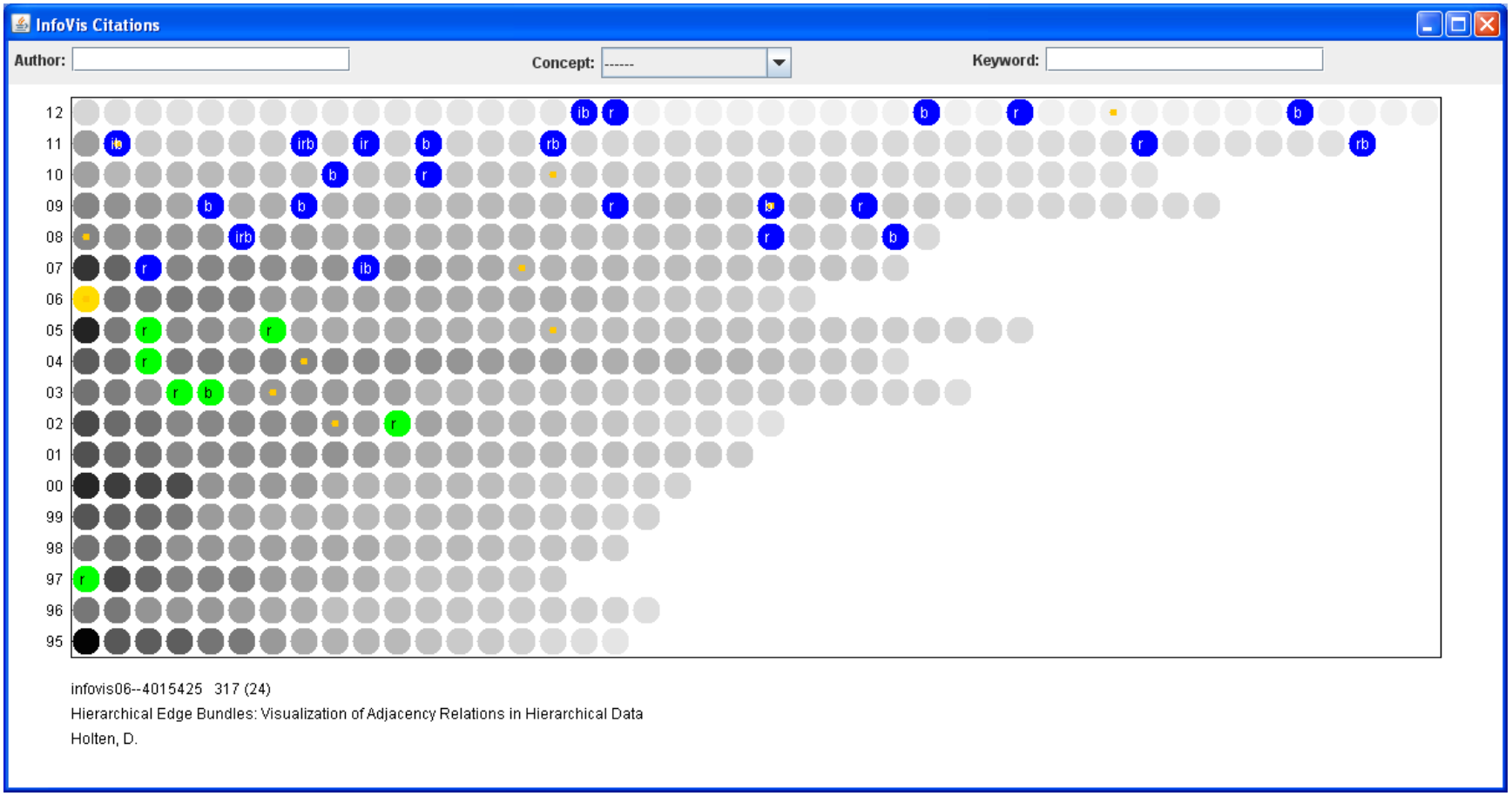
Can we do more?



Employ interaction in a more fundamental manner to strengthen the power of visualization

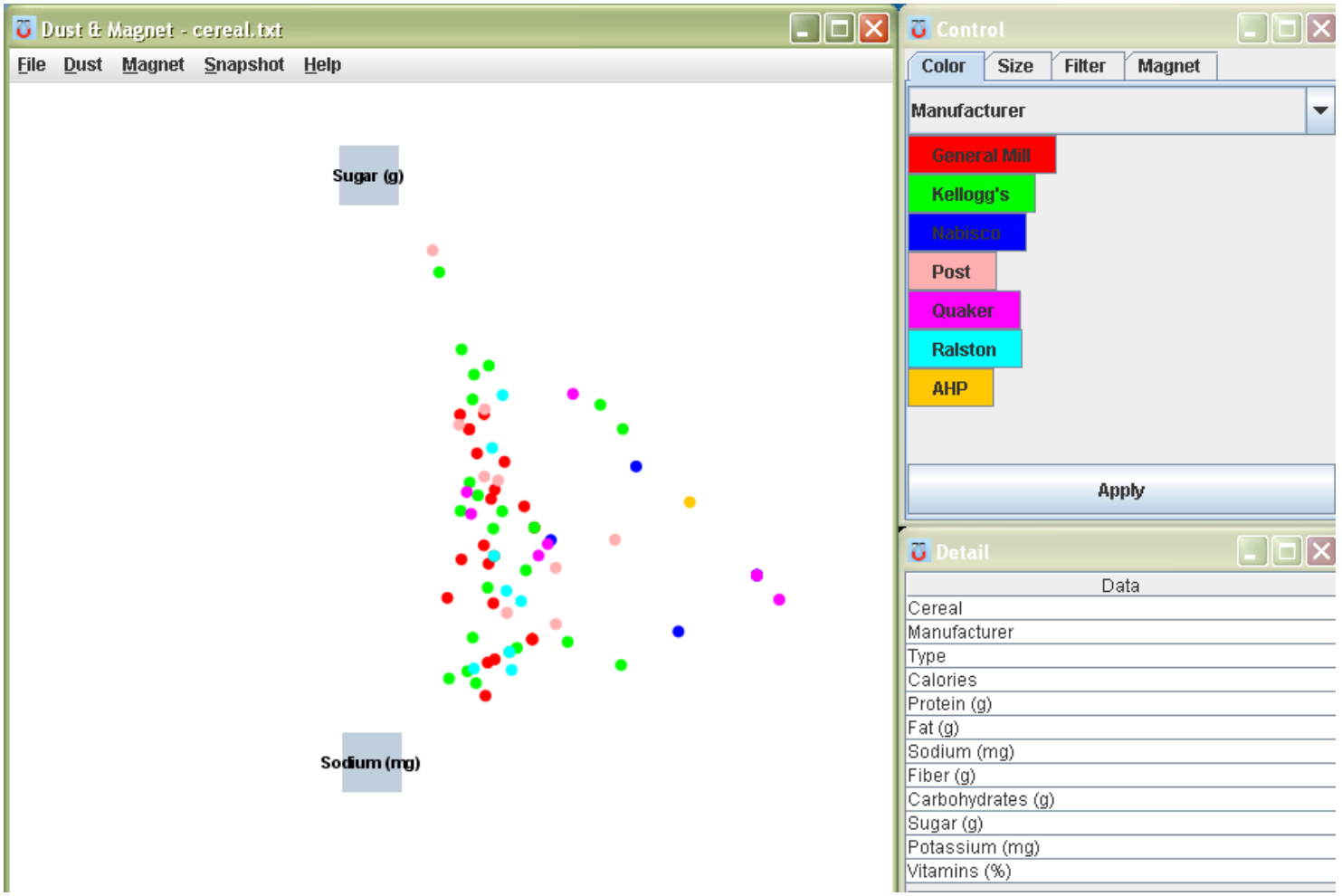
CiteVis

http://www.cc.gatech.edu/gvu/ii/citevis



Stasko, Choo, Han, Hu, Pileggi, Sadana & Stolper
InfoVis poster '13

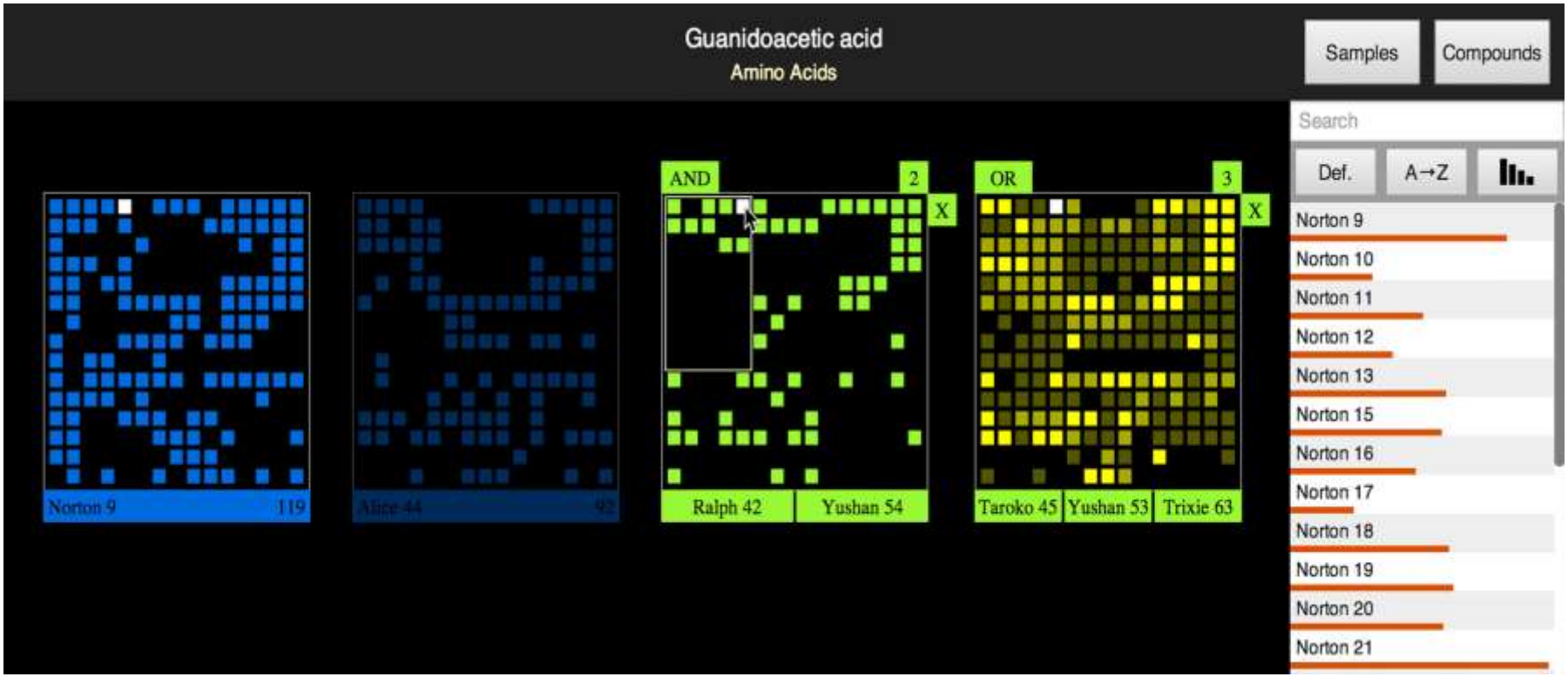
Dust and Magnet



Yi, Melton, Stasko & Jacko
Information Visualization '05

OnSet

http://www.cc.gatech.edu/gvu/ii/setvis/



Demo

Sadana, Major, Dove, & Stasko

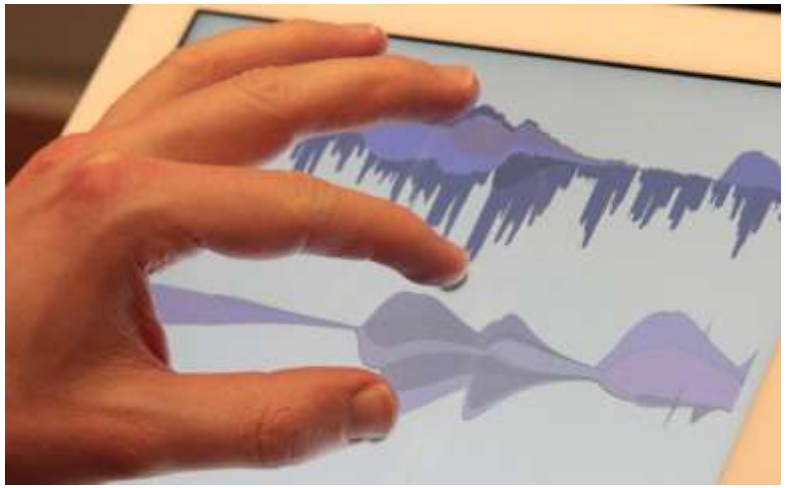
Interaction

What are the tools of interaction?

Traditional – Desktop: keyboard, mouse

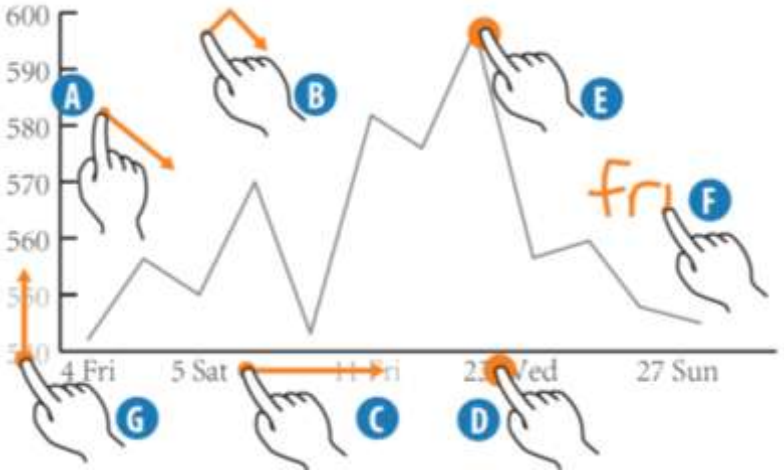
New – Tablet: fingers and multitouch

TouchWave



Baur, Lee & Carpendale
ITS ‘12

Selection Gestures

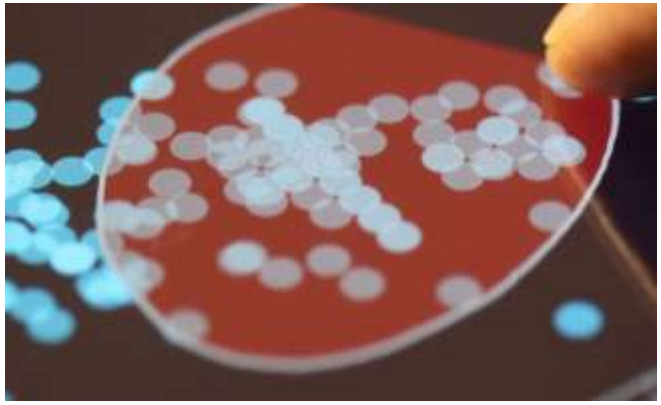


Willett, Lan & Isenberg
EuroVis ‘14

AVI 2014

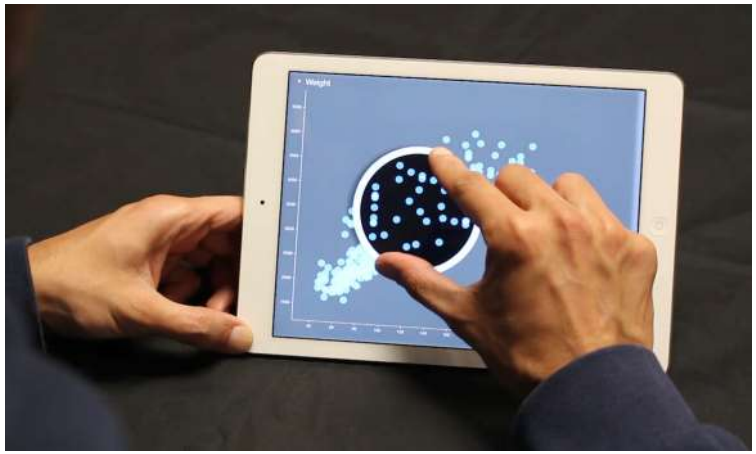


Moving to Tablets



Scatterplot







Video

Sadana & Stasko
AVI ‘14

Key Open Problems in Visualization

Key Open Problems in Visualization

1. Assess and communicate **value**

$$V_{alue} = T + I + E + C$$

Key Open Problems in Visualization

1. Assess and communicate value

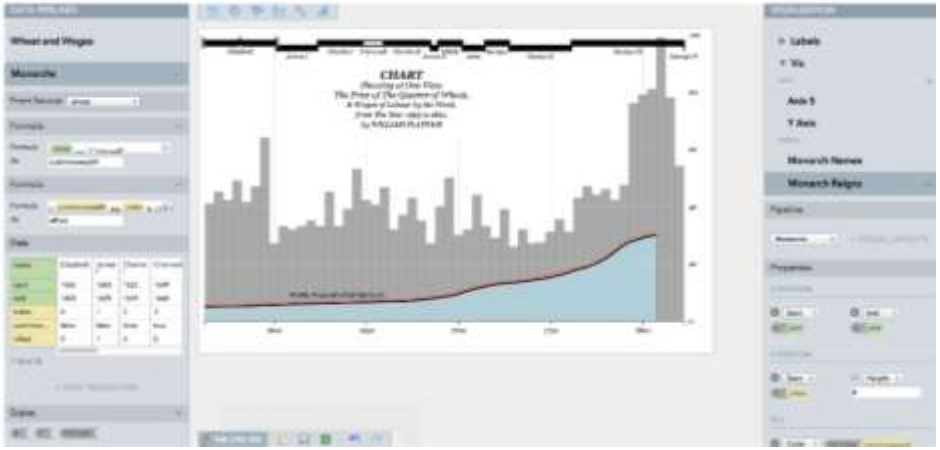
2. Make the construction of visualizations easier

SketchStory



Lee, Kazi & Smith
TVCG (InfoVis) ‘13

Lyra



Satyanarayan & Heer
EuroVis ‘14

Key Open Problems in Visualization

1. Assess and communicate value
2. Make the construction of visualizations easier
3. Address real world “big” problems







Take Aways

Visualization is not for a few precise, concrete tasks
  Exploratory data analysis

Presentation & analysis, related but different

Value = T + I + E + C

Interaction provides power, use it

And finally…


jigsaw
GINGERBREAD
ALE
Middle Sister
Limoncello di Capri
Yuengling
PORTER

Thank you

Acknowledgments

- Supported by the DHS Center of Excellence in Command, Control & Interoperability (VACCINE Center)



- Supported by CCF-0808863 (FODAVA lead), NSF IIS-0915788, NSF IIS-1320537



- Supported by DARPA's XDATA program

April 26, 2014

The Value of Visualization for Exploring and Understanding Data

John Stasko
School of Interactive Computing
Georgia Institute of Technology
stasko@cc.gatech.edu

Exploratory Data Analysis Workshop at
SIAM Intl. Conf. on Data Mining

Many Data Analysis Approaches

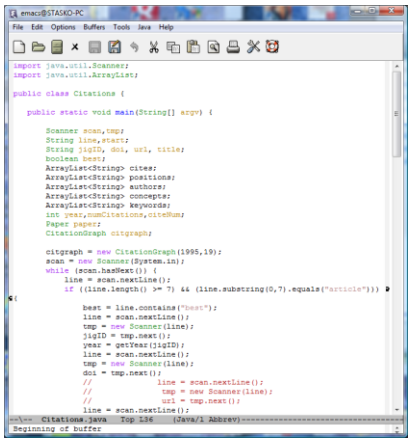
Statistics
Database & information retrieval
Data mining
Machine learning
Visualization





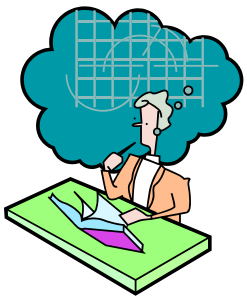
Computational Analysis Approaches

Statistical analysis
Data mining
Machine learning

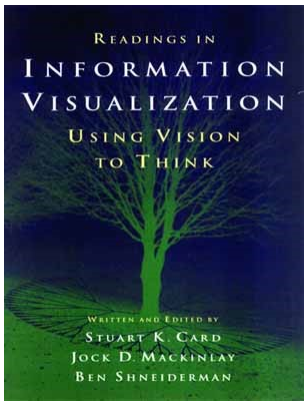


Purpose

Identification

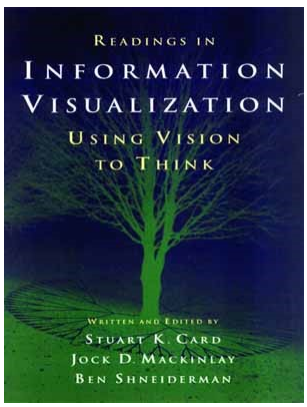


What happens if the analysis isn't any good?



Visualization
"The use of computer-supported, interactive visual representations of data to amplify cognition"

5

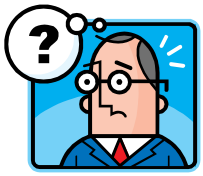


Visualization
"The use of computer-supported, **interactive** visual representations of data to amplify cognition"

6

Visualization most useful in **exploratory data analysis**

- Don't know what you're looking for
- Don't have a priori questions
- Want to know what questions to ask



7

"Contained within the data of any investigation is information that can yield conclusions to questions not even originally asked. That is, there can be surprises in the data…To regularly miss surprises by failing to probe thoroughly with visualization tools is terribly inefficient because the cost of intensive data analysis is typically very small compared with the cost of data collection."

W. Cleveland
The Elements of Graphing Data

8

Preconceptions about Visualization Utility

- Answering specific questions and accomplishing specific analytic tasks
- Generating unexpected, serendipitous discoveries and insights
  "Finding a needle in a haystack"



Yes, but not what it's best for

9

Visualization Utility

Awareness
Inquiry
Exploration

10

1. Visualization is more than just answering specific questions about data (as is often the case for automated analysis methods)
It also facilitates the investigative analysis process, which supports analysts in developing **awareness** of, learning about, and generating trust in their data, its domain, and its context.

Learning, awareness, trust, context

11

2. Visualization, primarily through its interactive capabilities, promotes a dialog of **inquiry** between analysts and their data by allowing a diverse and flexible set of questions to be asked and answered about a data collection and by spurring the generation of new questions.

Q & A inquiry dialog through interaction

12

3. Visualization rapidly and efficiently facilitates flexible **exploration** of data to foster both broad and deep understanding of the information contained therein.

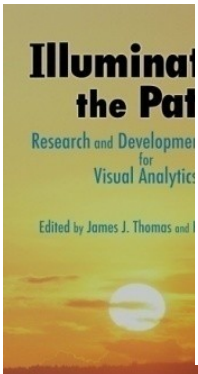
Exploration for broad & deep understanding

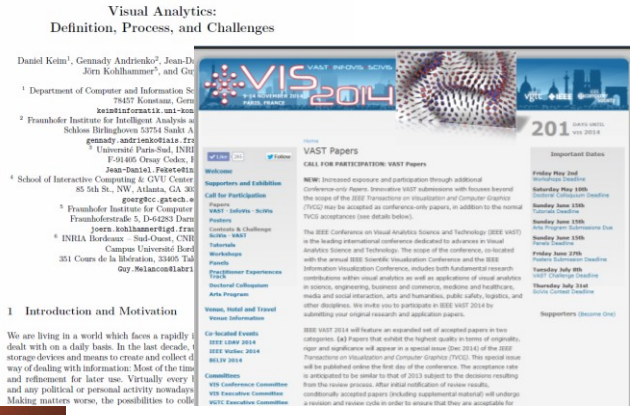
13

Computational analysis methods

\+ Human-directed interactive visualization

Visual Analytics




Visual Analytics:
Definition, Process, and Challenges

1 Introduction and Motivation

VIS 2014

VAST Papers

14

How do we integrate computational analysis methods with interactive visualization smoothly, fluidly, and effectively?

Our experiences with Jigsaw

15

Help “investigators” explore, analyze and understand large document collections

Academic Papers

Health Forums

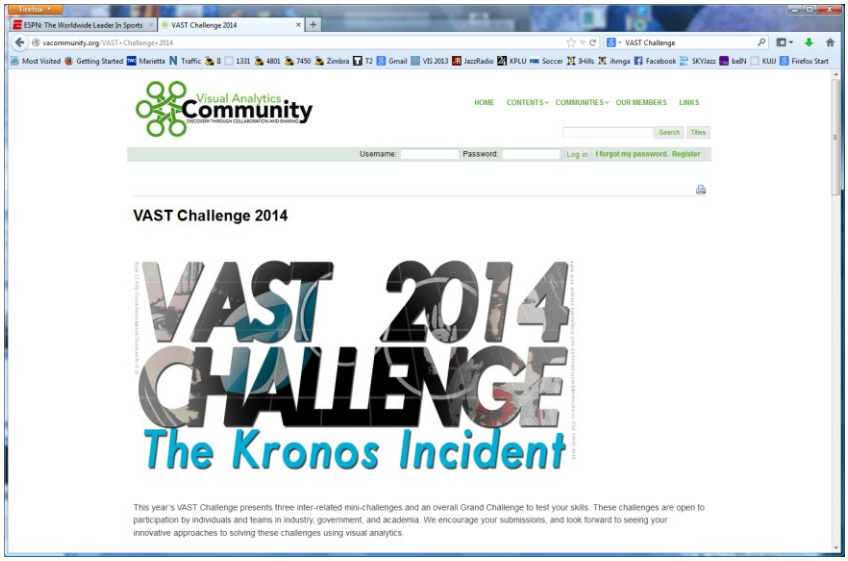
Product Reviews

Police Reports



16

http://vacommunity.org/VAST+Challenge+2014



17

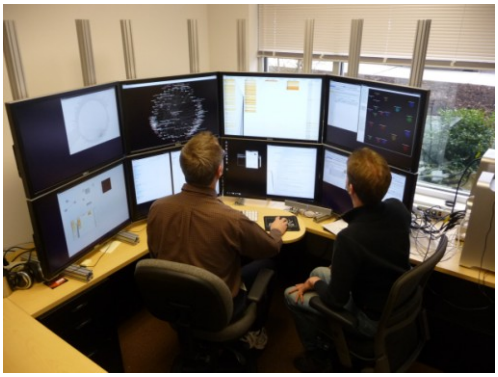
"Putting the pieces together"

Jigsaw

Multiple visualizations of documents, analysis results, entities, and their connections

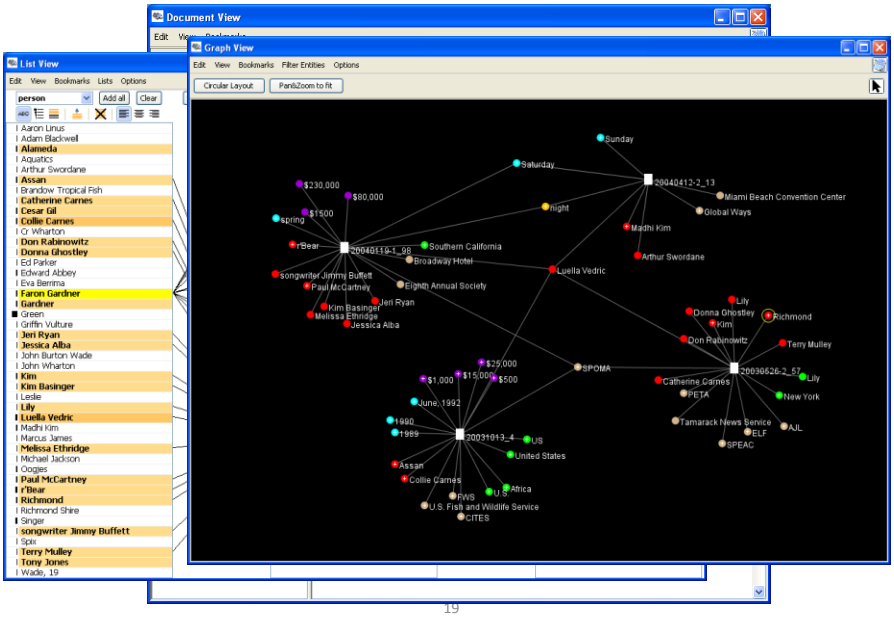
Views are highly coordinated





18

Jigsaw's early capabilities



19

Computational Analysis (text mining)

Document Summarization
one sentence summaries, word clouds, and keyword summaries

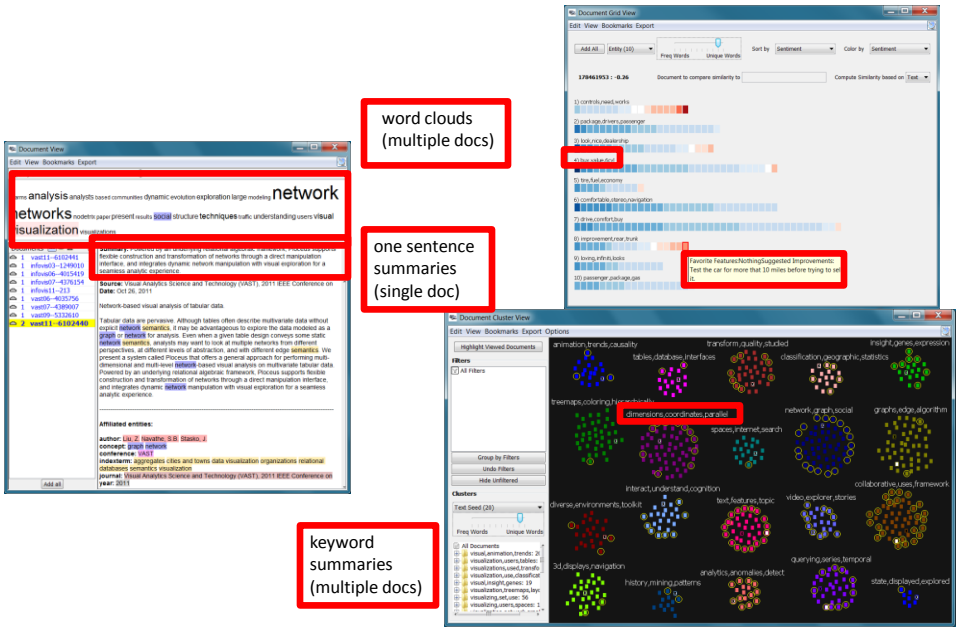
Document Similarity
based on text or entities

Document Clustering
based on text or entities

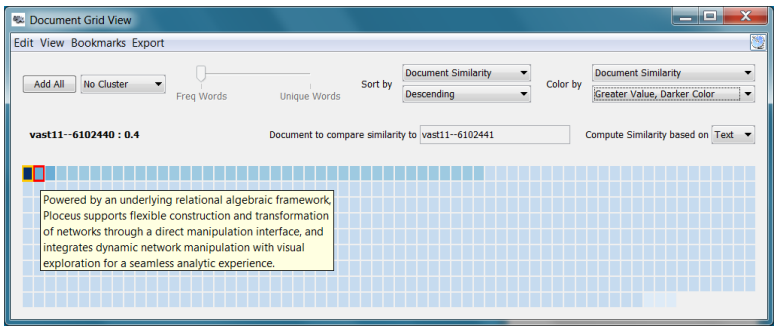
Sentiment Analysis
per document

20

Document Summarization

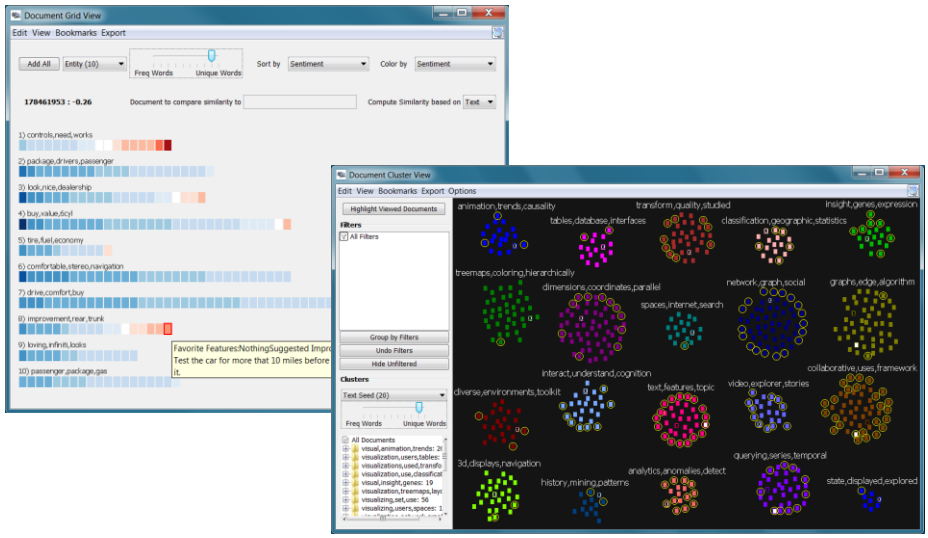


Document Similarity



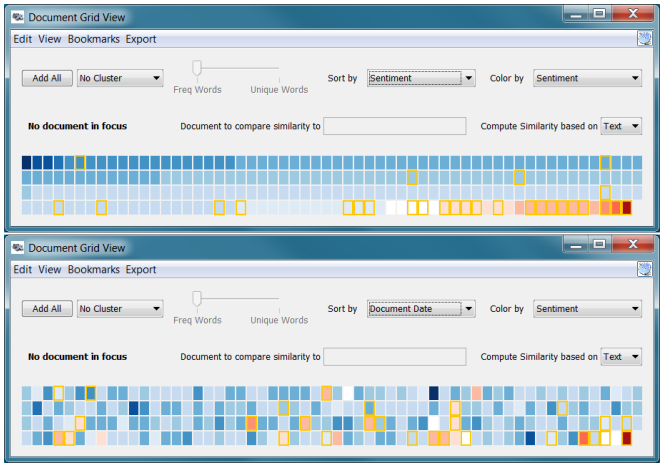
Documents ordered and colored by similarity to selected reference document

Document Clustering



Sentiment Analysis

Documents ordered and colored by sentiment



Documents ordered by date and colored by sentiment

Fluid integration

Meld text mining and visualization together fluidly and smoothly



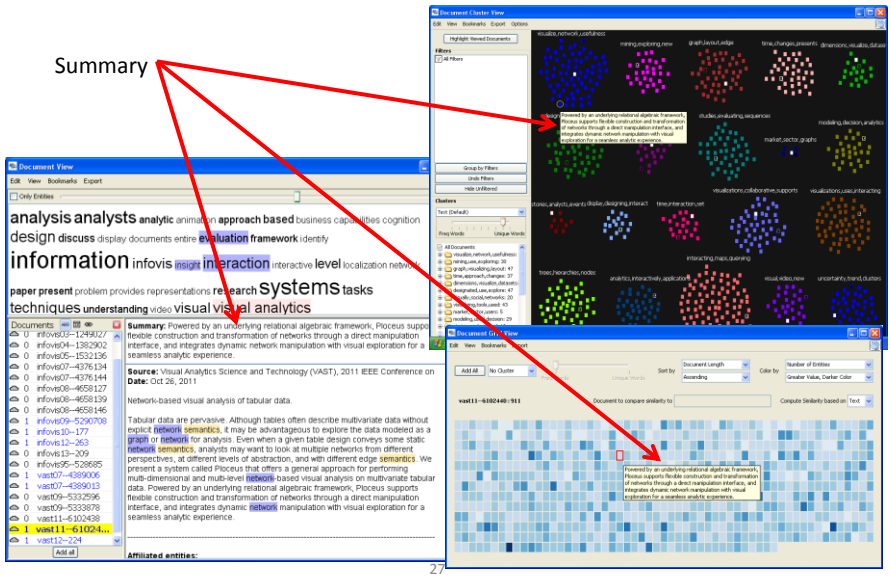
25

Fluidity Lessons/Recommendations

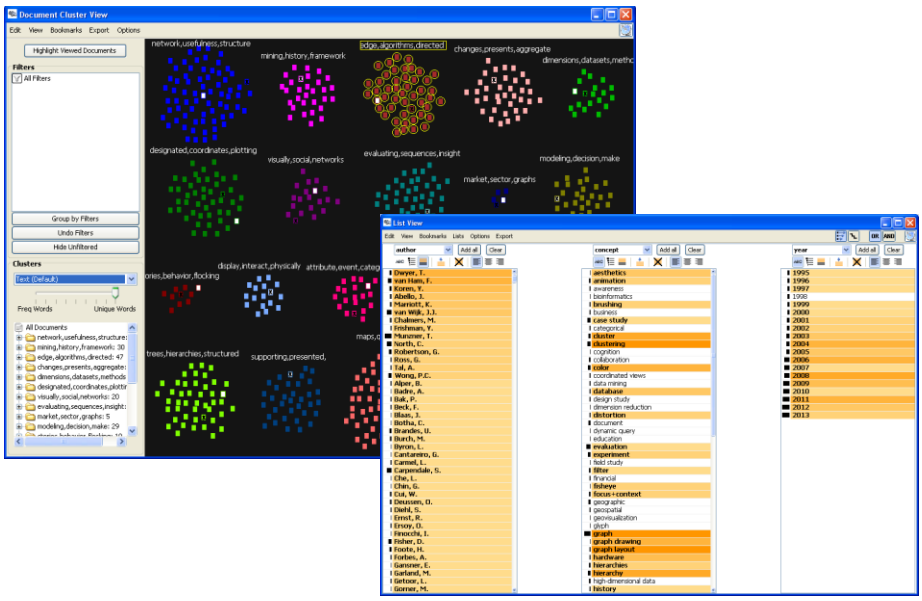
1. Make different computational analysis results available throughout the system in a variety of different contexts and views
2. Flexibly allow analysis output also to be used as input
3. Integrate different, independent computational analysis measures through interactive visualization to extend functionality and power
4. Provide computational support for both analysis directions: narrowing down as well as widening the scope of an investigation
5. Expose algorithm parameters in an interactive user accessible way

26

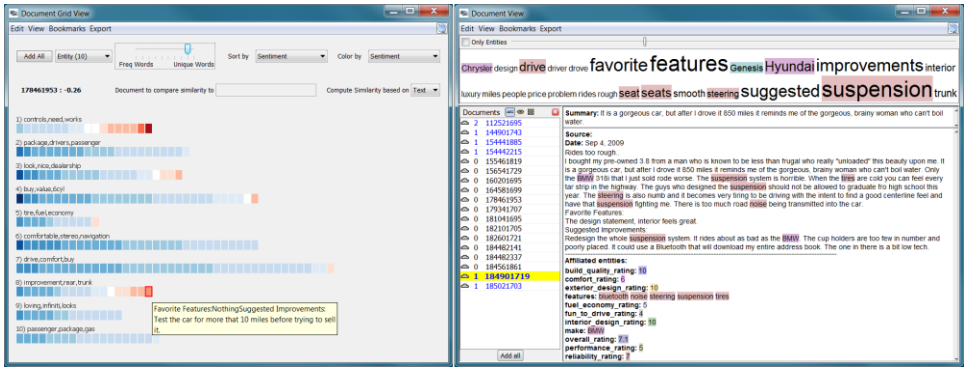
1. Make different computational analysis results available throughout the system in a variety of different contexts and views



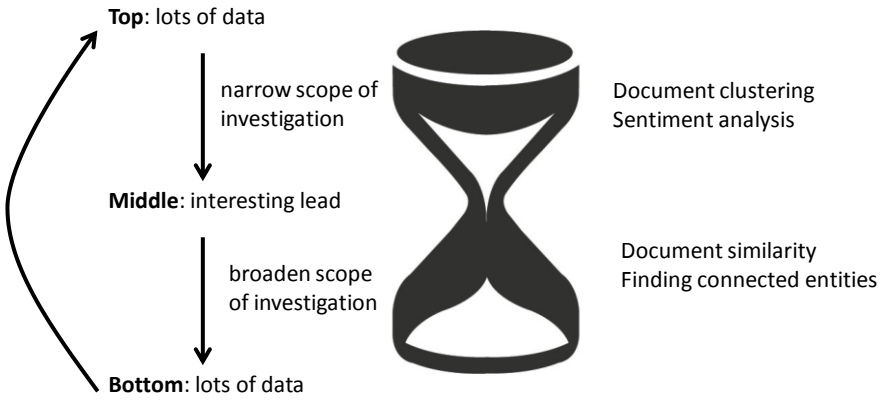
2. Flexibly allow analysis output also to be used as input



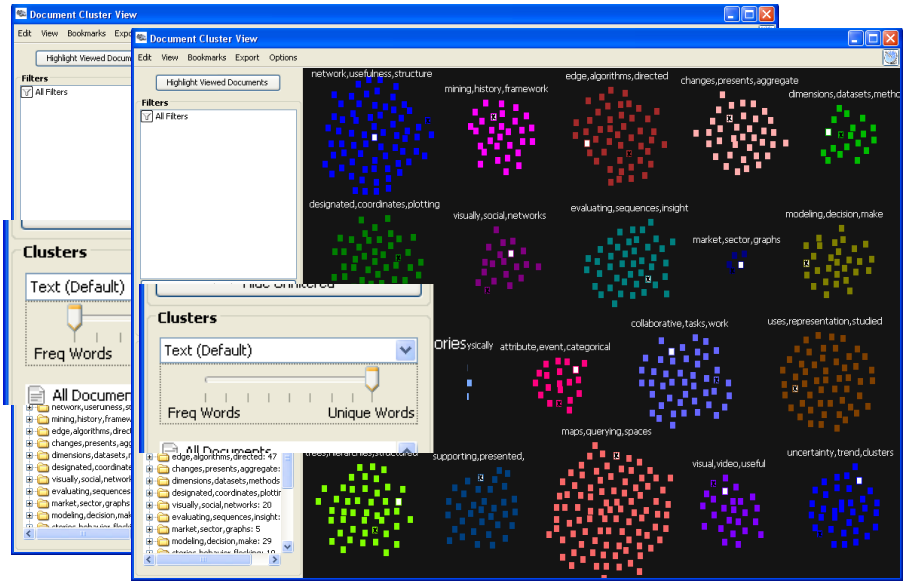
3. Integrate different, independent computational analysis measures through interactive visualization to extend functionality and power



4. Provide computational support for both analysis directions: narrowing down as well as widening the scope of an investigation



5. Expose algorithm parameters in an interactive user accessible way



31

1646 IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 19, NO. 10, OCTOBER 2013

Combining Computational Analyses and Interactive Visualization for Document Exploration and Sensemaking in Jigsaw

Carsten Görg, *Member, IEEE*, Zhicheng Liu, Jaeyeon Kihm, Jaegul Choo, *Student Member, IEEE*, Haesun Park, *Member, IEEE*, and John Stasko, *Senior Member, IEEE*

Abstract—Investigators across many disciplines and organizations must sift through large collections of text documents to understand and piece together information. Whether they are fighting crime, curing diseases, deciding what car to buy, or researching a new field, inevitably investigators will encounter text documents. Taking a visual analytics approach, we integrate multiple text analysis algorithms with a suite of interactive visualizations to provide a flexible and powerful environment that allows analysts to explore collections of documents while sensemaking. Our particular focus is on the process of integrating automated analyses with interactive visualizations in a smooth and fluid manner. We illustrate this integration through two example scenarios: An academic researcher examining InfoVis and VAST conference papers and a consumer exploring car reviews while pondering a purchase decision. Finally, we provide lessons learned toward the design and implementation of visual analytics systems for document exploration and understanding.

Index Terms—Visual analytics, information visualization, sensemaking, exploratory search, information seeking, document analysis

1 Introduction

Everyday, analysts and investigators confront large collections of data as they make decisions, solve problems, or simply seek to understand a situation better. Frequently, the data collections include text documents or documents with key text components. While numerical or structured data are more amenable to statistical and computational analysis, text data are conversely often messy and noisy, requiring a very sequential, slow processing

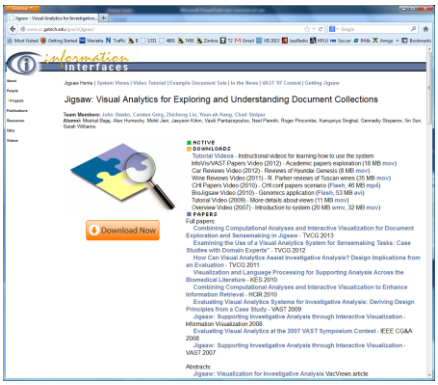
- An academic researcher moves into a new area and seeks to understand the key ideas, topics, and trends of the area, as well as the set of top researchers, their interests, and collaborations.
- A consumer wants to buy a new car but encounters a large variety of possible models to choose from, each of which has 10 to 20 "professional" reviews and a web forum with hundreds of postings

Free access

32

Download for free

http://www.cc.gatech.edu/gvu/ii/jigsaw





33

Further benefits of visualization

Ease of specifying queries
Opportunistic discovery of relevant data
Spurs the generation of new questions

34



Take Aways

Visualization's benefits

Awareness, inquiry, exploration

Fluidly integrate computational analysis & visualization

Example techniques

Acknowledgments

- Work conducted as part of the Southeastern Regional Visualization and Analytics Center, supported by DHS and NVAC and the DHS Center of Excellence in Command, Control & Interoperability (VACCINE Center)

- Supported by NSF IIS-0414667, CCF-0808863 (FODAVA lead), NSF IIS-0915788

This material is supported in part by: National Science Foundation *Grant #0414667*

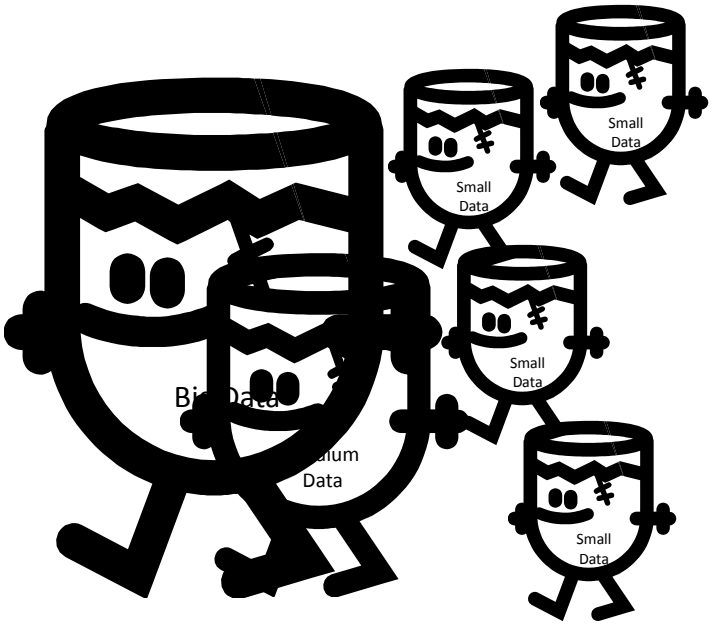
November 21, 2013

The Value of Visualization for Exploring and Understanding Data

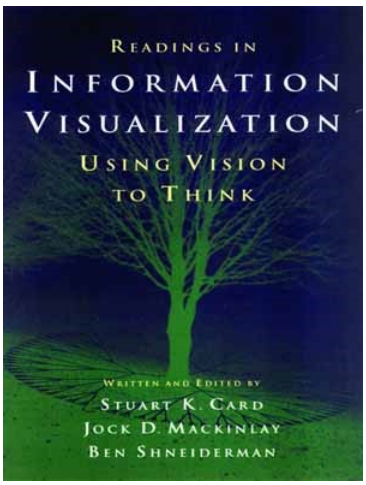
John Stasko

School of Interactive Computing
Georgia Institute of Technology
stasko@cc.gatech.edu





3



Visualization

"The use of computer-supported, interactive visual representations of data to amplify cognition"

4

Visualization

~~Making pretty pictures~~

A cognitive process

Internalize an understanding



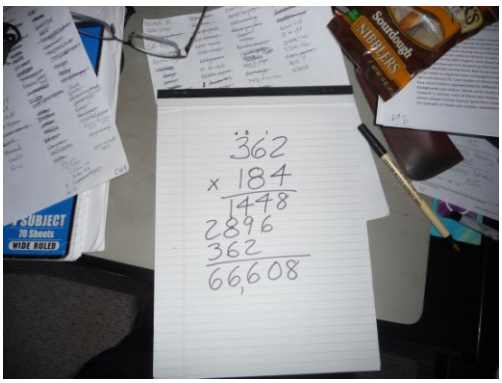
5

Visuals help us think

Provides a frame of reference, temporary storage area

Cognition → Perception

Pattern matching



6

Visualization for Big Data?

Maybe…

Petabyte →



10^{15} → 10^{6}

7

Big Data Tasks

Identification
- Data mining

Prediction
- Machine learning

Visualization
- Let's see…

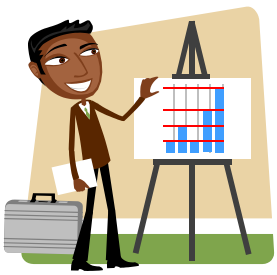
8

Applications of Visualization

Presentation

Analysis

9



1. Presentation

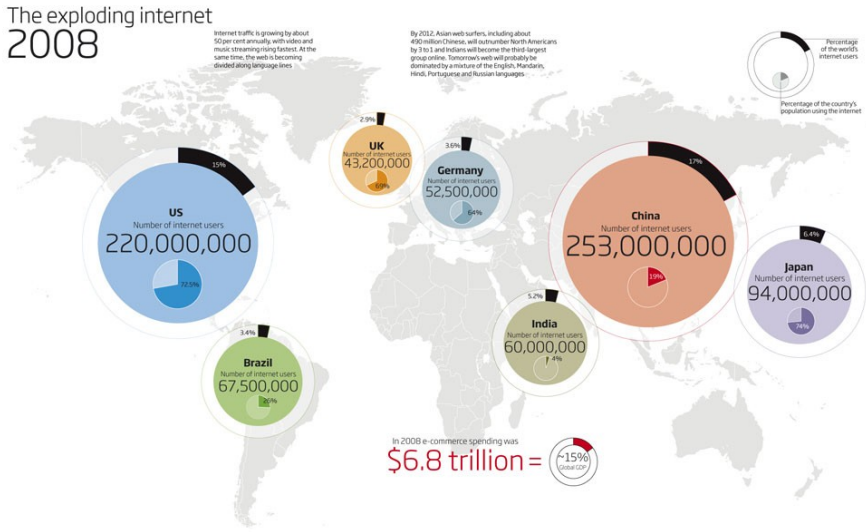
Communicate data and ideas

Explain and inform

Influence and persuade

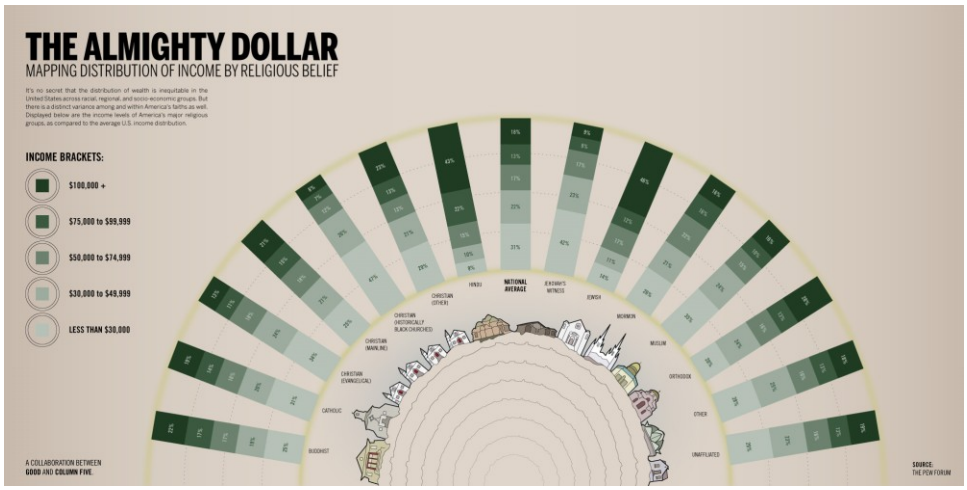
Provide evidence and support

10



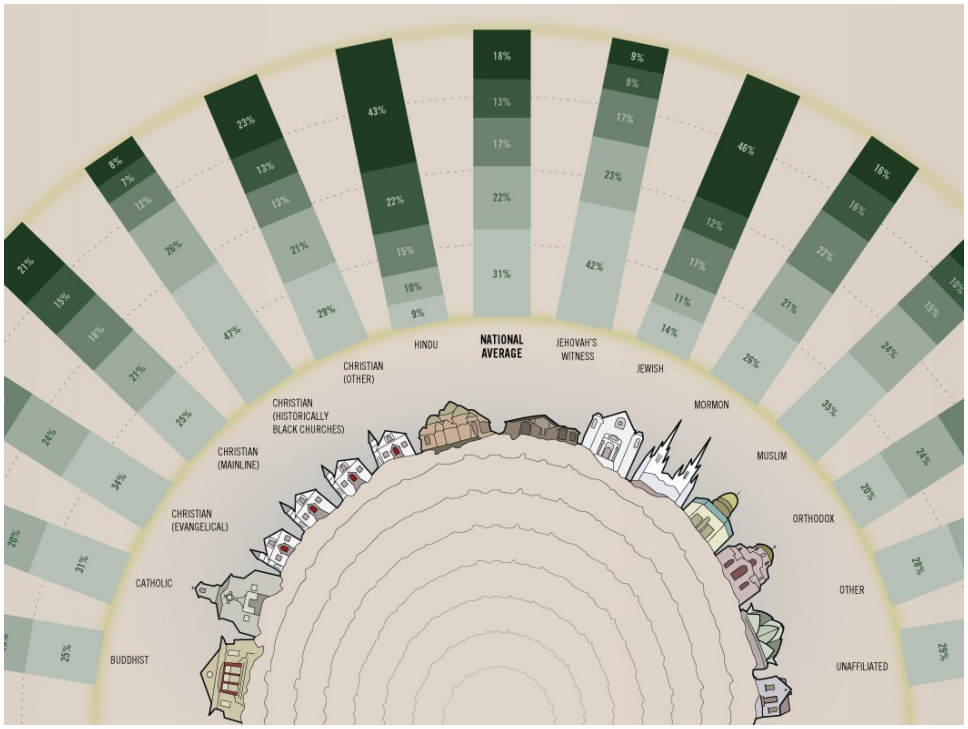
http://www.newscientist.com/data/images/ns/cms/mg20227062.200/mg20227062.200-6_1000.jpg

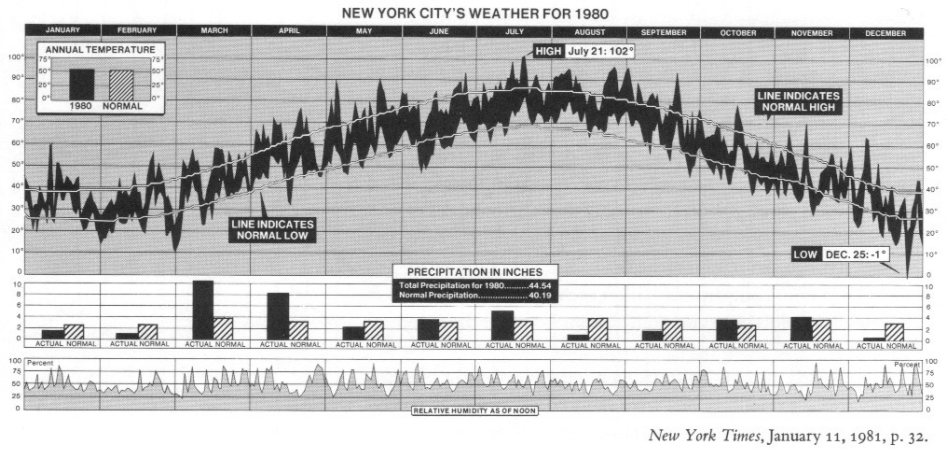
11



http://awesome.good.is.s3.amazonaws.com/transparency/web/1002/almighty-dollar/transparency.jpg

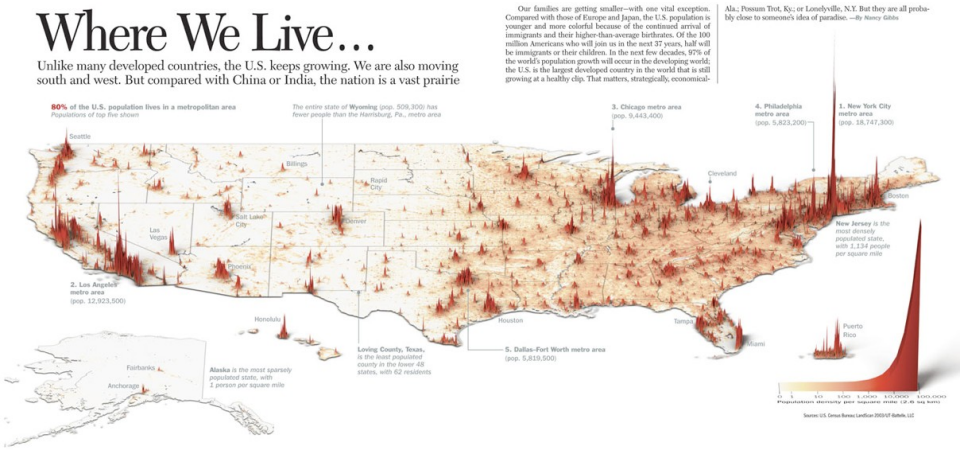
12





New York Times, January 11, 1981, p. 32.

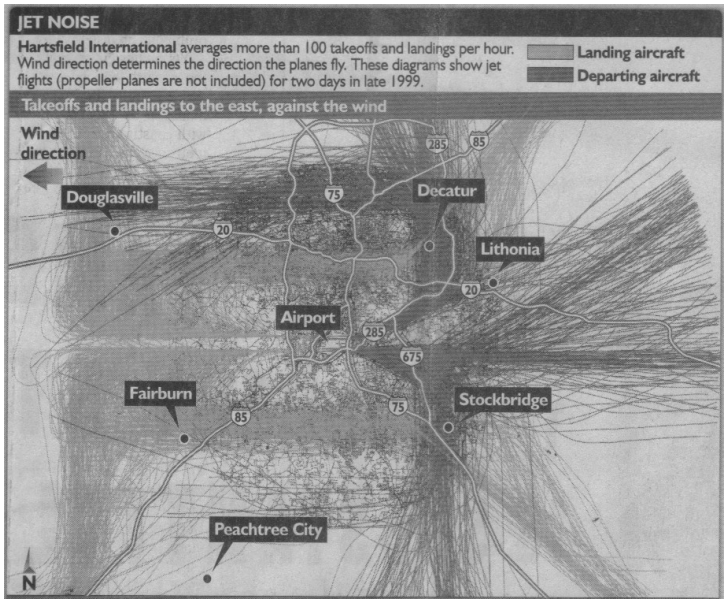
Tufte, Vol. 1



http://infographicsnews.blogspot.com/2009/04/mantras-joe-lertolas-maps.html

15

Atlanta Journal
April 30, 2000



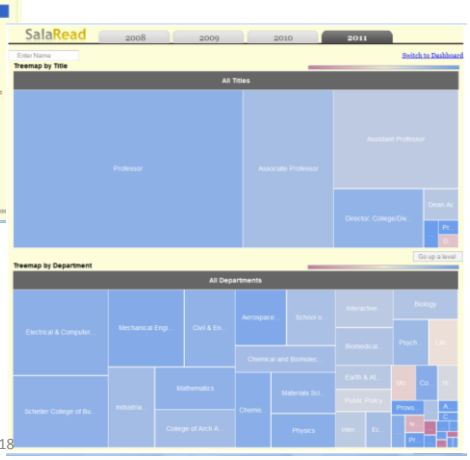
16

Simply presenting data ***visually*** can have a profound impact

17



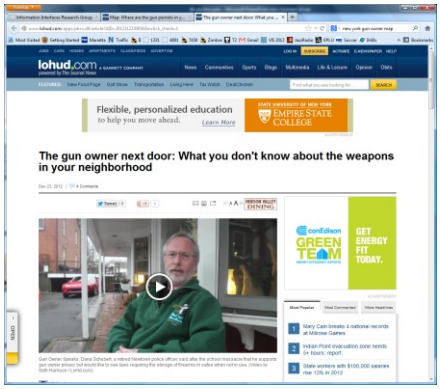
CS 7450
Fall '12
Design project

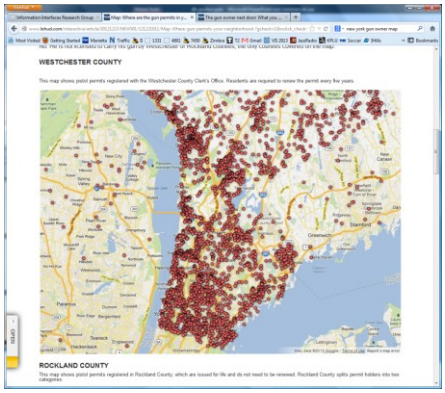


Nate Osborne
Nitya Noronha
Ameya Zambre
Pratik Zaveri

18

Gun ownership in New York counties





http://www.lohud.com/apps/pbcs.dll/article?AID=2012312230056&nclick_check=1

http://www.lohud.com/interactive/article/20121223/NEWS01/121221011/Map-Where-gun-permits-your-neighborhood-?gcheck=1&nclick_check=1

19

Frequent presentation goals

Clarify
Focus
Highlight
Simplify

May just show a few variables
and/or a subset of the data cases

20

2. Analysis

understand, compare, decide, judge, evaluate, assess, determine, …



21

Many Data Analysis Approaches

Statistics
Database & information retrieval
Data mining
Machine learning





22

"Contained within the data of any investigation is information that can yield conclusions to questions not even originally asked. That is, there can be surprises in the data…To regularly miss surprises by failing to probe thoroughly with visualization tools is terribly inefficient because the cost of intensive data analysis is typically very small compared with the cost of data collection."

W. Cleveland
The Elements of Graphing Data

23

Visualization Techniques for Analysis

When the standard charts work, use them

24



http://infosthetics.com/archives/2008/09/funniest_pie_chart_ever.html

25



26

Frequent analysis goals

Show many variables
Illustrate overview and detail
Facilitate comparison

Display may not be easy to interpret at first

27

Preconceptions about Visualization Utility

Answering specific questions and accomplishing specific analytic tasks
Generating unexpected, serendipitous discoveries and insights
"Finding a needle in a haystack"



Yes, but not what it's best for

28

1. Visualization is more than just answering a specific question (as is often the case for automated analysis methods)
It also is about the investigative analysis process, which helps us to learn about, develop awareness of, and generate trust in the data, its domain, and its context.

Learning, awareness, trust, context

29

Many Data Analysis Approaches
Statistics
Database & information retrieval
Data mining
Machine learning





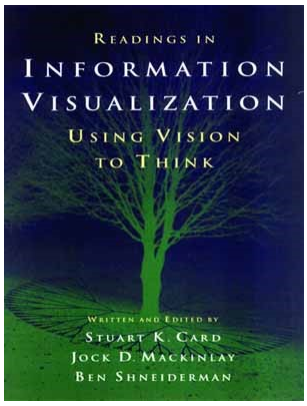
Use them when they get the job done!

30

2. Visualization, primarily through its interactive capabilities, promotes a dialog between analysts and their data by allowing a diverse and flexible set of questions to be asked and answered about a data collection and by spurring the generation of new questions.

Q & A dialog through interaction

31



Visualization
"The use of computer-supported, **interactive** visual representations of data to amplify cognition"

32

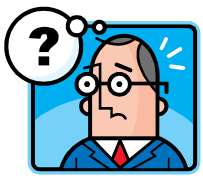
3. Visualization rapidly and efficiently facilitates flexible exploration to foster both broad and deep understanding of the information in a data collection.

Broad and deep understanding quickly

33

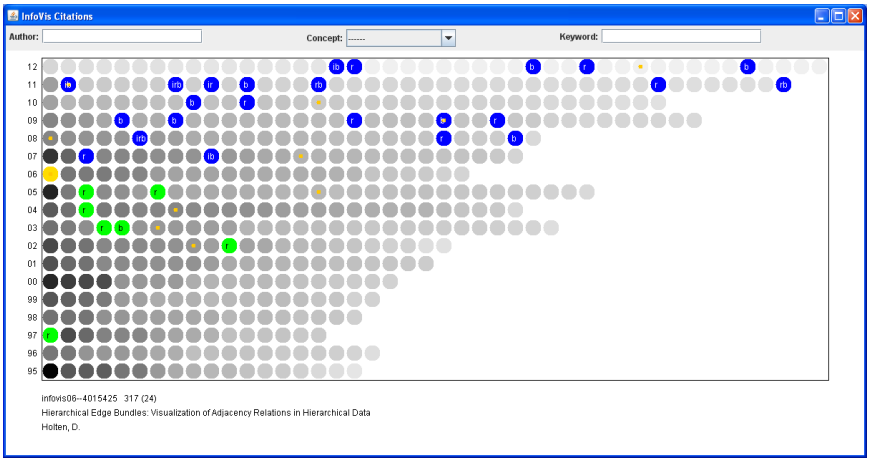
Visualization most useful in **exploratory data analysis**

- Don't know what you're looking for
- Don't have a priori questions
- Want to know what questions to ask



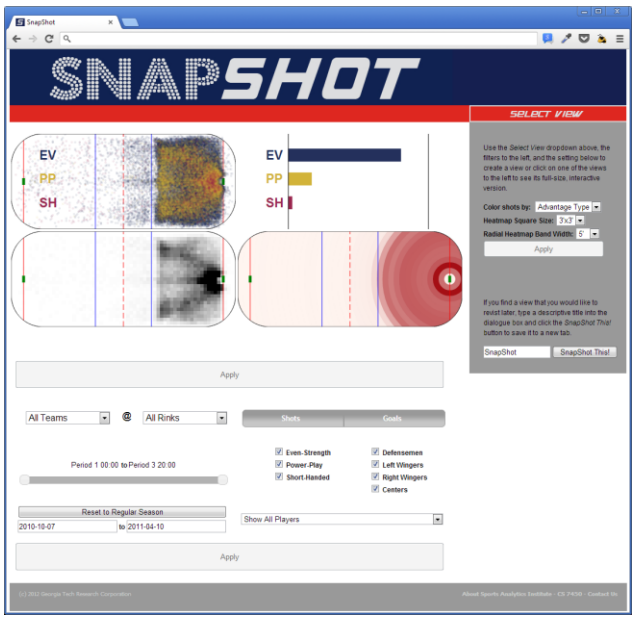
34

http://www.cc.gatech.edu/gvu/ii/citevis



CiteVis

Stasko et al
InfoVis poster '13



SnapShot

Pileggi, Stolper, Boyle, Stasko
TVCG (InfoVis) '12

"Putting the pieces together"

Jigsaw

Computational analysis of documents' text

Entity identification, document similarity, clustering, summarization, sentiment

Multiple visualizations of documents, analysis results, entities, and their connections

Views are highly coordinated



37



38





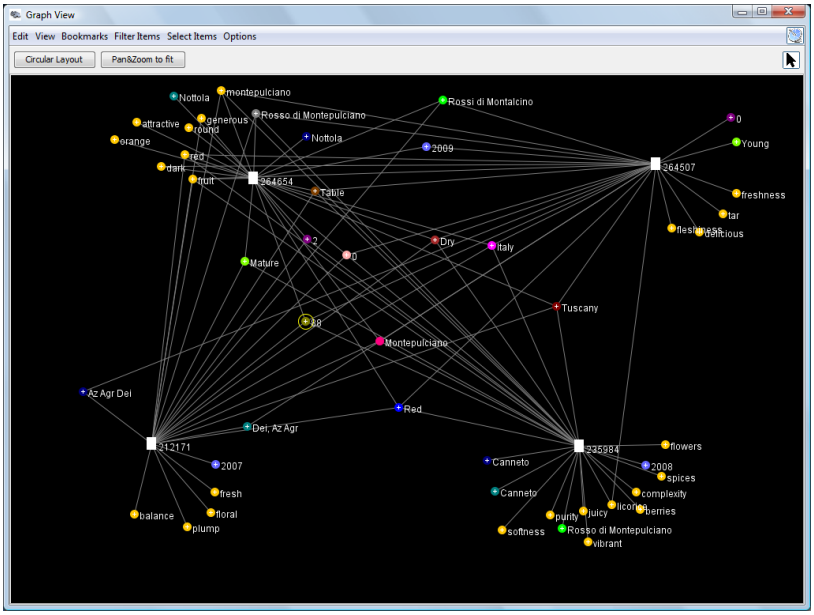
Example

Reviews of wines from Washington state from '07-'10
Text: review narrative
Entities: variety, producer, rating, vintage, color, location, producer, "descriptor", …
Descriptor terms (~9000), eg: abrasive, oaky, cherry, mocha, textured
1323 reviews
from database of 150,000 reviews

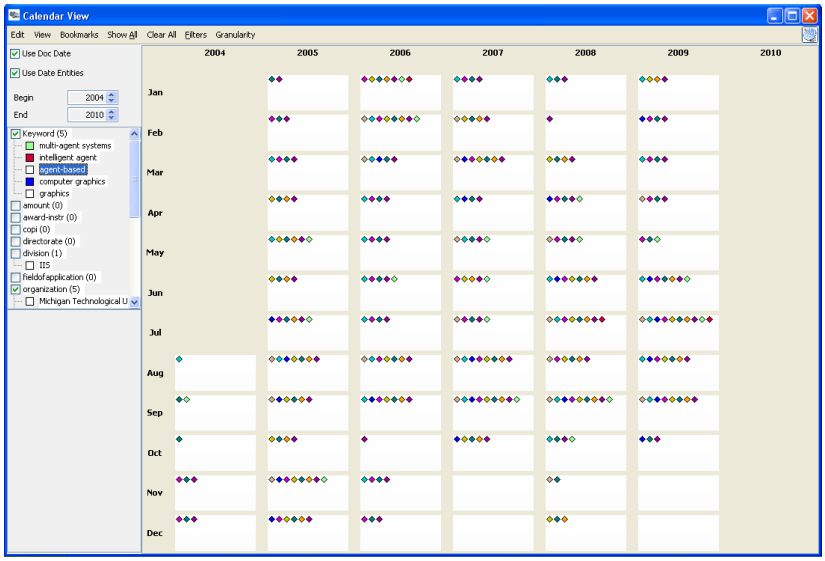




Graph View

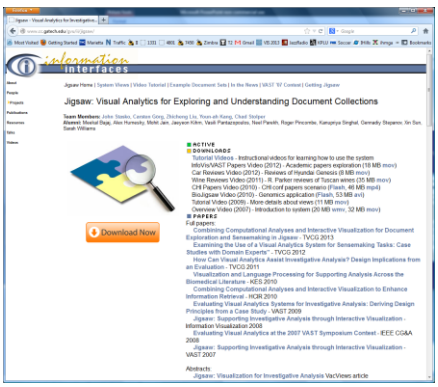


Calendar View



Download for free

http://www.cc.gatech.edu/gvu/ii/jigsaw





Applications

- Intelligence & law enforcement
  - Police cases
  - Won 2007 VAST Contest
  - Stasko et al, *Information Visualization* '08
- Academic papers, PubMed
  - All InfoVis & VAST papers
  - CHI papers
  - Görg et al, KES '10
- Investigative reporting
- Fraud
  - Finance, accounting, banking
- Grants
  - NSF CISE awards from 2000
- Topics on the web (medical condition)
  - Autism
- Consumer reviews
  - Amazon product reviews, edmunds.com, wine reviews
  - Görg et al, HCIR '10
- Business Intelligence
  - Patents, press releases, corporate agreements, …
- Emails
  - White House logs
- Software
  - Source code repositories
  - Ruan et al, SoftVis '10

43

1. Visualization is more than just answering a specific question (as is often the case for automated analysis methods)
It also is about the investigative analysis process, which helps us to learn about, develop awareness of, and generate trust in the data, its domain, and its context.

Learning, awareness, trust, context

Data cleaning example

44

2. Visualization, primarily through its interactive capabilities, promotes a dialog between analysts and their data by allowing a diverse and flexible set of questions to be asked and answered about a data collection and by spurring the generation of new questions.

Q & A dialog through interaction

Repeated interactive queries

45

3. Visualization rapidly and efficiently facilitates flexible exploration to foster both broad and deep understanding of the information in a data collection.

Broad and deep understanding quickly

Varieties, colors, reviewers, producers, concepts, …

46

Insights

No great aha! – More learning & knowledge building
Domain knowledge (wines) matters

47

Further benefits of visualization

Ease of specifying queries
Opportunistic discovery of relevant data
Spurs the generation of new questions

48

Back to tasks

What do you want to do with big data?
Search, identification, answer questions, prediction, exploration, learning, …



Take Aways

Presentation & analysis

Interaction provides the power

Exploring & developing questions

Acknowledgments

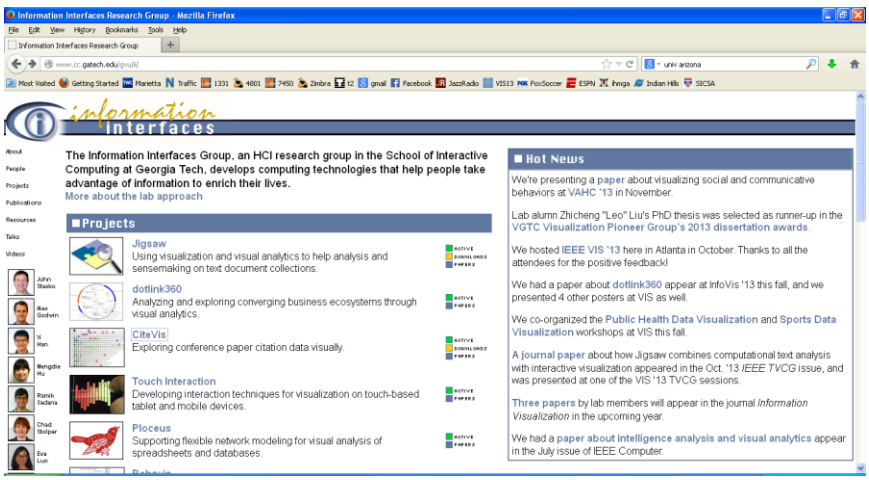
- Work conducted as part of the Southeastern Regional Visualization and Analytics Center, supported by DHS and NVAC and the DHS Center of Excellence in Command, Control & Interoperability (VACCINE Center)

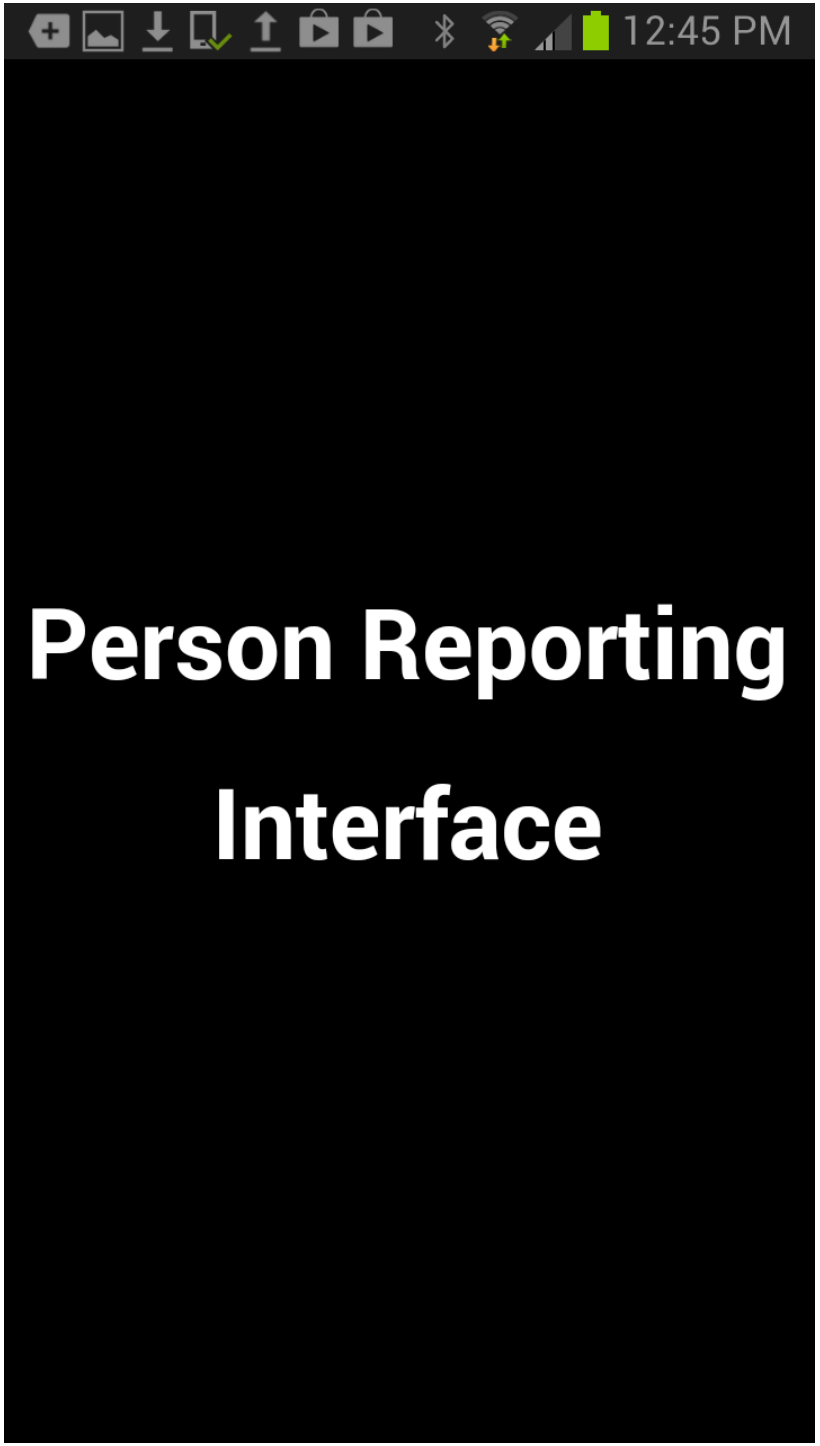
- Supported by NSF IIS-0414667, CCF-0808863 (FODAVA lead), NSF IIS-0915788

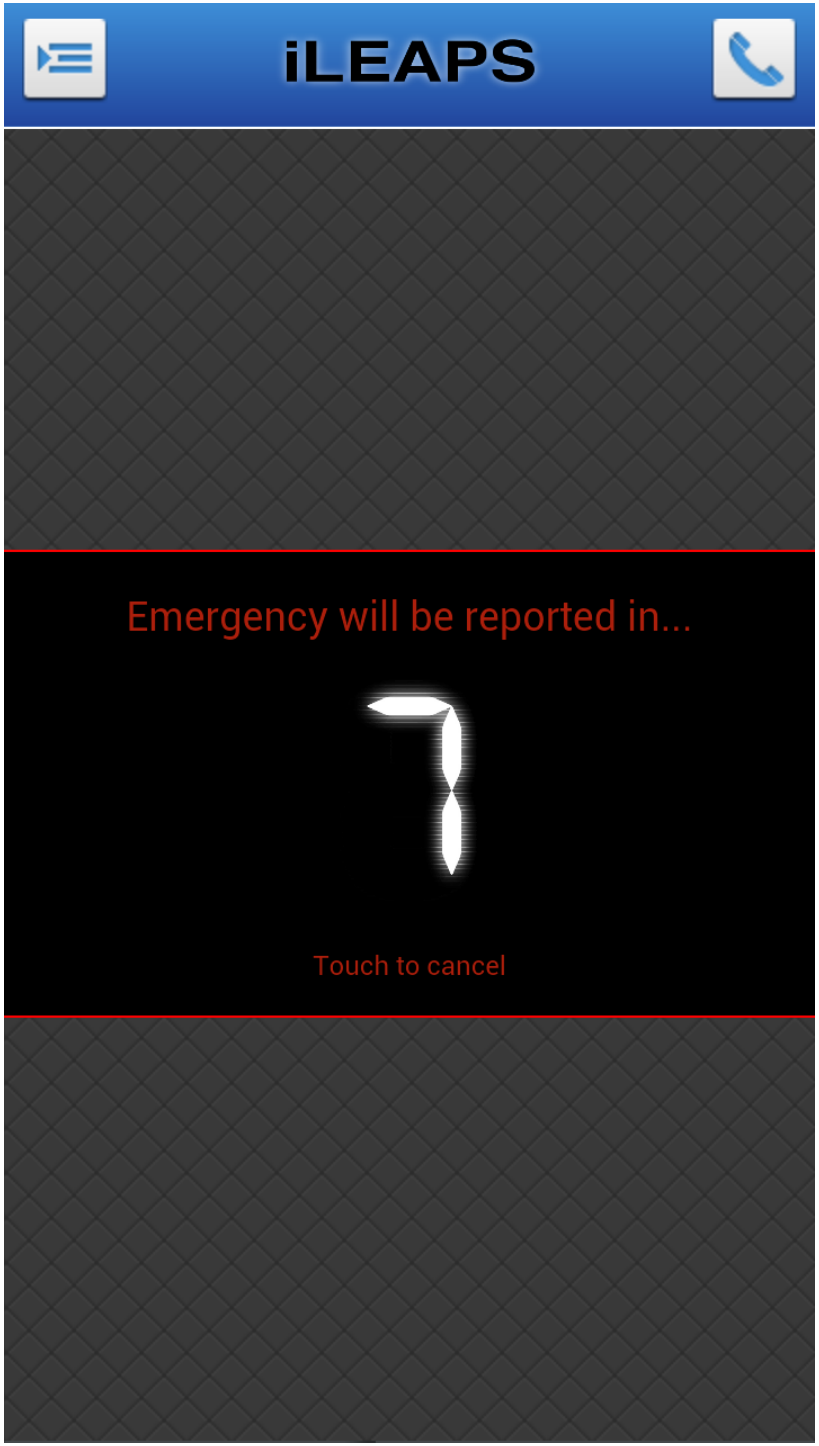
This material is supported in part by:
National Science Foundation
Grant #0414667

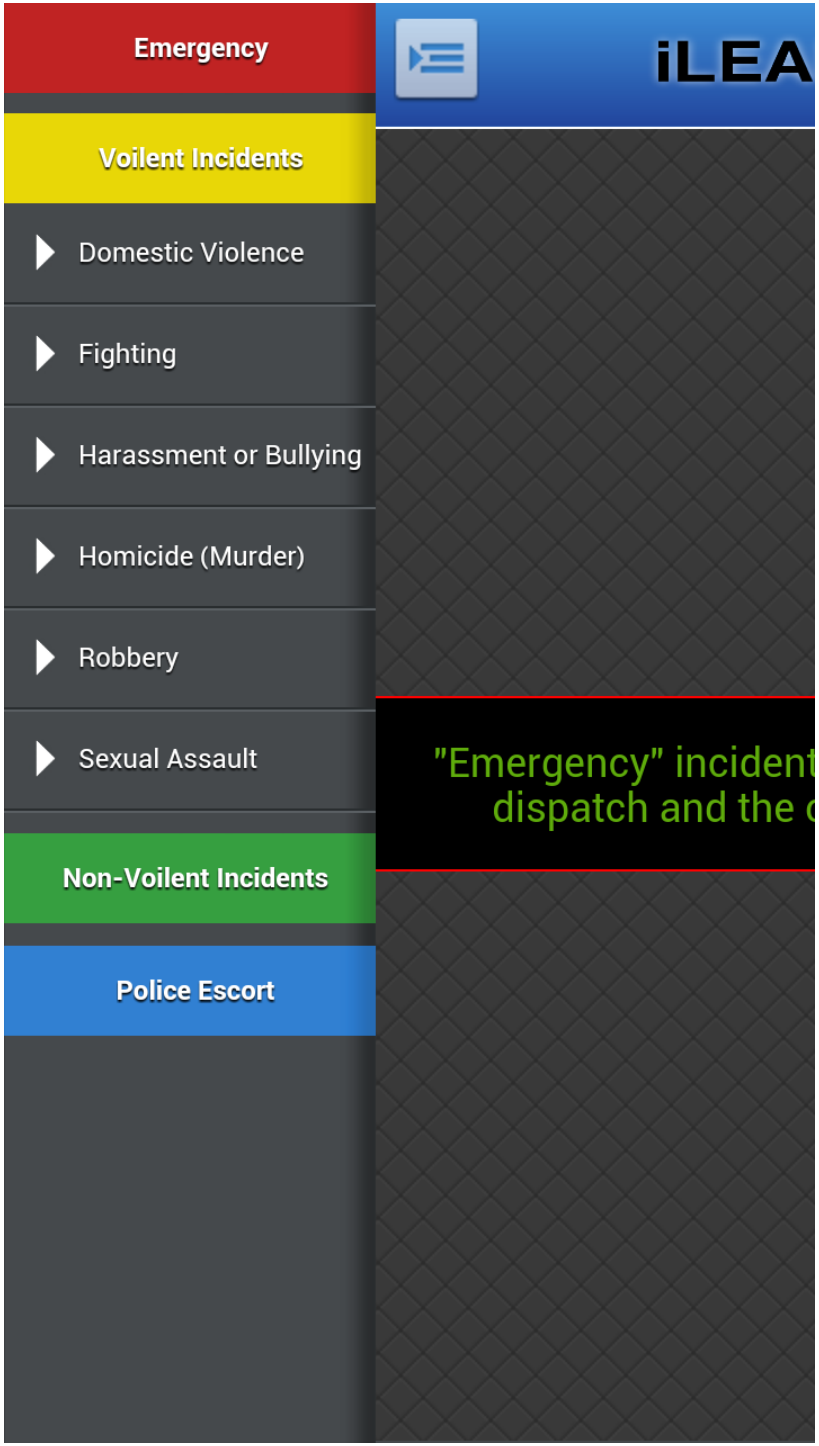
My Research

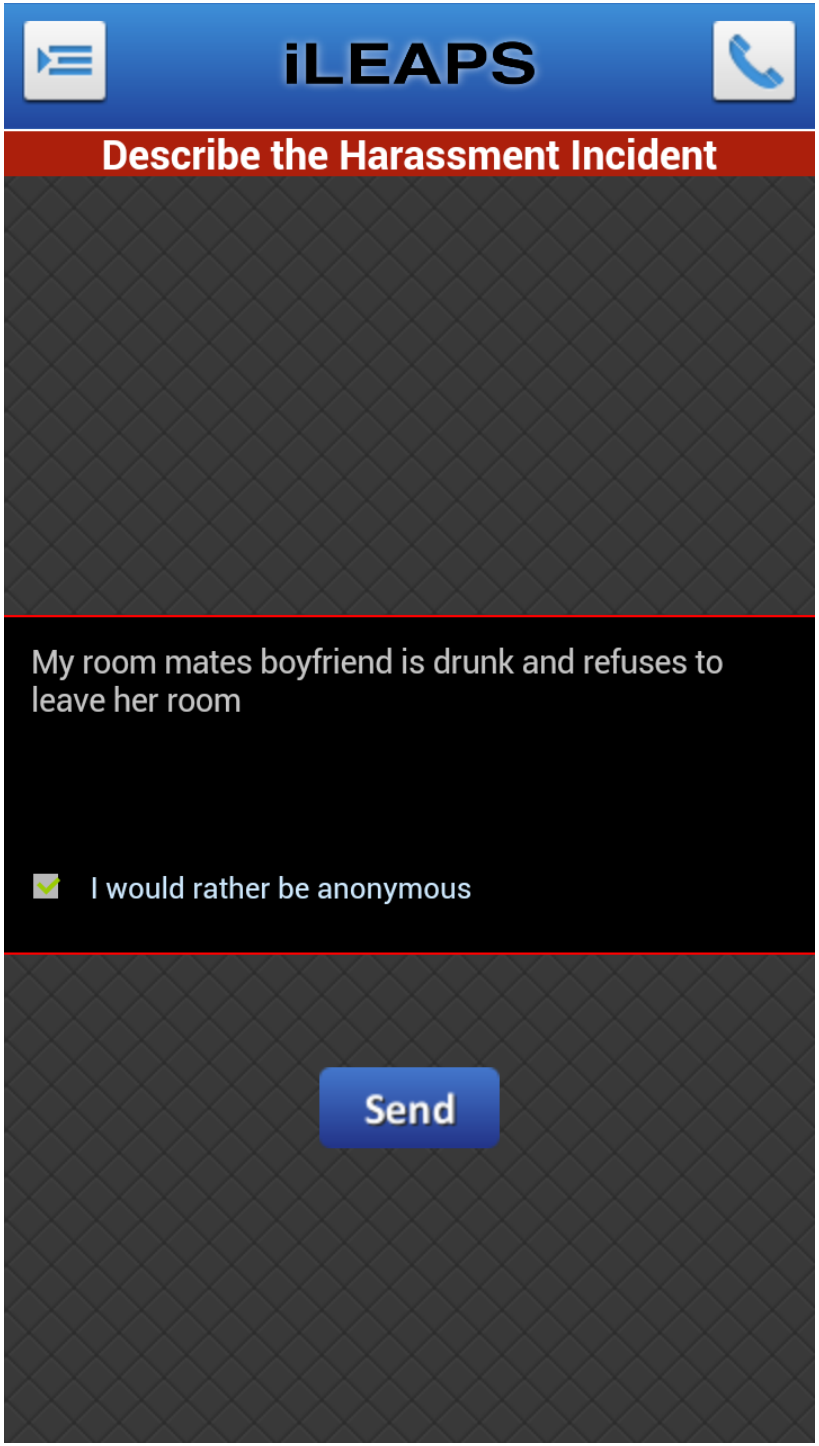
http://www.cc.gatech.edu/gvu/ii

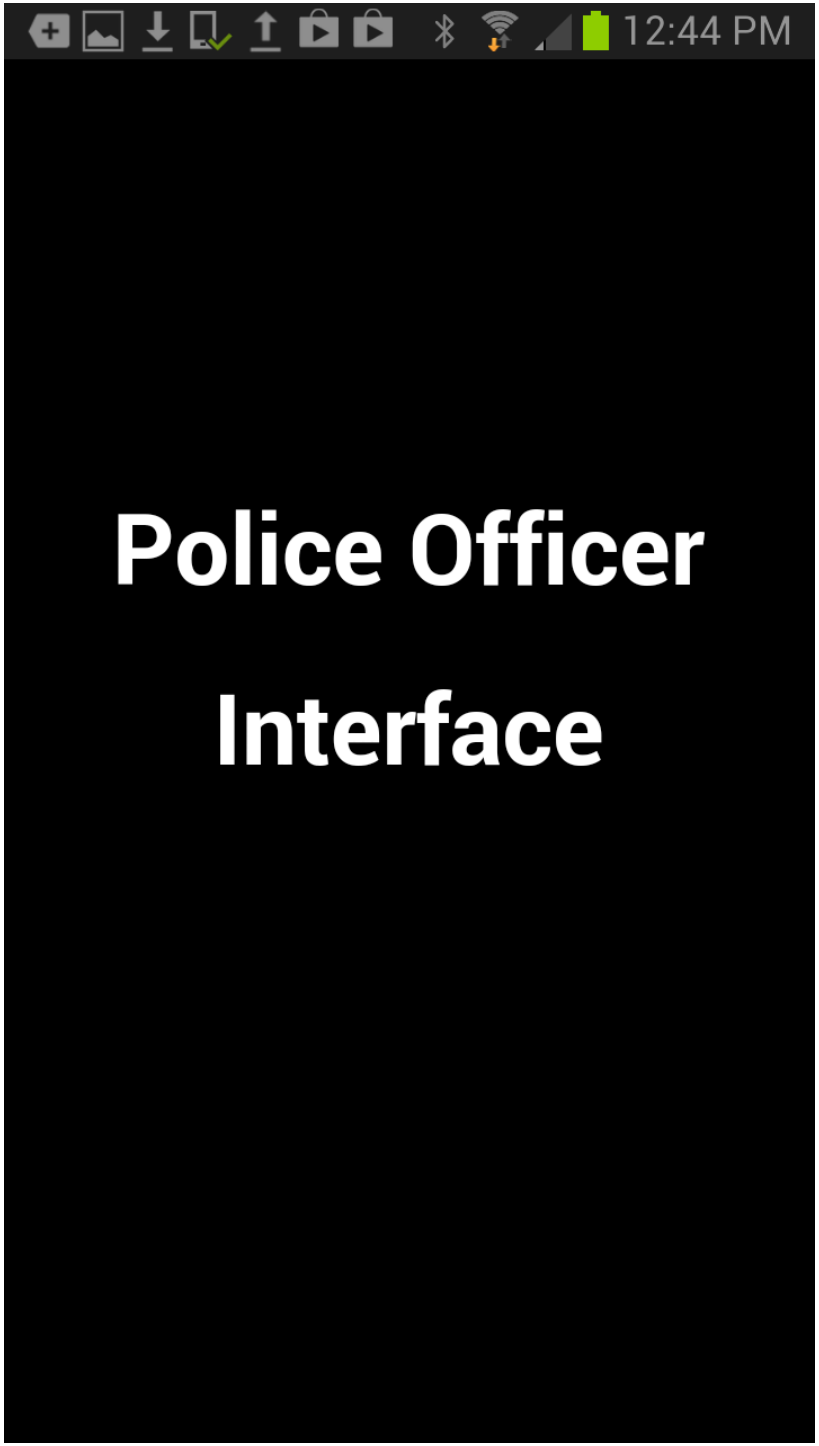


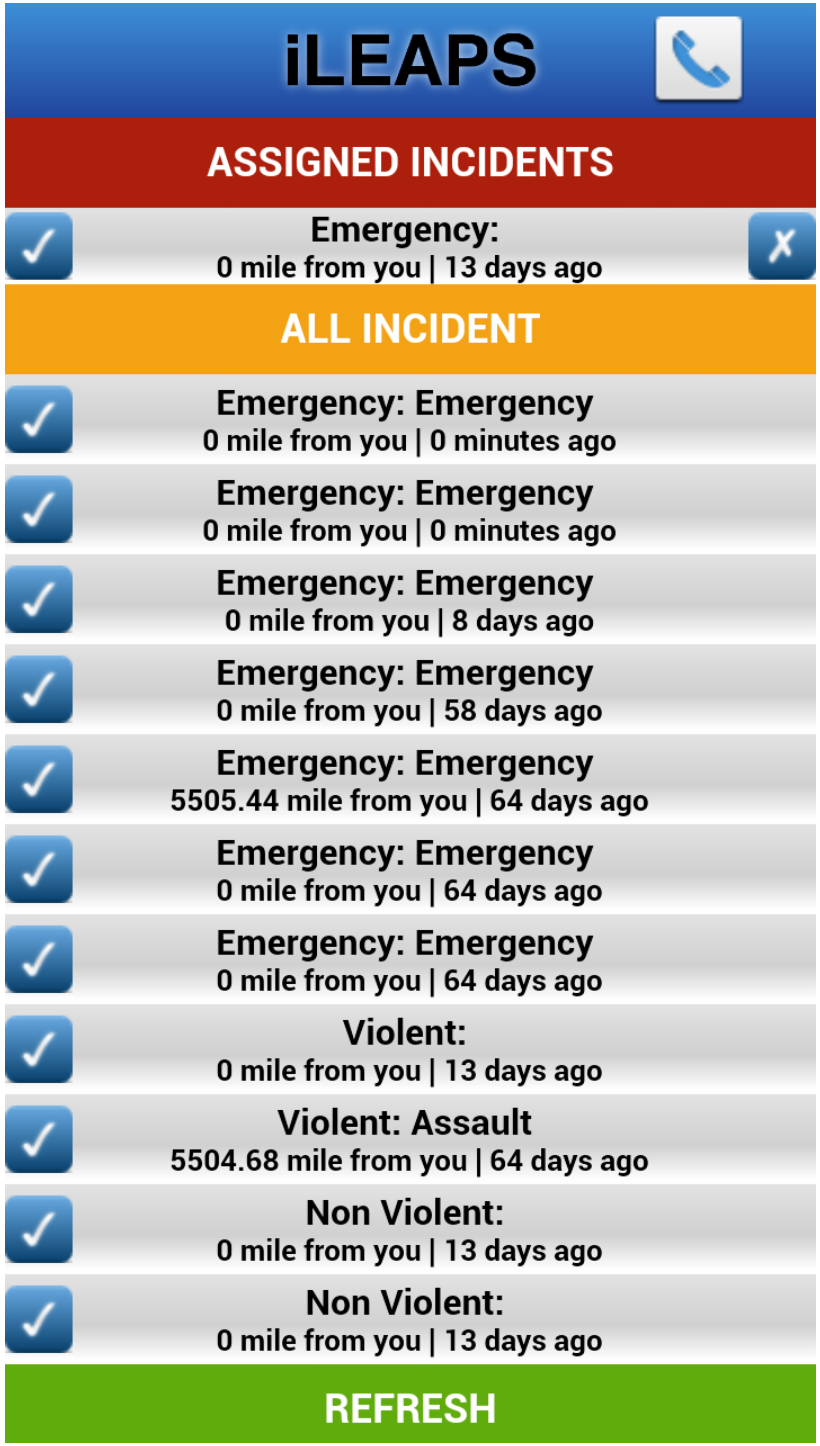

12:45 PM
Person Reporting
Interface

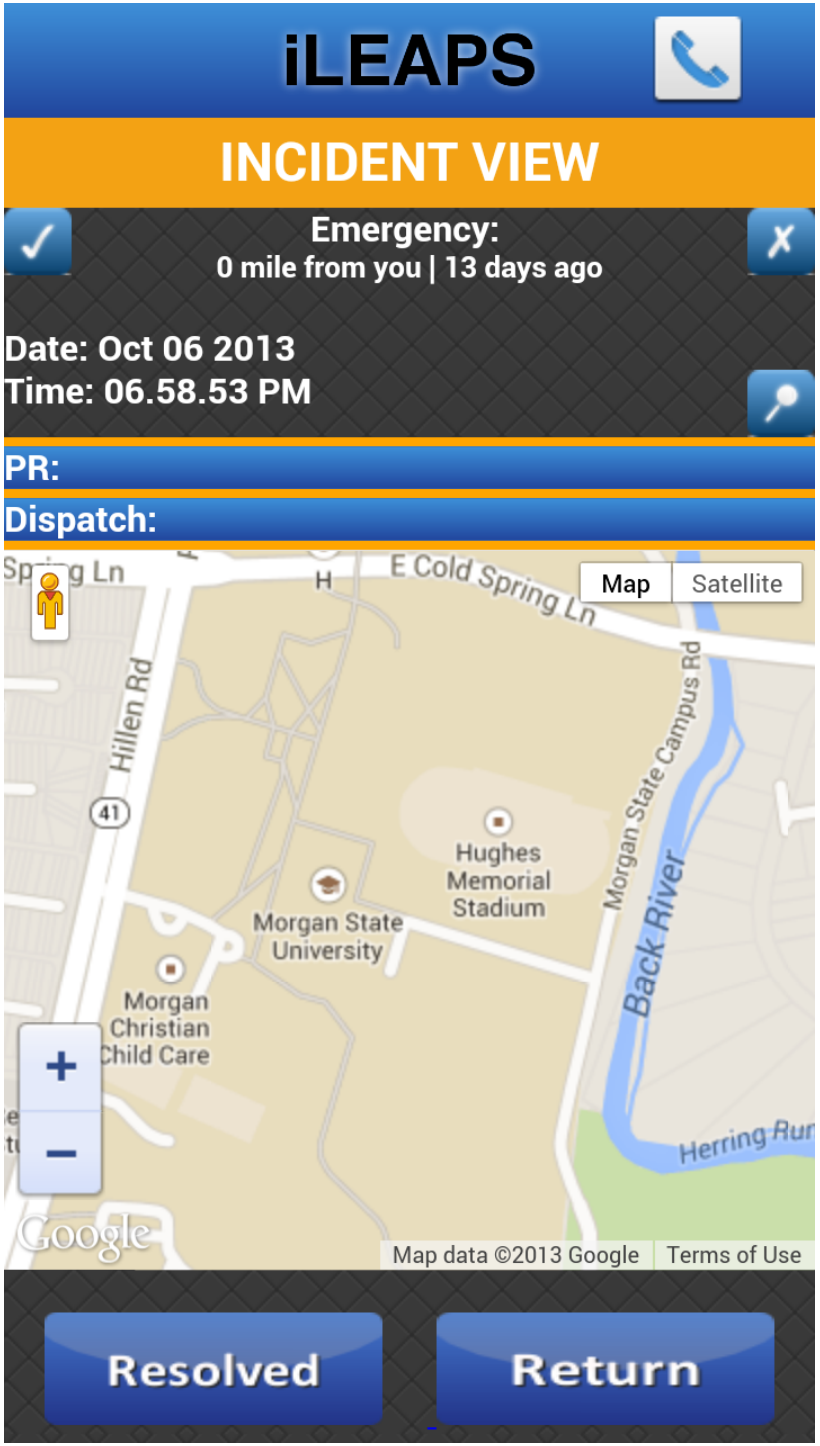

iLEAPS
Emergency will be reported in...
7
Touch to cancel

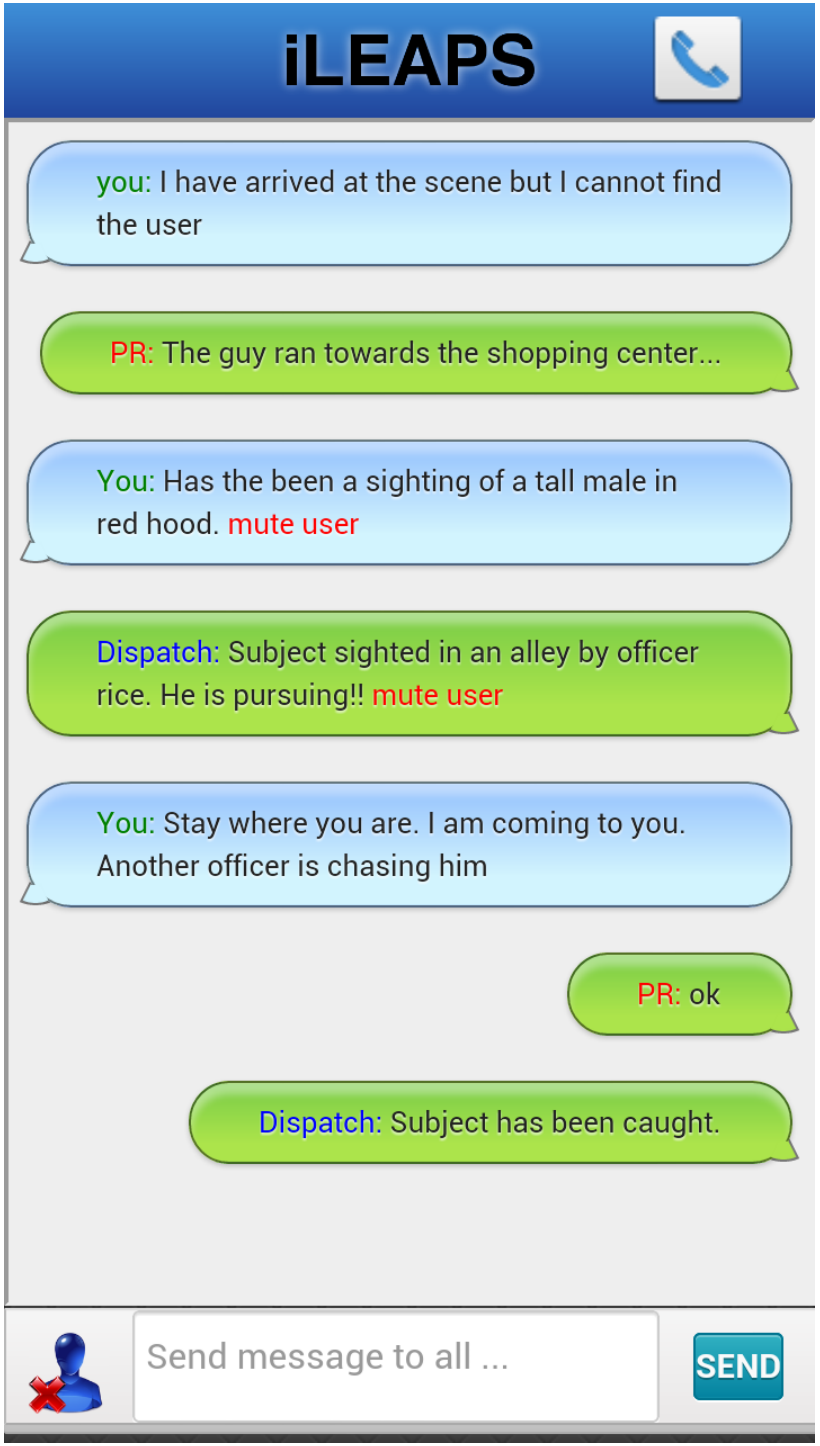

Emergency
Voilent Incidents
Domestic Violence
Fighting
Harassment or Bullying
Homicide (Murder)
Robbery
Sexual Assault
Non-Voilent Incidents
Police Escort
iLEA
"Emergency" incident
dispatch and the


iLEAPS
Describe the Harassment Incident
My room mates boyfriend is drunk and refuses to leave her room
I would rather be anonymous
Send


12:44 PM
Police Officer
Interface


iLEAPS
ASSIGNED INCIDENTS
Emergency:
0 mile from you | 13 days ago
ALL INCIDENT
Emergency: Emergency
0 mile from you | 0 minutes ago
Emergency: Emergency
0 mile from you | 0 minutes ago
Emergency: Emergency
0 mile from you | 8 days ago
Emergency: Emergency
0 mile from you | 58 days ago
Emergency: Emergency
5505.44 mile from you | 64 days ago
Emergency: Emergency
0 mile from you | 64 days ago
Emergency: Emergency
0 mile from you | 64 days ago
Violent:
0 mile from you | 13 days ago
Violent: Assault
5504.68 mile from you | 64 days ago
Non Violent:
0 mile from you | 13 days ago
Non Violent:
0 mile from you | 13 days ago
REFRESH


iLEAPS
INCIDENT VIEW
Emergency:
0 mile from you | 13 days ago
Date: Oct 06 2013
Time: 06.58.53 PM
PR:
Dispatch:
Map
Satellite
E Cold Spring Ln
Hillen Rd
41
Morgan State Campus Rd
Hughes Memorial Stadium
Morgan State University
Morgan Christian Child Care
Back River
Herring Run
Google
Map data ©2013 Google
Terms of Use
Resolved
Return


iLEAPS
you: I have arrived at the scene but I cannot find the user
PR: The guy ran towards the shopping center...
You: Has the been a sighting of a tall male in red hood. mute user
Dispatch: Subject sighted in an alley by officer rice. He is pursuing!! mute user
You: Stay where you are. I am coming to you. Another officer is chasing him
PR: ok
Dispatch: Subject has been caught.
Send message to all ...
SEND

Coordinating Views within Interactive Web Geovisualization (#1380)

F. Hardisty , A. Savelyev
Pennsylvania State University Geography, 302 Walker Bld, University Park, PA 16802, United States

Geovisualization methods focus on the interactive, exploratory mapping of phenomena represented as numerical and textual data. Geovisualization often requires the coordination of multiple views of this data. However, it is not clear what the best methods are for developing coordinated views in JavaScript, the programming language present in the largest number of computing devices, including mobile phones. We present here a new open source library for easing the development of coordinated views called GeoVizCoordinator.js. GeoVisCoordinator.js combines current best practices in JavaScript with past research into geographic visualization methods in Java in order to ease the development and coordination of multiple kinds of views in a web context.

Past research on coordinated views for geovisualization lead to the development of a number of object patterns, including the meta-pattern called "introspective observer coordination". This type of coordination relies on introspection and reflective invocation to allow for extensible, efficient coordination between views, including coordinating selections, spatial extents, and visual variables. However, the differing semantics of JavaScript mean that the Java methods cannot be directly applied to JavaScript.

We undertook a comprehensive review of current best practices in JavaScript coordination libraries in order to develop the best ap-proach for GeoVizCoordinator.js, including libraries that use model-view, publish-subscribe, and other object and function patterns that can be used for coordination. Model-View techniques include Model-View-Controller (MVC), Model-View-ViewModel (MVVM) and Model-View-Presenter (MVP) patterns. These widely used patterns help separate the data from the presentation of the data, without the various components needing to be aware of each other. A related technique that shares this goal is the publish-subscribe or observer pattern, which is implemented in a number of leading JavaScript frameworks. We abstract the core of these libraries for use in GeoVizCoordinator.js.

For future work, we also consider how new kinds of coordination are made possible in a web context. These include coordination between remote actors, coordination between servers and clients, and coordination between clients and web services represented as REST endpoints. Coordination between remote actors is possible with current approaches, but cumbersome to develop. We examine how GeoVizCoordinator.js can be extended using techniques developed in other libraries such as GeoJabber, and using standard communication protocols like XMPP to enable these kinds of coordination.

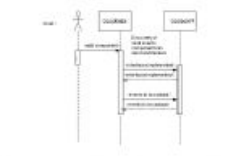


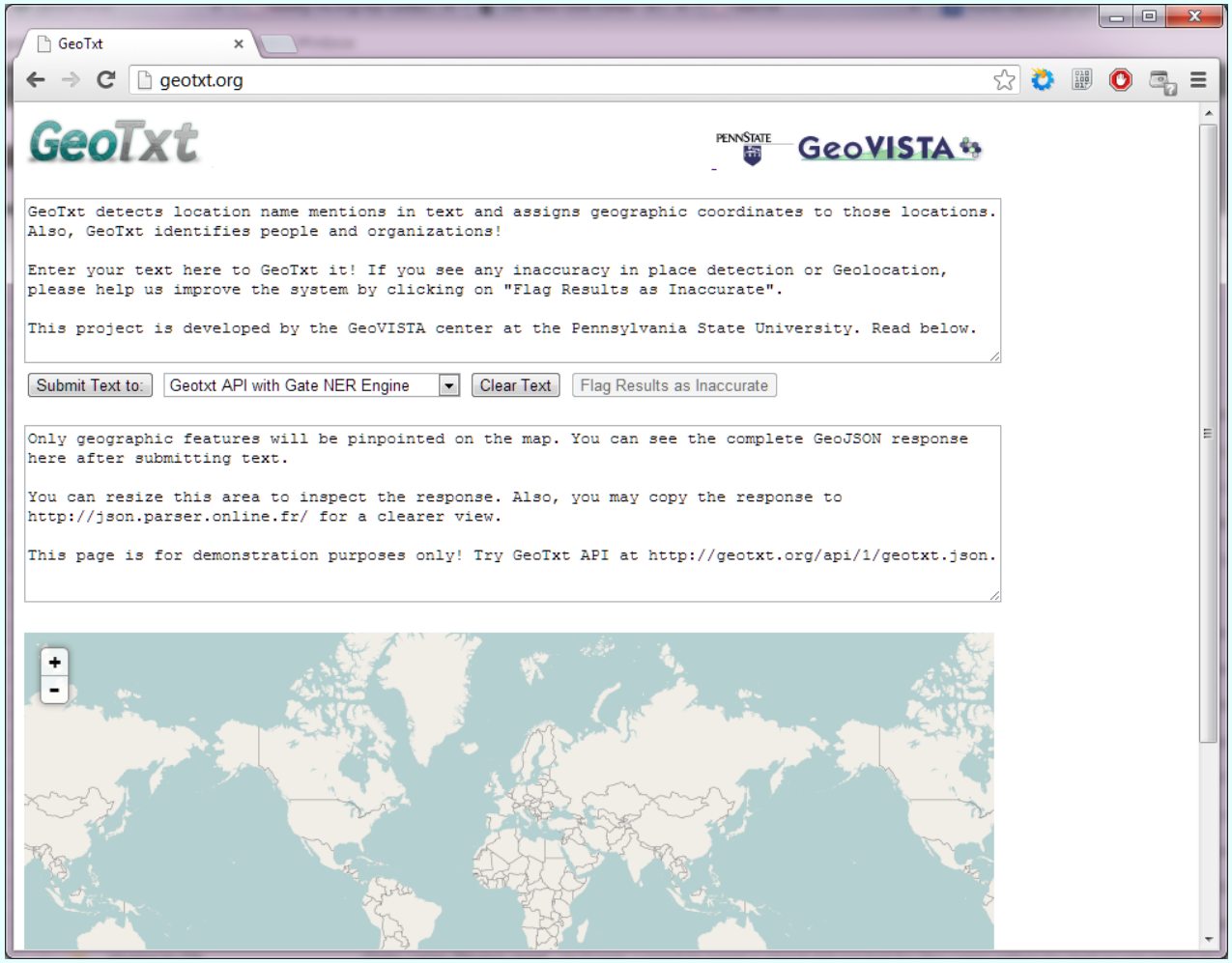
Figure 1:
UML diagram illustrating coordinator program flow

Keywords: geovisualisation, geovisualization, JavaScript Models

Designing a Web Service to Geo-Locate Subjects of Volunteered, Textual Geographic Information

R. Mullins, F. Hardisty, S. Pezanowski, S. Das, A. Savelyev, A. MacEachren, P. Mitra, A. Jaiswal, J. Wallgrün, M. Karimzadeh

GeoTxt.org


GeoTxt
geotxt.org
GeoTxt
PENNSTATE
GeoVISTA
GeoTxt detects location name mentions in text and assigns geographic coordinates to those locations.
Also, GeoTxt identifies people and organizations!
Enter your text here to GeoTxt it! If you see any inaccuracy in place detection or Geolocation,
please help us improve the system by clicking on "Flag Results as Inaccurate".
This project is developed by the GeoVISTA center at the Pennsylvania State University. Read below.
Submit Text to:
Geotxt API with Gate NER Engine
Clear Text
Flag Results as Inaccurate
Only geographic features will be pinpointed on the map. You can see the complete GeoJSON response
here after submitting text.
You can resize this area to inspect the response. Also, you may copy the response to
http://json.parser.online.fr/ for a clearer view.
This page is for demonstration purposes only! Try GeoTxt API at http://geotxt.org/api/1/geotxt.json.
+
-

Outline

What GeoTxt Does
How it works
Demo + API
Problems
Going forward
Alternatives
Thanks

Turn text into spatial locations

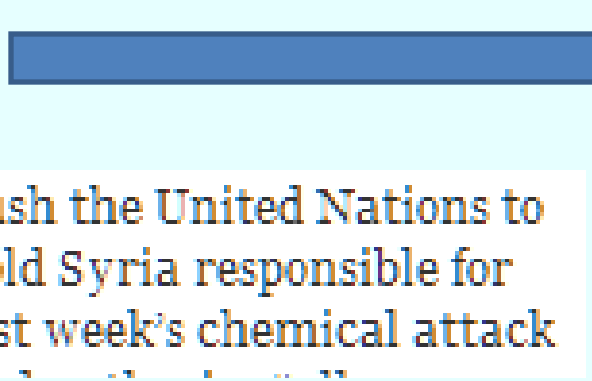
Processing pipeline

Named Entity Recognition (NER)

Find the locations in the text

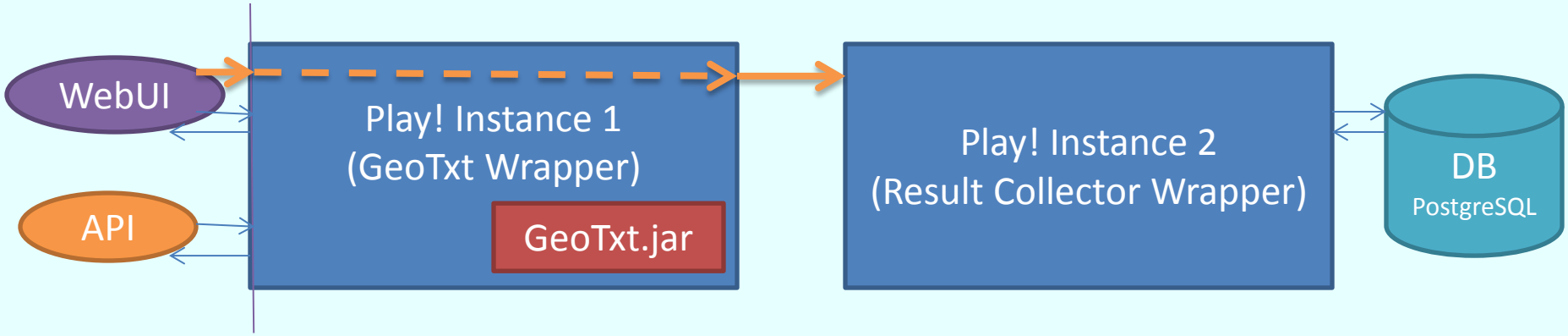
Geocoding Disambiguation Gazetteer

Turn the locations into coordinates

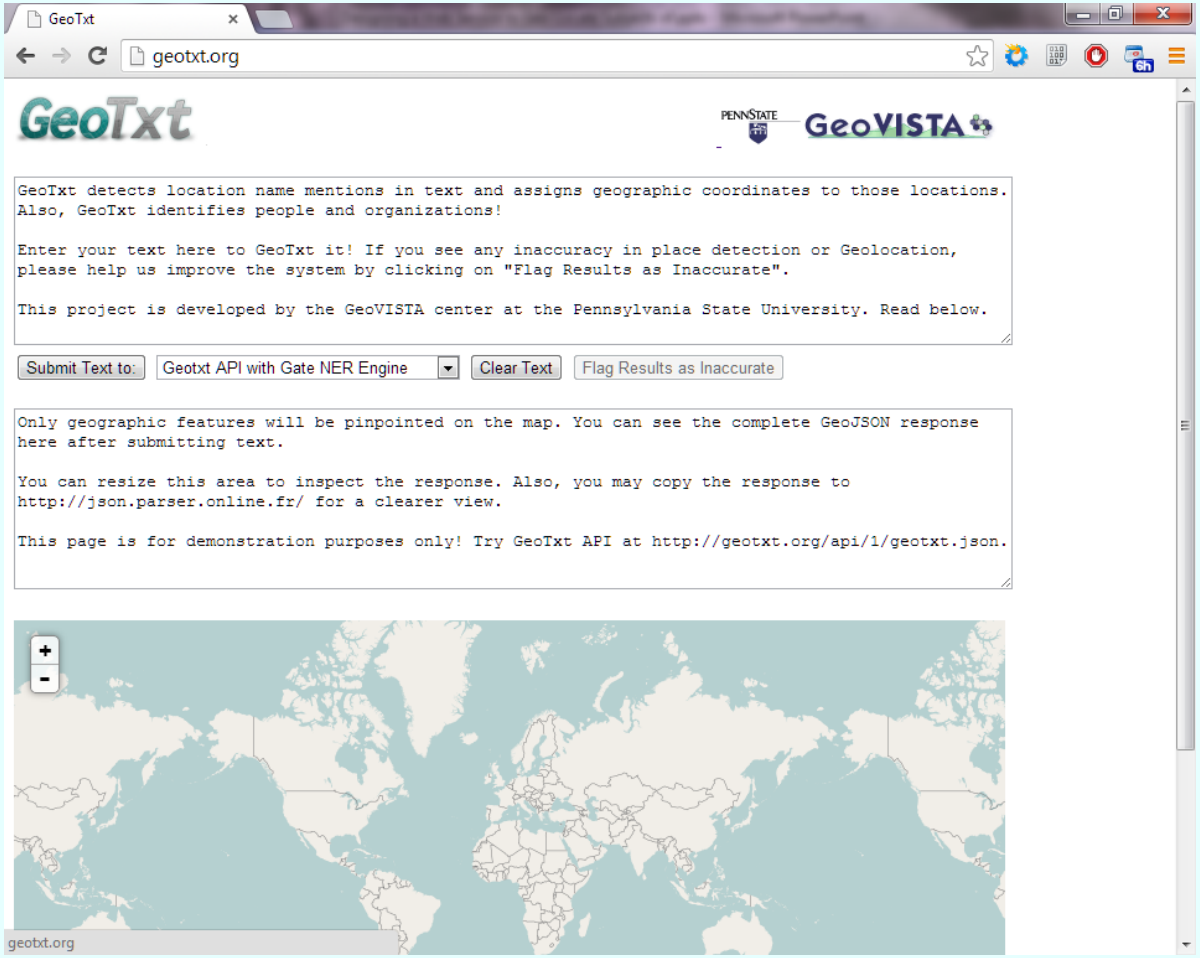




Information architecture



Demo + API


GeoTxt
geotxt.org
GeoTxt
PENNSTATE
GeoVISTA
GeoTxt detects location name mentions in text and assigns geographic coordinates to those locations.
Also, GeoTxt identifies people and organizations!
Enter your text here to GeoTxt it! If you see any inaccuracy in place detection or Geolocation,
please help us improve the system by clicking on "Flag Results as Inaccurate".
This project is developed by the GeoVISTA center at the Pennsylvania State University. Read below.
Submit Text to:
Geotxt API with Gate NER Engine
Clear Text
Flag Results as Inaccurate
Only geographic features will be pinpointed on the map. You can see the complete GeoJSON response
here after submitting text.
You can resize this area to inspect the response. Also, you may copy the response to
http://json.parser.online.fr/ for a clearer view.
This page is for demonstration purposes only! Try GeoTxt API at http://geotxt.org/api/1/geotxt.json.
+
-
geotxt.org

Good things

- It works!
- It scales – for more throughput, add more machines
- Extensible architecture
- Play Framework is easy-(ish)
- Akka under the hood (possibility for actor-based programming)

Problems

Definition
Disambiguation
Accuracy

Definition:
Which texts count as locations?

Dresden, Germany
Europe
Bayern Munich football team
Mayor of Dresden
Angela Merkel
U.S. Dollar

Disambiguation:
Which location do you mean?



Photo credit: http://www.flickr.com/photos/foilman/2803261256/

Disambiguation:
Which location do you mean?

London -- ?

London, UK
London, Ontario

San José (or San Jose)

How Many?
At least 1,716 (on FIPS55 and GEOnet in 2006)

Accuracy

Sometimes, it doesn’t work!
NER problems – Twitter is different
Disambiguation problems

Going forward – context!

Context – other places
Context – mentions of people, organizations
Context – sources with location built in
Context – language use???

Going forward – better NER!

Stanford NER is a Conditional Random Field (CRF) classifier
Training data is super important
Developing Twitter training data using PI labor, and Amazon Mechanical Turk

Going forward – more gazetteers!

GeoNames is good but it’s not everything

OpenStreetMap

Wikipedia

???

Alternatives

- Two steps: NER and Geocoding
- NER
  - Stanford NLP
  - GATE
- Geocoding
  - GeoNames webservice
  - GeoNames local db
- Clavin (especially on Amazon Cloud)
- Google, Yahoo APIs if you can respect TOS (or are willing to ignore them).
- OpenCalais

Thanks to co-authors

At Penn State:

- Geography
- Information Science and Technology

Thanks to sponsors

- VACCINE (Visual Analytics for Command, Control, and Interoperability Environments)
- NGA University Research Initiative (NURI)
- Penn State Online GIS Programs
- Department of Geography
- GeoVISTA Center

Thanks to you!

Exploratory Visualization of Movement Patterns in Twitter Data

Alexander (Sasha) Savelyev
Alan MacEachren

The Pennsylvania State University

Introduction

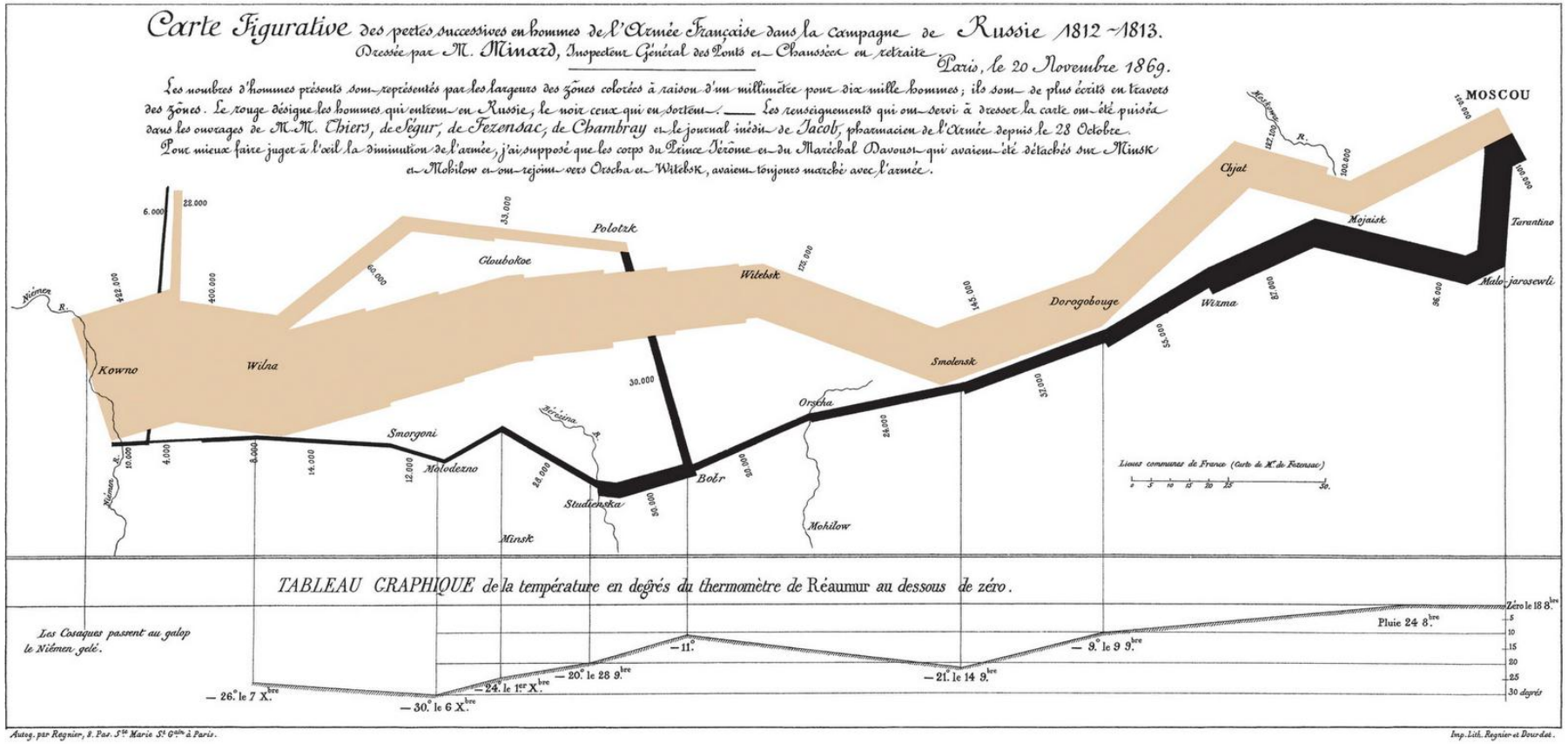
I work at GeoVISTA Center at PennState.

GeoVISTA's main emphasis is geovisualization.

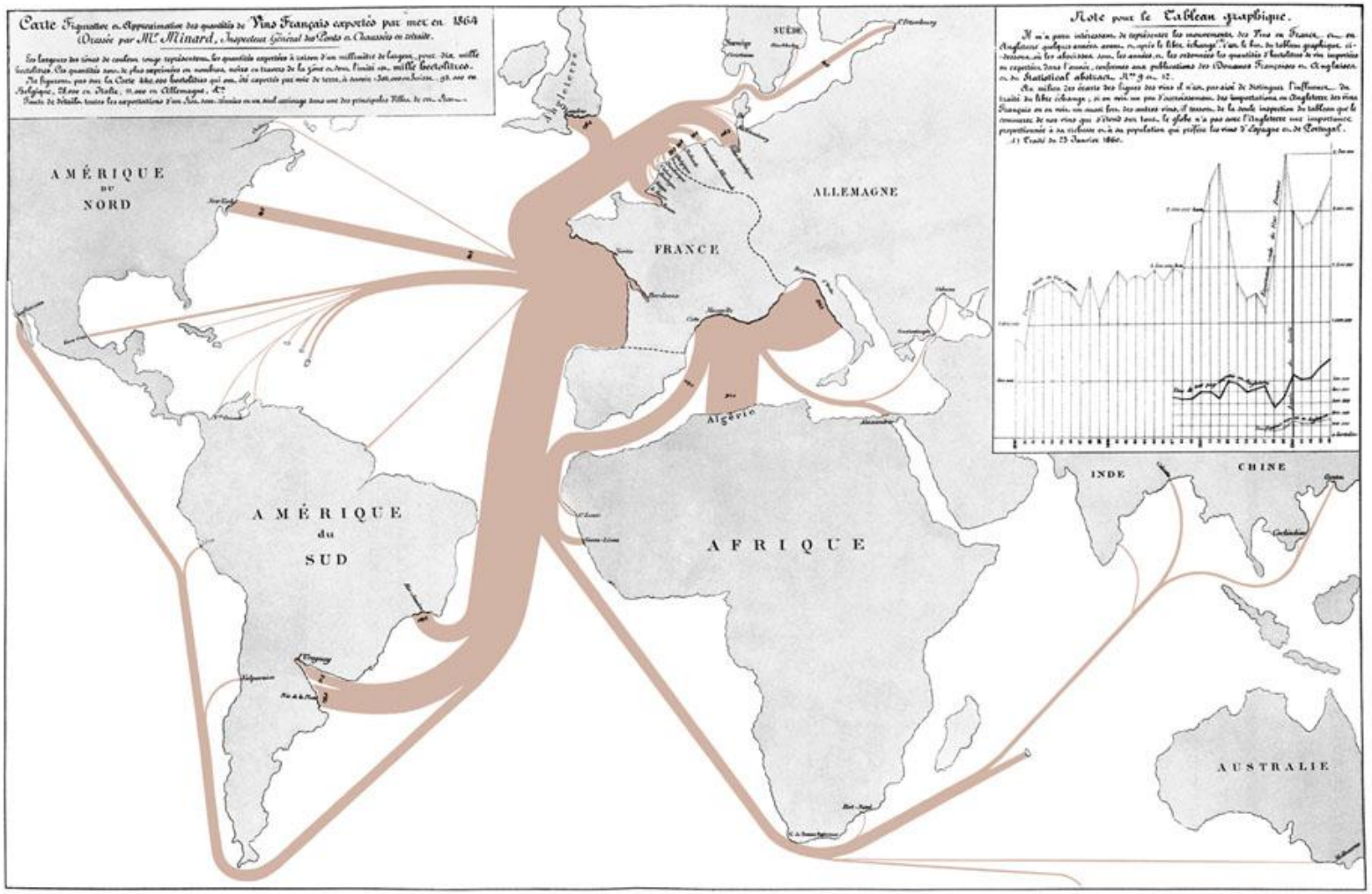
We have multiple collaboration projects with folks with all kinds of technical and theoretical backgrounds.

A recent project set our focus on analysis and visualization of *movement data*.

Maps of Movement



Maps of Movement



Charles Joseph Minard, *Tableaux Graphiques et Cartes Figuratives de M. Minard*, 1845-1869, a portfolio of his work held by the Bibliothèque de l'École Nationale des Ponts et Chaussées, Paris.

Movement Data Galore

Mass computerization led to explosive growth in the volume of recorded information.

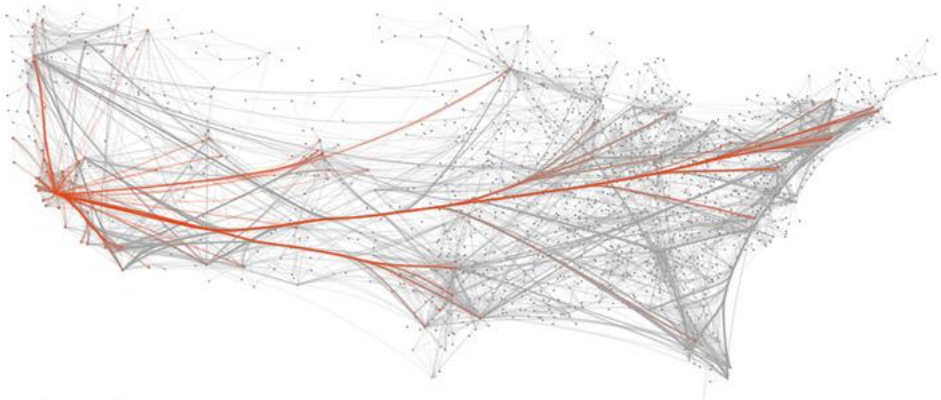
- A fitting metaphor is “digital exhaust”.
- A lot of it is geographic in nature.

Most popular example is Social Media.

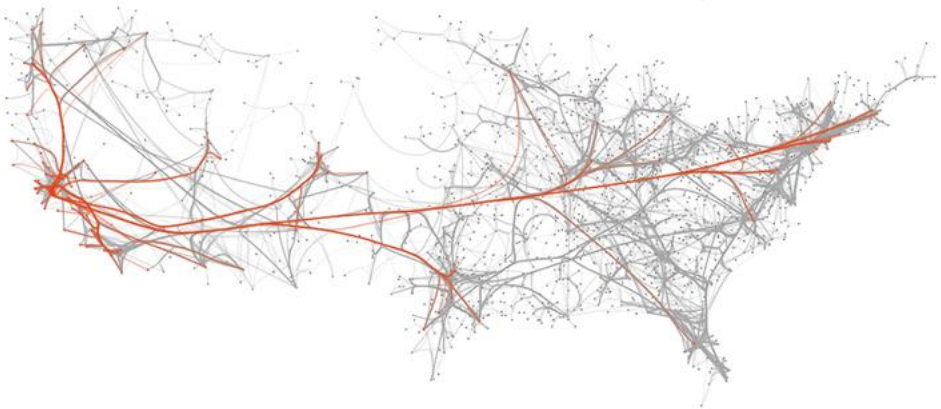
- E-ZPass and license plate readers.
- Mass transit passes and bicycle rentals.
- Loyalty programs etc.

Maps of Movement Redux









Danny Holten, Jarke J. van Wijk

Implications of Limitations

Evaluating the degree of "truth" that can be attributed to such data is a discipline of its own.

Most obvious limitations:

- Spatial inaccuracy.
- Implicit social exclusion.
- Sampling divide between digital and real worlds.

Learning the Affordances

Choice of visualization metaphor relies on solid understanding of your data and its affordances.

Philosophies on the relationship between data and the map are plenty.

> Understanding the phenomena you are mapping is generally considered a boon.

Such understanding can be achieved in multiple ways. I opted for *visual* exploration of movement data as the approach of choice.

Movement Traces in Twitter Data

If visual exploration of Twitter data is the goal, what are the project requirements?

1. Support obscene volumes of data.
2. Support simple data formats (text is best).
3. Use Graphical User Interface (GUI).

The latter two points make this tool a little less exclusive to interested third parties.

Proposed Framework

Two stages – collection and visualization.

Collection implemented in *Node.js*:

- A total of 50 lines of code.
- Successfully tested on geography and sociology students.
- Quite scalable (up to 1 billion out of the box).

Visualization implemented in *Processing*.

More on that in a second.

Data Collection

1. Connect to Twitter Streaming API.
2. Keep only the latest tweet for each user.
3. Whenever the same person tweets again,
    1. add a new movement record to the output file,
    2. switch to the latest tweet.

1367853020000 5 1 1 39.718876 -75.077048 39.718875 -75.077047

"3:10 PM May 6, 2013" "5 minutes", "about 1 MPH", "about 1 mi"

Visualization Framework

Mainstream GIS are a poor fit for the task:

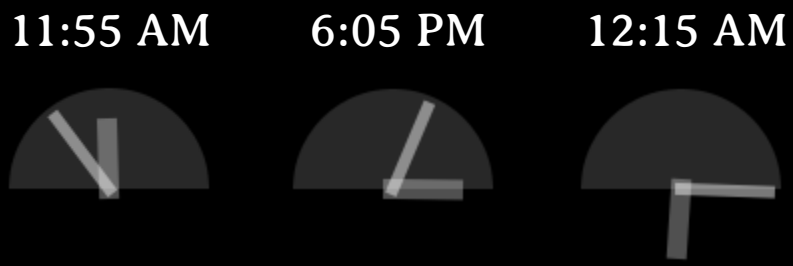
Today's *small* demo has 250 GB of data, or about 100,000,000 movement records total.

Solution inspired by the VHS tape player:

- Data "tape" can be of any length, as only the current "frame" is shown.
- The number of movement records shown "per frame" has no impact on performance.

User Manual, Unabridged

24-hour clock:



Adjust-A-Dial:



Demo!

Steve Jobs would have been proud of this slide.

GeoTxt: A Web API to Leverage Place References in Text

Morteza Karimzadeh[1,2], Wenyi Huang[3], Siddhartha Banerjee[1,3], Jan Oliver Wallgrün[1,2], Frank Hardisty[1,2], Scott Pezanowski[1,2], Prasenjit Mitra[1,3] and Alan M. MacEachren[1,2]

1) GeoVISTA Center, 302 Walker Building, University Park, PA, 16802 +1-814-865-3433 {karimzadeh, wallgrun, hardisty}@psu.edu

2) Department of Geography, The Pennsylvania State University, 302 Walker Building, University Park, PA 16802 +1-814-865-3433 {scottpez,maceachren}@psu.edu

3) College of Information Sciences and Technology, 332 Info Science and Tech University Park, PA 16802, USA +1-814-865-3528 {wzh112,sub253,pmitra}@ist.psu.edu

ABSTRACT

Associating place name mentions in unstructured text with their actual references in geographic space is vital to enable spatial queries and analysis. In this paper, we introduce GeoTxt, a web API plus human-usable web tool designed and implemented to tackle three components of place-reference processing from text, namely: extraction, disambiguation, and geolocation of place names mentioned in unstructured text. Current GeoTxt development is focused particularly on support for processing short microblog posts.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithm, Design, Performance

Keywords

Geocoding, Geographic Information Retrieval, Natural Language Processing, Geographic Information Systems

1. INTRODUCTION

While the volume of textual data with explicit geo-location is increasing rapidly due to GPS-enabled devices and sensors of many kinds, an even larger source of place-based information exists in text artifacts ranging from microblogs, through news stories and press releases, to scientific publications. The GeoTxt API has been designed to support extraction, disambiguation, and geolocation of place entities in text submitted to the API from other applications. The primary focus is on extracting place references from microblog posts, partly similar to the goals of [2; 6; 8] as opposed to most other efforts such as [3-5; 9] that address similar issues but focus on longer and more grammatically correct text artifacts. This is a challenging task due to the limited context in these short posts (e.g., 140 character limit in Twitter), the related use of abbreviations, and the non-standard syntax often used (e.g., words are often not capitalized as they would be in standard text). Nevertheless, users will be able to indicate the nature of query text to get the best results for either microblog posts or longer text articles. Also, GeoTxt includes a human-usable interface for processing individual text artifacts and testing the methods.

Below, we outline the GeoTxt API approach and system specifications, detail how it works, and point to future work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
GIR'13, November 05 2013, Orlando, FL, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2241-6/13/11$15.00.
http://dx.doi.org/10.1145/2533888.2533942

2. GEOTXT ARCHITECTURE AND CAPABILITIES

GeoTxt has been designed and implemented as an easy to use RESTful Web API. It identifies mentions of place names in unstructured text, and assigns geographic coordinates to those place names. Trusted applications are able to query the service with HTTP GET requests and receive the responses as GeoJSON FeatureCollection objects containing geocoded place names along with persons and organizations identified in free form text.

Figure 1 shows the schematic architecture of the GeoTxt API. GeoTxt, written in Java, processes input text in two separate steps, which work independently in the current release. At the first step, Named Entity Recognition is performed to extract place names. Second, those place names found in text are disambiguated and geolocated to their respective geographic coordinates by the GeoCoder module.

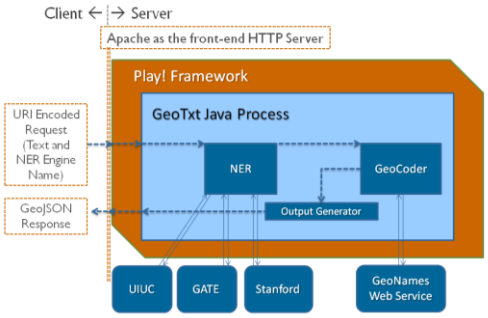


Figure 1. GeoTxt API Architecture.

The Illinois Named Entity tagger [7] (depicted as UIUC in Figure 1) has been tested and integrated with the system; however, its current beta release has been considerably slower than GATE ANNIE [1] and Stanford NER in terms of computation time. Because GeoTxt is designed as a web API backend for big data processing, the two faster NER engines i.e. Stanford NER and Gate ANNIE are used in the current release.

After the identification of place names mentioned in text, GeoTxt uses the GeoNames geographic database (http://geonames.org) to

help geolocate those names. Although the GeoNames database is open source and available for free, its ranking mechanism is not public. Also, the GeoNames Search Web Service does not always come up with the ranking a human agent expects, e.g. it ranks *Colorado springs* and *San Luis Río Colorado* higher than *Colorado* (the State) when it is queried with the text *Colorado*. Therefore, we use (a) the geographic level, e.g. country, province, city of the place name in text when provided by the NER engine as best guess to initially rank and distinguish between candidates in the database, and then (b) the Levenshtein Distance of the name mentioned in text and the candidate's name (which indicates how close the two strings are) to choose the candidate with the least distance, and when multiple candidates have the same shortest distance to the name in question (c) the population of potential candidates with higher priority given to places with higher population to choose the best candidate.

The biggest challenge in geolocation is to disambiguate between places with identical names. The current implementation of GeoTxt leverages spatial logic to overcome such ambiguity in case more than one place name is mentioned within the same document. In such circumstances, GeoTxt retrieves the top five candidates for each place name mentioned in text, and for each candidate, retrieves all entities higher up in the geographical hierarchy. Then, for each name, the candidate for which an entity higher in its hierarchy matches any other place name mentioned in the text is picked. For example, in the tweet "*Finally landing in London. I love Canada*!", London will be geolocated to London, Ontario instead of London, UK; although the latter has higher population and stands higher in GeoNames ranking.



Figure 2. GeoTxt web User Interface.

Human users of GeoTxt have access to the same system through a visual web interface (see Figure 2). Users are able to paste in a piece of text, select the NER engine, and see the raw GeoJSON response in a text box and also geocoded locations overlayed on a base map. Each location is labeled with both the name that appears in text and the one picked by the GeoCoder module (in case they are different, whether due to a mistake or due to name abbreviation etc.) to help the user detect inaccurate results. Users are then able to flag any erroneous results as inaccurate. Such results are being used to monitor possible mistakes and improve the performance of GeoTxt. In the current release of GeoTxt, users (application and human) are able to select between two Named Entity Recognition engines of Stanford NER and GATE ANNIE to extract locations, people and organizations. But, the system has been designed to allow for additional NER engines to be added and compared, and also for versions of individual NER engines trained on different kinds of text to be selected. Enabling user feedback and the comparison of multiple NER engines are two unique features of GeoTxt compared to other efforts such as [3; 5; 9].

Play! Framework (http://www.playframework.com/) is used to expose GeoTxt functionality as a web API and to render User Interface and documentation web pages. The API provides versioned web services to guarantee backward compatibility.

3. FUTURE WORK

Plans for the future include customizing methods and NER tools to more comprehensibly utilize context in text and also context specific to microblog platforms (e.g. Twitter generated metadata) for the purpose of disambiguation, experimenting with ensemble approaches that combine the best of multiple methods, and utilizing detailed user feedback to improve the results. We are also starting to build a corpus of hand annotated microblogging posts to train the NER tools on such posts and to assess the overall system.

4. ACKNOWLEDGMENTS

This material is based in part upon work supported by the U.S. Department of Homeland Security under Award #2009-ST-061-CI0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

5. REFERENCES

[1] Cunningham, H., 2002. GATE, a general architecture for text engineering. *Computers and the Humanities 36*, 2, 223-254.

[2] Kitamoto, A. and Sagara, T., 2012. Toponym-based geotagging for observing precipitation from social and scientific data streams. In *Proceedings of the Proceedings of the ACM multimedia 2012 workshop on Geotagging and its applications in multimedia* (Nara, Japan2012), ACM, 2390799, 23-26. DOI= http://dx.doi.org/10.1145/2390790.2390799.

[3] Leetaru, K.H., 2012. Fulltext Geocoding Versus Spatial Metadata for Large Text Archives: Towards a Geographically Enriched Wikipedia. *D-Lib Magazine 18*, 9, 5.

[4] Leidner, J.L., 2007. Toponym resolution in text.

[5] Lieberman, M.D. and Samet, H., 2012. Adaptive context features for toponym resolution in streaming news. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* ACM, 731-740.

[6] Lingad, J., Karimi, S., and Yin, J., 2013. Location extraction from disaster-related microblogs. In *Proceedings of the 22nd international conference on World Wide Web companion* International World Wide Web Conferences Steering Committee, 1017-1020.

[7] Ratinov, L. and Roth, D., 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* Association for Computational Linguistics, 147-155.

[8] Schulz, A., Hadjakos, A., Paulheim, H., Nachtwey, J., and Muhlhäuser, M., 2013. A Multi-Indicator Approach for Geolocalization of Tweets. In *Seventh International AAAI Conference on Weblogs and Social Media*.

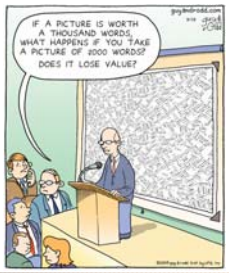
[9] Speriosu, M. and Baldridge, J. Text-Driven Toponym Resolution using Indirect Supervision.

Geo-Visual Analytics & 'Big' Data: Leveraging Geo-Information in Unstructured Data

Alan M. MacEachren

GeoVISTA Center
Department of Geography
Penn State University
maceachren@psu.edu

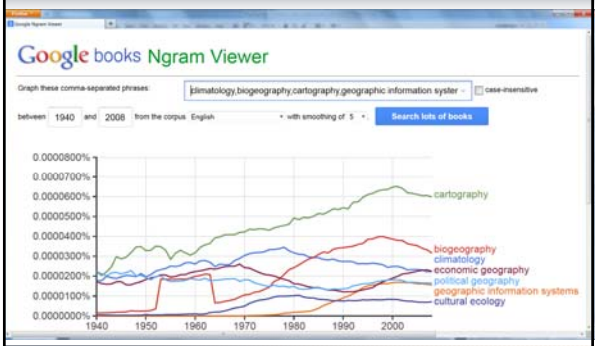
Geo-Visual Analytics & 'Big' Data: Leveraging Geo-Information in Unstructured Data



Alan M. MacEachren

GeoVISTA Center
Department of Geography
Penn State University
maceachren@psu.edu

…. and what happens with 500 billion words? 'mining' books for sub-discipline trends



'mining' books: sensitivity to queries



Outline

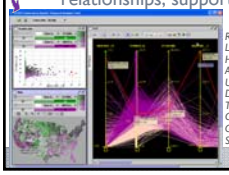
1) **Geo-visual analytics** (a)**, Big Data** (b)**, & reasoning** (c)
2) **a key challenge:** finding and extracting geographic information from unstructured "data" (with a focus on place references in text)
   a) statements about movement: GeoCAM
   b) leveraging fragmentary microblog posts: SensePlace2
3) **closing ideas**

1.a: GeoVisual Analytics

from cartography, through geovisualization, to geovisual analytics:

data → info → knowledge → knowledge application

- (late '80s) **Visualization in Scientific Computing (VisSC)**: prompting shift from **cartography → geovisualization** –
- ('90s) dynamic **geovisualization** – (Cartography + VisSC + EDA + HCI + InfoVis): prompting hypotheses &enabling insight
- (2000s) **geovisualization + computation** – (…) + KDD & statistical methods: turning data to info, uncovering patterns and relationships, supporting knowledge construction



Robinson, A.C., Chen, J., Lengerich, E.J., Meyer, H.G. and MacEachren, A.M. 2005: Combining Usability Techniques to Design Geovisualization Tools for Epidemiology. Cartography and Geographic Information Science 32, 243-255.



Chen, J., MacEachren, A.M. and Guo, D. 2008: Supporting the Process of Exploring and Interpreting Space-Time Multivariate Patterns: The Visual Inquiry Toolkit. Cartography and Geographic Information Science 35, 33-50.

1.a: GeoVisual Analytics

from cartography, through geovisualization, to geovisual analytics:

data → info → knowledge → knowledge application

- (late '80s) **Visualization in Scientific Computing (VisSC)**: prompting shift from **cartography → geovisualization** –
- ('90s) dynamic **geovisualization** – (Cartography + VisSC + EDA + HCI + InfoVis): prompting hypotheses &enabling insight
- (2000s) **geovisualization + computation** – (…) + KDD: turning data to info, uncovering patterns and relationships, supporting knowledge construction
- **(geo)visual analytics →**

Geovisual Analytics

- *Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces* (quote from: Thomas, J.J. and Cook, K.A., editors 2005: *Illuminating the Path: The Research and Development Agenda for Visual Analytics. Los Alamos, CA: IEEE Computer Society.)*

  … thus, the focus is on assembling evidence, generating inferences and explanations from evidence, comparing / assessing those inferences and explanations, and reporting results
- "Visual analytics combines automated analysis techniques with interactive visualisations for effective understanding, reasoning and decision making on the basis of very large and complex datasets" *(quote from Keim, Kohlhammer, Ellis and Mansmann, editors, 2010, Mastering the Information Age: Solving Problems with Visual Analytics, Eurographics Association)*
- ***Geo***visual analytics focuses on visual interfaces to analytical methods that support *reasoning* with and about (big, *heterogeneous, messy*) geo-information – *to enable insights and decisions about something for which place matters*

Geovisual analytics: 3 challenges

- *scaling humans to cope with big data*
- *extracting meaningful information from messy, unstructured, and/or heterogeneous data*
- *understanding and supporting the analytical reasoning processes needed to utilize the data*

Scaling analysts to work w/ big data



1.b: Big Data

- **Volume** – as of 2012, 2.5 exabytes (2.5×10^{18})/day
- **Variety** – heterogeneous (diverse sources & kinds of information)
- **Velocity** – real-time streaming
- **Variability** – inconsistencies
- **Complexity** – of connections among data & problems addressed

1,250,000,000,000 (1.25 trillion) pdf pages

EU Commission, Brussels, 07 November 2013:"Every minute the world generates 1.7 million billion bytes of data, equivalent to 360,000 standard DVDs"



source: Oracle, 2012 – growth of business data

Geovisual analytics: scaling analysts *to cope with information deluge*

Big Data



The New York Times **SundayReview** | The Opinion Pages

The Age of Big Data

GOOD with numbers? Fascinated by data? The sound you hear is opportunity knocking.

global development

Global Pulse, a new initiative by the United Nations, wants to leverage Big Data for global development. The group will conduct so-called sentiment analysis of messages in social networks and text messages — using natural-language deciphering software — to help predict job losses, spending reductions or disease outbreaks in a given region. The goal is to use digital early-warning signals to guide assistance programs in advance to, for example, prevent a region from slipping back into poverty.

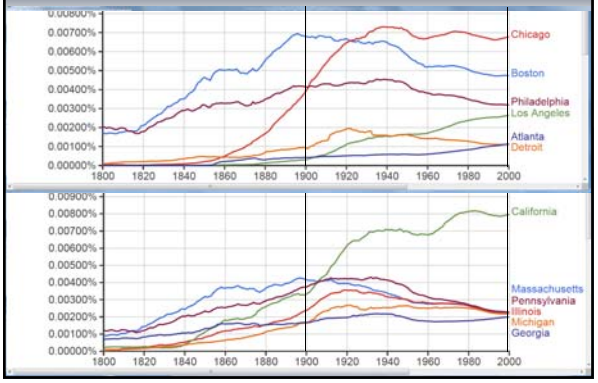
real estate & economy

In economic forecasting, research has shown that trends in increasing or decreasing volumes of housing-related search queries in Google are a more accurate predictor of house sales in the next quarter than the forecasts of real estate economists. The Federal Reserve, among others, has taken notice. In July, the National Bureau of Economic Research is holding a workshop on "Opportunities in Big Data" and its implications for the economics profession.

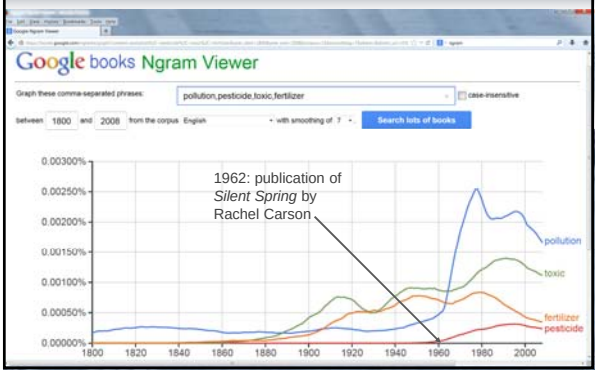
‘mining’ books for place references: cities



‘mining’ books for place references: cities versus states



‘mining’ books for risk-result relationships



1.c: reasoning: an example



Dilbert reasoning

1.c: ‘Reasoning’: in the context of geographic, phenomena, processes, and behavior

‘Reasoning’ research foci:

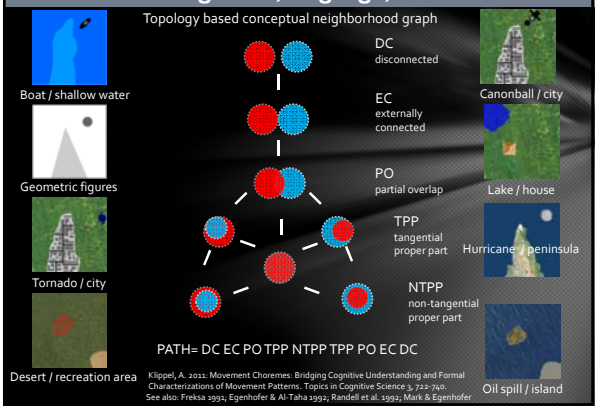
- enabling human spatial/geographic reasoning to extract data and derive information and knowledge (from unstructured sources)
- applying automated reasoning to retrieve or analyze geographic data (from unstructured sources) and contextualize the data to create information

2.a: Geographic Conceptualization for Accounts of Movement (GeoCAM)

- GeoCAM goal: *develop, implement, and assess an integrated approach to identifying, extracting, interpreting linguistic accounts of movement—in context*
- A key to the solution, we believe, is to *not* treat the problem as a natural language processing (NLP) problem *in isolation*.
- Integrated approach – *build on spatial-cognitive theory* – use geo-databases and related info iteratively, together with NLP methods, and visual analytics tools to interpret geo-references in ***place, time, and thematic context***.

Jaiswal, A., Zhang, X., Mitra, P., Pezanowski, S., Turton, I., Xu, S., Klippel, A. and MacEachren, A.M. 2011: GeoCAM: A Geovisual Analytics Workspace to Contextualize and Interpret Statements about Movement. Journal of Spatial Information Science 3, 65-101.

Motivation: Cognition, language, and movement



Motivation: Historical accounts of travel "South Africa, by Henry Balfour, 1910"

- Sat. 25 June - Left Oxford by 9.10 train + came on by 10.25 from Reading (via Newbury + Winchester). Secured cabin 18 on board "Kildonan Castle", met Lady Peck + her daughter who were seeing Peck + Benson off. Sailed soon after 4.30 p.m. Wind moderate, ship pitching somewhat in the evening. Before leaving S. Hampton, wrote to E. and to Lewis, Jo, Mother + K.B. + wired to Jo + E.
- Sun. 26 - Rounded Ushant before 8.30 a.m. Moderate sea, slight motion which increased in the evening. Ships very numerous.
- Mon. 27 - Dull, drizzling morning, clearing later. Passed Finisterre before breakfast, but did not sight the cape. Sea slight.
- Tues. 28 - Warm, sea slight, drifting sea mists + sun. Had long talk with Barry, the champion sculler + his trainer Phelps. Wrote to E. and to Jo-+-Lewis for posting in Funchal. Passed homeward mail ("Briton") at about 9.0 pm.
- Wed. 29 - Arrived Funchal at about 6 a.m. Cloudy day but fine. Ship coaling. Went ashore at 7.0 + walked out to Reid's Palace Hotel where I breakfasted. Very nice hotel in splendid situation. Walked about Funchal + wired to Hoernle at Cape Town about lectures. Had received letter from him + later one from E. Came on board at 10.30. Sailed from Funchal at noon. Ship rolled a good deal in afternoon + evening.

http://www.prm.ox.ac.uk/manuscripts/balfourdiaries1910.html

Motivation: From Norfolk to the Congo … but will Martin the cuckoo ever make it home?

- Britain's cuckoo population is in sharp decline, so ornithologists have tracked five migrating birds to try to find out why; *Robin McKie, The Observer, Saturday 21 April 2012 08.10 EDT*
- He has **flown thousands of miles across Africa and Europe** and enjoyed a winter break **in the Congo rainforest**. But now Martin the cuckoo is missing and scientists are waiting, with mounting concern, for news of the errant bird. Martin is one of a group of five cuckoos – the first to be fitted with satellite tags – which have revealed extraordinary details about the birds' **migration routes**. These **journeys** have taken them, over the past 12 months, **from Norfolk to France, Spain and Italy** and then, **by a variety of routes that have included crossings of the Sahara and voyages through west Africa, to the equatorial forests of Congo**, where the birds spent the winter. The variety and length of these **journeys** have astonished ornithologists who had little idea about the complexity and range of cuckoo migration until then.
- But now this band of feathered brothers is struggling to **make its way back to Britain** and has suffered casualties. "We have already lost one cuckoo," …
- Martin's transmitter has gone silent, leaving the trust's Cuckoo Project with only three known survivors: Lyster and Kasper, who are **holed up in Algeria**, while Chris has **made it to Italy**. …


Very little is known about cuckoo migration patterns. Photograph: Jonathan Gale/Getty Images

http://www.theguardian.com/environment/2012/apr/21/ornithologists-track-cuckoo-africa

GeoCAM: framing components of the approach

continuum of accounts of movement in text

wide range of other documents (e.g. Martin the Cuckoo) — situation report — travel journal/blog (e.g. Henry Balfour) — travel narrative — **written directions** — navigation service



a typical example

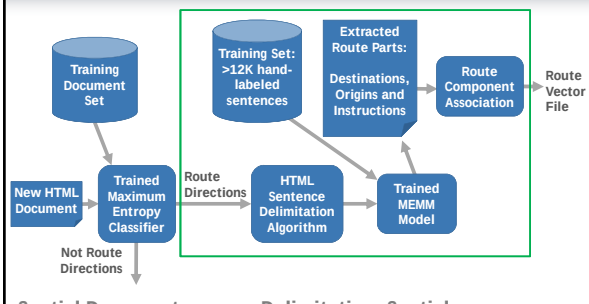
focus on constrained problem – *human-generated route directions*;

constrained:
- *linguistically* by focus on travel directions
- *geographically* by focus on directions within a network

Meaningful Route Parts

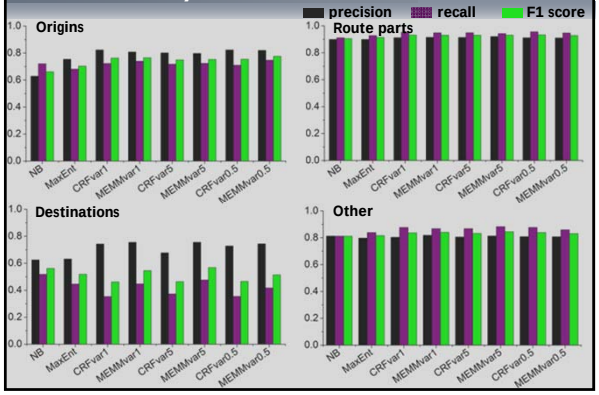
- Key distinctions important for interpreting route directions (for a human interpreter as well as for an algorithm):
  - Destination
    - Where the route ends, usually the location of the business or organization providing the direction info
    - e.g., "Directions to ***the Campus***"
  - Origin
    - Where the route starts (Can be to be subjective, inexact)
    - e.g., "***From New York***", "From I-80 East", "***From the North***"
  - Instructions
    - A set of sequential actions the traveler should take in order to get to the destination from the origin
    - e.g. "***turn left*** **onto** ***College Ave***.", "***Merge*** **onto** ***US-220 South***"

GeoCAM System Architecture

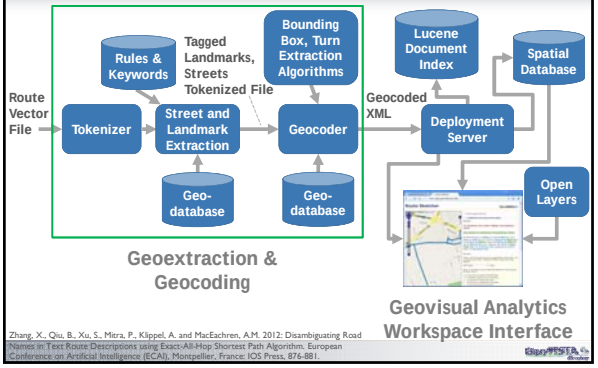


Zhang, X., Mitra, P., Klippel, A. and MacEachren, A.M. 2011: Identifying Destinations Automatically from Human Generated Route Directions. ACM-SIGSPATIAL Poster Paper, Chicago, IL: ACM, 373-376.

Experimental Results: precision, recall, F1 score for 12K manually labeled sentences in 100 documents



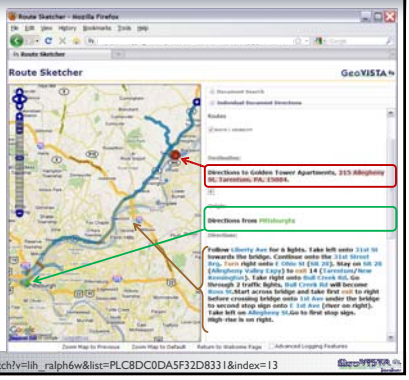
GeoCAM System Architecture



Zhang, X., Qiu, B., Xu, S., Mitra, P., Klippel, A. and MacEachren, A.M. 2012: Disambiguating Road Names in Text Route Descriptions using Exact-All-Hop Shortest Path Algorithm. European Conference on Artificial Intelligence (ECAI), Montpellier, France: IOS Press, 876-881.

RouteSketcher

- direction document processed and mapped



Jaiswal, A., Zhang, X., Mitra, P., Pezanowski, S., Turton, I., Xu, S., Klippel, A. and MacEachren, A.M. 2011: GeoCAM: A Geovisual Analytics Workspace to Contextualize and Interpret Statements about Movement. *Journal of Spatial Information Science 3, 65-101.*

video at: http://www.youtube.com/watch?v=lih_ralph6w&list=PLC8DC0DA5F32D8331&index=13

RouteSketcher



2.b: SensePlace2 – Supporting Situational Awareness w/ Public Social Media

- Traditional map-based situational awareness methods for assessing a situation based on reports from the field



Potential of social media for SA



Heading off disaster, one tweet at a time

By Jim Spellman, CNN
September 22, 2010 1:13 p.m. EDT | Filed under: Social Media

CNN iReport

Denver, Colorado (CNN) -- When word came of a huge explosion in San Bruno, California, FEMA chief Craig Fugate grabbed his cell phone and logged on to Twitter.

"I got out my little Android phone and went on Twitter and pulled up the grid to search for 'explosion' and got tweets coming out of the area," he said.

The Boulder, Colorado, wildfire triggered a strong response from iReport, often CNN's first source for disaster images.

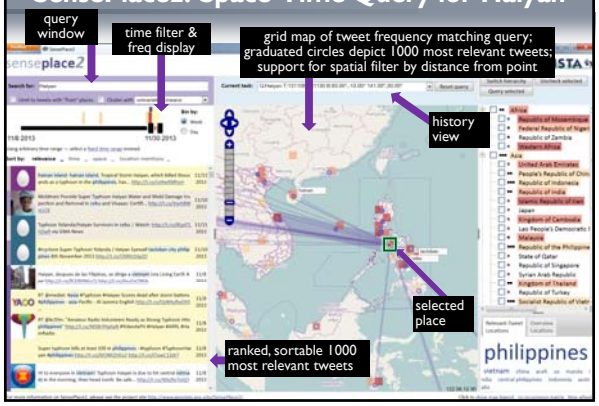
After a few minutes he says he determined the gas explosion and ensuing fire, though horrible, was a localized event and wouldn't spread to other communities.

STORY HIGHLIGHTS

- Government officials are usually wary of information on Twitter and other social media
- FEMA chief Craig Fugate uses the data for "situational awareness"

"I got better situational awareness [from Twitter] before we got official word," he said. "Four or five years ago I wouldn't have gotten that quality of information."

SensePlace2: Space-Time Query for Haiyan



SensePlace2: Computational process

- Collecting and processing Tweets:
  - query Tweets
  - analyze tweets for where they are from
    - GPS coordinates – accurate & precise but only 1-2%
    - "location" setting – often imprecise and inaccurate
  - extract named-entities from tweet text; e.g., locations, organizations, persons, hashtags, URLs, etc (what and where tweets are about)
  - index tweets by place, time, organizations, hashtags, etc.: Lucene index provides fast query and relevance ranks based on "boosting" (e.g., for tweets w/ places, hashtags, organizations, time, money, etc.)
  - store extracted entities: PostGres DB w/ GIS extension

MacEachren, A.M., Robinson, A.C., Jaiswal, A., Pezanowski, S., Savelyev, A., Blanford, J. and Mitra, P. 2011: Geo-Twitter Analytics: Applications in Crisis Management, *25th International Cartographic Conference, Paris, France.* MacEachren, A.M., Jaiswal, A., Robinson, A.C., Pezanowski, S., Savelyev, A., Mitra, P., Zhang, X. and Blanford, J. 2011: SensePlace2: GeoTwitter Analytics Support for Situational Awareness. In Miksch, S. and Ward, M., editors, *IEEE Conference on Visual Analytics Science and Technology, Providence, RI: IEEE, 181 - 190.*

GeoTxt: Building an open access, open source geoparser API and web interface

- The GeoText API (application programming interface) will support extraction, disambiguation, and geolocation of place entities in text submitted to the API from other applications or by users of a web interface
- Initial focus: microblog and other short text entities, with plans to extend to all text
- API being built upon existing open source tools and will be open source as well.
- API designed to take advantage of limited contextual information (e.g., multiple tweets on the same topic or tweets in sequence) to improve precision of geocoding

Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J.O., Hardisty, F., Pezanowski, S., Mitra, P. and MacEachren, A.M. 2013: GeoTxt: A Web API to Leverage Place References in Text. *7th ACM SIGSPATIAL Workshop on Geographic Information Retrieval. Orlando, FL: ACM.*

Sample Query and Map + GeoJSON Response

Query:

Transportation bill gets final OK, heads to Corbett. Prices at gas pumps will rise, November 21, 2013 11:14 PM

HARRISBURG -- From highway interchanges in **Erie**, to repairs of **Pittsburgh**'s Liberty Bridge, to additional funds for Port Authority buses and mass transit in **Philadelphia**, all corners of the state will share in the $2.3 billion transportation package approved this week by the Legislature, Gov. **Tom Corbett** said Thursday night.

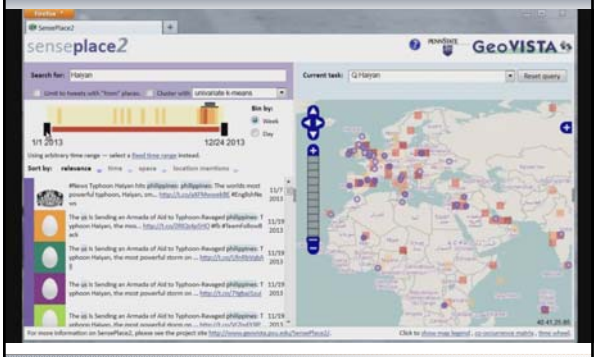


{"type":"Point","coordinates":[-79.99589, 40.44062]}},{"type":"Feature","properties":{"type":"location","name":"philadelphia","toponym":"Philadelphia","geoNameId":4560349,"locationType":null,"positions":[266]},"geometry":
{"type":"Feature","properties":{"type":"person","name":"tom corbett","personType":null,"gender":null,"kind":null,"positions":[403]},"geometry":null}]}

automated identification of places: some (seemingly) simple questions

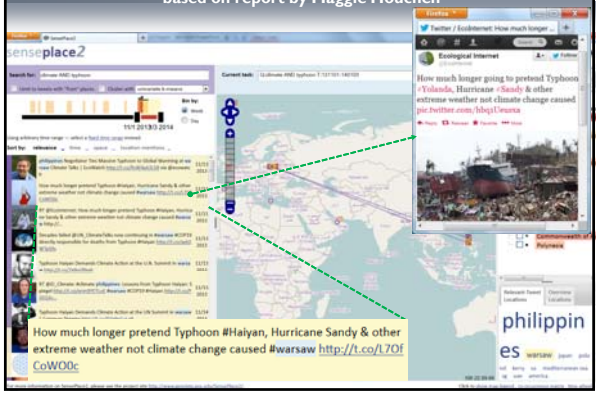
- what is a place name:
  - State College, PA
  - Happy Valley
  - Central PA
  - Starbucks
- what is a place reference: a prototypical category
  - ATH hopes to be in Eugene for the Oregon spring game
  - Apparently, Milan parked the bus and left the windows open!
  - I'm heading to Starbucks.
  - As AC Milan, we win and loose as together.
  - Riding around in the back of a pickup truck. Living the Texas dream.
- what is a "place"

SensePlace2: Place-Time Web Analytics



SensePlace2: Query for climate AND typhoon

based on report by Maggie Houchen

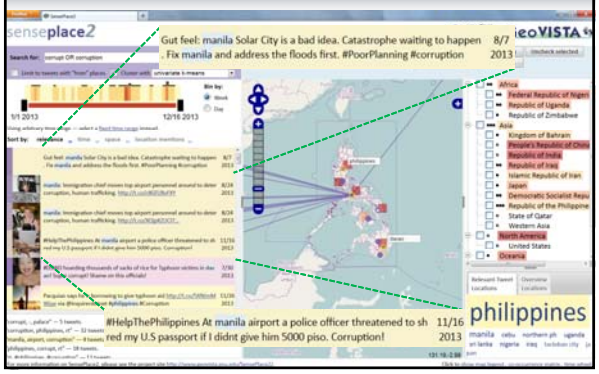


SensePlace2: Query for climate AND typhoon w/ clustering



SensePlace2: Query for corrupt OR corruption

based on report by Sterling Quinn



3. A few closing ideas

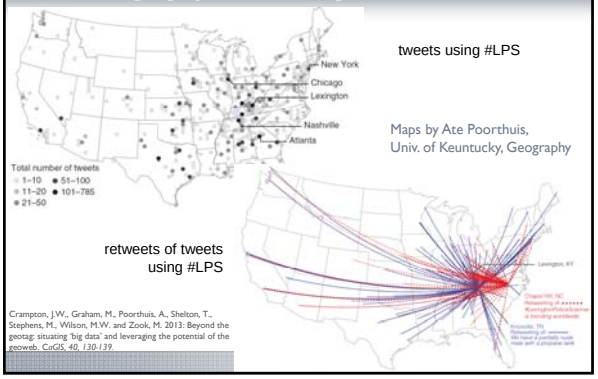
Crampton, et al: going 'beyond the geotag'

"potential impacts of 'big data' for the practice of critical human geography"

- going beyond social media that is explicitly geographic – where a tweet is posted is not everything that matters geographically
- going beyond spatialities of the 'here and now' – consider evolution over time
- going beyond the proximate; e.g., consider social networks
- going beyond the human to data produced by bots and automated systems
- going beyond the geoweb itself, by leveraging these sources against ancillary data, such as news reports and census data.

Crampton, J.W., Graham, M., Poorthuis, A., Shelton, T., Stephens, M., Wilson, M.W. and Zook, M. 2013: Beyond the geotag: situating 'big data' and leveraging the potential of the geoweb. *Cartography and Geographic Information Science 40, 130-139.*

analyzing events after University of Kentucky Wildcats men's basketball team victory in the NCAA championship game The geographies of #LexingtonPoliceScanner



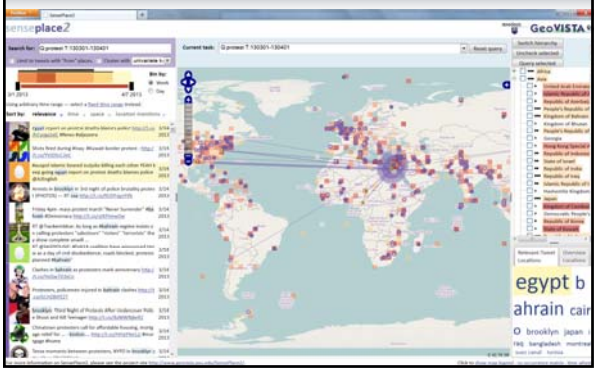
Movement Analysis – 3 days in the life of Tweeters:

Alexander (Sasha) Savelyev

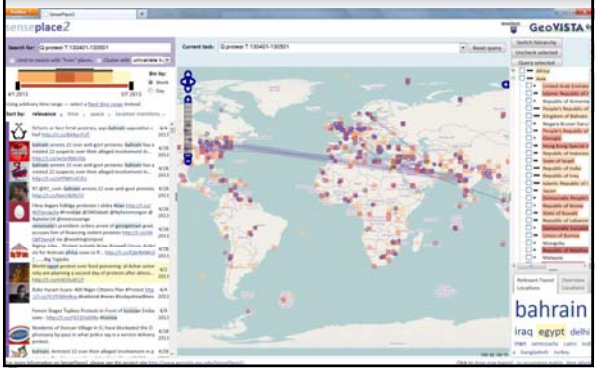
Exploratory visualization of movement patterns in Twitter data, presented at NACIS, 2013



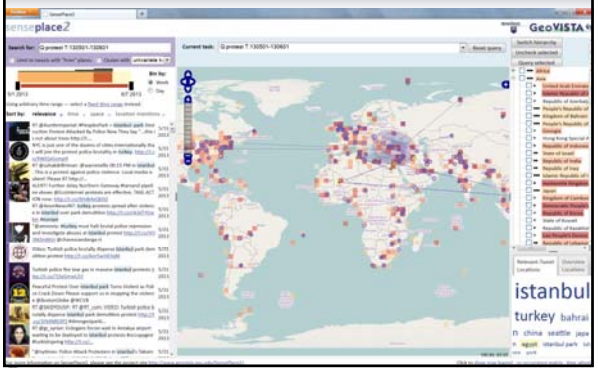
3 months of "protest" – March, 2013



3 months of "protest" – April, 2013



3 months of "protest" – May, 2013

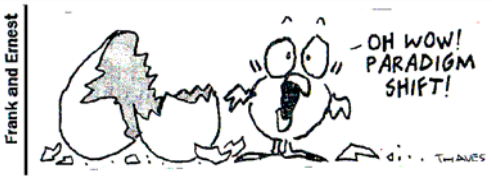


'Reasoning' about "what's next"

A few options for what might be next in geovisual analytics for unstructured data (in addition to the 4 components of reasoning support cited earlier):

- ***formalizing the concept of a "place reference"*** *– what counts a "geographic" information*
- ***identifying and understanding movement as reflected in unstructured data*** – 'moving' beyond processing directions
- ***integrating heterogeneous forms of information*** *(e.g., "authoritative" with "non-authoritative" – imagery and social media)*
- ***analyzing and understanding geosocial networks*** *– relating* geographic space & place + social space
- ***understanding dynamic behavior/process*** – (e.g., applying complexity science methods)
- **addressing issues of geo/social media data privacy**

That's all: What's Next?



Thanks

maceachren@psu.edu
www.geovista.psu.edu

Acknowledgements: Parts of the work represented here were supported by funds from: National Science Foundation: EIA-0306845 & 9983451 & DGE-1144860 (IGERT-Big Data Social Science); Gates Foundation funding of the Vaccine Modeling Initiative (https://www.vaccinemodeling.org/); National Visualization and Analytics Center, a U.S. Dept. of Homeland Security program operated by the Pacific Northwest National Laboratory (PNNL), and the Dept of Homeland Security Visual Analytics for Command, Control, and Interoperability Environments Center of Excellence.

Moving forward: Reasoning about movement and other things dynamic

Alan M. MacEachren

GeoVISTA Center
Department of Geography
Penn State University
maceachren@psu.edu



"moving forward"

- progressing spatially in a forward direction, defined from an ego-centric viewpoint related to the human body and eyes, a vehicle and its "front", etc.
- progressing conceptually *toward* a goal (e.g., developing methods to support *Knowledge Discovery from Moving Objects; or … economic recovery*)
- progressing conceptually *into the future from some waypoint*, thus moving beyond an initial goal or other change, whether good or bad – thus, deciding "what next?"





Europe Moving Forward (cartoon by Chappatte), International Herald Tribune

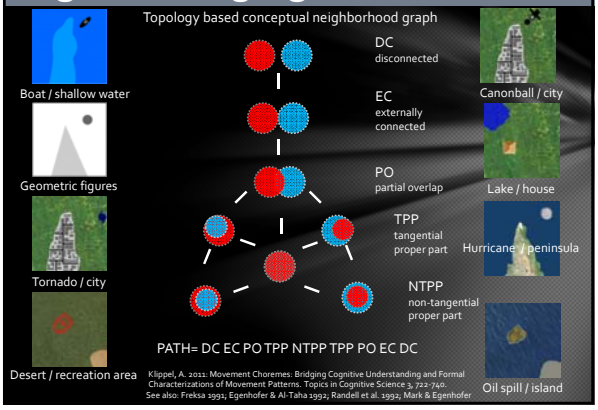
"reasoning" about movement and other things dynamic

- interpretation of analytical results – explaining patterns of movement, change, behaviors, etc
- understanding how humans reason (cognitively and linguistically) about movement and behavior
- applying automated reasoning to retrieve or analyze movement info
- enabling human spatial/geographic reasoning to derive knowledge about: movement, change, behavior and things dynamic
- speculating about progress and future objectives

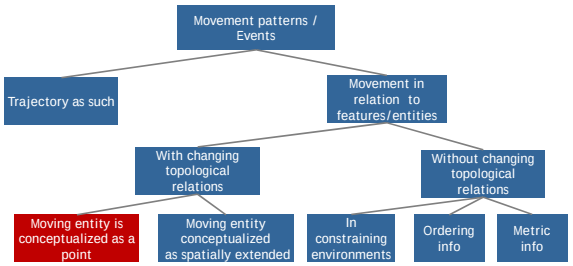
Outline

- **cognition, language, and movement**: how movement is conceptualized, articulated, and recognized in/extracted from articulations
- **visual analytics and reasoning**: enabling (and automating) reasoning about movment, change, behavior, and things dynamic
- **reasoning about "moving forward scientifically"**

Cognition, language, and movement



Formal Basis of Conceptual Primitives of Movements



Klippel, A. (2011). Movement choremes: Bridging cognitive understanding and formal characterization of movement patterns. Topics in Cognitive Science, 3 (4), 722–740

Topological DLine-Region Relations

26 line-region relations as identified by the 9+ - Intersection model

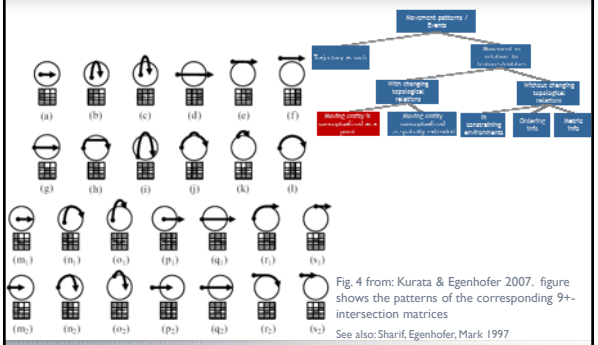


Fig. 4 from: Kurata & Egenhofer 2007. figure shows the patterns of the corresponding 9+-intersection matrices

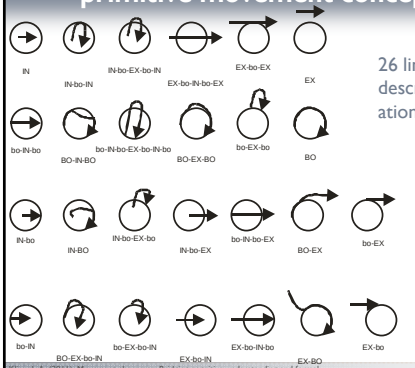
See also: Sharif, Egenhofer, Mark 1997

Modeling the same relations with primitive movement concepts (mc)

- Only 4 movement concepts are necessary
  - IN for movement that occurs within an area
  - EX for movement that occurs outside an area
  - BO and bo for movement that occurs on the boundary either extended (BO) or restricted (bo)
- The important thing to keep in mind is that complex paths/routes are build up from simple movement concepts

Klippel, A. (2011). Movement choremes: Bridging cognitive understanding and formal characterization of movement patterns. Topics in Cognitive Science, 3 (4), 722–740

Modeling the same relations with primitive movement concepts (mc)



26 line-region relations described by combin-ations of 4 primitives.

… can form basis of methods to link movements to their geographic context and to support "reasoning" (both human and computational) about processes underling movements.

Klippel, A. (2011). Movement choremes: Bridging cognitive understanding and formal characterization of movement patterns. Topics in Cognitive Science, 3 (4), 722–740

Geographic Conceptualization for Accounts of Movement (GeoCAM)

- GeoCAM goal: *develop, implement, and assess an integrated approach to identifying, extracting, interpreting linguistic accounts of movement—in context*
- A key to the solution, we believe, is to *not* treat the problem as a natural language processing (NLP) problem *in isolation*.
- Integrated approach – *build on spatial-cognitive theory* – use geo-databases and related info iteratively, together with NLP methods, and visual analytics tools to interpret geo-references in ***place, time, and thematic context***.

Jaiswal, A., Zhang, X., Mitra, P., Pezanowski, S., Turton, I., Xu, S., Klippel, A. and MacEachren, A.M. 2011: GeoCAM: A Geovisual Analytics Workspace to Contextualize and Interpret Statements about Movement. Journal of Spatial Information Science 3, 65-101.

Motivation: Historical accounts of travel "South Africa, by Henry Balfour, 1910"

- Sat. 25 June - Left Oxford by 9.10 train + came on by 10.25 from Reading (via Newbury + Winchester). Secured cabin 18 on board "Kildonan Castle", met Lady Peck + her daughter who were seeing Peck + Benson off. Sailed soon after 4.30 p.m. Wind moderate, ship pitching somewhat in the evening. Before leaving S. Hampton, wrote to E. and to Lewis, Jo, Mother + K.B. + wired to Jo + E.
- Sun. 26 - Rounded Ushant before 8.30 a.m. Moderate sea, slight motion which increased in the evening. Ships very numerous.
- Mon. 27 - Dull, drizzling morning, clearing later. Passed Finisterre before breakfast, but did not sight the cape. Sea slight.
- Tues. 28 - Warm, sea slight, drifting sea mists + sun. Had long talk with Barry, the champion sculler + his trainer Phelps. Wrote to E. and to Jo-+-Lewis for posting in Funchal. Passed homeward mail ("Briton") at about 9.0 pm.
- Wed. 29 - Arrived Funchal at about 6 a.m. Cloudy day but fine. Ship coaling. Went ashore at 7.0 + walked out to Reid's Palace Hotel where I breakfasted. Very nice hotel in splendid situation. Walked about Funchal + wired to Hoernle at Cape Town about lectures. Had received letter from him + later one from E. Came on board at 10.30. Sailed from Funchal at noon. Ship rolled a good deal in afternoon + evening.

http://www.pmt.ox.ac.uk/manuscripts/balfourdiaries1910.html

framing components of initial approach

continuum of accounts of movement in text

wide range of other documents | situation report | travel journal/blog | travel narrative | **written directions** | navigation service



a typical example

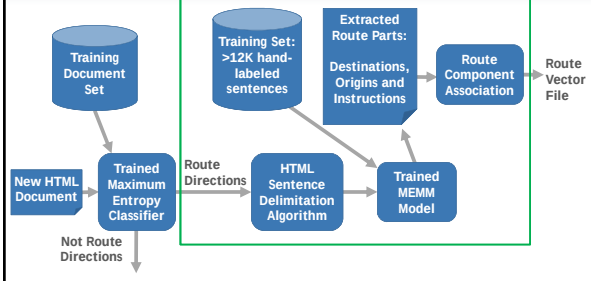
focus on constrained problem – *human-generated route directions*;

constrained:
- *linguistically* by focus on travel directions
- *geographically* by focus on directions within a network

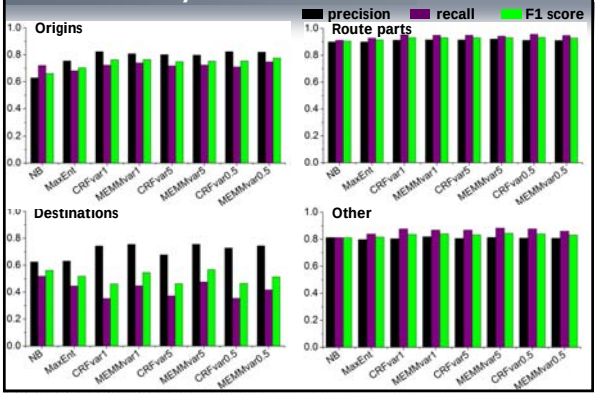
Meaningful Route Parts

- Key distinctions important for interpreting route directions (for a human interpreter as well as for an algorithm):
  - Destination
    - Where the route ends, usually the location of the business or organization providing the direction info
    - e.g., "Directions to ***the Campus***"
  - Origin
    - Where the route starts (Can be to be subjective, inexact)
    - e.g., "***From New York***", "From I-80 East", "***From the North***"
  - Instructions
    - A set of sequential actions the traveler should take in order to get to the destination from the origin
    - e.g. "***turn left*** **onto** ***College Ave***.", "***Merge*** **onto** ***US-220 South***"

GeoCAM System Architecture



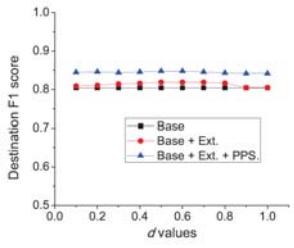
Experimental Results: precision, recall, F1 score for 12K manually labeled sentences in 100 documents



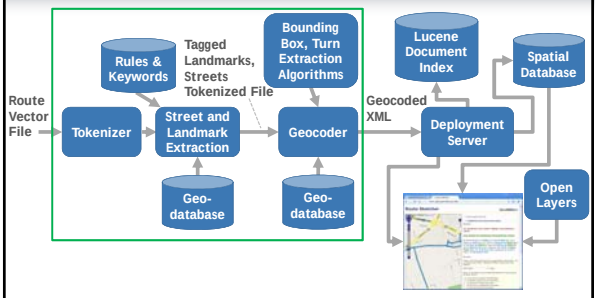
Experimental Results

Improving destination ID

- Find candidate sentences:
  - Examine "arrival information" sentences
  - Match language patterns
  - Tag nouns and proper nouns
  - Check type nouns (from pre-defined list)
  - Generate a set of candidates
- Combine the candidates and predicted destination sentences.
- Match type nouns and proper names in predicted destinations and "Other" sentences
- If matched, change labels from "Other" to "Dest."

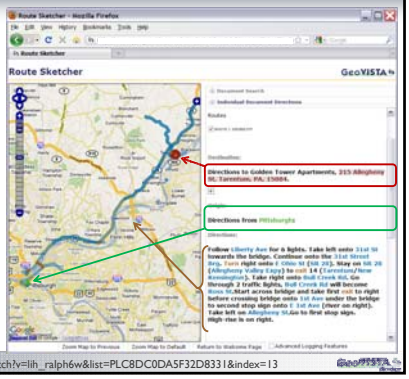


GeoCAM System Architecture



RouteSketcher

- direction document processed and mapped

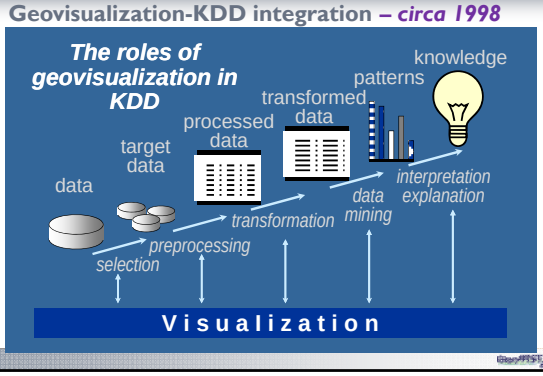


RouteSketcher



2) Visual analytics support for reasoning about movement and geographic behavior

Geovisualization-KDD integration – *circa 1998*



Geovisual Analytics

- *Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces (2004)*

  … thus, the focus is on a process of assembling data and evidence, deciding analytical parameters, generating hypotheses, inferences and explanations from evidence, and comparing / assessing those hypotheses, inferences and explanations

  http://nvac.pnl.gov/agenda.stm & http://www.vismaster.eu/book/

Geovisual Analytics

- *Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces (2004)*

  … thus, the focus is on a process of assembling data and evidence, deciding analytical parameters, generating hypotheses, inferences and explanations from evidence, and comparing / assessing those hypotheses, inferences and explanations

  http://nvac.pnl.gov/agenda.stm & http://www.vismaster.eu/book/

- **Geo**visual analytics focuses on visual interfaces to visual-analytical methods that support a process of *reasoning* with and about information having geographic components

Geovisual analytics: 3 (of many) challenges

- *cope with "big data" (large volume and/or complex)*
- *leverage heterogeneous geoinformation*
- *support the analytical reasoning processes*

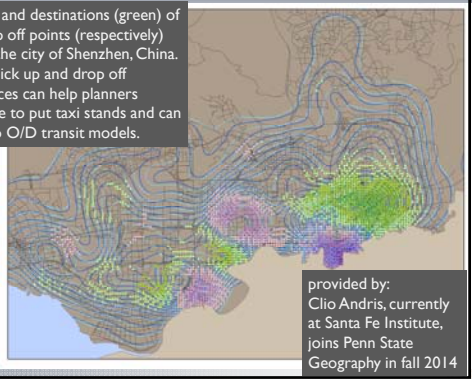
Scaling analysts to work w/ big data



OD Map of Taxi behavior in Shenzhen, China

daily operations of 3000 cabdrivers in over 48 million of trips

Origins (purple) and destinations (green) of pick up and drop off points (respectively) for Taxi trips in the city of Shenzhen, China. Understanding pick up and drop off attraction of places can help planners determine where to put taxi stands and can serve as input to O/D transit models.

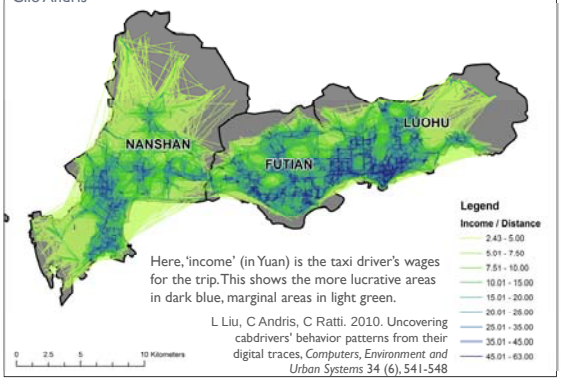


L Liu, C Andris, C Ratti. 2010. Uncovering cabdrivers' behavior patterns from their digital traces, *Computers, Environment and Urban Systems* 34 (6), 541-548

provided by:
Clio Andris, currently at Santa Fe Institute, joins Penn State Geography in fall 2014

provided by:
Clio Andris

Taxi incomes in Shenzhen

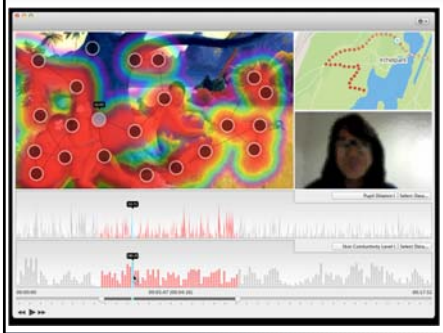


Here, 'income' (in Yuan) is the taxi driver's wages for the trip. This shows the more lucrative areas in dark blue, marginal areas in light green.

L Liu, C Andris, C Ratti. 2010. Uncovering cabdrivers' behavior patterns from their digital traces, *Computers, Environment and Urban Systems* 34 (6), 541-548

Heterogeneous Info: GeoMotion Explorer

Josh Stevens, IGERT Big Data Social Science

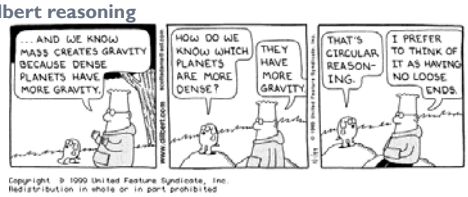


will combine data and multimedia from several sources to enable synchronous, real-time exploration and analysis within and across each data source (video, eye movement, pupil dilation, biometric data)

Summer '13 visitor with Sara Fabrikant at University of Zurich

Support for reasoning: about movement, geographic behavior, and things dynamic

Dilbert reasoning

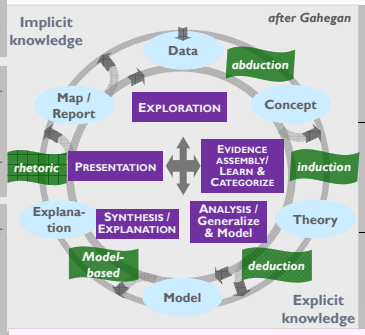


Framework for visually-enabled science reasoning w/geo-data → information → knowledge

abduction: *a method of reasoning from facts to the most likely explanation*

induction: *a method of reasoning by which one infers a generalization from a series of instances*

deduction: *a method of reasoning by which something must be true because it is a special case of a general principle known to be true*



ACTIVITY
geoscientific activities

Represent-ation
represented objects

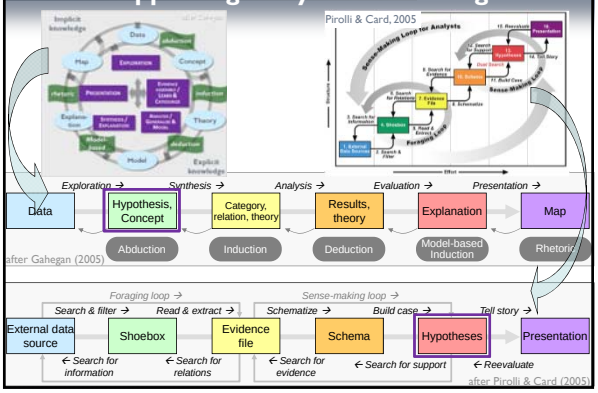
Reasoning
primary modes of inference

model-based reasoning: www.msu.edu/~tne/science/sciencelink_11.htm

Reasoning

- **deduction**: Birds can fly up to 600 miles per day during migration; GPS coordinates show 1900 miles of travel for a bird in a single day; ∴ there Bill deduces there was a GPS error.
- **induction**: Bill counts birds at a hawk watch every day for a month; during the period he counts 252 eagles pass by and all are immature birds (1st year); ∴ Bill infers that only immature birds move along this migration route.
- **abduction**: Female ruby-throated hummingbirds are rather non-descript and are common in Pennsylvania in the summer and early fall and other species are very rare; Bill sees a non-descript hummingbird in Pennsylvania in July;; ∴ Bill's interpretation is that he just saw a female ruby-throated hummingbird

Geovisual analytics: supporting analytical reasoning



Understanding the diversity of activities that rely on spatial skills and concepts:

The process of geological reasoning in the field

- At first, her (11) route takes her past (12) flat-lying sedimentary rocks that she (13) determines are sandstone. As she (14) continues toward the (15) mesa, she notices that the (16) orientation of the (17) sedimentary layers (18) begins to steepen until they are (19) nearly vertical. She makes (20) measurements of (21) these changing bedding orientations (22) using a compass and (23) records their locations on the map. After (24) walking past the (25) vertical sedimentary layers, she finds herself in an area with a dark, (26) massive, (27) crystalline rock with white elongate minerals. She (28) deviates from her planned traverse to walk (29) along the contact (30) between the (31) igneous and sedimentary rocks to (32) record its location on her map.

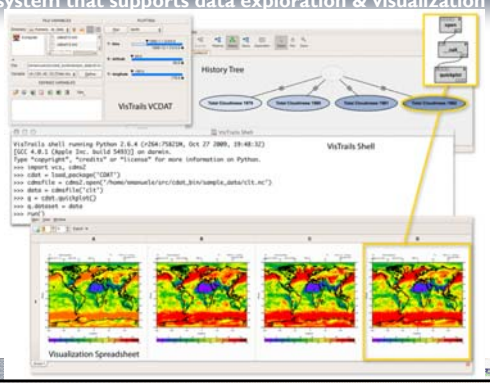
Liben, L.S. and Titus, S.J. 2012: The importance of spatial thinking for geoscience education: Insights from the crossroads of geoscience and cognitive science. *Geological Society of America Special Papers 486, 51-70.*

Support for reasoning in visual analytics: Ideas on what is needed

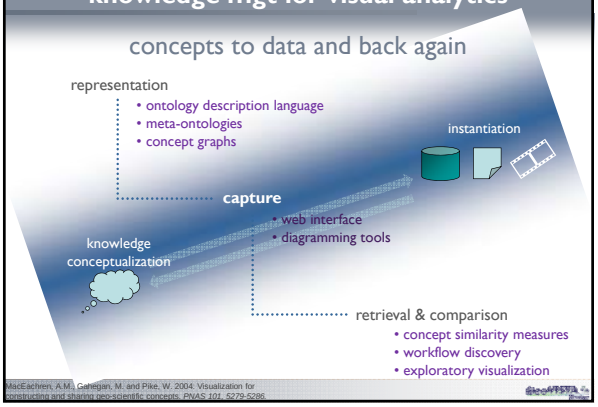
- understanding of how distributed/external cognition uses the environment or manipulable external artifacts (e.g., Liu, Z., Nersessian, N. and Stasko, J. 2008: Distributed cognition as a theoretical framework for information visualization. *IEEE TV&CG 14, 1173-1180.*)
- methods to semantically enrich entities (e.g., trajectories) extracted from data analysis (e.g., for moving objects and other dynamic phenomena) (e.g., Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V., Damiani, M.L., Gkoulalas-Divanis, A., Macedo, J. and Pelekis, N. 2013: Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR) 45, 42.)*
- methods to capture, save, and retrieve analysis process and insights (e.g., "pattern basket" Chen, J. and MacEachren, A.M. 2008: Resolution Control for Balancing Overview and Detail in Multivariate Spatial Analysis. *Cart. J. 45, 261-273.*)
- knowledge management methods to link analytical findings with other knowledge (e.g., Pike, W. and Gahegan, M. 2007: Beyond ontologies: Toward situated representations of scientific knowledge. *International Journal of Human-Computer Studies 65, 674-688;* Dadzie, A., Lanfranchi, V. and Petrelli, D. 2009: Seeing is believing: Linking data with knowledge. *Information Visualization 8, 197.)*

VisTrails: http://www.vistrails.org

An open-source scientific workflow & provenance management system that supports data exploration & visualization



knowledge mgt for visual analytics



MacEachren, A.M., Gahegan, M. and Pike, W. 2004: Visualization for constructing and sharing geo-scientific concepts. *PNAS 101, 5279-5286.*

3) Reasoning about "what's next"

A few options for what might be next (in addition to the 4 components of reasoning support):

- ***influencing*** movement / behavior
- ***understanding*** dynamic processes (including, but not limited to, movement)
- ***complexity & network science*** and dynamic geographic behavior

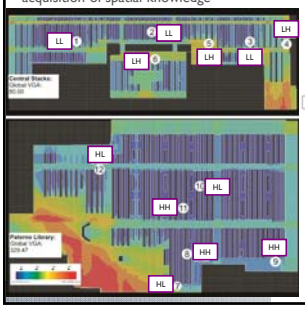
Influencing movement

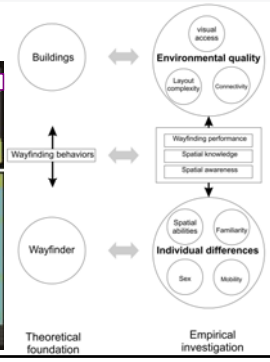


http://nostalgia049.wordpress.com/tag/funny-signs/

Influencing movement: A holistic (spatial cognition-based) framework of indoor wayfinding behaviors – Rui Li

Environmental quality matters for wayfinding performance, familiarity matters for acquisition of spatial knowledge





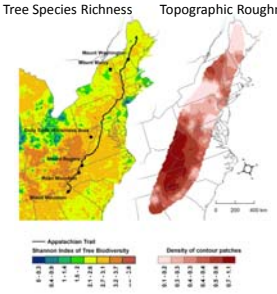
Influencing movement: Integrated Crowd Management



Van de Weghe, N., Bellens, R., De Jaeger, T., Gautama, S., Huybrechts, R., Meier, B. and Versichele, M. 2013: Towards an Integrated Crowd Management Platform. *Intelligent Systems for Crisis Management*. Springer, 301-308.

LEAPS Landscape Ecology at Penn State — understanding dynamic processes

Tree Species Richness | Topographic Roughness



Tree biodiversity is correlated with topographic complexity along the Appalachian Trail, a north-south gradient in the Eastern U.S.

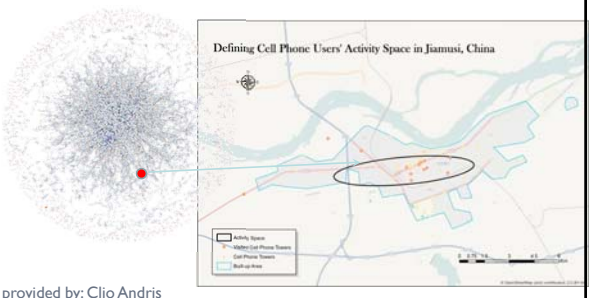
--Appalachian and Chinese forests are biologically similar and distinctive, having the world's richest temperate broadleaf forests

--implications for species migration under climate change; network analysis using Circuitscape. Circuitscape is a free, open-source program which borrows algorithms from electronic circuit theory to predict patterns of movement, gene flow, and genetic differentiation among plant and animal populations in heterogeneous landscapes. http://www.circuitscape.org/

Elizabeth Crisfield & Erica Smithwick

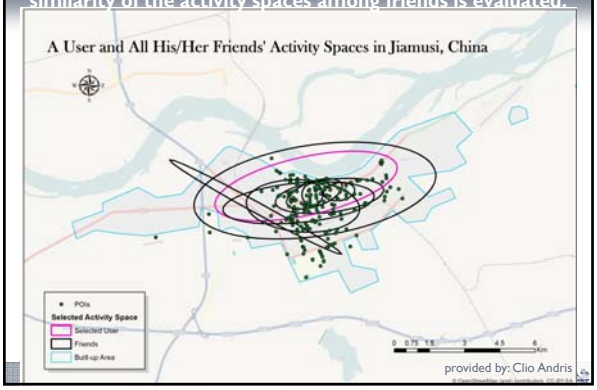
Complexity / network science

Network of cell phone users. Each node is given an activity space: an ellipse drawn around his or her most frequently used cell towers. (96 Cell towers total).



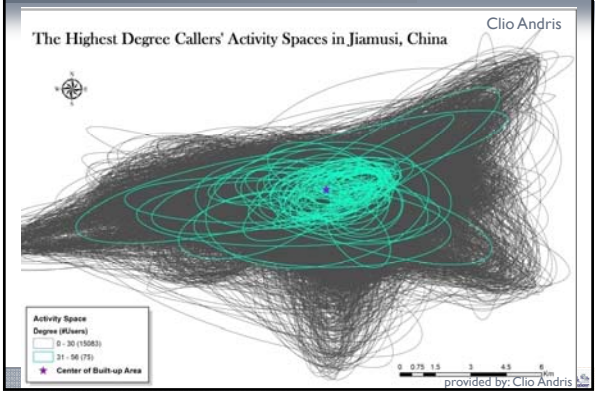
provided by: Clio Andris

One user is chosen out of the social network, and his or her friends are as well. Their activity spaces are mapped and the similarity of the activity spaces among friends is evaluated.

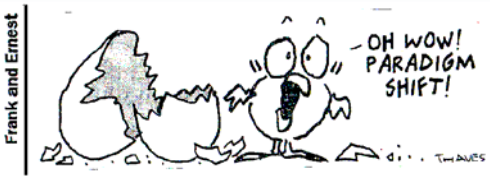


provided by: Clio Andris

The top 1% of callers with the most connections (highest degree) use the city center more often than the average users.



That's all: What's Next?



Thanks

maceachren@psu.edu
www.geovista.psu.edu

Acknowledgements: Parts of the work represented here were supported by funds from: National Science Foundation: EIA-0306845 & 9983451 & DGE-1144860 (IGERT-Big Data Social Science); Gates Foundation funding of the Vaccine Modeling Initiative (https://www.vaccinemodeling.org/); National Visualization and Analytics Center, a U.S. Dept. of Homeland Security program operated by the Pacific Northwest National Laboratory (PNNL), and the Dept of Homeland Security Visual Analytics for Command, Control, and Interoperability Environments Center of Excellence.

Questions Facing Map Design in the Age of Mobility and Siri

Ryan S Mullins
GeoVISTA Center
Department of Geography
The Pennsylvania State University

DEPARTMENT OF GEOGRAPHY
COLLEGE OF EARTH AND MINERAL SCIENCES

PENNSTATE 1855

GeoVISTA Center

Driving Questions

- What is a mobile user?
- What is a mobile device?
- What roles do maps play on mobile devices?
- What kind of decisions are made using mobile maps?
- How is uncertainty understood in the decision making process?
- How can uncertainty be mitigated?
- How do we know if mitigation strategies work?

What is a Mobile User?

- Simple definition: anyone using a mobile device
- Demographics
  - Any age
  - Any ethnicity
  - Any education level
  - Any income level

Cell Phone and Smartphone Ownership

% of American adults within each group who own a cell phone and the % who own a smartphone

		Cell Phone	Smartphone
All adults (n=2,252)		**91%**	**56%**
a	Men (n=1,029)	93[b]	59[b]
b	Women (n=1,223)	88	53
Race/ethnicity			
a	White, Non-Hispanic (n=1,571)	90	53
b	Black, Non-Hispanic (n=252)	93	64[a]
c	Hispanic (n=249)	88	60
Age			
a	18-29 (n=404)	97[cd]	80[bcd]
b	30-49 (n=577)	95[cd]	67[cd]
c	50-64 (n=641)	89[d]	45[d]
d	65+ (n=570)	76	18
Education attainment			
a	No high school diploma (n=168)	83	36
b	High school grad (n=630)	88	46[a]
c	Some College (n=588)	92[ab]	60[ab]
d	College + (n=834)	95[abc]	70[abc]
Household income			
a	Less than $30,000/yr (n=580)	86	43
b	$30,000-$49,999 (n=374)	90[a]	52[a]
c	$50,000-$74,999 (n=298)	96[ab]	61[a]
d	$75,000+ (n=582)	98[ab]	78[abc]
Urbanity			
a	Urban (n=763)	92[c]	59[c]
b	Suburban (n=1,037)	91[c]	59[c]
c	Rural (n=450)	85	40

Source: Pew Internet Spring Tracking Survey, April 17 – May 19, 2013. N=2,252 adults ages 18+. Interviews were conducted in English and Spanish and on landline and cell phones. Margin of error is +/- 2.3 percentage points for results based on all adults.

Note: Columns marked with a superscript letter ([a]) or another letter indicate a statistically significant difference between that row and the row designated by that superscript letter. Statistical significance is determined inside the specific section covering each demographic trait.

Data reported Feb 2012.
http://pewinternet.org/Commentary/2012/February/Pew-Internet-Mobile.aspx

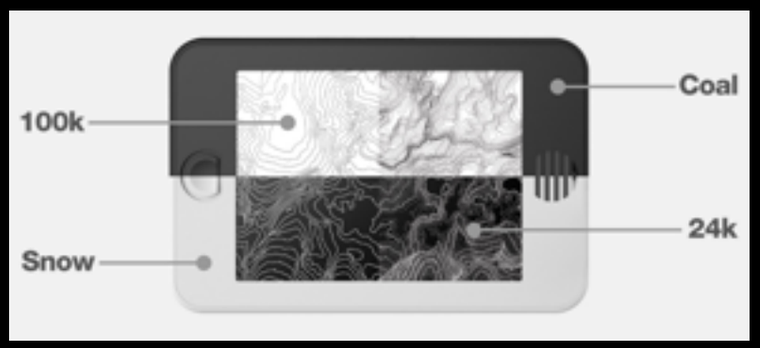
What is a Mobile Device?

- Computing system that...
  - Easily transported
  - Held in the hand
  - Physically small
  - Access to various data networks
  - Power to last all day
  - Multi-modal interaction

What is a Mobile Device?

- Smart phones
- Tablets
- eReaders





Role of Mobile Maps

- Way finding and navigation
- Destination and route planning
- Location sharing via social media (i.e. FourSquare, Facebook, Yelp)
- Taking and viewing pictures or other geo-referenced data

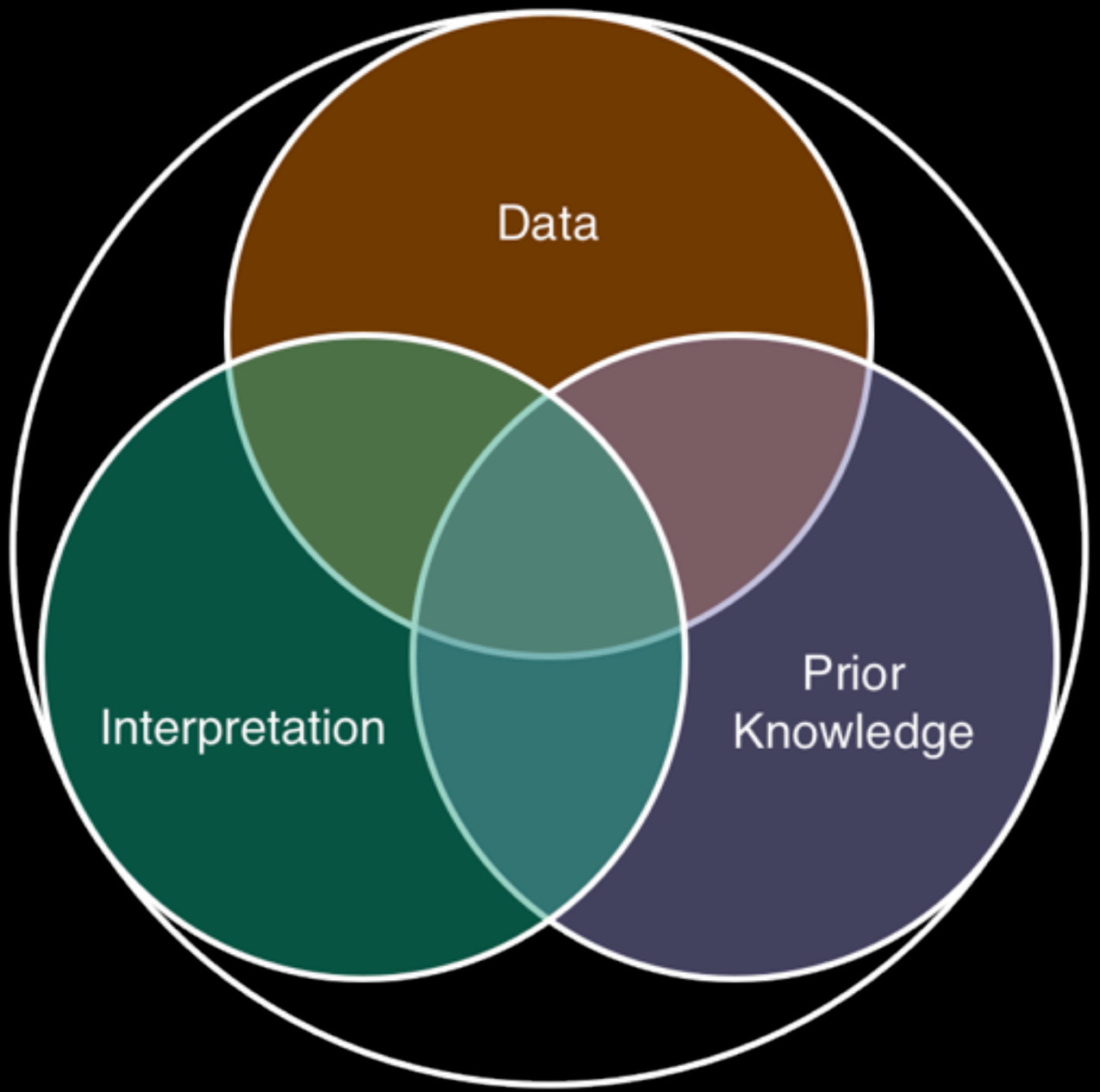
- All of these tasks can be related to...

Decisions

- Maps are used to support decisions about...
  - Where to visit/eat/stay
  - Which way to go
  - Where to tell others to go, meet, etc.
  - Which mode of travel to use

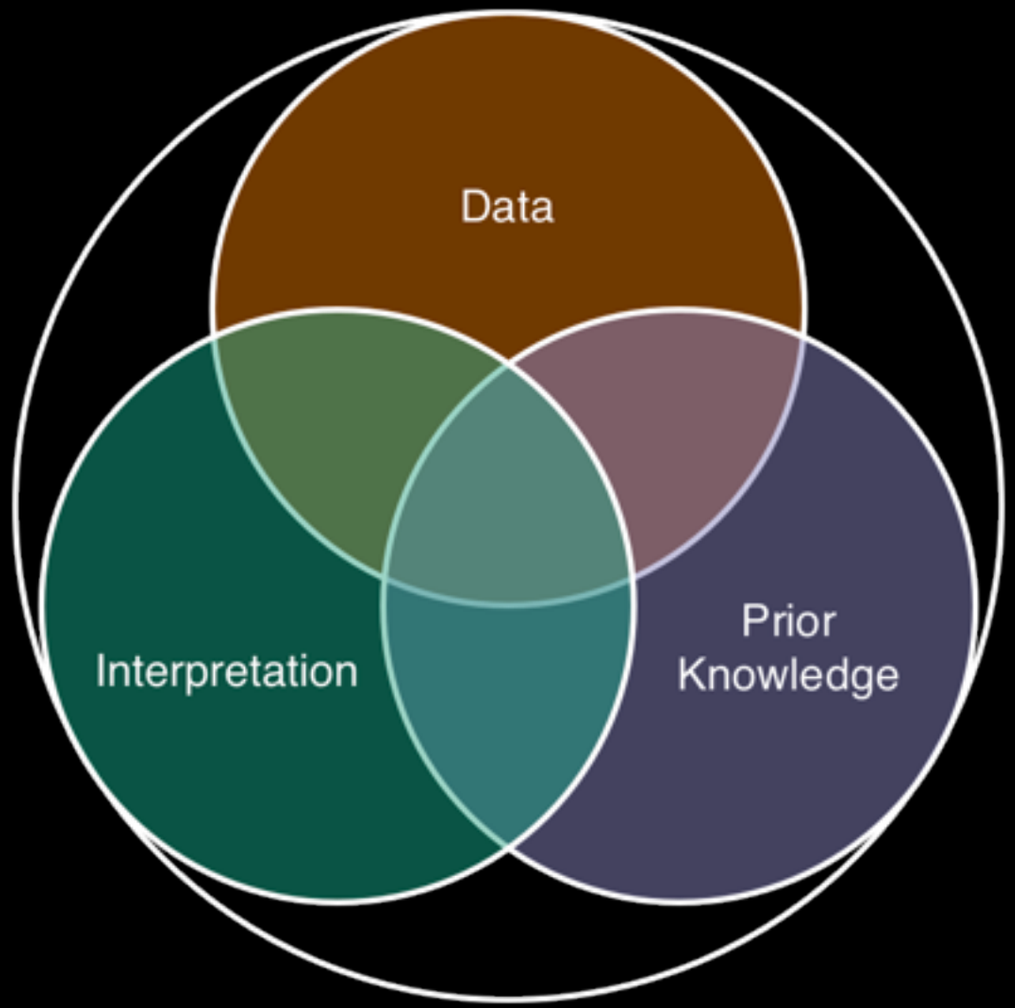
- All decisions have some...

Uncertainty in Decisions



Uncertainty in Decisions

- Data
  - Error, accuracy, precision, etc.[1,2]
- Interpretation of visuals
  - Perceived meaning[3], ambiguity/vagueness, context
- Prior Knowledge
  - Career[4], initial/prior solutions[4], subjective valuation of information[5,6]



[1] MacEachren, A. M. (1992) Visualizing uncertain information. *Cartographic Perspectives*. 13. 10-19.
[2] MacEachren, A.M. (1994) SOME truth with maps: A primer on symbolization and design. *Association of American Geographers*, Washington, D. C.
[3] MacEachren, A.M. (1995) *How Maps Work: Representation, Visualization, and Design*. Guilford Press.
[4] Kobus, D., S. Proctor, and Holste. S. (2001) Effects of experience and uncertainty during dynamic decision making. International Journal of Industrial Ergonomics 28. 275-290.
[5] Reece, W., and Matthews, L. (1993) Evidence and uncertainty in subjective prediction: Influences on optimistic judgment. Psychological Reports 2(72). 435-439.
[6] Cohen, B. L., and Wallsten, T. S. (1992) The effect of constant outcome value on judgments and decision making given linguistic probabilities. Journal of Behavioral Decision Making 1(5). 53

Mitigating Uncertainty

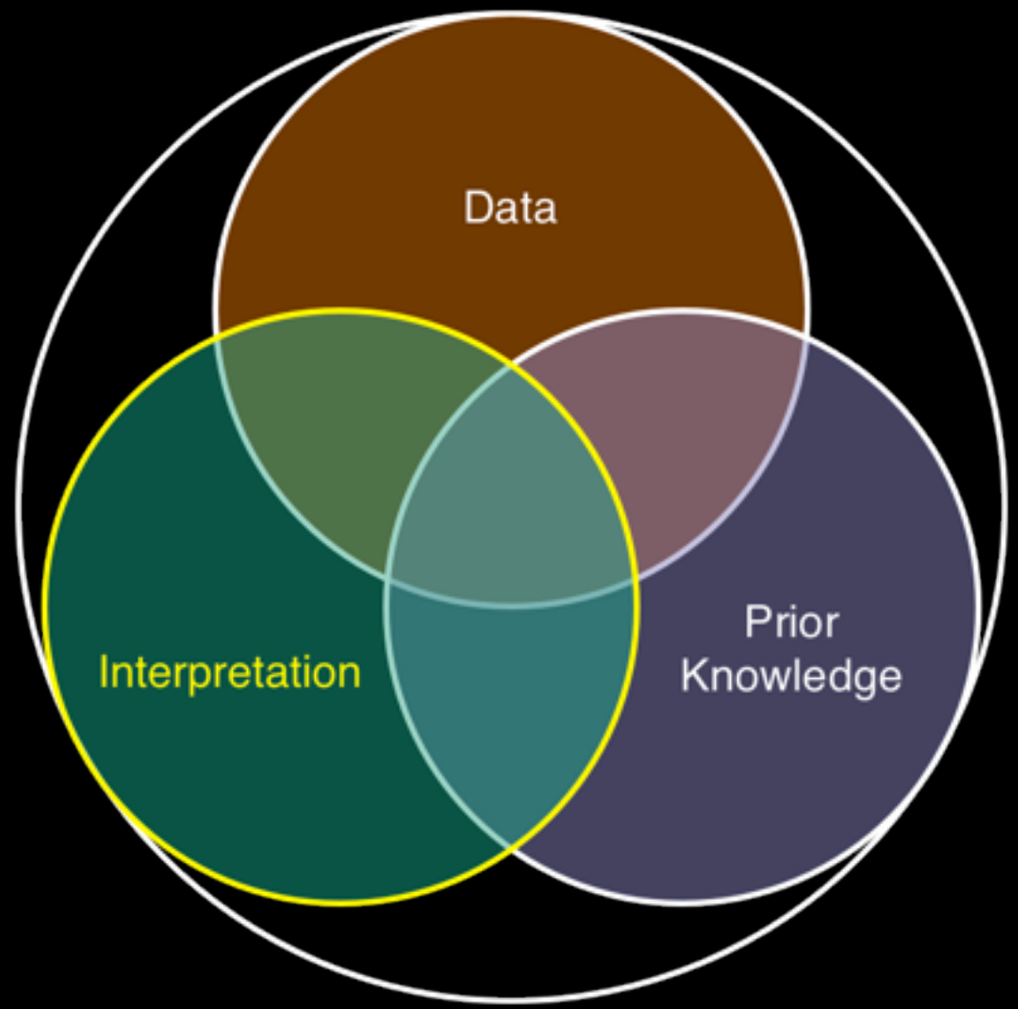
- Common, data driven, strategies[1]
  - Presenting good aspects of certainty (accuracy/precision focus)
  - Presenting bad aspects of uncertainty (error focus)
- Design symbology to convey uncertainty[2,3]
- Design interaction to provide access to uncertainty[4,5]

- More work can be done here...

[1] Buttenfield, B. P. (1993) Representing data quality. *Cartographica* (special content, Mapping Data Quality). 30(2&3). 1-7.
[2] Pang, A. (2001) Visualizing uncertainty in geo-spatial data. In: *Proceedings of the Workshop on the Intersections between Geospatial Information and Information Technology*. National Academies Committee of the Computer Science and Telecommunications Board, Washington, D.C.
[3] Wittenbrink, C.M., Saxon, E.., Furman, J.J., Pang, A., and Lodha, S. (1995) Glyphs for visualizing uncertainty in environmental vector fields. *IEEE Translations on Visualization and Computer Graphics*. 2(3). 266-79.
[4] Fisher, P. (1993) Visualizing uncertainty in soil maps by animation. *Cartographica.* 30(2+3). 20-27.
[5] Howard, D., and MacEachren., A.M.(1996) Interface design for geographic visualization: Tools for representing reliability. *Cartography and Geographic Information Systems*. 23(2). 59-77.

Limiting Scope

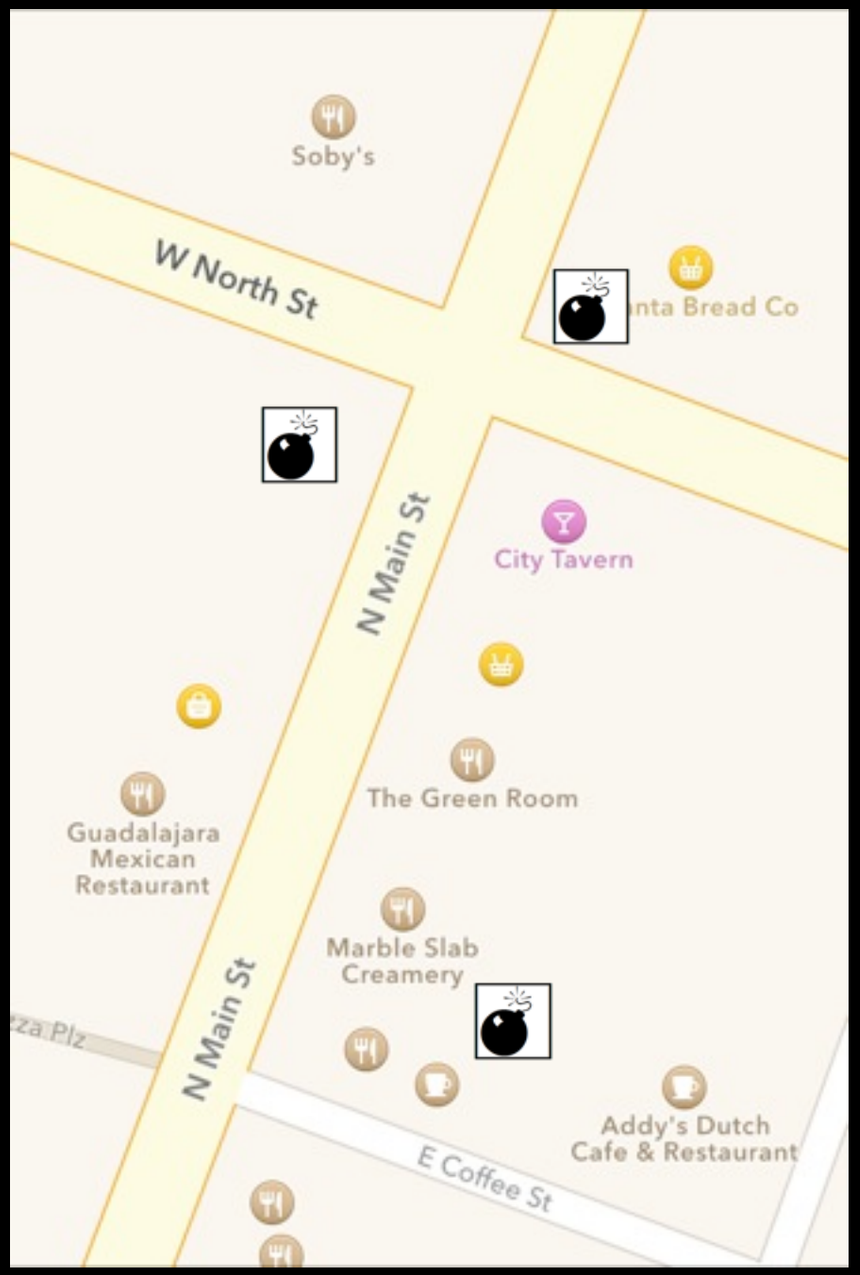
- Focus on Interpretation
  - Specifically on ambiguity in visual depictions
- Restricting to point symbols
  - Many decisions relate to point features
- Three entry vectors
  - Type of phenomenon
  - Attributes
  - Location

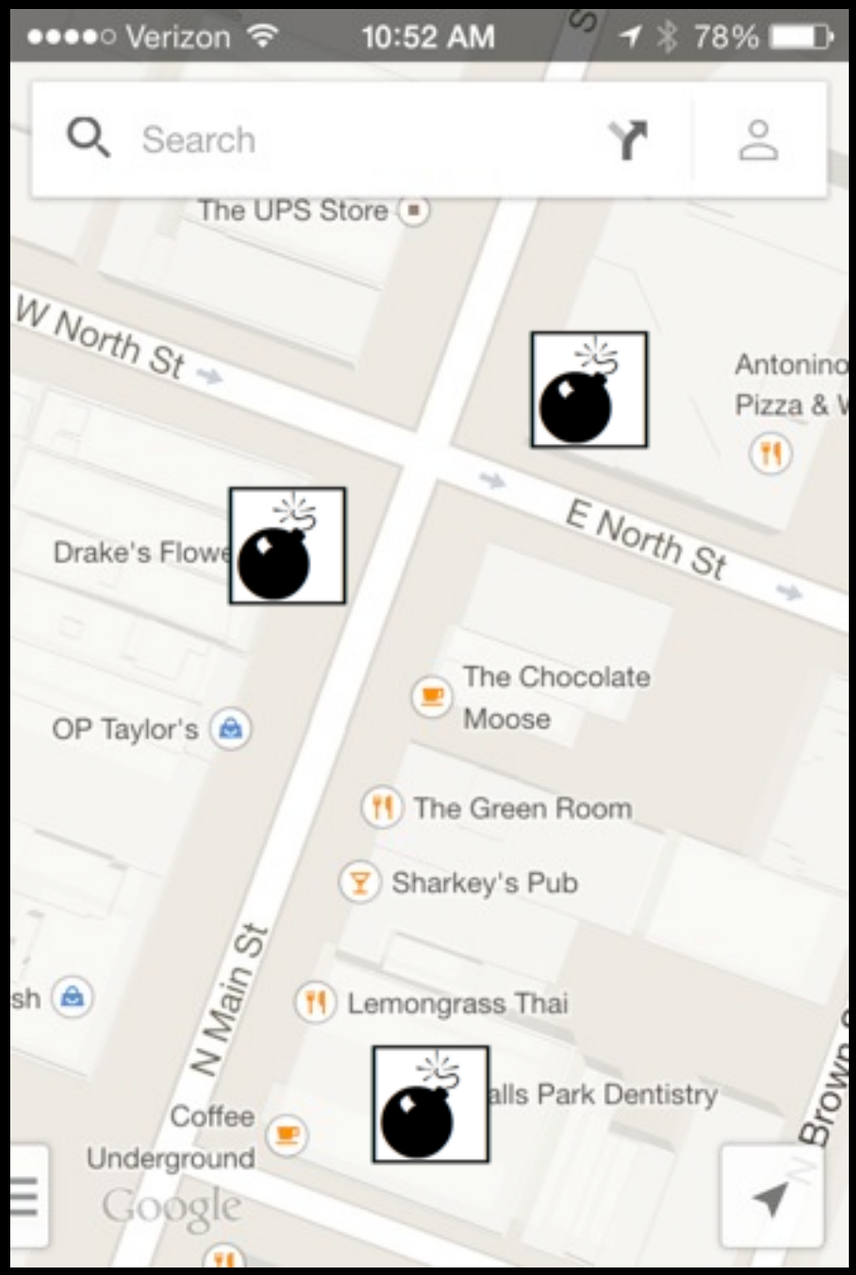


Ambiguity – Point Symbols – Mobile

- Ambiguity of visual depiction driven by physical aspects of device
  - Smaller screen → more generalization
  - Smaller screen → fewer/smaller symbols
  - Lower resolution → more aliasing/pixelation
- Benefit of mobile: real-time verification using environment

An Example


Soby's
W North St
nta Bread Co
N Main St
City Tavern
The Green Room
Guadalajara Mexican Restaurant
Marble Slab Creamery
N Main St
E Coffee St
Addy's Dutch Cafe & Restaurant


Verizon
10:52 AM
78%
Search
The UPS Store
W North St
E North St
Drake's Flowe
The Chocolate Moose
OP Taylor's
The Green Room
Sharkey's Pub
N Main St
Lemongrass Thai
Coffee Underground
Google

Evaluating Point Symbols

- Long history of evaluating point symbols
  - Tourist maps[1,2]
  - Digital displays[3]
  - Application domain[4,5]
- Limited evaluations on mobile devices

[1] Forrest, D. and Castner, H.W. (1985) The design and perception of point symbols for tourist maps. *The Cartographic Journal*. 22. 11-19.
[2] Clarke, L.M. (1989) An experimental investigation of the communicative efficiency of point symbols on tourist maps. *The Cartographic Journal*. 26. 105-110.
[3] Morrison, C. and Forrest, D. (1995) A study of point symbol design for computer based large scale tourist mapping. The Cartographic Journal. 32. 126-136.
[4] Akella, M.K. (2009) First Responders and Crisis Map Symbols: Clarifying Communication. *Cartography and Geographic Information Science*. 36. 19-28.
[5] Bianchetti, R.A., Wallgrün, J.O., Yang, J., Blanford, J.I., Robinson, A.C., and Klippel, A. (2012) Free classification of Canadian and American emergency management map symbol standards. *The Cartographic Journal*. 49(4). 350-360.

Evaluating Symbols on Mobile Devices

- Creating an evaluation tool for mobile devices
  - Design and run user studies on mobile devices
  - Define user study as series of tasks
    - Tasks test for different interpretation problems
    - Record/report results to researchers/designers
    - Automated bias correction
- Free and open source

Future Work

- Evaluating multi-modal interfaces
  - Transitions between modes
  - Visual cues indicating interaction type
- Relationship to accessibility

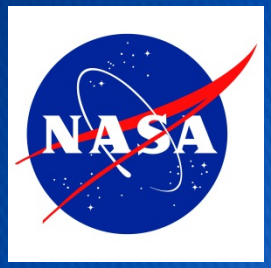
Thanks

- Dr. Alan MacEachren
- Dr. Anthony Robinson

Questions or Comments?

Symbol Sharing Tools for Emergency Management Cartography at NASA

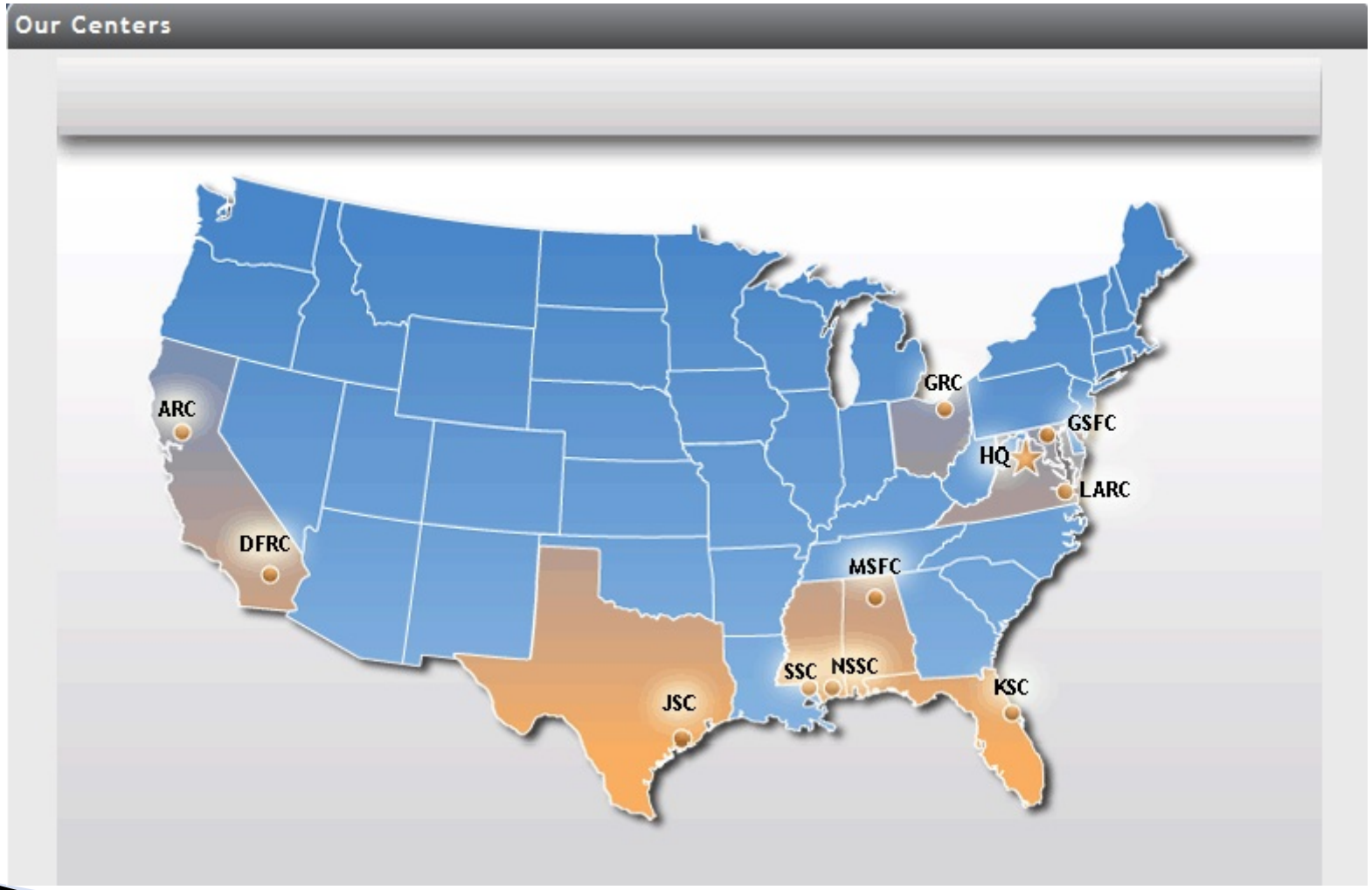
Susan A. Hulsey – Capstone Proposal
GEOG 596A



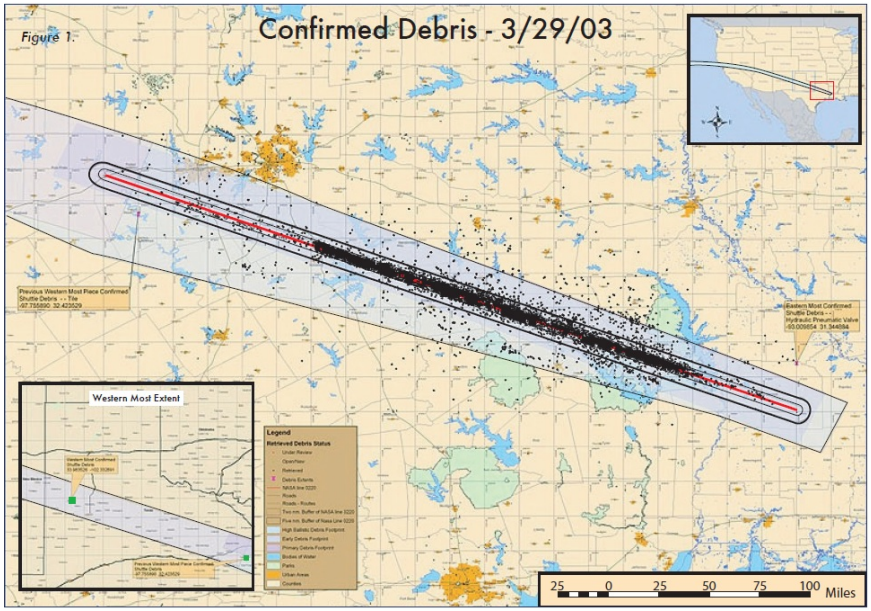
Outline

- Introduction
- Background
- Study Objectives
- Proposed Methodology
- Proposed Project Timeline
- Anticipated Results
- References

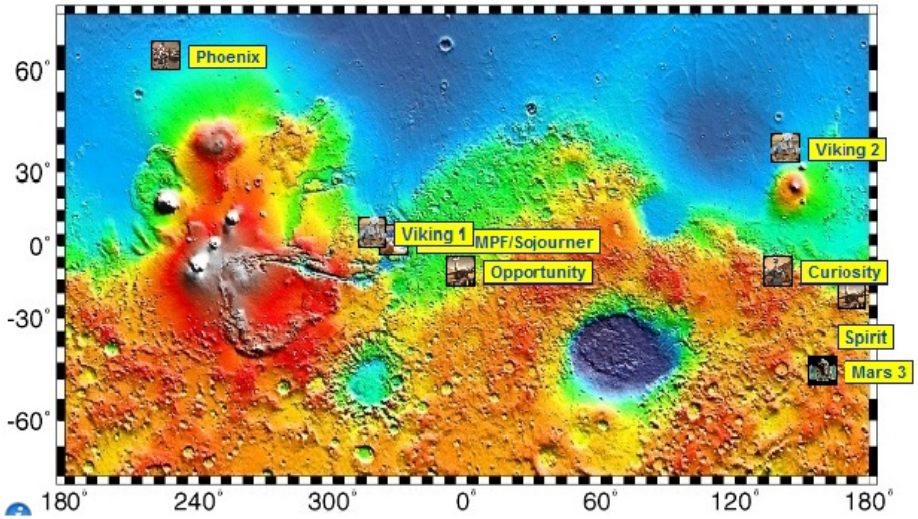
Introduction



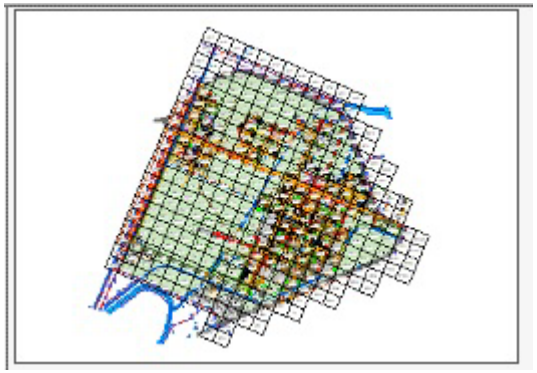
Introduction



Columbia debris recovery mapping

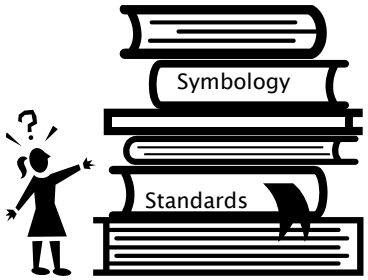


Mars rover placement mapping

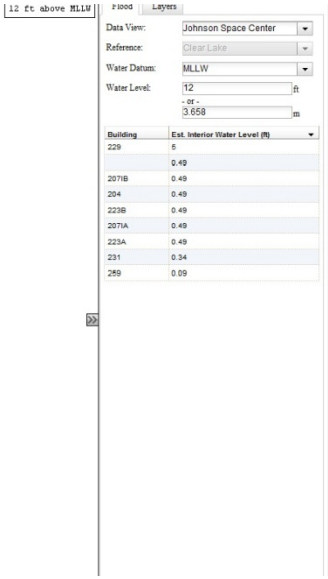


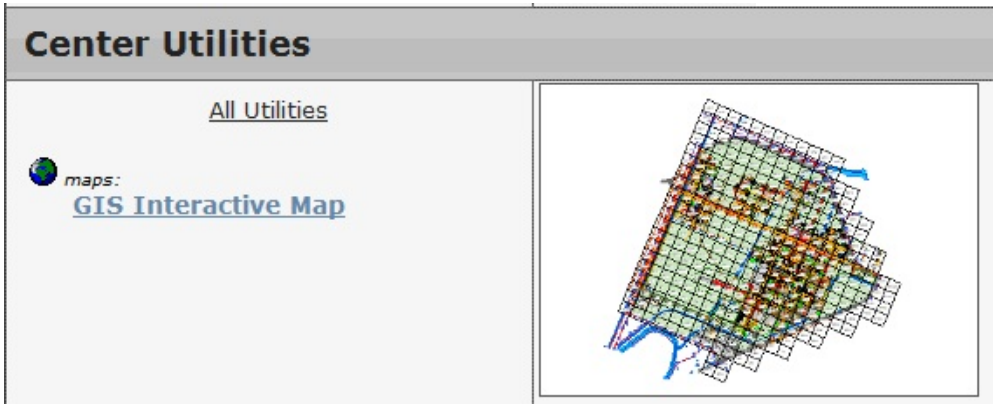
Utility mapping

Introduction

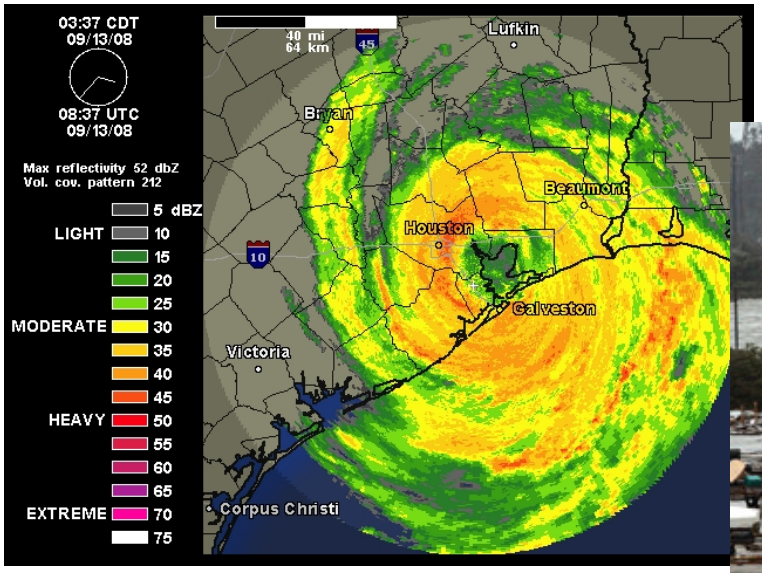








Introduction

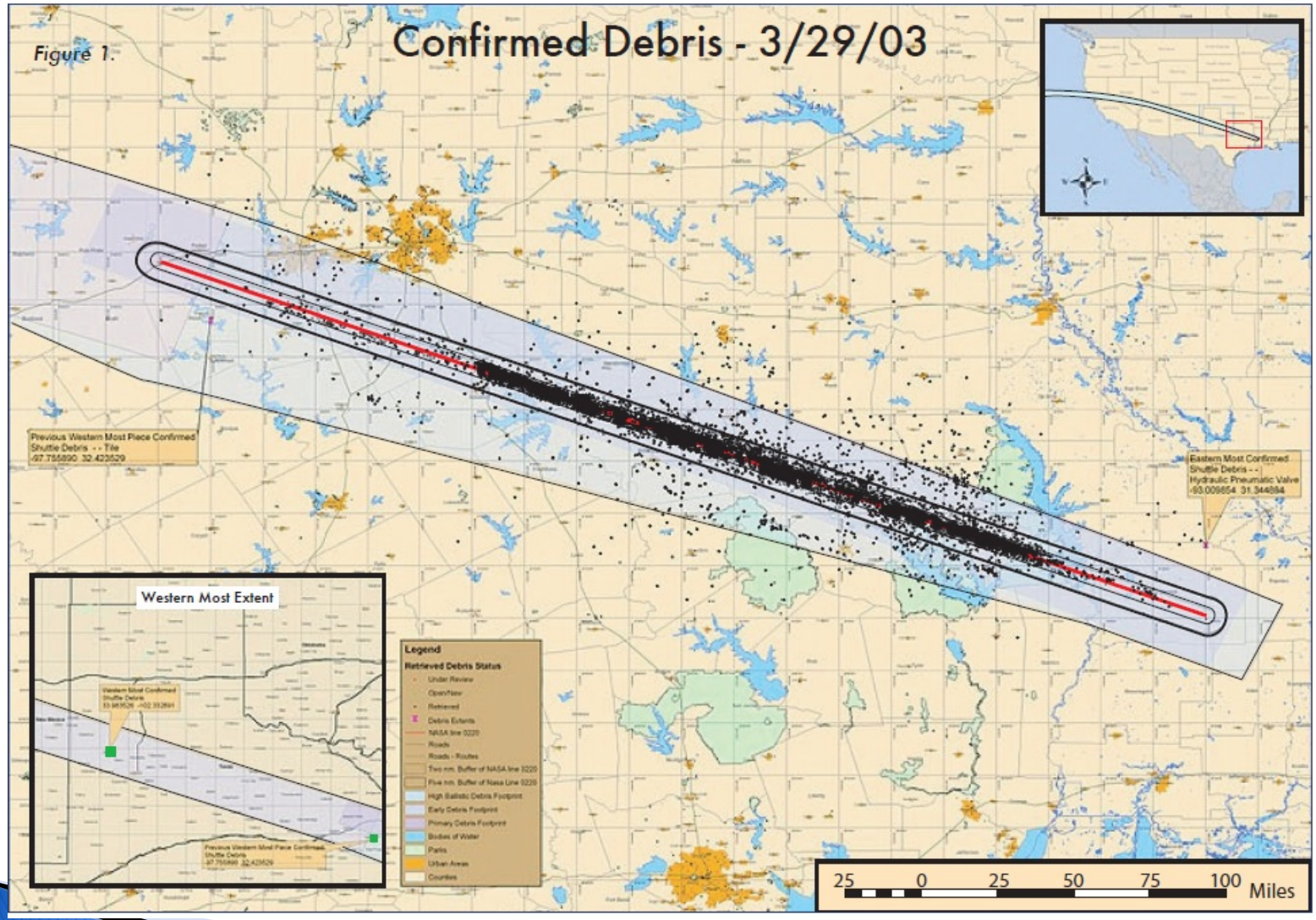




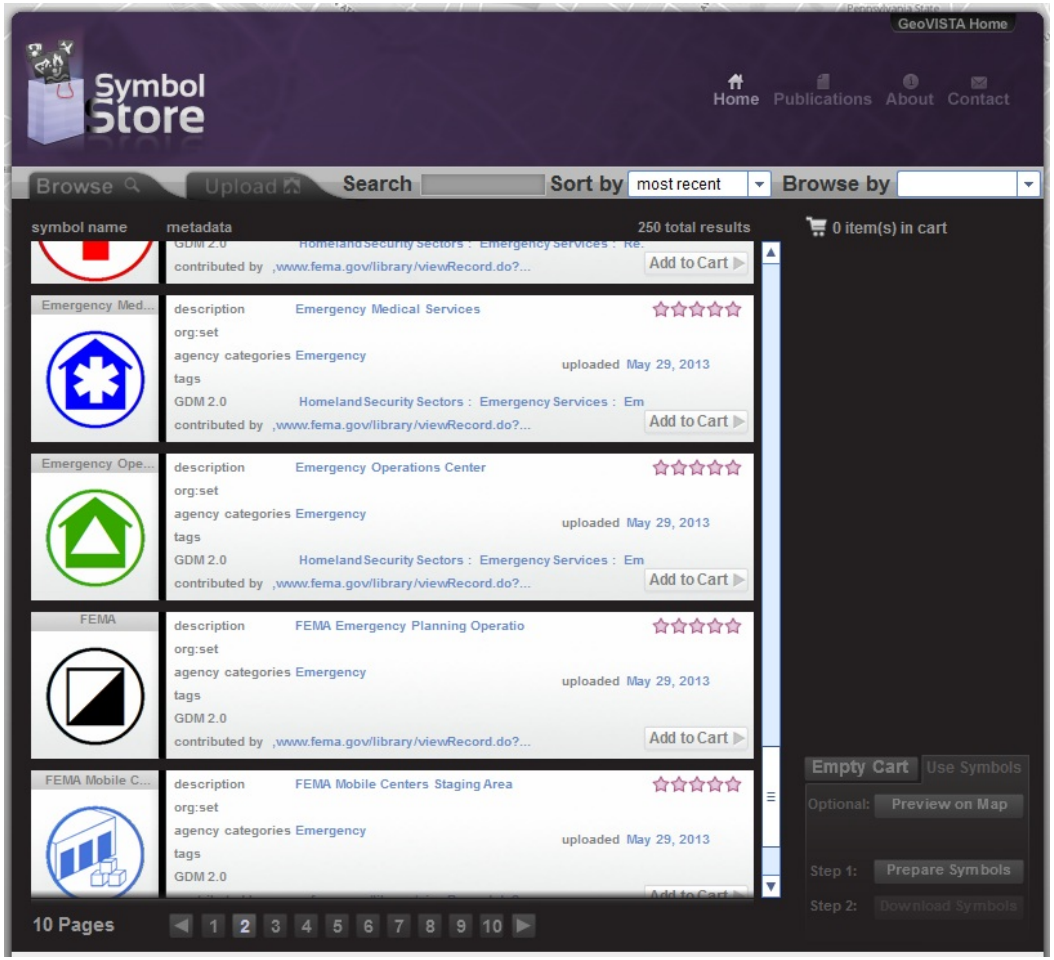


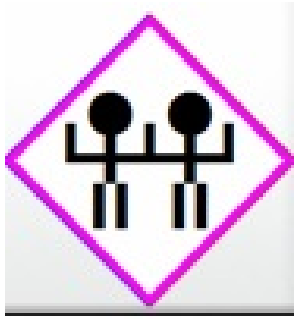


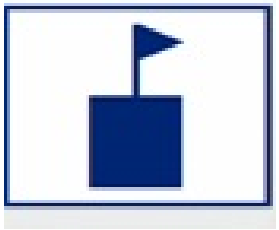
Introduction



Background

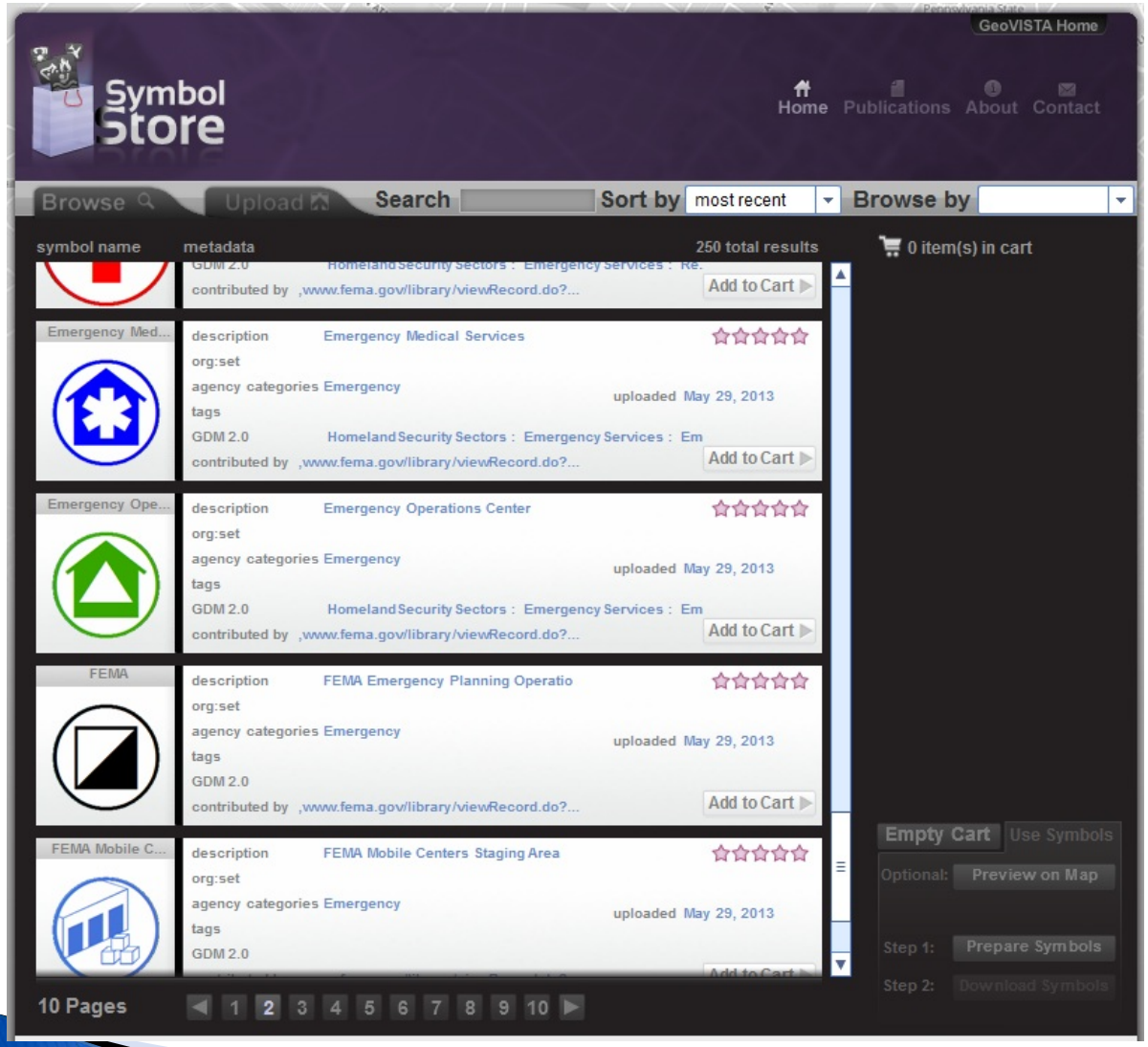

GeoVISTA Home
Symbol Store
Home
Publications
About
Contact
Browse
Upload
Search
Sort by
most recent
Browse by
symbol name
metadata
250 total results
0 item(s) in cart
Add to Cart
Emergency Med...
description Emergency Medical Services
org:set
agency categories Emergency
tags
GDM 2.0
uploaded May 29, 2013
Emergency Ope...
description Emergency Operations Center
FEMA
description FEMA Emergency Planning Operatio
FEMA Mobile C...
description FEMA Mobile Centers Staging Area
Empty Cart
Use Symbols
Optional:
Preview on Map
Step 1:
Prepare Symbols
Step 2:
Download Symbols
10 Pages







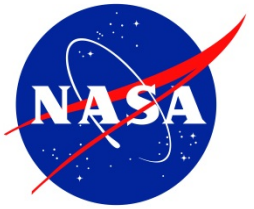
Study Objectives





- SEARCH
- RETRIEVE
- REVIEW
- DOWNLOAD

Proposed Methodology



Objectives:

- Determine if users are capable of successfully uploading symbols;
- Determine if users are capable of successfully categorizing symbols;
- Determine if users are capable of successfully adding metadata for said symbols;
- Determine if users are capable of successfully reviewing symbols after they have been uploaded and;
- Determine if there are any recurring issues or difficulties with the tool that the users experience.

Proposed Methodology

Quality metrics to determine:

- Success rate;
- Time required;
- Error rate;
- Subjective satisfaction/dissatisfaction with the symbology sharing tool.



Proposed Methodology

Breakdown – 4 Parts:

1. Structured interview;
2. Written tasks and semi-structured survey;
3. Perform tasks and complete Likert scale;
4. Compile/analyze results.



Likert Scales

Please fill in the number that represents how you feel about the computer software you have been using

I am satisfied with it

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

It is simple to use

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

It is fun to use

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

It does everything I would expect it to do

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

I don't notice any inconsistencies as I use it

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

Proposed Methodology

Quality metric to determine:

- Success rate;
- Time required;
- Error rate;
- Subjective satisfaction/dissatisfaction of the symbology sharing tool.



Proposed Project Timeline

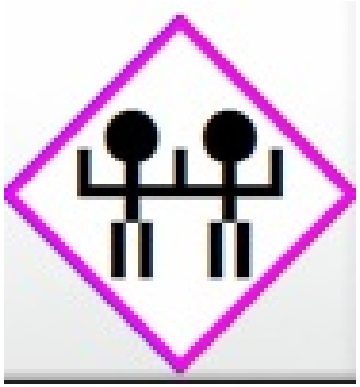




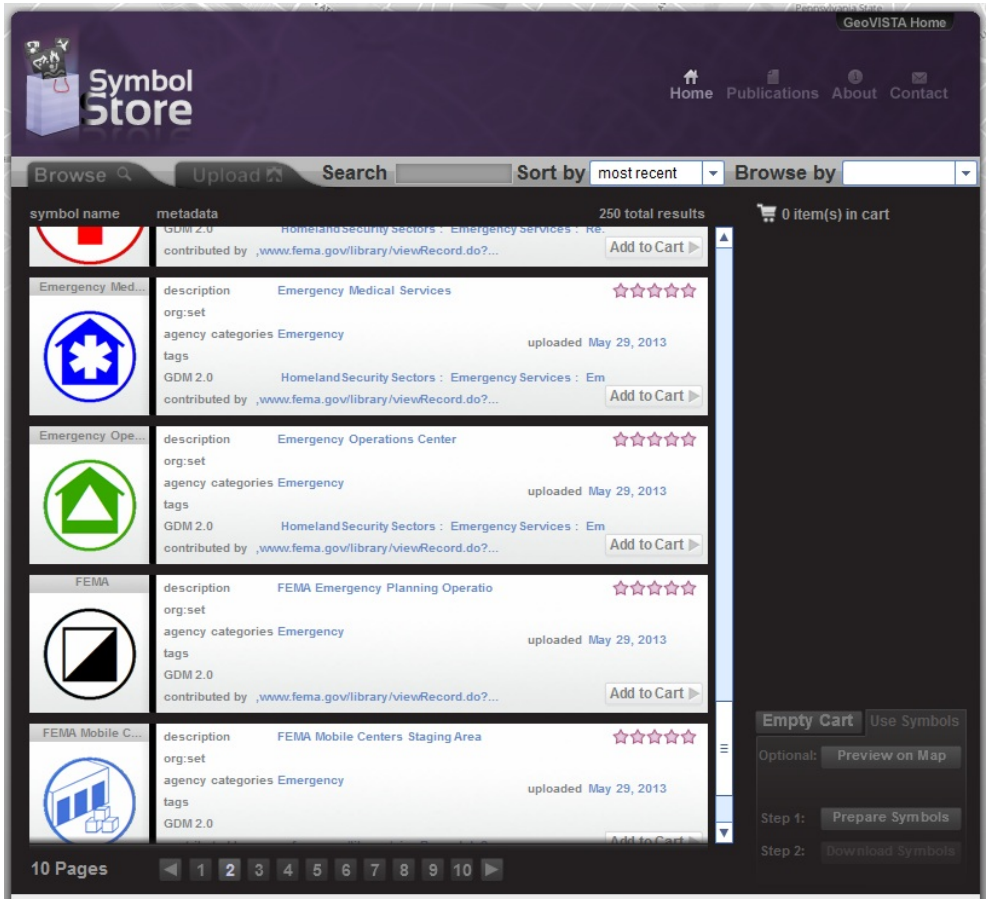
10 weeks



Anticipated Results







Resources

Akella, Mamata Kumari (2009). "*First Responders and Crisis Map Symbols: Clarifying Communication*" Cartography and Geographic Information Science, Vol. 36, No.1, pp. 19-28.

Dymon, Ute J., et.al (2011). "*Preparing an ANSI Standard for Emergency and Hazard Mapping Symbology*". Department of Geography, Kent State University.

Federal Geographic Data Committee (2013) "*The Federal Geographic Data Committee*" http://www.fgdc.gov accessed June, 2013

Pierce College (2013) "Homeland Security Emergency Management" http://wp.pierce.ctc.edu/blog/coe/2012/08/19/the-symbol-store-over-1200-map-symbols/ accessed June, 2013

Robinson, Anthony, et.al (2010) "*Challenges for Map Symbol Standardization in Crisis Management*"
Proceedings of the 7th International ISCRAM Conference, Seattle WA, May 2010.

Robinson, Anthony, et.al (2011) "*Understanding User Needs for Map Symbol Standards in Emergency Management*" Journal of Homeland Security and Emergency Management: Vol. 8: Iss. 1, Article 33. DOI: 10.2202/1547-7355.1811

Robinson, Anthony, et.al (2012) "*Developing map symbol standards through an iterative collaboration process*" Environment and Planning B: Planning and Design 2012, volume 39, pages 1034-1048.

Robinson, Anthony, et. al (2013) "*Symbol Store: sharing map symbols for emergency management*" Cartography and Geographic Information Science, DOI: 10.1080/15230406.2013.803833

VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence



Homeland Security

Science and Technology

PURDUE

UNIVERSITY

VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

Panel 1: Big Data, Visualization, and Risk Based Decision Making

David S. Ebert

CVADA

Center for Visual Analytics and Data Analytics

A U.S. Department of Homeland Security Center of Excellence

November 2013

Panelists

- Deborah Durham, DHS
- Starnes Walker, Mitre??
- Ben Maule, USCG
- Fred Roberts, CCICADA (Rutgers)
- Abish Malik, VACCINE (Purdue)

Setting the Stage

Big Data: Solution to Global Challenges?



- **Big Data Definition:** A phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex and diverse types of data. (TechAmerica Foundation)
  Big Data is often defined along three dimensions:
  - **Volume** – size
  - **Velocity** – rate of input, update, change
  - **Variety** – different types, sources, variables
- Need:
  - Advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of information (TechAmerica Foundation)
  - **Enable effective, efficient analysis, decision making, planning, and action**

Big Data Challenges

- Heterogeneity
- Unprecedented volume and/or complexity
- Multivariate, Multidimensional
- Streaming
- Qualitative, Quantitative, Categorical, Temporal,…
- Distributed Sources: sensors, text, images, maps, …
- Inconsistent and Uncertain
- Different Scales and Cross Scale

Big Data Analytics – The Solution?

- Advanced data analytics for big data can help manage big data
  - Machine learning
  - Advanced statistical analysis
  - Hadoop-deployed analytics
- However, it falls short in several aspects

What Big Data Analytics Can't Do
(inspired and adapted from David Brooks, New York Times, 2/18/2013)

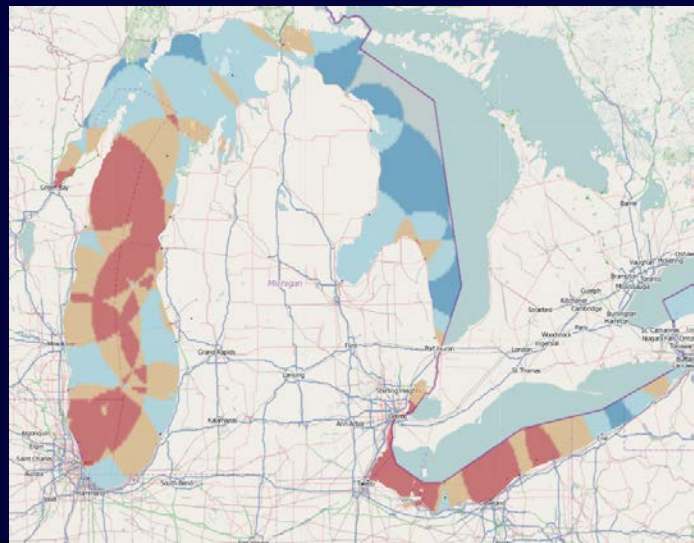
- Qualitative, fuzzy, and social data
  - Preferences, significance of one relationship over another
  - Trust
- Context
  - Data is rarely complete, nor does it incorporate all relevant information for decision making
  - Humans have extensive information and experience that never make it into the collected data

What Big Data Analytics Can't Do
(inspired and adapted from David Brooks, New York Times, 2/18/2013)

- Spurious vs. Significant
  - Big data means more statistically significant events and correlations, but they may not have any relevance
  - Increases noise to signal ratio
- Big problems
  - Complex, multifaceted, multiparameter big challenges with unquantified dependencies
  - Exactly the topic of this meeting!

The Power of Visual Analytics

- Turn the abstract into the concrete
- Enable rapid perception and visual cognition of information to increase efficient understanding and decision making
- Goal: Transform *data* into *information and knowledge*
- Data size doesn't matter,
  size of insights matters
- Visualization is a key method to increase understandability of data and analytical results



VA Approaches to Big Data

- Don’t display all the data!
- Extract relevant information at the appropriate natural scales using analytics and statistical models to reduce data space so that it is mappable to the visual space
- User interaction to guide above process
- Adapt to available hardware for the problem
  - Should we use big iron or cell phones?

Visual Analytics Uses for Risk-Based Decision Making



- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Interactive, multi-day, month, type investigation
- Multisource, multimedia data integration & analysis



VACCINE

Visual Analytics for Command, Control and Interoperability Environments
A U.S. Department of Homeland Security
Science and Technology Center of Excellence

National Strategic Maritime Risk Stakeholders Alliance Meeting

Data-Driven Risk-Based Decision Making in the Maritime Domain

November 20th and 21st, 2013

Ross Ade Stadium Buchannan Club
Purdue University I West Lafayette, IN

VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence



Homeland Security

Science and Technology

PURDUE

UNIVERSITY

VACCINE

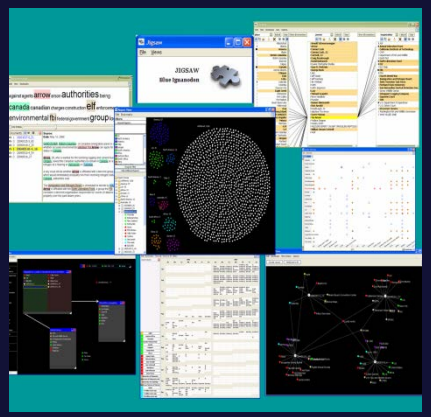
Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

COAST Case Study

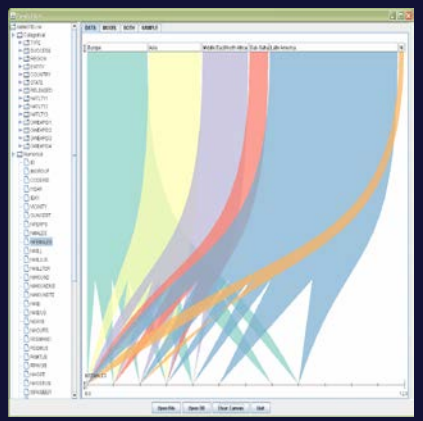
DHS Data Workshop

Dr. David Ebert

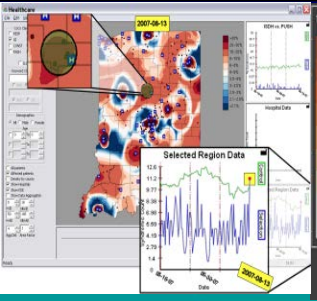
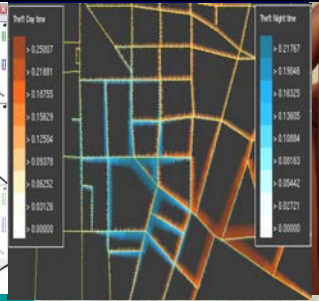




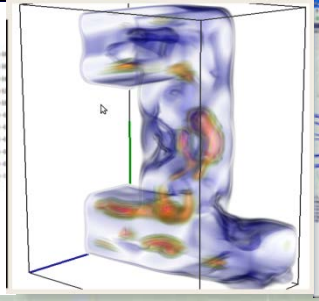
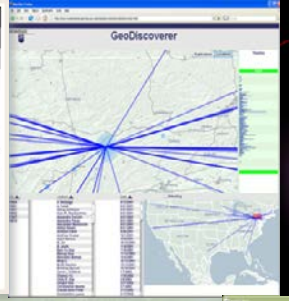
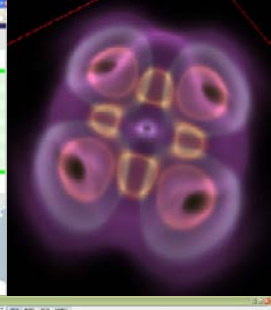
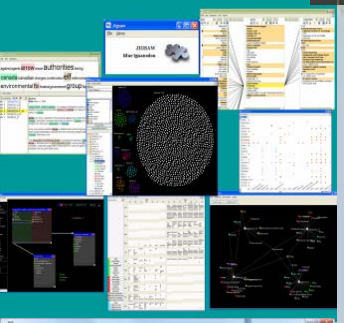
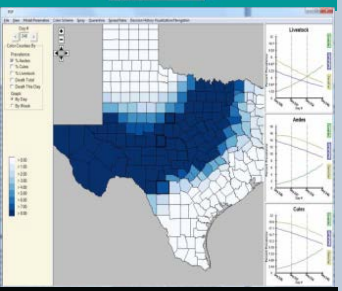
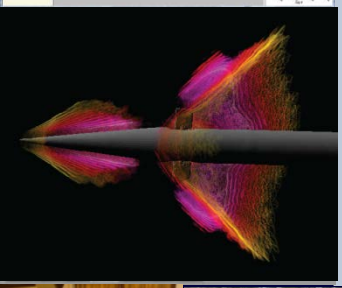
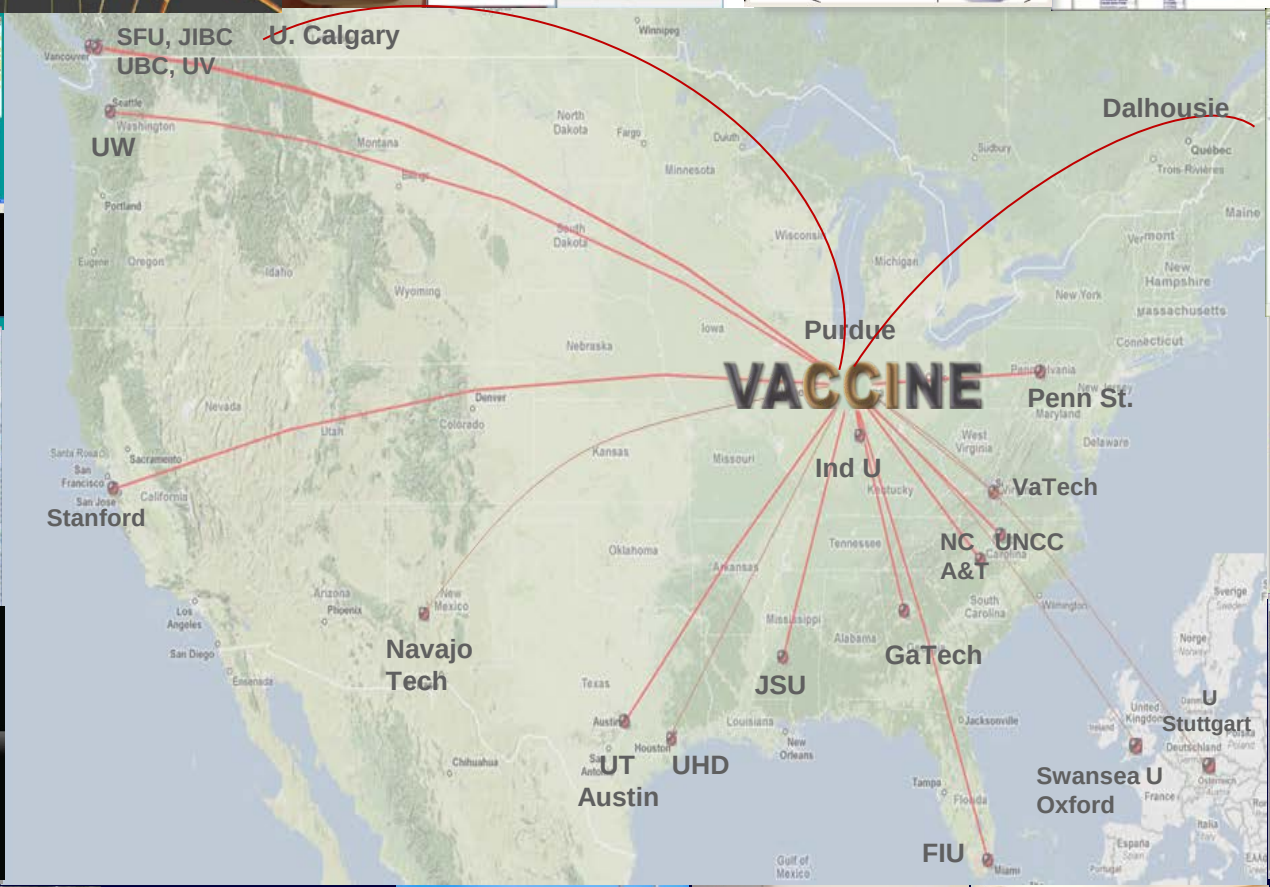
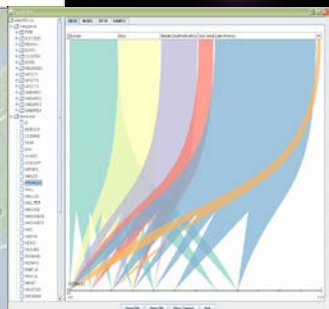


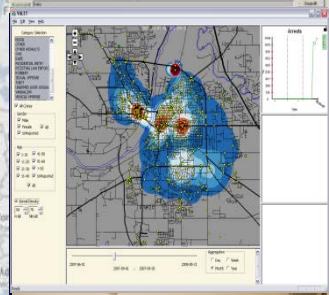





Selected Region Data


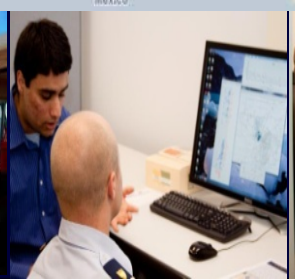




GeoDiscoverer





SFU, JIBC
UBC, UV
U. Calgary
UW
Dalhousie
Purdue
VACCINE
Penn St.
Ind U
VaTech
NC A&T
UNCC
Stanford
Navajo Tech
JSU
GaTech
UT Austin
UHD
U Stuttgart
Swansea U
Oxford
FIU






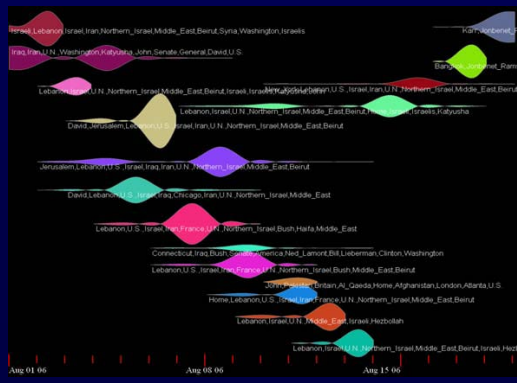
VACCINE
PURDUE





VACCINE Mission





- Provide visual analytic and scalable solutions for all 2.3 million extended homeland security personnel
  - 185,000 DHS personnel, 350,000 law enforcement personnel, 750,000 homeland security practitioners, 2.3 million extended personnel
- Achieve excellence in visual analytics and visualization sciences
- Educate homeland security stakeholders and the next generation of talent



VACCINE

VACCINE: What We Do – Visual Analytic Solutions

We enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable user to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

Visual Analytic Solutions: What We Offer

- **Improved Effectiveness**: We enable users to be more effective through innovative, interactive visualization, analysis, and decision making tools
  - Provide the right information, in the right format, within the right time to solve the problem
  - Enable user to be more effective from planning to detection to response to recovery
  - Enable effective communication of information
- **Innovative Fielded Solutions:** We provide innovative visual analytic and scalable solutions to the extended homeland security community
- **People and Partnerships**
  - Interdisciplinary world-leading team of researchers and students

"cgSARVA has proven its worth time and again, providing key analytic information for decision makers for large scale projects…"

VADM Robert Parker, 2012 MRS Keynote Address

VADM Robert Parker with VACCINE student researchers (cgSARVA, COAST, iOPAR)



COAST

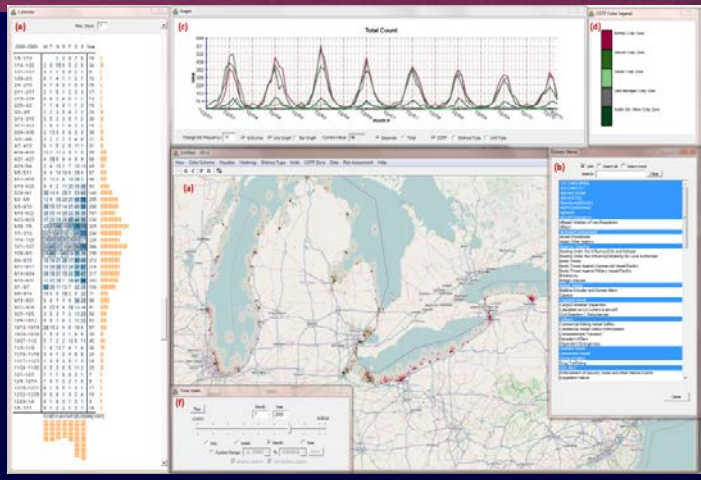
- Coast Operations Analysis Suite of Tools
  - This suite of tools includes visualization, simulation, mathematical modeling, and optimization techniques
  - VACCINE focused specifically on resource allocation to aid decision making – took the form of a visual analytics tool, cgSARVA
  - cgSARVA utilizes MISSILE data in order to quickly and easily understand operational redundancies, capabilities, historical mission demands,costs, and assist in optimizing efficiency.

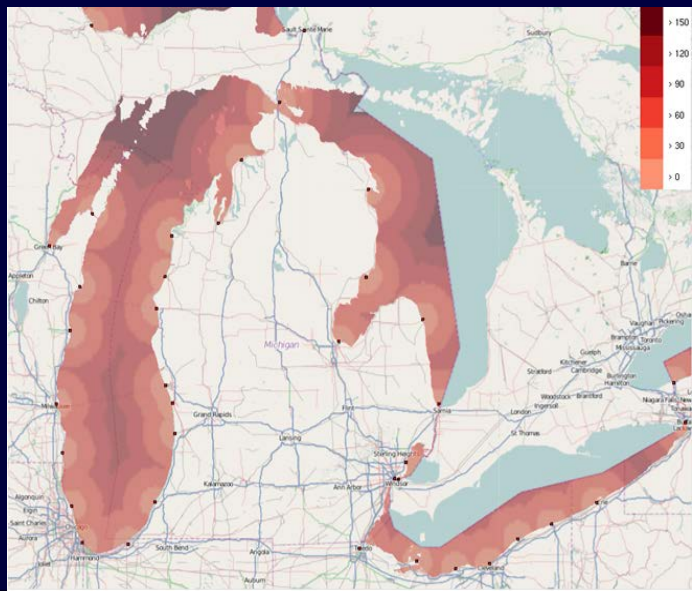
U.S. Coast Guard Search and Rescue VA (cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771

IMPACTS:

- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Key component of USCG D9 and CG-wide reallocation plan based ondecreased budget
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ

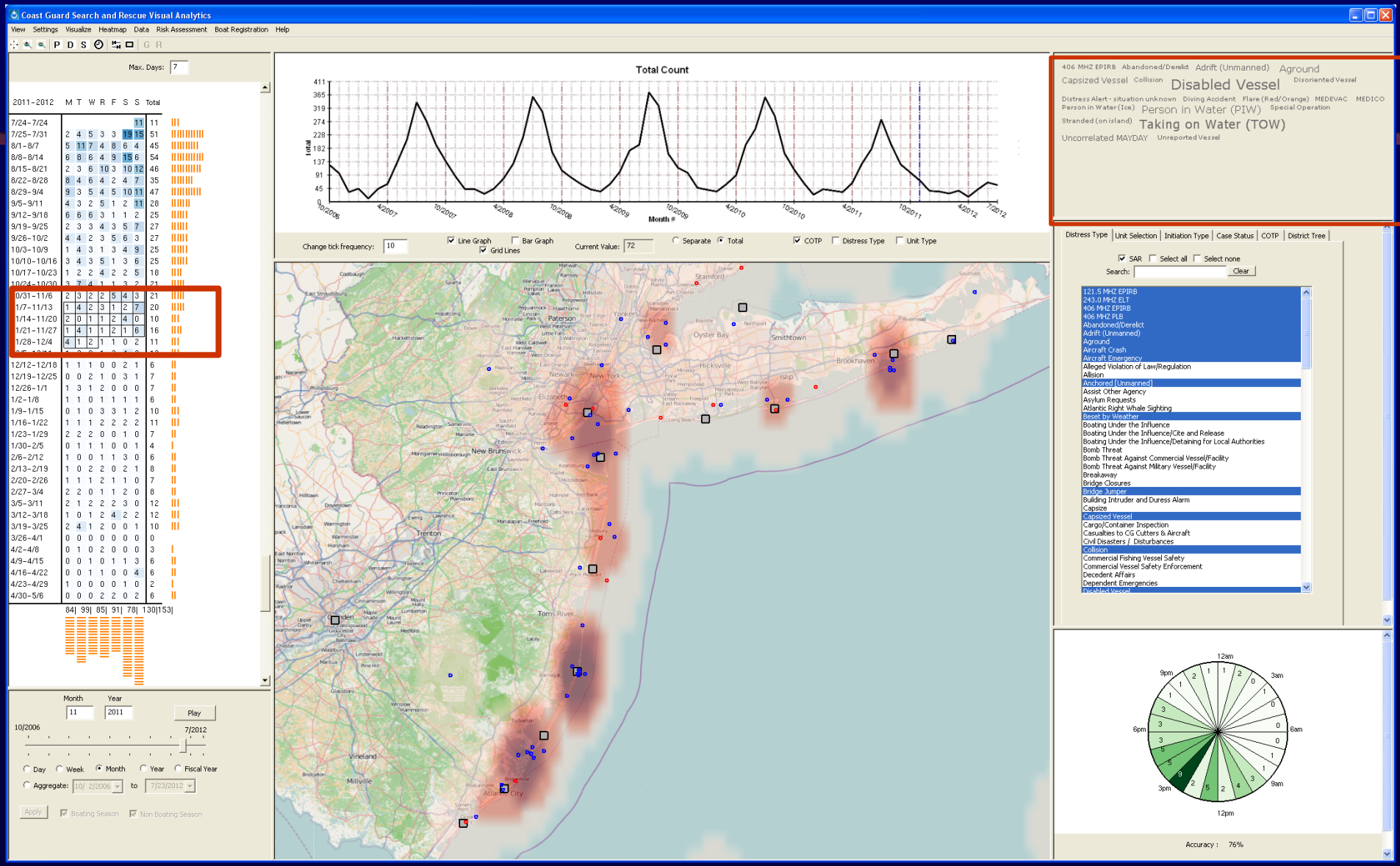




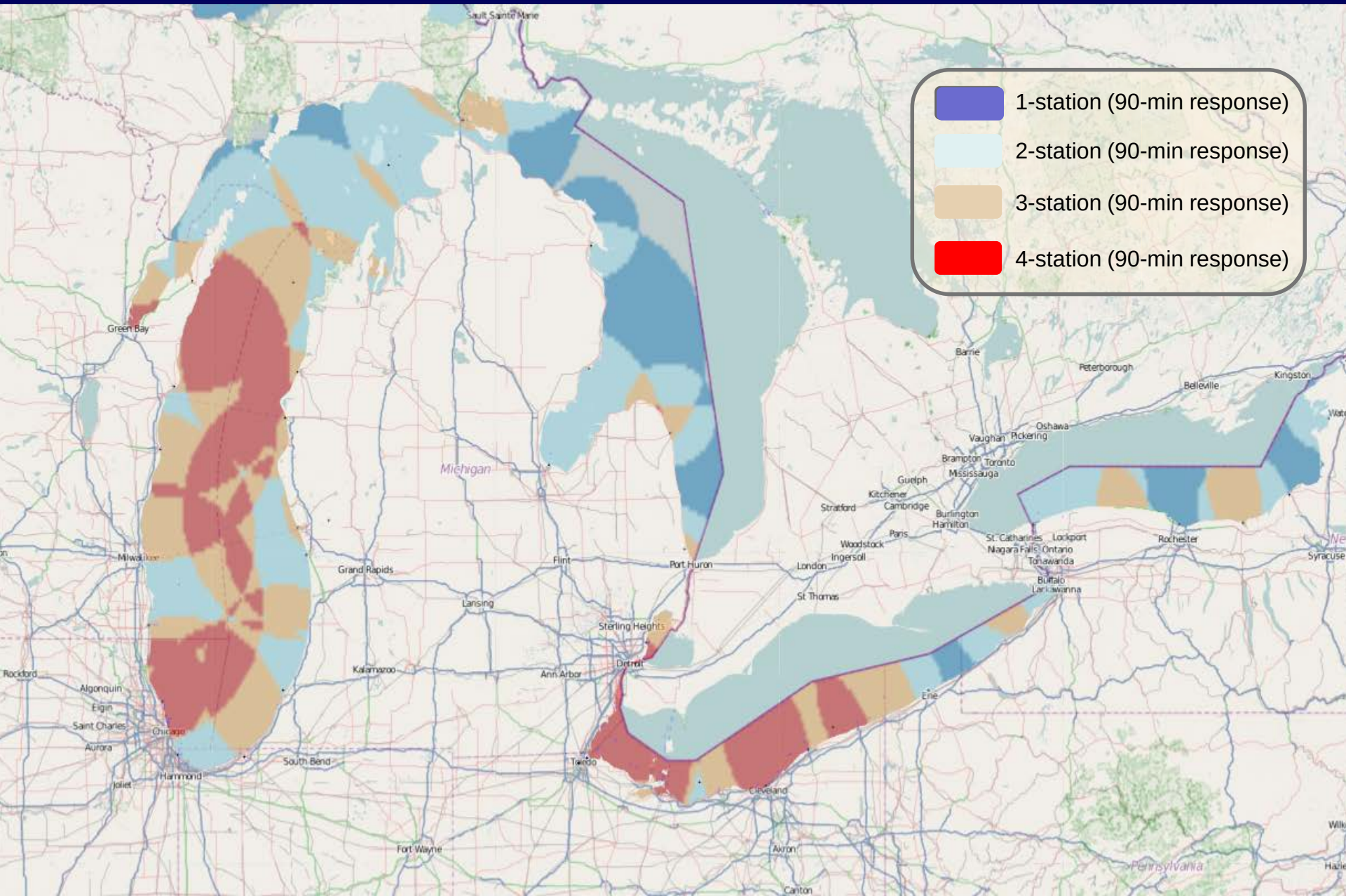
VACCINE

cgSARVA: Risks and Consequences From Sandy:

SAR Cases November 2011 NJ/NYC Area



Response Efficiency



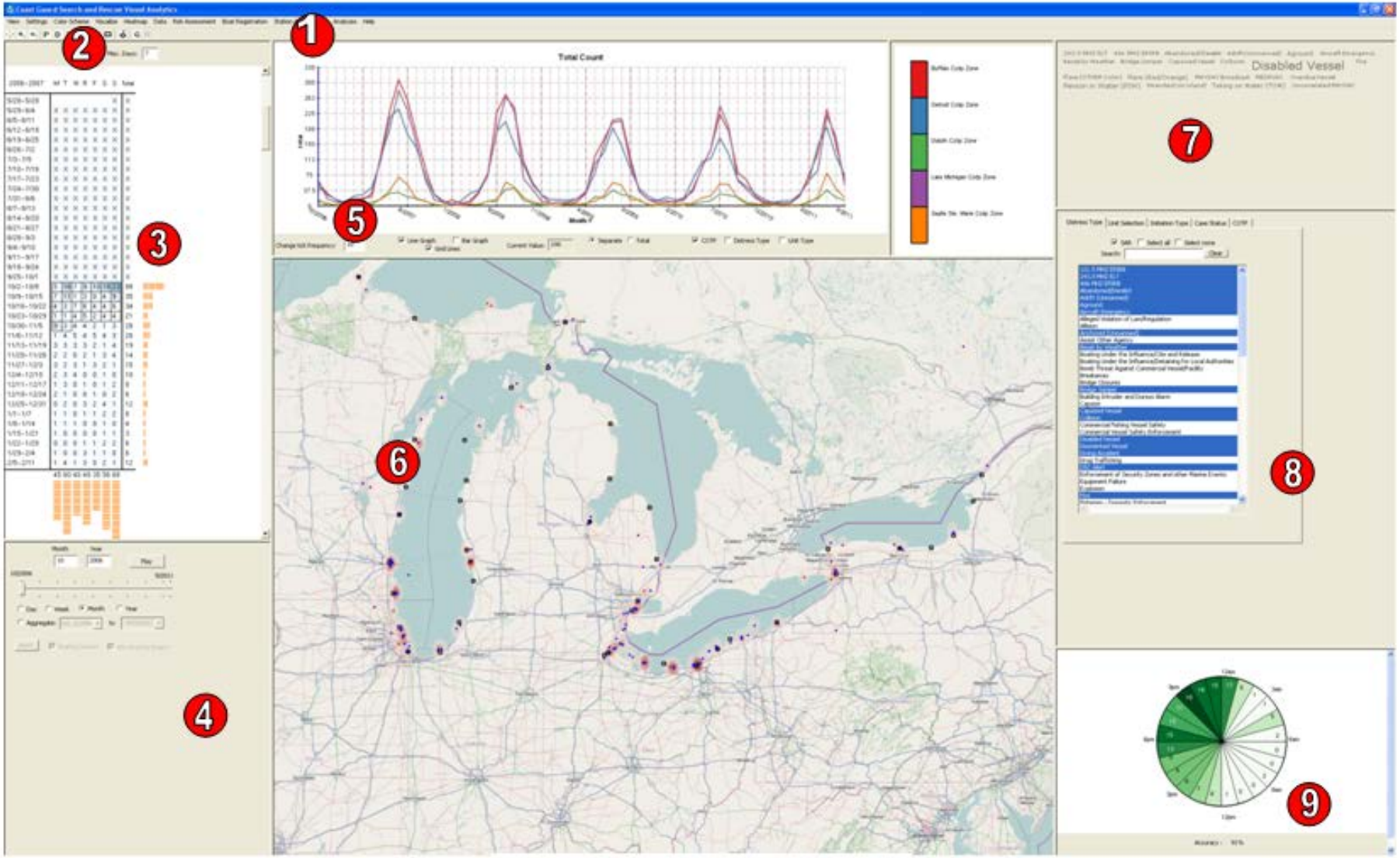


Figure 3: *A screenshot of the cgSARVA system*

1. Menu
2. Toolbar
3. Calendar View
4. Time Slider Widget
5. Time Series View
6. Map View
7. Word Cloud
8. Interactive Selection Menu
9. Clock View

Clarity of Mission Objectives

- A few key stakeholders and decision makers had buy-in moving the project forward
- Initial question looked at D9 search and rescue missions in the Great Lakes
- A well defined problem with a common understanding between researchers and stakeholders of the goal



Authorization and Access Control

- MISSILE database is USCG mission database with eleven categories of mission, e.g.,:
  - Search and Rescue
  - Marine Safety
  - Law Enforcement
- Actual CG data was provided.
  - Exported from CGBI to USCG local system
  - Copied to laptop on unclass network
  - Sent to us
- NDA was signed

Process, Procedures, and Communication

- With stakeholder support, the project was conducted with limited points of contact – actual analysts who knew the right questions to ask and could help us design a better system
- It was key to have their constant/weekly feedback.
- Data was shared as needed – USCG provided initial dataset to use as a model.

Technology and Execution

- Quickest path forward was to receive and manage FOUO data.
- Started with SAR data that has no intelligence or sensitive information
- There are “quality” issue with most data, but mostly from a reporting errors and lack of complete information (e.g., lat/lon, time of day, origin of case)
- Care needed to make decisions based on incomplete information (mission hours vs. reduction in risk, e.g., Detroit River Bridge)

Resolution

- cgSARVA official underwent the VV&A process through the Coast Guard Research and Development Center
  - The R&D Center tested the software under a number of parameters to verify the validity of the tool.
  - This process certified it for Coast Guard-wide use for decision making
  - The first tool from a COE to undergo this process

Lessons Learned

- Clearly define the problem
- Obtain sample data set ASAP
- Work with a highly motivated end-user who is tenacious
  (Quick start result of CDR Ben Maule)
- Be certain to have regular contact with end user/ stakeholder and remember you are serving them – know your audience!



VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

Empowering Discovery and Decision Making Through Visual Analytics

David S. Ebert

Purdue University

ebertd@purdue.edu

CVADA

Center for Visual Analytics and Data Analytics

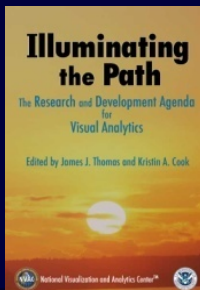
A U.S. Department of Homeland Security Center of Excellence

May 2014

What We Do: Visual Analytics

Visual Analytics[1] is the science of analytical reasoning facilitated by interactive visual interfaces

Interactive visualization, data analysis, exploration, and decision making with human in the loop!



1. *Illuminating the Path: The R&D Agenda for Visual Analytics, Editors: Thomas and Cook*

Big Data: Solution to Global Challenges?



- Big Data is often defined along three dimensions:
  - **Volume** – size
  - **Velocity** – rate of input, update, change
  - **Variety** – different types, sources, variables
- Need:
  - Advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of information (TechAmerica Foundation)
  - **Enable effective, efficient analysis, decision making, planning, and action**

Why VA For Big Data?

(Or What Big Data Analytics Can't Do)

(inspired and adapted from David Brooks, New York Times, 2/18/2013)

- Qualitative, fuzzy, and social data
  - Preferences, significance of one relationship over another; Trust
- Context
  - Data is rarely complete, nor incorporates all relevant information Humans have extensive information and experience that never make it into the collected data
- Spurious vs. Significant
  - Big data means more statistically significant events and correlations, but they may not have any relevance
  - Increases noise to signal ratio
- Big problems
  - Complex, multifaceted, multiparameter big challenges with unquantified dependencies

VA Approaches to Big Data

- Don’t display all the data!
- Extract relevant information at the appropriate natural scales using analytics and statistical models to reduce data space so that it is mappable to the visual space
- User interaction to guide this process
- Adapt to available hardware for the problem

What Our Visual Analytic Solutions Offer

We enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable user to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

Our Teams' Benefits

- **Improved Effectiveness**: We enable users to be more effective through innovative, interactive visualization, analysis, and decision making tools
  - Provide the right information, in the right format, within the right time to solve the problem
  - Enable user to be more effective from planning to detection to response to recovery
  - Enable effective communication of information
- **Innovative Fielded Solutions:** We provide innovative visual analytic and scalable solutions to the extended homeland security community
- **People and Partnerships**
  - Interdisciplinary world-leading team of researchers and students

"cgSARVA has proven its worth time and again, providing key analytic information for decision makers for large scale projects…"

VADM Robert Parker, 2012 MRS Keynote Address

VADM Robert Parker with VACCINE student researchers (cgSARVA, COAST, iOPAR)



Our Visual Analytics Process

(extended and adapted from Bertin, Normal, Tversky)

Process:

- Define problem / question
- Determine:
  - Workflow and interaction
  - Relevant data
  - Data characteristics (e.g., types, quality)
  - Abstraction/representation level
  - Natural scales
- Map to appropriate visual representation for task

- Utilize iterative refinement with user feedback from the start

Guiding Principles:

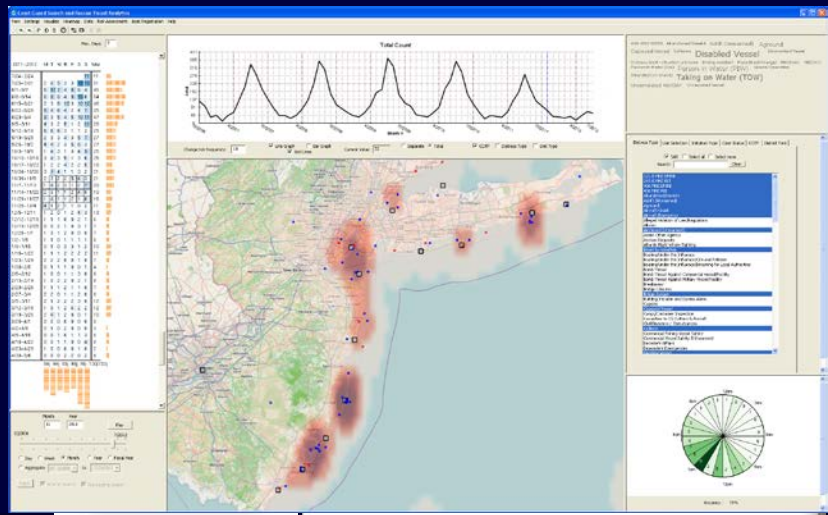
- Appropriateness principle
  - Neither more nor less
- Naturalness principle
  - Match with cognitive models
- Matching principle
  - Match the task, offer action affordances
- Principle of apprehension
  - Content should be accurately and easily perceived

Example of Users With a Problem: Law Enforcement - Tasks and Uses

- Police Chief/Sheriff - strategic goals
  - Resource allocation, return on investment
  - Yearly effectiveness (Are we safer?)
  - Emerging trends
- Police Captain – short-term effectiveness
  - Force allocation for the day/week
  - Emerging trends and anomalies
- Officer – improved situational awareness
  - What happened near me around the same time in the past?
  - What will make me better prepared when I respond to this call?
- Crime analyst/investigator – connections
  - Correlations, dependencies, trends

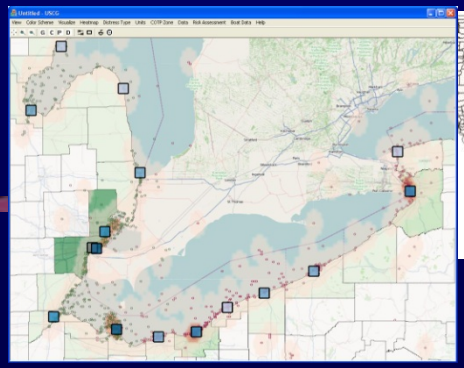
Research Area Overview and Example Projects

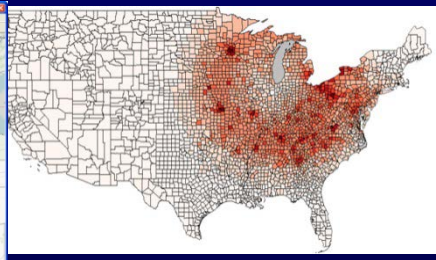
- Research applications areas
- Example projects
  - Public safety and law enforcement
  - Social media visual analytics
  - US Coast Guard
  - Additional projects

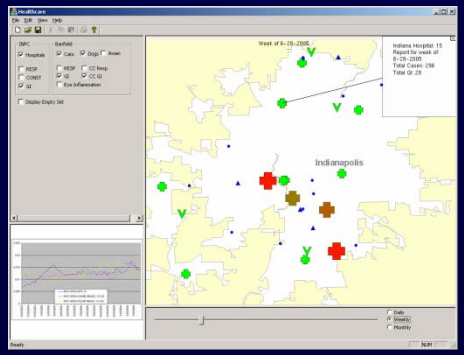


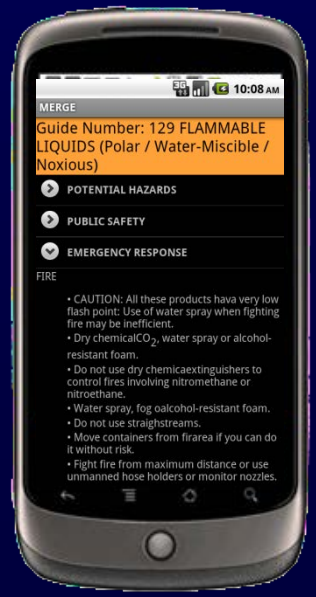
Public Safety & Health

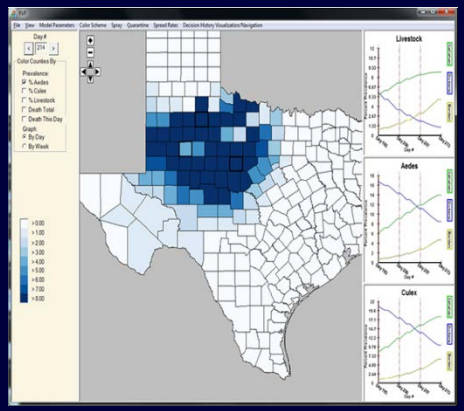
- Public safety visual analytics
  - cgSARVA
    - Coast Guard Search And Rescue Visual Analytics
  - MERGE
    - Mobile Emergency Response Guide
- Public health visual analytics
  - LAHVA
    - Linked Animal-Human Health Visual Analytics
  - RVF
    - Rift Valley Fever
    - Decision support environment for epidemic modeling and responses
  - PanVis
    - Pandemic influenza modeling and visualization tool
  - Cancer Care Engineering











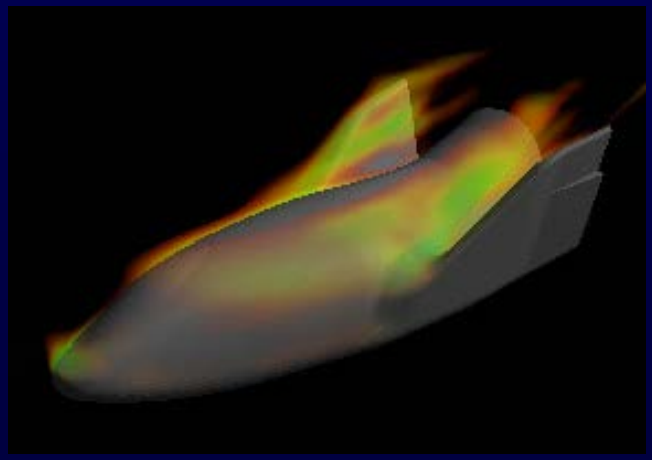


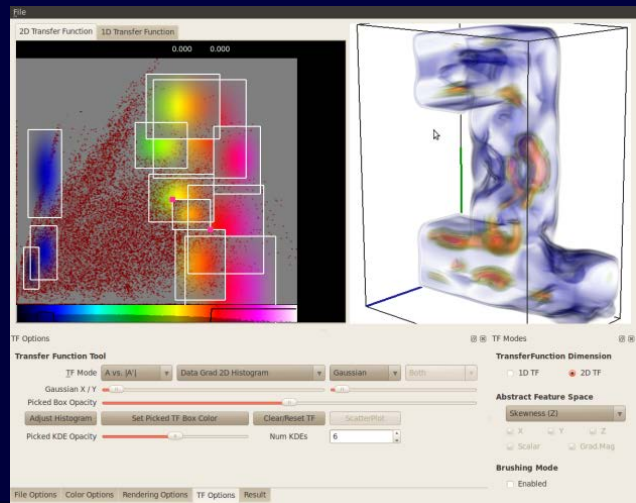
VACCINE

Scientific Discovery

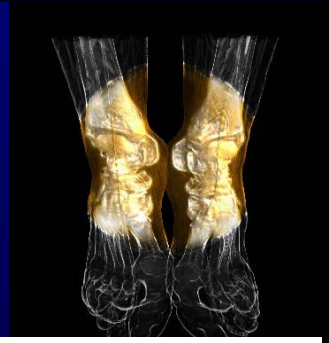


- Flow dynamics visualization (Purdue, TACC)
  - Providing insights on large flow data
  - Visualization linked with simulations
  - Innovative feature visualization
- Nanohub
  - Information-assisted data analysis and visualization of nanoelectronic models

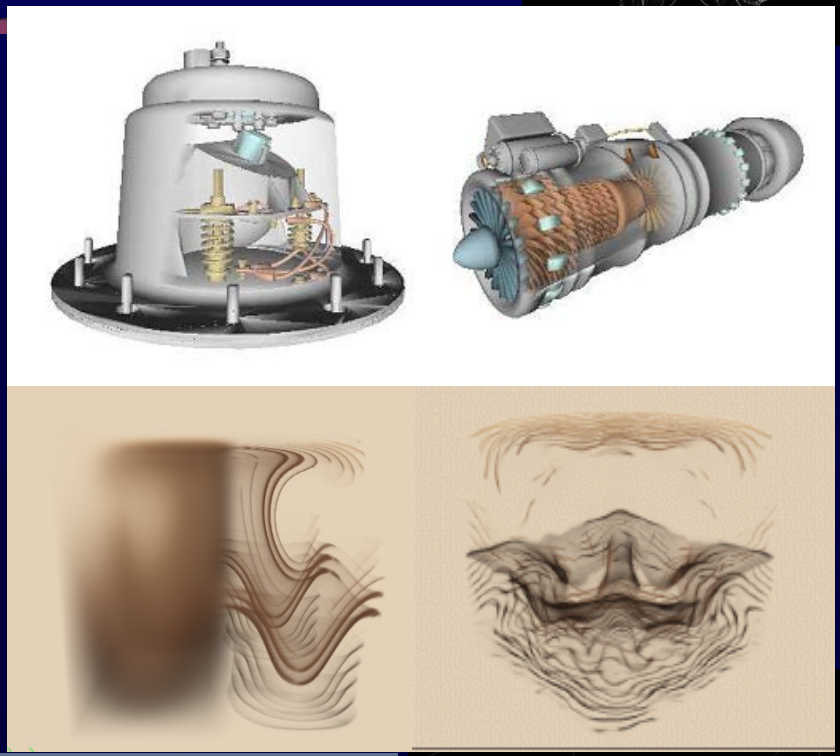


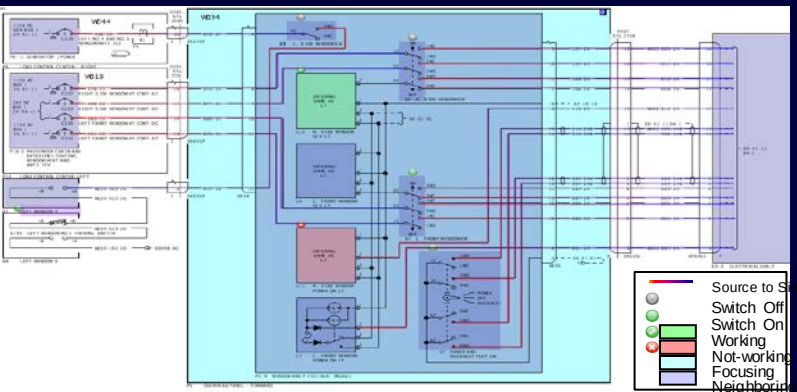


Scientific Discovery

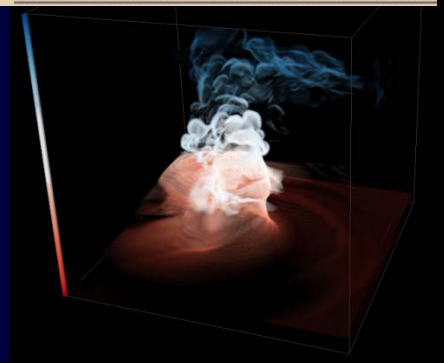


- Illustrative Visualization
  - Mechanical assemblies
  - Schematics and wiring diagrams
  - Illustrative flow visualization
  - Illustrative medical visualization
- Interactive multivariate atmospheric science data visual analytics



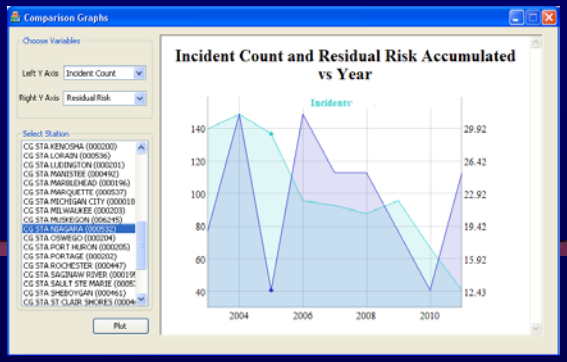


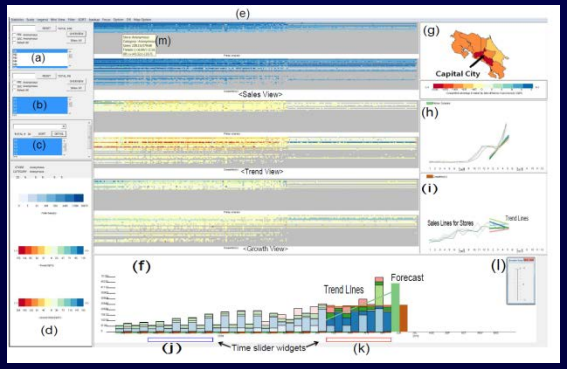




Business Visual Analytics

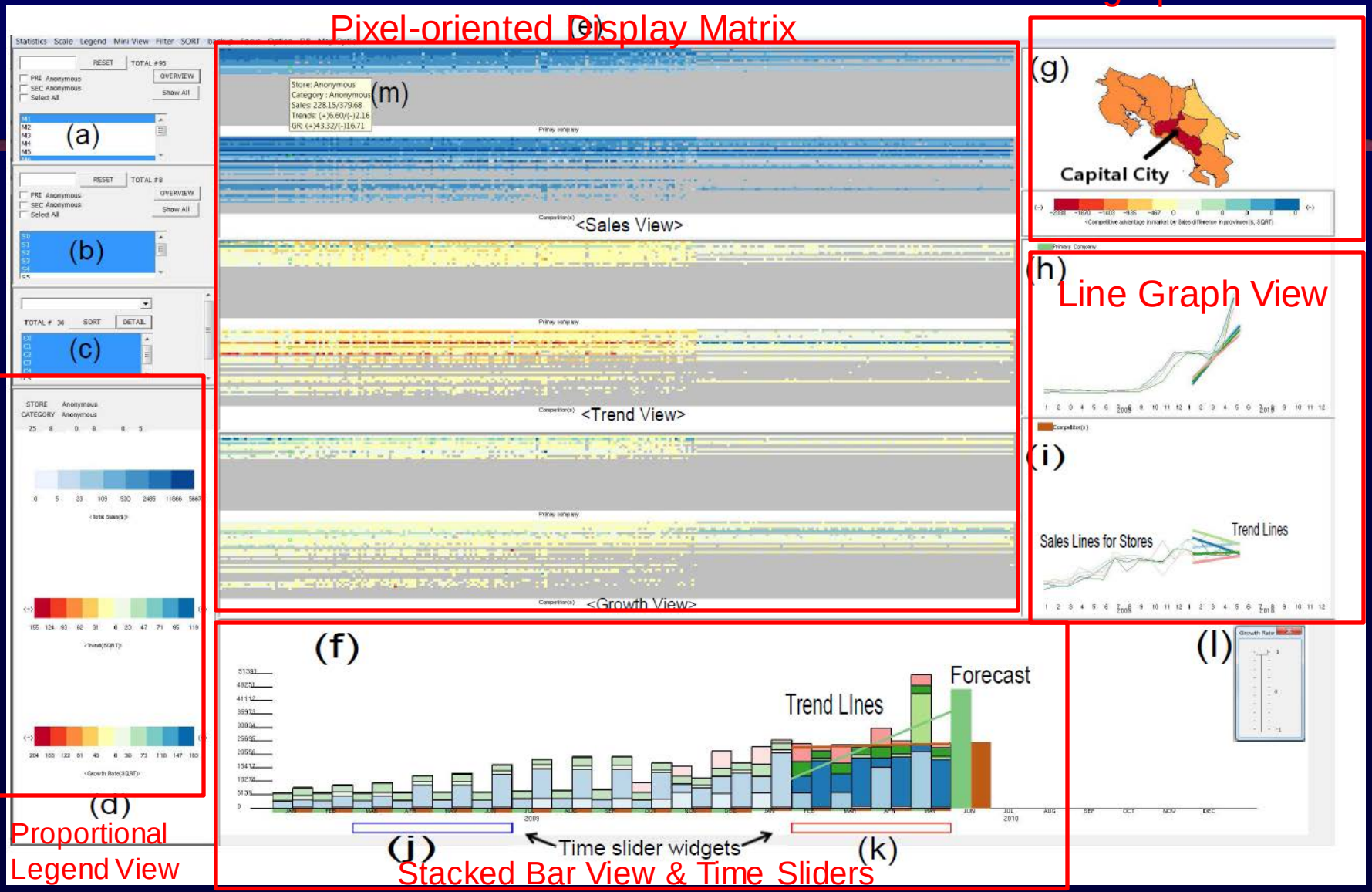
- Risk-based decision making and resource allocation
  - Coast guard operational risk assessment model
  - Helping to prioritize efforts to minimize risk
- Competitive Intelligence
  - Visual analytics system for business intelligence
- EconVIS
  - Visual analytics in various economic problems
  - Improving decision making and identifying key motivations in knowledge creation







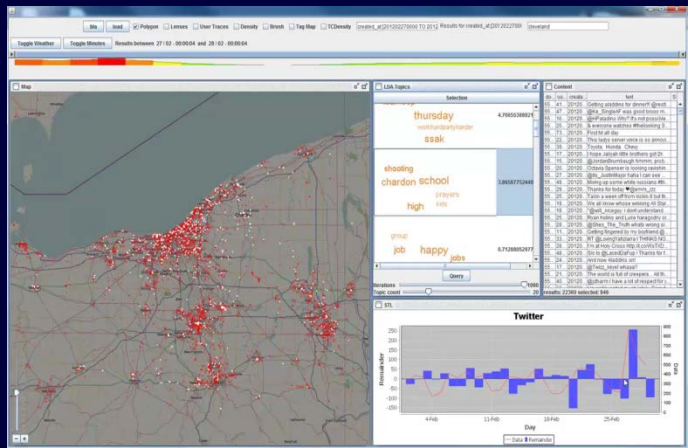
Market Analyzer



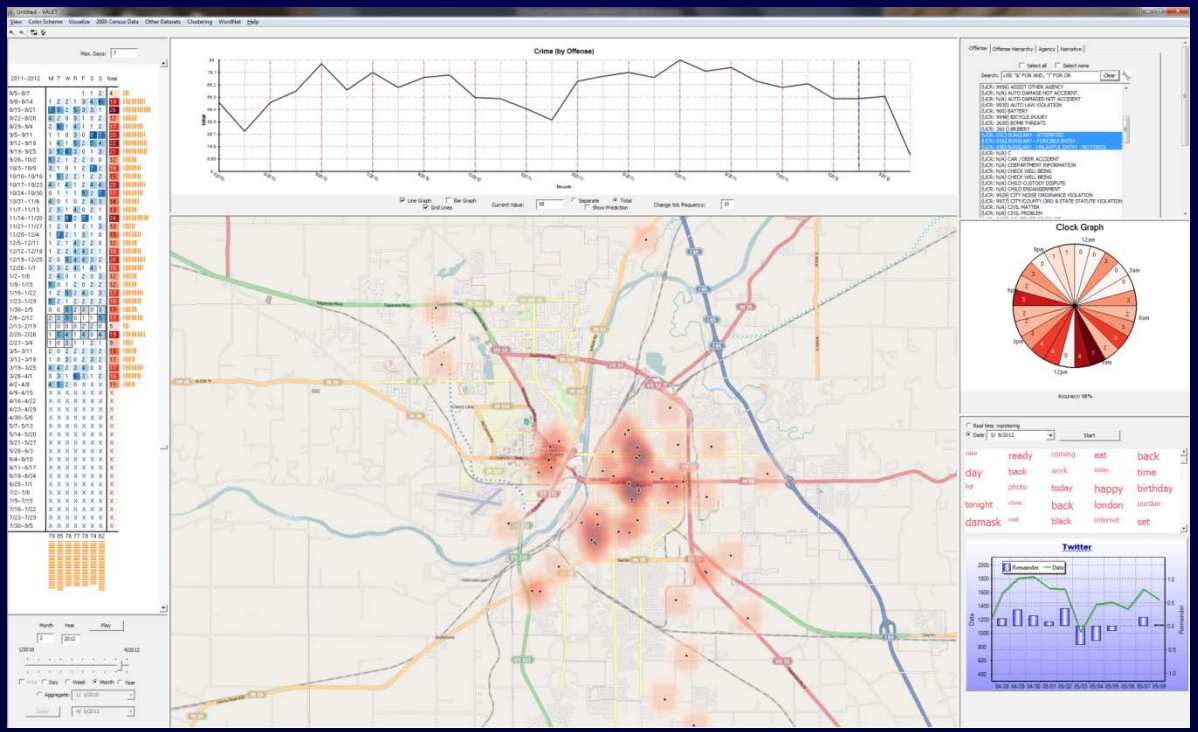
Visual Analytics Uses for Public Safety

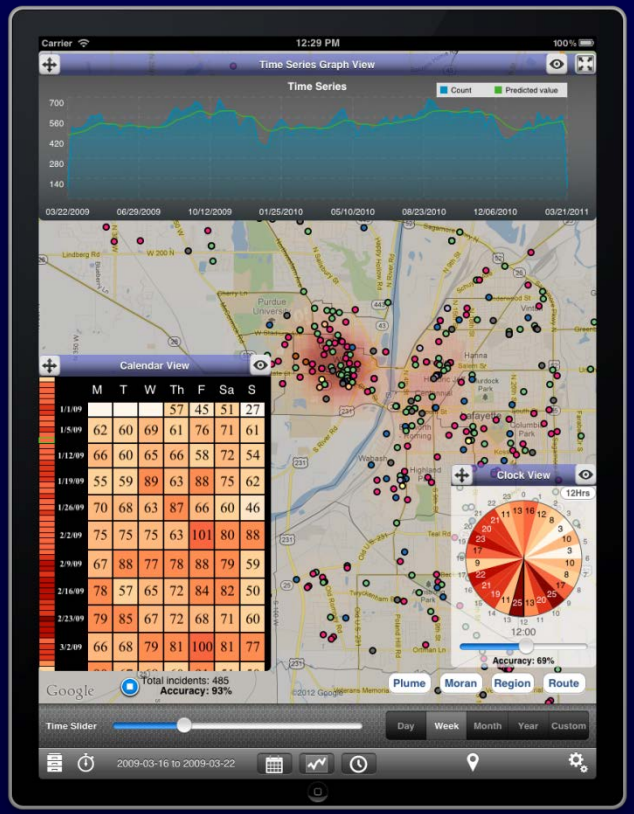


- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Multisource, multimedia massive data integration & analysis
- Purpose: Planning for resiliency, training, detection, investigation, response, recovery, remediation



Visual Analytics Law Enforcement Toolkit (i)VALET



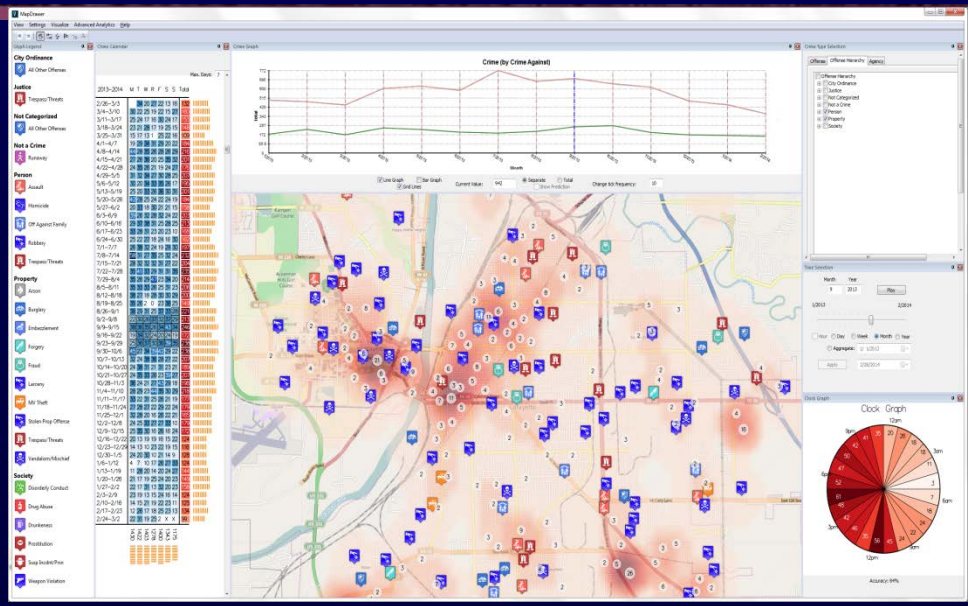


VACCINE

Visual Analytics Law Enforcement Toolkit (VALET, iVALET)

Impacts:

- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public (June 2014) for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

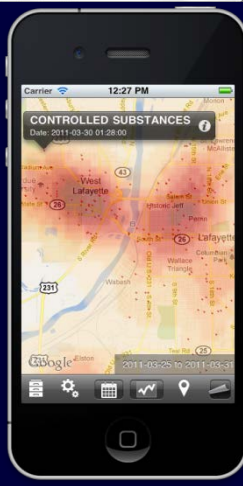


VALET delivered:

- Spring 2011: WL, Lafayette Police
- Fall 2013: NYPD

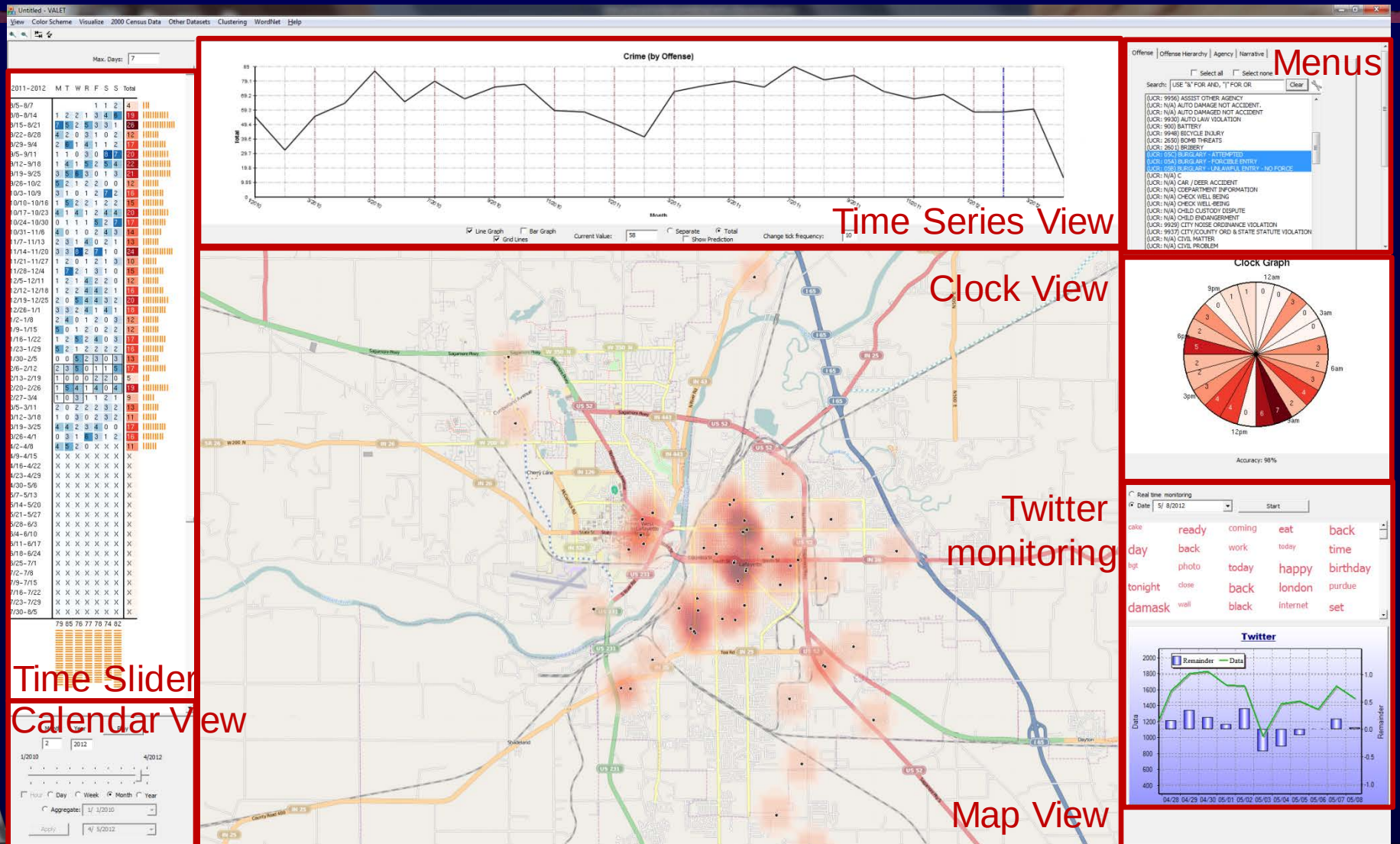
iVALET delivered:

- October 2011: Purdue, WL Police

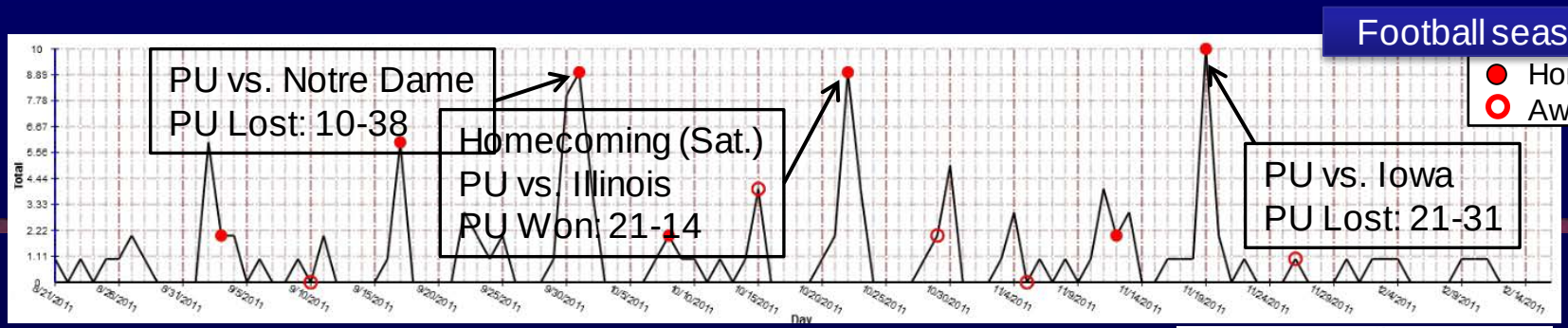


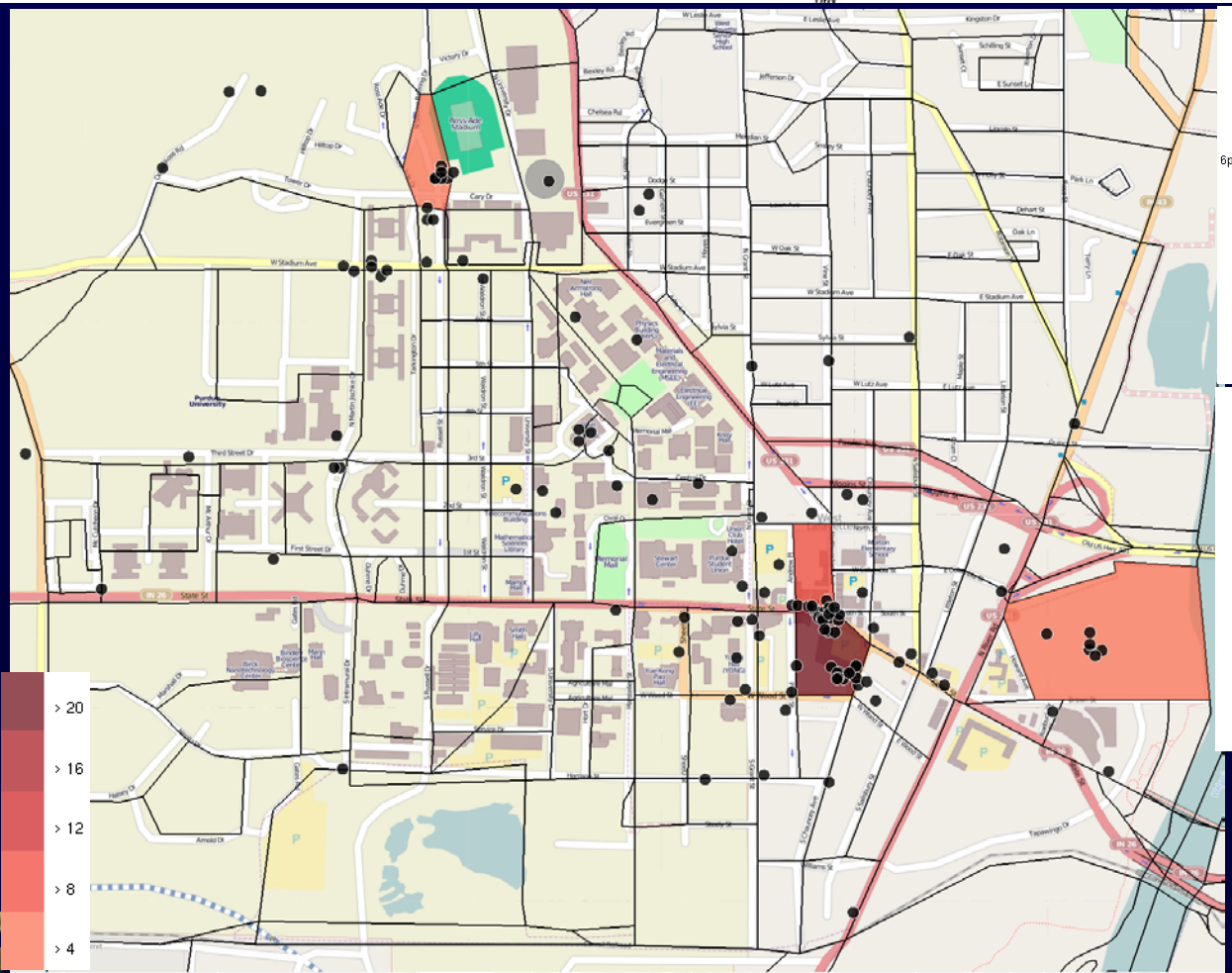
VACCINE

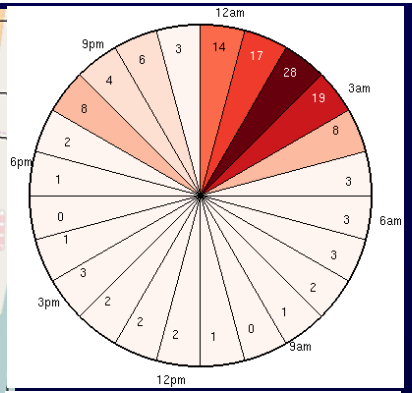
VALET Overview

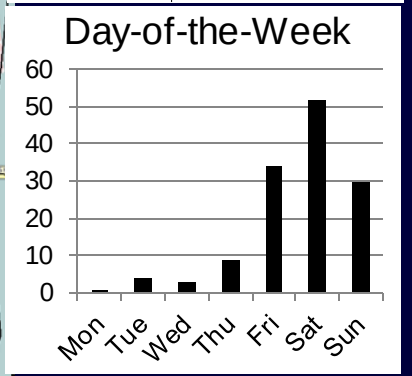


Example: Drunkenness / Public Intoxication



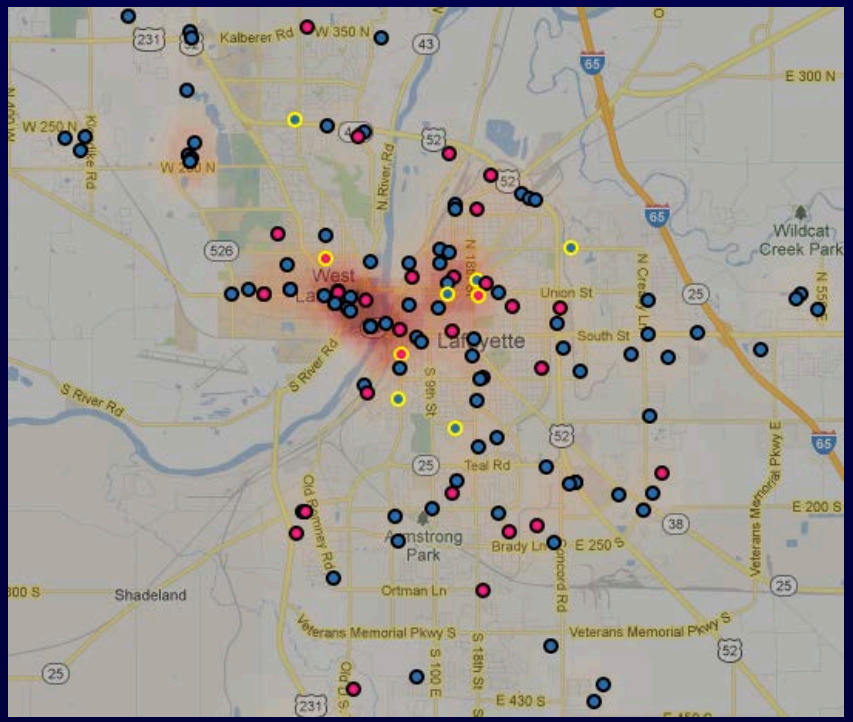




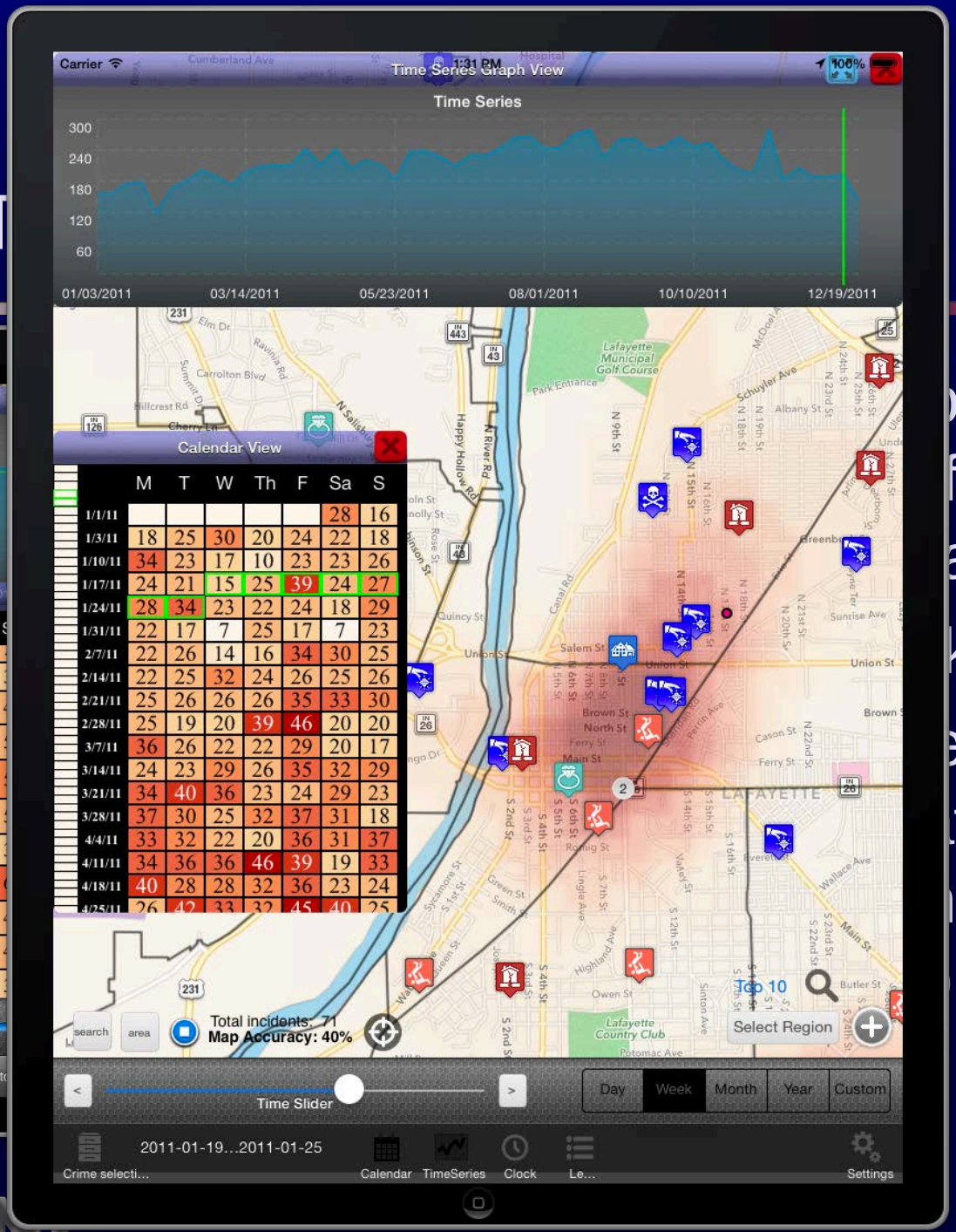


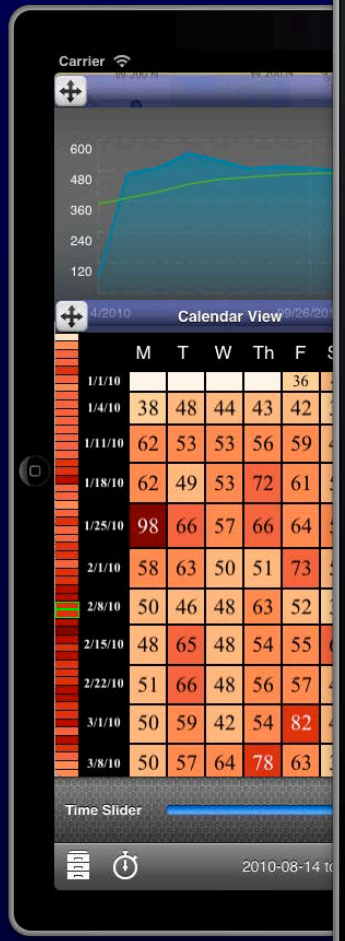
Top 10 Hot Incidents

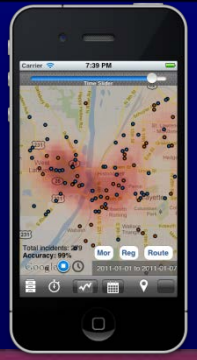
- Identify unusual localized high-frequency patterns of crimes (near repeats)
- Each data entry is checked for other crimes with similar properties within a 1 block radius of the incident location and a 14-day time period;
- Top 10 incidents with the most number of related incidents in this space-time window are highlighted

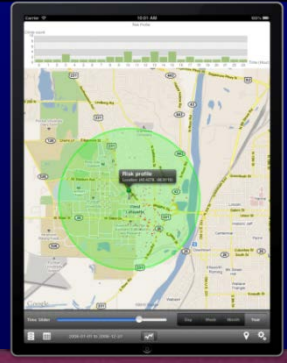


iVALET











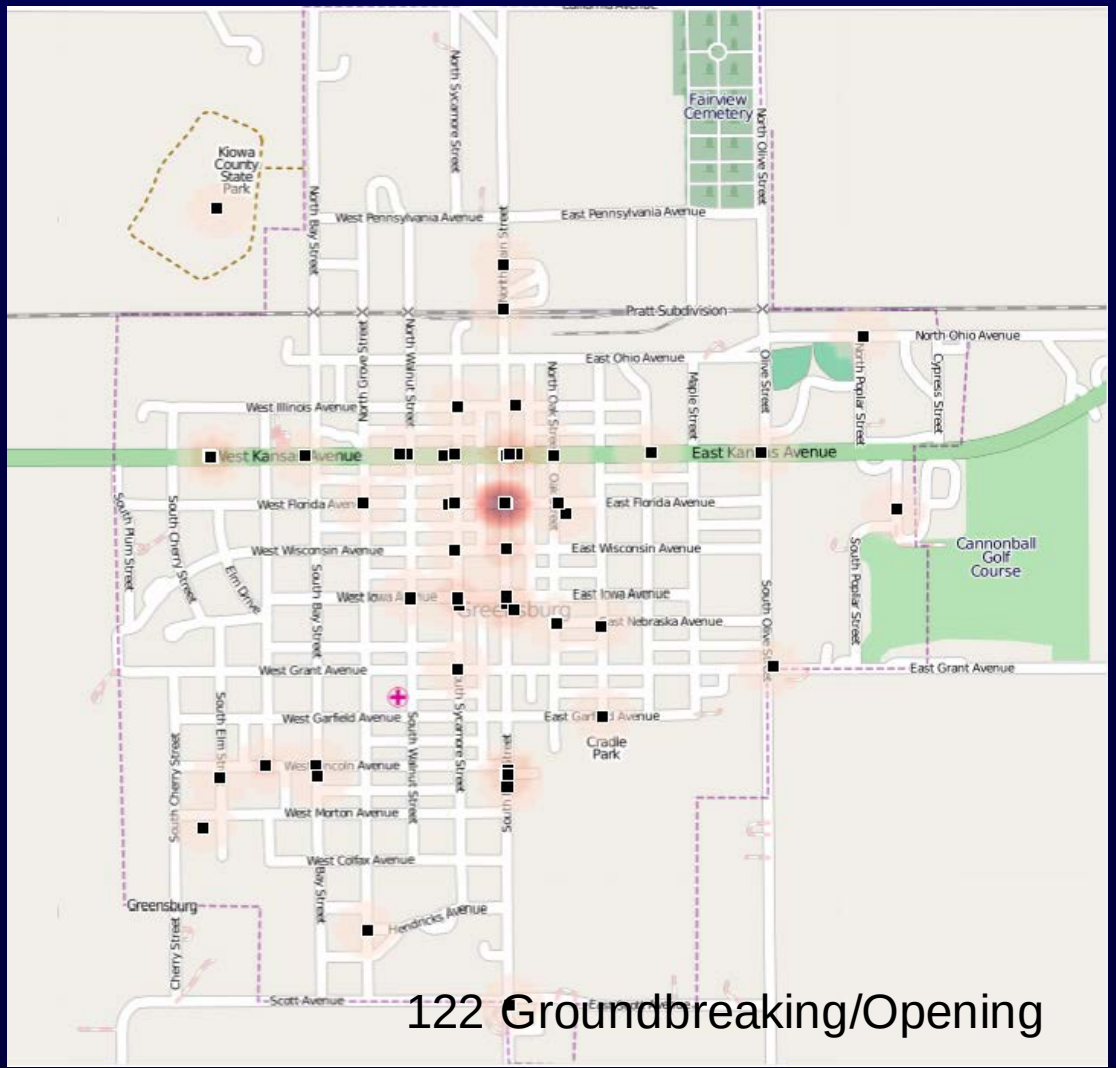
lore criminal,
fic and civil
a on-the-go
k assessment
e current
tial +
poral context
analysis

VACCINE

Educational Use of VALET: Greensburg, KS City Rebuilding (2007 EF5 Tornado)

95% of town levelled

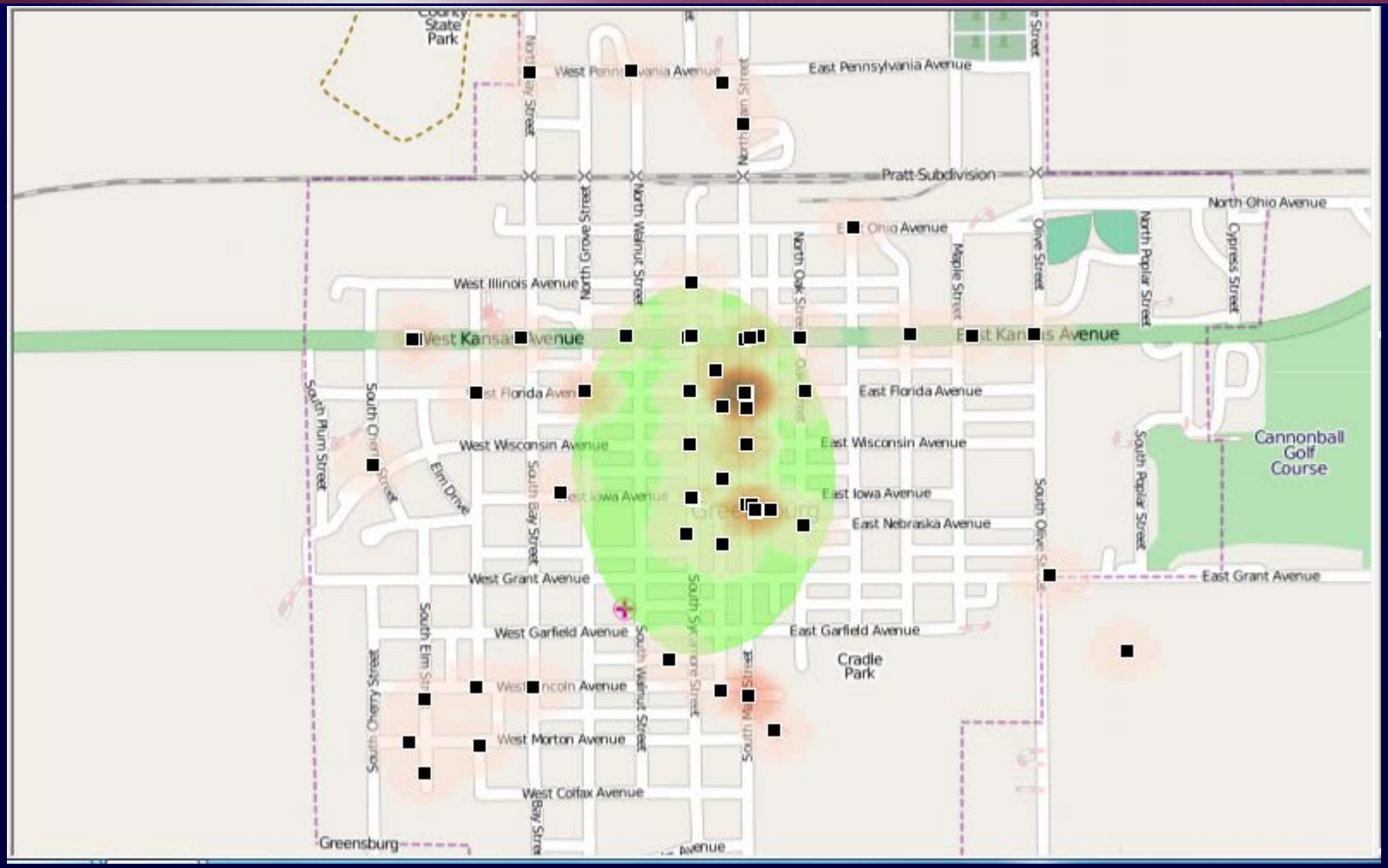
Groundbreaking data



122 Groundbreaking/Opening

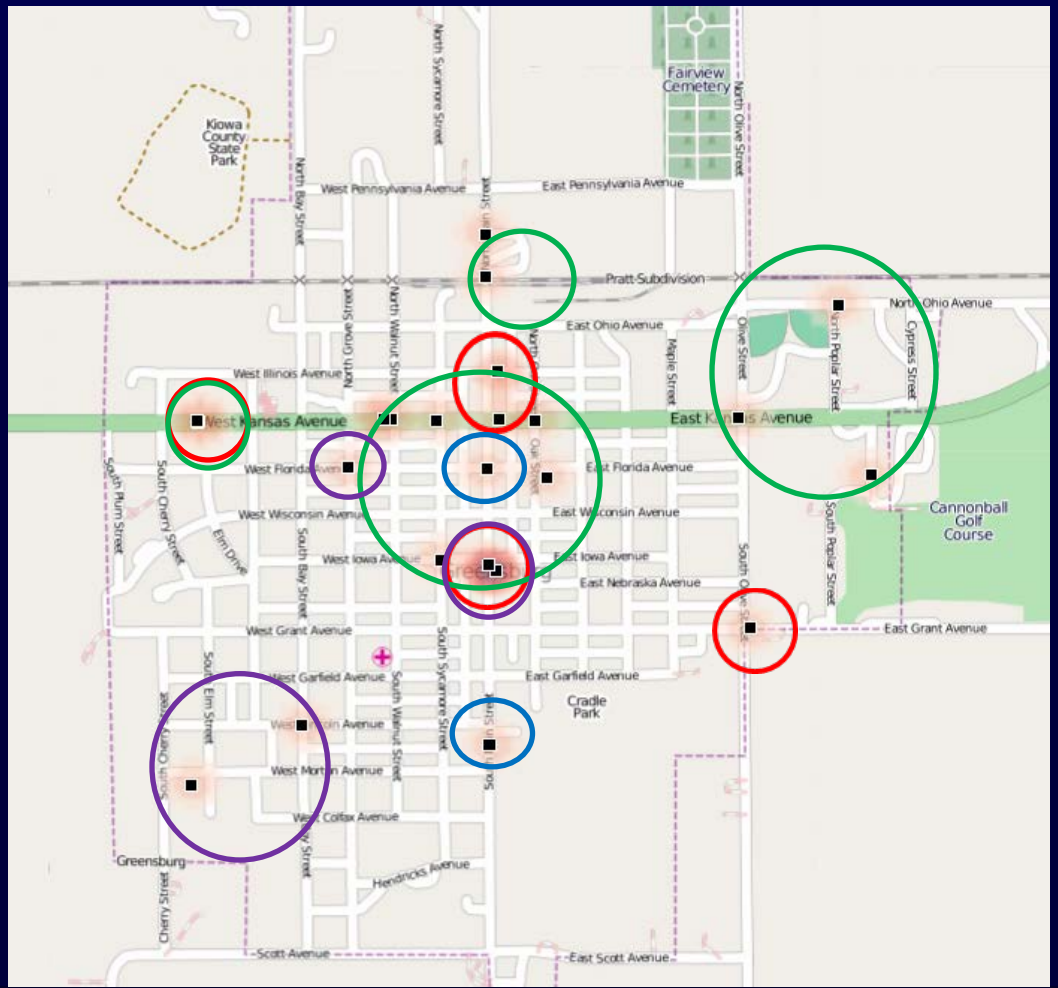
VACCINE

Statistical Analysis



2007

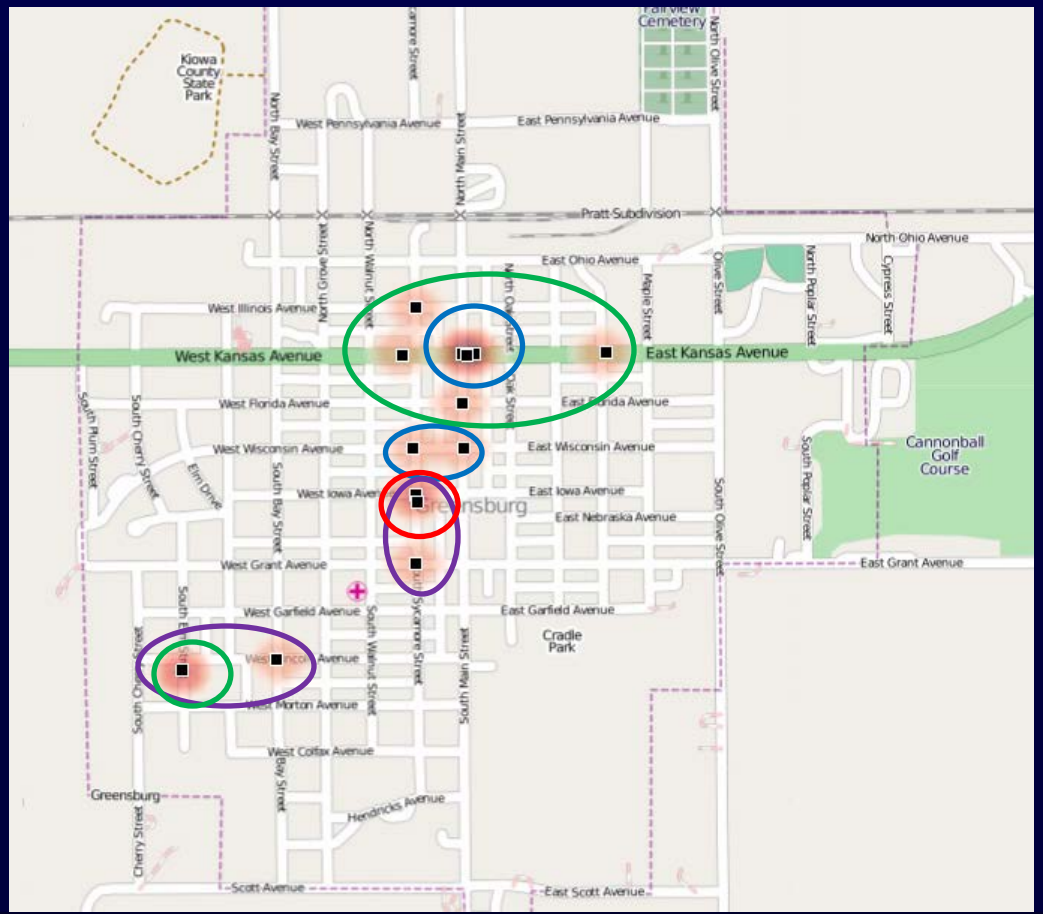
- Basic and Social Services
  - 7
- Business
  - 16
- Home
  - 7
- Third Space
  - 2



VACCINE

2008

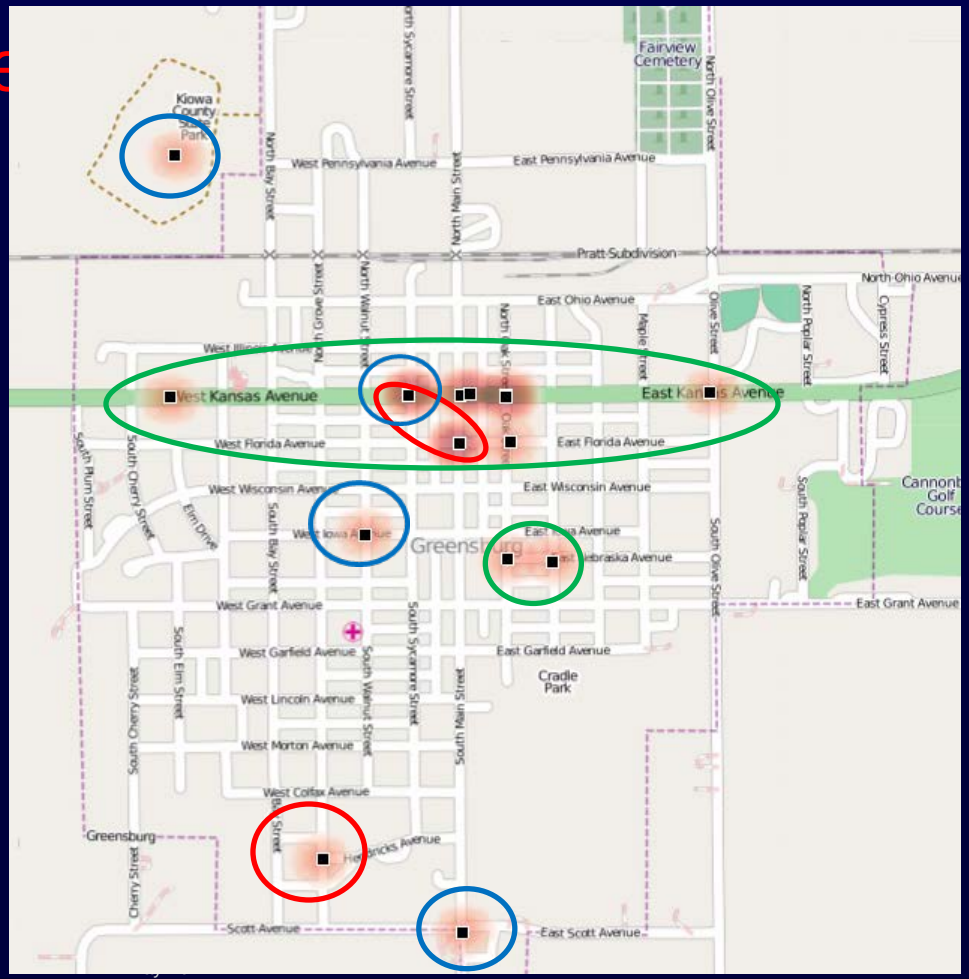
- Basic and Social Services
  - 1
- Business
  - 8
- Home
  - 3
- Third Space
  - 3



VACCINE

2009

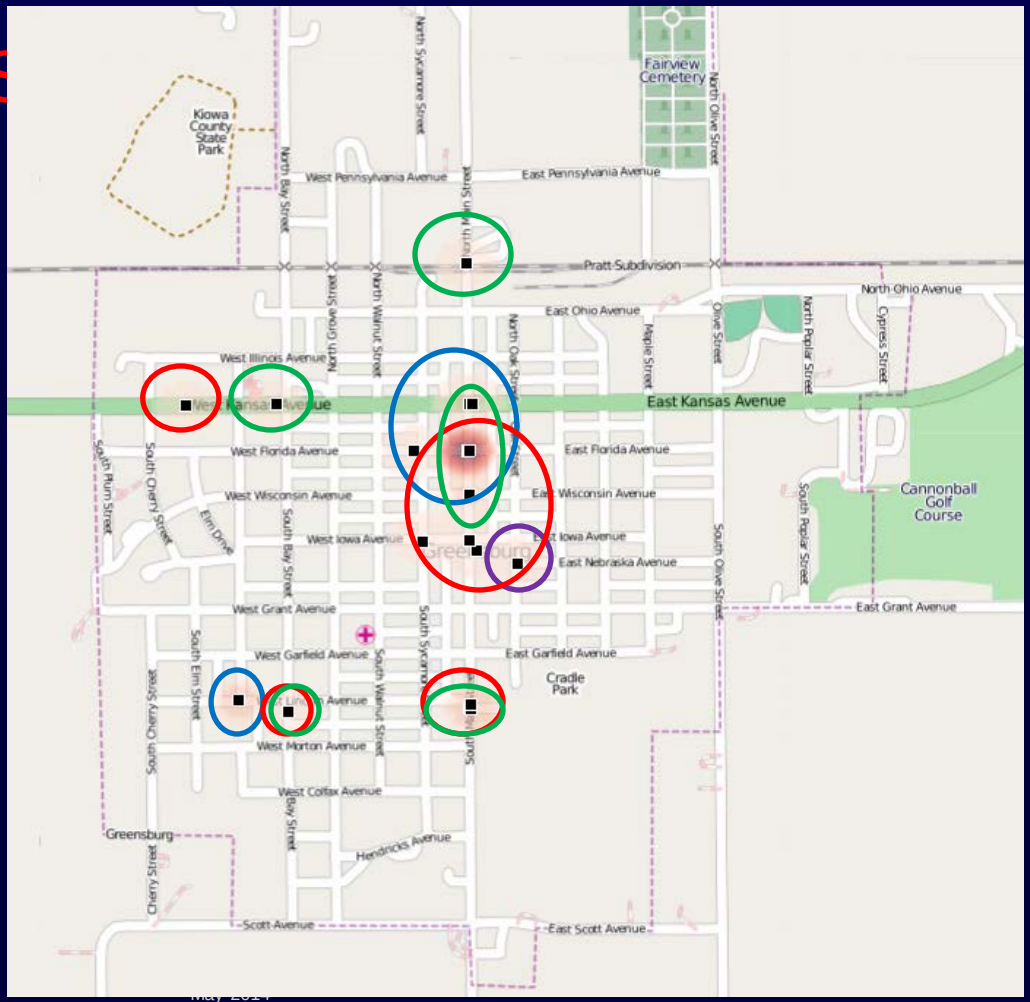
- Basic and Social Se
  - 3
- Business
  - 10
- Home
  - 0
- Third Space
  - 4



VACCINE

2010

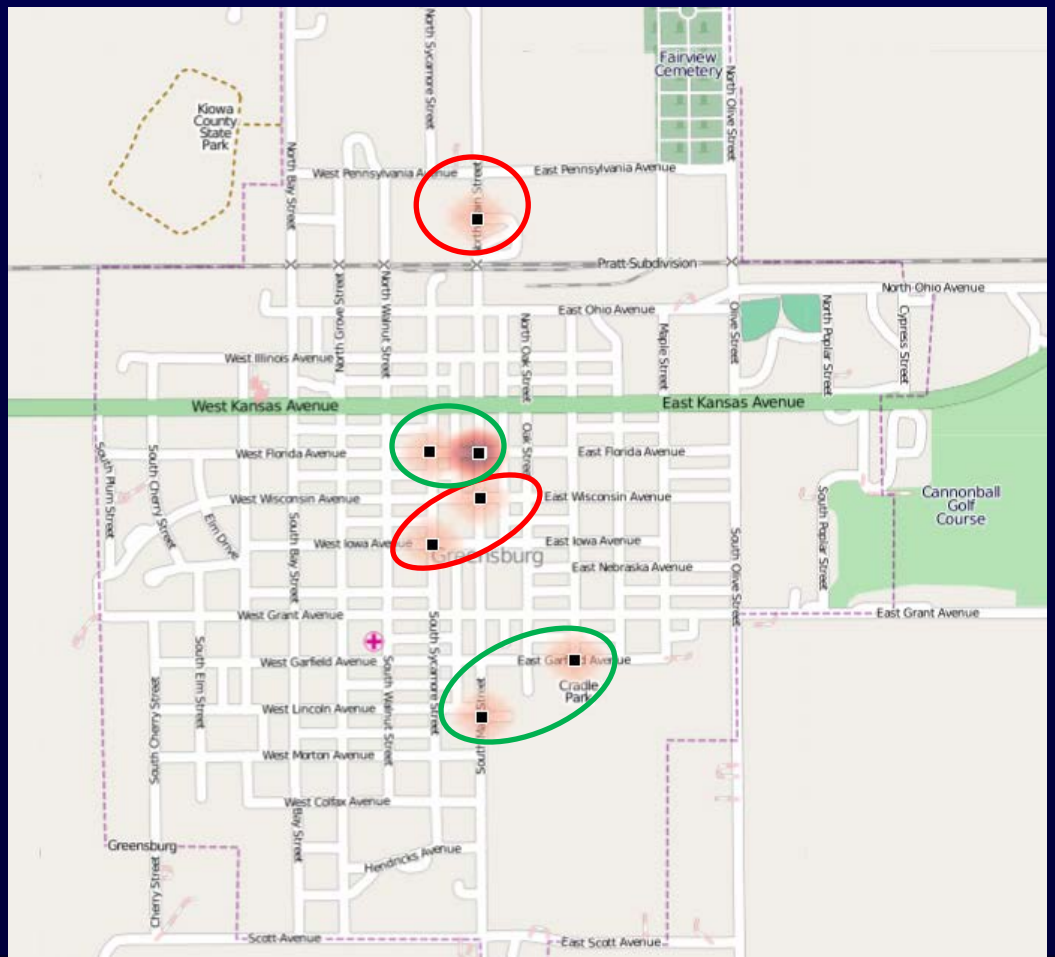
- Basic and Social S
  - 14
- Business
  - 26
- Home
  - 1
- Third Space
  - 5



VACCINE

2011

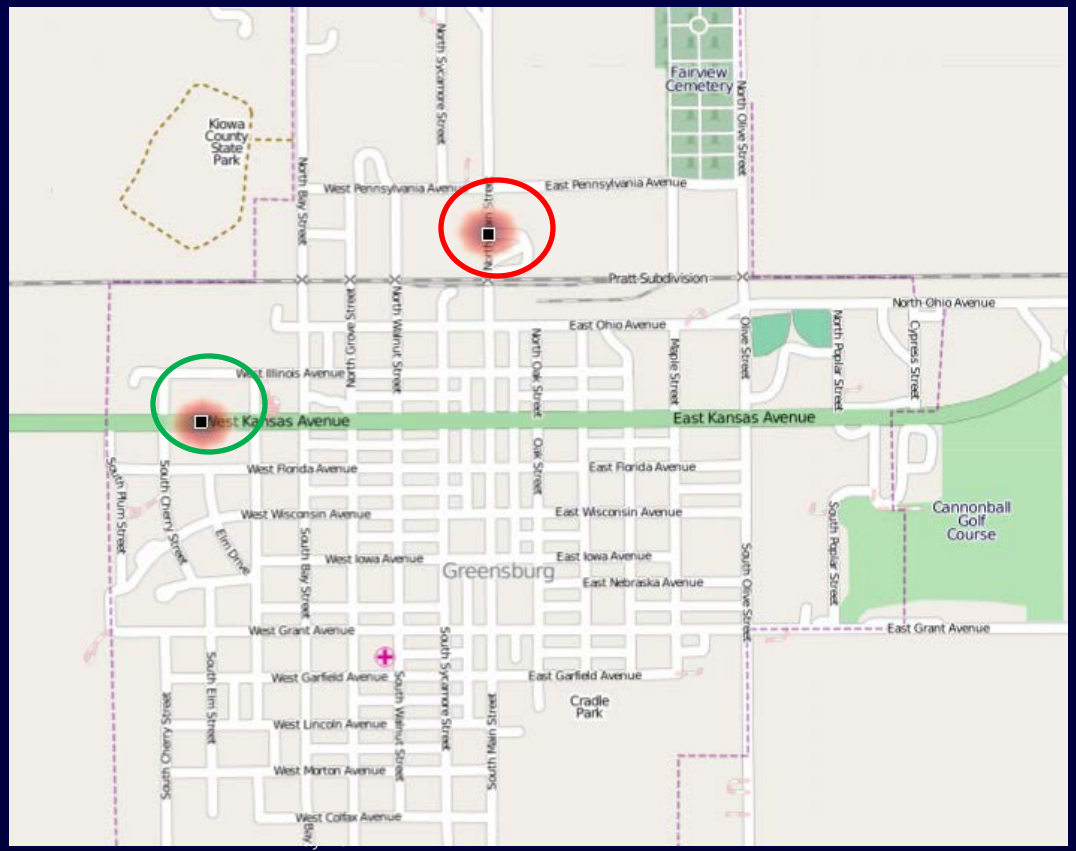
- Basic and Social Services
  - 4
- Business
  - 6
- Home
  - 0
- Third Space
  - 0



VACCINE

2012

- Basic and Social Services
  - 1
- Business
  - 1
- Home
  - 0
- Third Space
  - 0



VACCINE

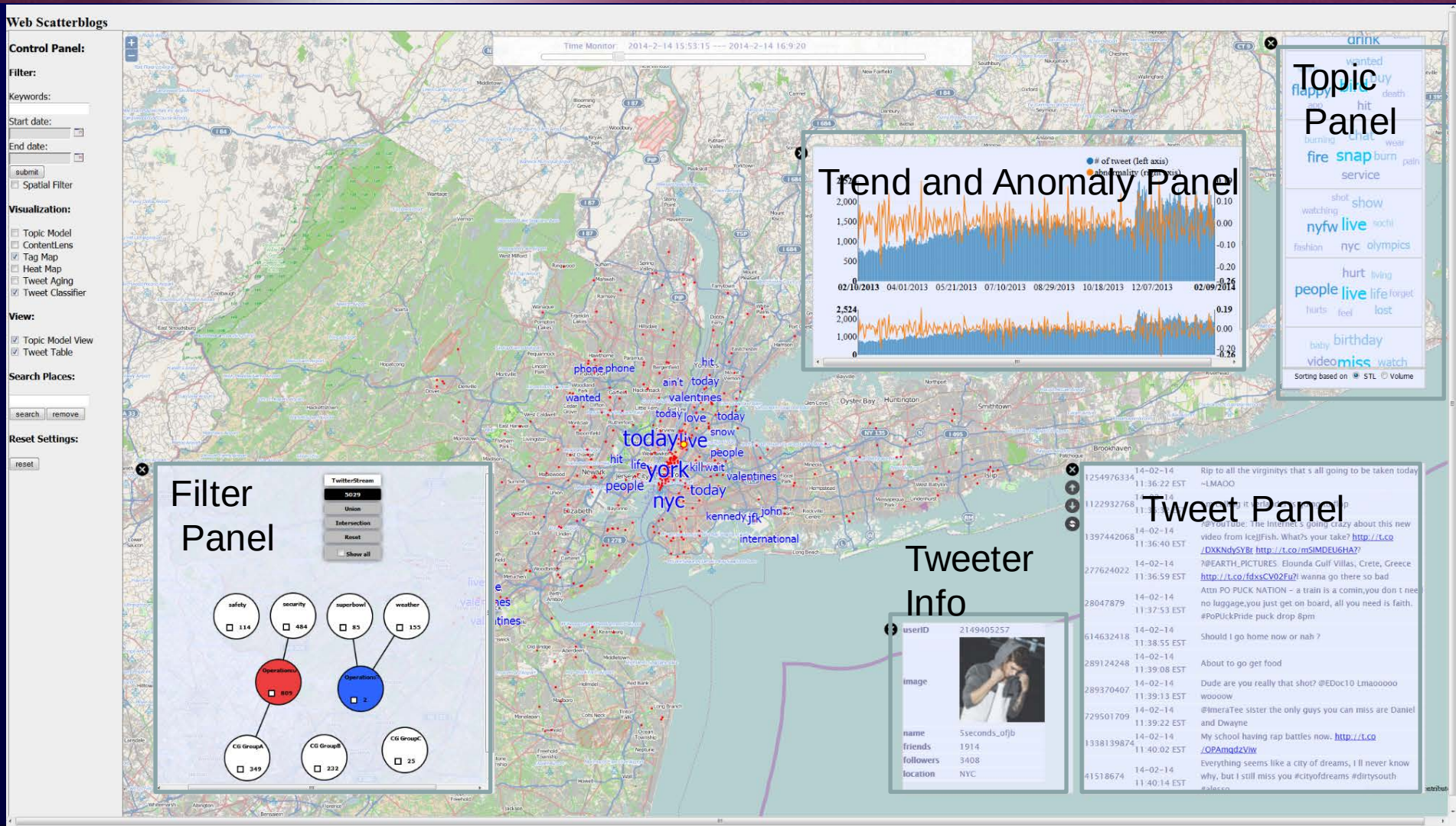
Social Media: Real-time Twitter Visual Analytics

(Purdue, Stuttgart, Penn St.)

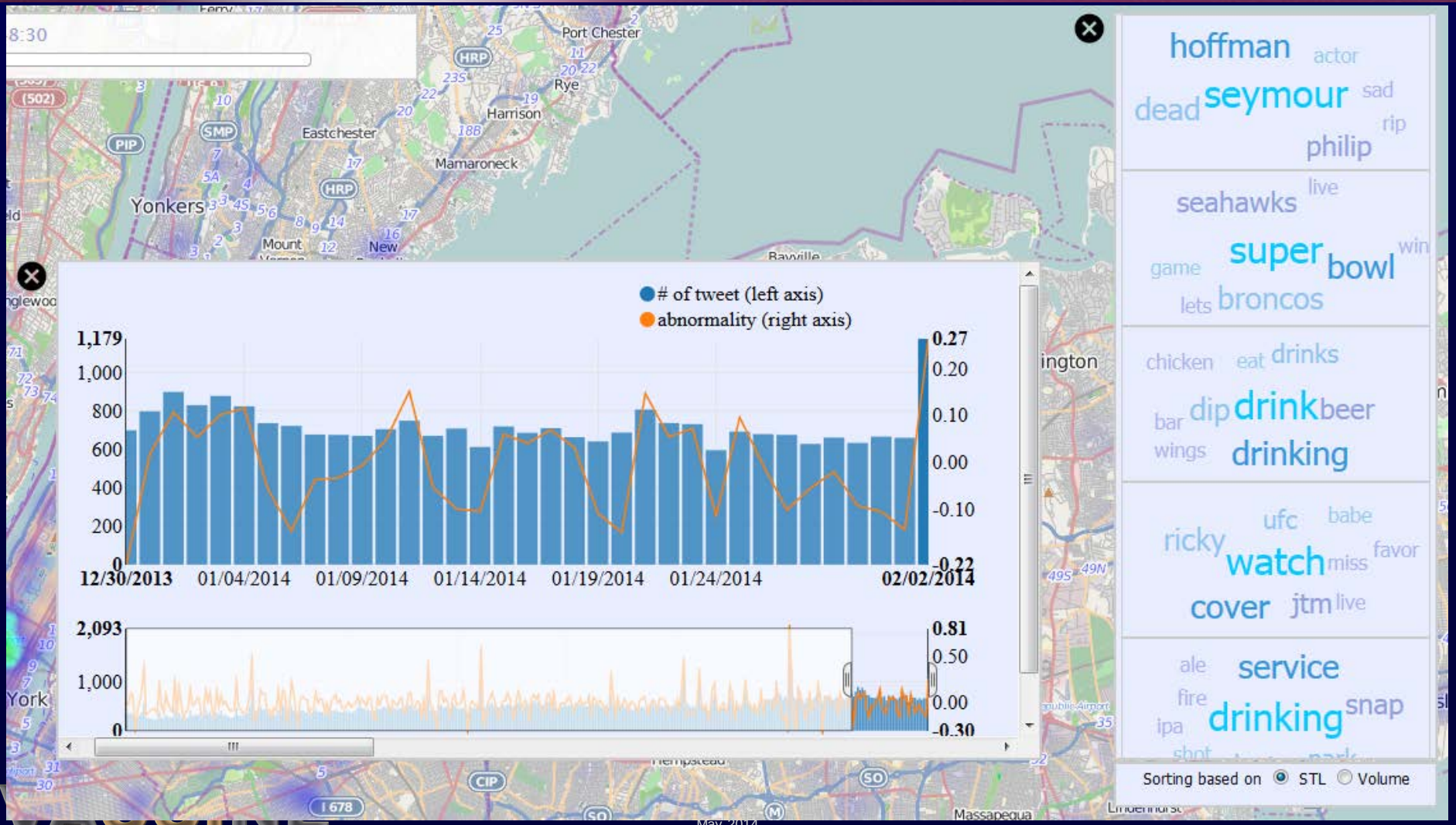
- Anomalous topic extraction using LDA and novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud



Social Media Visual Analytics: Scatterblogs

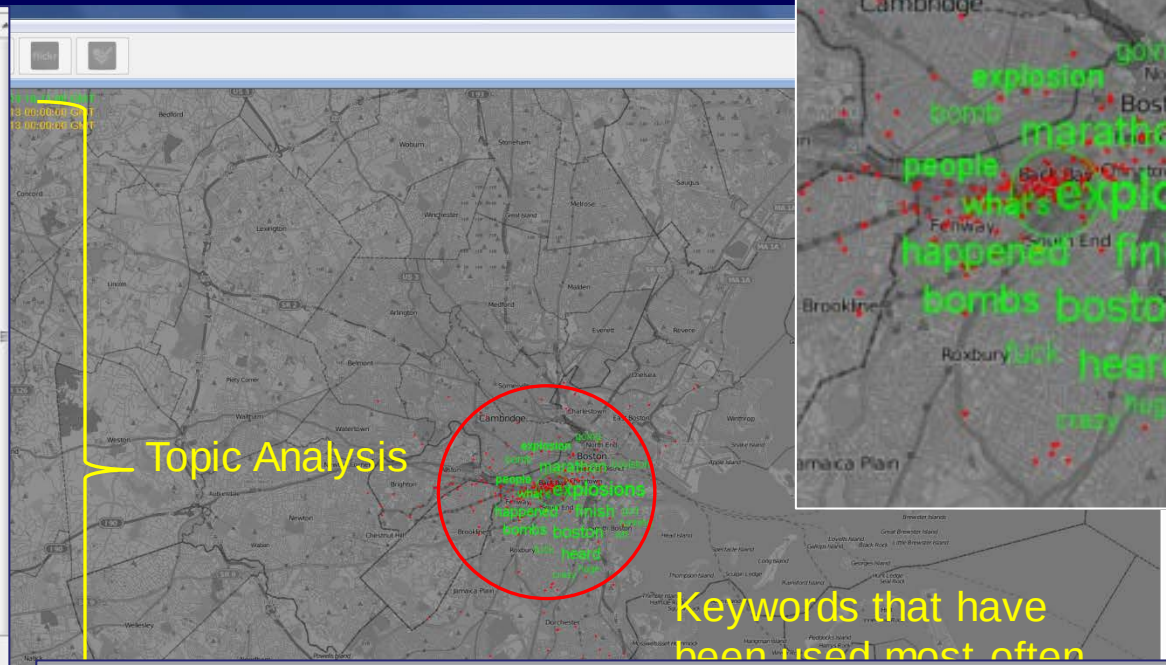


Anomaly Detection



Explosion Area in Boston





Keywords that have been used most often

Tweets

Text
Oh my god what just happened
Something happened at the end of the Boston Marathon. Something bad and there is a lot of chatter on Twitter. What's going on?
Multiple people are injured near the Boston Marathon finish line after two explosions. The #BostonMarathon has been stopped.
two bombs just went off on boylston
BREAKING NEWS: Two powerful explosions detonated in quick succession right next to the Boston Marathon finsh line this afternoon.
What the FUCK was that
Back in Sept, @croon1 solicited me for $2000. He now has a music video with William Shatner. If you watch it (god forbid) keep that on mind.
Literally what the fuck get me out of here
@DTenenbaum my office right next to it
@FRANCESCalciO I figured 3 people would get the joke
Two explosions just rocked the finish line of the Boston Marathon. Sirens galore. People running in fear. Wonder what happened.
This is crazy i seen that blow up #bostonmarathon
can someone tell me what that explosion was!? #boston #bostonmarathon

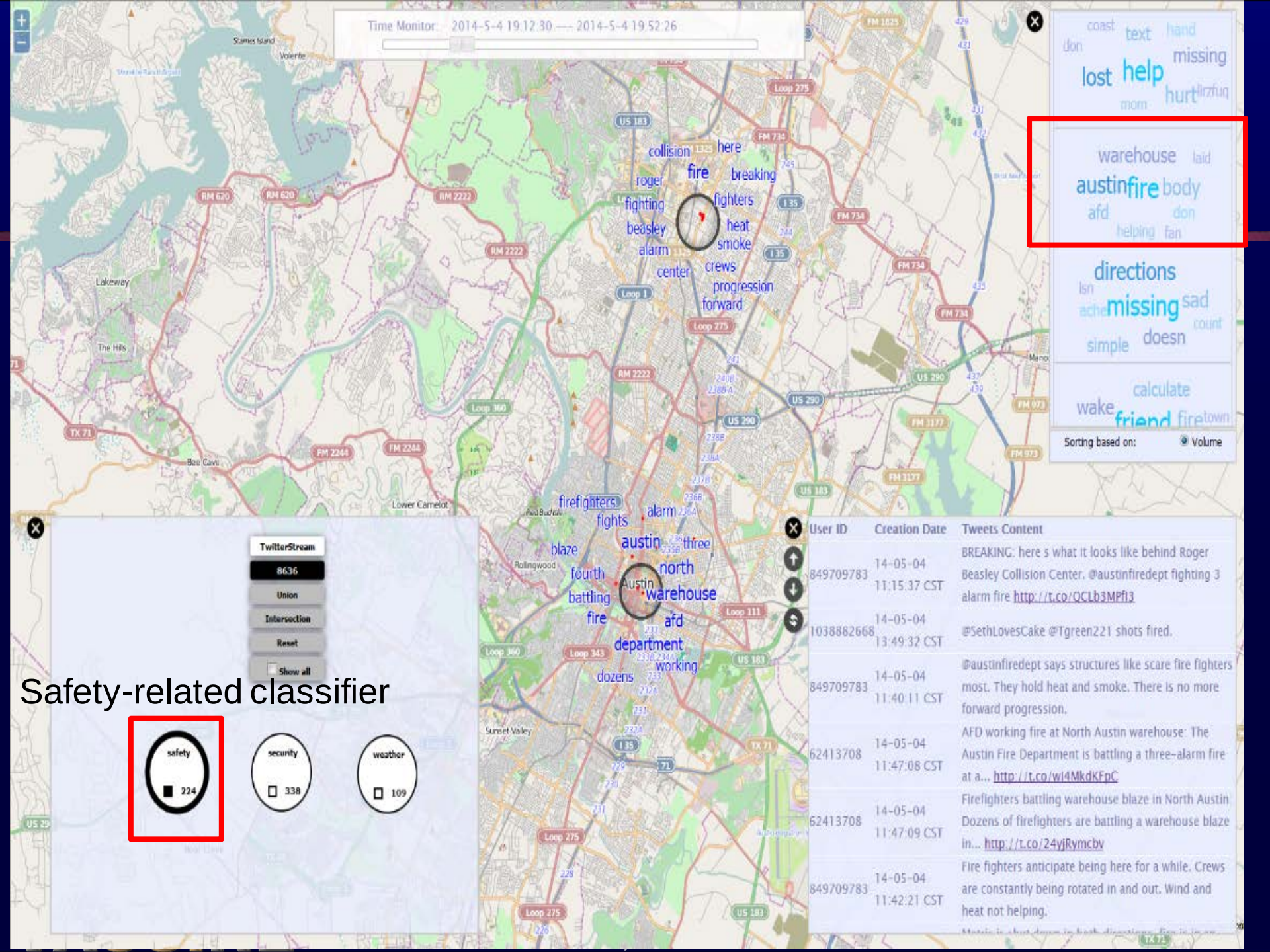
First Response (Tweet & Picture)



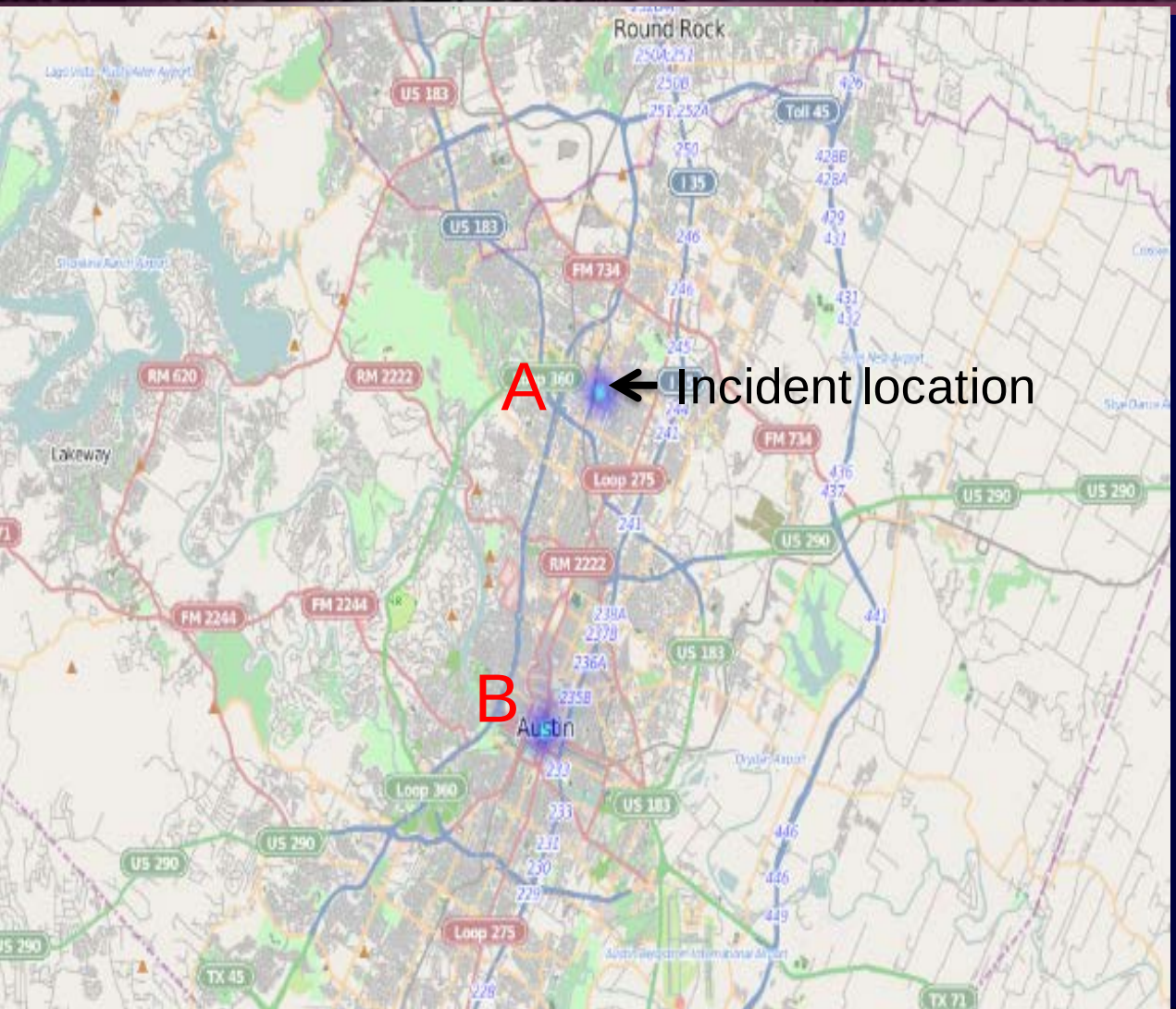
Austin, Tx Sunday Morning
May 4, 2014

Safety Classifier: Fire Event Detected

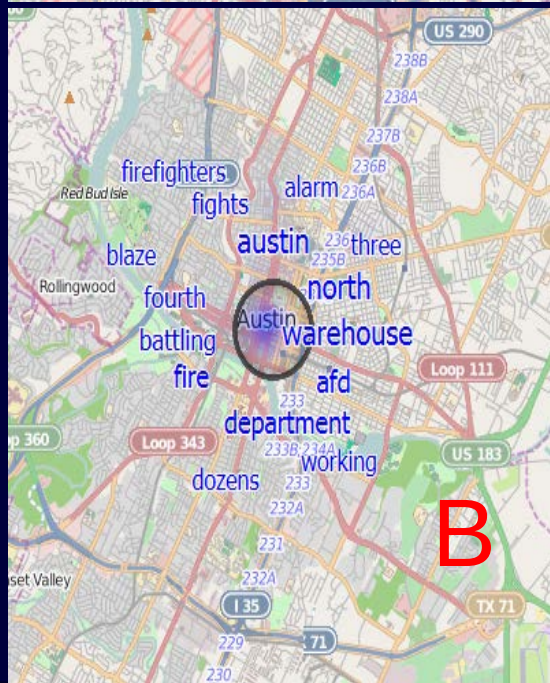
Location: Roger Beasley Collision Center


Time Monitor: 2014-5-4 19:12:30 — 2014-5-4 19:52:26
coast
text
hand
don
missing
lost
help
hurt
mom
warehouse
laid
austinfire
body
afd
don
helping
fan
directions
isn
missing
sad
count
simple
doesn
calculate
wake
friend
fire
town
Sorting based on:
Volume
collision
here
fire
breaking
roger
fighting
fighters
beasley
heat
alarm
smoke
center
crews
progression
forward
firefighters
fights
alarm
austin
three
blaze
north
fourth
Austin
warehouse
battling
fire
afd
department
working
dozens
TwitterStream
8636
Union
Intersection
Reset
Show all
Safety-related classifier
safety
224
security
338
weather
109
User ID
Creation Date
Tweets Content
849709783
14-05-04 11:15:37 CST
BREAKING: here s what it looks like behind Roger Beasley Collision Center. @austinfiredept fighting 3 alarm fire http://t.co/QCLb3MPfI3
1038882668
14-05-04 13:49:32 CST
@SethLovesCake @Tgreen221 shots fired.
849709783
14-05-04 11:40:11 CST
@austinfiredept says structures like scare fire fighters most. They hold heat and smoke. There is no more forward progression.
62413708
14-05-04 11:47:08 CST
AFD working fire at North Austin warehouse: The Austin Fire Department is battling a three-alarm fire at a... http://t.co/wI4MkdKFpC
62413708
14-05-04 11:47:09 CST
Firefighters battling warehouse blaze in North Austin Dozens of firefighters are battling a warehouse blaze in... http://t.co/24yjRymcbv
849709783
14-05-04 11:42:21 CST
Fire fighters anticipate being here for a while. Crews are constantly being rotated in and out. Wind and heat not helping.

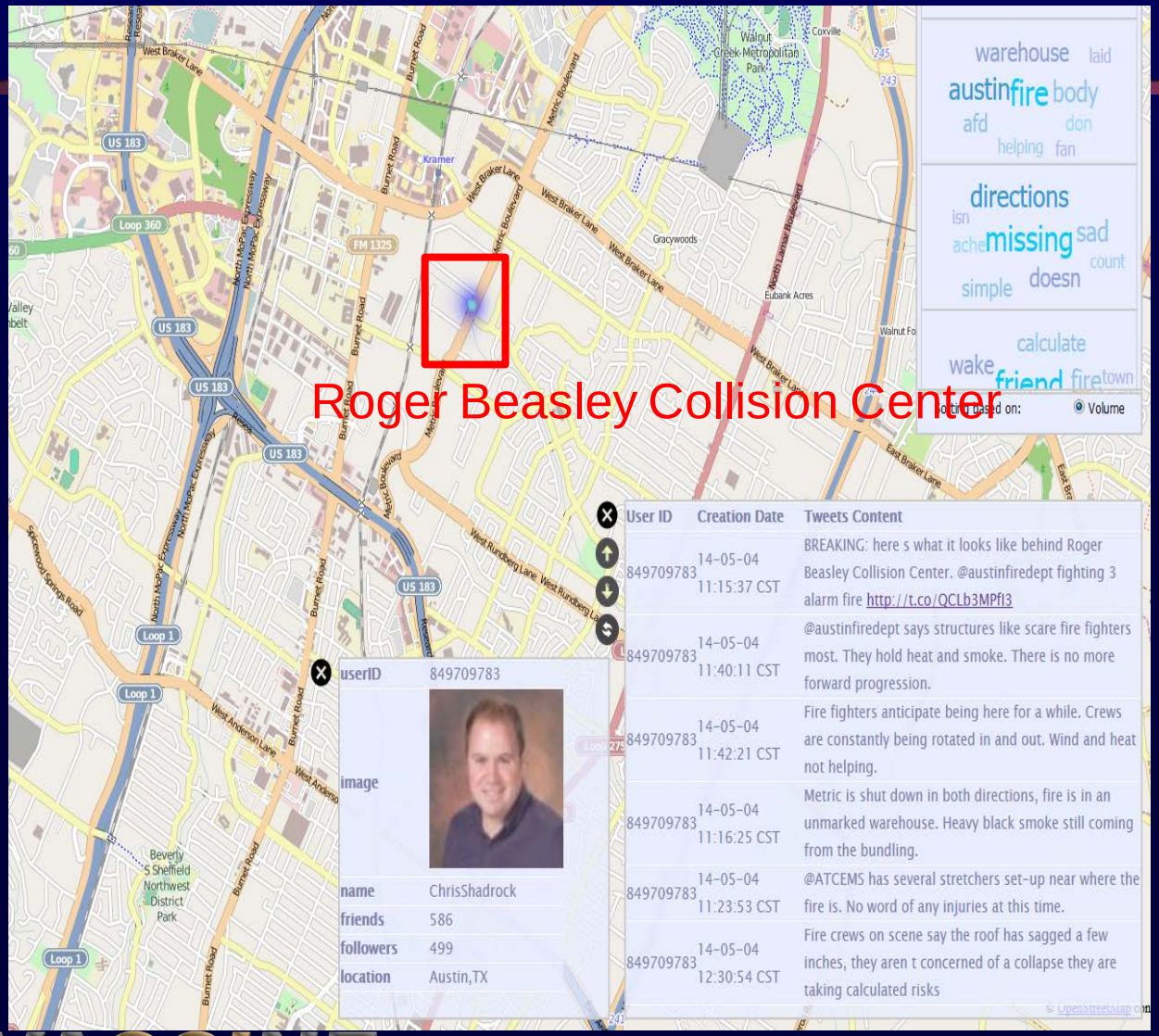
Spatial Distribution (Heat Map) and Keywords (Content Lens) of Fire-related Tweets





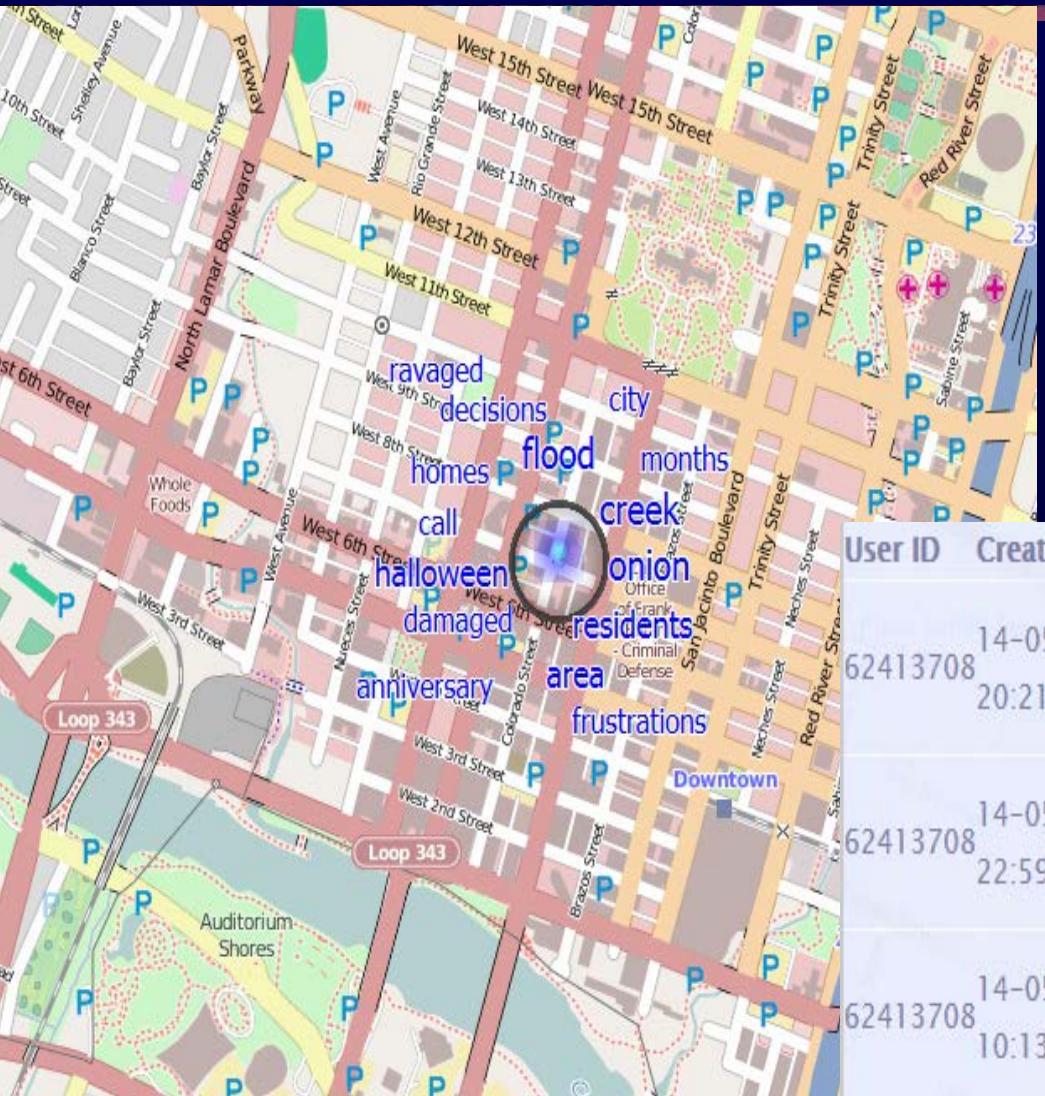


A Local Journalist Kept Updating The Event On The Spot



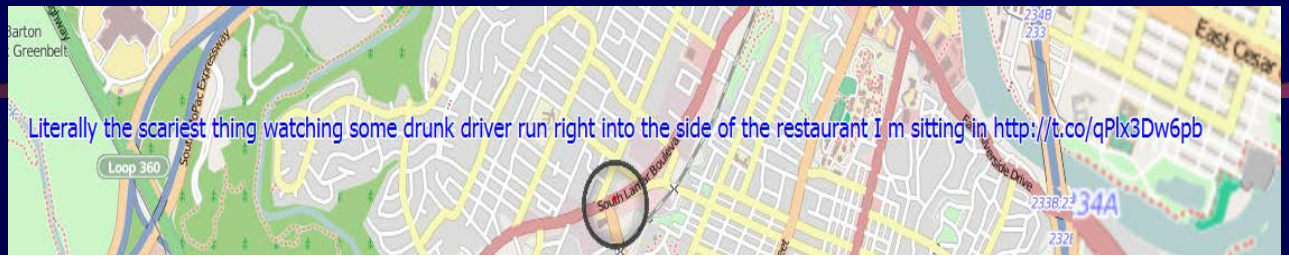


Also Detected This Weekend: Halloween Flood Related Tweets



User ID	Creation Date	Tweets Content
62413708	14-05-03 20:21:25 CST	Onion Creek residents call for city to make decisions on flood-damaged homes: The frustrations of Onion Creek... http://t.co/gPADzk6ppd
62413708	14-05-03 22:59:04 CST	6 month anniversary of the Halloween floods: It s been six months since flood waters ravaged the Onion Creek area... http://t.co/BEIGDV2ODY
62413708	14-05-04 10:13:05 CST	South Austin post-flood needs survey: On Saturday, residents in the Onion Creek area hit hard by the Halloween... http://t.co/xwMQ8PhVHl

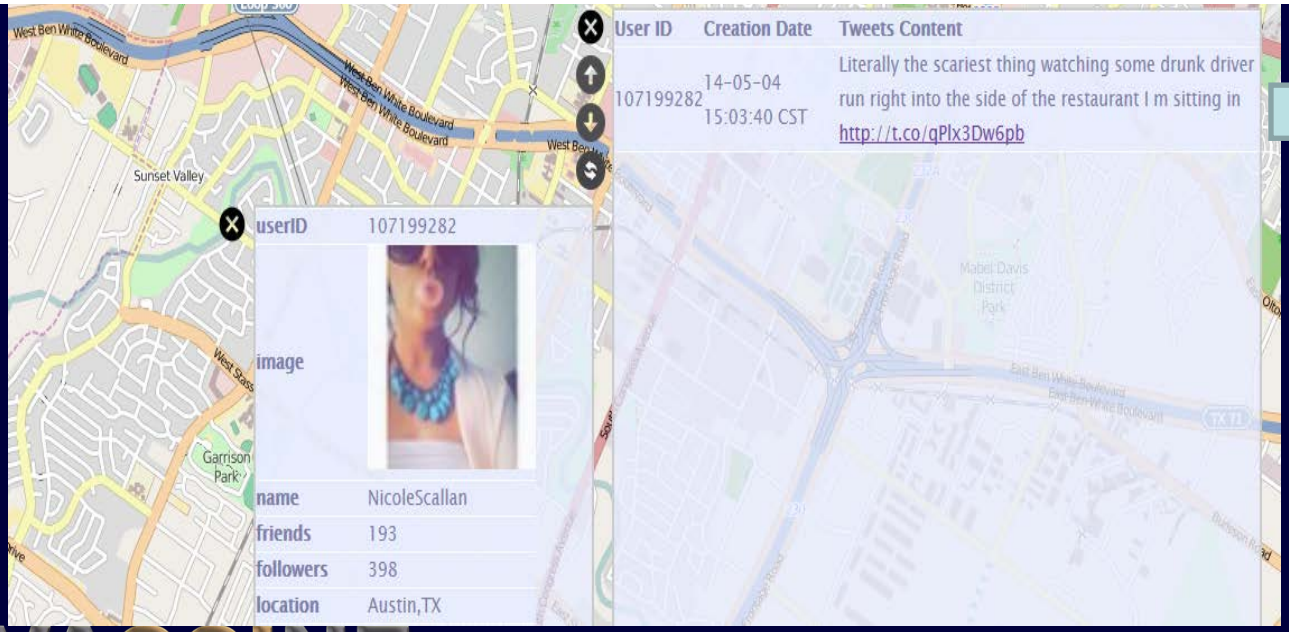
Also Detected This Weekend: Drunk Driving Accident

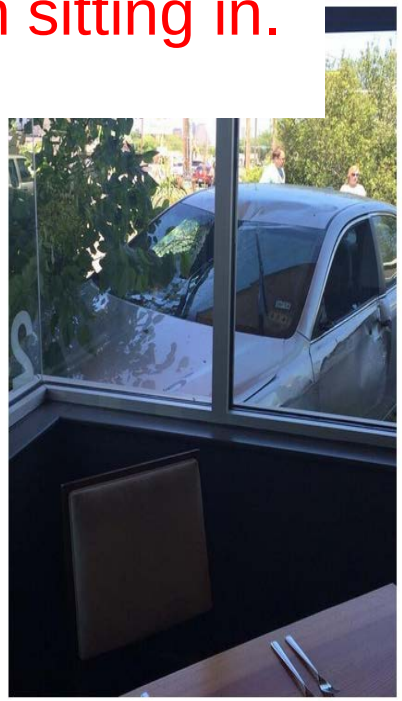


Nicole Scallan
@NicoleScallan

Literally the scariest thing watching some drunk driver run right into the side of the restaurant I'm sitting in

Tweet: Literally the scariest thing watching some drunk driver run right into the side of the restaurant I'm sitting in.





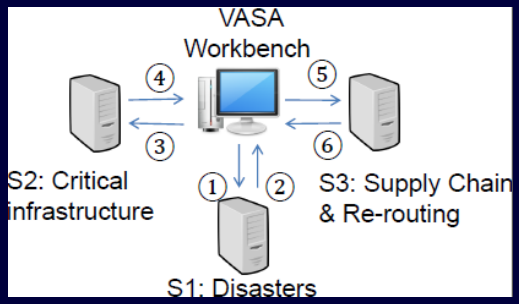
Cascading Critical Infrastructure Resiliency Modeling and Analytics (VASA)



- Purpose: Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.
- Collaborating Institution(s): Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Texas (TACC), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities
- End-User(s): Power Suppliers (e.g., Duke Energy), Cyber Community (e.g., Cisco), Quick Service Restaurants and suppliers, food supply

Purdue

UNCC

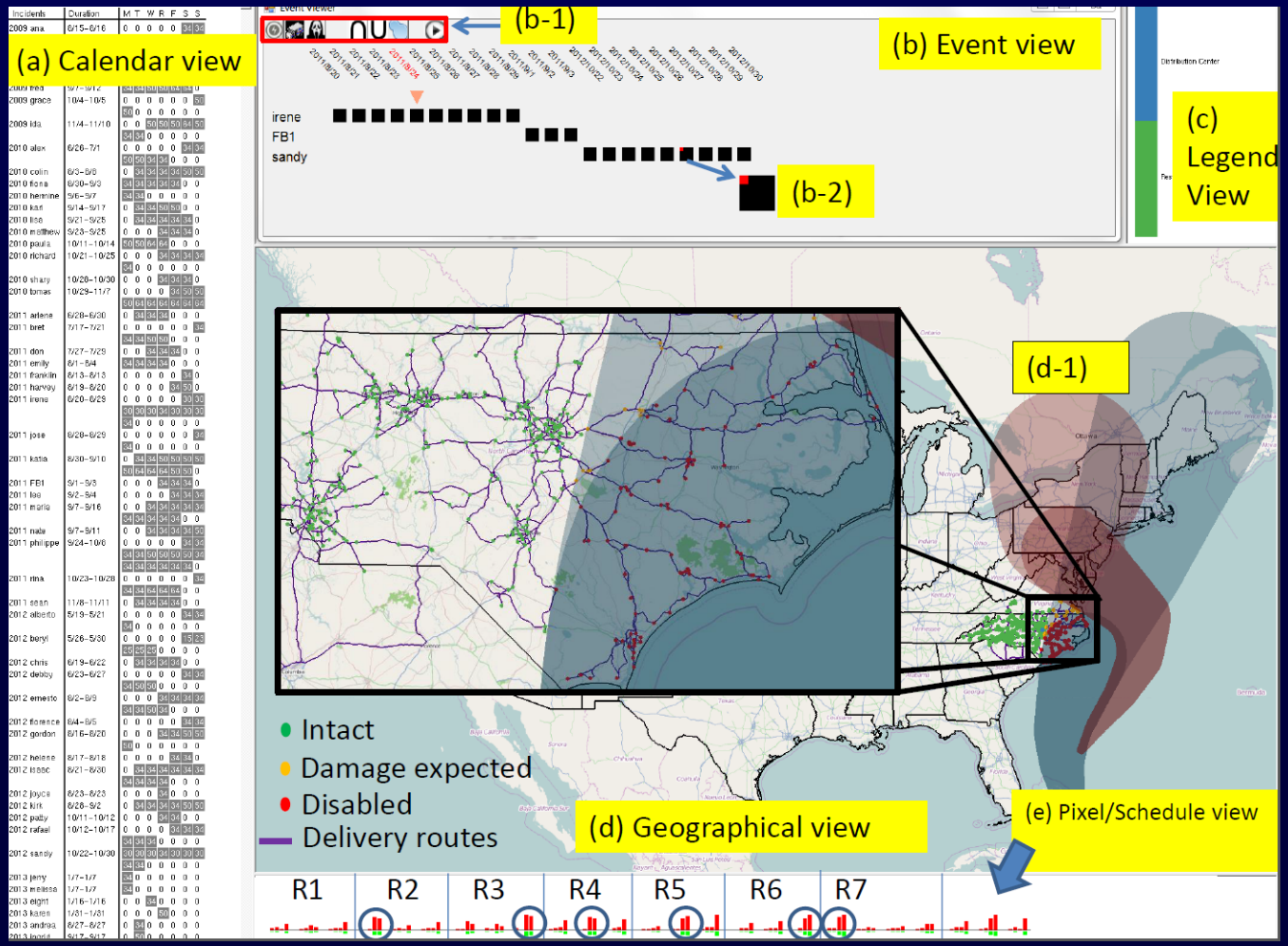


U Minn

TACC

VACCINE

System Interface



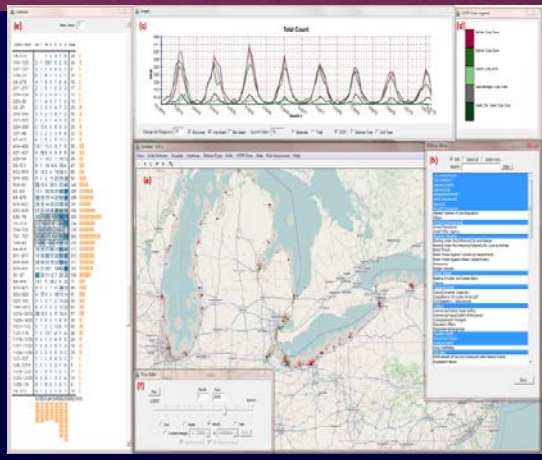
Effective Risk-based Decision Making and Resource Allocation Visual Analytics

- Evaluate current and historical mission area:
  - Demands
  - Risks (total, mitigated, residual)
  - Resource allocation
  - Return on investment
- Evaluate courses of action
- Evaluate above at both Strategic and Tactical/Operational level



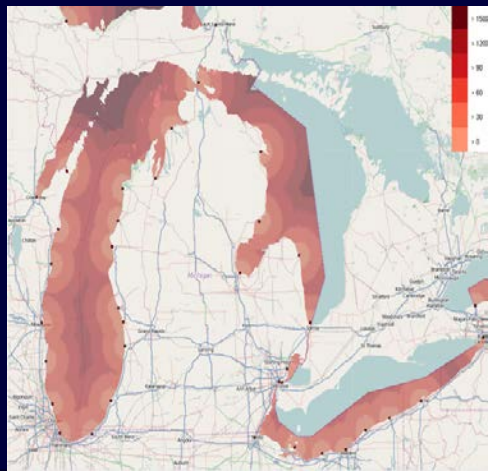
U.S. Coast Guard Search and Rescue VA (cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771



IMPACTS:

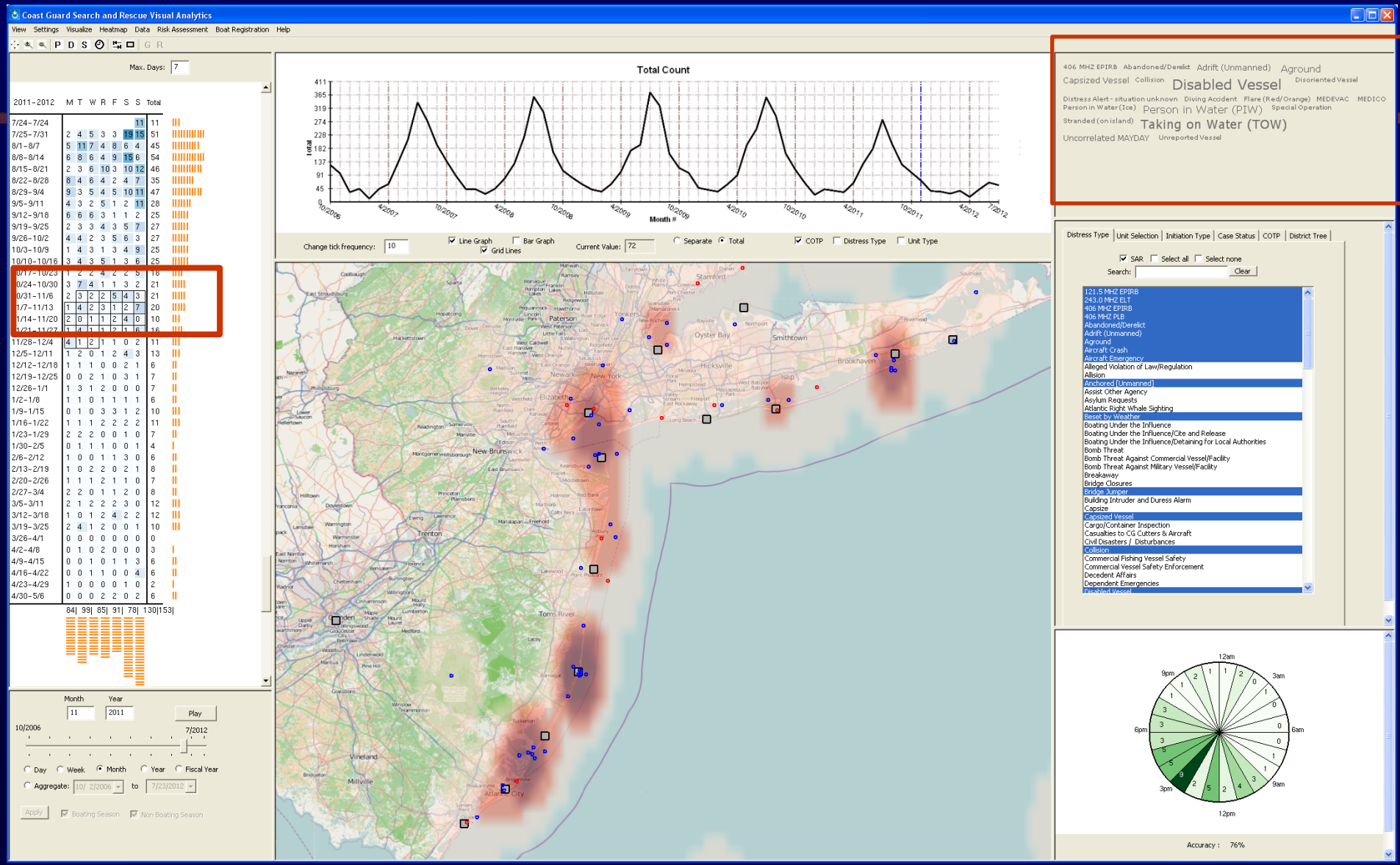
- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Hurricanes Sandy and Irene resource allocation decisions based on cgSARVA analysis and visualization
- Informed Commandant's budget testimony to Congress
- Key component of USCG D9 reallocation plan for 2011-12
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ



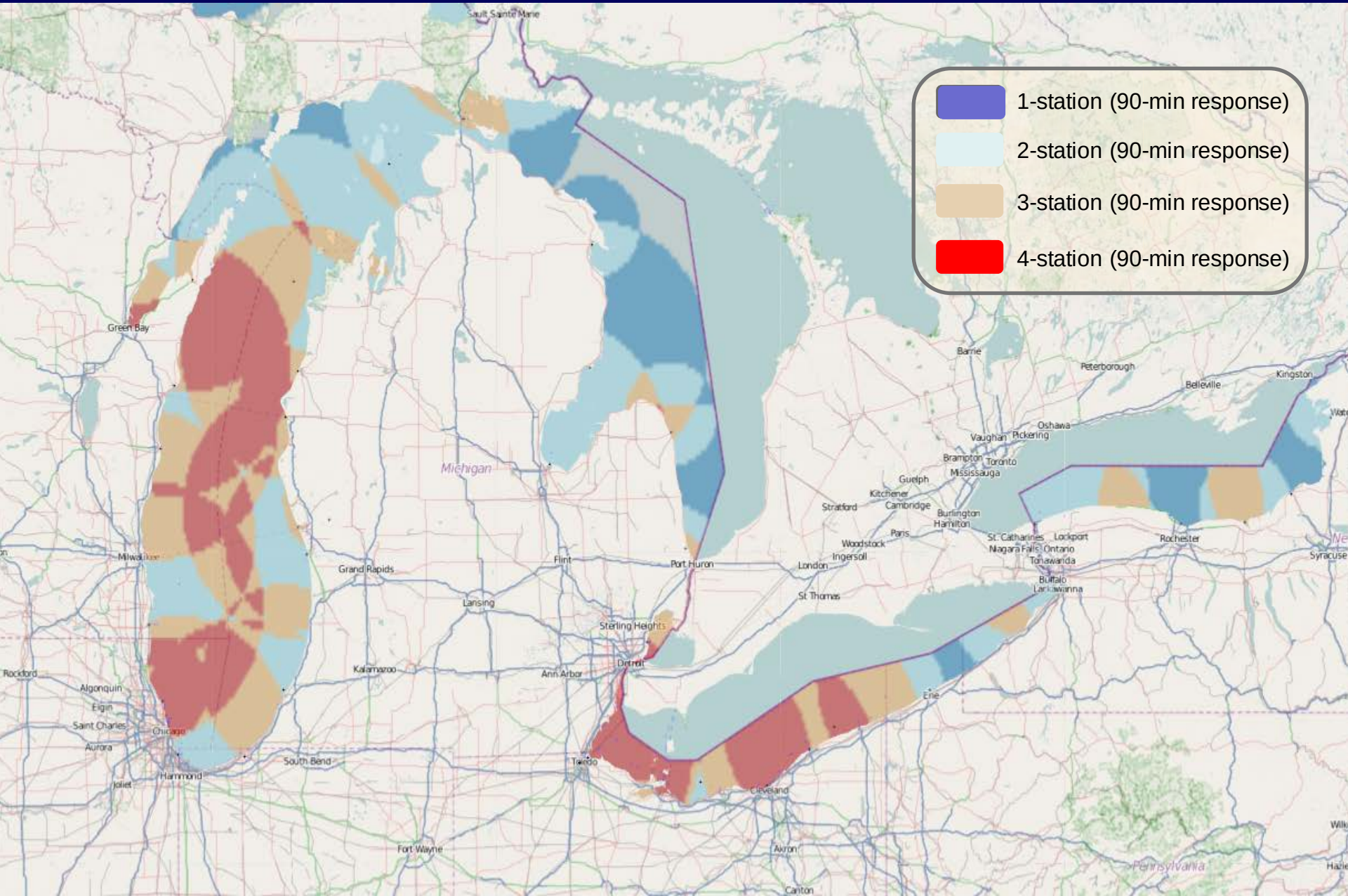
VACCINE

Example: Risks and Consequences From Sandy:

SAR Cases November 2011 NJ/NYC Area



Response Efficiency – Potential Future Assets

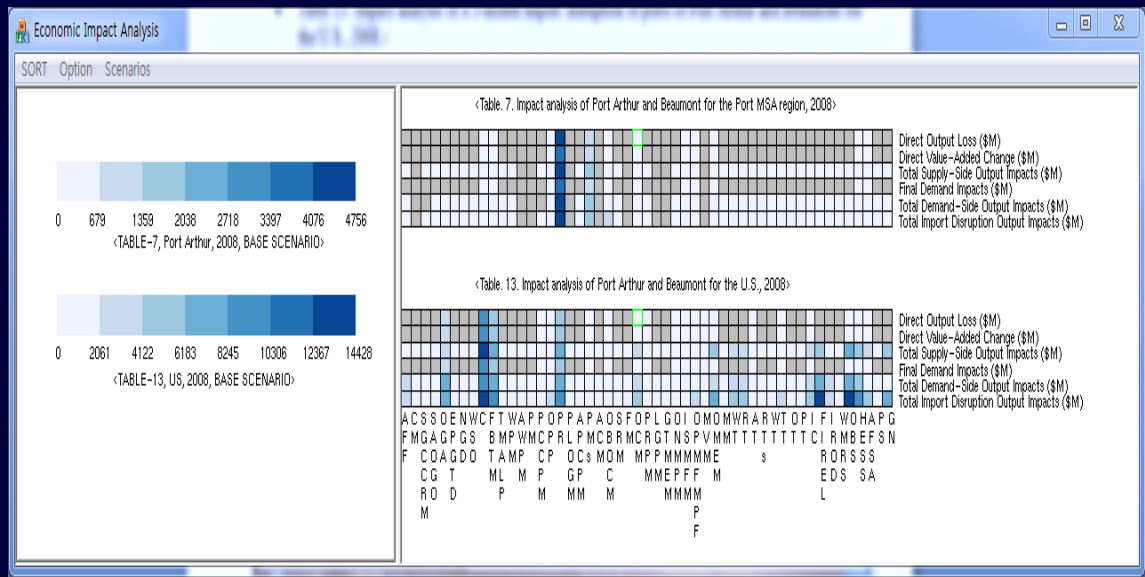


USCG Port Closure Economic Impact VA

Partners: USC CREATE, USCG RDC, USCG D7, USCG LANT

IMPACT:

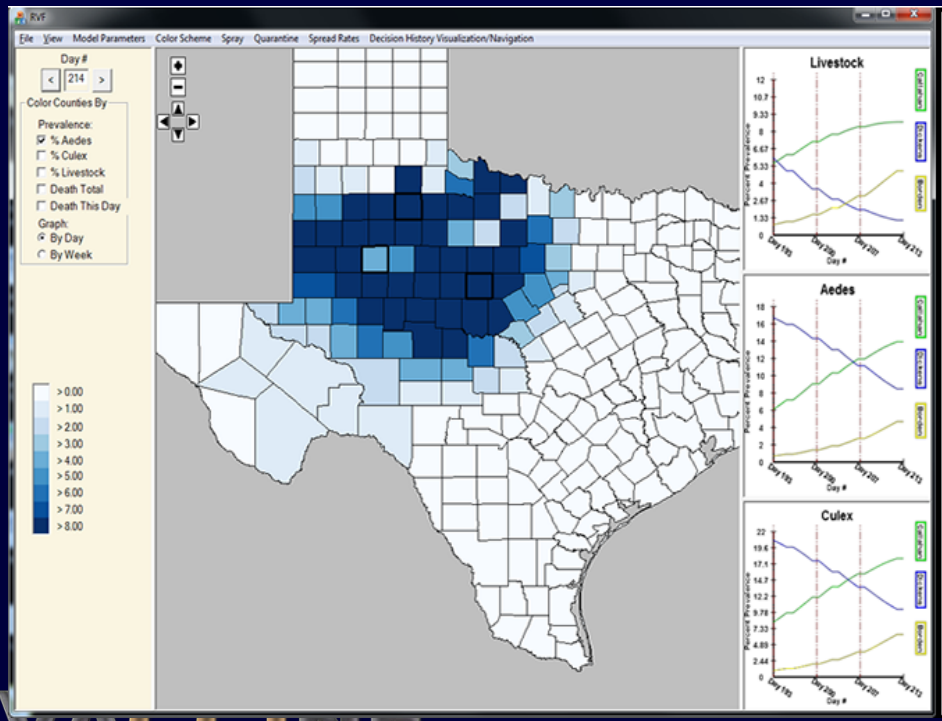
- Provided tool for use analysis and planning for impact of port closure in Port Arthur, Tx
- Economic sector impact, local and national impact
- Impact and effectiveness of alternative mitigation strategies

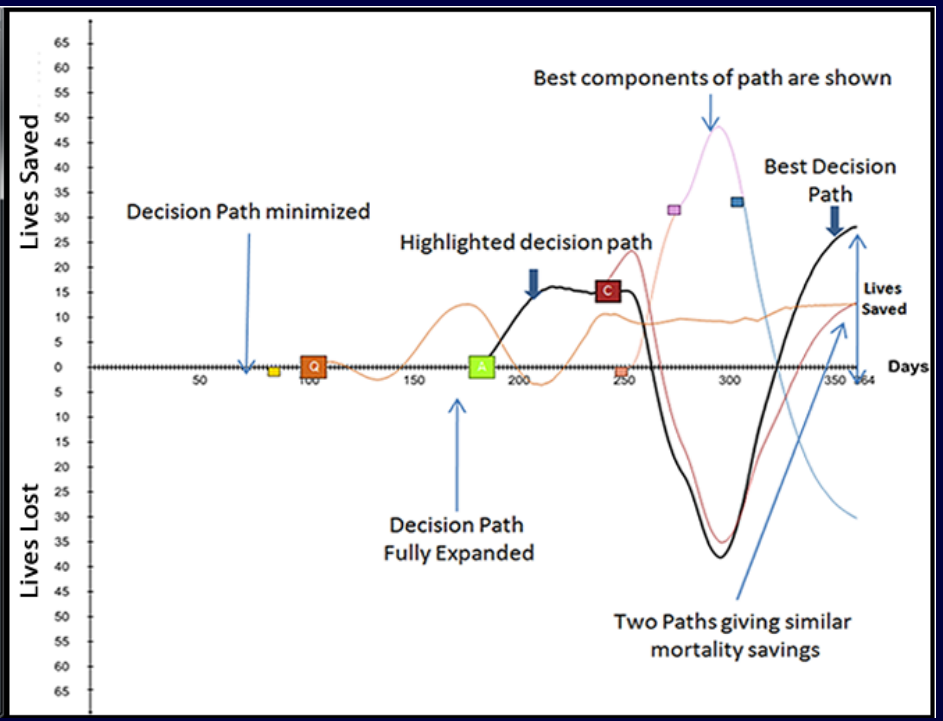


VACCINE

Complex Decision Making: Advanced Decision Support Tools (e.g., Rift Valley Fever)

- Integrated simulation models
- Explore alternative courses of action in decision space and real space



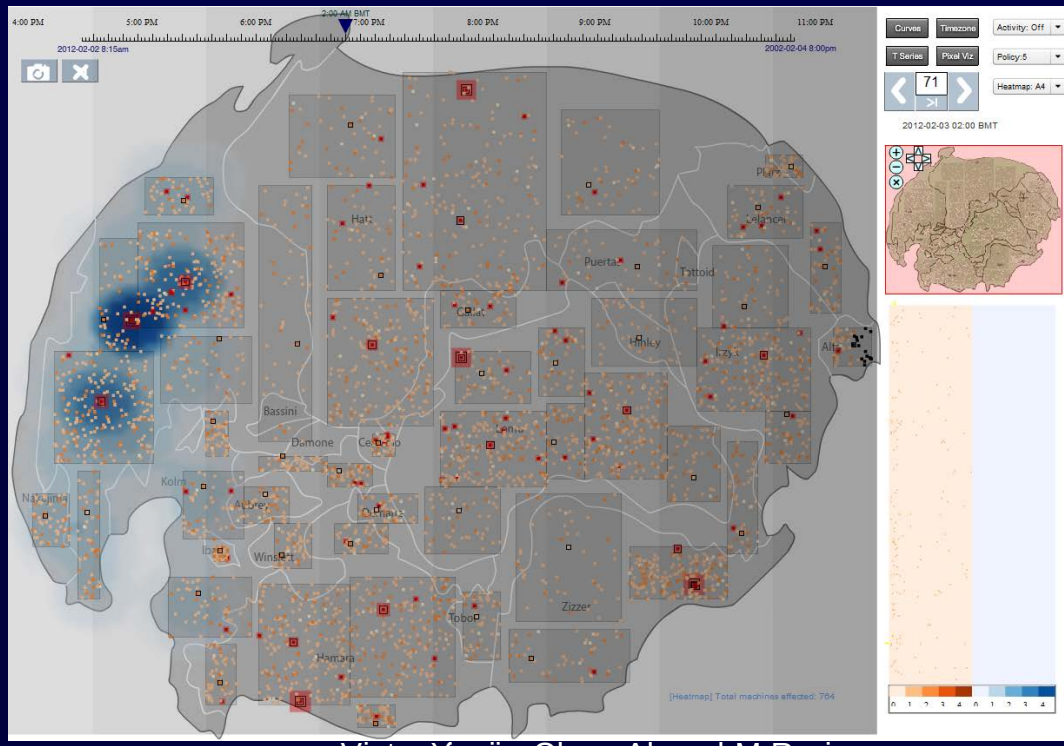


SemanticPrism: A Multi-aspect View of Large High-dimensional Data (Purdue University)

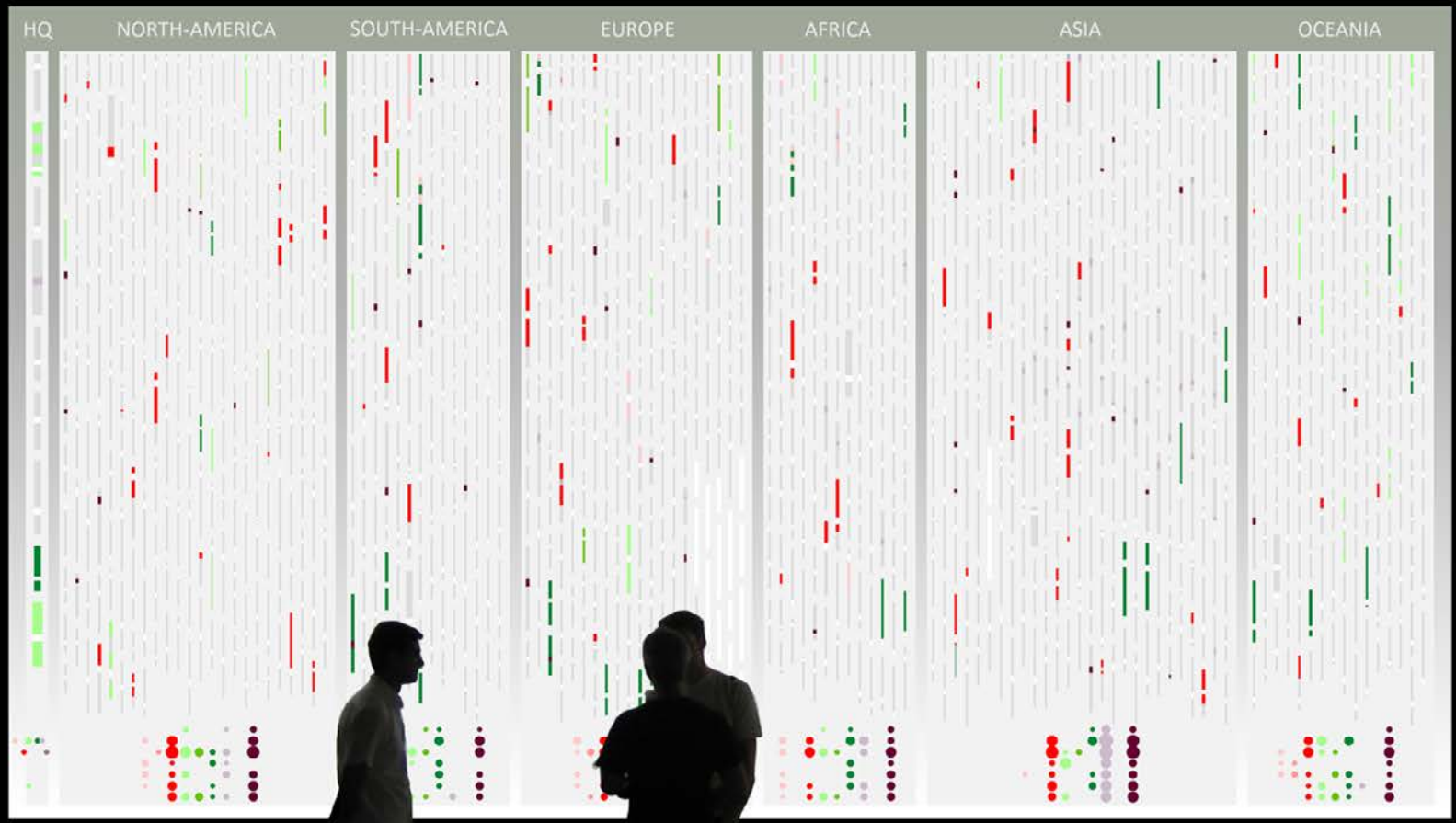
- **VAST 2012 Mini Challenge 1 Award: Outstanding Integrated Analysis and Visualization**

- **Geo-Temporal**
- **Time-serial**
- **Pixel-based**

- **Semantic Zoom**



Victor Yingjie Chen, Ahmad M Razip, Sungahn Ko, Cheryl Zhenyu Qian, David S.Ebert



2013 VAST Challenge MC2 Award

Outstanding Creative Design

www.interactiondesign.us/vast2013/SpringRain

Spring Rain

A Visual Analytics System with an Ambient Information Display +

Visual Analytics: Remember…

- We need to be cognizant of parameters for visual representations
- Appropriate analysis can guide users to interesting features in the data
- Refined analysis through user interaction and their domain knowledge can help discover hidden problems
- There is no single catch-all visual representation or analysis

Keys for Success

- User and problem driven
- Balance human cognition and automated analysis and modeling
  - Often applied on-the-fly for specific components identified by the user
- Interactivity and easy interaction
  - Utilizing HPC and novel analysis approaches
- Understandability of why predicted value is what it is
- Intuitive visual cognition
- Not overloaded with features

For Further Information

www.VisualAnalytics-CCI.org



VACCINE

Visual Analytics for Command, Control, and Interoperability Environments
A U.S. Department of Homeland Security Center of Excellence

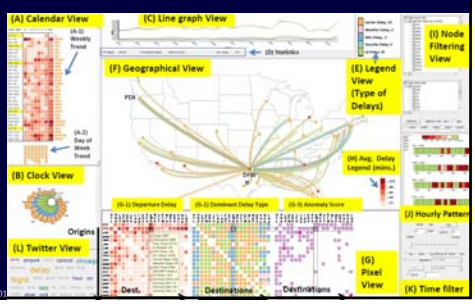
Empowering Effective Decision Making Through Visual Analytics

David S. Ebert
Purdue University
ebertd@purdue.edu

CVADA
Center for Visual Analytics and Data Analytics
A U.S. Department of Homeland Security Center of Excellence
April 2014

Overview

- VACCINE center
- Big Data challenges
- Visual analytics
- Research topics
- Example projects
- Path forward



VACCINE

April 201

DHS Centers of Excellence:

Developing … "A Coordinated University-Based System"

- Multi-COE, multi-division, multi-disciplinary projects
  - Target cross-cutting areas of uncertainty and critical need
  - Integrate National Labs in joint COE projects
- Multi-national, multi-disciplinary workshops in U.S., EU, Asia
- Integration of Education with COEs and S&T divisions
- Integration of MSIs into COEs
- Placement of COE students, scholars and fellows with DHS and Federal labs, industry, universities, foreign institutions
- Transition of research results to end-users

VACCINE

April 2014

Homeland Security
Science and Technology

The 9 DHS Centers of Excellence

1. Center for Risk & Economic Analysis of Terrorism Events (CREATE): University of Southern California
2. National Center for Zoonotic & Animal Disease Defense (ZADD): Kansas State University and Texas A&M University
3. National Center for Food Protection & Defense (NCFPD): University of Minnesota
4. National Consortium for the Study of Terrorism & Responses to Terrorism (START): University of Maryland
5. **The Center for Visual and Data Analytics (CVADA): Purdue University and Rutgers University**
6. The Center for Awareness and Location of Explosives-Related Threats (ALERT): Northeastern University
7. The National Center for Border Security and Immigration (NCBSI): University of Arizona & University of Texas at El Paso
8. The Center for Maritime, Island, Remote and Extreme Environment Security (MIREES): University of Hawaii & Stevens Institute of Technology
9. Coastal Hazards Center (CHC): University of North Carolina at Chapel Hill and Jackson State University

VACCINE

April 2014

Homeland Security
Science and Technology

Who We Are:
International Team of Experts
75+ Faculty, 24 Institutions



- Purdue University
- Georgia Institute of Technology
- Pennsylvania State University
- Stanford University
- University of North Carolina at Charlotte
- University of Washington
- Simon Fraser University
- University of British Columbia
- Ontario Institute of Technology
- Dalhousie University
- University of Manitoba
- Carleton University
- University of Houston, Downtown
- Virginia Tech
- Indiana University
- Florida International University
- University of Texas at Austin
- Morgan State University
- Arizona State University
- University of Stuttgart
- University of Swansea
- Oxford University
- University of Calgary
- University of Victoria

VACCINE

April 2014

Overview

- VACCINE center
- **Big Data challenges**



Big Data: Solution to Global Challenges?



- **Big Data Definition:** A phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex and diverse types of data. (TechAmerica Foundation)
  Big Data is often defined along three dimensions:
  - **Volume** – size
  - **Velocity** – rate of input, update, change
  - **Variety** – different types, sources, variables
- Need:
  - Advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of information (TechAmerica Foundation)
  - **Enable effective, efficient analysis, decision making, planning, and action**

VACCINE

April 2014

Big Data Analytics – The Solution?

- Advanced data analytics for big data can help manage big data
  - Machine learning
  - Advanced statistical analysis
  - Hadoop-deployed analytics
- However, it falls short in several aspects

VACCINE

April 2014

Why VA or What Big Data Analytics Can't Do

(inspired and adapted from David Brooks, New York Times, 2/18/2013)

- Qualitative, fuzzy, and social data
  - Preferences, significance of one relationship over another; Trust
- Context
  - Data is rarely complete, nor incorporates all relevant information Humans have extensive information and experience that never make it into the collected data
- Spurious vs. Significant
  - Big data means more statistically significant events and correlations, but they may not have any relevance
  - Increases noise to signal ratio
- Big problems
  - Complex, multifaceted, multiparameter big challenges with unquantified dependencies

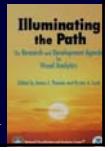
Beyond Visualization to Visual Analytics

Visual Analytics[1] is the science of analytical reasoning facilitated by interactive visual interfaces

Interactive visualization, data analysis, exploration, and decision making with human in the loop!

People use visual analytics tools and techniques to

- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
- Detect the expected and discover the unexpected
- Provide timely, defensible, understandable assessments
- Communicate assessment effectively for action



1. *Illuminating the Path: The R&D Agenda for Visual Analytics, Editors: Thomas and Cook*

VA Approaches to Big Data

- Don't display all the data!
- Extract relevant information at the appropriate natural scales using analytics and statistical models to reduce data space so that it is mappable to the visual space
- User interaction to guide above process
- Adapt to available hardware for the problem
  - Should we use big iron or cell phones?

What We Do – Visual Analytic Solutions

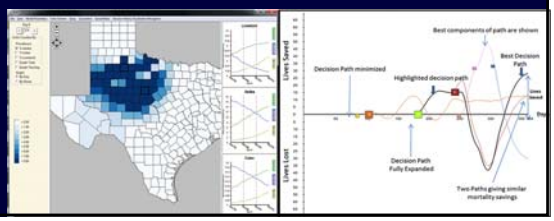
We enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable user to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

Overview

- VACCINE center
- Big Data challenges
- Visual analytics
- Research topics
- Example projects
- Path forward



VACCINE

Our Visual Analytics Process
(extended and adapted from Bertin)

- Define the problem / question
- Determine the data: process / extract the relevant data
  - Abstraction
- Determine
  - Characteristics of the relevant data
  - Types of data: nominal, ordinal, interval, ratio
  - Quality of the data
  - Size, dimensionality & number of data items per sample
- Map the data characteristics to an appropriate visual representation for the task at hand
- Results: Enable rapid, actionable, defensible, reasoned decisions and action

VACCINE

April 2014

Our Guiding Principles for Effective Visual Representations (Norman, Tversky)

- Appropriateness principle:
  - Visual representation should provide neither more nor less information than that needed for the task at hand
- Naturalness principle:
  - Experiential cognition most effective when representation most closely matches the information being represented
  - New visual metaphor's must match the user's cognitive model of the information
- Matching principle
  - Representations much match the task to be performed
  - Visual representations should present affordances suggestive of appropriate actions
- Principle of apprehension:
  - The content of the representation should be accurately and easily perceived

Example of Users With a Problem: Law Enforcement (VALET, iVALET): Tasks and Uses

- Police Chief/Sheriff - strategic goals
  - Resource allocation, return on investment
  - Yearly effectiveness (Are we safer?)
  - Emerging trends
- Police Captain – short-term effectiveness
  - Force allocation for the day/week
  - Emerging trends and anomalies
- Officer – improved situational awareness
  - What happened near me around the same time in the past?
  - What will make me better prepared when I respond to this call?
- Crime analyst/investigator – connections
  - Correlations, dependencies, trends

Our Spatiotemporal Visual Analytics Work

- Techniques:
  - New anomaly and prediction techniques
  - New interactive correlative analysis
  - Natural scale work
  - Social medial visual analytics
- Uses:
  - Anomaly detection
  - Situational awareness
  - Predictive policing
  - Coast Guard visual analytics
    - Search and rescue
    - Risk-based resource allocation
    - Risk-based operational performance

VACCINE

April 2014

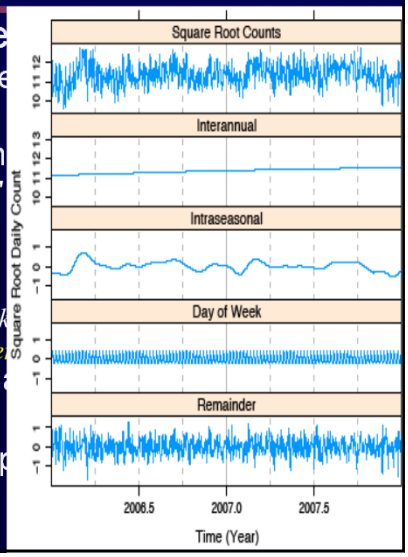
Event/Anomaly Detection - STL

- Temporal trend analysis and pre
  - Seasonal Trend decomposition base weighted regression) (STL)
  - Time series data viewed as the sum

$$Y = T$$

Y : original time series
T : trend component
S : seasonal (daily/week
R : remainder compone

- Large value of *R* indicates substanti
- Filter out the noise component
- If value varies by 2 SD from STL exp



VACCINE

April 2014

Time Series Prediction Using STL

- Seasonal Trend decomposition based on Loess (STL)
  - Apply a sequence of *loess* smoothing operations
  - De-synthesize signal $\sqrt{Y}$ into its various components:

$$\sqrt{Y_v} = T_v + S_v + D_v + R_v$$

where: *v* is the time step
T is the inter-annual component
S is the yearly-seasonal component
D is the day-of-the-week effect
R is the remainder component

Time Series Prediction Using STL

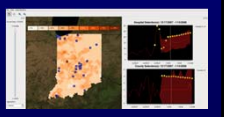
- Prediction using STL:
  - Utilize features of *loess* operator that fitted values ($\hat{Y}$) are linear transformations of input (Y)

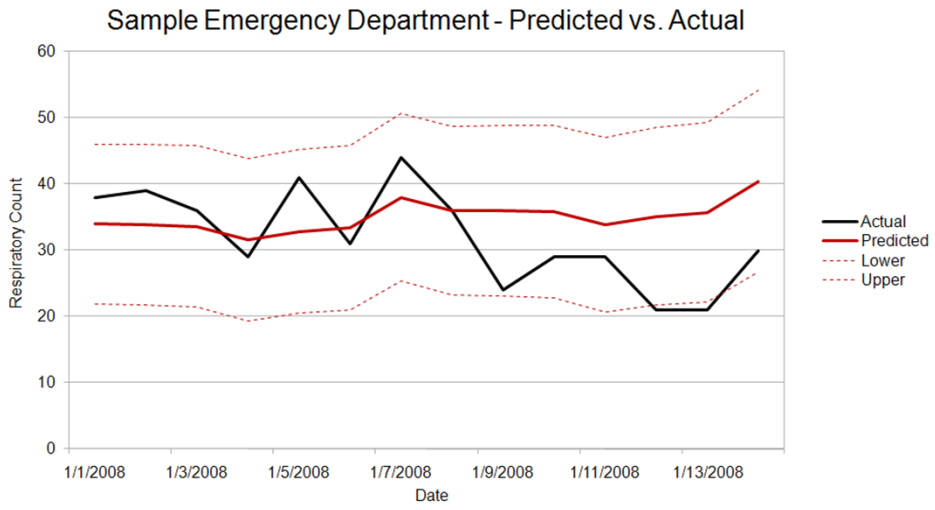
$$\hat{y}_i = \sum_{i=1}^{n} h_{ij} y_j \Rightarrow \hat{Y} = HY$$

  - To predict ahead by *n* days, use operator matrix *H*

$$\hat{y}_{n+1} = \sum_{i=1}^{n} H_{n+1,i}\, Y_i$$

Predictive Visual Analytics



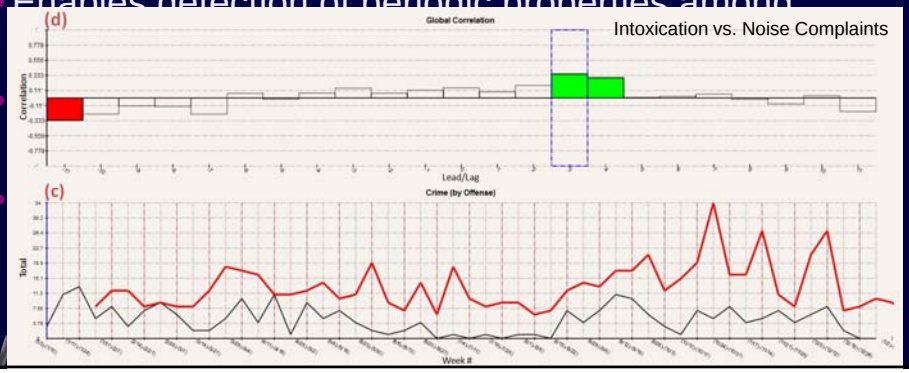


Visual Correlative Analytics: Value of Visual Correlative Exploration?

- Understanding the relationships between spatial and temporal trends
  - Acts as precursor to creating predictive models from data
- Provides insights into the workings of real world environments
- Different forces that determine nature of relationships
  - E.g., one event precursor to another
  - Requires exploration of leads/lags between datasets
  - Determine if one dataset is related to another at a different temporal location
- Aids analysts and decision makers in hypothesis generation

Multivariate Correlative Predictive Analytics:

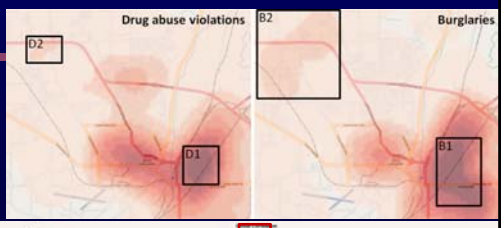
- Two Key Features:
  - Automatic correlation computation against lead/lags
  - Temporal and spatial windowing
- Enables detection of periodic properties among

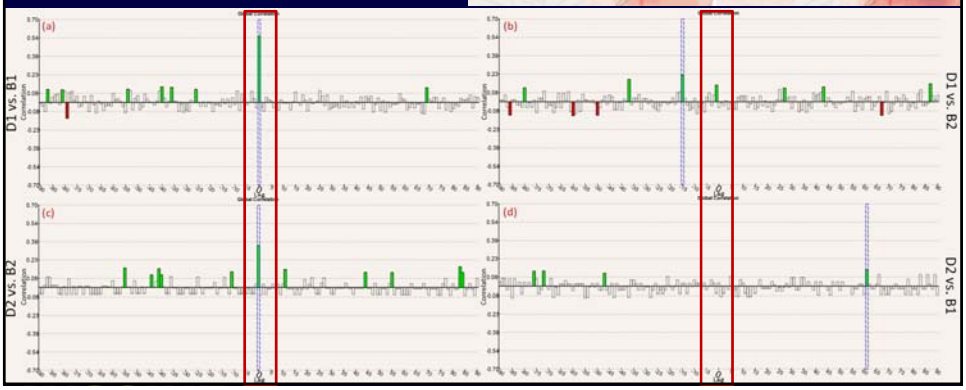


Intoxication vs. Noise Complaints

Spatial Windowing: Drug Abuse Violations vs. Burglaries

- Over geospatial hotspots





The Scale Problem

- Data scale:
  - Coarse scale analysis is quick but can have too much variance or contain combinations of multiple signals reducing effectiveness
  - Fine scale analysis is fast but may limit applicability because of fine resolution and can have data sparseness issues
- User/problem scale
  - Users perform analysis at scales that are natural to them and the task/problem
  - Characteristics might not be well articulated or available digitally

Natural Scale Templates

- Spatial and temporal visualization and analysis templates that balance data scale issues and human decision maker scale issues
- Some important questions:
  - How do we facilitate appropriate analysis for the casual expert?
  - How can casual expert improve performance of analytical results?
  - How do we effectively communicate the result, accuracy, and certainty?

Dynamic Covariance Kernel Density Estimation (DCKDE)

- Factors in properties of certain crime types:
  - Crime occurring at a location can be correlated with recent crimes at nearby locations
  - Removes stochastic variation from width of kernel
- Using KDE gives continuous function with fewer zeroes for prediction and analysis
- An example problem…

VACCINE

April 2014

Predicting For Hotspots On Tuesday 3/11/2014



Predicting using standard KDE



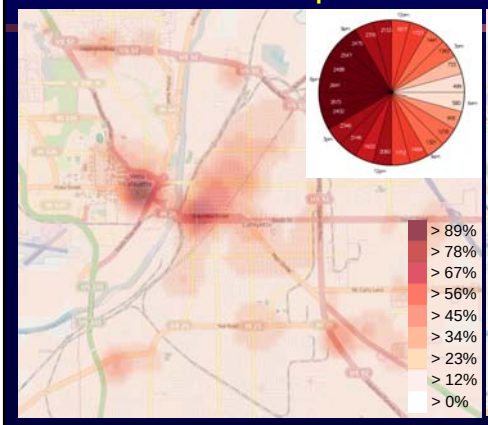
Predicting using DCKDE

VACCINE

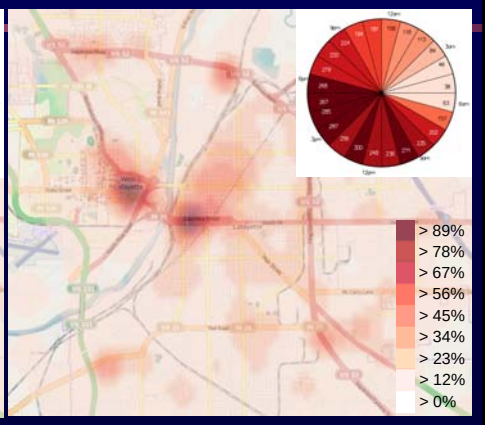
April 2014

Overall Historical Distribution: Spatial and Temporal Variations

Note areas of sparse data



Past 2 Years
3/11/2012 – 3/10/2014

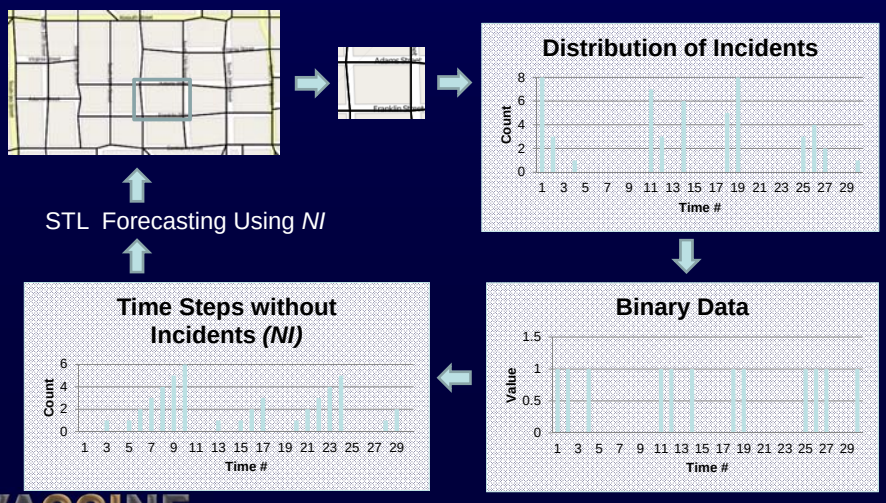


March + Tuesday Filter
(Past 10 Years)

VACCINE

April 2014

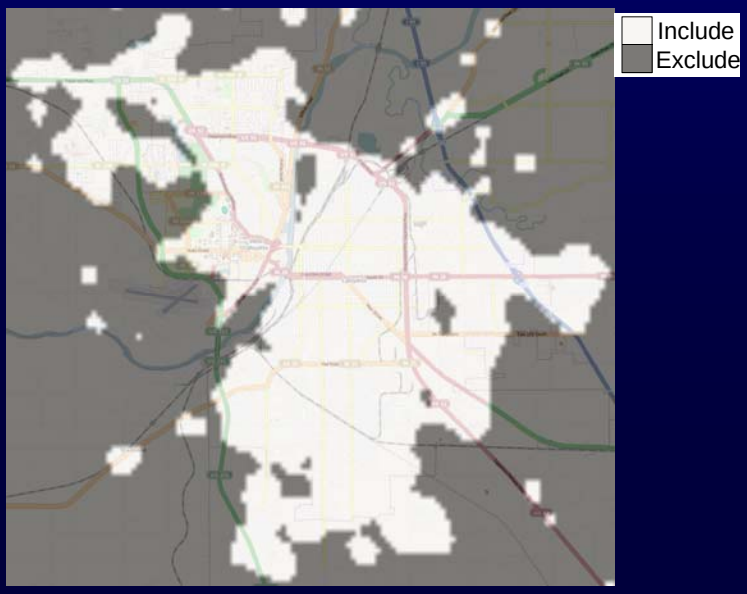
Geospatial Natural Scale Template For Sparse Data



VACCINE

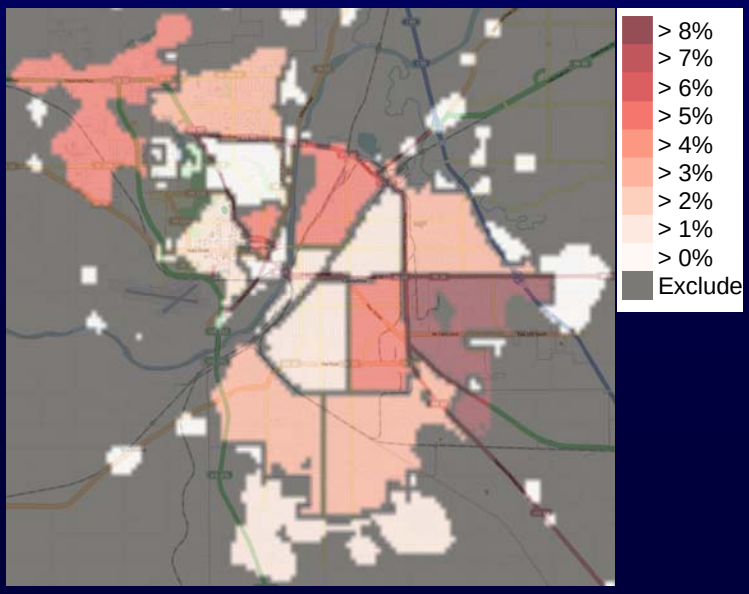
April 2014

Geospatial Natural Scale Template



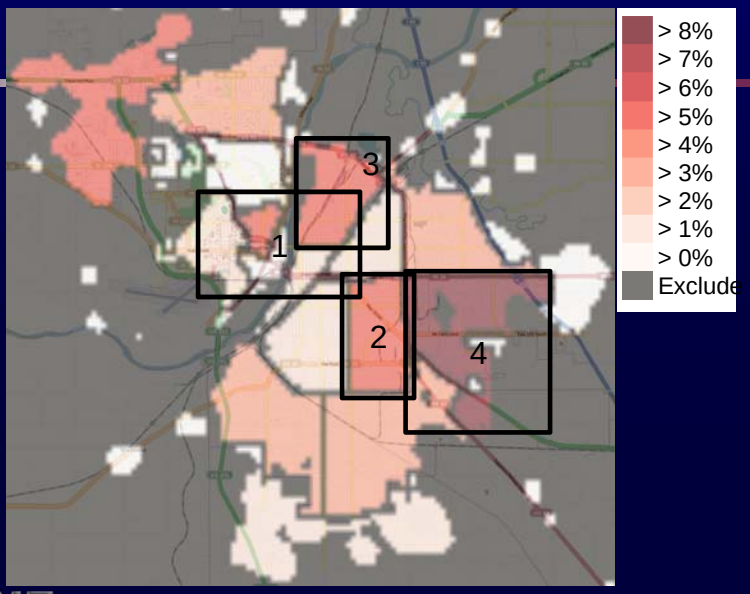
April 2014

Predicting and Allocating Resources by Police Beats for 3/11/2014



April 2014

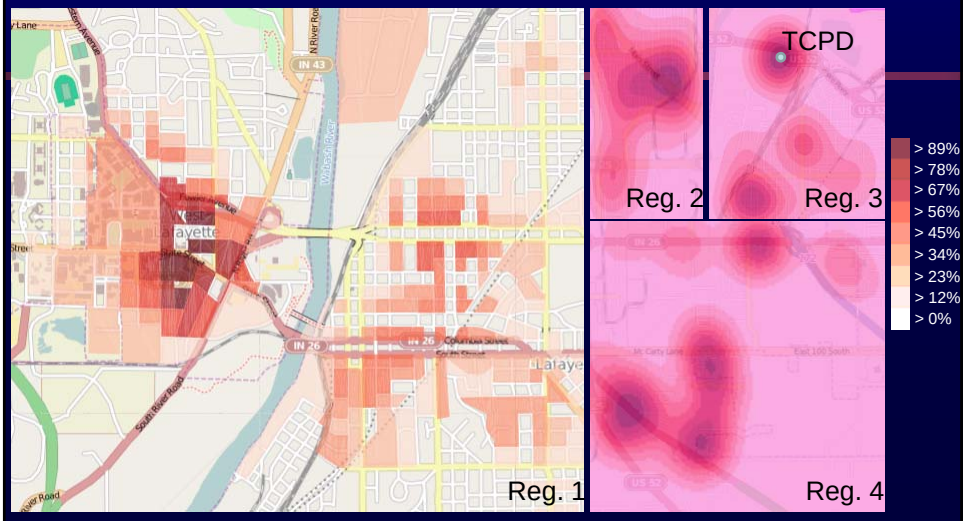
User Refinement Using Domain Knowledge



VACCINE

April 2014

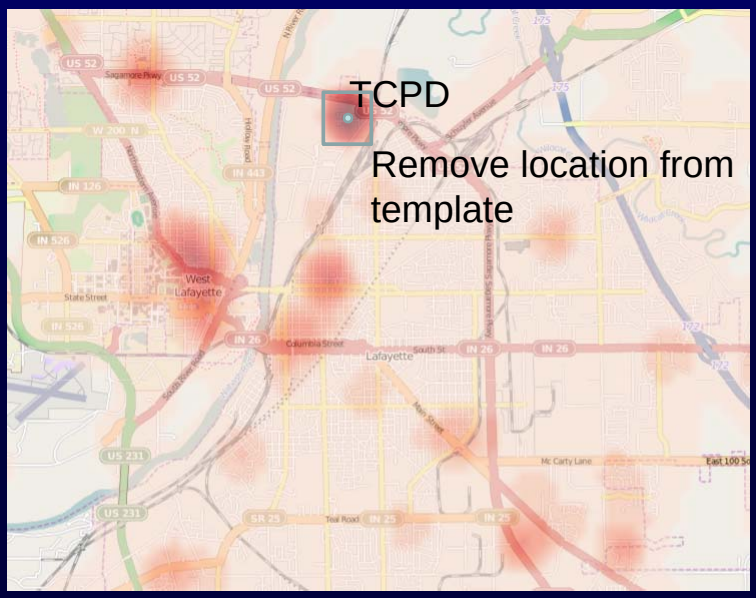
User Refinement Using Domain Knowledge



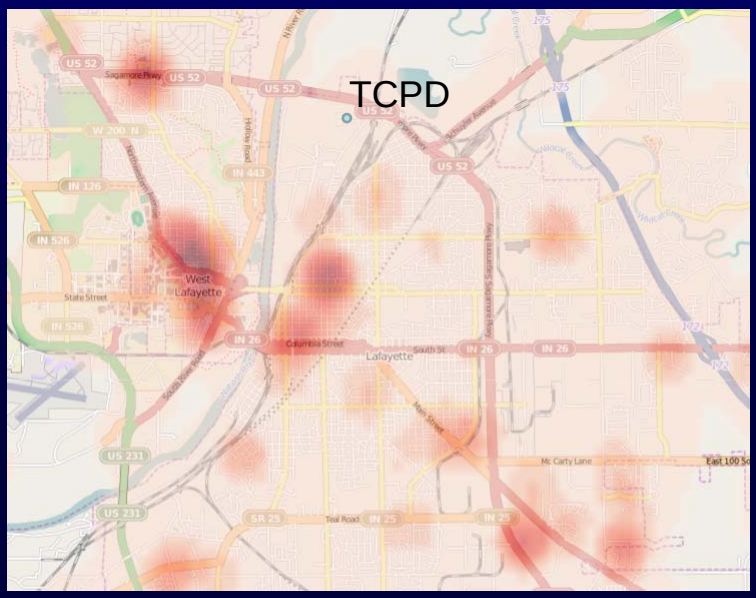
VACCINE

April 2014

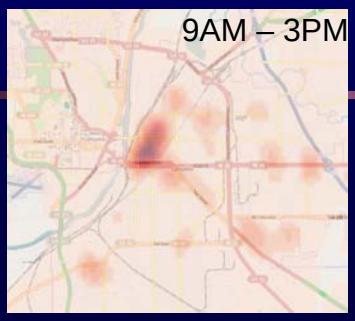
Hotspots Prediction For Tippecanoe County

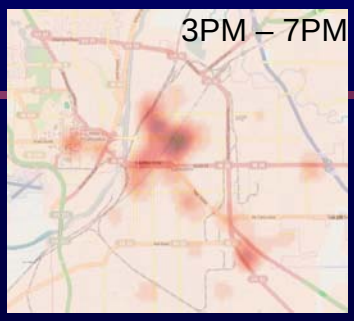


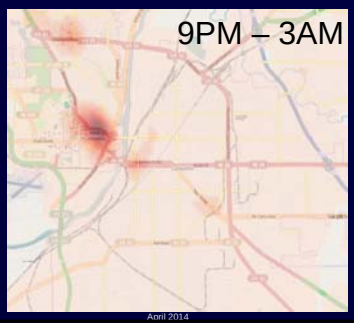
Hotspots Prediction For Tippecanoe County



Applying Natural Time Templates





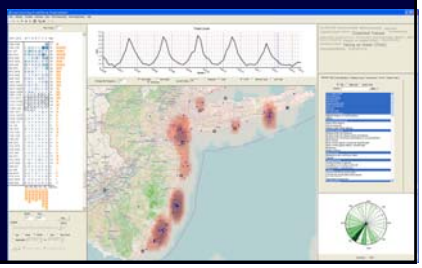


VACCINE

April 2014

Overview

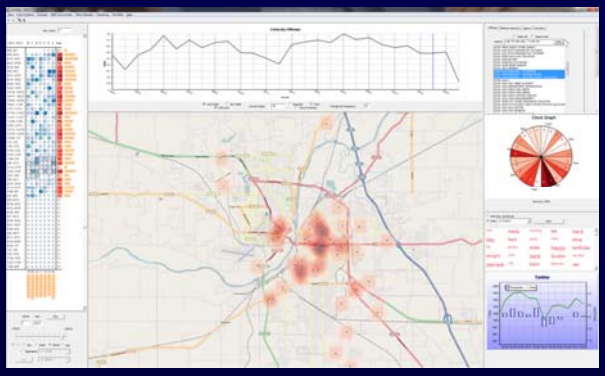
- VACCINE center
- Big Data challenges
- Visual analytics
- Research topics
- Example projects
  - VALET system result
  - Social Media
  - US Coast Guard
- Path forward

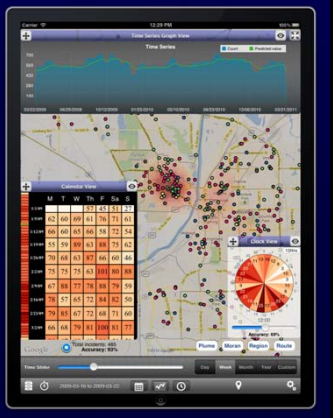


VACCINE

April 2014

Visual Analytics Law Enforcement Toolkit (i)VALET





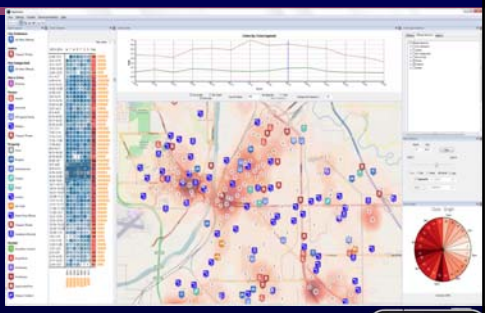
VACCINE

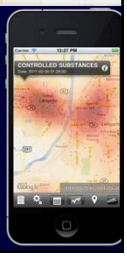
April 2014

Visual Analytics Law Enforcement Toolkit (VALET, iVALET)

Impacts:

- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public (June 2014) for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts





VALET delivered:

- Spring 2011: WL, Lafayette Police
- Fall 2013: NYPD

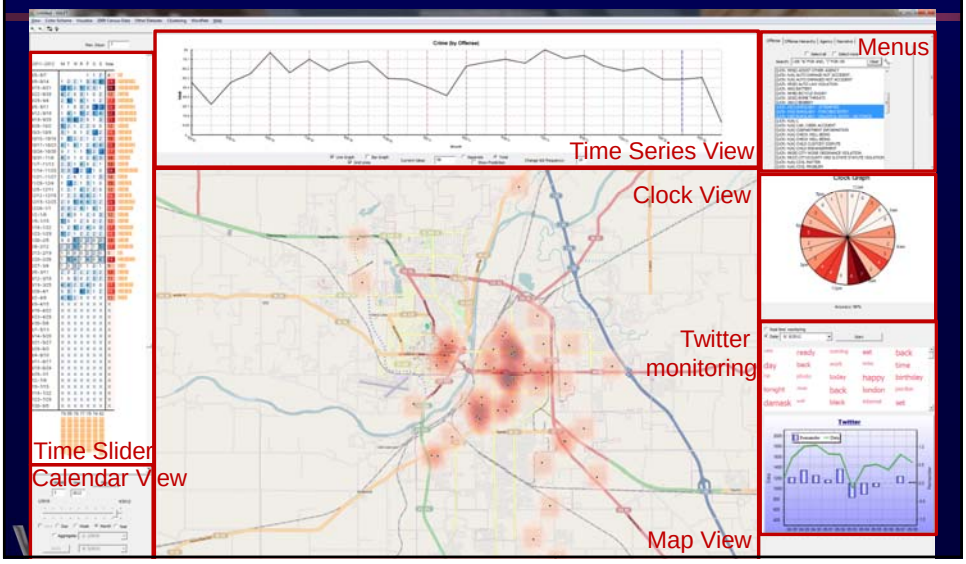
iVALET delivered:

- October 2011: Purdue, WL Police

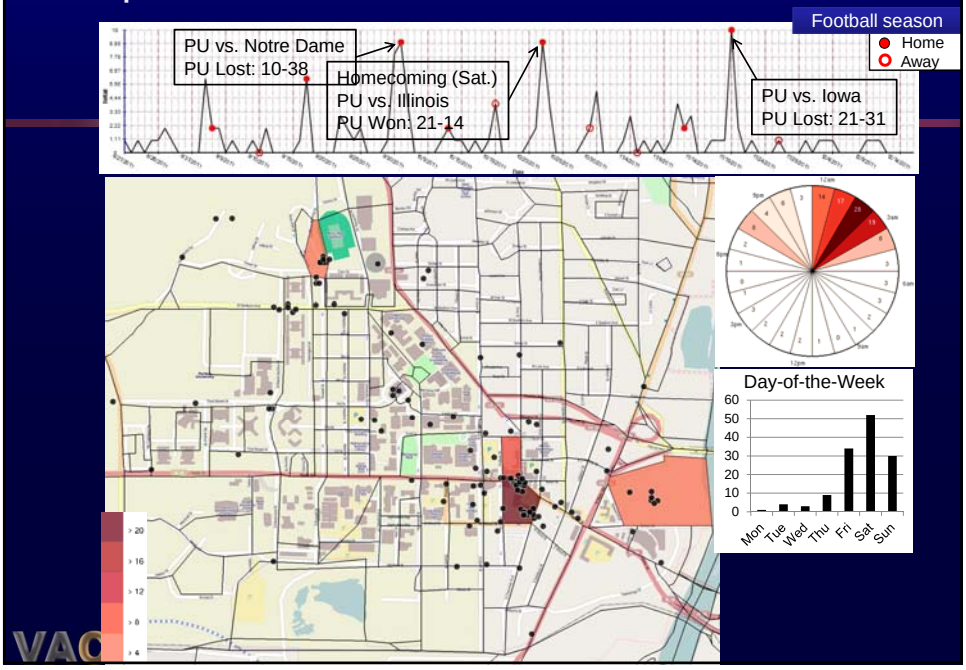
VACCINE

April 2014

VALET Overview

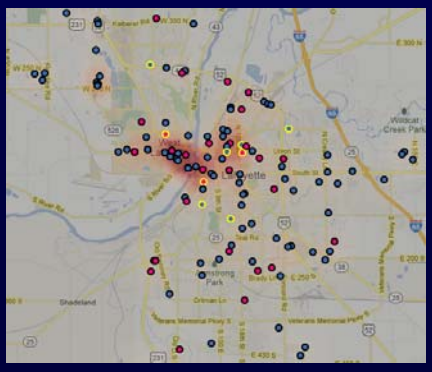


Example: Drunkenness / Public Intoxication



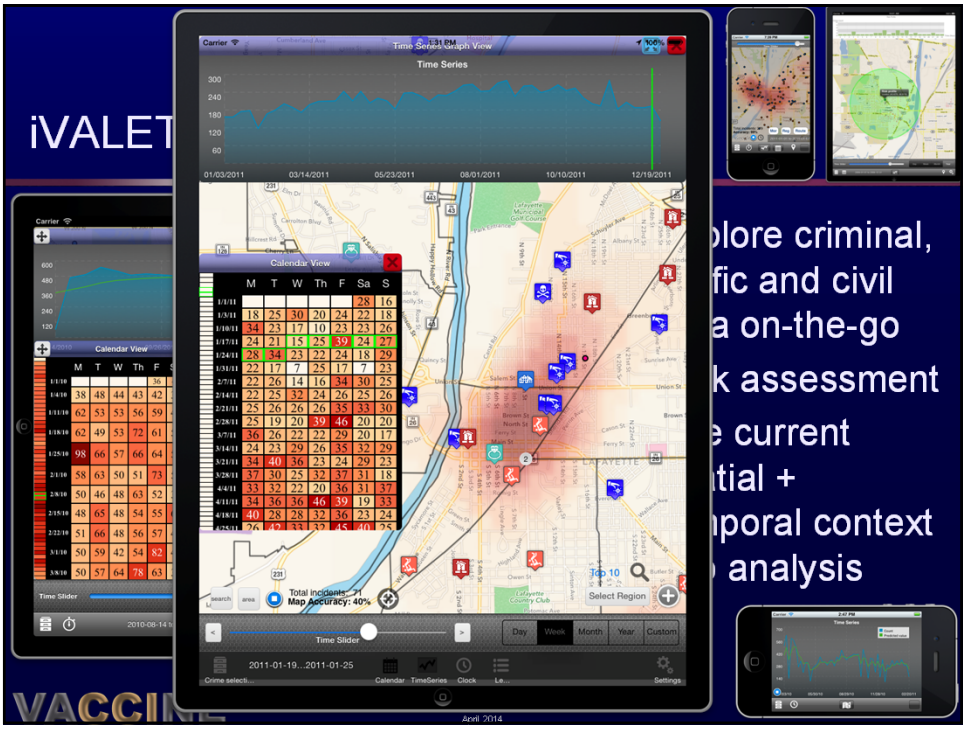
Top 10 Hot Incidents

- Identify unusual localized high-frequency patterns of crimes (near repeats)
- Each data entry is checked for other crimes with similar properties within a 1 block radius of the incident location and a 14-day time period;
- Top 10 incidents with the most number of related incidents in this space-time window are highlighted



VACCINE

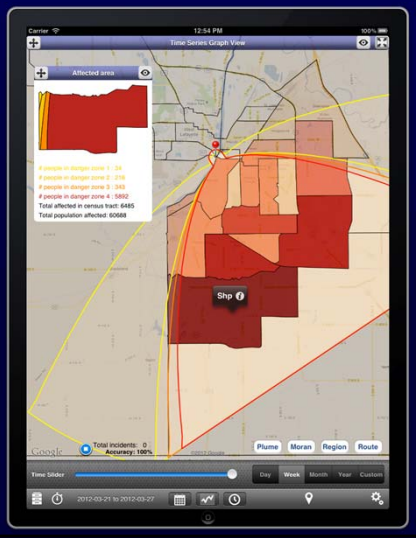
April 2014



iVALET Interactive Plume Visualization and Evacuation Planning

- Chemical release plume modeling identifies census tracts with the highest number of expected people affected



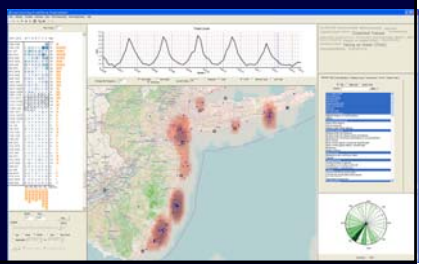


VACCINE

April 2014

Overview

- VACCINE center
- Big Data challenges
- Visual analytics
- Research topics
- Example projects
  - Social media
  - US Coast Guard
- Path forward



VACCINE

April 2014

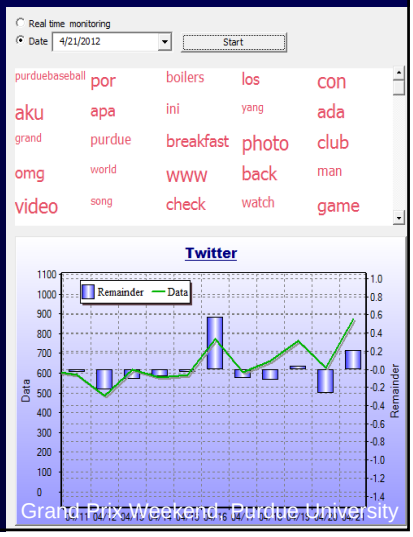
Example Project:
Social Media Visual Analytics

VACCINE

April 2014

Social Media: Real-time Twitter Visual Analytics
(Purdue, Stuttgart, Penn St.)

- Anomalous topic extraction using LDA and novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud



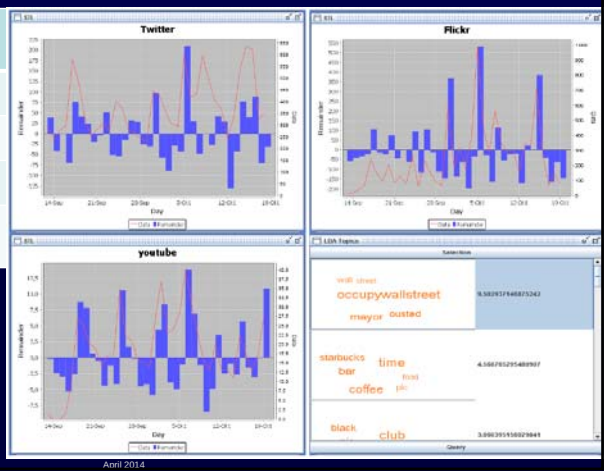
VACCINE

April 2014

Correlation Between Multiple Social Media

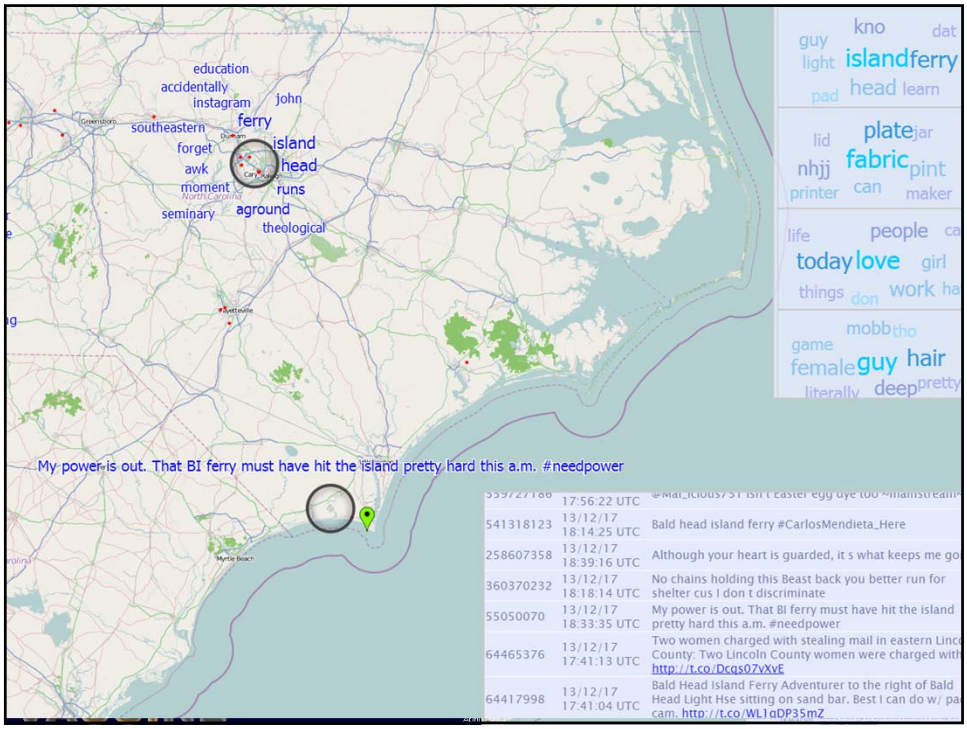
- Occupy Wall Street event

Date	Twitter (z-score)	Flickr (z-score)	Date	YouTube (z-score)
9/17, 2011	2.53	0.97	9/18, 2011	5.98
9/30, 2011	3.20	3.50	10/2, 2011	1.38
10/5, 2011	3.17	3.36	10/6, 2011	2.75
10/15, 2011	4.90	4.53	10/15, 2011	2.19

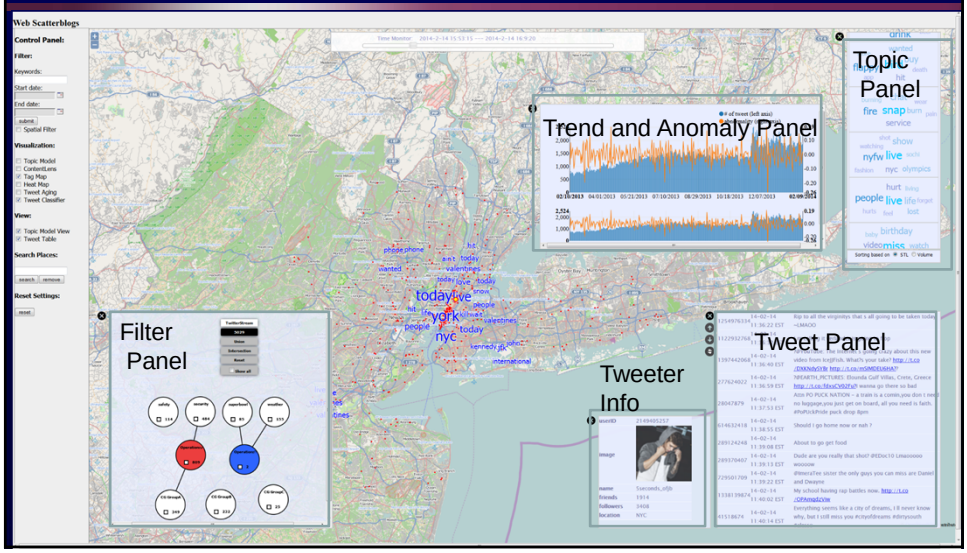


VACCINE

April 2014



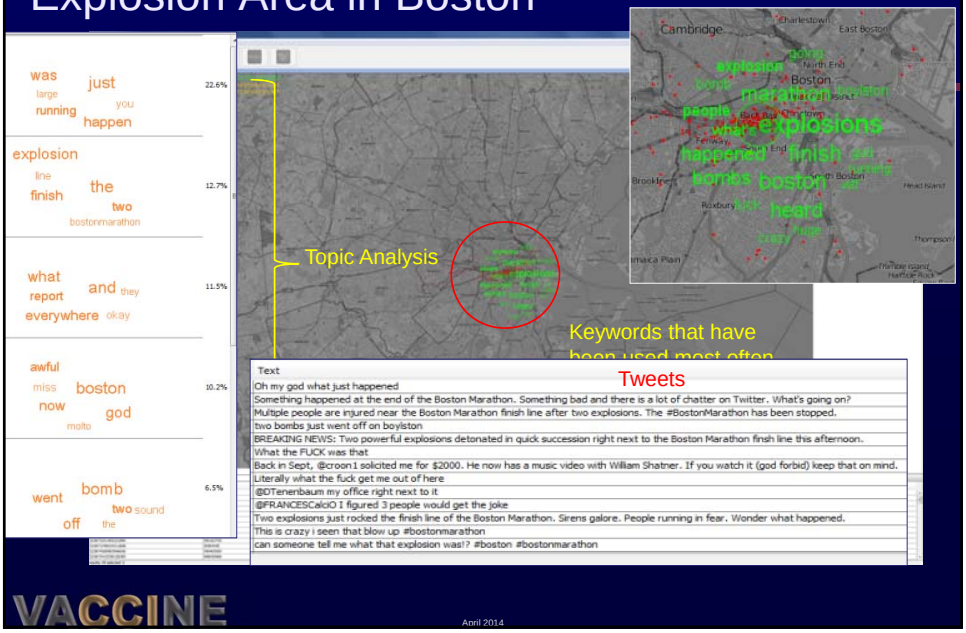
Scatterblogs – Web Version Overview



Anomaly Detection



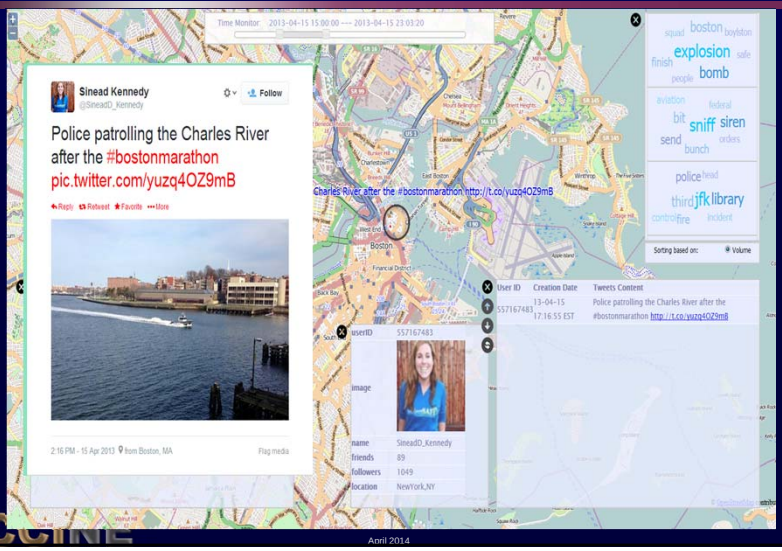
Explosion Area in Boston



First Response (Tweet & Picture)



Police Patrolling The River



VACCINE

April 2014

Cascading Critical Infrastructure Resiliency Modeling and Analytics (VASA)



- Purpose: Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.
- Collaborating Institution(s):
  Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Texas (TACC), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities
- End-User(s): Power Suppliers (e.g., Duke Energy), Cyber Community (e.g., Cisco), Quick Service Restaurants and suppliers, food supply

Purdue

UNCC



U Minn

TACC

VACCINE

April 2014

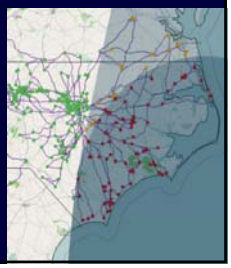
VASA: Visual Analytics for Simulation-Based Action

Collaborating Institution(s): Purdue, Minnesota, UTexas, UNCC + German universities
End-User(s): Fast-food restaurant chain, emergency management and planning personnel

Impacts and Accomplishments:



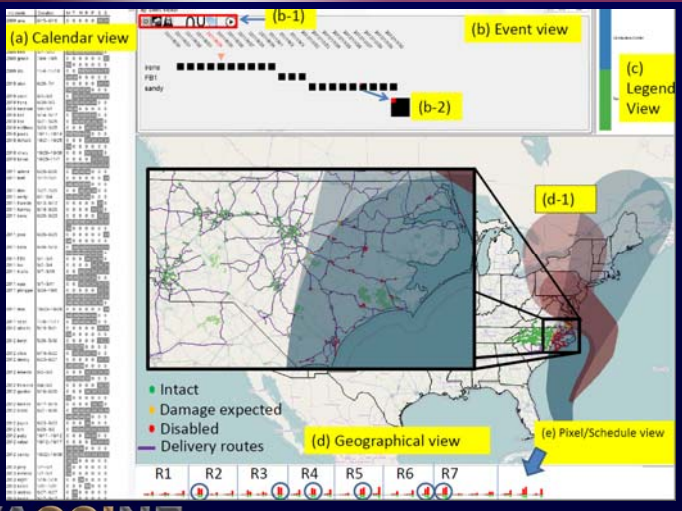
- Support decision-making for extreme weather and natural disaster scenarios
  - Combine real and simulation data
  - Allow "what-if" exploration
- **System of systems**: binds together multiple simulations models from collaborators into coherent whole
  - **Minnesota**: food distribution model
  - **Texas**: simulated and historical weather (hurricanes, storms)
  - **UNCC**: critical infrastructure
  - **Purdue**: roads + interaction visual analytics tool
- **Challenge**:
  combine interactive VA with complex simulation models



VACCINE

April 2014

System Interface

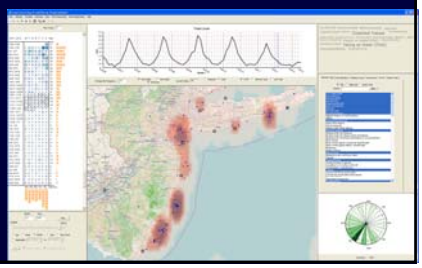


VACCINE

April 2014

Overview

- VACCINE center
- Big Data challenges
- Visual analytics
- Research topics
- Example projects
- Path forward



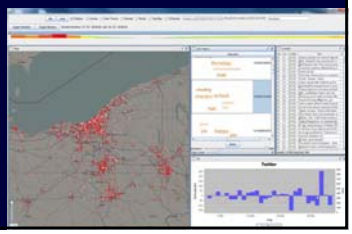
VACCINE

April 2014

Visual Analytics Uses for Public Safety



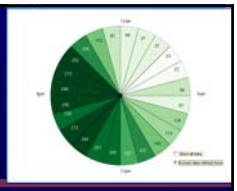
- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Multisource, multimedia massive data integration & analysis
- Purpose: Planning for resiliency, training, detection, investigation, response, recovery, remediation



VACCINE

April 2014

Risk-Based Allocations

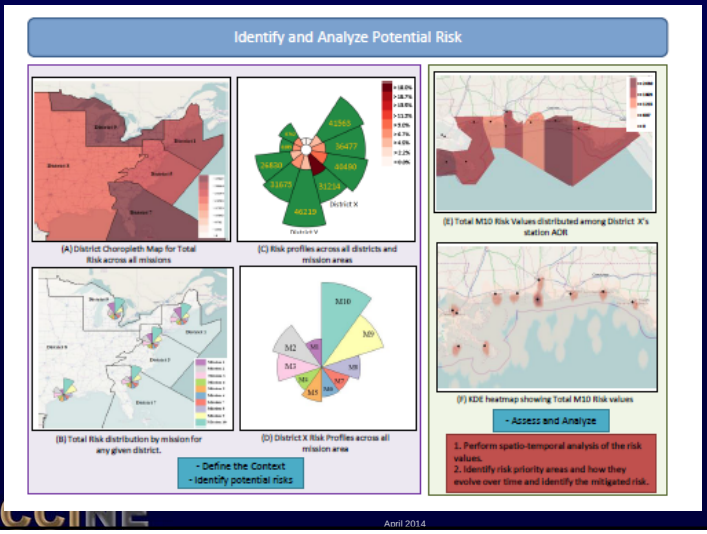


- Comparative visual analysis of mission cases/hours vs. staffing hours
- Comparative visualization of resources vs. risk
- Trend visual analytics
  - Increase/decrease in resource allocation
  - Increase/decrease in risk (total, mitigated, residual)
  - Increase/decrease in incidents
- Exploration of alternatives and effect on risk
- Predictive analytics based on historical data (STL and EWMA)

VACCINE

April 2014

Multi-Visualization, Multivariate VA for Risk-Based Decision Making



VACCINE

April 2014

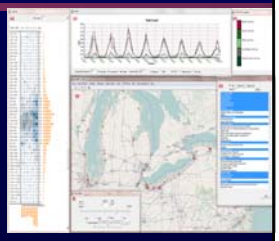
U.S. Coast Guard Search and Rescue VA

(cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771

IMPACTS:

- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Hurricanes Sandy and Irene resource allocation decisions based on cgSARVA analysis and visualization
- Informed Commandant's budget testimony to Congress
- Key component of USCG D9 reallocation plan for 2011-12
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ



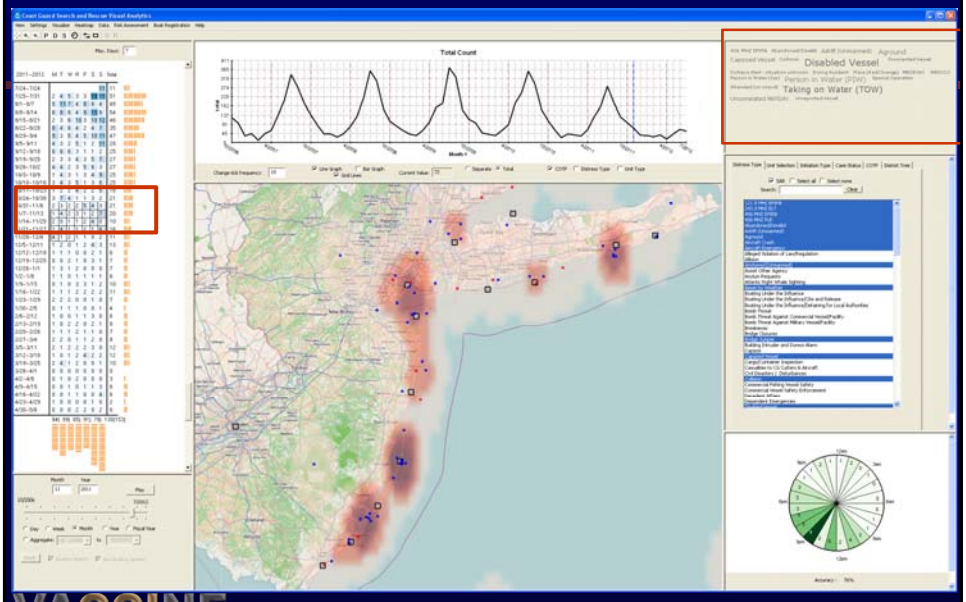


VACCINE

April 2014

Example: Risks and Consequences From Sandy:

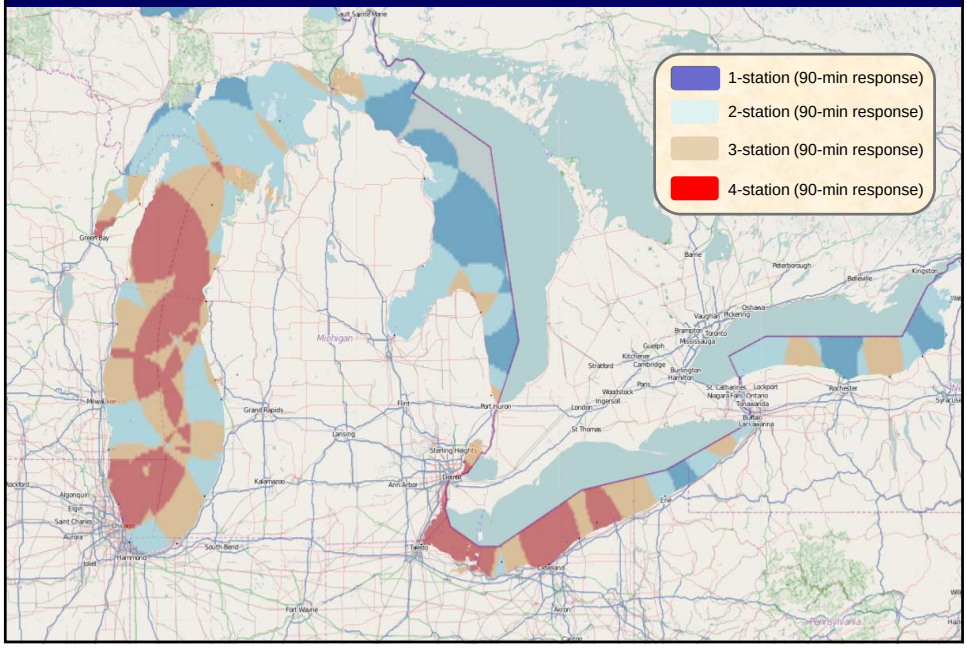
SAR Cases November 2011 NJ/NYC Area



VACCINE

April 2014

Response Efficiency – Potential Future Assets



Directions Forward, Keys to Success & Challenges

VACCINE

April 2014

Seven Challenges for Proactive & Predictive Visual Analytics

1. Creating computer-human visual cognition environments
2. Integrating interactive simulation and analytics
3. Solving specific scale issues and cross-scale issues
4. Managing uncertainty and time
5. Enabling risk-based decision making environments
6. Developing the Science of Interaction for Visual Analytics
7. Defining appropriate evaluation methodologies for visual analytics

VACCINE

April 2014

Visual Analytics: Remember…

- We need to be cognizant of parameters for visual representations
- Appropriate analysis can guide users to interesting features in the data
- Refined analysis through user interaction and their domain knowledge can help discover hidden problems
- There is no single catch-all visual representation or analysis

VACCINE

April 2014

Keys for Success

- User and problem driven
- Balance human cognition and automated analysis and modeling
  - Often applied on-the-fly for specific components identified by the user
- Interactivity and easy interaction
  - Utilizing HPC and novel analysis approaches
- Understandability of why predicted value is what it is
- Intuitive visual cognition
- Not overloaded with features

VACCINE

April 2014

For Further Information

www.VisualAnalytics-CCI.org



VAC

VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

Evaluation: Multistep and Multi-Level



David S. Ebert

CVADA

Center for Visual Analytics and Data Analytics

A U.S. Department of Homeland Security Center of Excellence

October 2013

Is This What We Should Do?



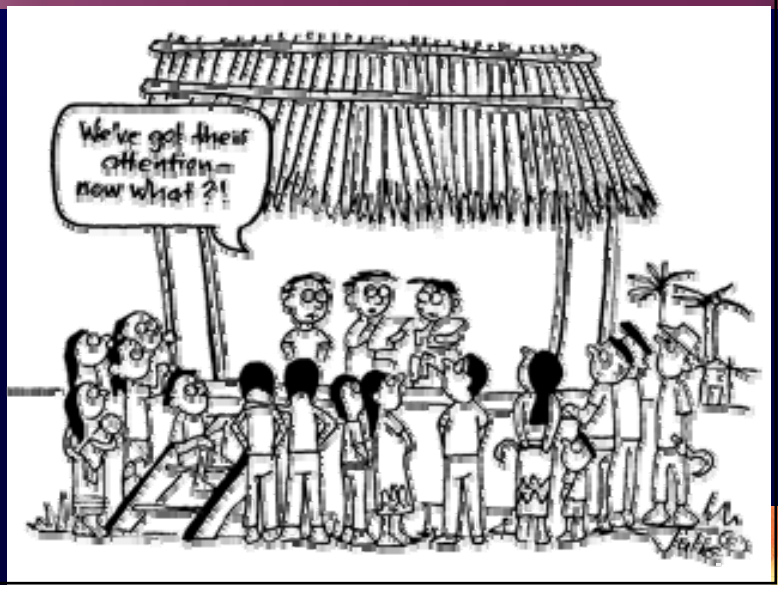
VACCINE

How Much Evaluation Should We Have?

- Originality, Rigor, and Significance easy to hard to judge? (test of time)
- So how much should we have?
- It all depends…
  - I've criticized a paper for too little evaluation
  - I've criticized a paper for too much evaluation
- One size doesn't fit all…

Where Are We?

- The visualization community has been emphasizing/requiring evaluation in the review process
- Community has responded – for good or bad



The Problems

- Much of the evaluation does not answer the important questions:
  - Is this better than previous contributions?
  - Is the system effective and useful?
- Frequently have unnoticed confounding factors
- Sometime use wrong scientific evaluation approach: social science approaches more appropriate – data driven science (Science 2.0?)



Traditional User Studies

- Simple user studies with students that show statistically significant performance on toy tasks often do not work



What is Needed?

- Evaluation and feedback need to be part of the iterative development process and occur at many levels from
  - Efficiency and accuracy to
  - Usability up to
  - Effectiveness



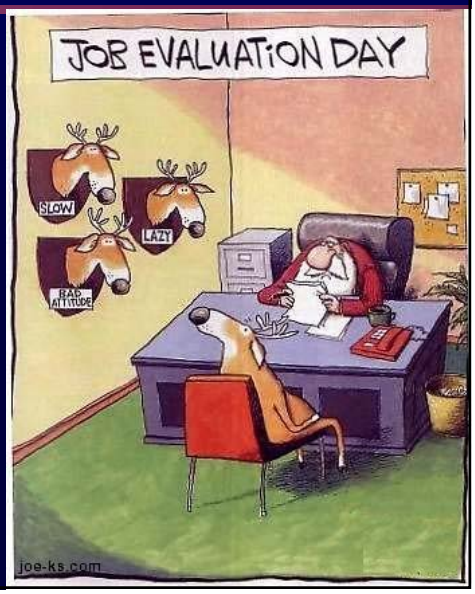
What is Needed?

- Design of evaluation from the start with targeted audience involved
- Iterative refinement with users as development continues until and after deployment
- Without ever conducting a formal lab user study with statistical significance!



Takeaways

- Publications without user studies are fine and sometime better!
- Evaluation results are important and significant and should be publishable on their own!



For Further Information

www.VisualAnalytics-CCI.org



vaccine@purdue.edu
ebertd@purdue.edu

Homeland Security
Science and Technology

Homeland Security University Programs
Today's Research & Education, Tomorrow's Security

DHS University Centers of Excellence Innovation Showcase

FIRST RESPONDER TECHNOLOGY DEMONSTRATION



VACCINE

Visual Analytics for Command, Control, and Interoperability Environments
A U.S. Department of Homeland Security Center of Excellence

Mission:

- **Create visual analytic tools to analyze and manage vast amounts of information (from social media to databases) for timely decision making.**

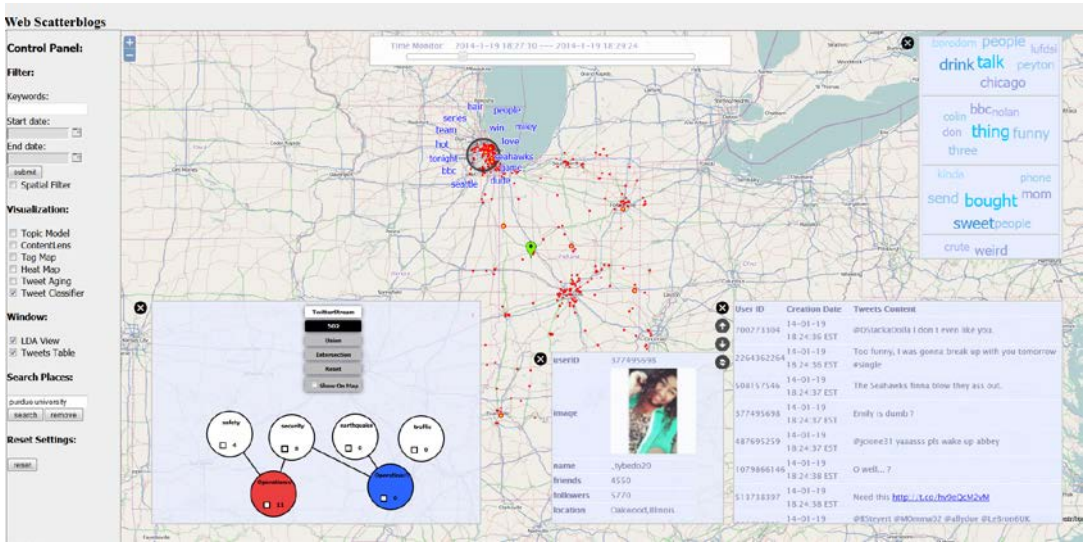
Dr. David S. Ebert
Director, VACCINE

Purdue University

02.11.14

ScatterBlogs

- What is ScatterBlogs?
  - Social media analysis tool
  - Provides analysts with scalable analysis and visualization of social media post using topic extraction, combinations of key word filters, word cluster examination, and unusual event detection to provide situational awareness
- How does it work? What are its benefits?
  - Filters Twitter (and other social media sources) during or after a disaster by comparing them to historical information and trends of previous, similar events (i.e., hurricanes and other natural disasters, terrorist attacks, and traffic.)
  - Identifies trends and anomalies

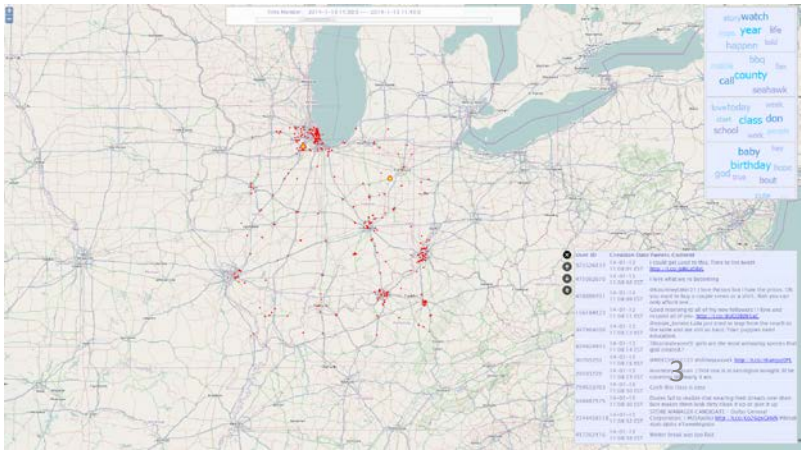


Homeland Security
Science and Technology

Homeland Security University Programs
Today's Research & Education, Tomorrow's Security

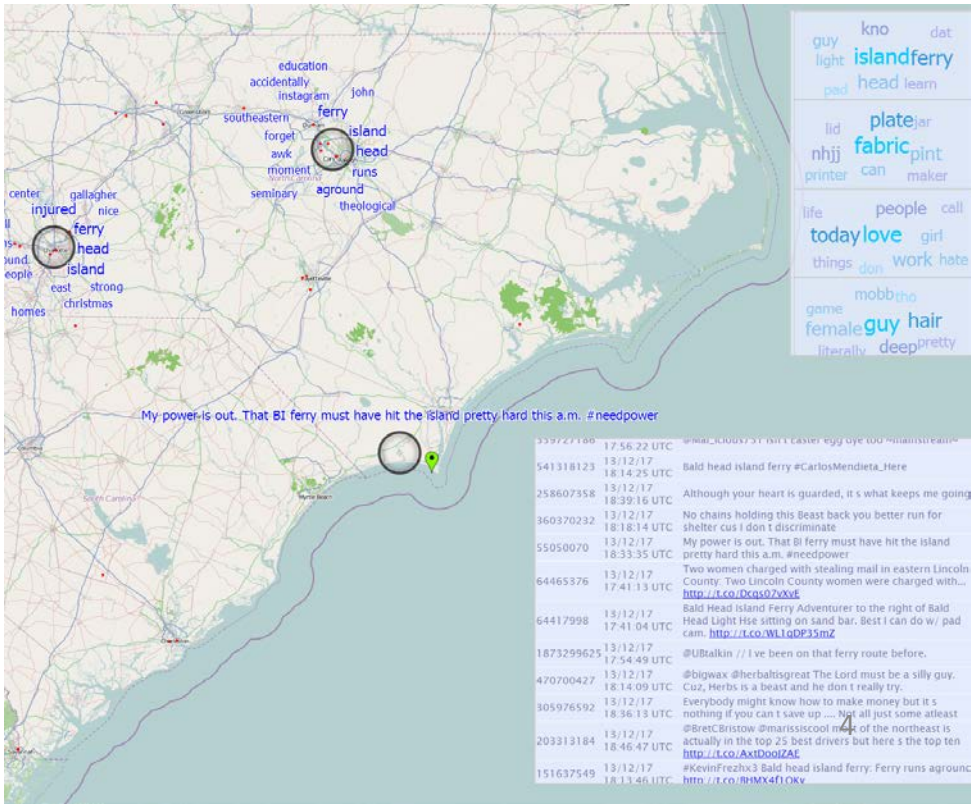
ScatterBlogs

- Impact/Relevance of ScatterBlogs to First Responders
  - When a disaster strikes, the Internet is flooded with microblogs, tweets, Facebook messages, and other social media posts with critical information
  - Used correctly, this information can shape the way public safety agencies handle the response to and recovery from major events
  - **ScatterBlogs enhances time-critical decision making during disasters and emergencies**
  - Example:
    - Analysts used ScatterBlogs as an early warning system during the 2013 Boy Scouts of America Jamboree in West Virginia to capture Tweets related to a variety of events, including severe weather, security, safety, and health



ScatterBlogs

- Next Steps
  - ScatterBlogs currently operates in a desktop and web based version
  - Field pilots will continue and researchers will upgrade the system based on end-user feedback
    - The Ohio State Highway Patrol and the U.S. Coast Guard are currently testing ScatterBlogs.
    - A number of police departments have also expressed interest in the tool.
  - Current focus has been on Twitter
    - Incorporates Flickr, YouTube, and google trends.
    - Additional social media sources are also being examined.



CVADA

Center for Visual Analytics and Data Analytics

A U.S. Department of Homeland Security Center of Excellence

VACCINE and Visual Analytics Solutions

David Ebert, Kaethe Beck

Purdue University

VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

May 2014

Who We Are:
International Team of Experts
75+ Faculty, 26 Institutions



- **Purdue University**
- **Georgia Institute of Technology**
- **Pennsylvania State University**
- **Stanford University**
- **University of North Carolina at Charlotte**
- **University of Washington**
- **Arizona State University**
- **Simon Fraser University**
- **University of British Columbia**
- **Justice Institute of British Columbia**
- **Ontario Institute of Technology**
- **Dalhousie University**
- **University of Houston, Downtown**
- **Virginia Tech**
- **Indiana University**
- **Florida International University**
- **University of Texas at Austin**
- **Morgan State University**
- **Navajo Technical College**
- **University of Stuttgart**
- **University of Swansea**
- **Oxford University**
- **University of Calgary**
- **University of Manitoba**
- **Carleton University**
- **University of Victoria**

VACCINE

Some of Our Engaged End-Users

- Federal Operating Components:
  - US Coast Guard
  - US Transportation Security Agency
  - US Citizenship and Immigration Service
  - US Federal Emergency Management Agency
  - US Customs and Border Protection
  - US CERT
- Law Enforcement
  - Over 40 local and state agencies (IN, IL, OH, PA, NC, NY)
- State Fusion Centers
  - Ohio (SAIC), Indiana (IIFC)
- Companies
  - Motorola, Kimberly Clark, Banfield, Bank of America





VACCINE: What We Do – Visual Analytic Solutions

We enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable user to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

Visual Analytic Solutions: What We Offer

- **Improved Effectiveness**: We enable users to be more effective through innovative, interactive visualization, analysis, and decision making tools
  - Provide the right information, in the right format, within the right time to solve the problem
  - Enable user to be more effective from planning to detection to response to recovery
  - Enable effective communication of information
- **Innovative Fielded Solutions:** We provide innovative visual analytic and scalable solutions to the extended homeland security community
- **People and Partnerships**
  - Interdisciplinary world-leading team of researchers and students

"cgSARVA has proven its worth time and again, providing key analytic information for decision makers for large scale projects…"

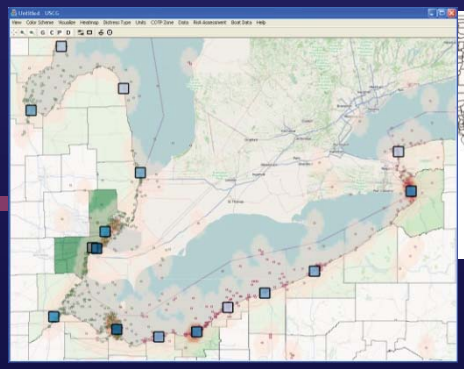
VADM Robert Parker, 2012 MRS Keynote Address

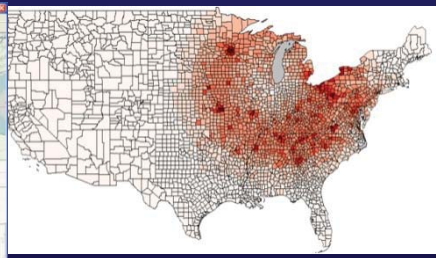
VADM Robert Parker with VACCINE student researchers (cgSARVA, COAST, iOPAR)

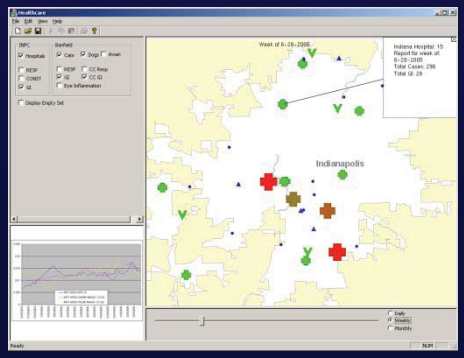


Public Safety & Health

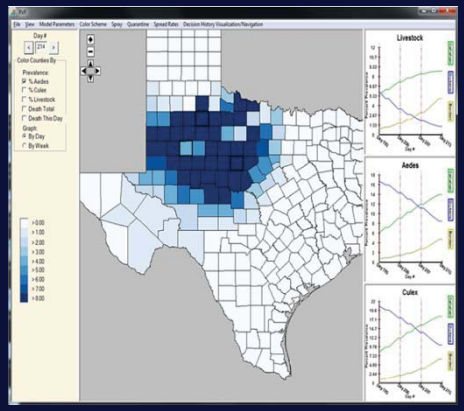
- Public safety visual analytics
  - cgSARVA
    - Coast Guard Search And Rescue Visual Analytics
  - MERGE
    - Mobile Emergency Response Guide
- Public health visual analytics
  - LAHVA
    - Linked Animal-Human Health Visual Analytics
  - RVF
    - Rift Valley Fever
    - Decision support environment for epidemic modeling and responses
  - PanVis
    - Pandemic influenza modeling and visualization tool
  - Cancer Care Engineering











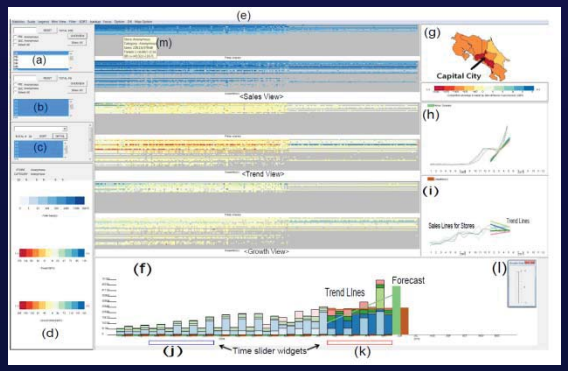


VACCINE

Business Visual Analytics

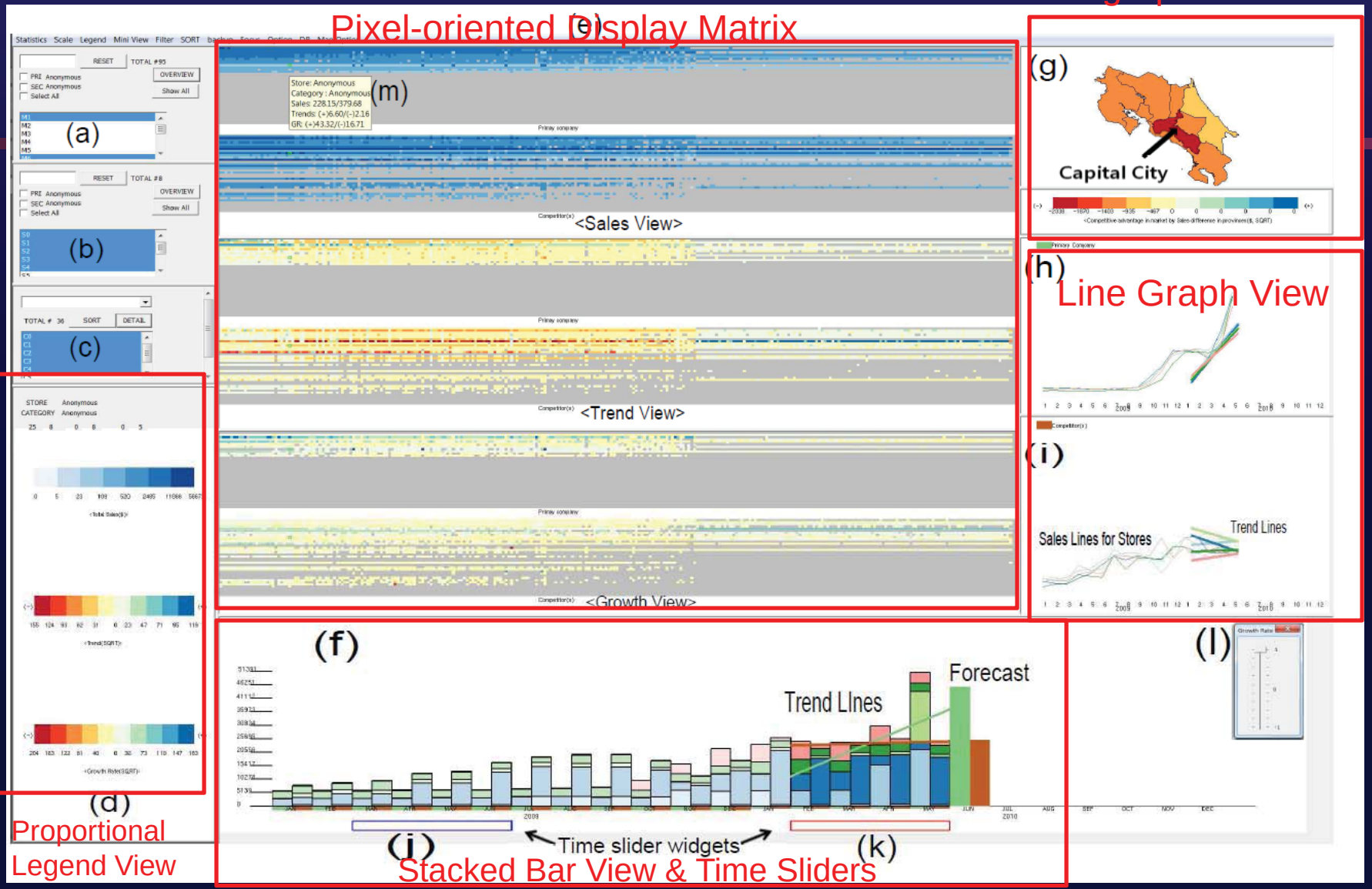
- Risk-based decision making and resource allocation
  - Coast guard operational risk assessment model
  - Helping to prioritize efforts to minimize risk
- Competitive Intelligence
  - Visual analytics system for business intelligence
- EconVIS
  - Visual analytics in various economic problems
  - Improving decision making and identifying key motivations in knowledge creation







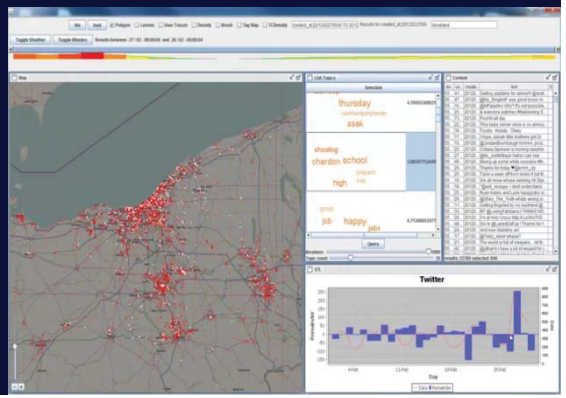
Market Analyzer



Visual Analytics Uses for Public Safety

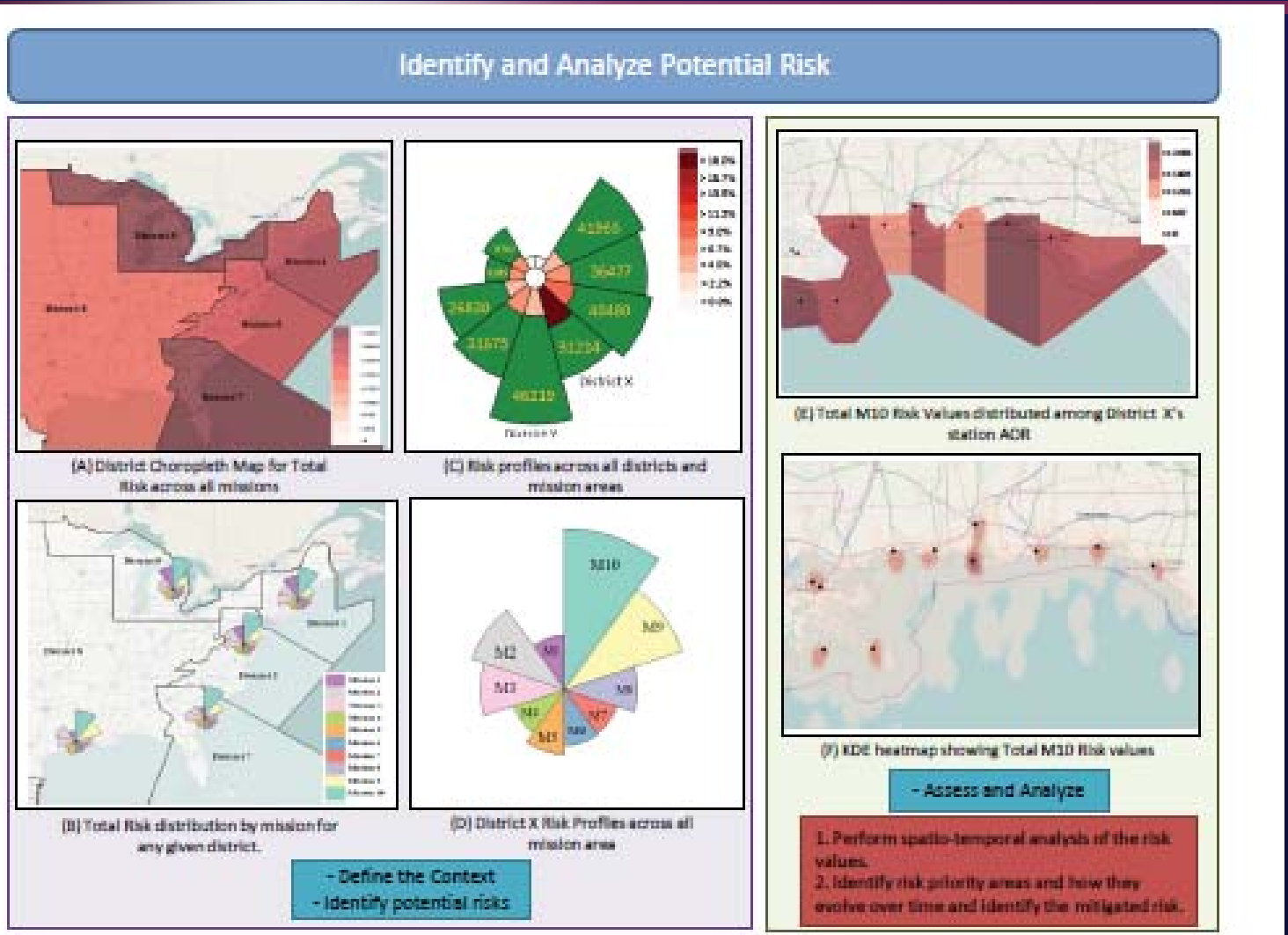


- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Data-driven policy analysis
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Multisource, multimedia massive data integration & analysis
- Purpose: Planning for resiliency, training, detection, investigation, response, recovery, remediation



VACCINE

Multi-Visualization, Multivariate VA for Risk-Based Decision Making

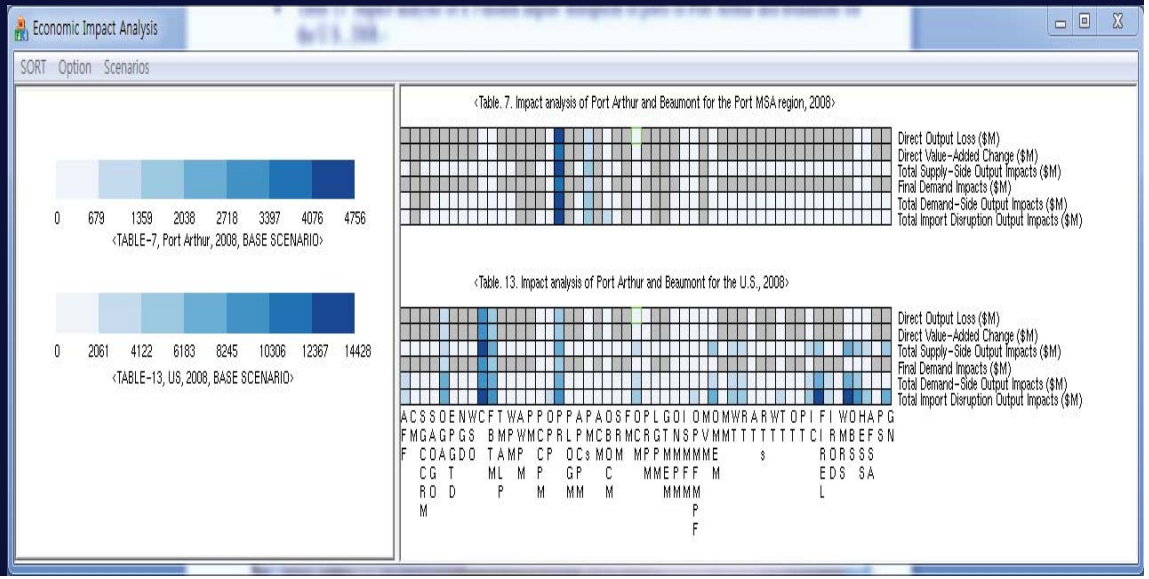


USCG Port Closure Economic Impact VA

Partners: USC CREATE, USCG RDC, USCG D7, USCG LANT

IMPACT:

- Provided tool for use analysis and planning for impact of port closure in Port Arthur, Tx
- Economic sector impact, local and national impact
- Impact and effectiveness of alternative mitigation strategies



VACCINE

Effective Risk-based Decision Making and Resource Allocation Visual Analytics

- Evaluate current and historical mission area:
  - Demands
  - Risks (total, mitigated, residual)
  - Resource allocation
  - Return on investment
- Evaluate courses of action
- Evaluate above at both Strategic and Tactical/Operational level

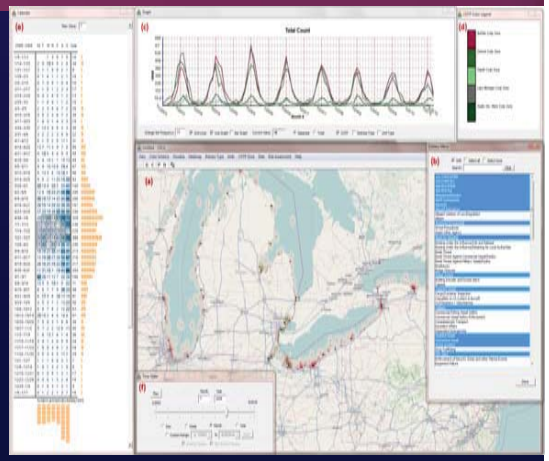


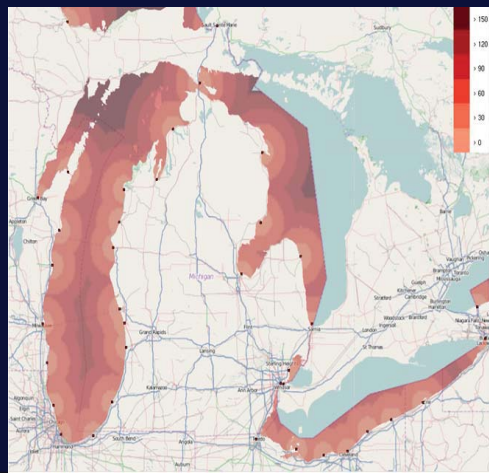
U.S. Coast Guard Search and Rescue VA (cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771

IMPACTS:

- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Hurricanes Sandy and Irene resource allocation decisions based on cgSARVA analysis and visualization
- Informed Commandant's budget testimony to Congress
- Key component of USCG D9 reallocation plan for 2011-12
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ

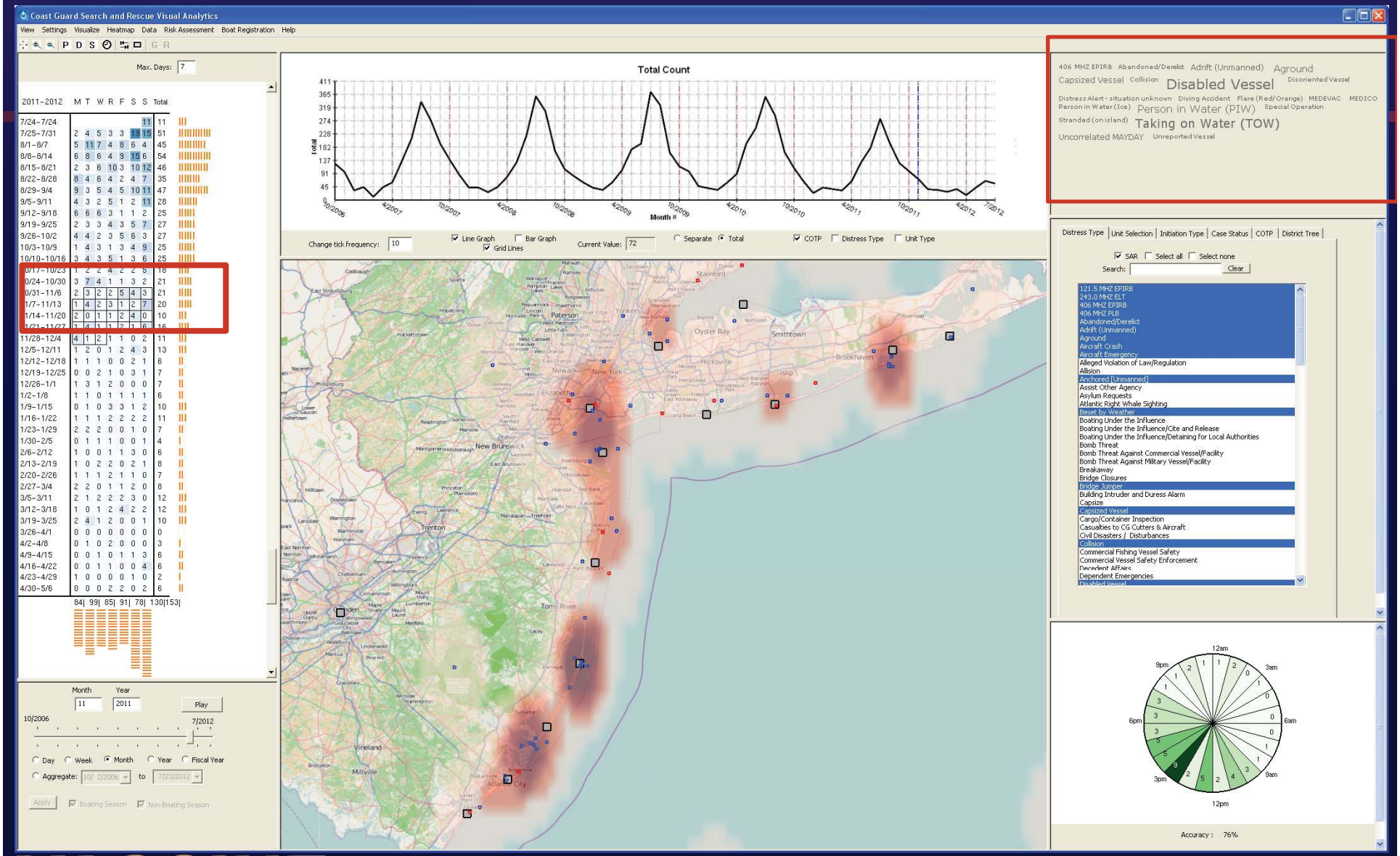




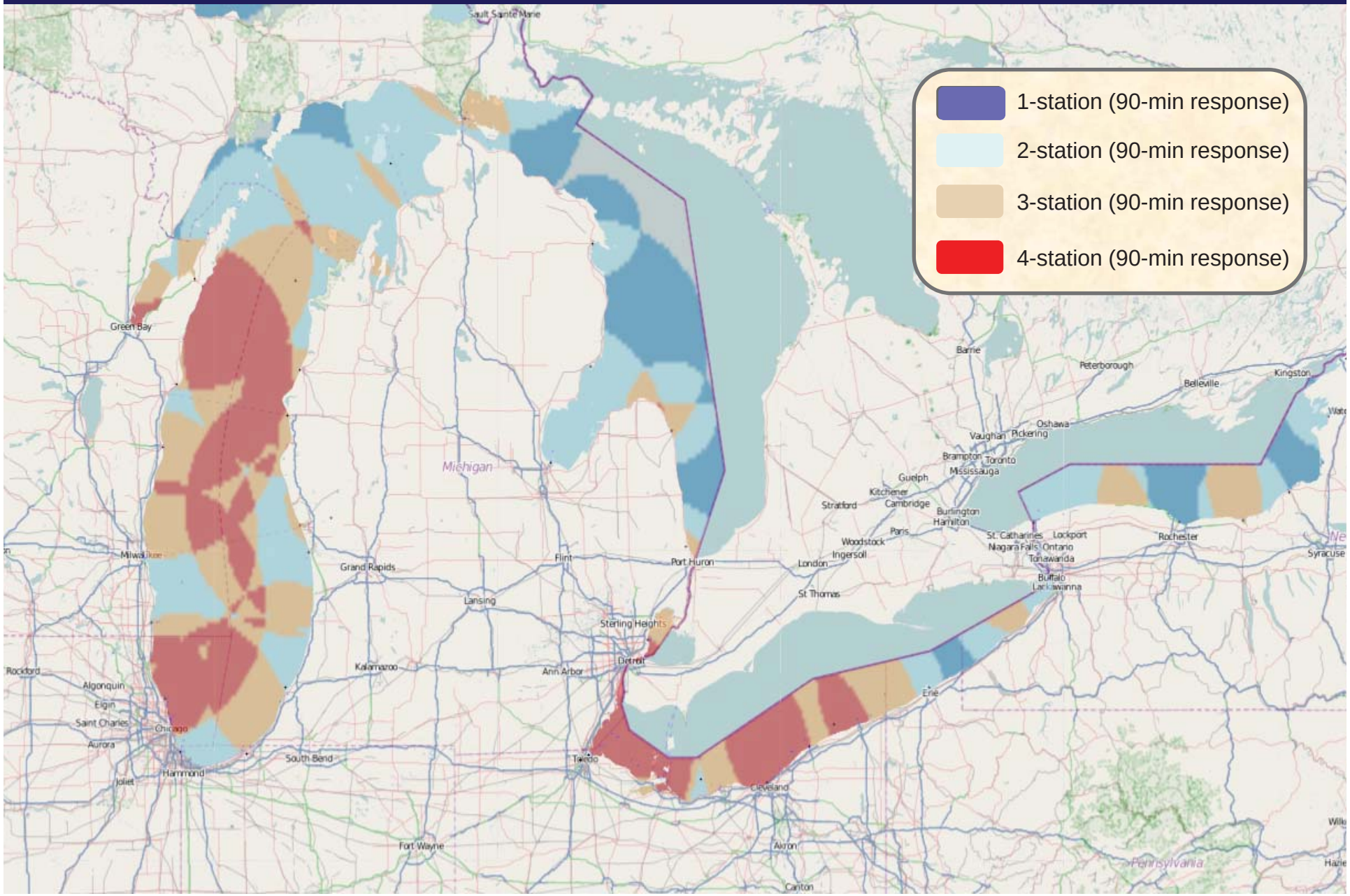
VACCINE

Example: Risks and Consequences From Sandy:

SAR Cases November 2011 NJ/NYC Area



Response Efficiency – Potential Future Assets



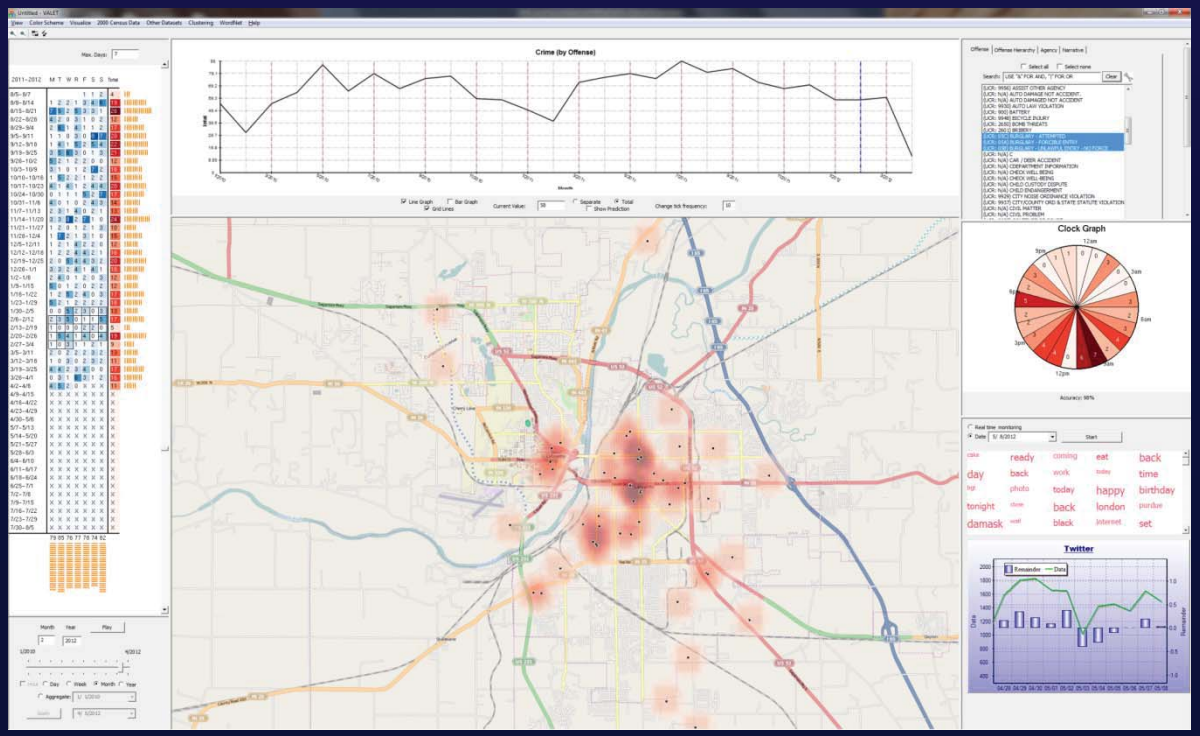
Interactive Operational Performance Assessment Report (iOPAR)

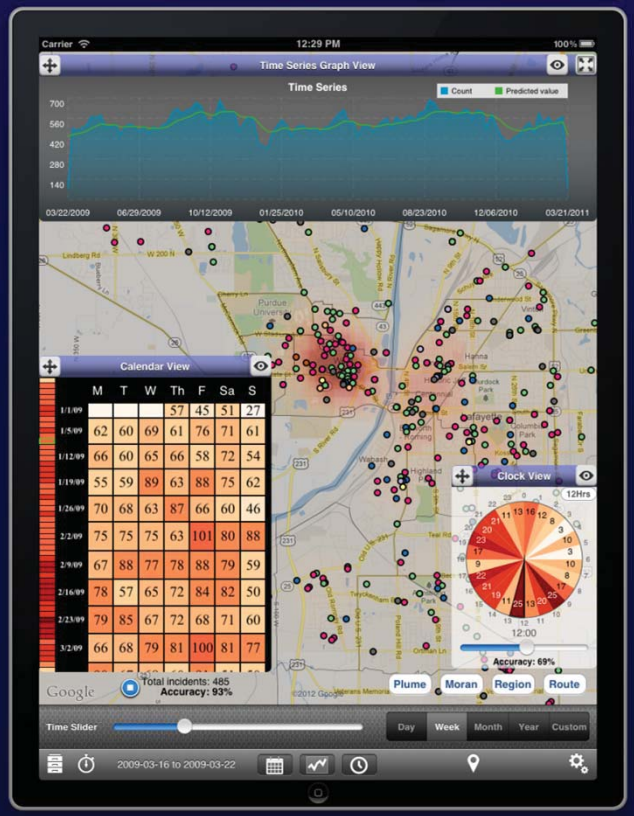
- OPAR
  - Standard report within the Coast Guard Business Intelligence (CGBI) system
  - Displays resource use and performance by core CG mission areas
- iOPAR – interactive iPad version to allow interactive visualization and analysis and inform decision making
  - Anaylze/visualize performance, targets, seasonal trends, predictions by boat, aircraft, cutter type, region levels

VACCINE Spatiotemporal VA

- Coast Guard visual analytics
  - Search and rescue
  - Risk-based resource allocation
  - Risk-based operational performance
- Anomaly detection
- Situational awareness
- Predictive policing

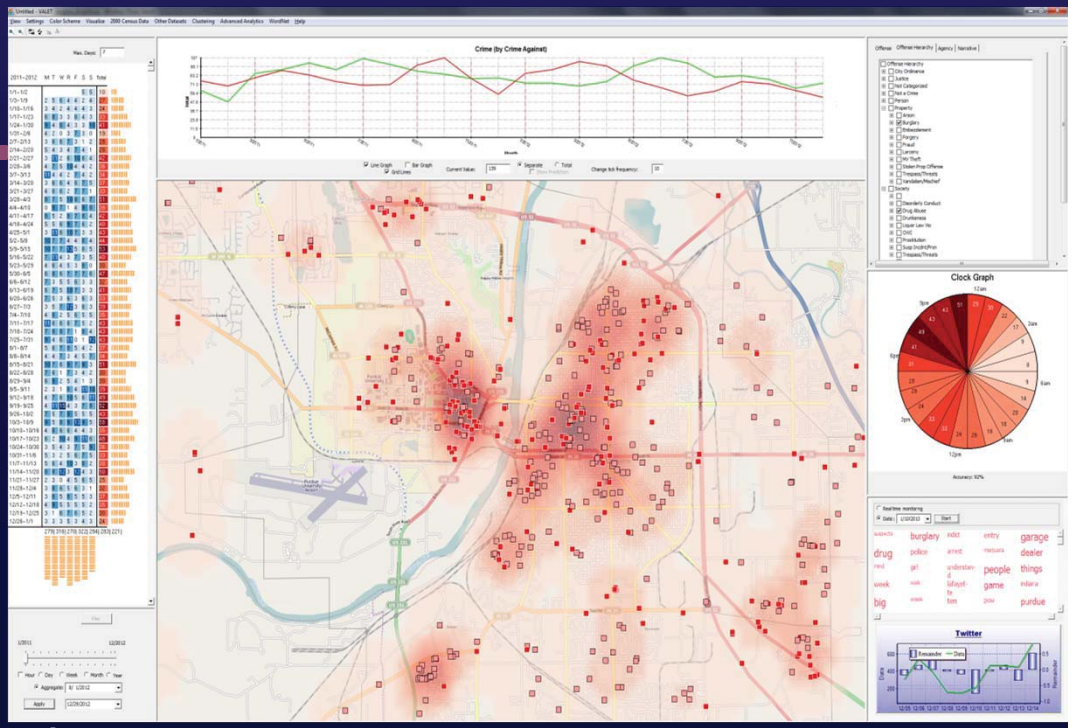
Visual Analytics Law Enforcement Toolkit (i)VALET





VACCINE

Visual Analytics Law Enforcement Toolkit (VALET, iVALET)



Impacts:

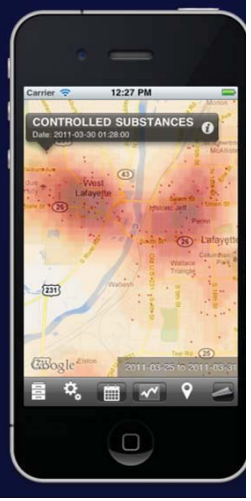
- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public (February 2013) for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

VALET delivered:

- Spring 2011: WL, Lafayette Police

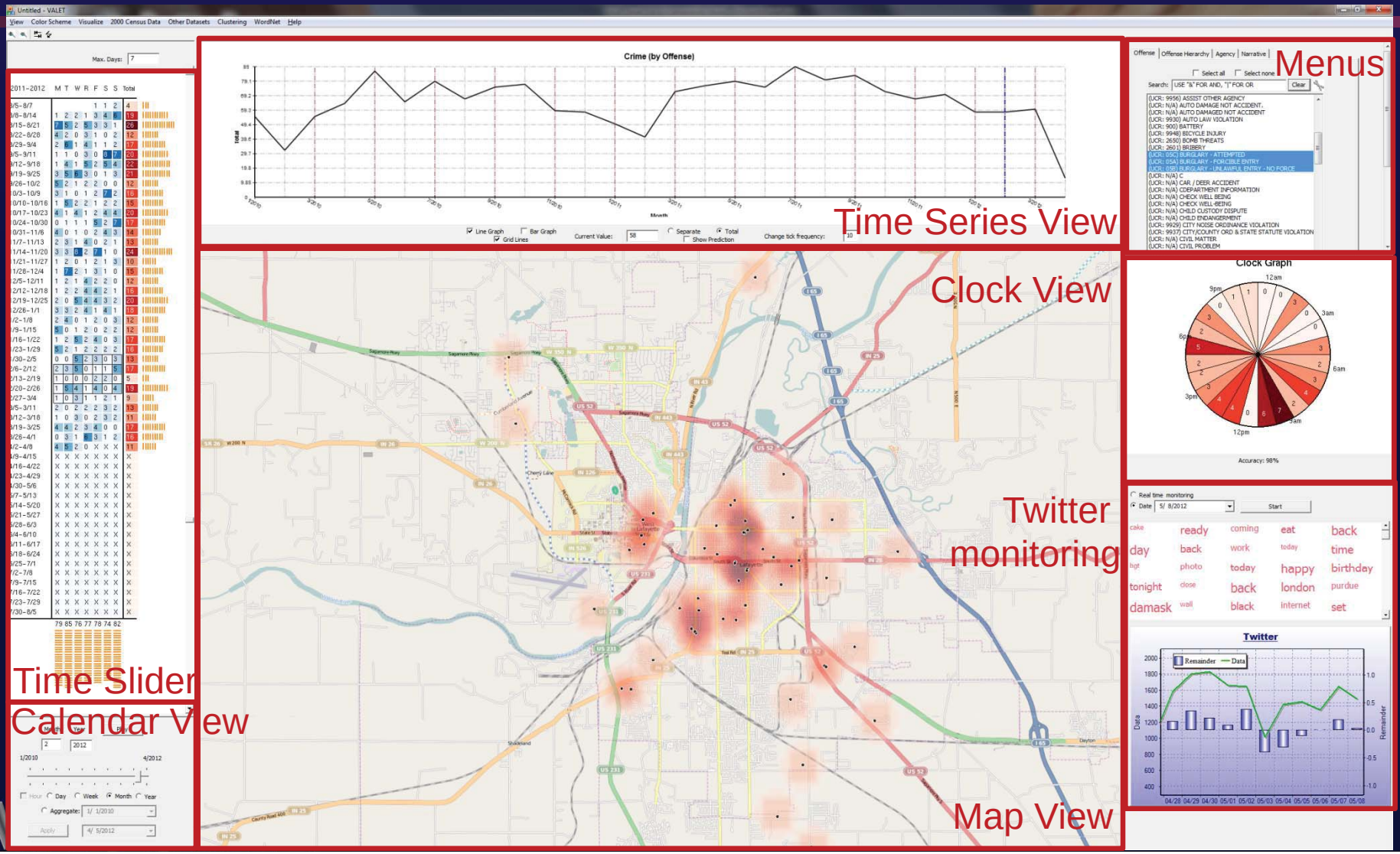
iVALET delivered:

- October 2011: Purdue, WL Police

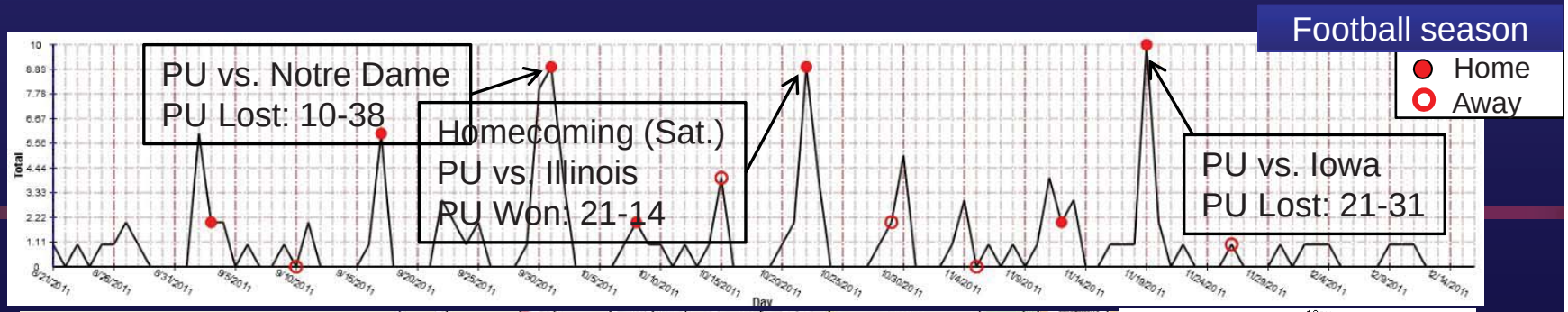


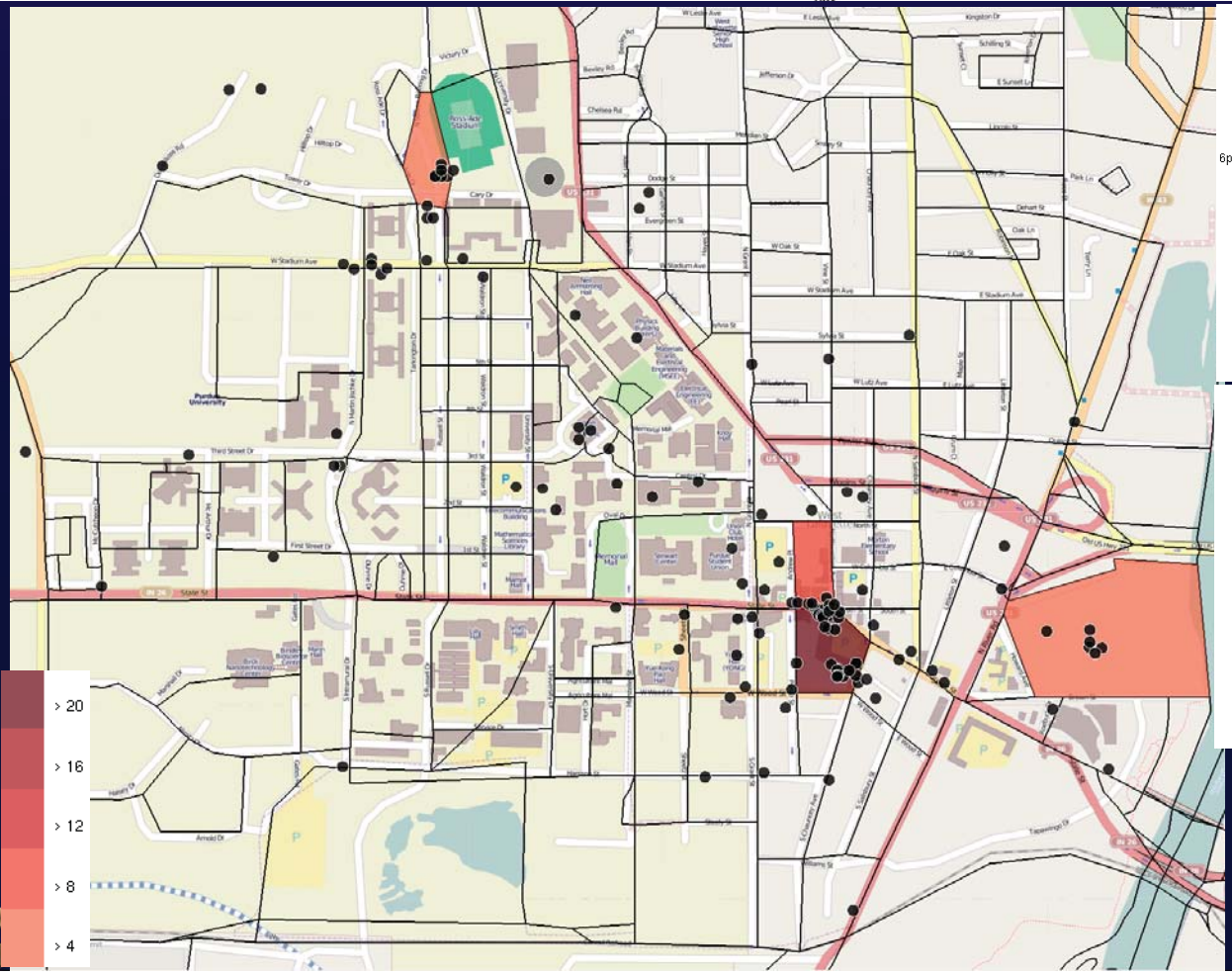
VACCINE

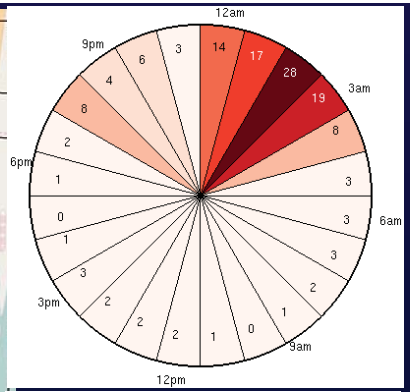
Example for Crime Analytics: VALET

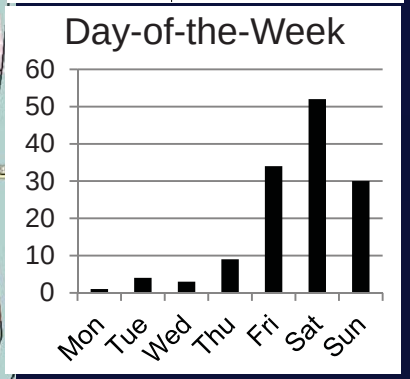


Example: Drunkenness / Public Intoxication



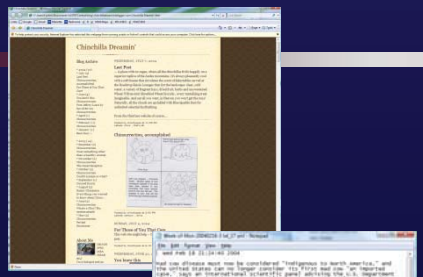


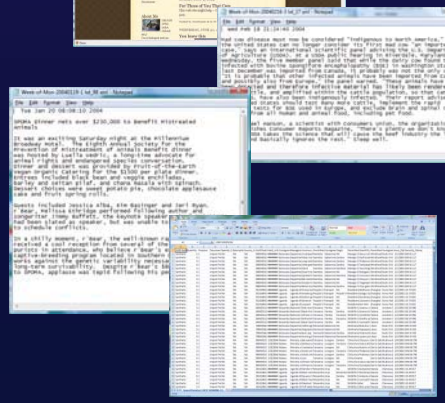




Jigsaw: Visual Analytics for Investigative Analysis

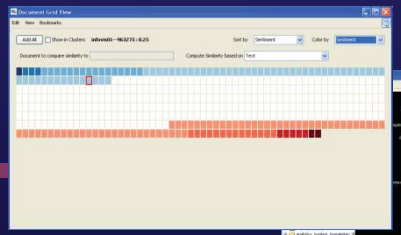
John Stasko

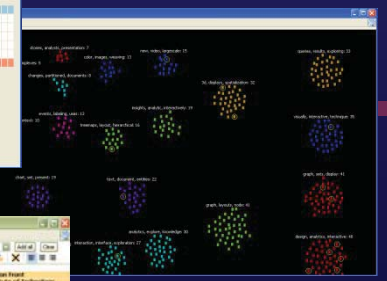


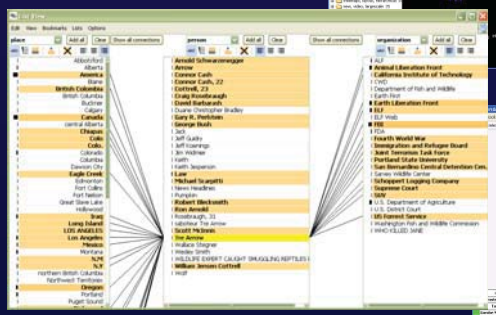


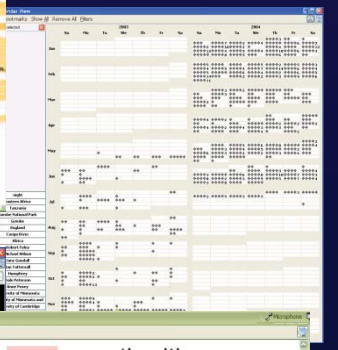
Documents in

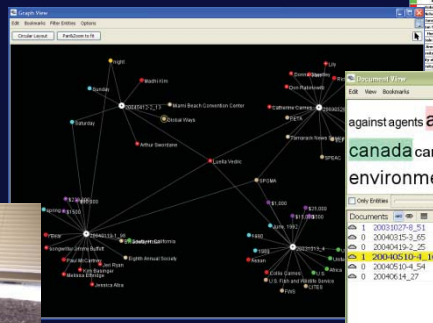
Knowledge out

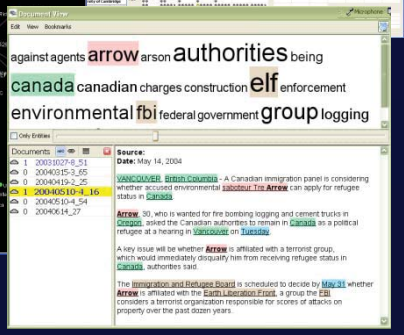


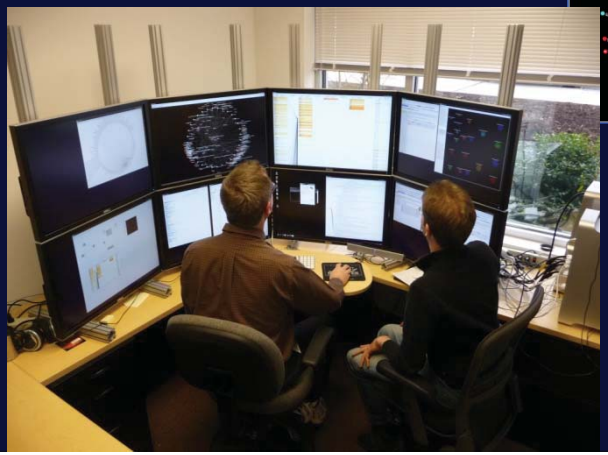








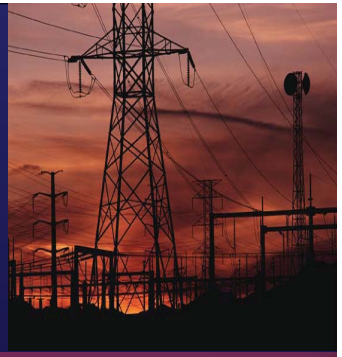




VACCINE

Georgia Tech College of Computing
School of Interactive Computing

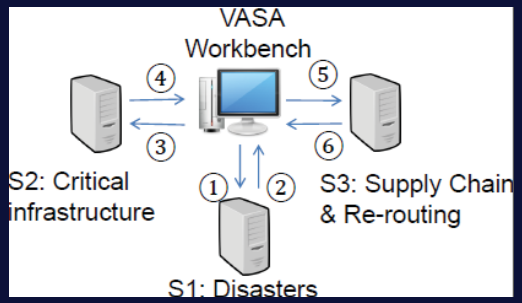
Cascading Critical Infrastructure Resiliency Modeling and Analytics (VASA)



- Purpose: Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.
- Collaborating Institution(s):
  Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Texas (TACC), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities
- End-User(s): Power Suppliers (e.g., Duke Energy), Cyber Community (e.g., Cisco), Quick Service Restaurants and suppliers, food supply

Purdue

UNCC



U Minn

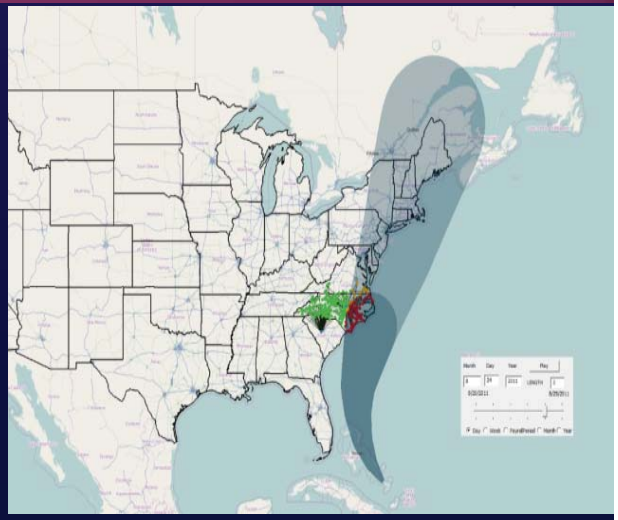
TACC

VACCINE

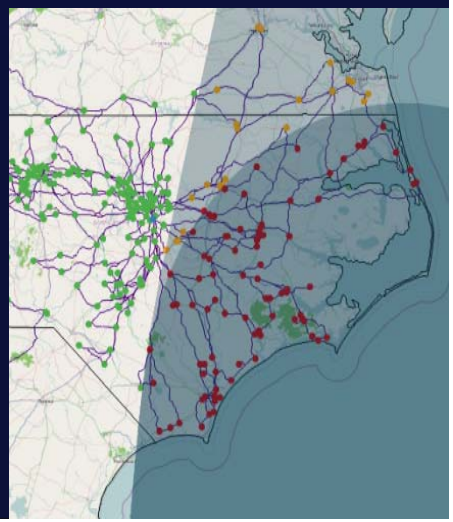
VASA: Visual Analytics for Simulation-Based Action

Collaborating Institution(s): Purdue, Minnesota, UTexas, UNCC + German universities
End-User(s): Fast-food restaurant chain, emergency management and planning personnel

Impacts and Accomplishments:

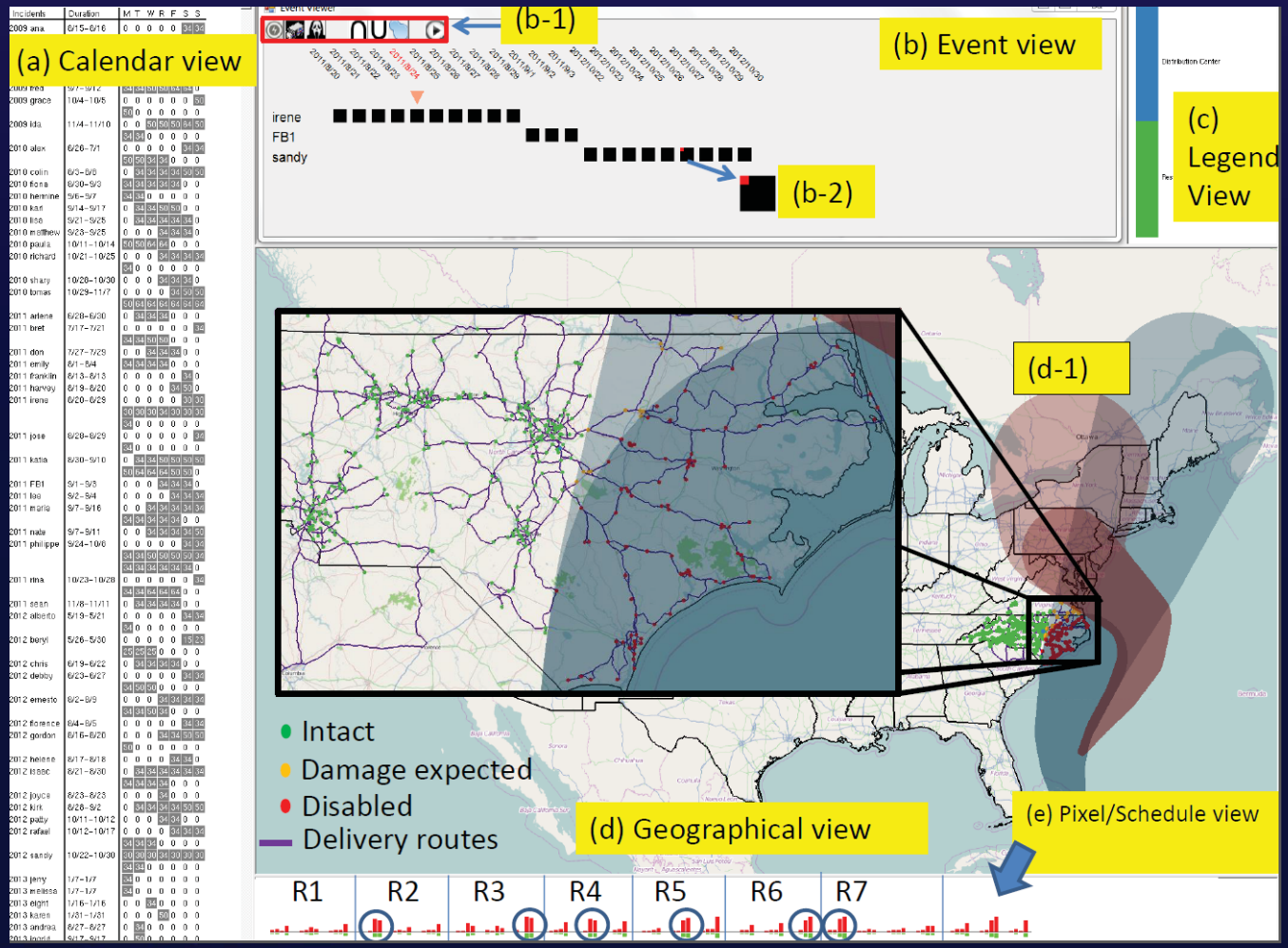


- Support decision-making for extreme weather and natural disaster scenarios
  - Combine real and simulation data
  - Allow “what-if” exploration
- **System of systems**: binds together multiple simulations models from collaborators into coherent whole
  - **Minnesota**: food distribution model
  - **Texas**: simulated and historical weather (hurricanes, storms)
  - **UNCC**: critical infrastructure
  - **Purdue**: roads + interaction visual analytics tool
- **Challenge**:
  combine interactive VA with complex simulation models

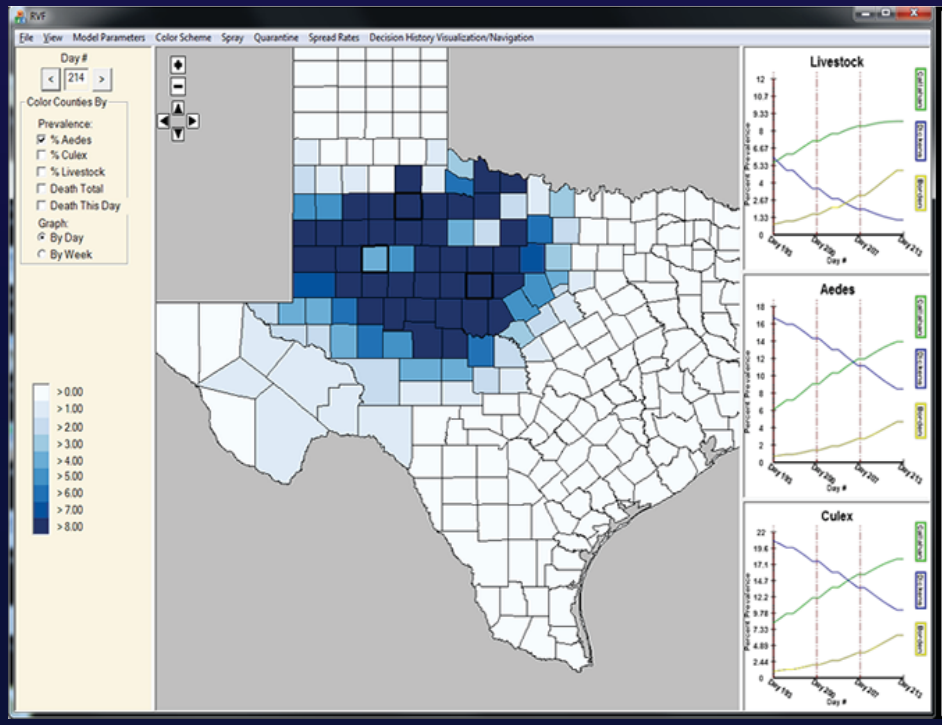


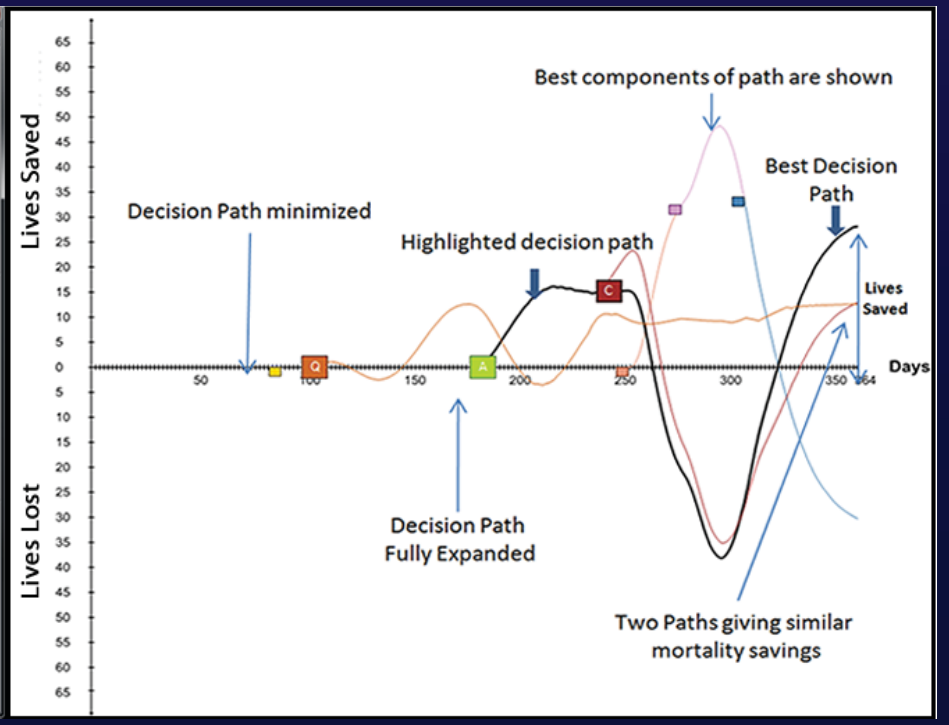
VACCINE

System Interface



Complex Decision Making: Advanced Decision Support Tools (e.g., Rift Valley Fever)

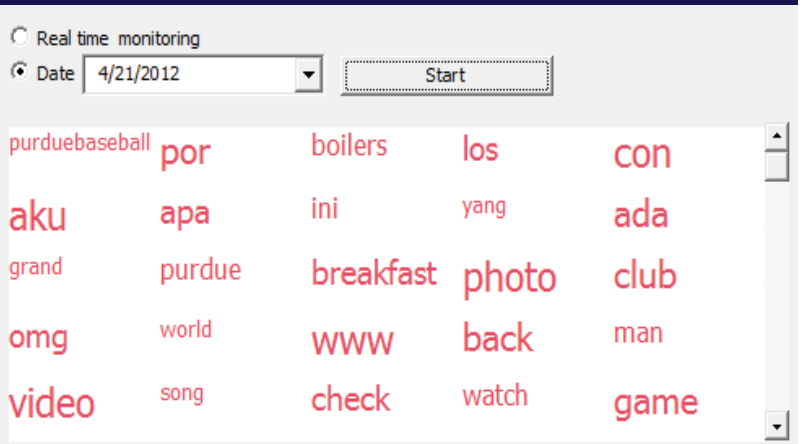


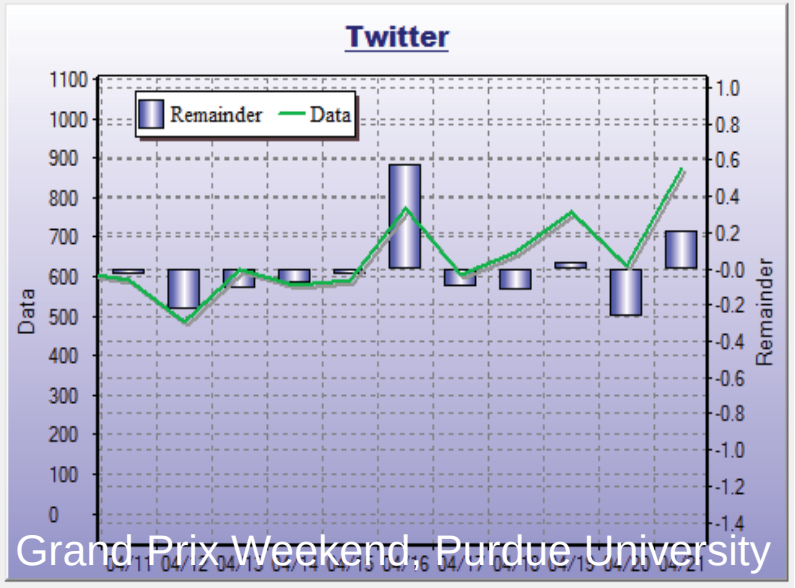


Social Media: Real-time Twitter Visual Analytics

(Purdue, Stuttgart, Penn St.)

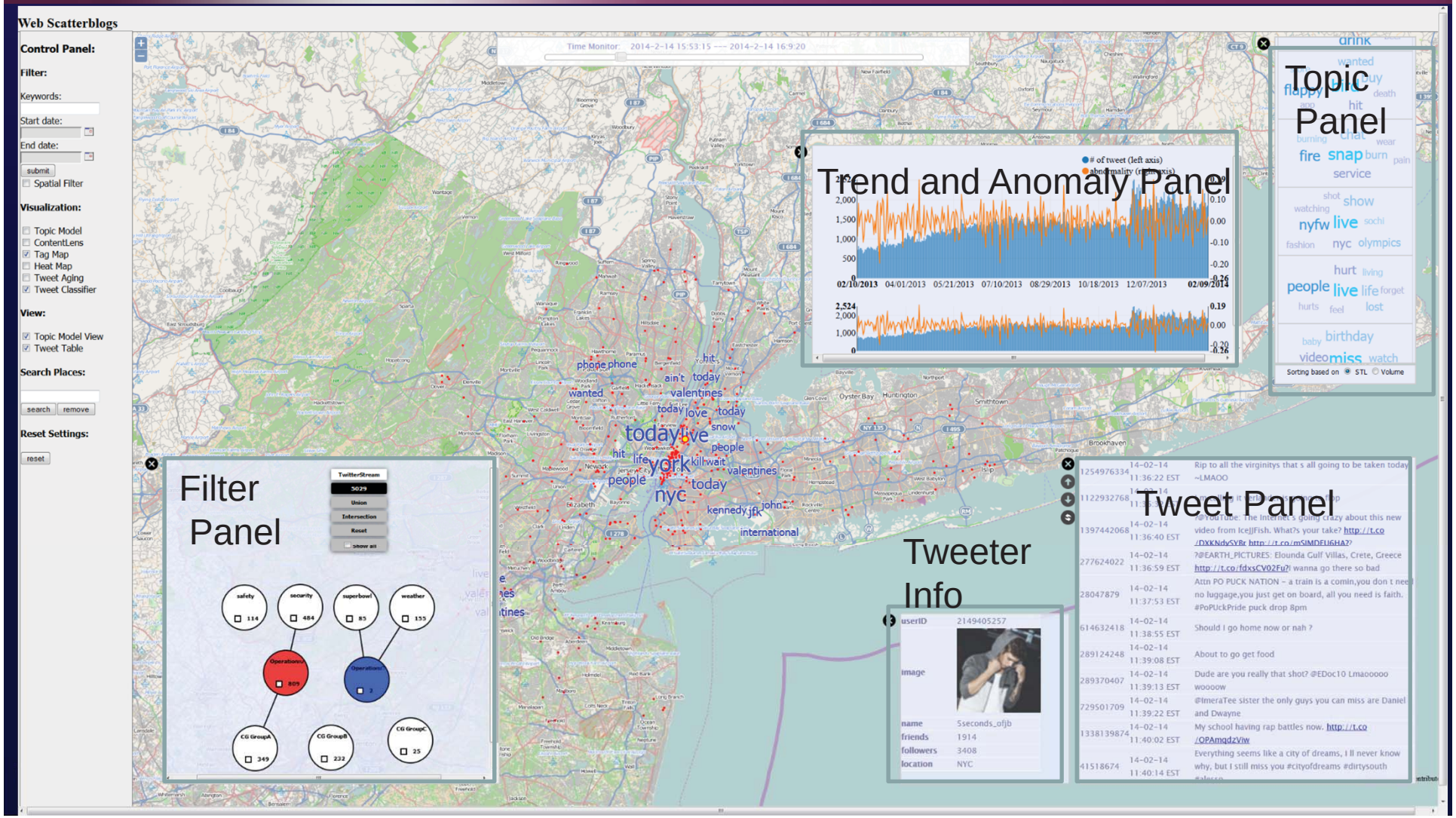
- Anomalous topic extraction using LDA and novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud



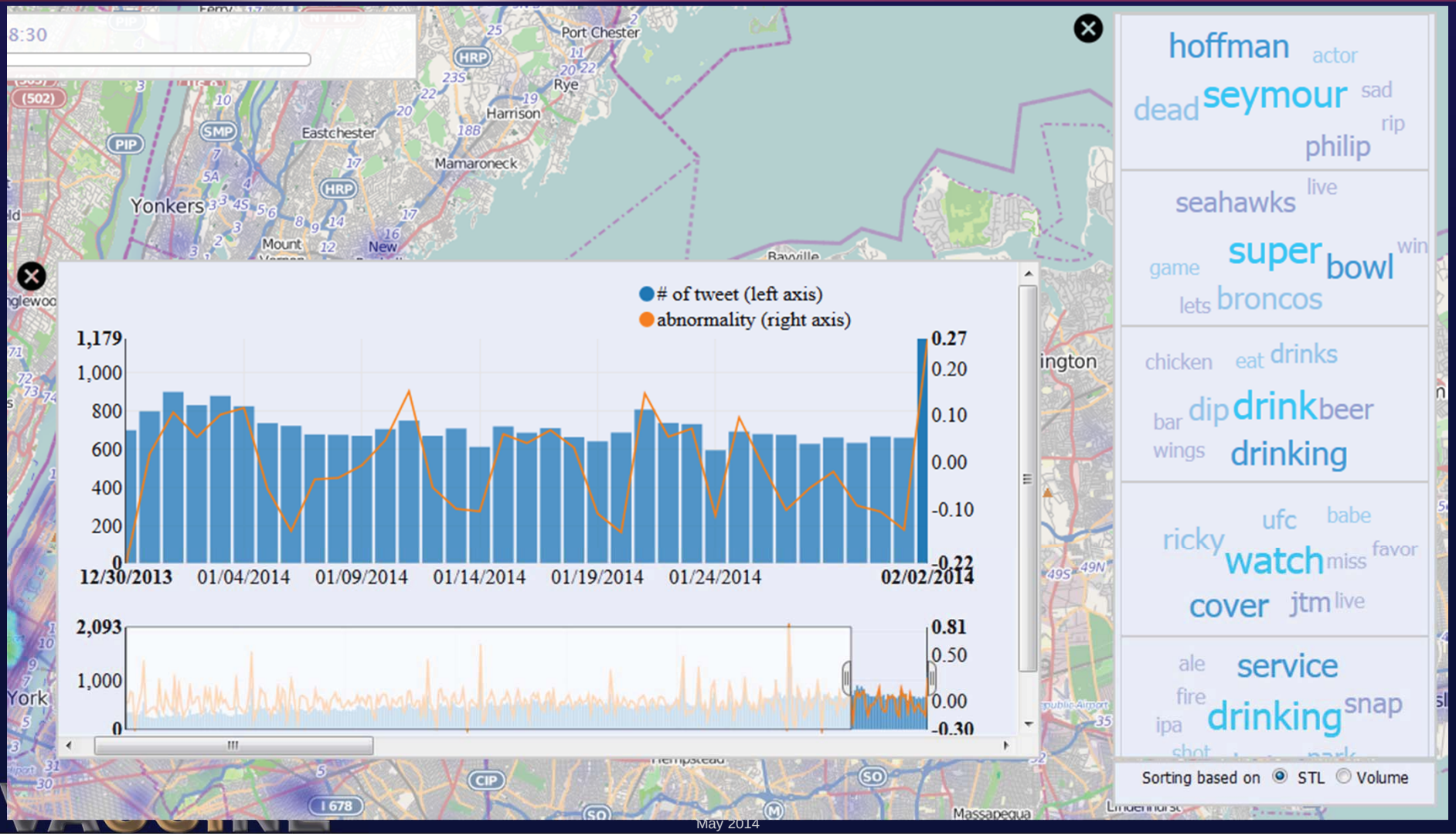


Grand Prix Weekend, Purdue University

Social Media Visual Analytics: Scatterblogs

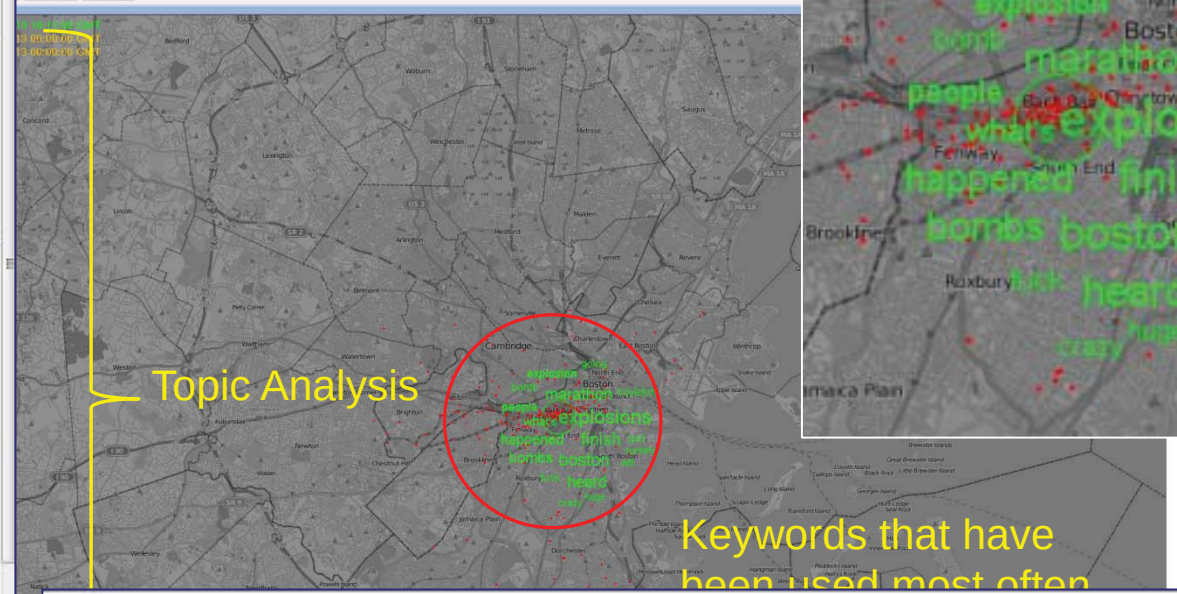


Anomaly Detection



Explosion Area in Boston







Keywords that have been used most often

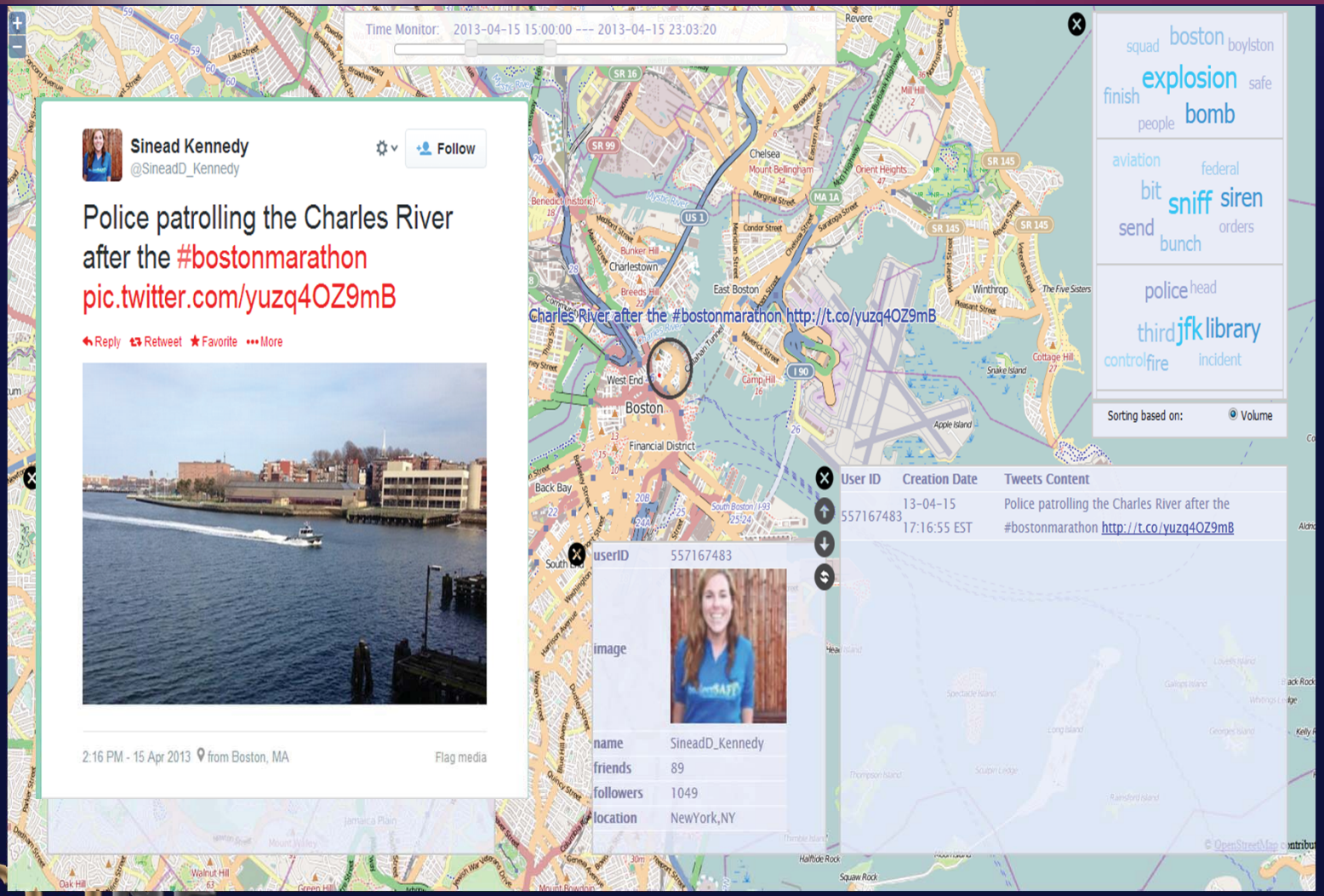
Tweets

Text
Oh my god what just happened
Something happened at the end of the Boston Marathon. Something bad and there is a lot of chatter on Twitter. What's going on?
Multiple people are injured near the Boston Marathon finish line after two explosions. The #BostonMarathon has been stopped.
two bombs just went off on boylston
BREAKING NEWS: Two powerful explosions detonated in quick succession right next to the Boston Marathon finsh line this afternoon.
What the FUCK was that
Back in Sept, @croon1 solicited me for $2000. He now has a music video with William Shatner. If you watch it (god forbid) keep that on mind.
Literally what the fuck get me out of here
@DTenenbaum my office right next to it
@FRANCESCalcIO I figured 3 people would get the joke
Two explosions just rocked the finish line of the Boston Marathon. Sirens galore. People running in fear. Wonder what happened.
This is crazy i seen that blow up #bostonmarathon
can someone tell me what that explosion was!? #boston #bostonmarathon

First Response (Tweet & Picture)



Police Patrolling The River

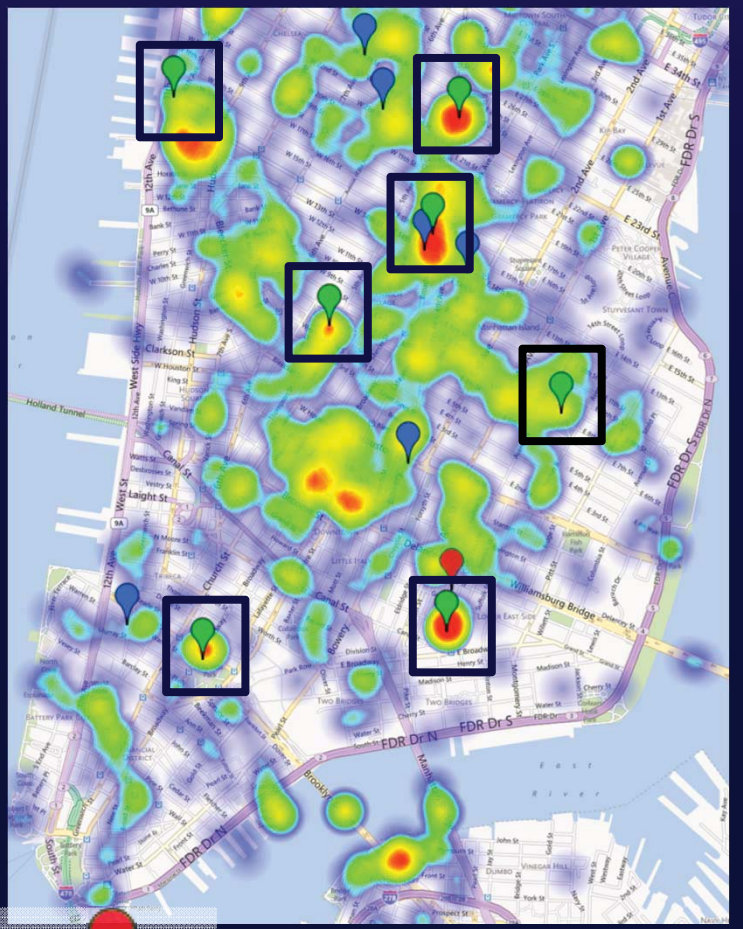


Visual Analytics of Activity During Hurricane Sandy

Two weeks before Sandy
10/14 (Sunday), 12:00 ~ 16:00

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00





Supermarket Park Shelter

VACCINE

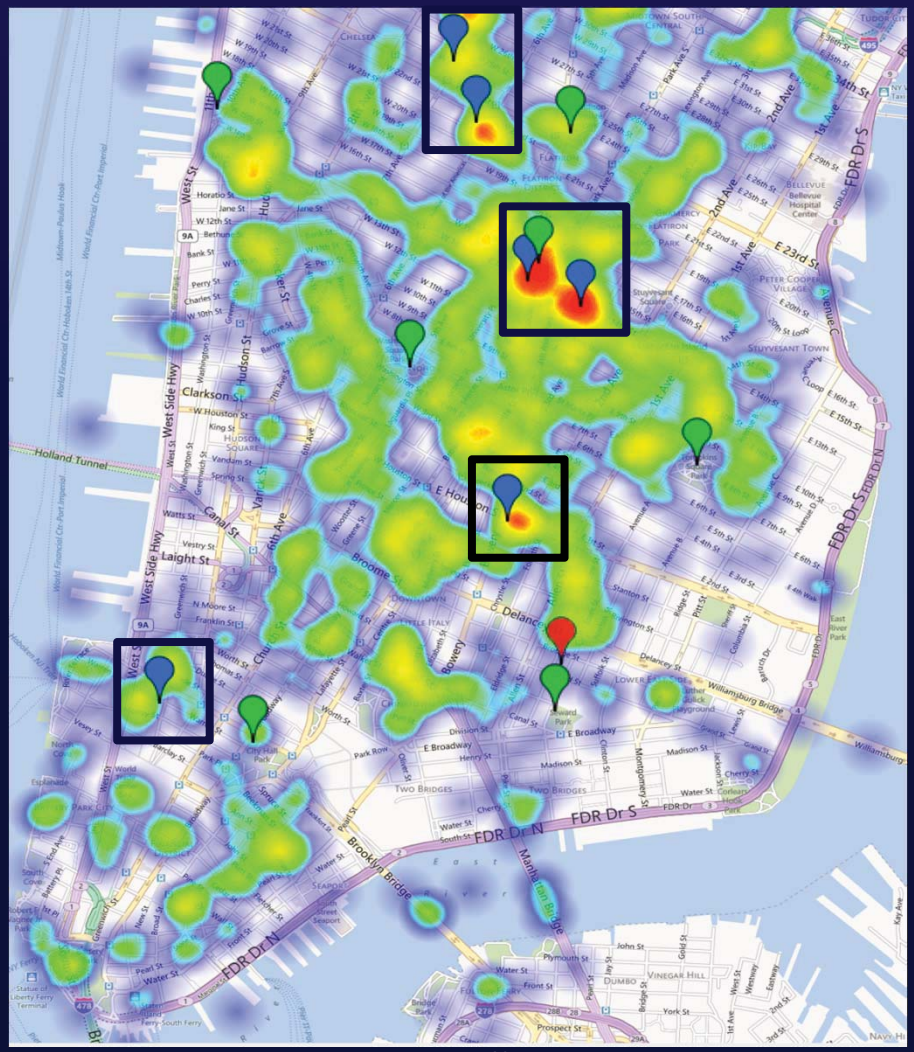
May 2014

Visual Analytics of Activity During Hurricane Sandy

After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00

Evacuation order:
10/28, 10:30 AM

Hurricane Sandy's
Arrival at NYC:
10/29, 8:00 PM



VACCINE

May 2014

For Further Information

www.VisualAnalytics-CCI.org



VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

Visual Analytics: The Power To Solve Big Data Challenges Through Interaction And Visual Reasoning

David S. Ebert

Purdue University

ebertd@purdue.edu

CVADA

Center for Visual Analytics and Data Analytics

A U.S. Department of Homeland Security Center of Excellence

May 2014

Big Data Challenges



Big Data: Solution to Global Challenges?



- Big Data is often defined along three dimensions:
  - **Volume** – size
  - **Velocity** – rate of input, update, change
  - **Variety** – different types, sources, variables
- Need:
  - Advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of information (TechAmerica Foundation)
  - **Enable effective, efficient analysis, decision making, planning, and action**

Why VA For Big Data?

(Or What Big Data Analytics Can't Do)

(inspired and adapted from David Brooks, New York Times, 2/18/2013)

- Qualitative, fuzzy, and social data
  - Preferences, significance of one relationship over another; Trust
- Context
  - Data is rarely complete, nor incorporates all relevant information Humans have extensive information and experience that never make it into the collected data
- Spurious vs. Significant
  - Big data means more statistically significant events and correlations, but they may not have any relevance
  - Increases noise to signal ratio
- Big problems
  - Complex, multifaceted, multiparameter big challenges with unquantified dependencies

What Our Visual Analytic Solutions Offer

We enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable user to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

Our Visual Analytics Process

(extended and adapted from Bertin, Normal, Tversky)

Process:

- Define problem / question
- Determine:
  - Workflow and interaction
  - Relevant data
  - Data characteristics (e.g., types, quality)
  - Abstraction/representation level
  - Natural scales
- Map to appropriate visual representation for task

- Utilize iterative refinement with user feedback from the start

Guiding Principles:

- Appropriateness principle
  - Neither more nor less
- Naturalness principle
  - Match with cognitive models
- Matching principle
  - Match the task, offer action affordances
- Principle of apprehension
  - Content should be accurately and easily perceived

Example of Users With a Problem: Law Enforcement - Tasks and Uses

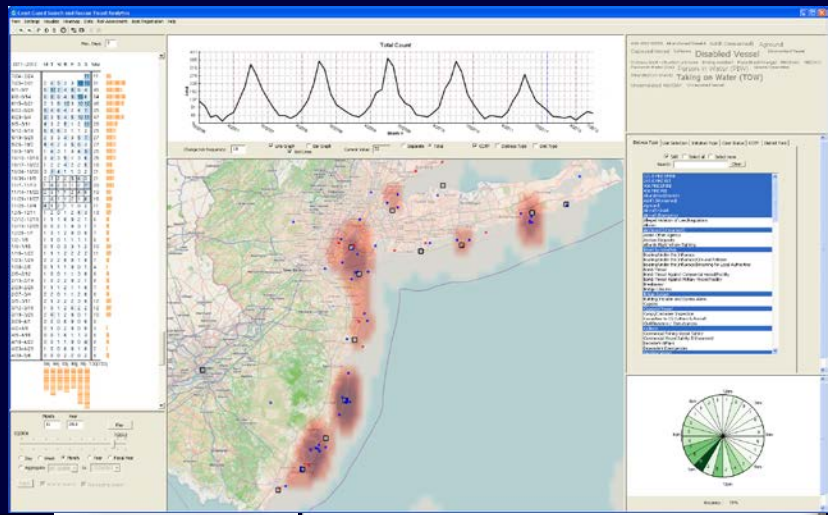
- Police Chief/Sheriff - strategic goals
  - Resource allocation, return on investment
  - Yearly effectiveness (Are we safer?)
  - Emerging trends
- Police Captain – short-term effectiveness
  - Force allocation for the day/week
  - Emerging trends and anomalies
- Officer – improved situational awareness
  - What happened near me around the same time in the past?
  - What will make me better prepared when I respond to this call?
- Crime analyst/investigator – connections
  - Correlations, dependencies, trends

Our Spatiotemporal Visual Analytics Work

- Techniques:
  - New anomaly and prediction techniques
  - New interactive correlative analysis
  - Natural scale work
  - Social medial visual analytics
- Uses:
  - Anomaly detection
  - Situational awareness
  - Predictive policing
  - Coast Guard visual analytics
    - Search and rescue
    - Risk-based resource allocation
    - Risk-based operational performance

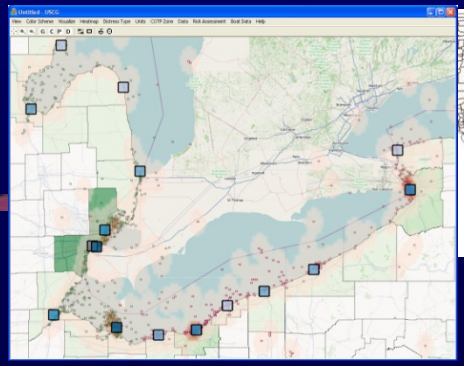
Research Area Overview and Example Projects

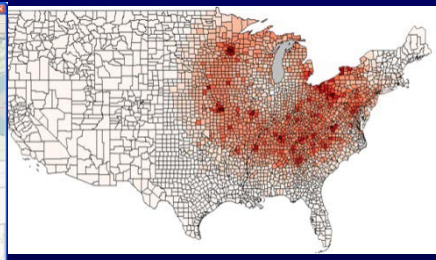
- Research applications areas
- Example projects
  - Public safety and law enforcement
  - Social media visual analytics
  - US Coast Guard
  - Additional projects

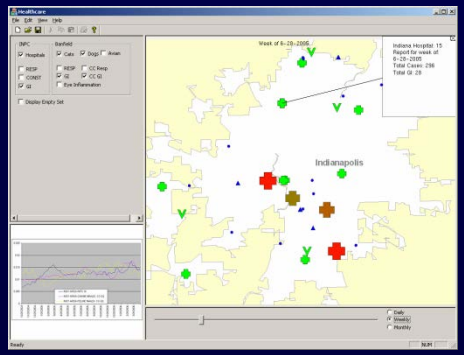


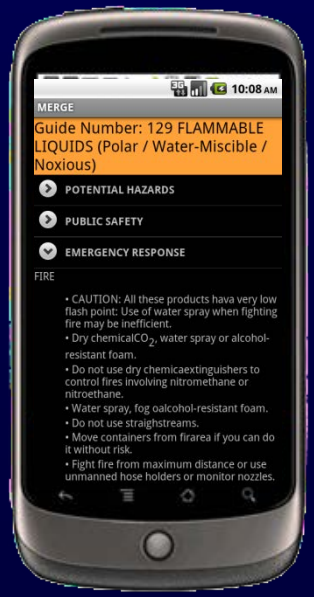
Public Safety & Health

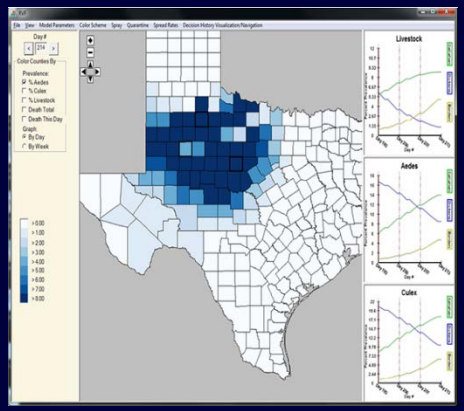
- Public safety visual analytics
  - cgSARVA
    - Coast Guard Search And Rescue Visual Analytics
  - MERGE
    - Mobile Emergency Response Guide
- Public health visual analytics
  - LAHVA
    - Linked Animal-Human Health Visual Analytics
  - RVF
    - Rift Valley Fever
    - Decision support environment for epidemic modeling and responses
  - PanVis
    - Pandemic influenza modeling and visualization tool
  - Cancer Care Engineering











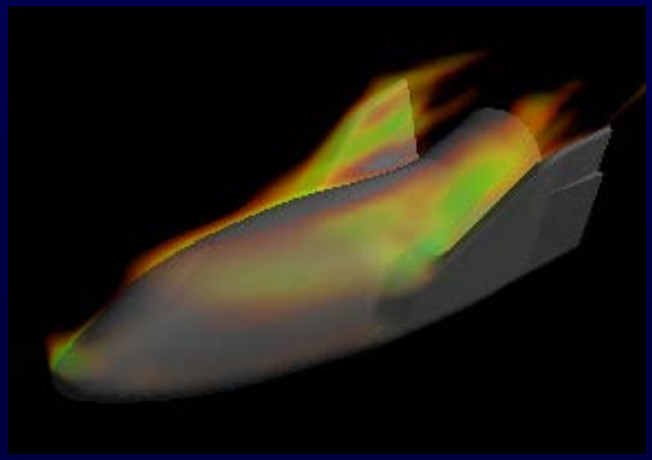


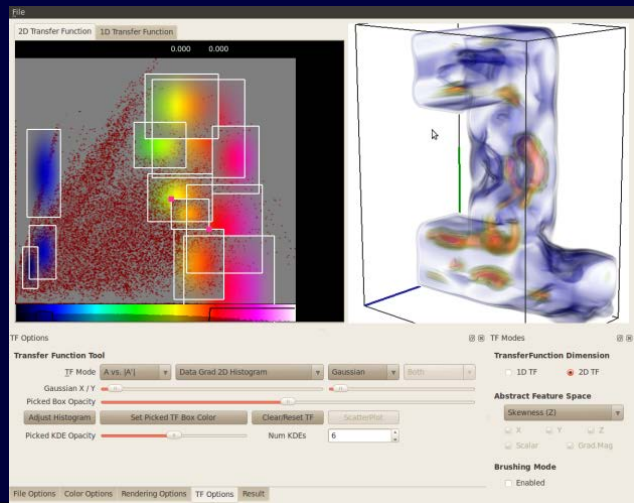
VACCINE

Scientific Discovery



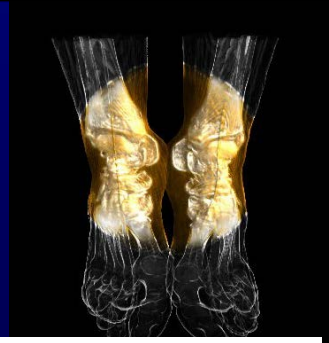
- Flow dynamics visualization (Purdue, TACC)
  - Providing insights on large flow data
  - Visualization linked with simulations
  - Innovative feature visualization
- Nanohub
  - Information-assisted data analysis and visualization of nanoelectronic models

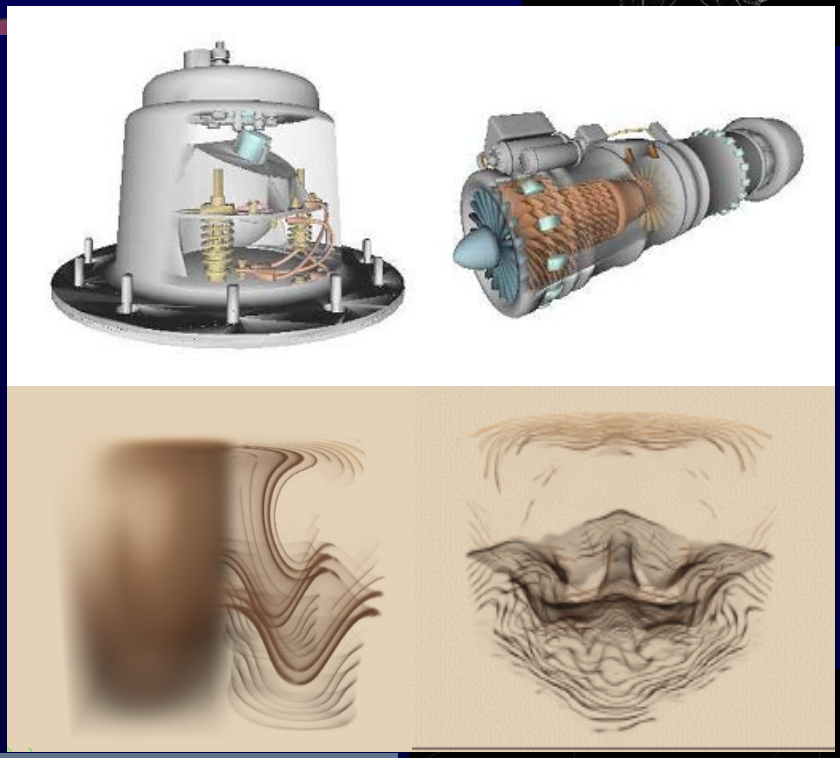


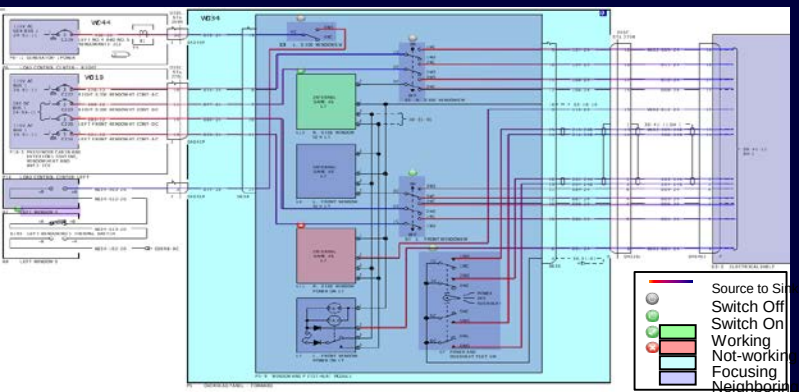


Scientific Discovery

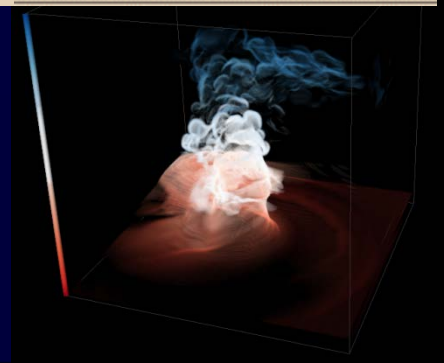
- Illustrative Visualization
  - Mechanical assemblies
  - Schematics and wiring diagrams
  - Illustrative flow visualization
  - Illustrative medical visualization
- Interactive multivariate atmospheric science data visual analytics





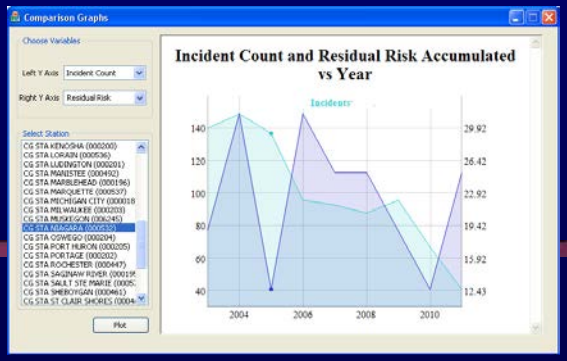


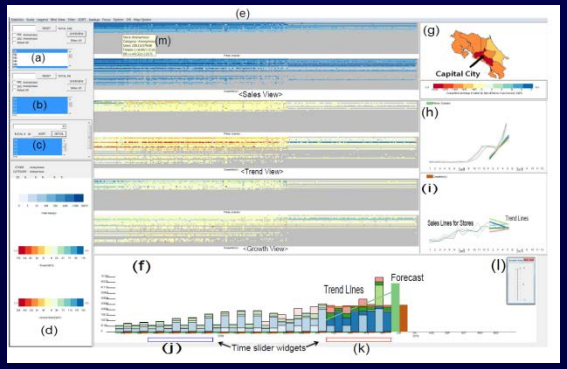




Business Visual Analytics

- Risk-based decision making and resource allocation
  - Coast guard operational risk assessment model
  - Helping to prioritize efforts to minimize risk
- Competitive Intelligence
  - Visual analytics system for business intelligence
- EconVIS
  - Visual analytics in various economic problems
  - Improving decision making and identifying key motivations in knowledge creation



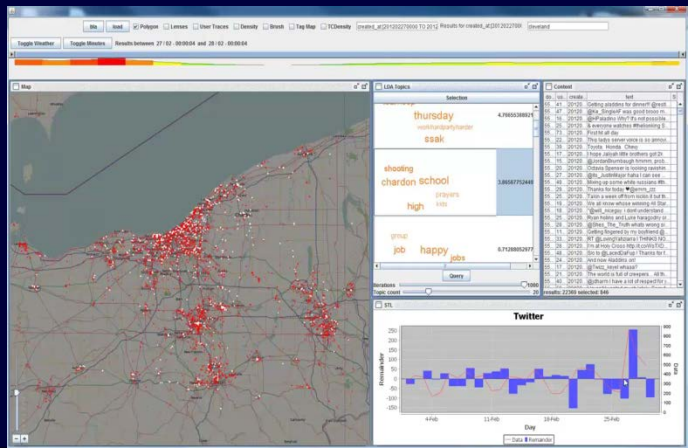




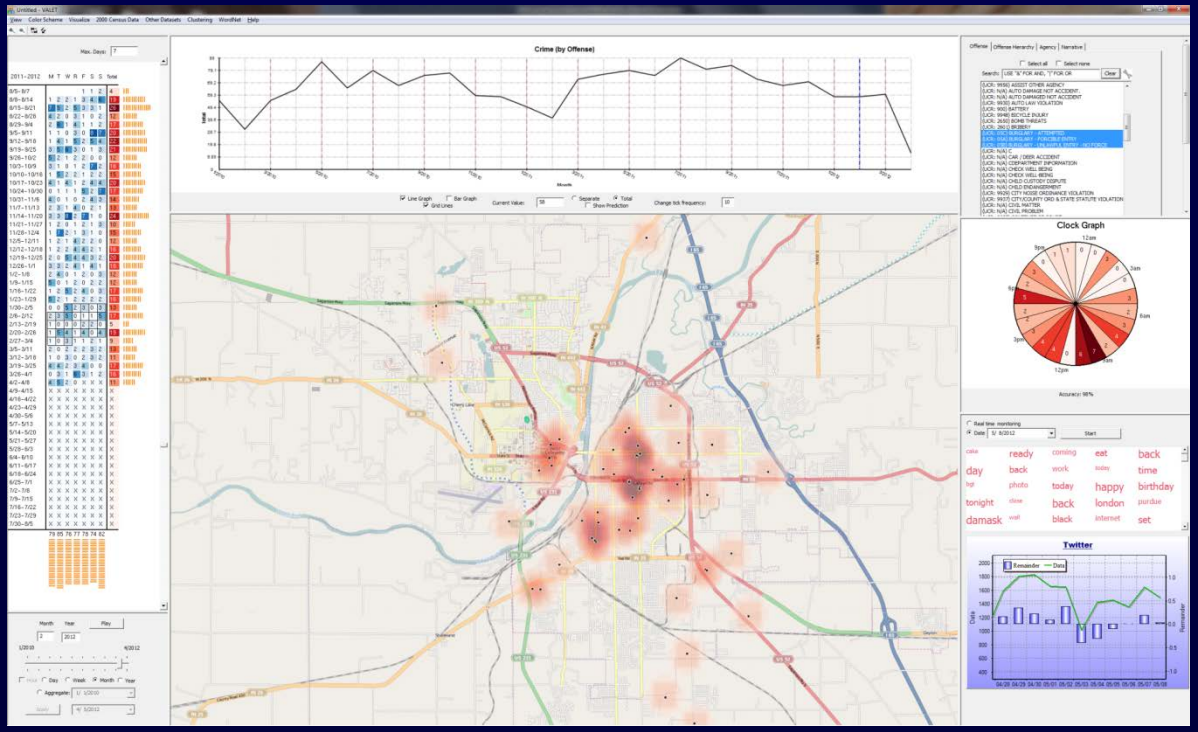
Visual Analytics Uses for Public Safety

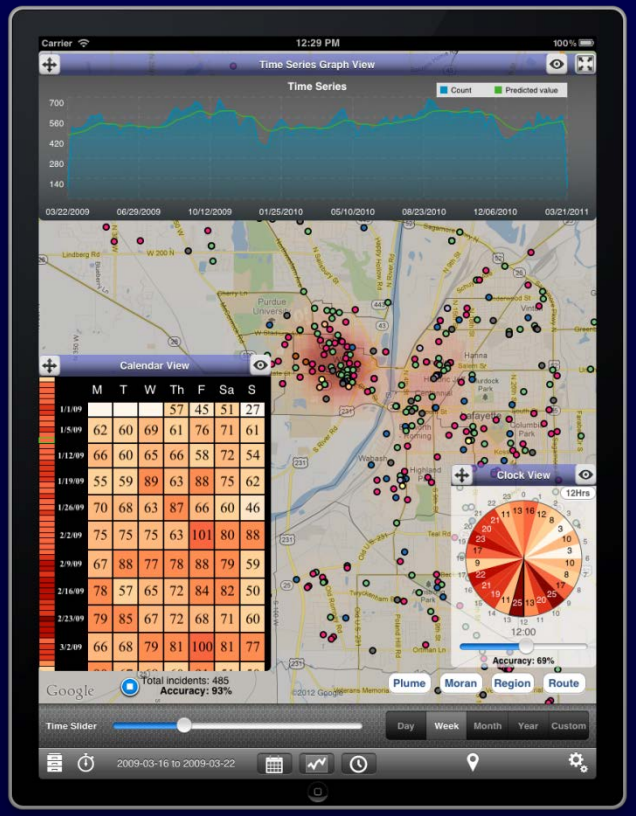


- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Multisource, multimedia massive data integration & analysis
- Purpose: Planning for resiliency, training, detection, investigation, response, recovery, remediation



Visual Analytics Law Enforcement Toolkit (i)VALET



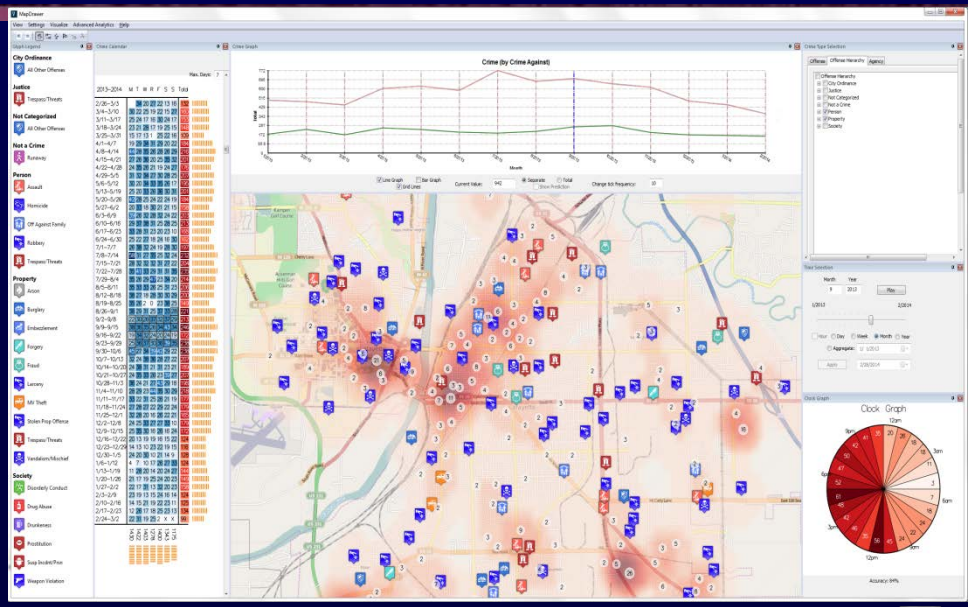


VACCINE

Visual Analytics Law Enforcement Toolkit (VALET, iVALET)

Impacts:

- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public (June 2014) for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

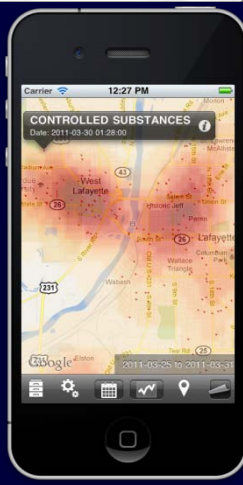


VALET delivered:

- Spring 2011: WL, Lafayette Police
- Fall 2013: NYPD

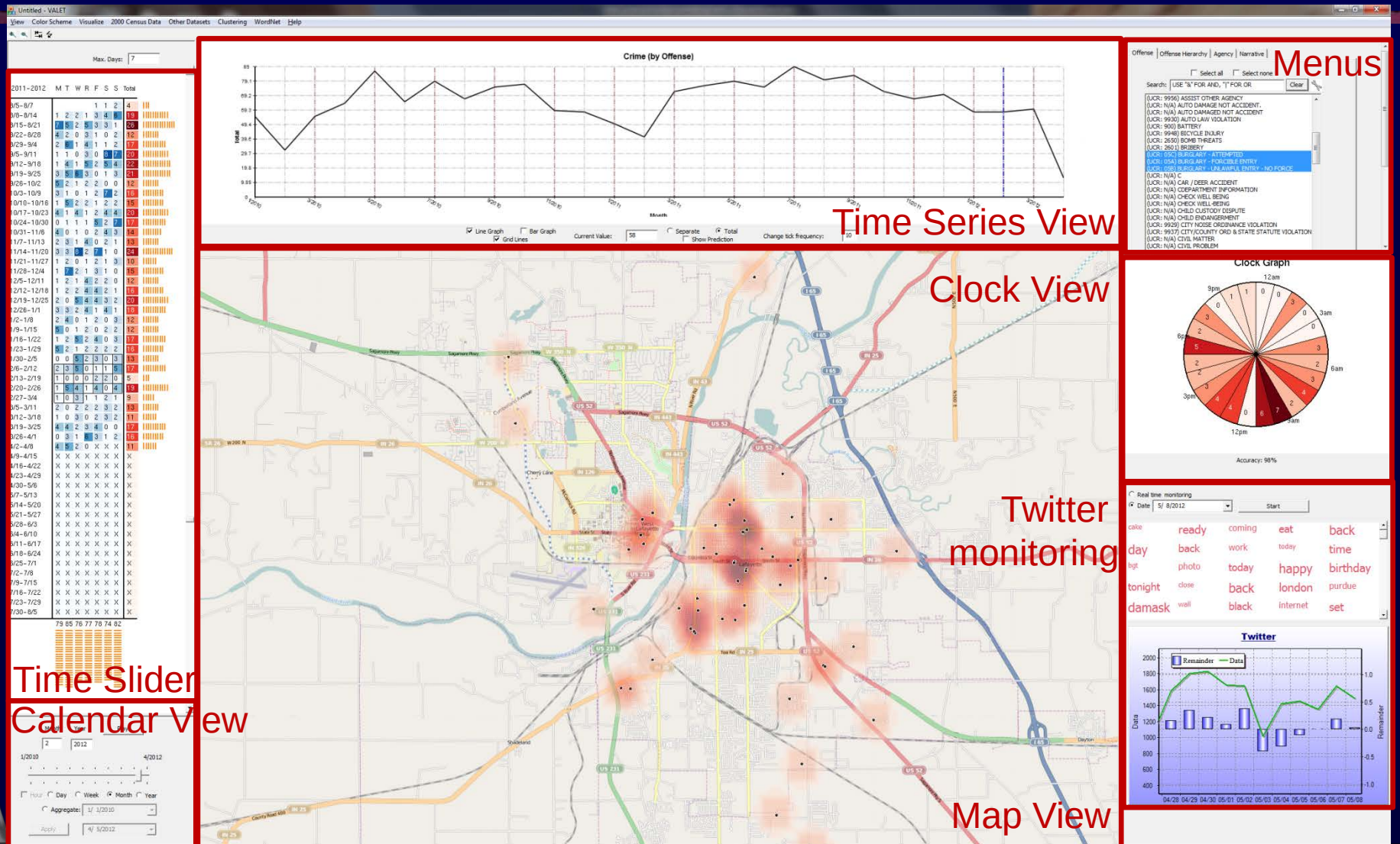
iVALET delivered:

- October 2011: Purdue, WL Police



VACCINE

VALET Overview



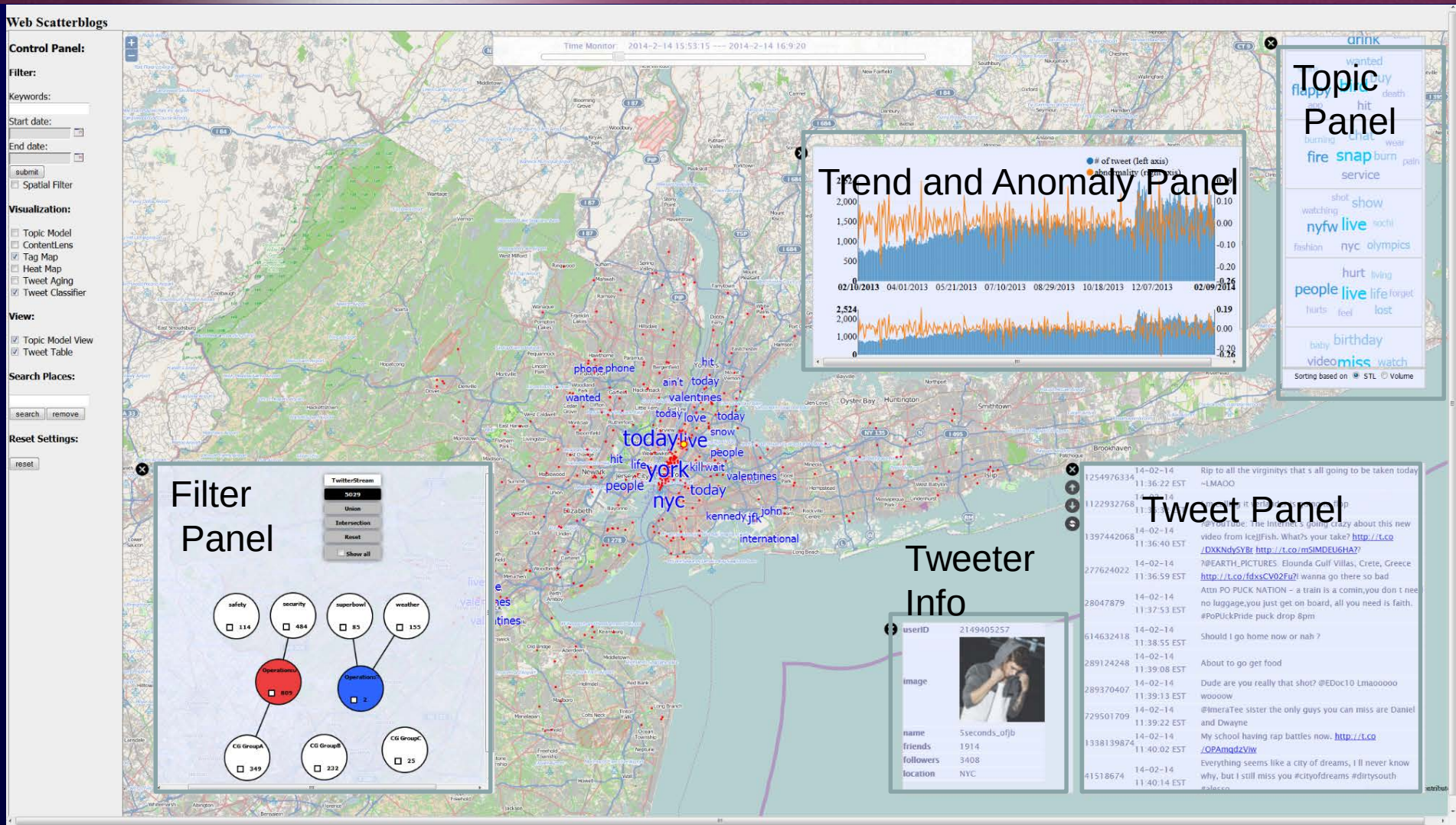
Social Media: Real-time Twitter Visual Analytics

(Purdue, Stuttgart, Penn St.)

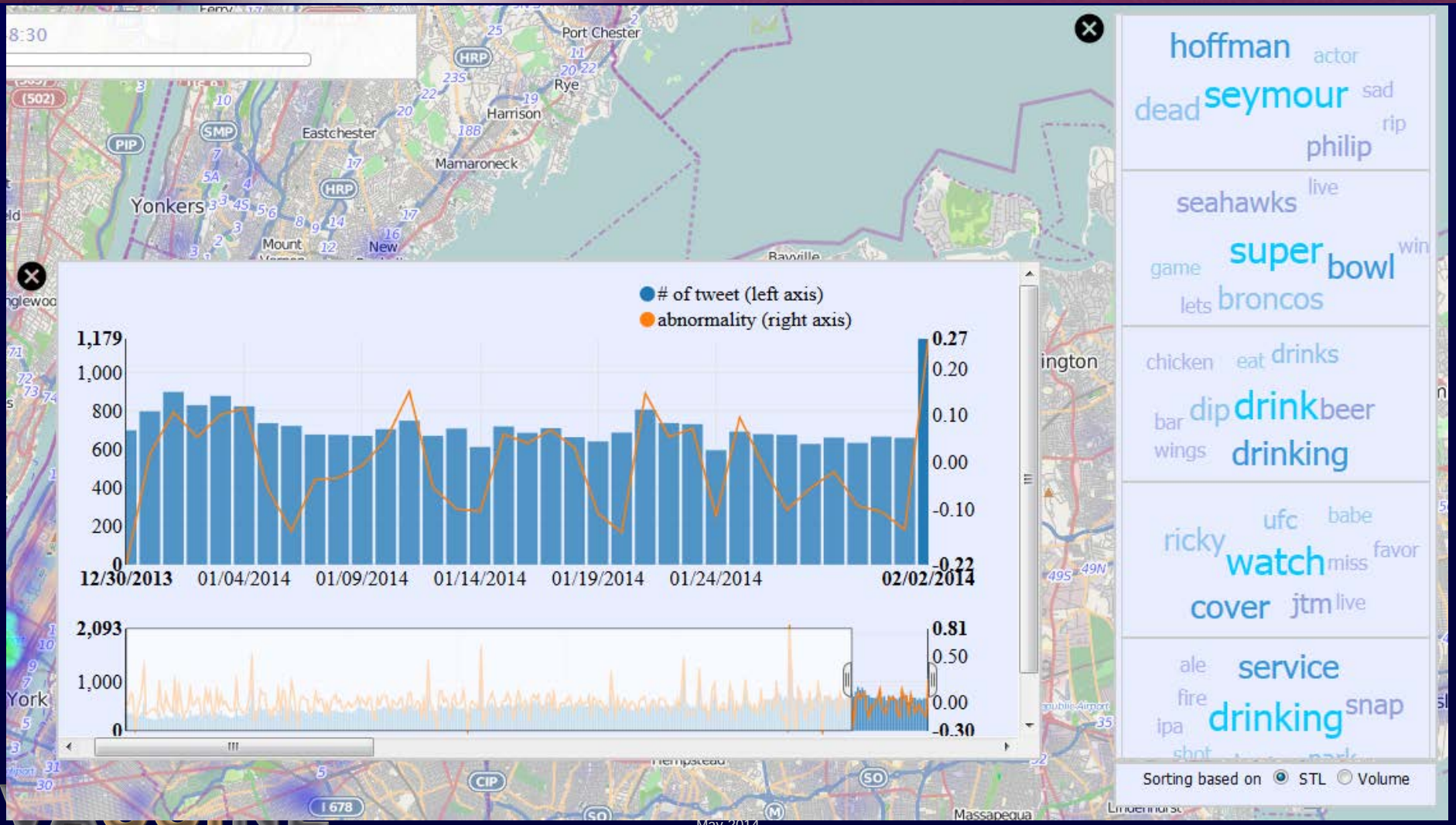
- Anomalous topic extraction using LDA and novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud



Social Media Visual Analytics: Scatterblogs



Anomaly Detection



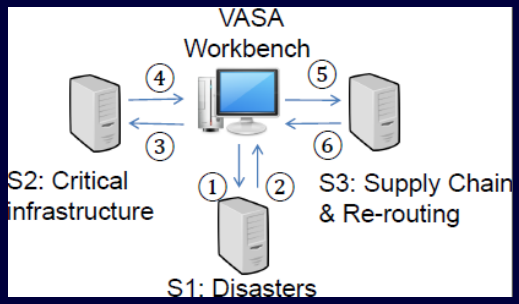
Cascading Critical Infrastructure Resiliency Modeling and Analytics (VASA)



- Purpose: Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.
- Collaborating Institution(s): Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Texas (TACC), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities
- End-User(s): Power Suppliers (e.g., Duke Energy), Cyber Community (e.g., Cisco), Quick Service Restaurants and suppliers, food supply

Purdue

UNCC

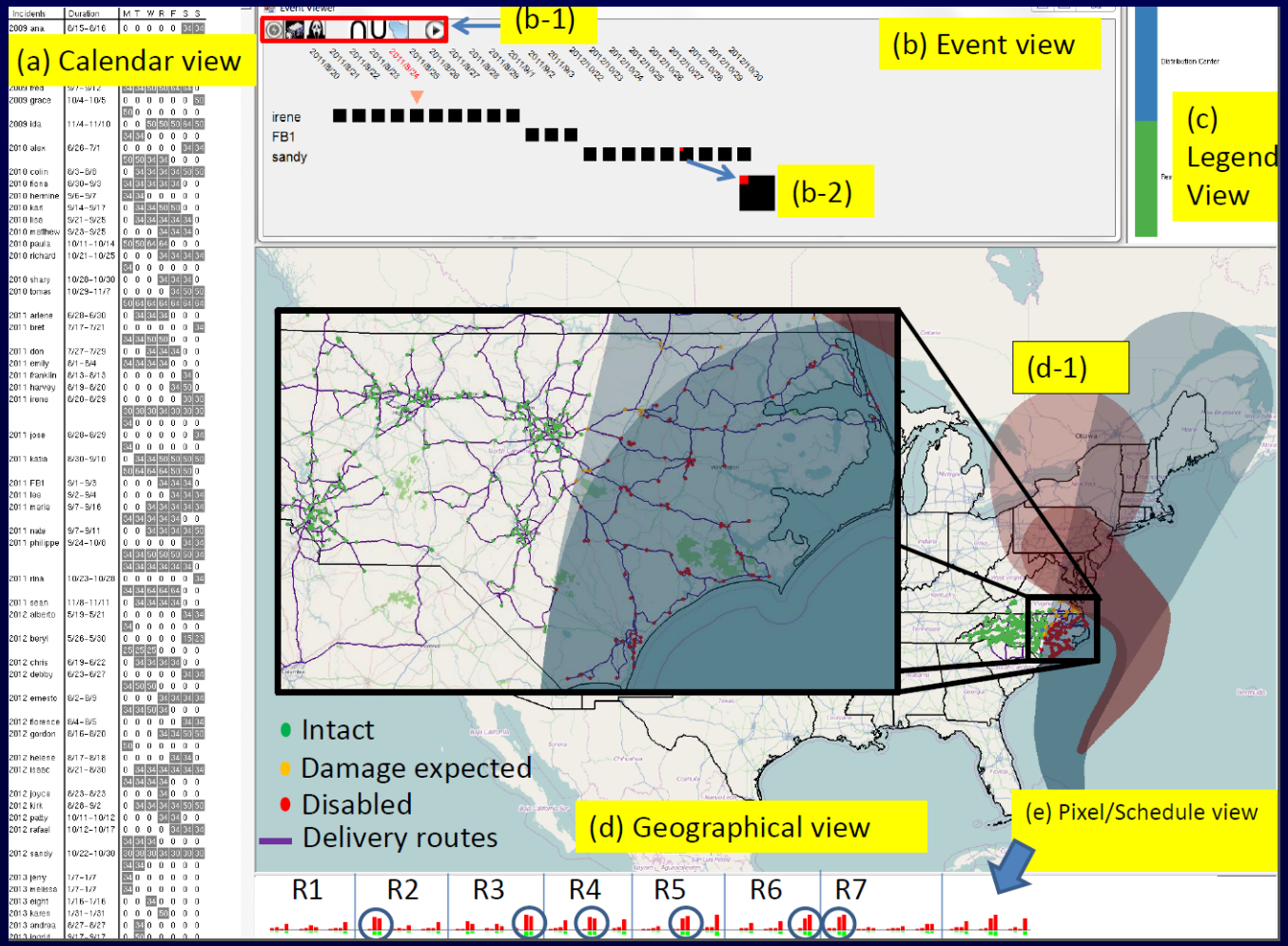


U Minn

TACC

VACCINE

System Interface



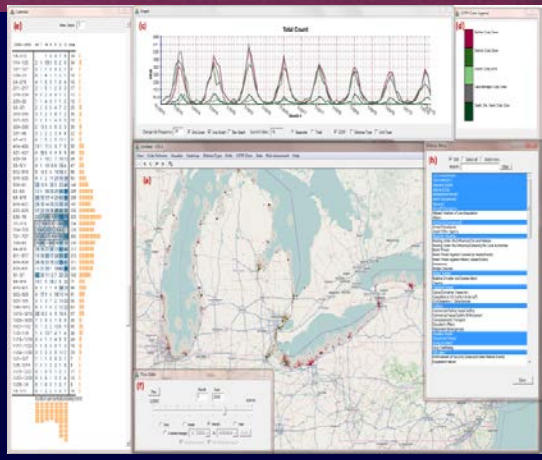
Effective Risk-based Decision Making and Resource Allocation Visual Analytics

- Evaluate current and historical mission area:
  - Demands
  - Risks (total, mitigated, residual)
  - Resource allocation
  - Return on investment
- Evaluate courses of action
- Evaluate above at both Strategic and Tactical/Operational level



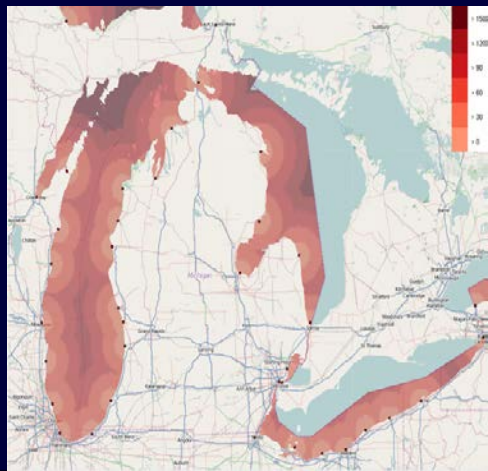
U.S. Coast Guard Search and Rescue VA (cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771



IMPACTS:

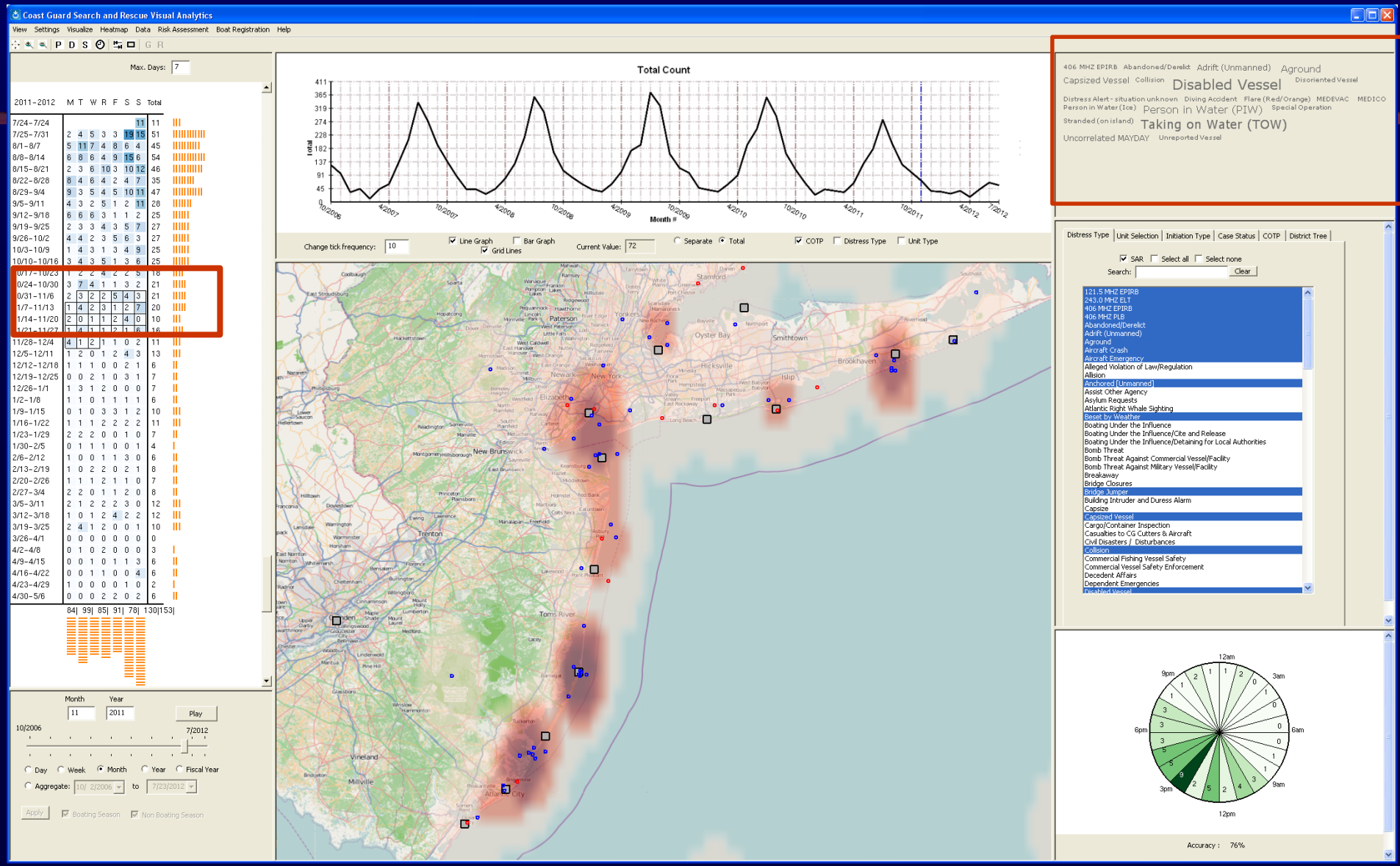
- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Hurricanes Sandy and Irene resource allocation decisions based on cgSARVA analysis and visualization
- Informed Commandant's budget testimony to Congress
- Key component of USCG D9 reallocation plan for 2011-12
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ



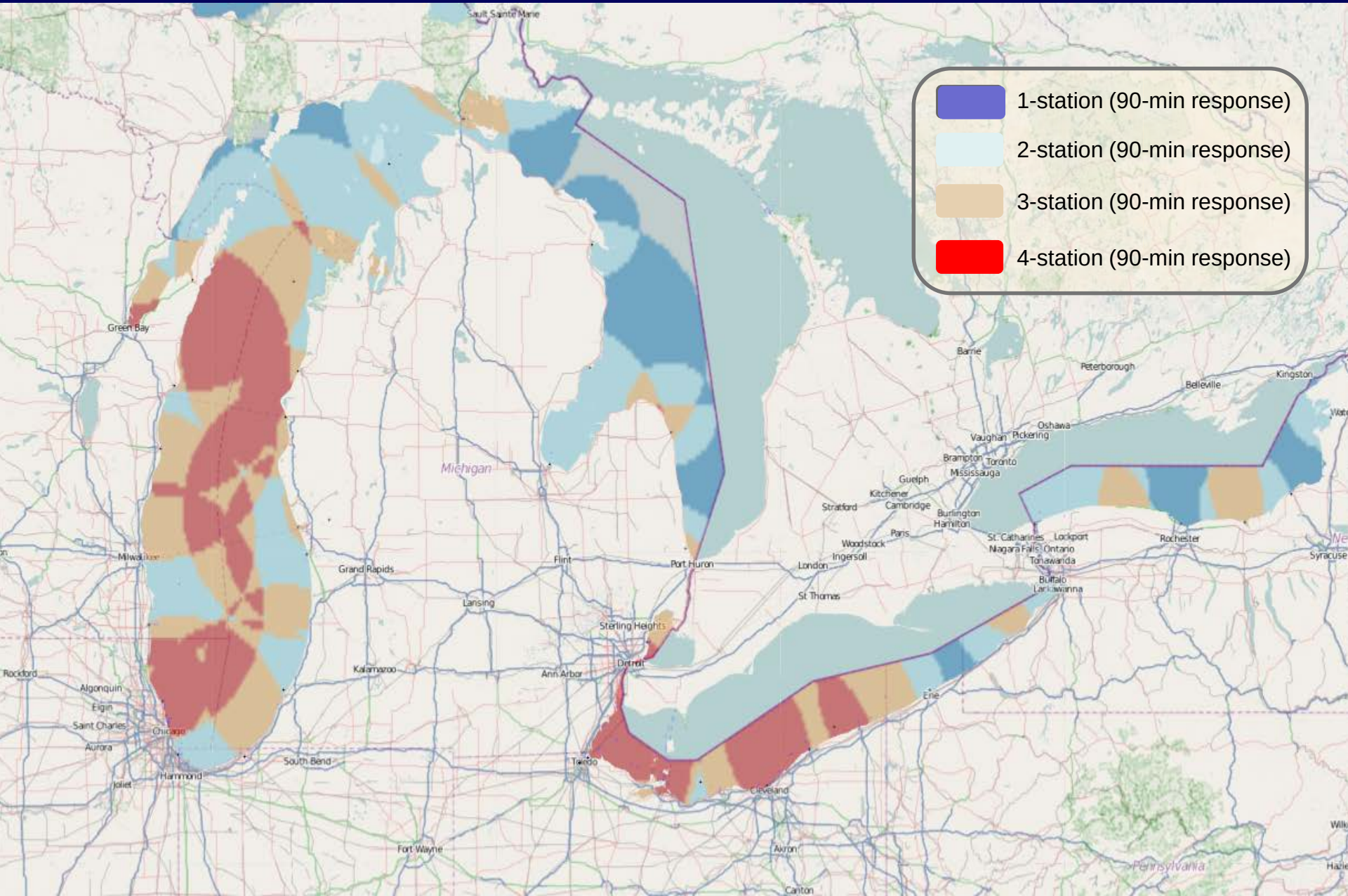
VACCINE

Example: Risks and Consequences From Sandy:

SAR Cases November 2011 NJ/NYC Area

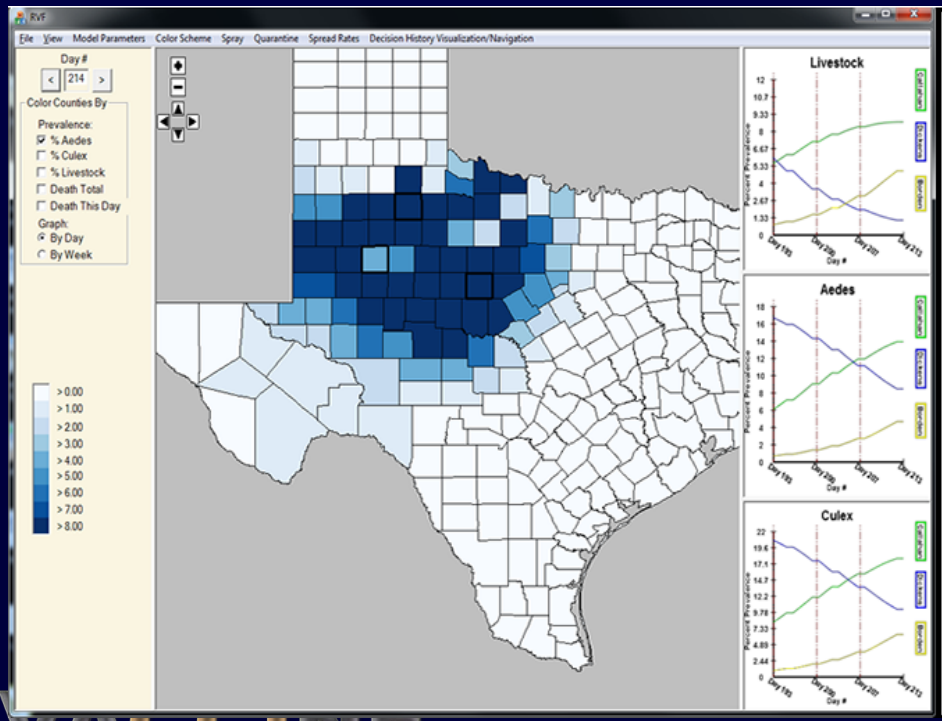


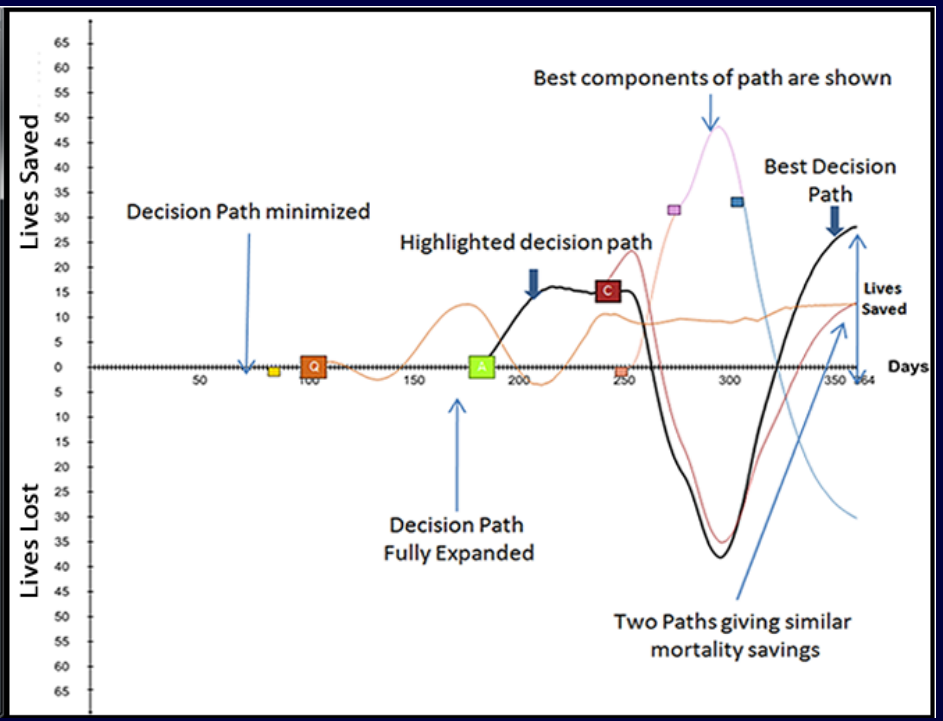
Response Efficiency – Potential Future Assets



Complex Decision Making: Advanced Decision Support Tools (e.g., Rift Valley Fever)

- Integrated simulation models
- Explore alternative courses of action in decision space and real space





Keys for Success

- User and problem driven
- Balance human cognition and automated analysis and modeling
  - Often applied on-the-fly for specific components identified by the user
- Interactivity and easy interaction
  - Utilizing HPC and novel analysis approaches
- Understandability of why predicted value is what it is
- Intuitive visual cognition
- Not overloaded with features

Seven Challenges for Proactive & Predictive Visual Analytics

1. Creating computer-human visual cognition environments
2. Integrating interactive simulation and analytics
3. Solving specific scale issues and cross-scale issues
4. Managing uncertainty and time
5. Enabling risk-based decision making environments
6. Developing the Science of Interaction for Visual Analytics
7. Defining appropriate evaluation methodologies for visual analytics

Cognitive Science Research Is Vital

- Cognitive science foundations and integration into interactive visual decision and policy making applications fertile area in need of new funded research programs
- Impact and opportunities are huge

For Further Information

www.VisualAnalytics-CCI.org



VACCINE

VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

Visual Analytics for CyberSecurity

David S. Ebert

CVADA

Center for Visual Analytics and Data Analytics

A U.S. Department of Homeland Security Center of Excellence

October 2013

Who We Are:
International Team of Experts
75+ Faculty, 26 Institutions



- Purdue University
- Georgia Institute of Technology
- Pennsylvania State University
- Stanford University
- University of North Carolina at Charlotte
- University of Washington
- Arizona State University
- Simon Fraser University
- University of British Columbia
- Justice Institute of British Columbia
- Ontario Institute of Technology
- Dalhousie University
- University of Houston, Downtown
- Virginia Tech
- Indiana University
- Florida International University
- University of Texas at Austin
- Morgan State University
- Navajo Technical College
- University of Stuttgart
- University of Swansea
- Oxford University
- University of Calgary
- University of Manitoba
- Carleton University
- University of Victoria

VACCINE Value

Our Value / Solution: Enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable users to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

VACCINE Value Part II



Our people and partnerships

- Interdisciplinary world-leading team of researchers
- Defining and extending the new science of visual analytics driven by real-world, real-scale problems of engaged partners (local, state, federal)







Visual Analytic Solutions: What We Offer

- **Improved Effectiveness**: We enable users to be more effective through innovative, interactive visualization, analysis, and decision making tools
  - Provide the right information, in the right format, within the right time to solve the problem
  - Enable user to be more effective from planning to detection to response to recovery
  - Enable effective communication of information
- **Innovative Fielded Solutions:** We provide innovative visual analytic and scalable solutions to the extended homeland security community
- **People and Partnerships**
  - Interdisciplinary world-leading team of researchers and students
  - Actively Engaged Partners

VADM Robert Parker with VACCINE student researchers (cgSARVA, COAST, iOPAR)

"cgSARVA has proven its worth time and again, providing key analytic information for decision makers for large scale projects…"

VADM Robert Parker, 2012 MRS Keynote Address



Engaged End-Users

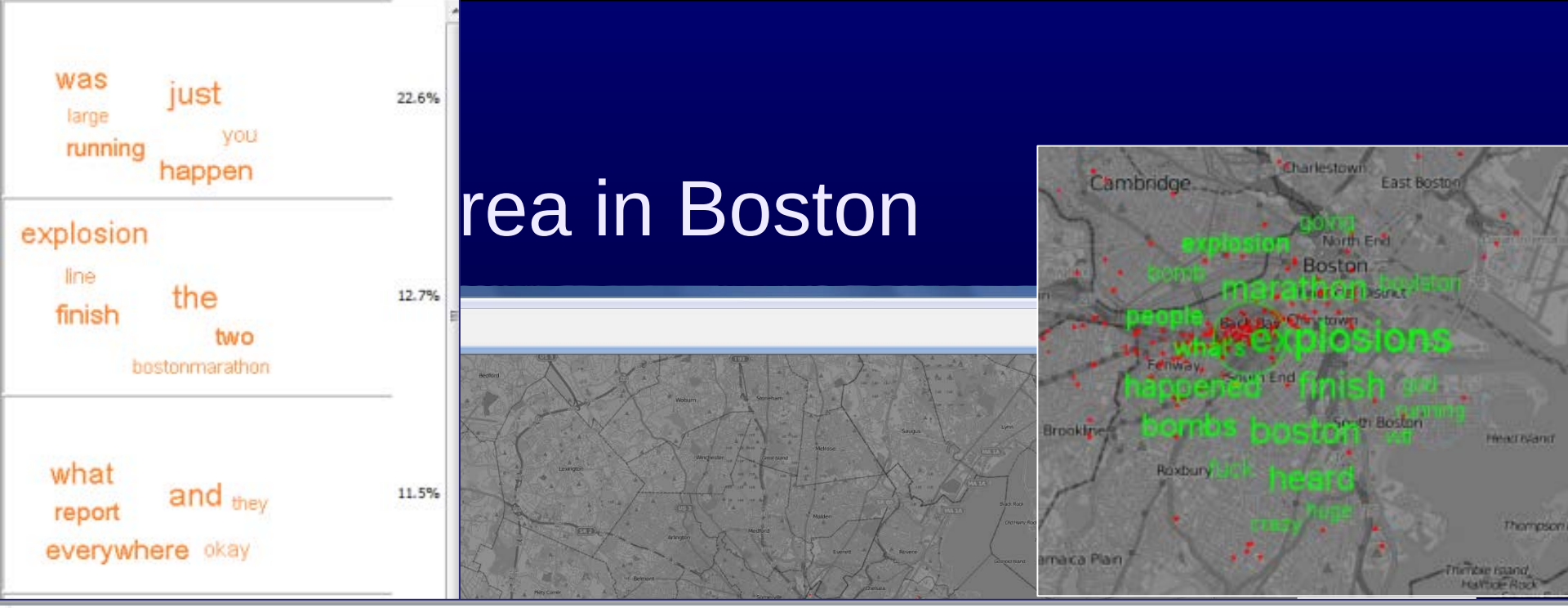
- Federal Operating Components:
  - US Coast Guard
  - US Transportation Security Agency
  - US Customs and Immigration Service
  - US Federal Emergency Management Agency
  - US Customs and Border Patrol
  - US CERT
  - US ICE (in progress)
- Law Enforcement
  - Over 40 local and state agencies (IN, IL, OH, SC, PA, NC, NY)
- Fusion Centers
  - Ohio (SAIC)
  - Indiana (IIFC)



Social Media: Real-time Twitter Monitoring and Integration into Tools (Purdue, Stuttgart, Penn St.)

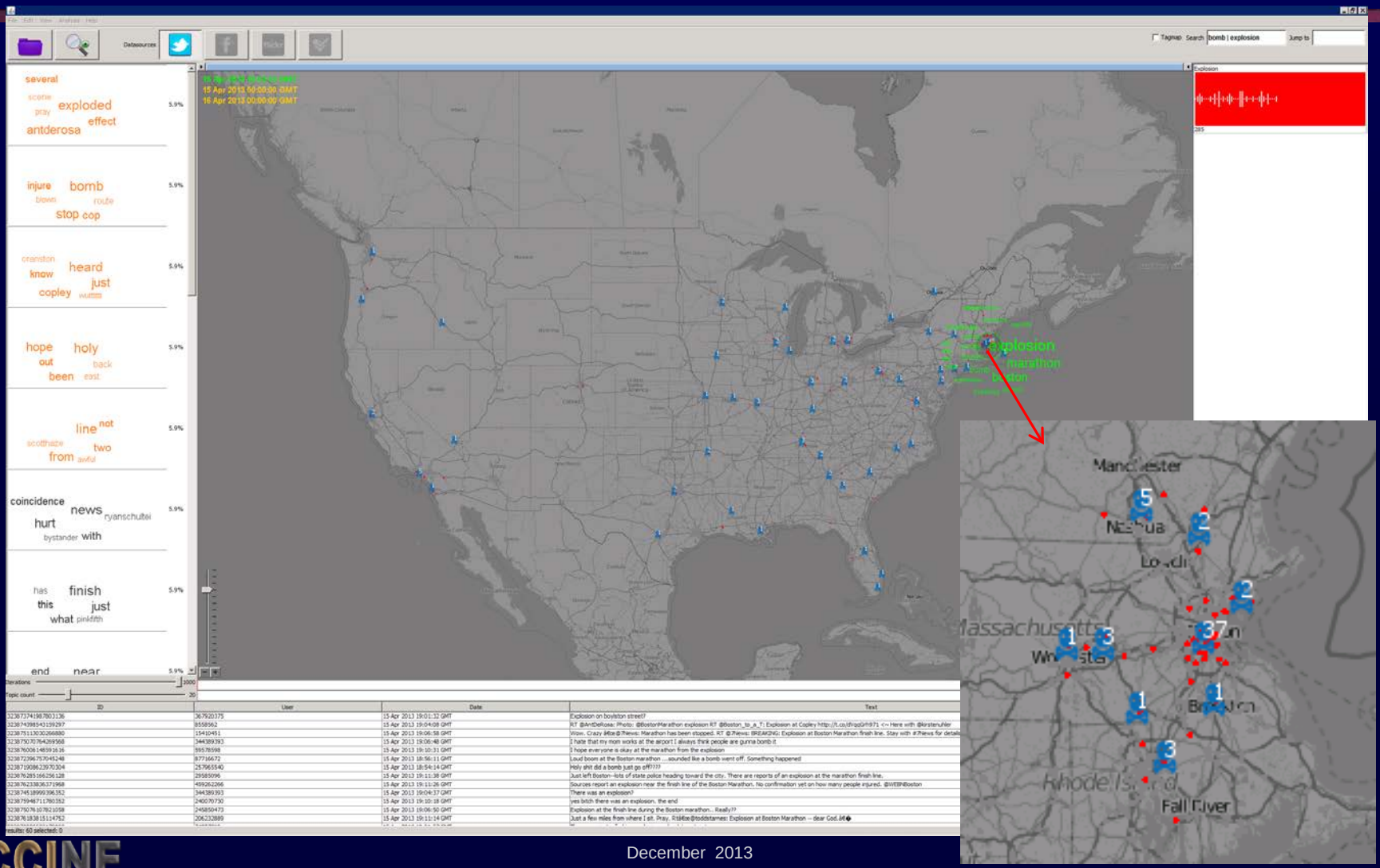
- Topic extraction using novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud





Text
Oh my god what just happened
Something happened at the end of the Boston Marathon. Something bad and there is a lot of chatter on Twitter. What's going on?
Multiple people are injured near the Boston Marathon finish line after two explosions. The #BostonMarathon has been stopped.
two bombs just went off on boylston
BREAKING NEWS: Two powerful explosions detonated in quick succession right next to the Boston Marathon finsh line this afternoon.
What the FUCK was that
Back in Sept, @croon1 solicited me for $2000. He now has a music video with William Shatner. If you watch it (god forbid) keep that on mind.
Literally what the fuck get me out of here
@DTenenbaum my office right next to it
@FRANCESCalciO I figured 3 people would get the joke
Two explosions just rocked the finish line of the Boston Marathon. Sirens galore. People running in fear. Wonder what happened.
This is crazy i seen that blow up #bostonmarathon
can someone tell me what that explosion was!? #boston #bostonmarathon

Detection using the Explosion Classifier



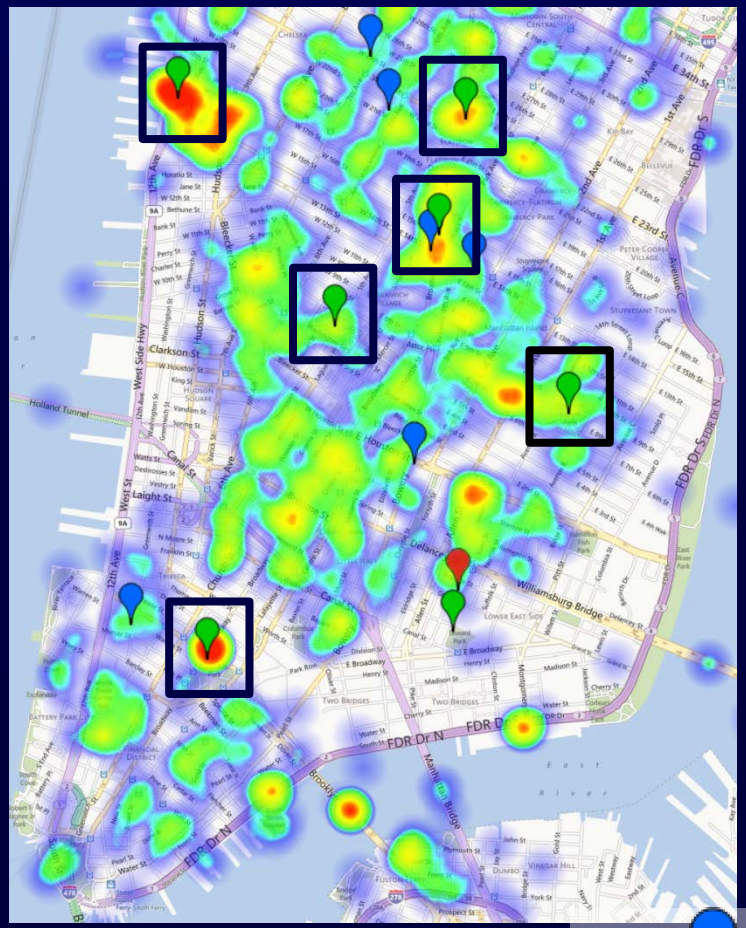
First Response (Tweet & Picture)

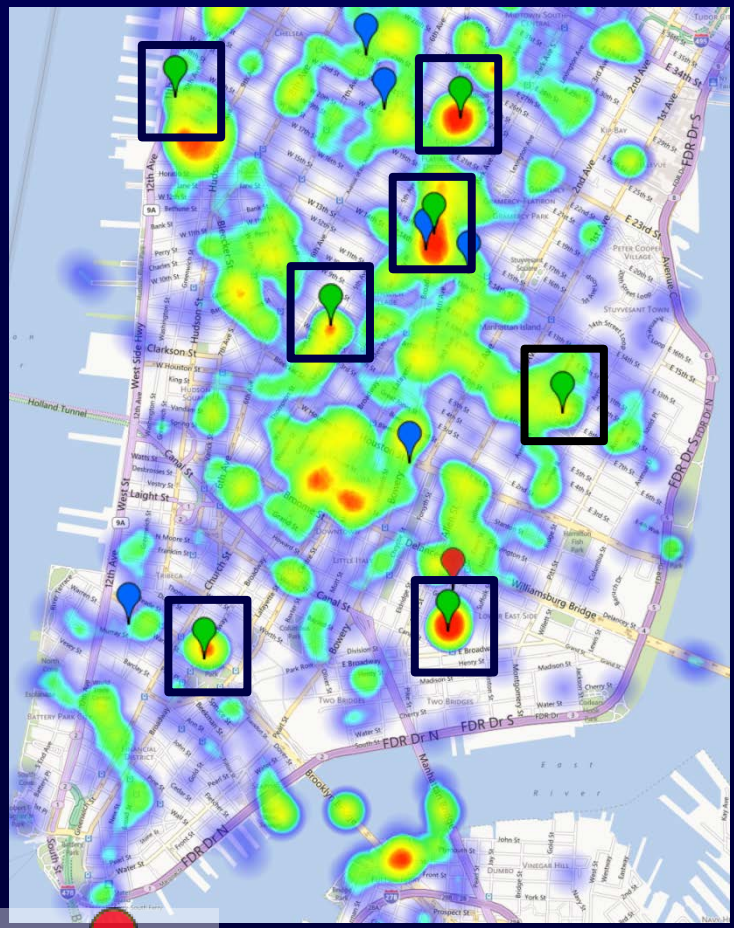


Visual Analytics of Activity During Hurricane Sandy

Two weeks before Sandy
10/14 (Sunday), 12:00 ~ 16:00

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00





Supermarket Park Shelter

December 2013

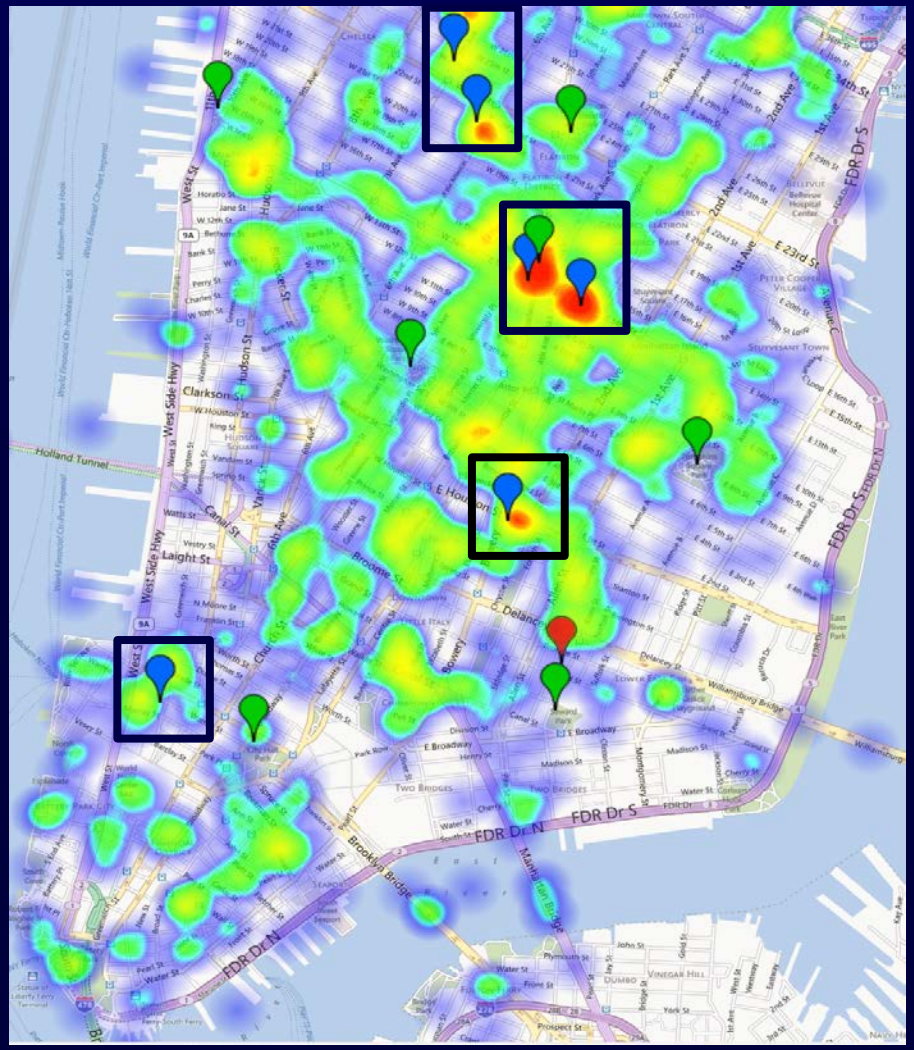
VACCINE

Visual Analytics of Activity During Hurricane Sandy

After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00

Evacuation order:
10/28, 10:30 AM

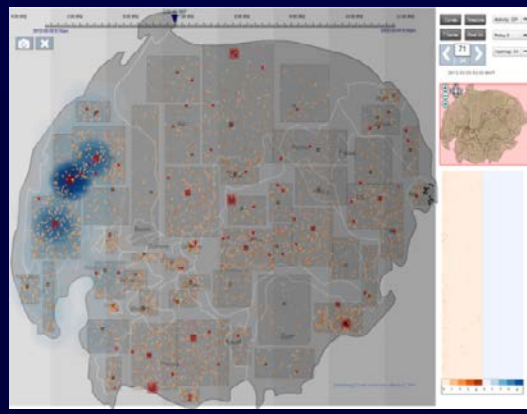
Hurricane Sandy's
Arrival at NYC:
10/29, 8:00 PM

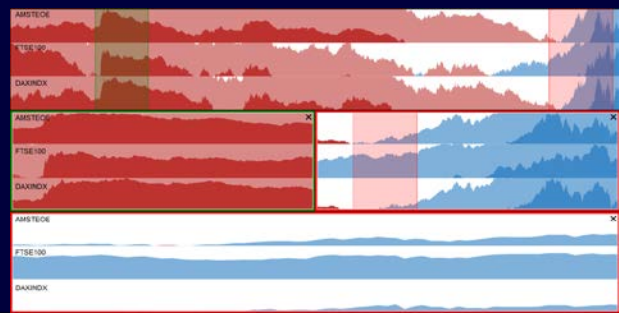


VACCINE

Example VACCINE Team Work in Cybersecurity

- Visual Analytics for Security Application (VASA)
- Corporate Insider Threat Detection (Oxford, Leicester, Cardiff)
- Sensor Forensics (Purdue)
- SemanticPrism (Purdue)
- Multiscreen, Multiview, Interactive Cyber Investigation (VaTech, PNNL)
- Log Visualization (Purdue)





Cascading Critical Infrastructure Resiliency Modeling and Analytics (VASA)

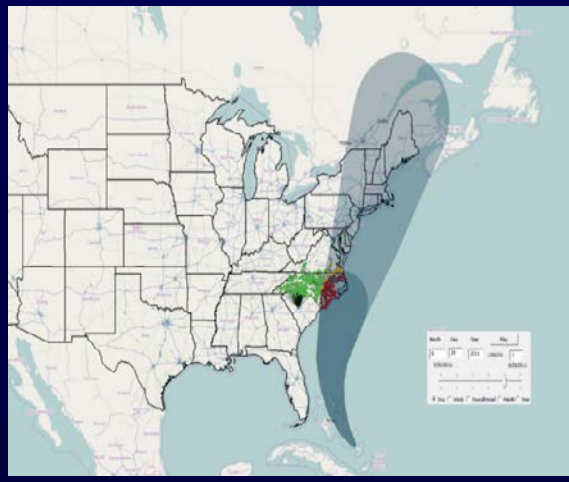


- **Purpose:** Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.
- **Collaborating Institution(s):** Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities
- **End-User(s):** Power Suppliers (e.g., Duke Energy), Cyber Community (e.g., Cisco), Quick Service Restaurants and suppliers, food supply

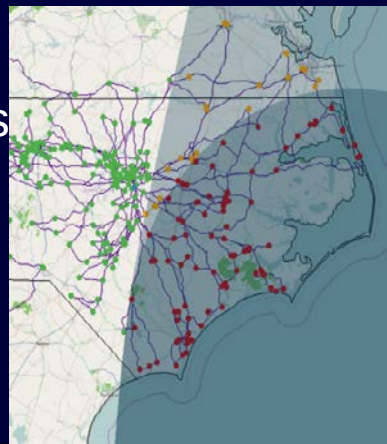
VASA: Visual Analytics for Security Applications

Collaborating Institution(s): Purdue, Minnesota, UTexas, UNCC + German universities
End-User(s): Fast-food restaurant chain, emergency management and planning personnel

Impacts and Accomplishments:



- Support decision-making for extreme weather and disaster (natural, man-made) scenarios
  - Combine real and simulation data
  - Allow “what-if” exploration
- **System of systems**: binds together multiple simulations models from collaborators into coherent
  - **Minnesota**: food distribution model
  - **Texas**: simulated and historical weather (hurricanes, storms)
  - **UNCC**: critical infrastructure
  - **Purdue**: roads + interaction visual analytics tool
- **Challenge**: combine interactive VA with complex simulation models for effective decision making



Sensor Forensics
(Purdue – Delp)

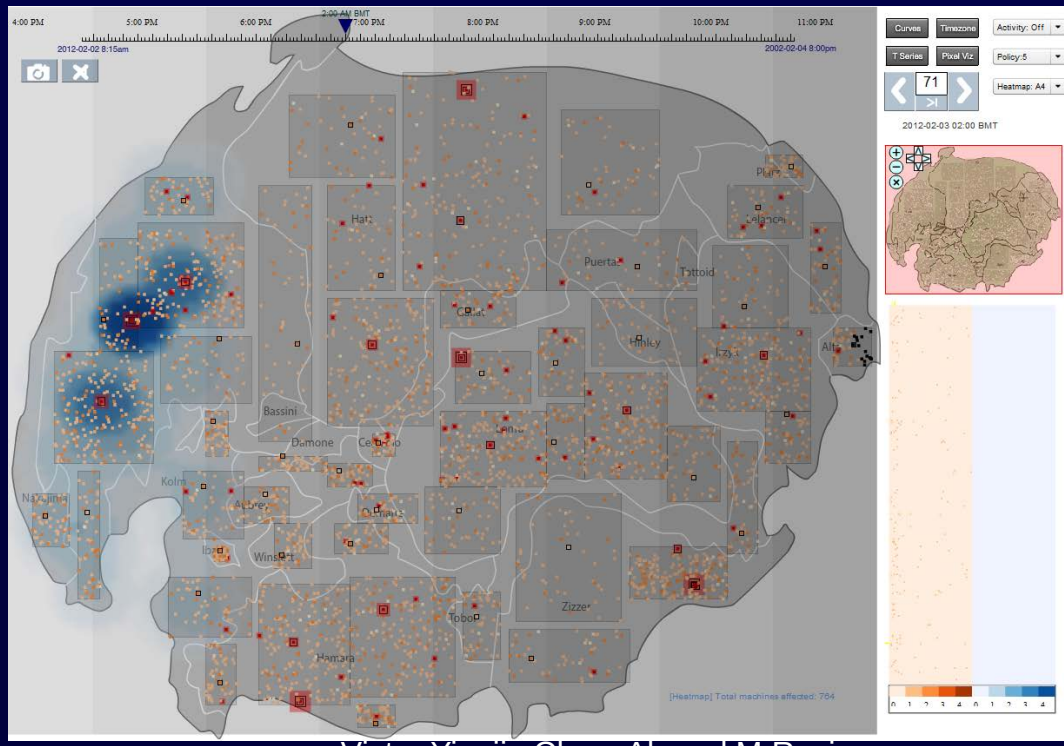
- Forensic characterization
  - Observe device output ➔ which device produced it?
  - Exploit how the device “makes” its output
- Device authentication
  - Performed using forensic characterization
  - Identify device type, make, model, configuration
  - Can the sensor be trusted?
- Detection of data forgery or alteration
- Fingerprint and trace

SemanticPrism: A Multi-aspect View of Large High-dimensional Data (Purdue University)

- **VAST 2012 Mini Challenge 1 Award: Outstanding Integrated Analysis and Visualization**

- **Geo-Temporal**
- **Time-serial**
- **Pixel-based**

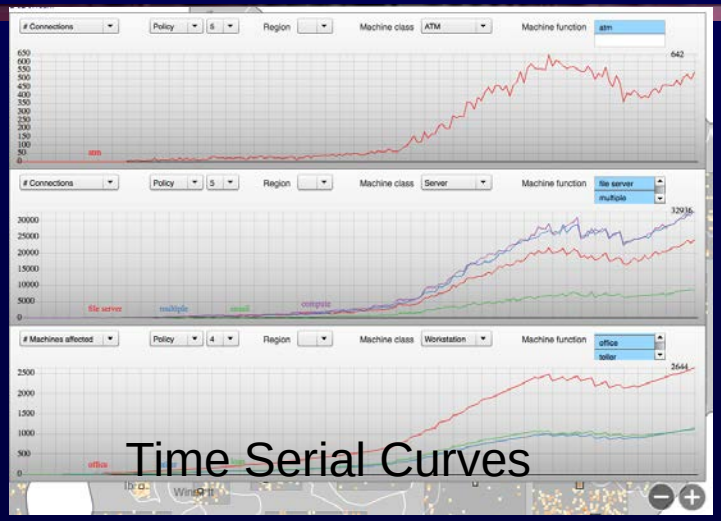
- **Semantic Zoom**



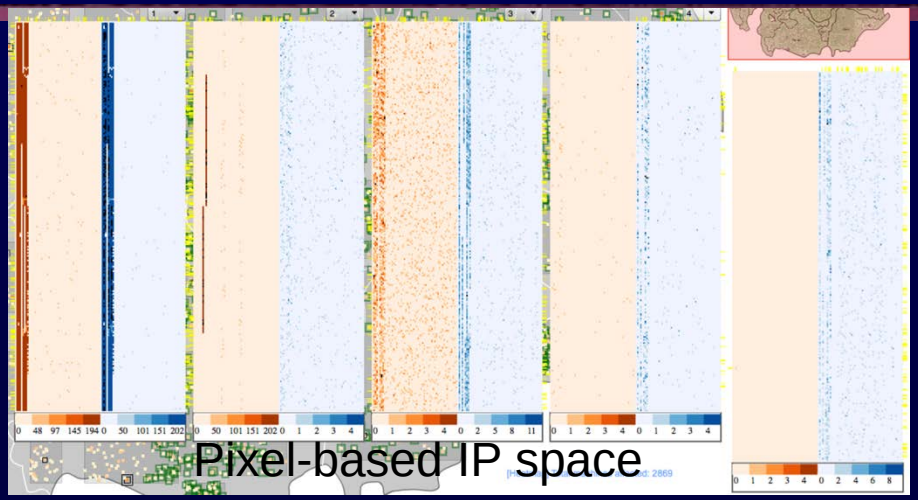
Victor Yingjie Chen, Ahmad M Razip, Sungahn Ko, Cheryl Zhenyu Qian, David S.Ebert

Computer Graphics Technology
VACCINE
Visual Analytics for Command, Control, and Interoperability Environments
A U.S. Department of Homeland Security Center of Excellence
iVD

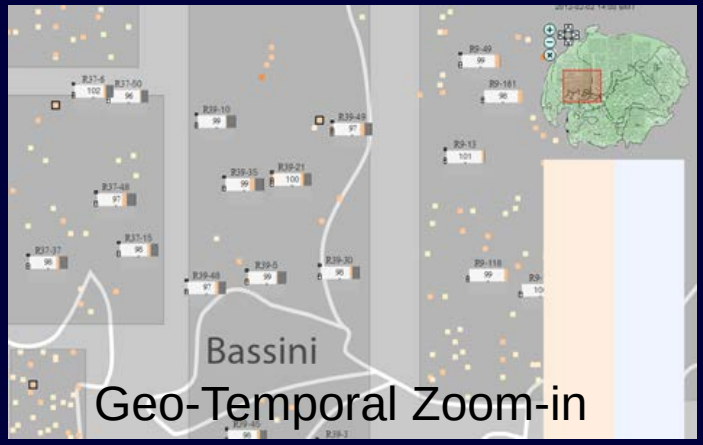
SemanticPrism



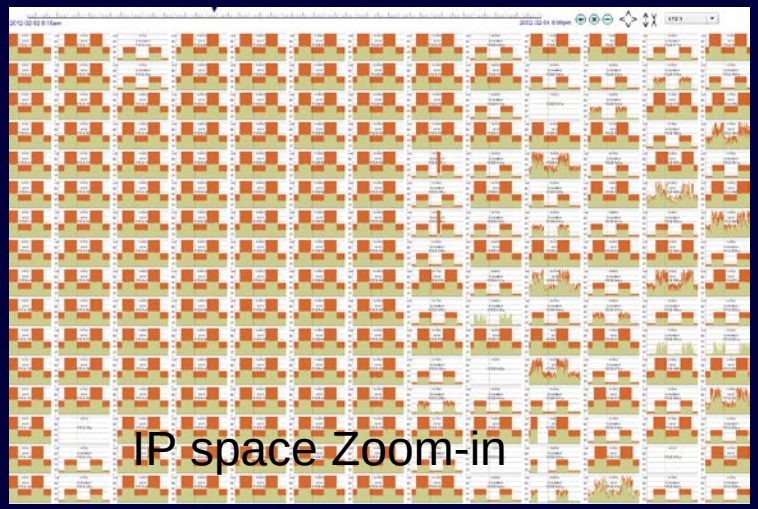
Time Serial Curves



Pixel-based IP space



Geo-Temporal Zoom-in

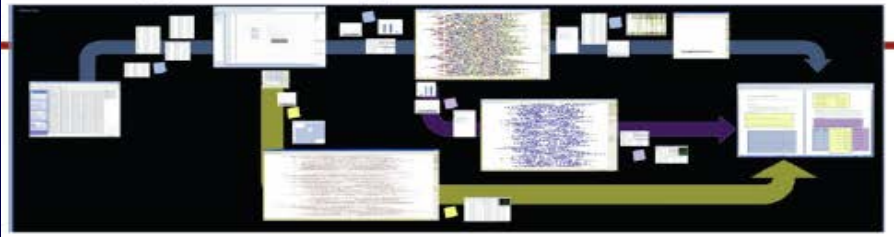


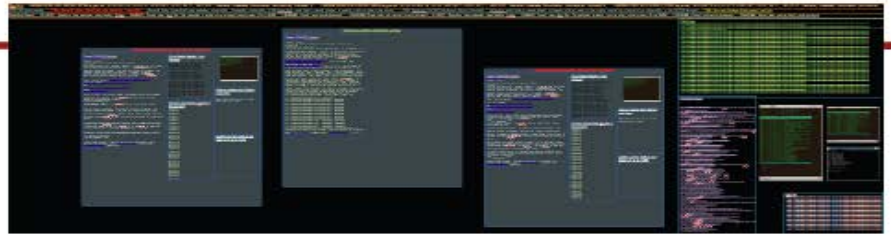
IP space Zoom-in

VA for Cybersecurity Analysts

(VaTech – North, Endert)

Large High-Res Workspaces for Analysts





History and Traceability

Multiple, Simultaneous Investigation cases

Large, High-Resolution Visualizations

De-Aggregate Vital Information

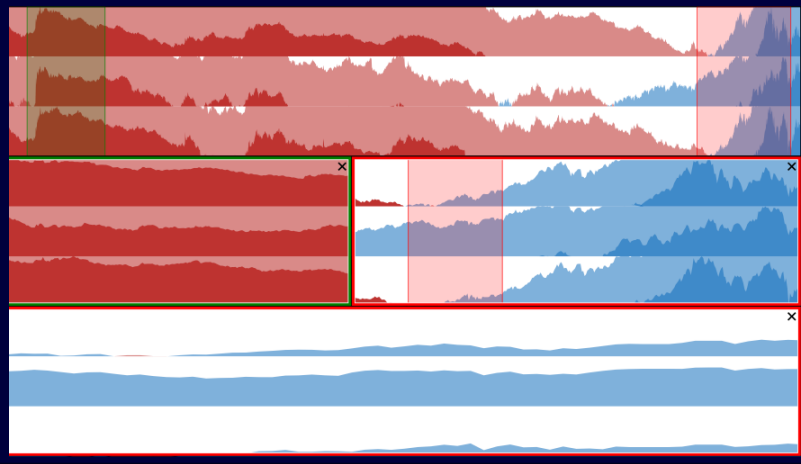


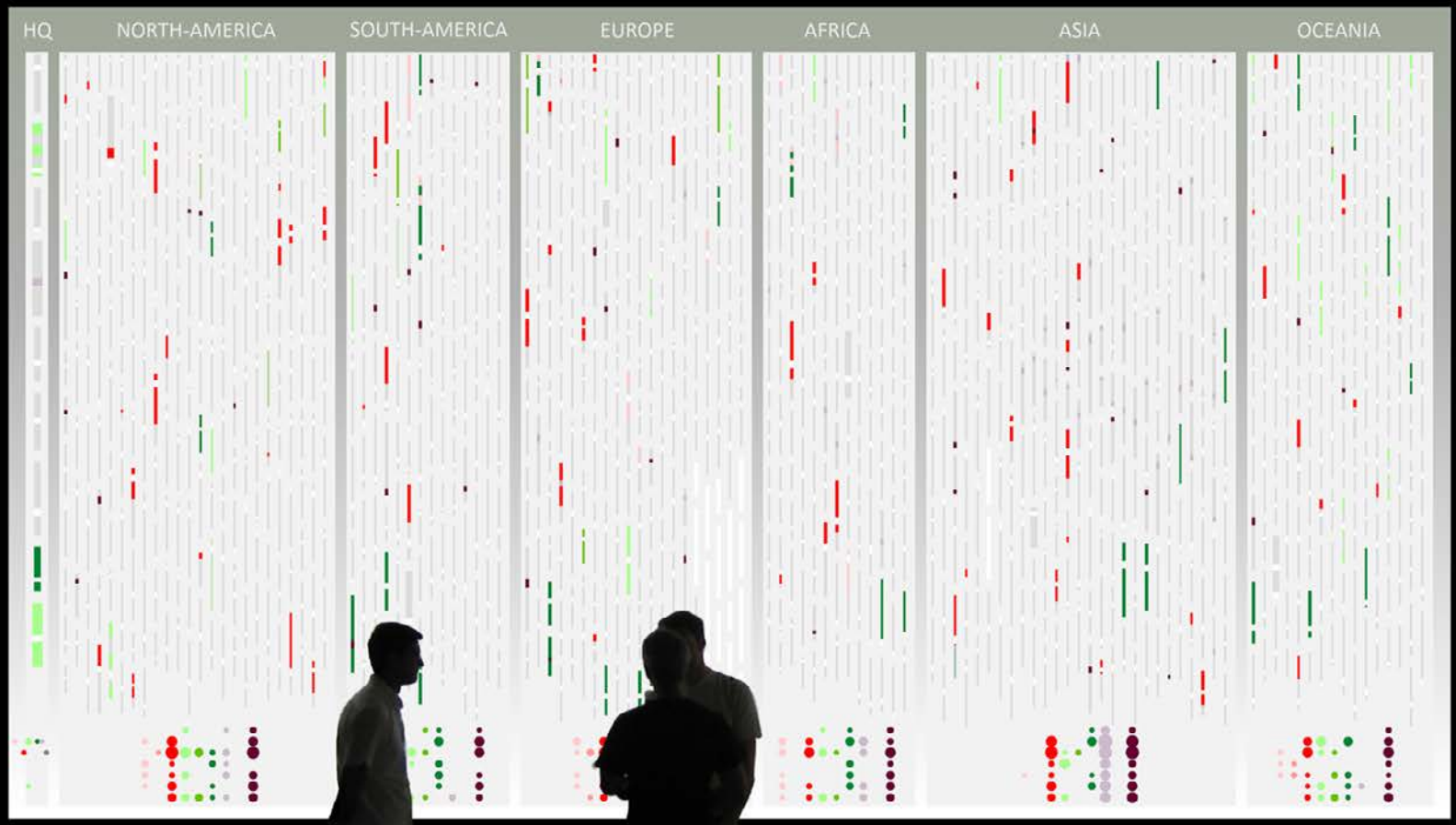


Server Cluster Log File Visualization

(Elmqvist, Purdue)

- Log file visualization for Purdue's ECN group
  - 200+ servers, 30 TB of storage, 6 million hits per month
  - User-centered, interview-based design
- Applied **stack zooming** to quantitative log data
  - CPU load, network hits, storage usage,
  - Users navigate in data
  - Very long time periods
- Limited deployment in Fall 2010
  - Very positive, powerful





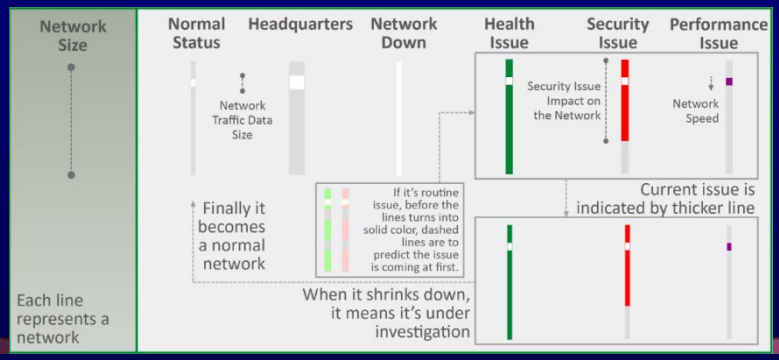
2013 VAST Challenge MC2 Award

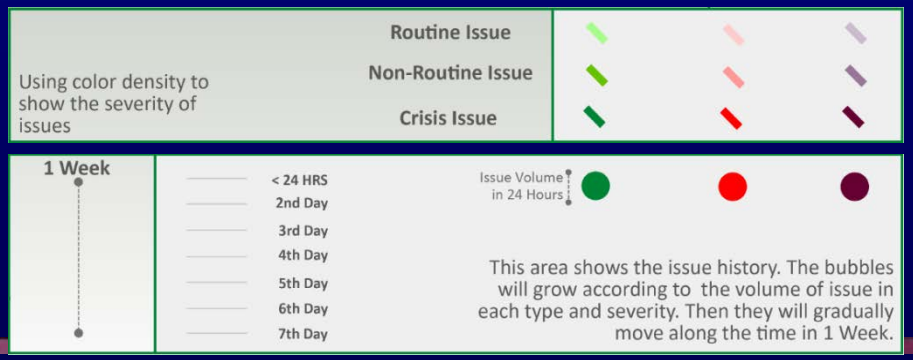
Outstanding Creative Design

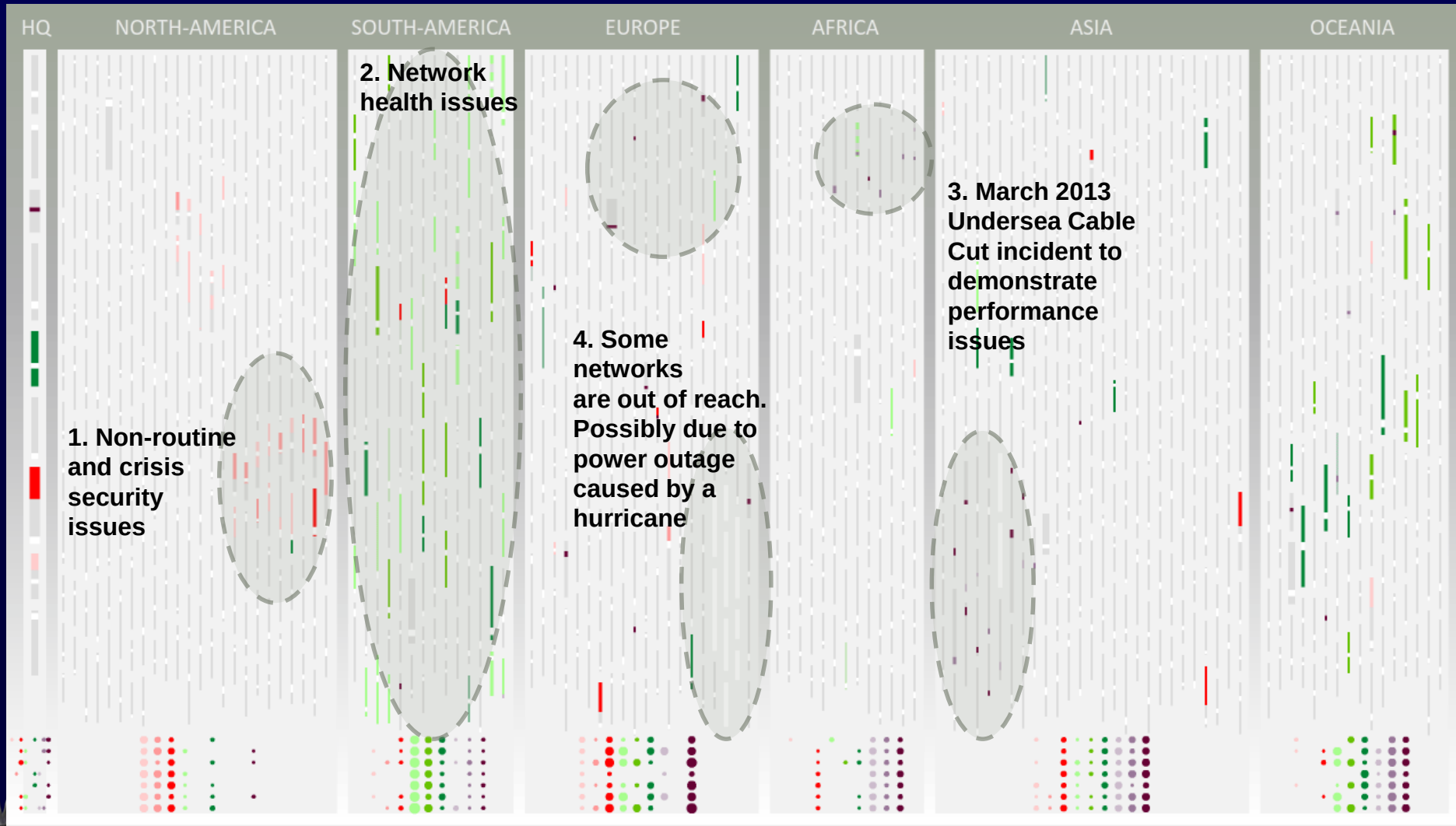
www.interactiondesign.us/vast2013/SpringRain

Spring Rain

A Visual Analytics System with an Ambient Information Display +


Network Size
Each line represents a network
Normal Status
Headquarters
Network Down
Health Issue
Security Issue
Performance Issue
Network Traffic Data Size
Security Issue Impact on the Network
Network Speed
Finally it becomes a normal network
If it's routine issue, before the lines turns into solid color, dashed lines are to predict the issue is coming at first.
Current issue is indicated by thicker line
When it shrinks down, it means it's under investigation


Using color density to show the severity of issues
Routine Issue
Non-Routine Issue
Crisis Issue
1 Week
< 24 HRS
2nd Day
3rd Day
4th Day
5th Day
6th Day
7th Day
Issue Volume in 24 Hours
This area shows the issue history. The bubbles will grow according to the volume of issue in each type and severity. Then they will gradually move along the time in 1 Week.


HQ
NORTH-AMERICA
SOUTH-AMERICA
EUROPE
AFRICA
ASIA
OCEANIA
1. Non-routine and crisis security issues
2. Network health issues
3. March 2013 Undersea Cable Cut incident to demonstrate performance issues
4. Some networks are out of reach. Possibly due to power outage caused by a hurricane

Application-Driven Visuali Cybersecurity

- Should we be driving the resear different users' goals are?
- Interesting survey article
  - Taxonomy of use-case classes:
    - Host-server monitoring
    - Internal/external monitoring
    - Port activity
    - Attack patterns
    - Routing behavior

Shiravi, et al., " A survey of visualization systems for netw security, *IEEE TVCG* 2012.

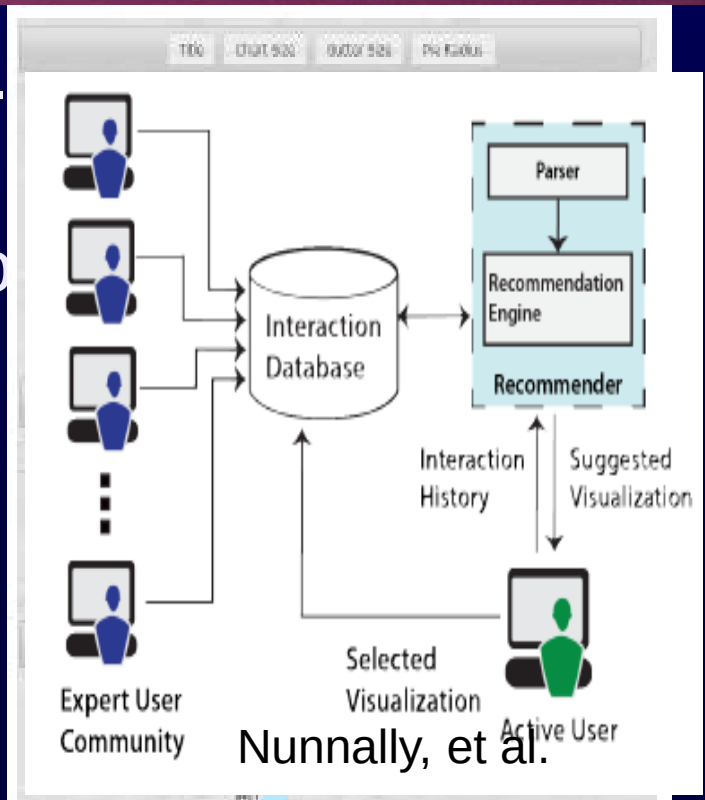


Visualization System	Visualization Technique(s)	Data Source(s)	Number of Citations
Host / Server Monitoring			
Erbacher et al. [4][5]	Glyph	Server Logs	106 | 7
Tudumi [6]	3D Node Link	Server Logs	38
NVisionIP [7,8]	Scatter Plot	NetFlows	145 | 20
Portall [9]	Node Link	Packet Traces	21
HoNe [10]	Node Link	Packet Traces	8
Perlman et al. [11]	Node Link | Glyph	Packet Traces	5
Radial Traffic [12]	Radial Panel	Packet Traces	23
Mansmann et al. [13]	Node Link	Packet Traces	2
Abdullah et al. [18]	Histogram	Packet Traces	30
Cube of Doom [19]	3D Scatter Plot	Packet Traces	99
PortVis [20]	Scatter Plot	NetFlows	112
NetBytes Viewer [21]	3D Scatter Plot	NetFlows	7
Existence Plots [22]	Scatter Plot	Packet Traces	3
Attack Patterns			
Giardin [29]	Color Map	Packet Traces	60
NIVA [30]	Node Link | Glyph	Intrusion Alerts	51
Snort View [31]	Scatter Plot | Glyph	Intrusion Alerts	67
IDGraphs [32]	Scatter Plot	NetFlows	29
IP Matrix [33]	Scatter Plot | Color	Intrusion Alerts	21
Visual Firewall [34]	Scatter Plot	Packet Traces	24
IDS Rainstorm [35]	Scatter Plot	Intrusion Alerts	60
Vizalert [36][37][38]	Radial Panel	Intrusion Alerts	38 | 35 | 29
Rumint [39][40]	Parallel Coordinates	Packet Traces	15 | 35
Ren et al. [41]	Flying Term	DNS Traces	10
Xiao et al. [42]	Scatter Plot	Packet Traces	23
Svision [43]	3D Scatter Plot	Packet Traces	9
Mansmann et al. [44]	Treemap	Packet Traces	20
SpiralView [45]	Radial Panel	Intrusion Alerts	5
NFlowVis [46]	Treemap	NetFlows	17
Avisa [49]	Radial Panel	Intrusion Alerts	2
Routing Behavior			
BGPlay [50]	Node Link	BGP Traces	22
Wong et al. [51]	Node Link	BGP Traces	9
LinkRank [52]	Node Link	BGP Traces	16
Teoh et al. [53][54][55]	Histogram | Node Link	BGP Traces	54 | 28 | 35
BGP Eye [56]	Color Map	BGP Traces	8

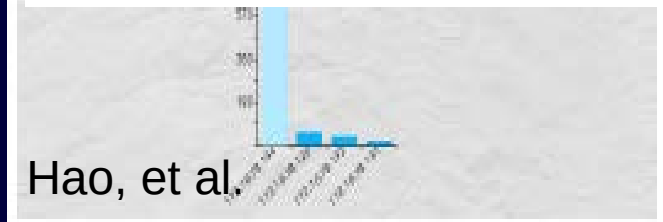
Do citation counts show real-world value?

VizSec 2013 Papers - My Analysis

- Expert users engaged: 3 of 9 (e.g. enforcement, network analysts)
- Evaluation – performance most co informal user studies
- Data – 4 with actual data, 1 using
- Users involved from the start – 1 paper
- Training of Novices based on experts – 1 paper



Nunnally, et al.



Hao, et al.

Directions Forward, Keys to Success & Challenges

Cybersecurity Education

- How do we train practitioners in the field?
  - Varied backgrounds, varied tasks, communication to public
  - Short time to learn
  - Visualization is key
- Approaches:
  - How people learn framework
  - Personalizing learning – Community of practice training
- How do we educate the public?
  - Again, visualization is key



Some Challenges for Cybersecurity

1. Understanding the task and workflow, access to expert users, actual problems, environments
2. Creating decision making environments for analysts with realtime data and decision making constraints at real-world scale (computer-human visual cognition environments)
3. Solving specific scale issues (scalability) and cross-scale issues (machine, intranet, internet)
4. Managing uncertainty and time
5. Enabling risk-based decision making environments

For Further Information

www.VisualAnalytics-CCI.org

vaccine@purdue.edu
ebertd@purdue.edu



VACCINE

VACCINE's Role

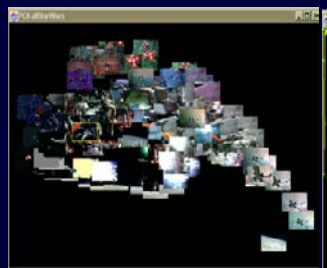


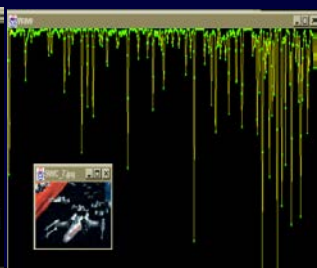
Problem: To solve current and future homeland security problems requires exploring, analyzing, and reasoning with massive, multi-source, multi-scale, heterogeneous, streaming data **–BIG DATA**

- Cuts across entire spectrum of homeland security needs

We provide tools to enable end users to get the relevant information they need during any situation to make a decision or take action

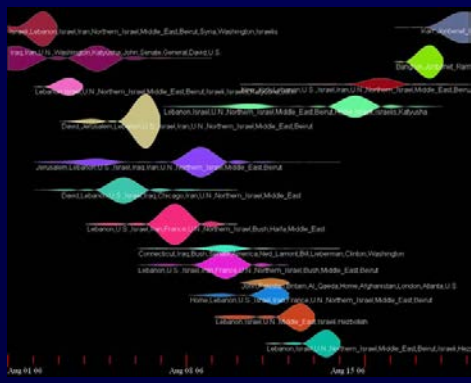






VACCINE Mission





- Provide visual analytic and scalable solutions to 2.3 million extended homeland security personnel
  - 185,000 DHS personnel, 350,000 law enforcement personnel, 750,000 homeland security practitioners
- Achieve excellence in visual analytics and visualization sciences
- Educate homeland security stakeholders and the next generation of talent



Big Data: Solution to Global Challenges?



- **Big Data Definition:** A phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex and diverse types of data. (TechAmerica Foundation)
  Big Data is often defined along three dimensions:
  - **Volume** – size
  - **Velocity** – rate of input, update, change
  - **Variety** – different types, sources, variables
- Need:
  - Advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of information (TechAmerica Foundation)
  - **Enable effective, efficient analysis, decision making, planning, and action**

Big Data Challenges

- Heterogeneity
- Unprecedented volume and/or complexity
- Multivariate, Multidimensional
- Streaming
- Qualitative, Quantitative, Categorical, Temporal,…
- Distributed Sources: sensors, text, images, maps, …
- Inconsistent and Uncertain
- Different Scales and Cross Scale

Big Data Analytics – The Solution?

- Advanced data analytics for big data can help manage big data
  - Machine learning
  - Advanced statistical analysis
  - Hadoop-deployed analytics
- However, it falls short in several aspects

What Big Data Analytics Can't Do
(inspired and adapted from David Brooks, New York Times, 2/18/2013)

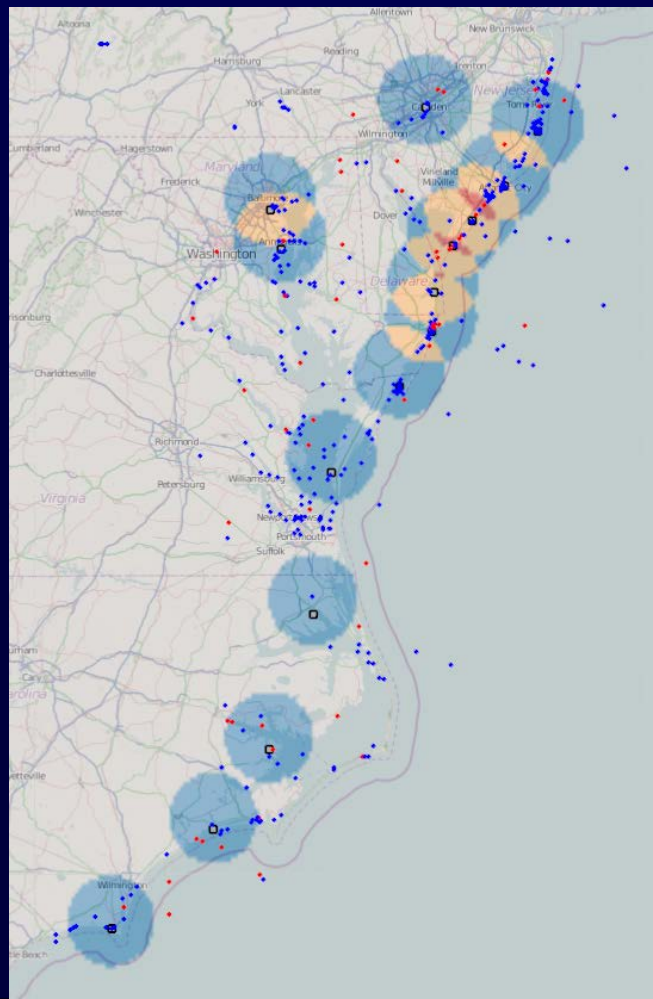
- Qualitative, fuzzy, and social data
  - Preferences, significance of one relationship over another
  - Trust
- Context
  - Data is rarely complete, nor does it incorporate all relevant information for decision making
  - Humans have extensive information and experience that never make it into the collected data

What Big Data Analytics Can't Do
(inspired and adapted from David Brooks, New York Times, 2/18/2013)

- Spurious vs. Significant
  - Big data means more statistically significant events and correlations, but they may not have any relevance
  - Increases noise to signal ratio
- Big problems
  - Complex, multifaceted, multiparameter big challenges with unquantified dependencies

Visualization, Data, and Decision Making

- At some level most decision making is driven by data
- What role does visual representation play?
- Good or bad – people believe charts and graphics – “seeing is believing”



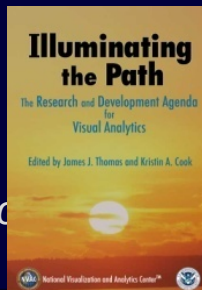
Beyond Visualization to Visual Analytics

Visual Analytics[1] is the science of analytical reasoning facilitated by interactive visual interfaces

Interactive visualization, data analysis, exploration, and decision making with human in the loop!

People use visual analytics tools and techniques to

- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
- Detect the expected and discover the unexpected
- Provide timely, defensible, understandable assessments
- Communicate assessment effectively for action



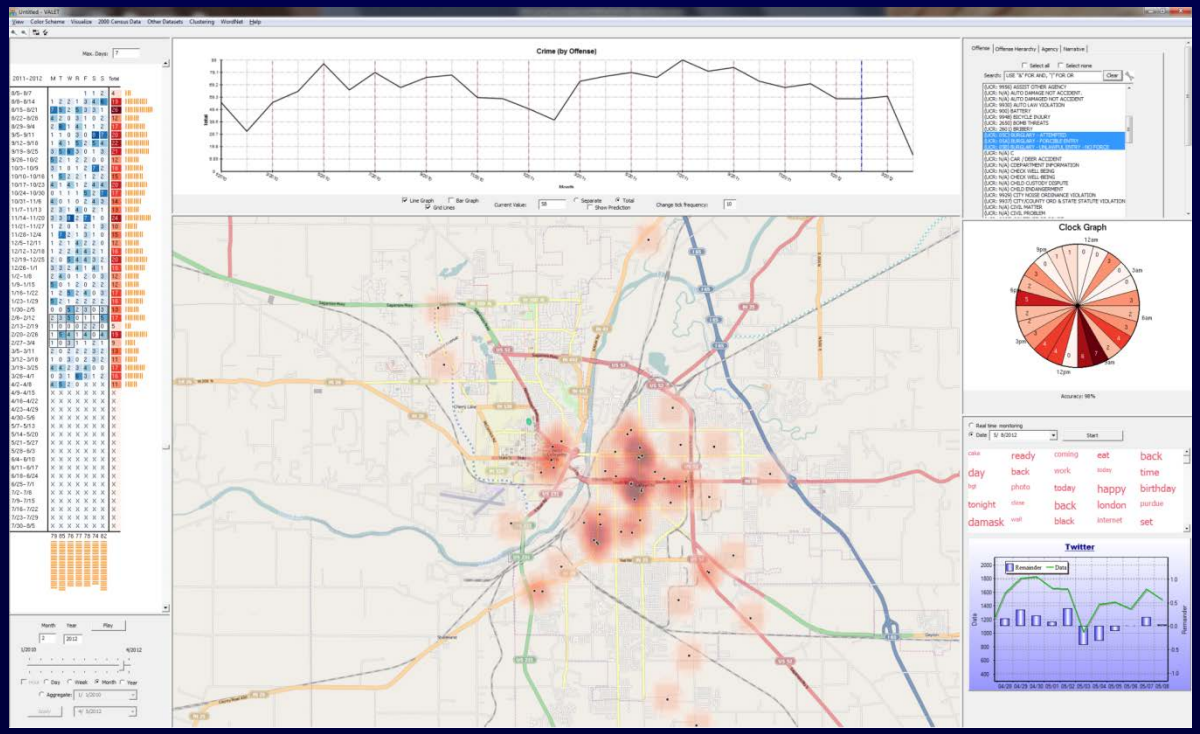
1. *Illuminating the Path: The R&D Agenda fo*
December 2005 *Visual Analytics*, Editors: Thomas and

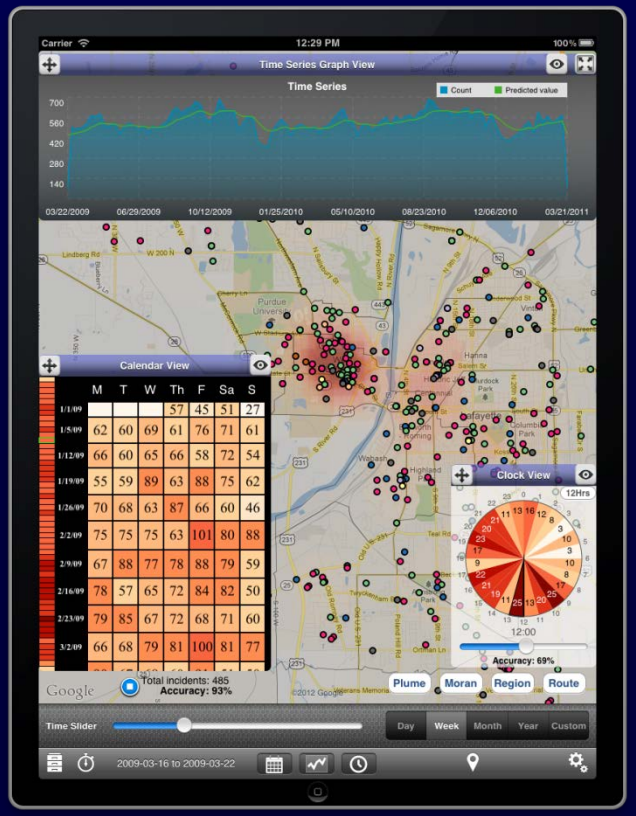
VACCINE

VA Approaches to Big Data

- Don’t display all the data!
- Extract relevant information at the appropriate natural scales using analytics and statistical models to reduce data space so that it is mappable to the visual space
- User interaction to guide above process
- Adapt to available hardware for the problem
  - Should we use big iron or cell phones?

Visual Analytics Law Enforcement Toolkit (i)VALET





VALET

Visual Analytics Law Enforcement Toolkit (VALET, iVALET)

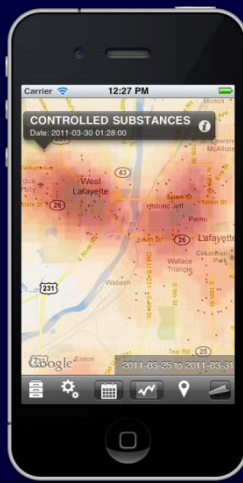
Impacts:

- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

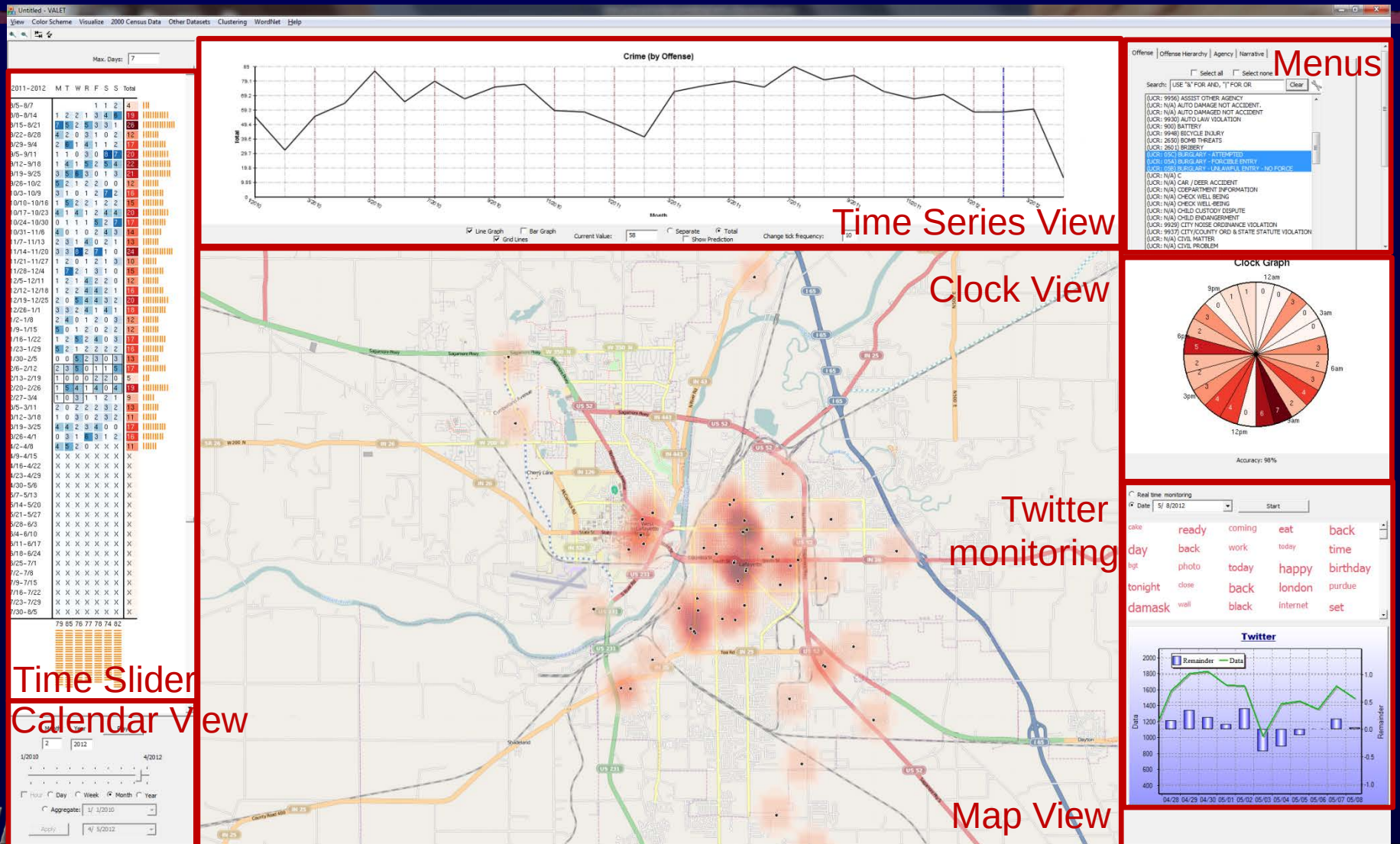
iVALET
Debriefing Example

VALET delivered:
- Spring 2011: WL, Lafayette Police

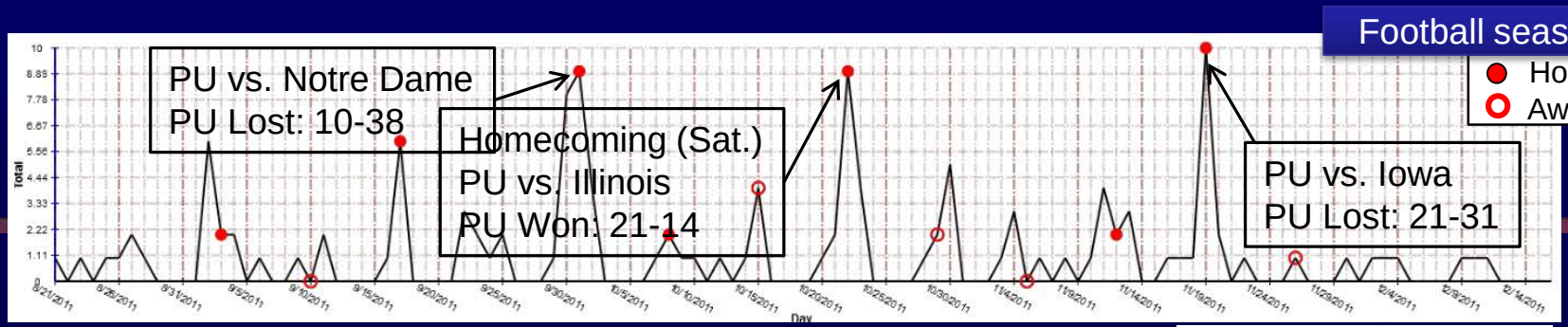
iVALET delivered:
- October 2011: Purdue, WL Police

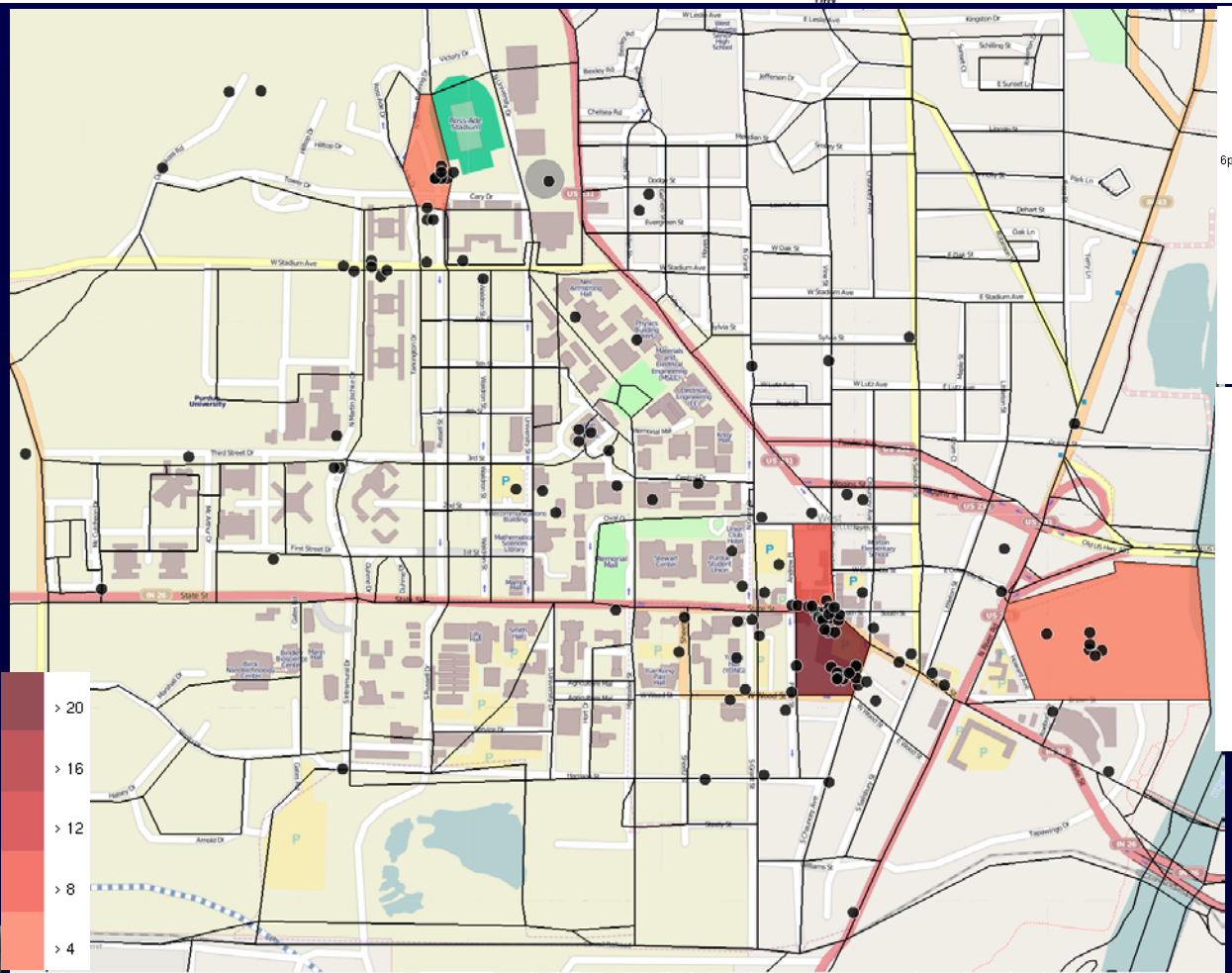


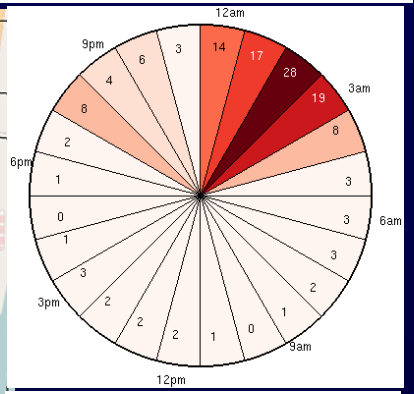
VALET Overview

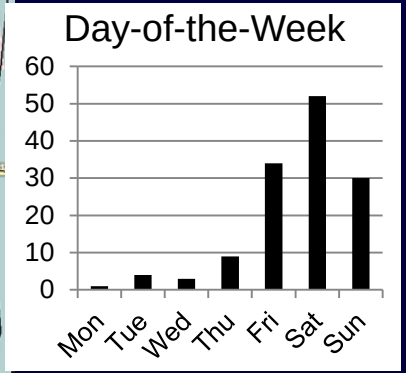


Example: Drunkenness / Public Intoxication



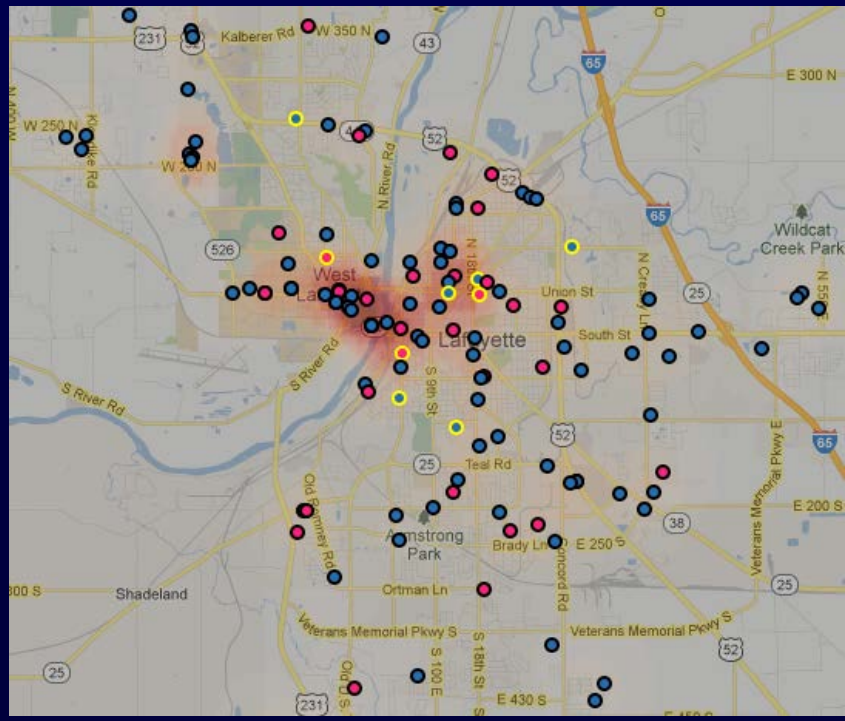






Top 10 Hot Incidents

- Identify unusual localized high-frequency patterns of crimes (near repeats)
- Each data entry is checked for other crimes with similar properties within a 1 block radius of the incident location and a 14-day time period
- Top 10 incidents with the most number of related incidents in this space-time window are highlighted



VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

Visual Analytics for Investigation and Law Enforcement:

David S. Ebert

CVADA

Center for Visual Analytics and Data Analytics

A U.S. Department of Homeland Security Center of Excellence

October 2013

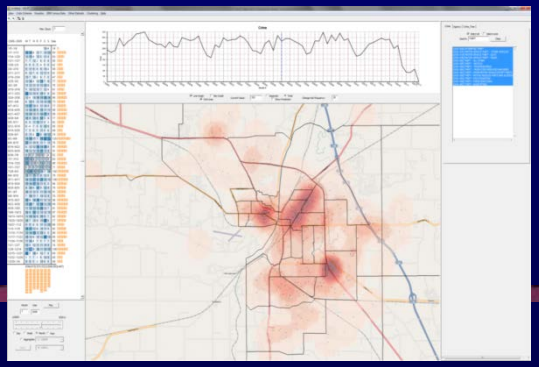
Engaged End-Users

- Federal Operating Components:
  - US Coast Guard
  - US Transportation Security Administration
  - US Citizenship and Immigration Services
  - US Federal Emergency Management Agency
  - US Customs and Border Protection
  - US CERT
  - US Immigration and Customs Enforcement (in progress)
- Law Enforcement
  - Over 40 local and state agencies (IN, IL, OH, SC, PA, NC, NY)
- Fusion Centers
  - Ohio (SAIC)
  - Indiana (IIFC)

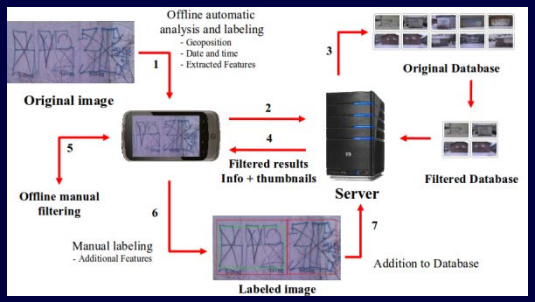


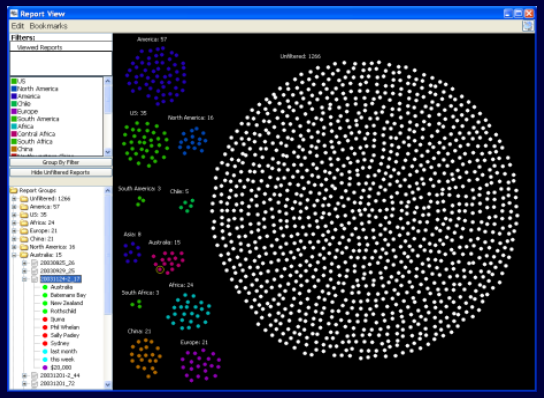
Law Enforcement

- Law enforcement visual analytics
  - VALET, iVALET (iPhone/iPad), CrimeViz
    - Visual Analytics Law Enforcement Toolkit
    - Analyzing crime patterns and time of day problems
- Gang activity analytics
  - GARI
    - Gang Graffiti Recognition and Interpretation using a mobile telephone
    - Allowing police to catalog and analyze gang graffiti images, better track and determine gang activities
- Document visual analytics
  - JIGSAW
    - Visualization for investigative analysis
    - Discovery of hidden relationship and threats across documents

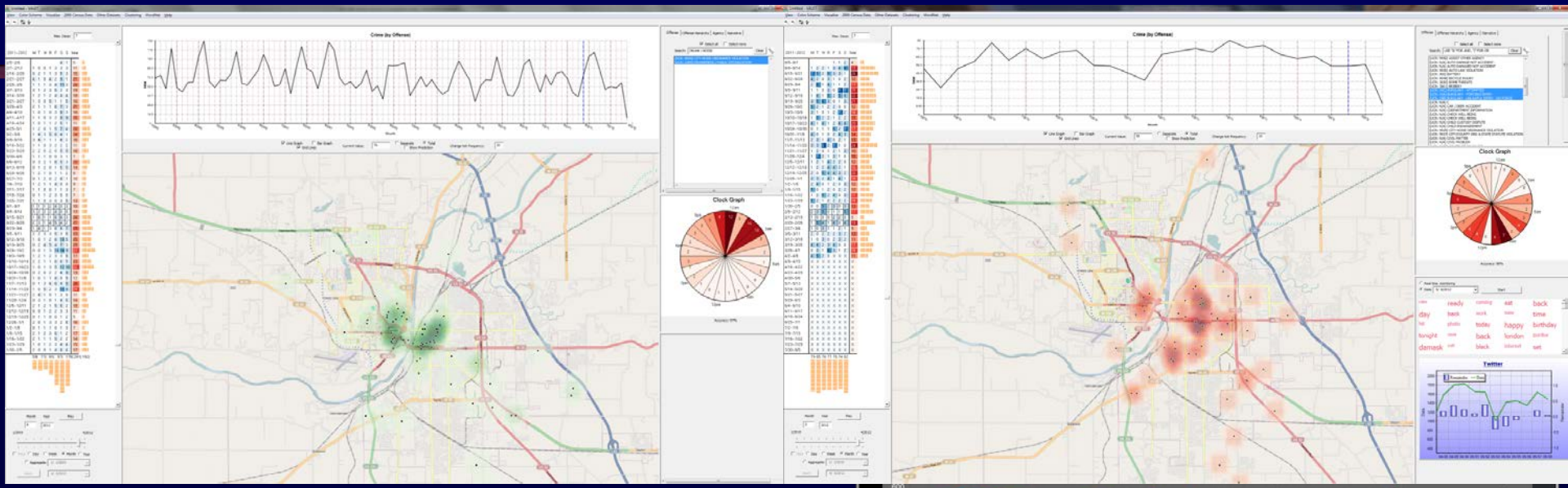




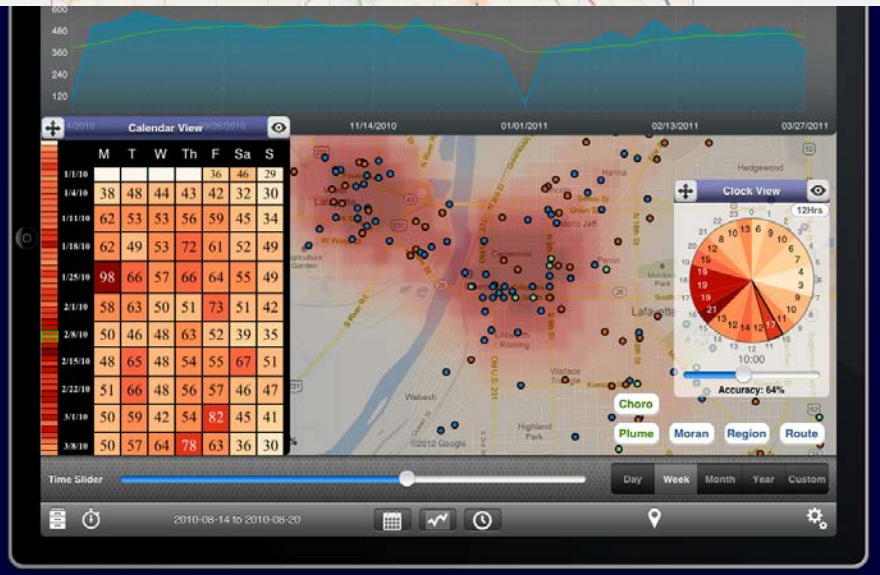




Visual Analytics Law Enforcement Toolkit (VALET)







VALET

Visual Analytics Law Enforcement Toolkit (VALET, iVALET)

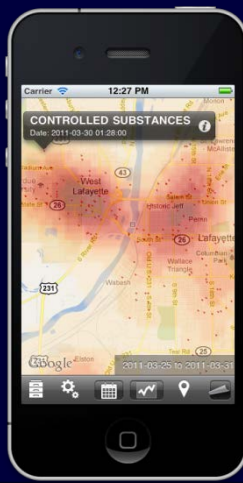
Impacts:

- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

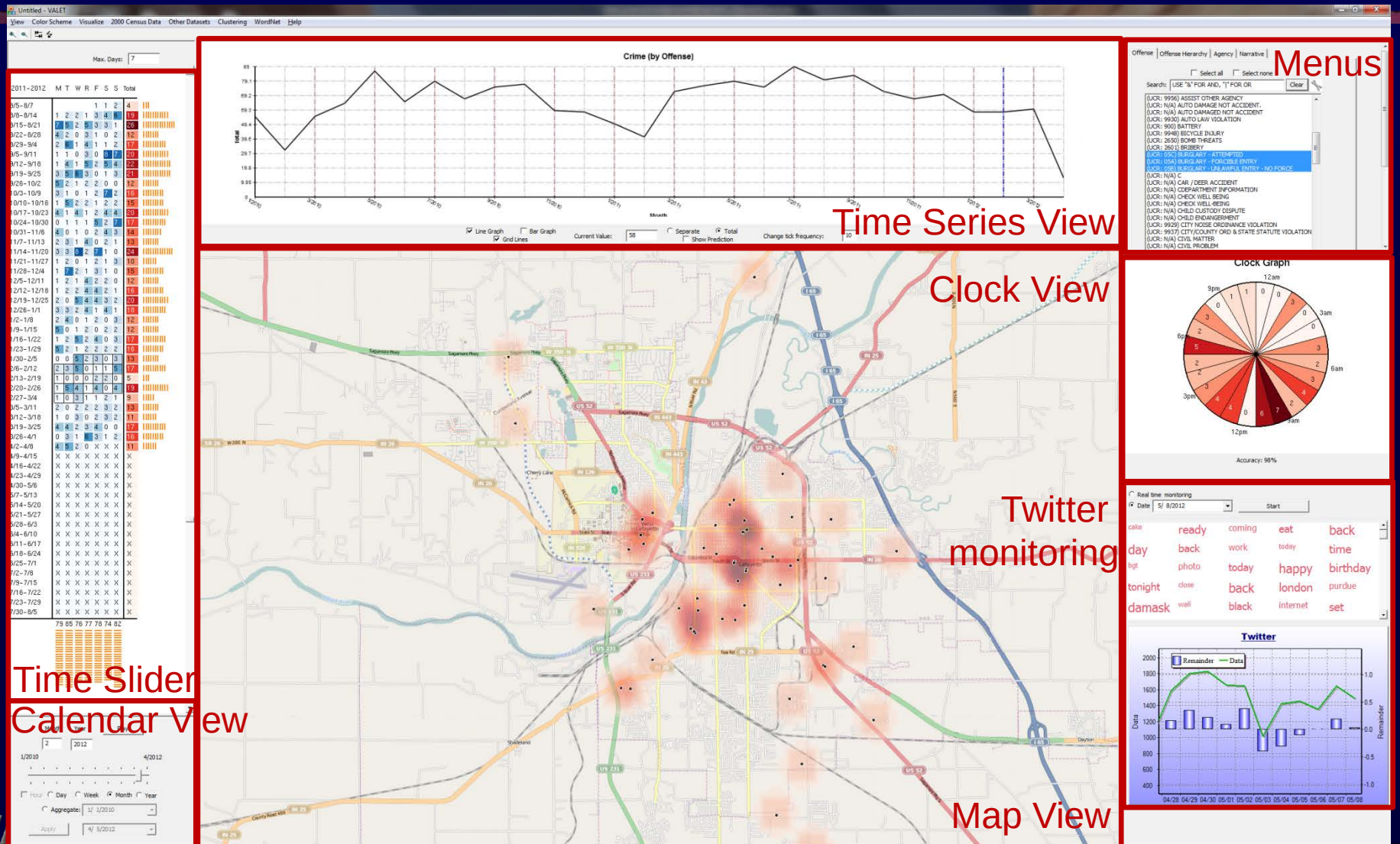
iVALET
Debriefing Example

VALET delivered:
- Spring 2011: WL, Lafayette Police

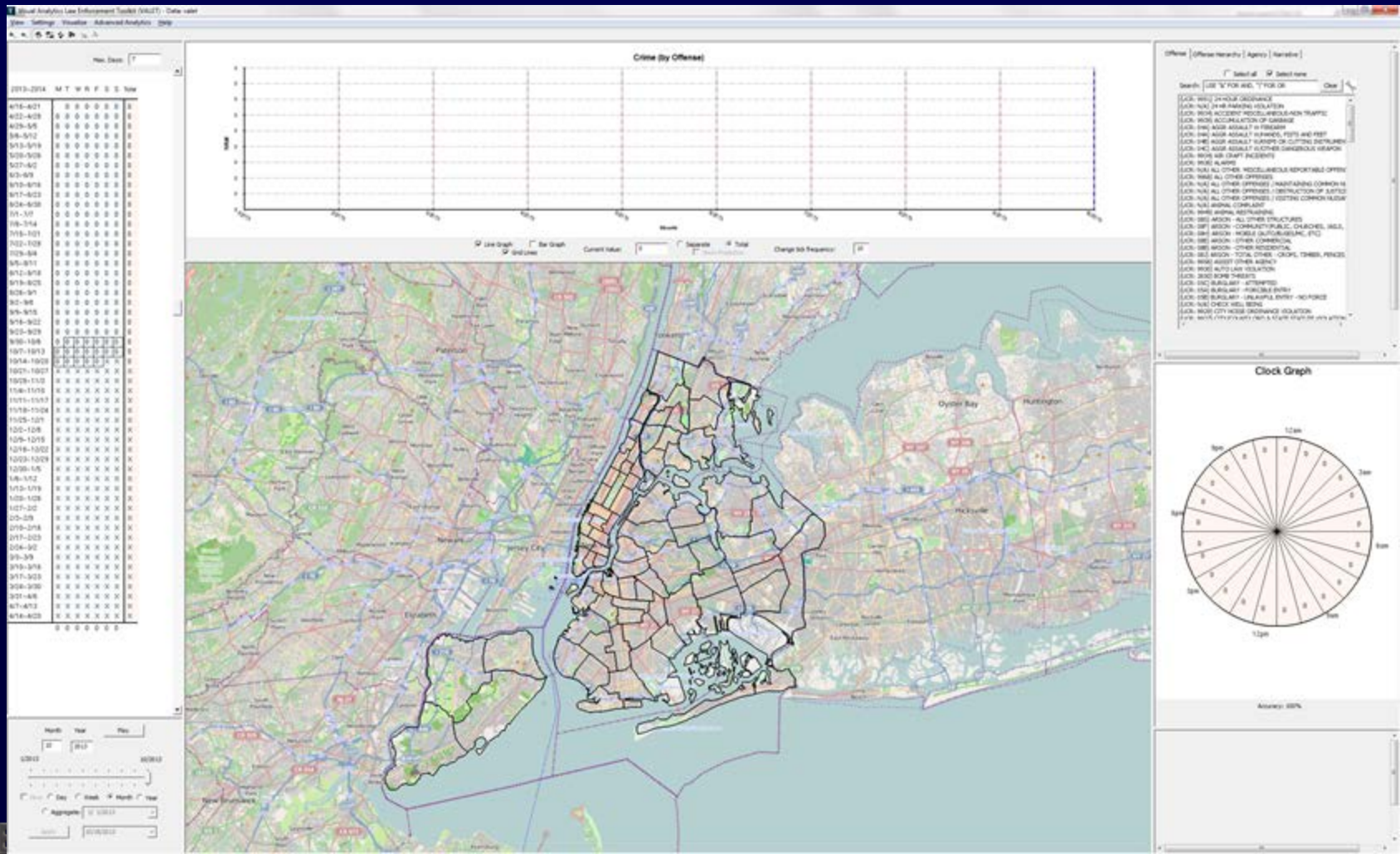
iVALET delivered:
- October 2011: Purdue, WL Police



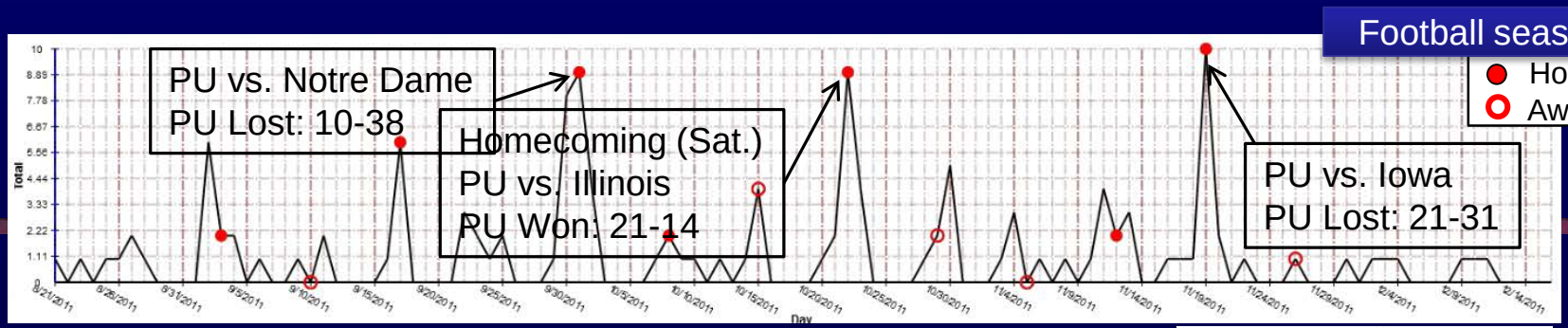
VALET Overview

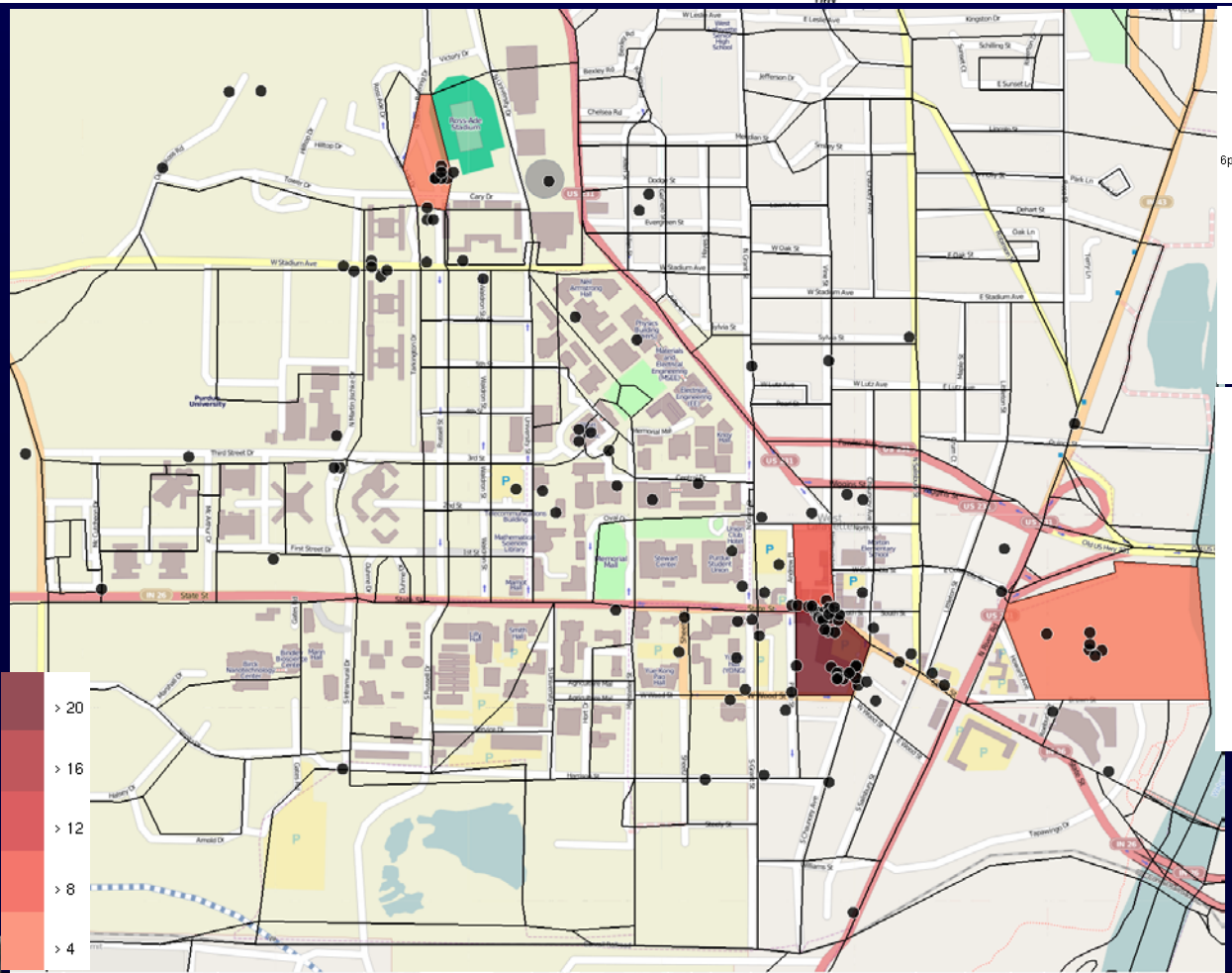


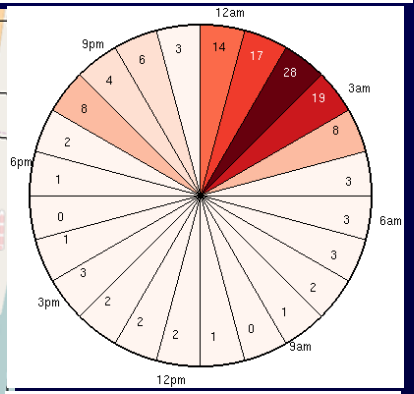
VALET – In Testing NYPD

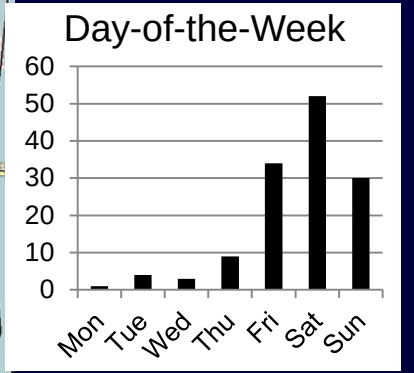


Example: Drunkenness / Public Intoxication

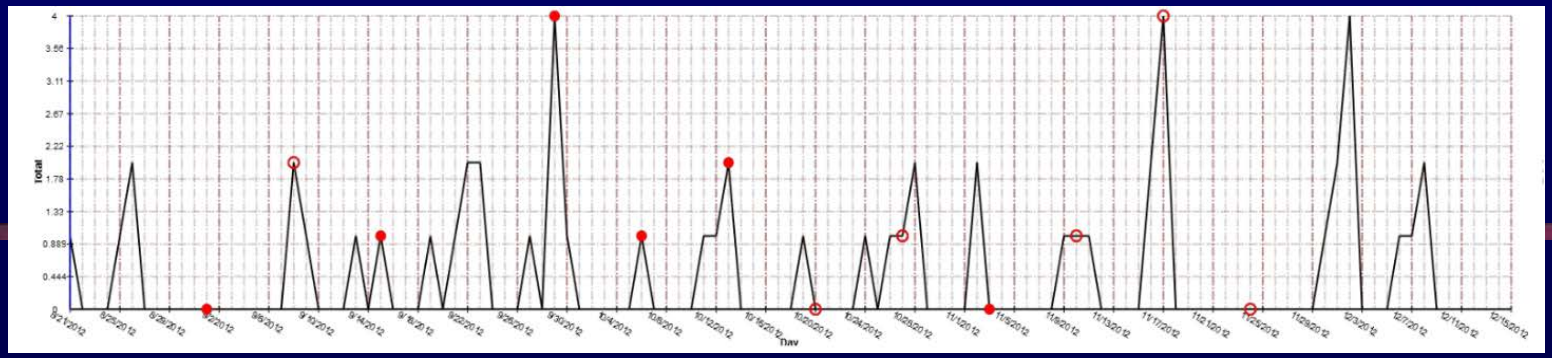


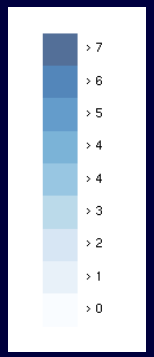


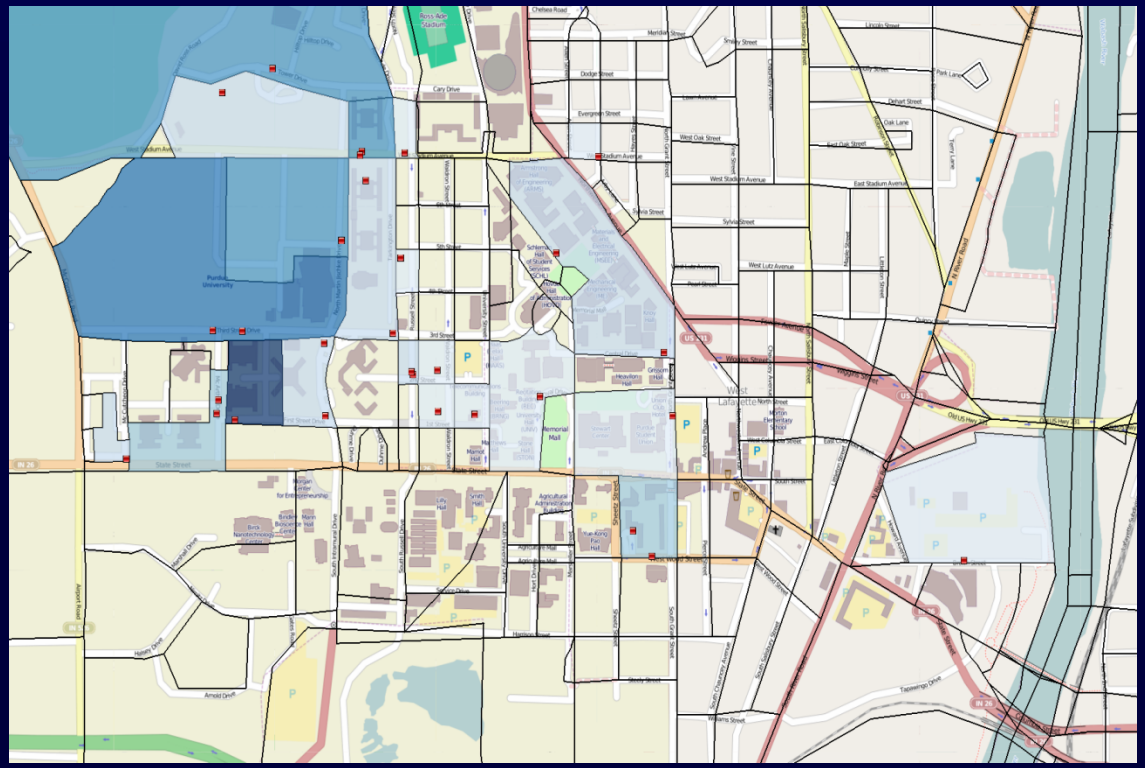


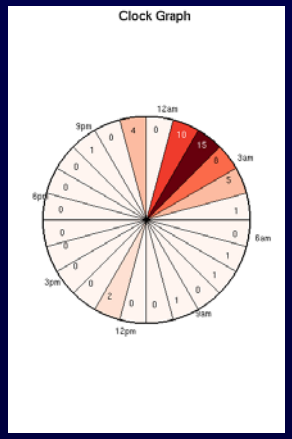


EMS Calls to Alcohol Related Incidents – Fall 2012

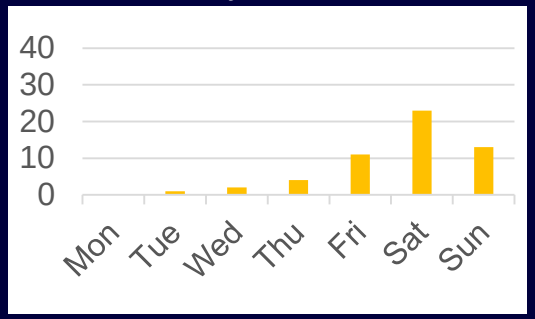




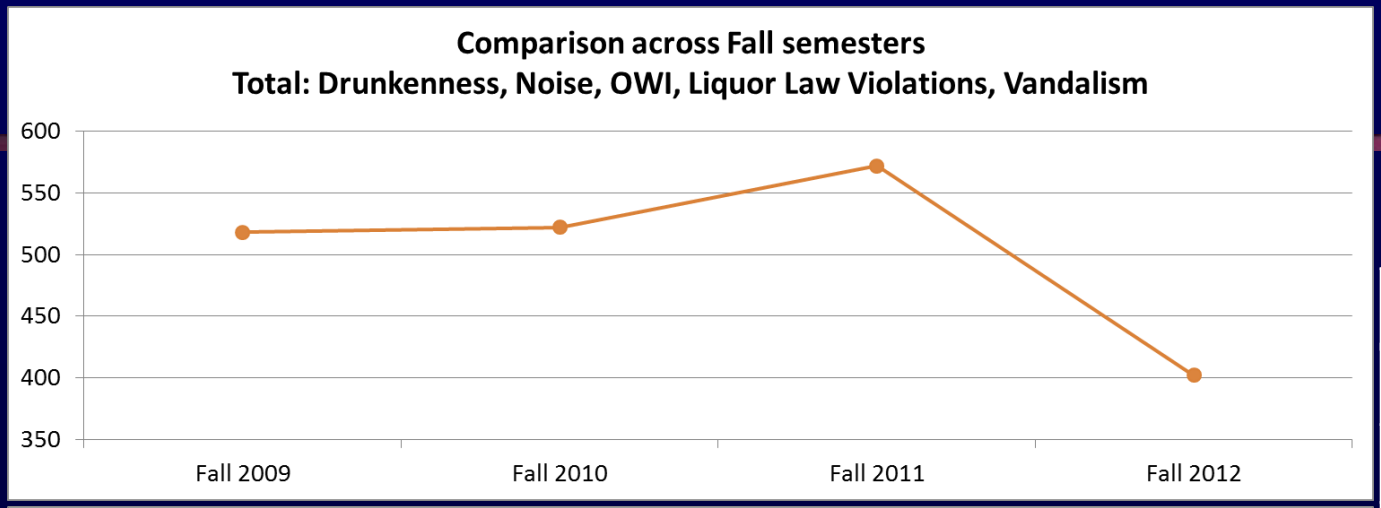


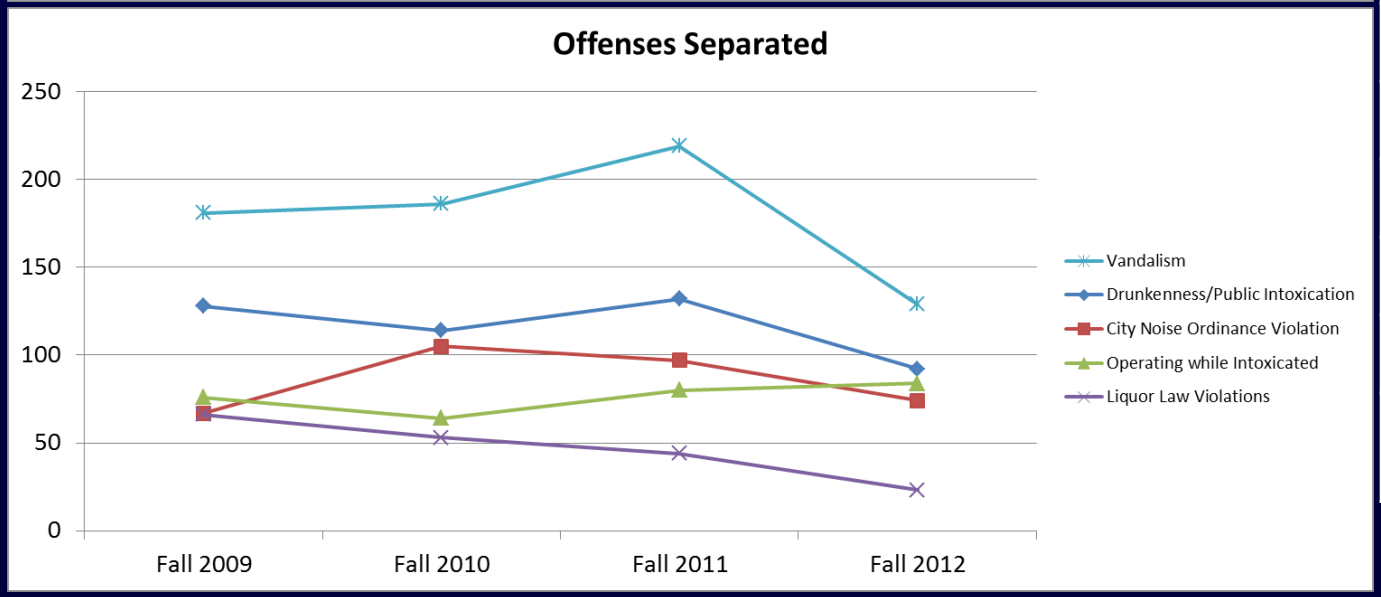


Accuracy 91%



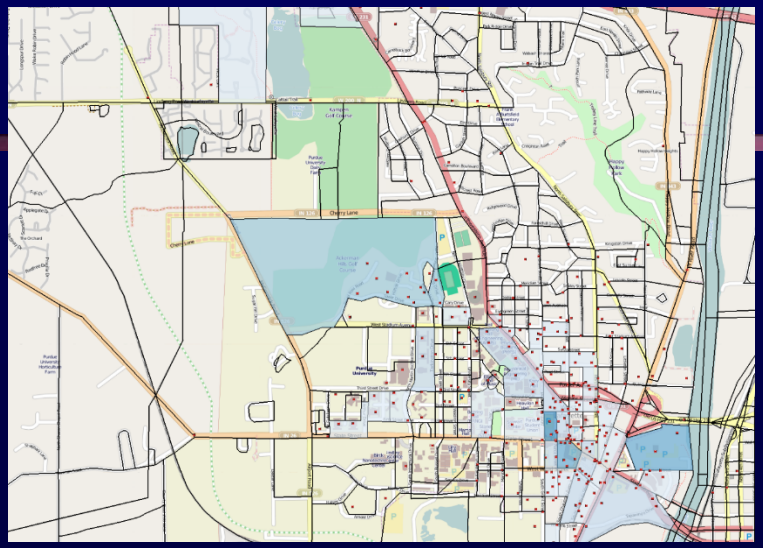
Comparison across Fall semesters



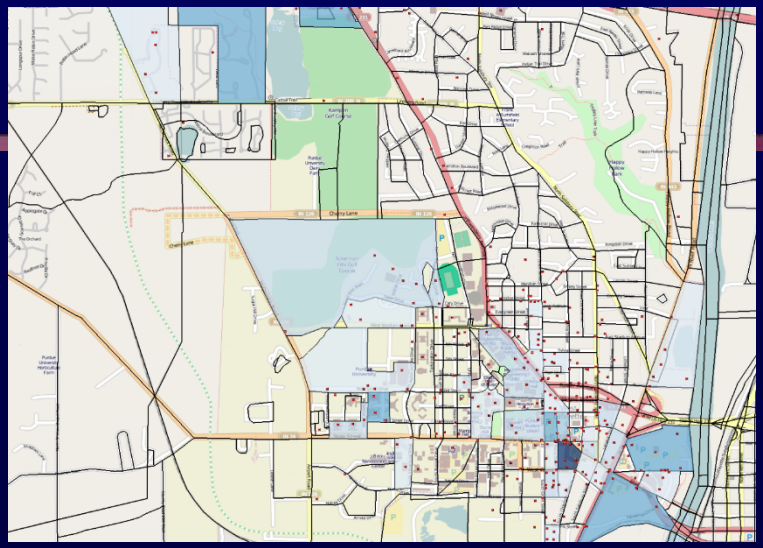


Fall 2011 to Fall 2012	
Offense	**% Change**
Total	−29.7%
Vandalism	−41.1%
Drunkenness	−30.3%
Noise	−23.7%
OWI	+5.0%
Liquor law violations	-47.7%

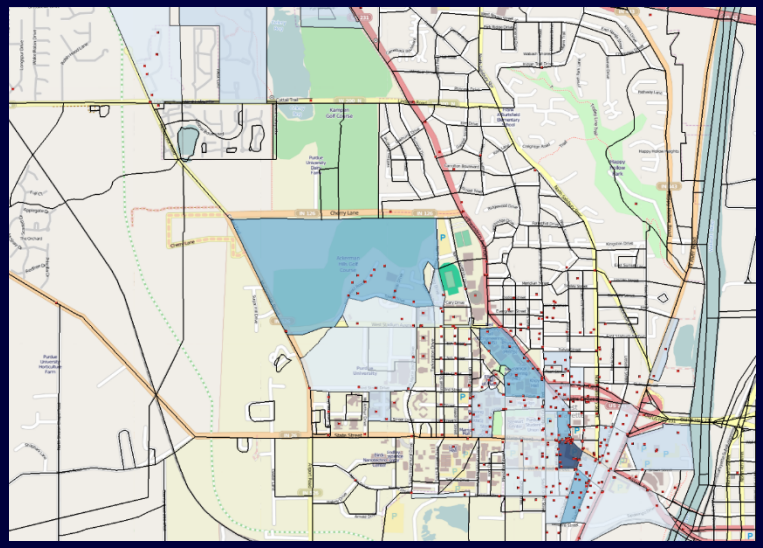
Comparison Across Fall Semesters



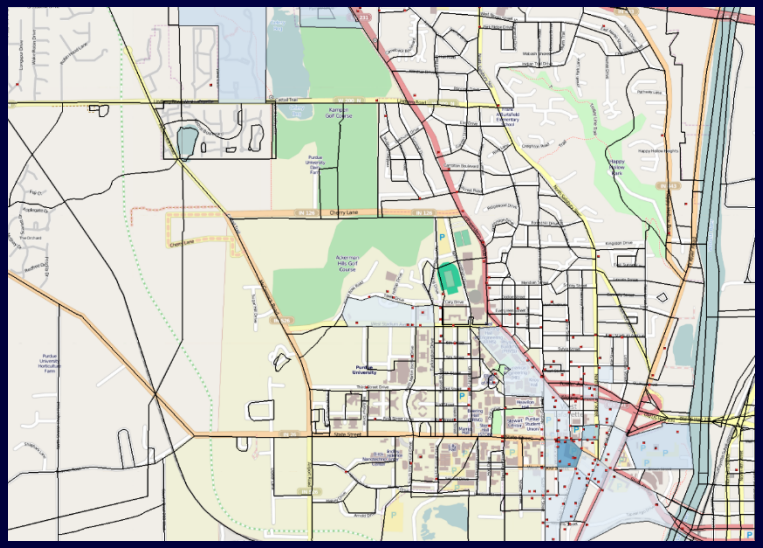
2009



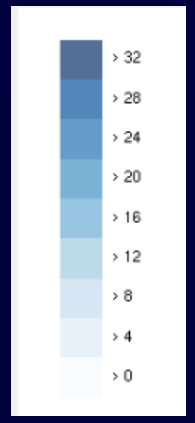
2010



2011

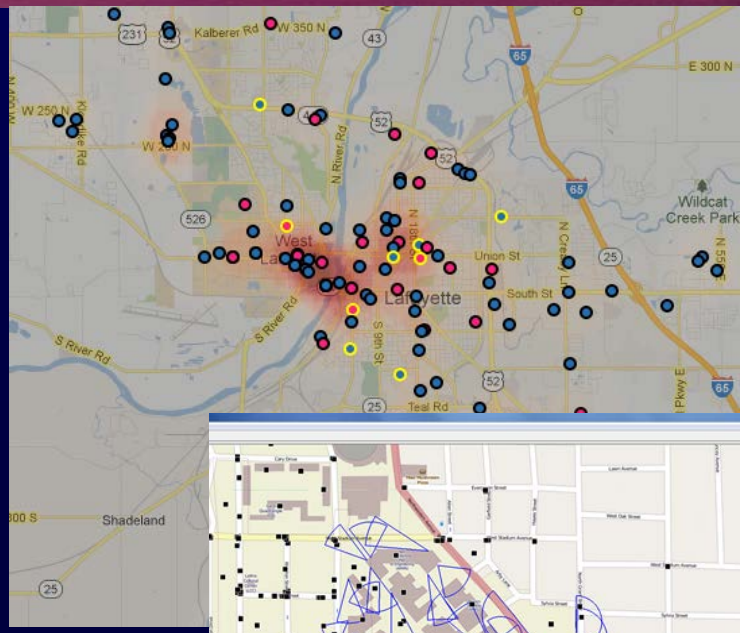


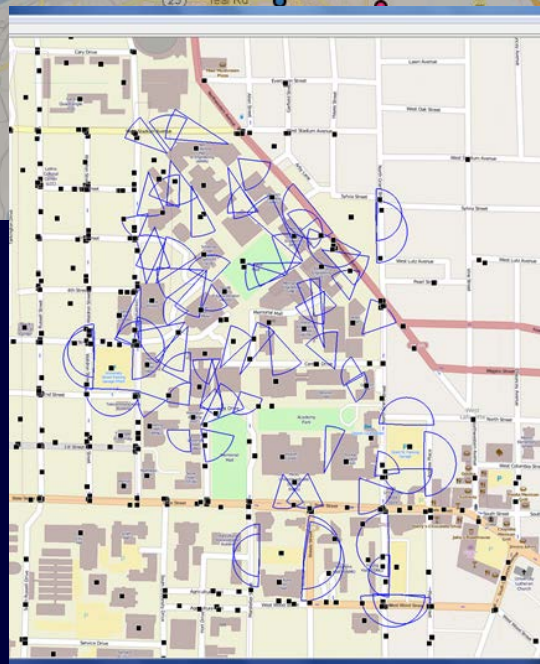
2012



Top 10 Hot Incidents, Trending, Cameras

- Identify unusual localized high-frequency patterns of crimes (near repeats)
- Each data entry is checked for other crimes with similar properties (within a 1 block radius and a 14-day time period)
- Top 10 incidents with the most number of related incidents in this space-time window are highlighted
- Trending: Case and search trends
- Cameras:
  - Where to put next one?
  - Which camera covered this crime?

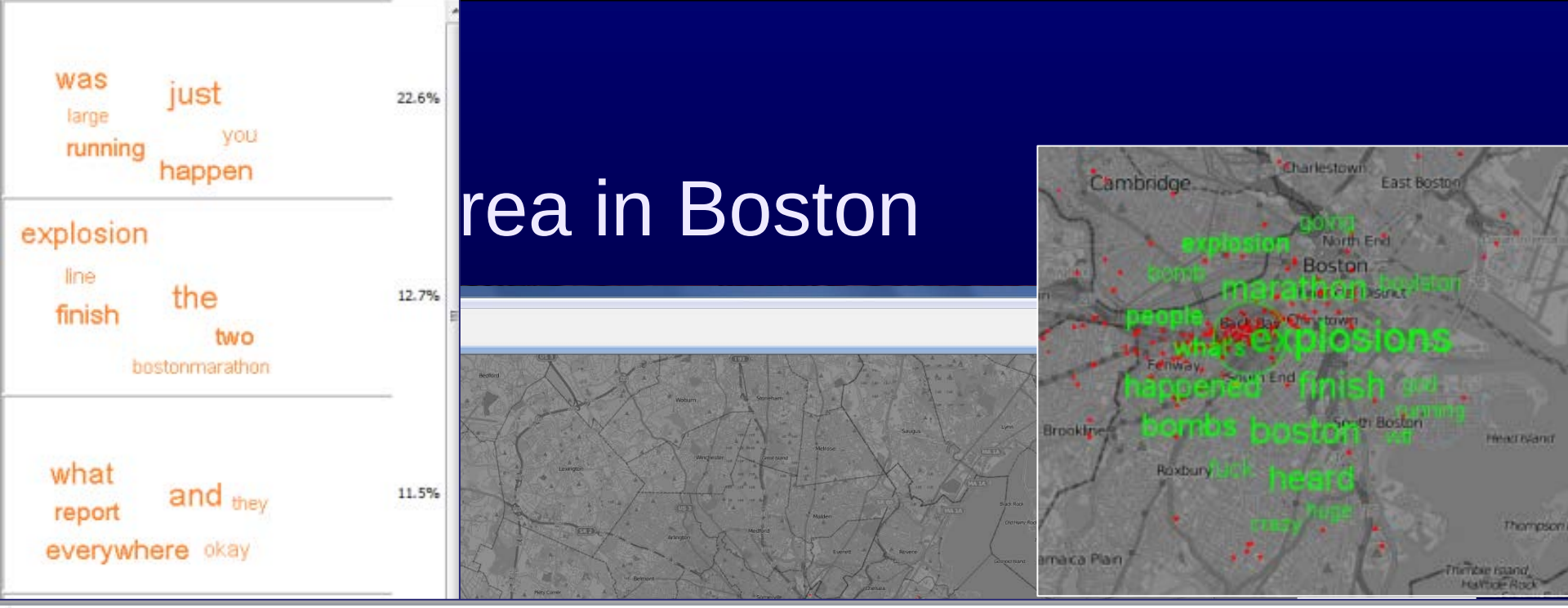




Social Media: Real-time Twitter Monitoring and Integration into Tools (Purdue, Stuttgart, Penn St.)

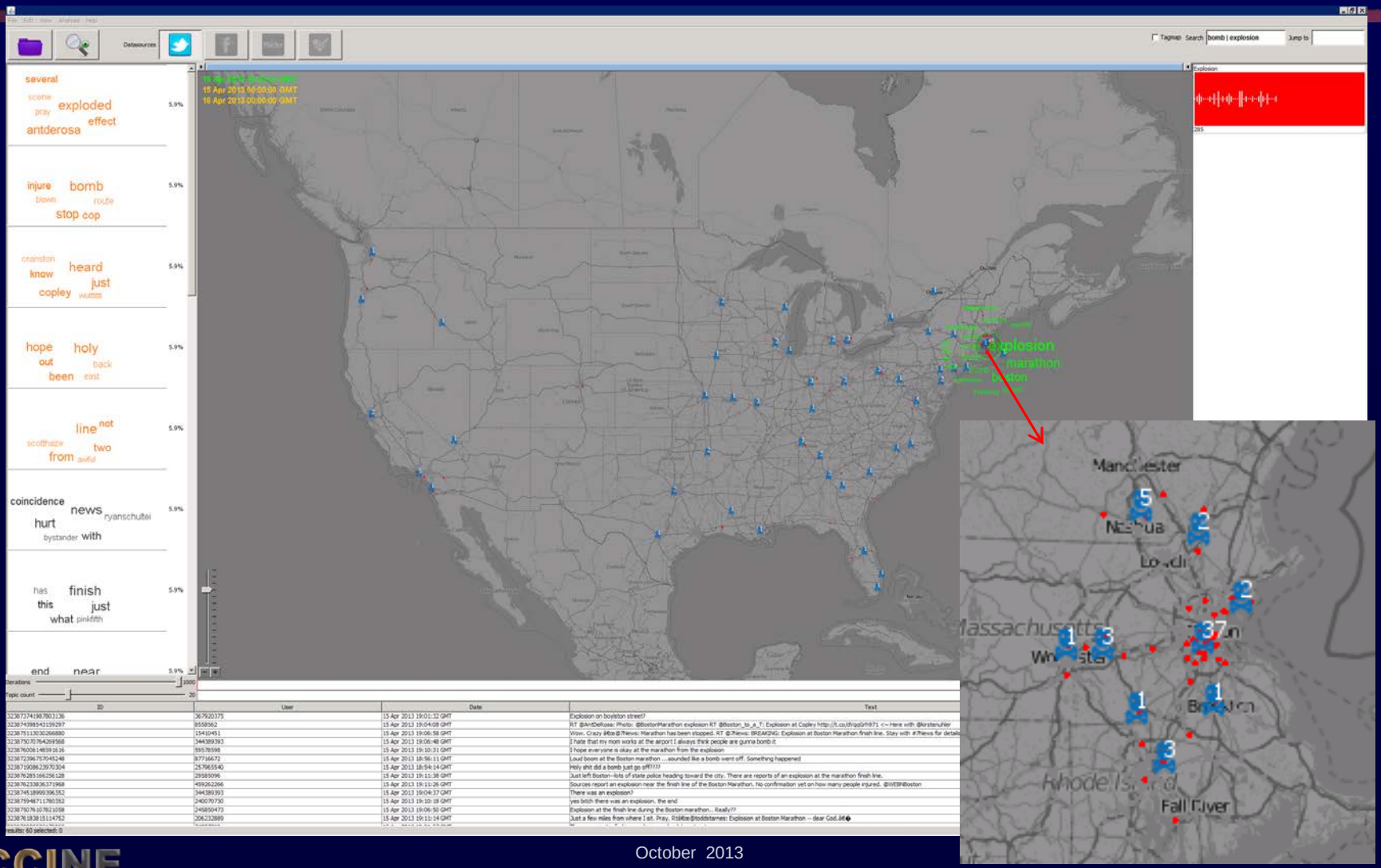
- Topic extraction using novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud





Text
Oh my god what just happened
Something happened at the end of the Boston Marathon. Something bad and there is a lot of chatter on Twitter. What's going on?
Multiple people are injured near the Boston Marathon finish line after two explosions. The #BostonMarathon has been stopped.
two bombs just went off on boylston
BREAKING NEWS: Two powerful explosions detonated in quick succession right next to the Boston Marathon finsh line this afternoon.
What the FUCK was that
Back in Sept, @croon1 solicited me for $2000. He now has a music video with William Shatner. If you watch it (god forbid) keep that on mind.
Literally what the fuck get me out of here
@DTenenbaum my office right next to it
@FRANCESCalciO I figured 3 people would get the joke
Two explosions just rocked the finish line of the Boston Marathon. Sirens galore. People running in fear. Wonder what happened.
This is crazy i seen that blow up #bostonmarathon
can someone tell me what that explosion was!? #boston #bostonmarathon

Detection using the Explosion Classifier



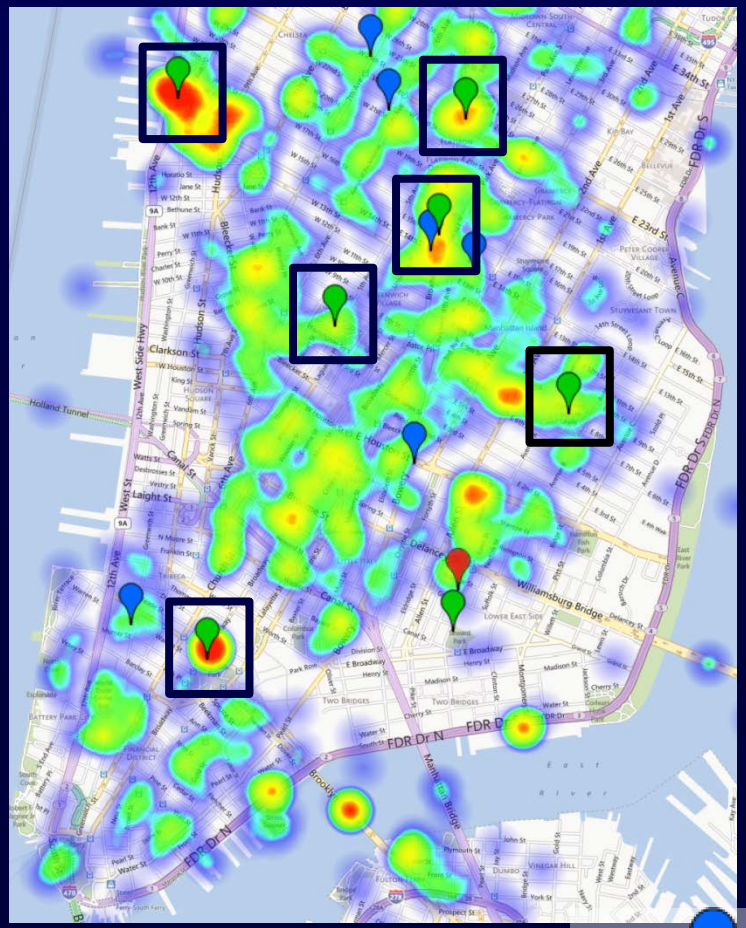
First Response (Tweet & Picture)

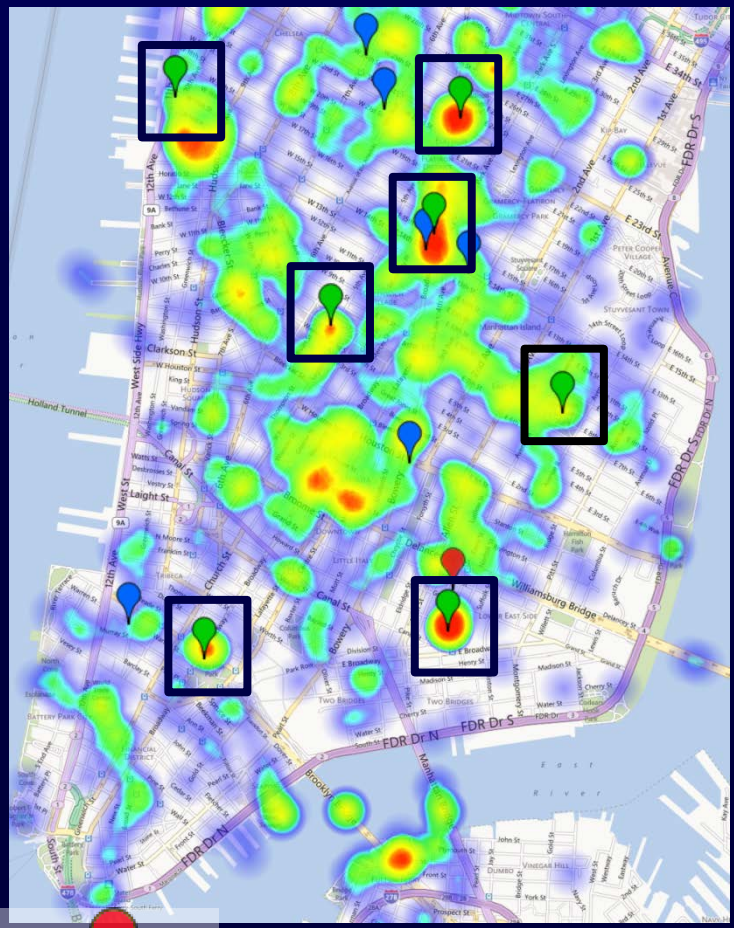


Visual Analytics of Activity During Hurricane Sandy

Two weeks before Sandy
10/14 (Sunday), 12:00 ~ 16:00

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00





Supermarket Park Shelter

October 2013

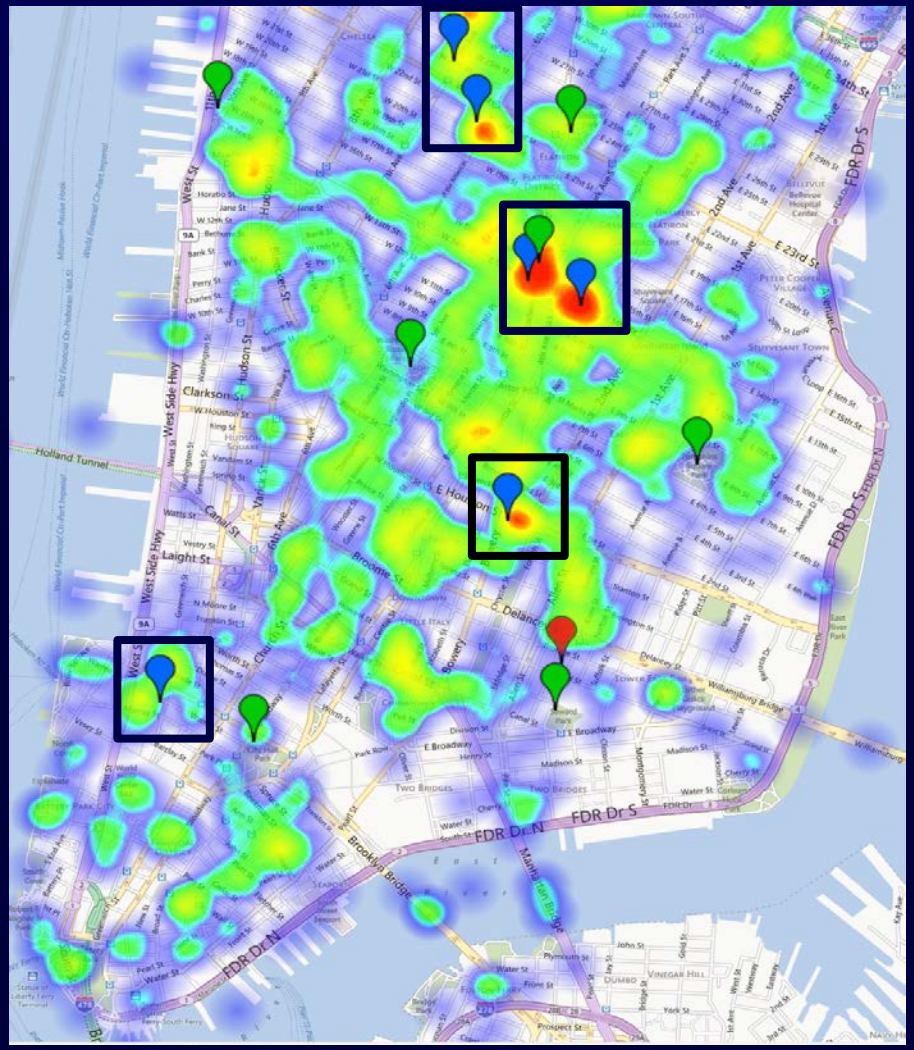
VACCINE

Visual Analytics of Activity During Hurricane Sandy

After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00

Evacuation order:
10/28, 10:30 AM

Hurricane Sandy's
Arrival at NYC:
10/29, 8:00 PM

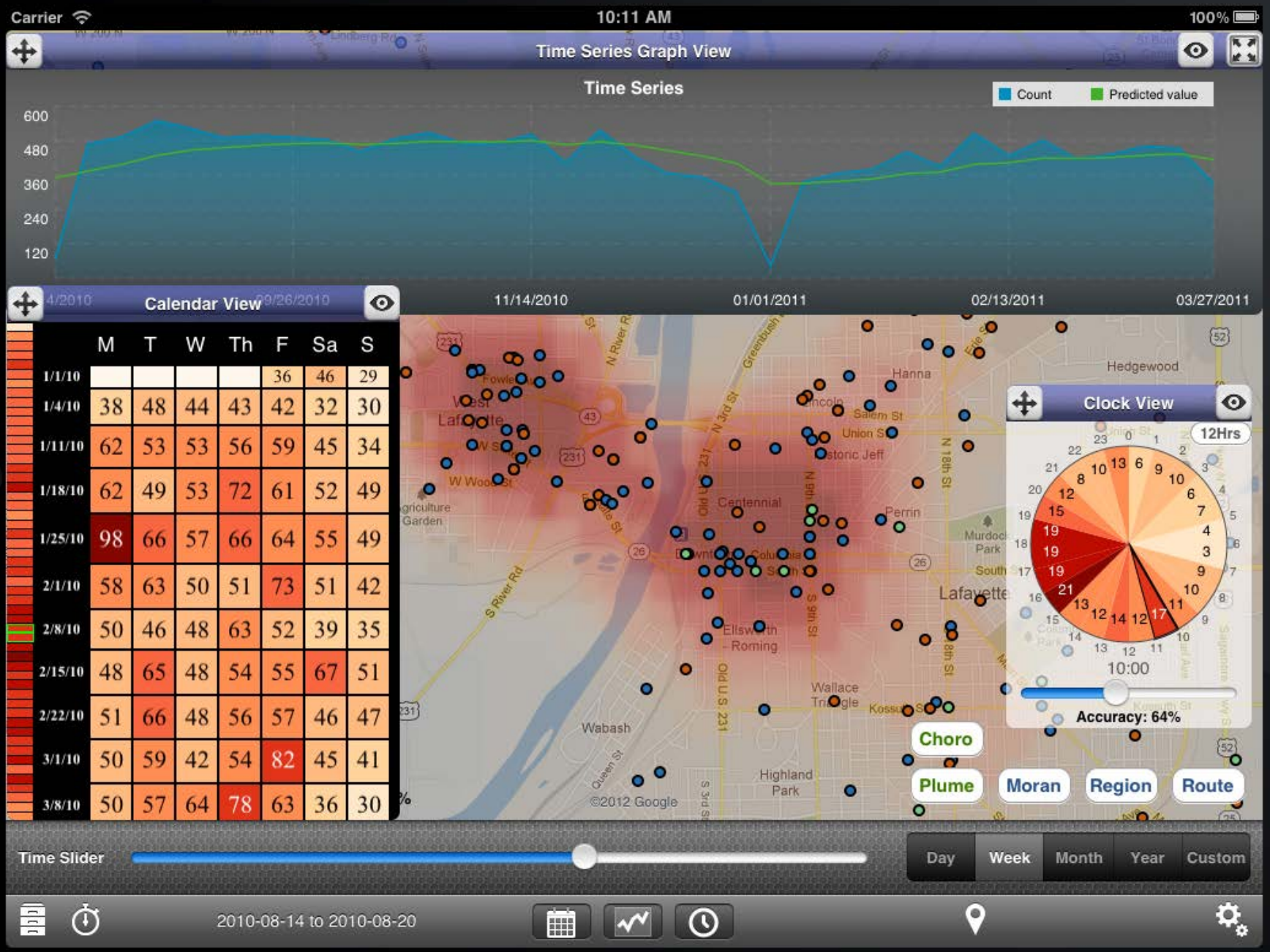


Supermarket

Park

Shelter

VACCINE


Carrier
10:11 AM
100%
Time Series Graph View
Time Series
Count
Predicted value
600
480
360
240
120
11/14/2010
01/01/2011
02/13/2011
03/27/2011
Calendar View
M T W Th F Sa S
1/1/10 36 46 29
1/4/10 38 48 44 43 42 32 30
1/11/10 62 53 53 56 59 45 34
1/18/10 62 49 53 72 61 52 49
1/25/10 98 66 57 66 64 55 49
2/1/10 58 63 50 51 73 51 42
2/8/10 50 46 48 63 52 39 35
2/15/10 48 65 48 54 55 67 51
2/22/10 51 66 48 56 57 46 47
3/1/10 50 59 42 54 82 45 41
3/8/10 50 57 64 78 63 36 30
Clock View
12Hrs
10:00
Accuracy: 64%
Choro
Plume
Moran
Region
Route
Time Slider
Day
Week
Month
Year
Custom
2010-08-14 to 2010-08-20
©2012 Google

Jigsaw

Visualization for Investigative Analysis across Document Collections

- Law enforcement & intelligence community
- Fraud (finance, accounting, banking)
- Academic research
- Journalism & reporting
- Consumer research

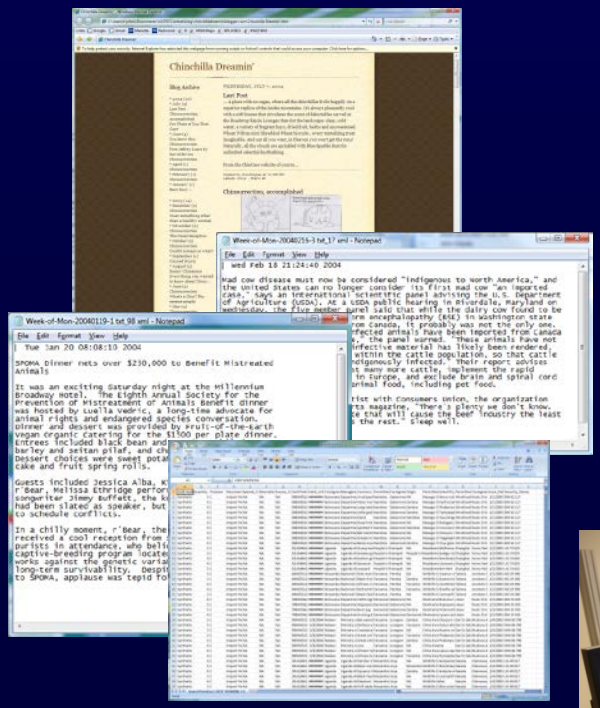
- Help “investigators” explore, analyze and understand large document collections



“Putting the pieces together”

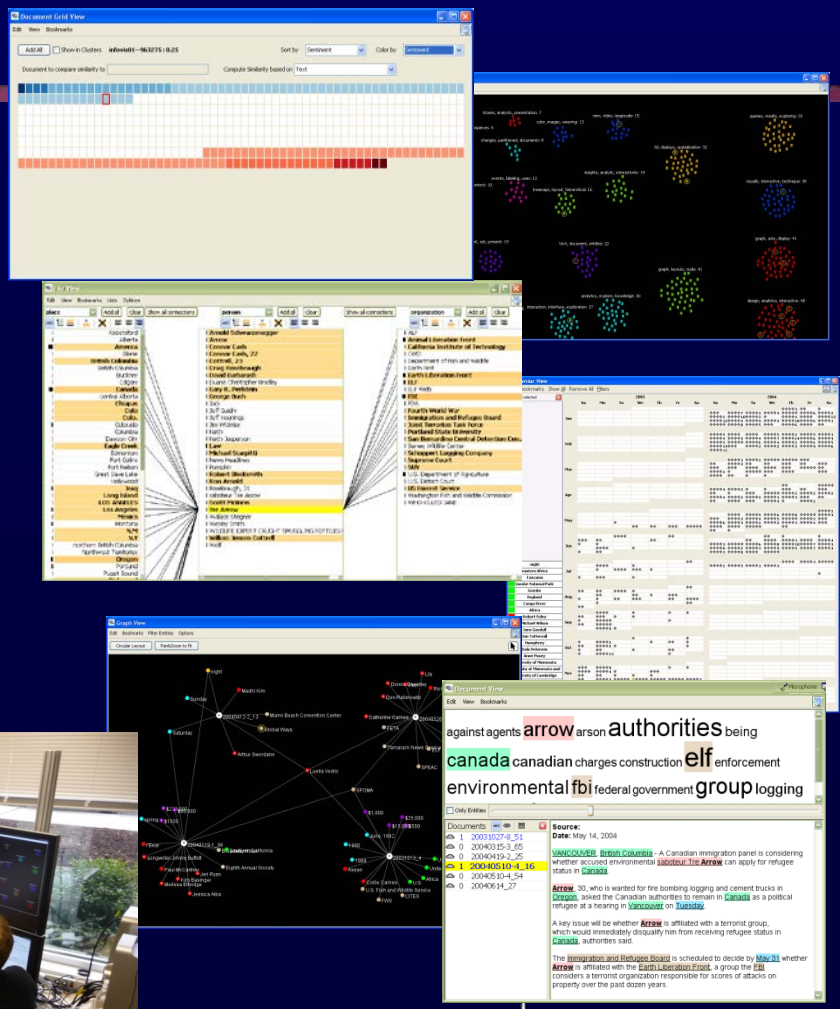
Jigsaw: Visual Analytics for Investigative Analysis

John Stasko



Documents in

Knowledge out





Georgia Tech College of Computing
School of Interactive Computing

VACCINE

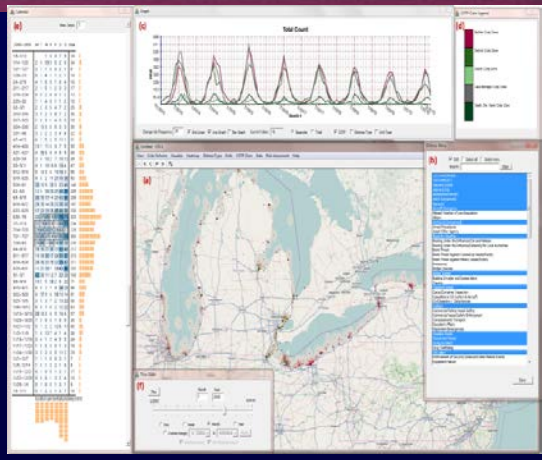
USCG: Effective Risk-based Decision Making and Resource Allocation Visual Analytics

- Evaluate current and historical mission area:
  - Demands
  - Risks (total, mitigated, residual)
  - Resource allocation
  - Return on investment
- Evaluate courses of action
- Evaluate above at both Strategic and Tactical/Operational level



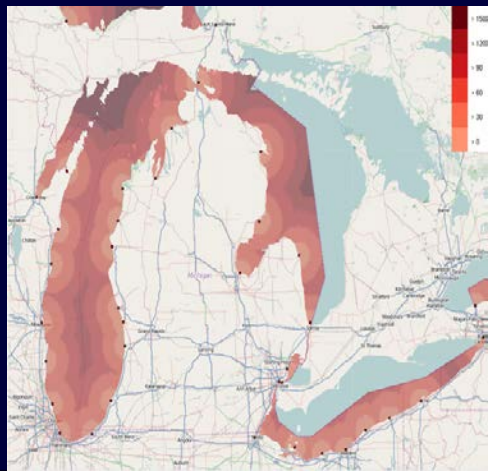
U.S. Coast Guard Search and Rescue VA (cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771



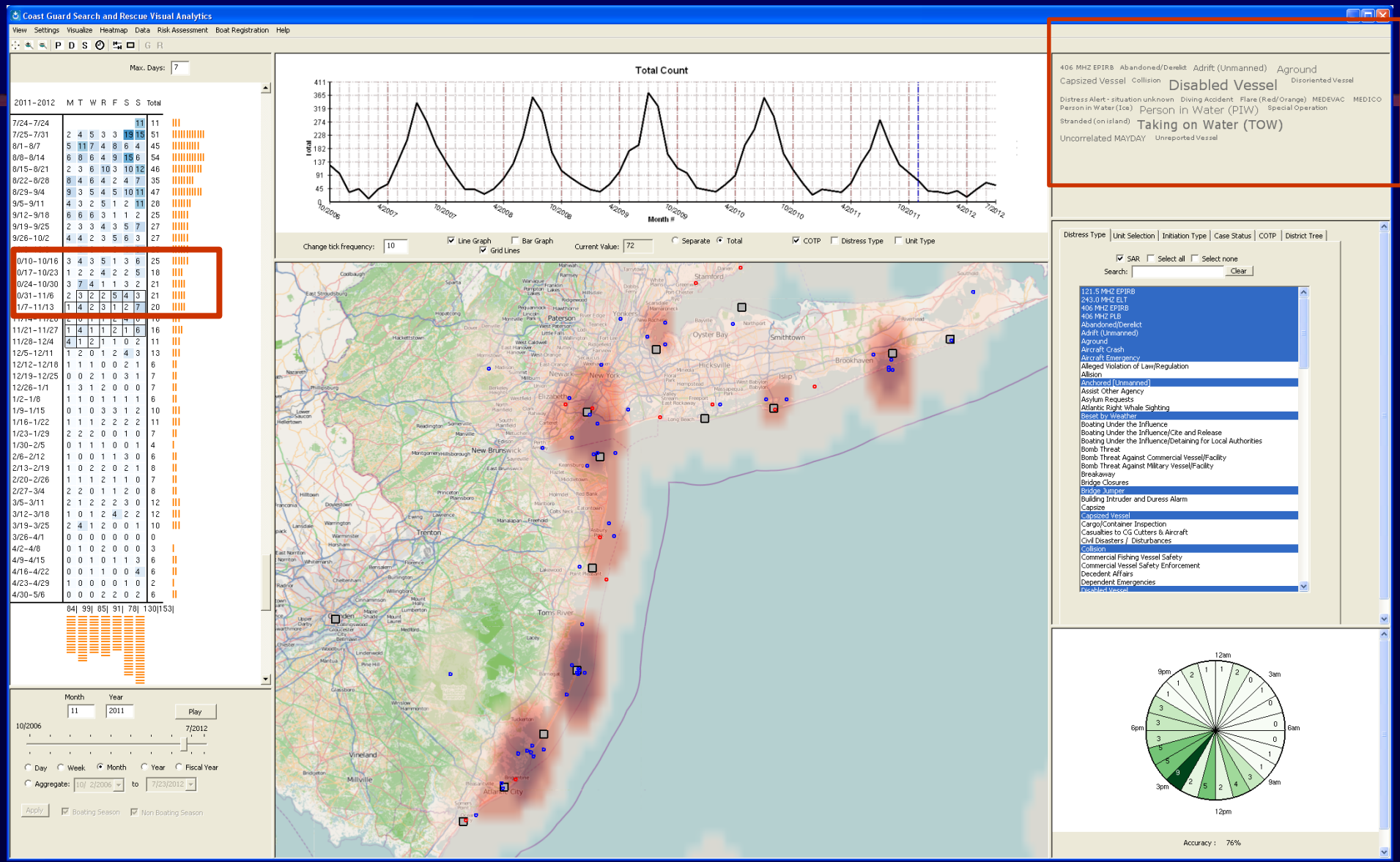
IMPACTS:

- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Hurricanes Sandy and Irene resource allocation decisions based on cgSARVA analysis and visualization
- Informed Commandant's budget testimony to Congress
- Key component of USCG D9 reallocation plan for 2011-12
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ

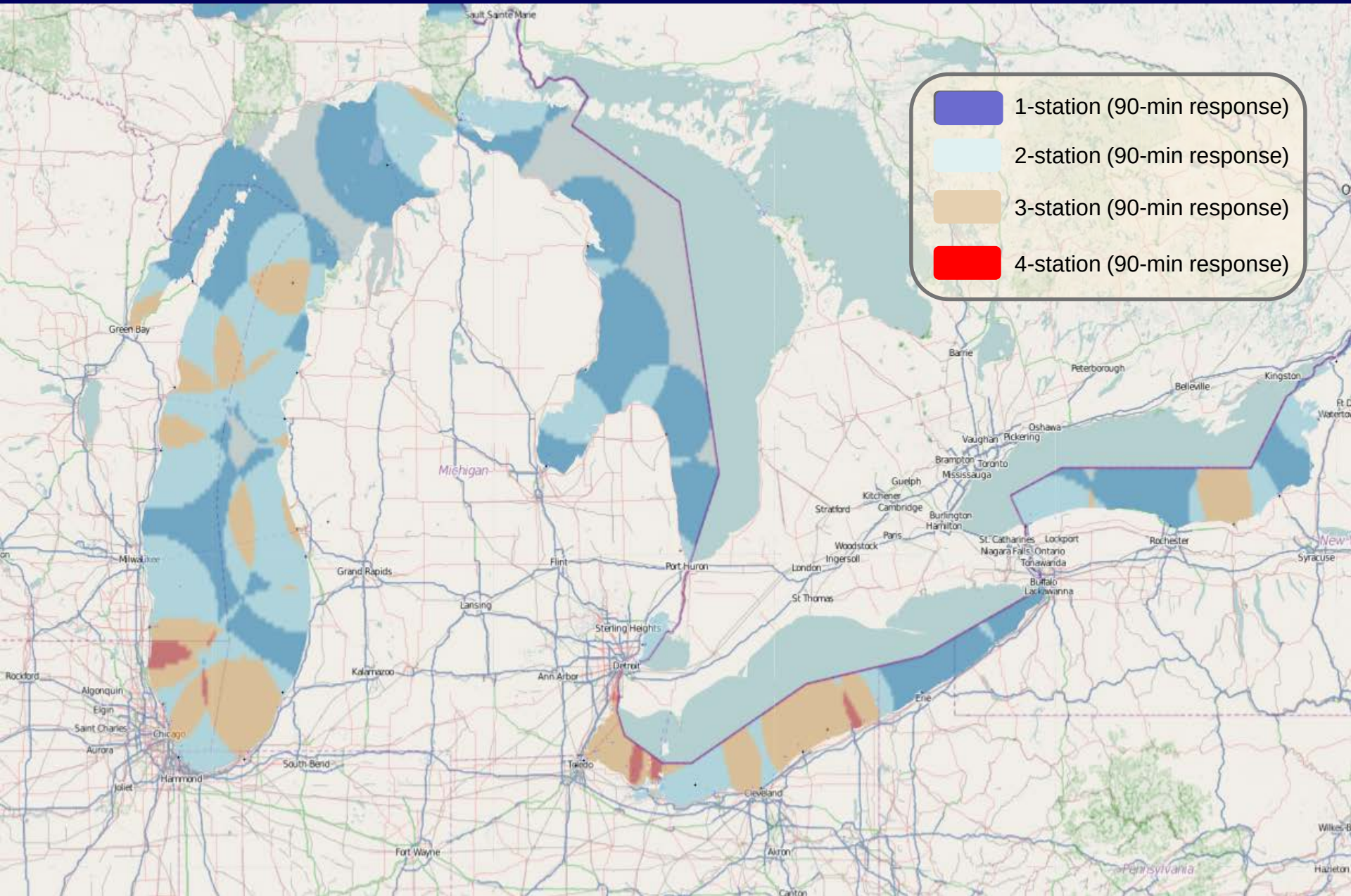


Example: Risks and Consequences From Sandy:

SAR Cases November 2011 NJ/NYC Area



Response Efficiency – Possible Asset Allocation

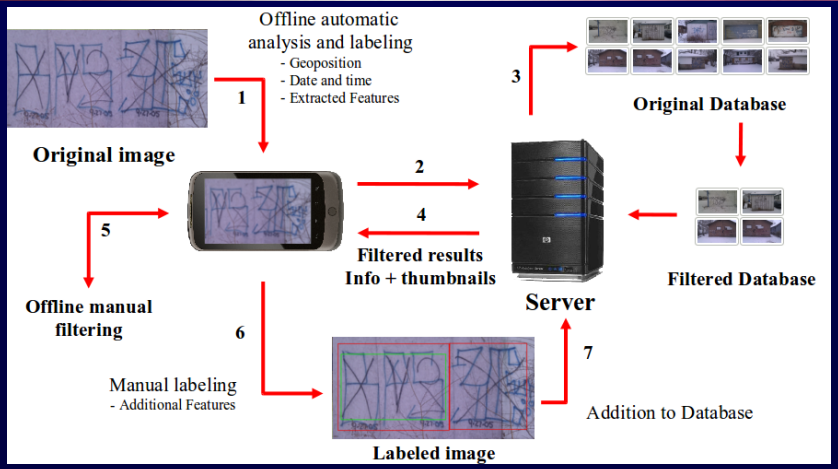


Software Accredited for Decision Making

- April 22, 2013 cgSARVA VV&A'd for US Coast Guard system-wide use



Gang Graffiti Analysis and Recognition Using a Mobile Telephone (GARI)





IMPACT:

- Allows police to catalog and analyze gang graffiti images into a database system to better track and determine gang activity throughout a region
- Will allow the graffiti images to be "interpreted"
- More than 140 users and 1300 graffiti images acquired

GARI delivered:

- Summer/Fall 2011:
  - IPD gang detectives
  - IPD at large
  - Ind Fusion Gang Task Force
  - Gang detectives across Indiana
- Spring/Summer 2013:
  - Cook County, IL
  - INGang ownership
  - Tattoo database added

For Further Information

www.VisualAnalytics-CCI.org

vaccine@purdue.edu
ebertd@purdue.edu



Who We Are:
International Team of Experts
75+ Faculty, 26 Institutions



- Purdue University
- Georgia Institute of Technology
- Pennsylvania State University
- Stanford University
- University of North Carolina at Charlotte
- University of Washington
- Arizona State University
- Simon Fraser University
- University of British Columbia
- Justice Institute of British Columbia
- Ontario Institute of Technology
- Dalhousie University
- University of Victoria
- University of Houston, Downtown
- Virginia Tech
- Indiana University
- Florida International University
- University of Texas at Austin
- Morgan State University
- Navajo Technical College
- University of Stuttgart
- University of Swansea
- Oxford University
- University of Calgary
- University of Manitoba
- Carleton University

VACCINE's Role

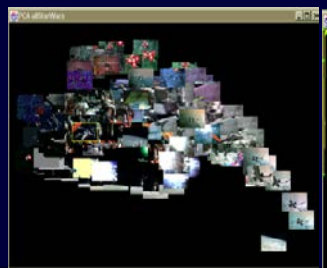


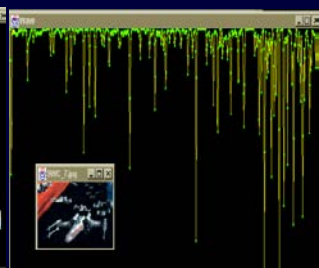
Problem: To solve current and future homeland security problems requires exploring, analyzing, and reasoning with massive, multi-source, multi-scale, heterogeneous, streaming data **–BIG DATA**

- Cuts across entire spectrum of homeland security needs

We provide tools to enable end users to get the relevant information they need during any situation to make a decision or take action

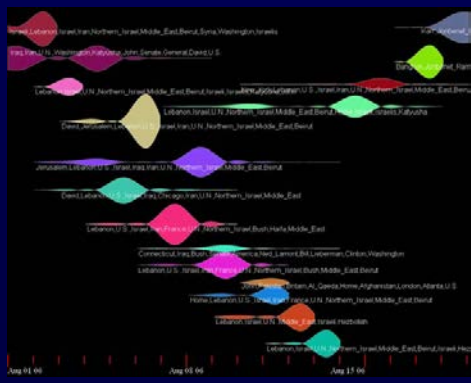






VACCINE Mission





- Provide visual analytic and scalable solutions to 2.3 million extended homeland security personnel
  - 185,000 DHS personnel, 350,000 law enforcement personnel, 750,000 homeland security practitioners
- Achieve excellence in visual analytics and visualization sciences
- Educate homeland security stakeholders and the next generation of talent



VACCINE Value

Our Value / Solution: Enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable users to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

VACCINE Value Part II



Our people and partnerships

- Interdisciplinary world-leading team of researchers
- Defining and extending the new science of visual analytics driven by real-world, real-scale problems of engaged partners (local, state, federal)







Visual Analytic Solutions: What We Offer

- **Improved Effectiveness**: We enable users to be more effective through innovative, interactive visualization, analysis, and decision making tools
  - Provide the right information, in the right format, within the right time to solve the problem
  - Enable user to be more effective from planning to detection to response to recovery
  - Enable effective communication of information
- **Innovative Fielded Solutions:** We provide innovative visual analytic and scalable solutions to the extended homeland security community
- **People and Partnerships**
  - Interdisciplinary world-leading team of researchers and students
  - Actively Engaged Partners

"cgSARVA has proven its worth time and again, providing key analytic information for decision makers for large scale projects…"

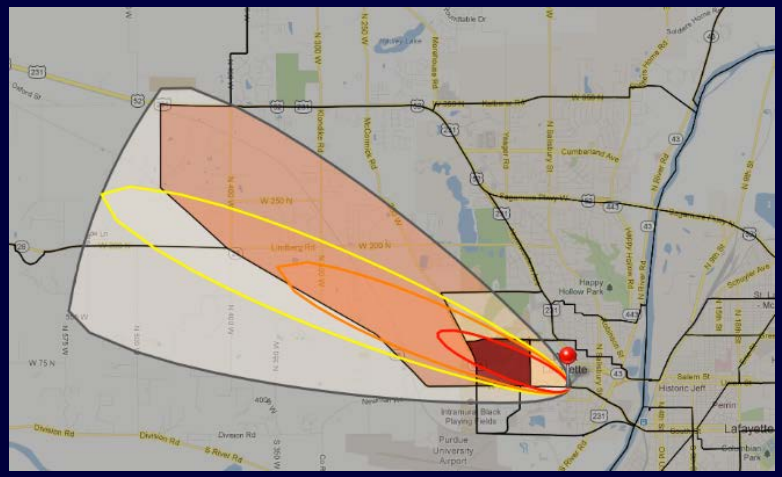
VADM Robert Parker, 2012 MRS Keynote Address

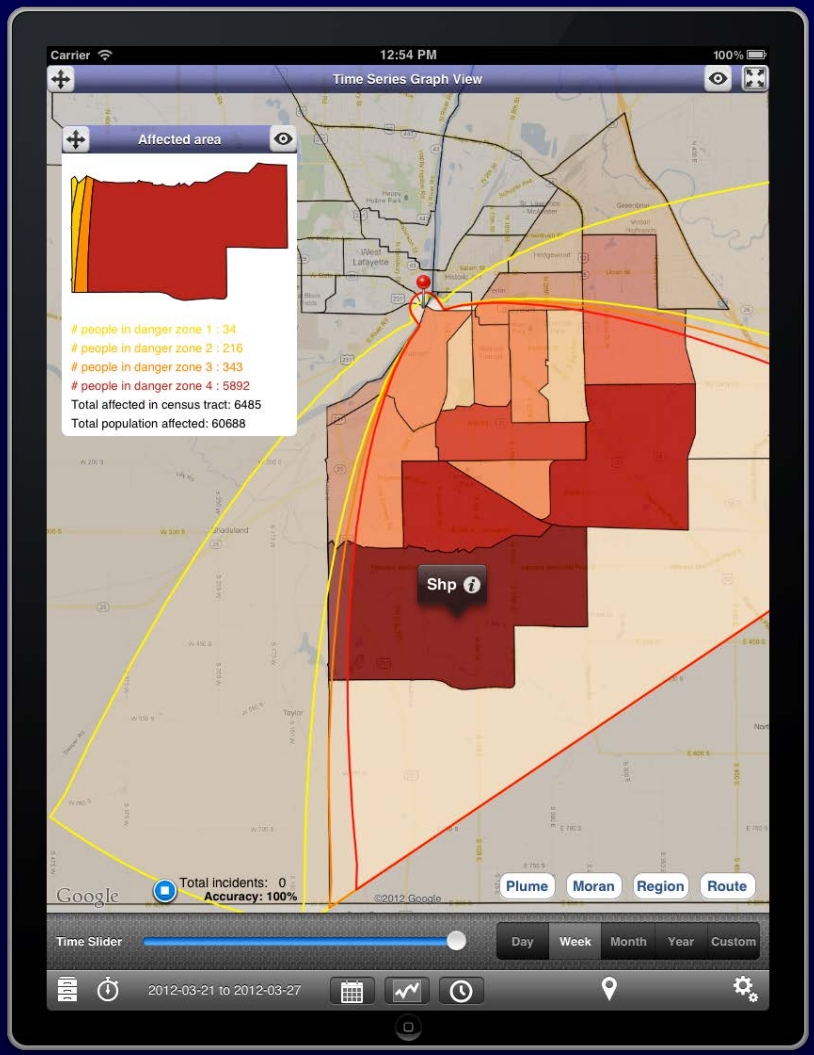


VADM Robert Parker with VACCINE student researchers (cgSARVA, COAST, iOPAR)

MERGE – iVALET Interactive Plume Visualization and Evacuation Planning

- Chemical release plume modeling identifies census tracts with the highest number of expected people affected





VACCINE
Visual Analytics for Command, Control, and Interoperability Environments
A U.S. Department of Homeland Security Center of Excellence

Visual Analytics for Security, Safety, and Privacy:

Approaches, Lessons Learned, Opportunities, and Challenges

David S. Ebert

CVADA
Center for Visual Analytics and Data Analytics
A U.S. Department of Homeland Security Center of Excellence

October 2013

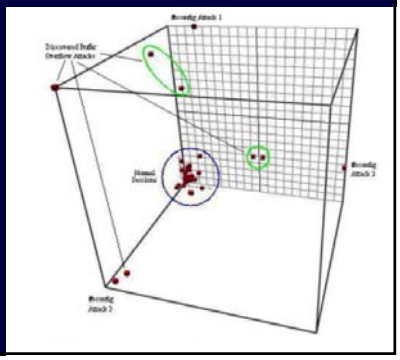
Overview

- Background: Why am I here?
- Challenges in developing effective deployed solutions
- Approaches: which one to choose?
- Some examples and lessons
- Path forward



Why Am I Here?

- My seminal paper from VisSym 2001?
  - Atkinson, T., Pensy, K., Nicholas, C., Ebert, D., Atkinson, A., Morris, C., "Case Study: Visualization and Information Retrieval Techniques for Network Intrusion Detection," *VisSym 2001:* Joint Eurographics - IEEE TCCG Symposium on Visualization, May 2001.
  - Interactive volume visualization of network attacks projected onto know attacks
- Or for my experience leading VACCINE?
  - Different safety and security (in general)
  - Cybersecurity enters many projects



VACCINE

October 2013



Who We Are:
International Team of Experts
75+ Faculty, 26 Institutions



- Purdue University
- Georgia Institute of Technology
- Pennsylvania State University
- Stanford University
- University of North Carolina at Charlotte
- University of Washington
- Arizona State University
- Simon Fraser University
- University of British Columbia
- Justice Institute of British Columbia
- Ontario Institute of Technology
- Dalhousie University
- University of Victoria
- University of Houston, Downtown
- Virginia Tech
- Indiana University
- Florida International University
- University of Texas at Austin
- Morgan State University
- Navajo Technical College
- University of Stuttgart
- University of Swansea
- Oxford University
- University of Calgary
- University of Manitoba
- Carleton University

VACCINE

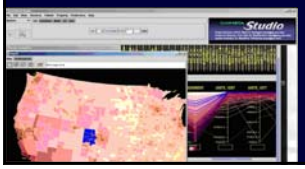
VACCINE's Role

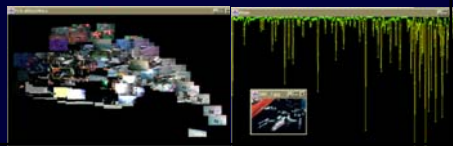


Problem: To solve current and future homeland security problems requires exploring, analyzing, and reasoning with massive, multi-source, multi-scale, heterogeneous, streaming data –**BIG DATA**

- Cuts across entire spectrum of homeland security needs

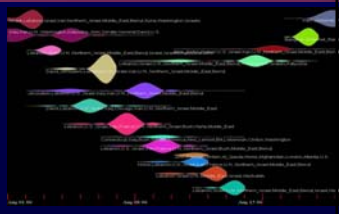
We provide tools to enable end users to get the relevant information they need during any situation to make a decision or take action





VACCINE Mission





- Provide visual analytic and scalable solutions to 2.3 million extended homeland security personnel
  - 185,000 DHS personnel, 350,000 law enforcement personnel, 750,000 homeland security practitioners
- Achieve excellence in visual analytics and visualization sciences
- Educate homeland security stakeholders and the next generation of talent



VACCINE Value

Our Value / Solution: Enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable users to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

VACCINE Value Part II



Our people and partnerships

- Interdisciplinary world-leading team of researchers
- Defining and extending the new science of visual analytics driven by real-world, real-scale problems of engaged partners (local, state, federal)







Visual Analytic Solutions: What We Offer

- **Improved Effectiveness**: We enable users to be more effective through innovative, interactive visualization, analysis, and decision making tools
  - Provide the right information, in the right format, within the right time to solve the problem
  - Enable user to be more effective from planning to detection to response to recovery
  - Enable effective communication of information
- **Innovative Fielded Solutions:** We provide innovative visual analytic and scalable solutions to the extended homeland security community
- **People and Partnerships**

"cgSARVA has proven its worth time and again, providing key analytic information for decision makers for large scale projects…"

VADM Robert Parker, 2012 MRS Keynote Address

VADM Robert Parker with VACCINE student researchers (cgSARVA, COAST, iOPAR)



Engaged End-Users

- Federal Operating Components:
  - US Coast Guard
  - US Transportation Security Agency
  - US Customs and Immigration Service
  - US Federal Emergency Management Agency
  - US Customs and Border Patrol
  - US CERT
  - US ICE (in progress)
- Law Enforcement
  - Over 40 local and state agencies (IN, IL, OH, SC, PA, NC, NY)
- Fusion Centers
  - Ohio (SAIC)
  - Indiana (IIFC)



VACCINE

October 2013

Challenges in Developing Effective Deployed Solutions: Crossing the Chasm



Challenges in Developing Effective Deployed Solutions

1. Understanding the situation
   - Task/problem
     - What are they trying to find, analyze, explore?
     - What is the final product of the system and task?
   - User
     - How do they conceptualize the problem?
     - What are the natural scales/aggregation levels, features?
   - Environment - time frame, solitary vs. collaborative, equipment
   - Language - developing a common language

Challenges in Developing Effective Deployed Solutions

2. Changing requirements
   - How to be effective in an agile software development environment
   - Avoiding feature creep
   - Clear end state, goals, and deliverables

Challenges in Developing Effective Deployed Solutions

3. Trust, polices, lack of standards
   - Trust
     - Will you deliver and follow-through?
     - Or, do you just want my data?
     - What can an academic really know about what I do?
   - Polices
     - Legal agreements and delays
     - Data and privacy
   - Standards - everyone has a different schema, RMS, etc.

One Potentially Useful Approach: Application-Driven Research & Development

- A full contact sport
- Increases rate of VA advances and application deployment to effectiveness
- Increases rate of application domain advances





Our Application-Driven Research Approach and Plan

- Evolving, effective, and enduring research by tight integration with stakeholders
- Driven by stakeholders – from initiation, through iterative development (agile software development), to deployment
- Visual Analytics research integrated:
  - Interactive visual/cognitive analytic environments based on novel research in visual analytics, algorithms, information transformation, cognitive and interaction science, creating precise information environments

Full-scale exercise February 2008



Research Motivation:

- Solving these real-world problems requires
  - Novel theories, techniques, approaches, and adaptations of algorithms
  - Integration of cross-disciplinary expertise
  - Overcoming the chasm from academic idea to deployed solution
- Solving these real-world problems provides
  - Compelling, publicly understandable value for your research
  - Advances in CS and in other disciplines
  - New publication opportunities
  - Great collaboration partners and proponents
  - Opportunities for new adventures



Examples of Overcoming the Chasm

- Public health syndromic surveillance
- Crime analytics
- US Coast Guard solutions

Solutions for Spatial Temporal Decision Making Environments – A Progression: The Long and Winding Road[1]



- Public health surveillance
  - Fusing apparently similar data that isn't (health data)
  - Dual domain decision making and real-world visualization and analysis for disease spread and interdiction
- Spatial and temporal visual analytics for law enforcement
- Search and rescue (SAR) and risk based visual analysis

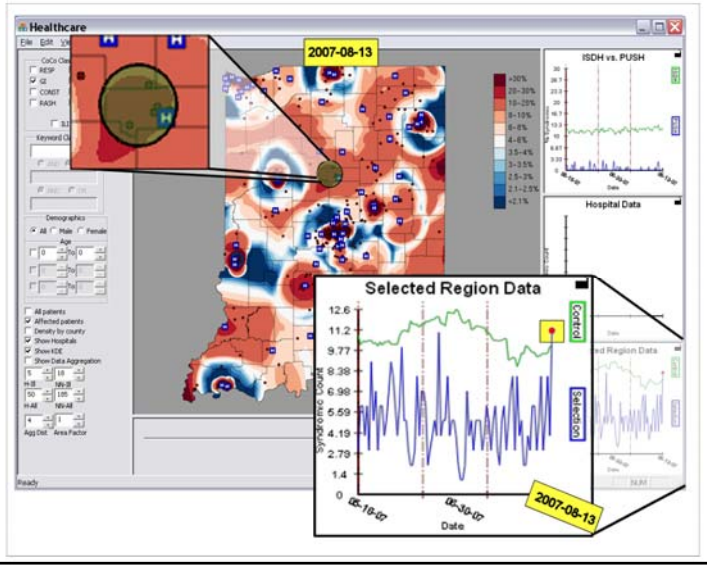
"The long and winding road
That leads to your door
Will never disappear" – P. McCartney

Improving Syndromic Surveillance

Interactive visual analytic environment for effective syndromic surveillance and response:

- System designed based on collaboration and feedback with state epidemiologists
- Integrated temporal, geospatial, multi-source, multi-scale analytic capability
- Density estimation for data exploration
- Syndromic control charts for temporal alerts
- Demographic filter controls for advanced analysis

Visual Analytics for Syndromic Surveillance: Hypothesis Generation and Exploration

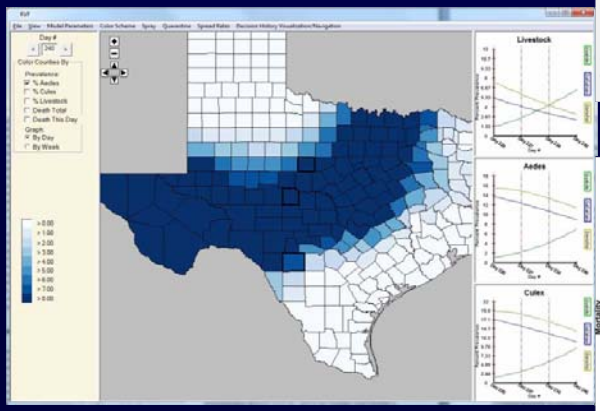


Project Design & Workflow Impetus: Indiana State Epidemiologist, EHR researcher

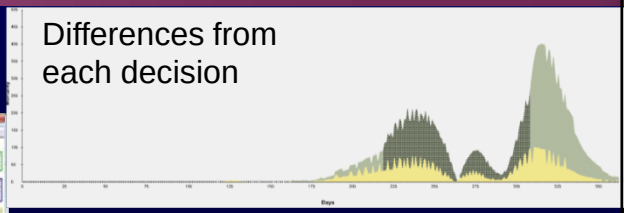
Best Paper Nominee, *IEEE Symposium on Visual Analytics Science and Technology (VAST),* October 2008, for "Understanding Syndromic Hotspots – A Visual Analytics Approach," (Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W., Grannis, S., Wade, M., Ebert, D.).

Example Decision Analysis Linked Displays – Example with 3 Decisions

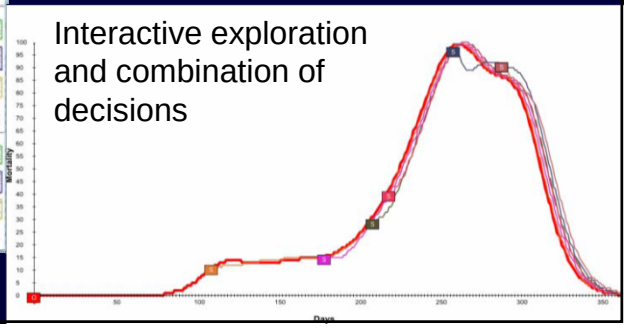
Partnership with FAZD



Differences from each decision



Interactive exploration and combination of decisions

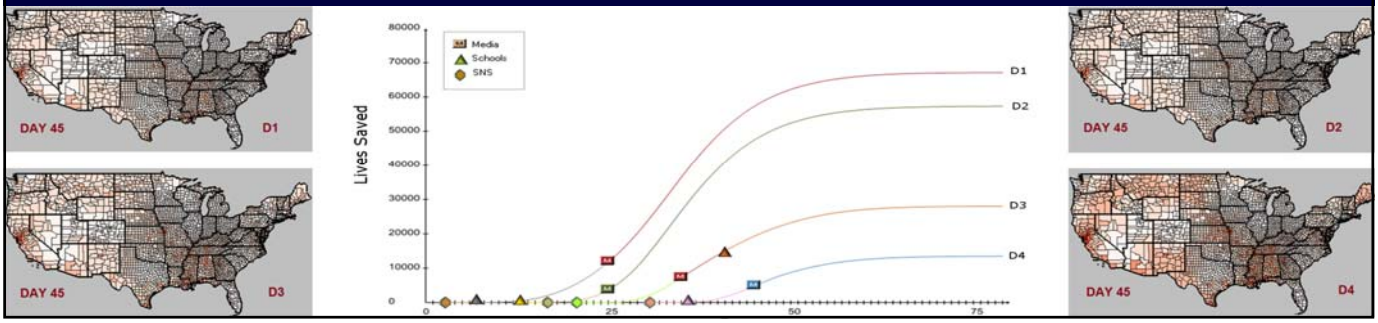


VACCINE

October 2013

Integrated Interactive Simulations and Analysis

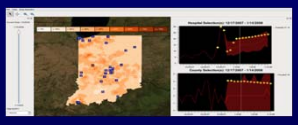
- Analysis and simulation must be interactive for integration into interactive environment
- Need novel computational & statistical models
- Goal: enable improved discovery, decision making,

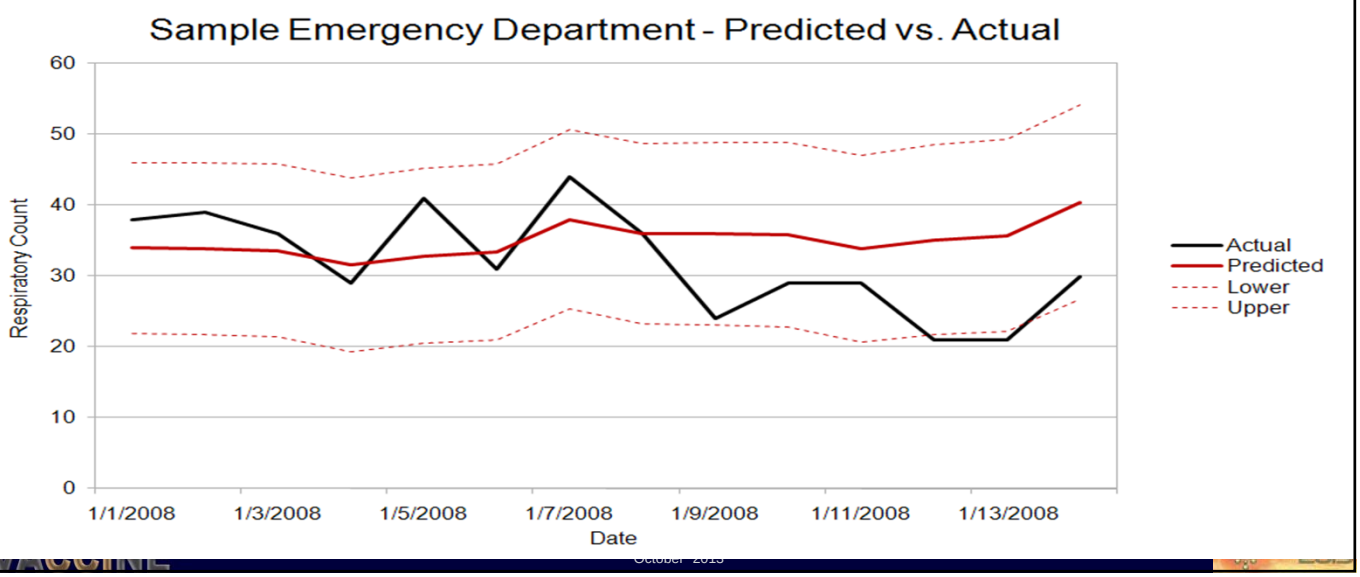


Situational Surveillance and Predictive Visual Analytics

- Focus is on categorical spatiotemporal event data
- Utilizing time series and density estimations we want to create an interactive environment for predicting future event magnitudes and locations
- We utilize seasonal trend decomposition with Loess smoothing
- 3D Kernel density estimation for spatiotemporal probability distributions

Predictive Visual Analytics





Crime VA – The Next Curve

- Sheriff wanted to know how to use integrated data across the county to see
  - If they are being more effective
  - If crime is being reduced
  - How officers to top-level officials can use this data for proactive and predictive policing
- Frequent meetings and continuous refinement of tools
- Now being tested by agencies in 4 states
  - NYPD, OSHP, Il.SP, LPD, WLPD, PUPD, TCSD

Visual Analytics Law Enforcement Toolkit (i)VALET





VALET



Visual Analytics Law Enforcement Toolkit
(VALET, iVALET)

Impacts:

- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

iVALET
Debriefing Example

VALET delivered:
- Spring 2011: WL, Lafayette Police

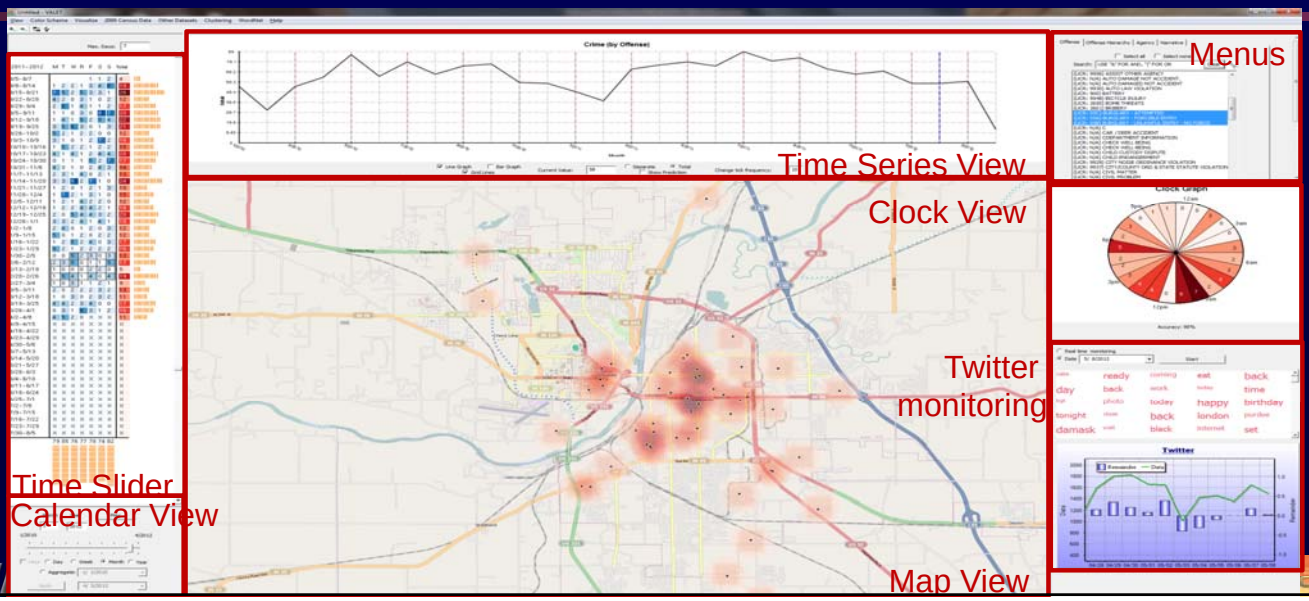
iVALET delivered:
- October 2011: Purdue, WL Police



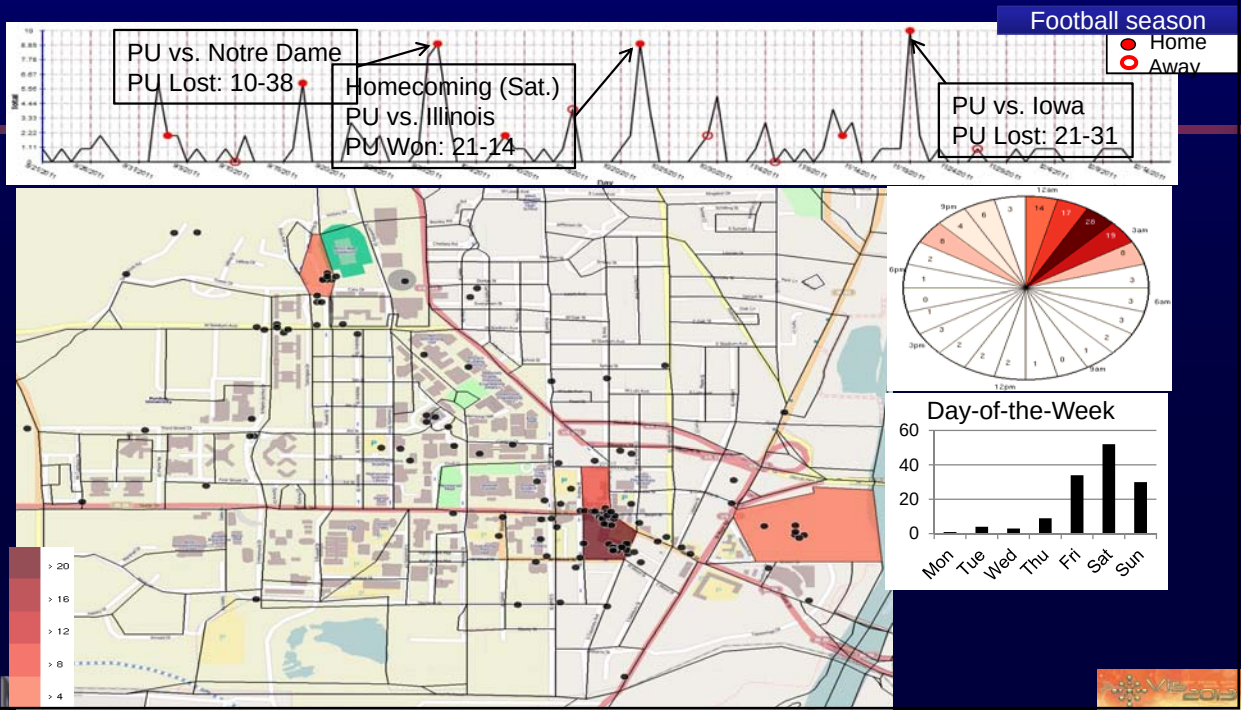
VACCINE

October 2013

VALET Overview

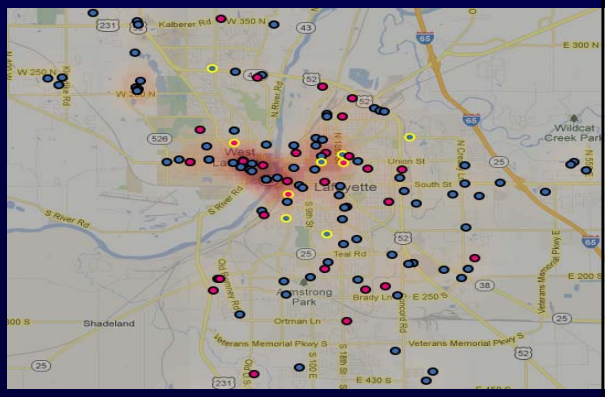


Example: Drunkenness / Public Intoxication



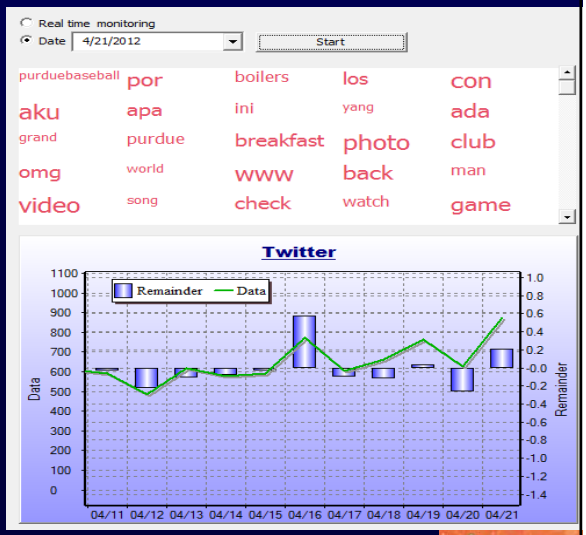
Top 10 Hot Incidents

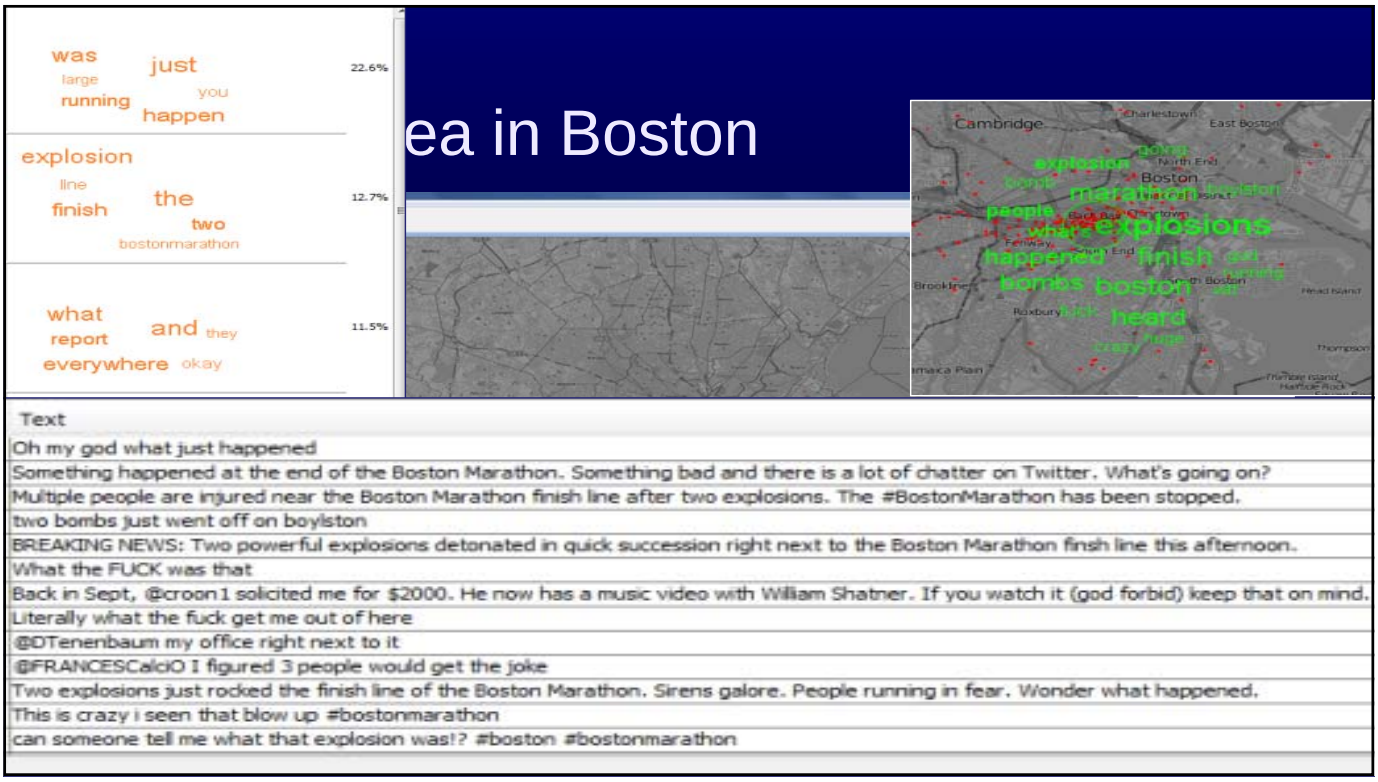
- Identify unusual localized high-frequency patterns of crimes (near repeats)
- Each data entry is checked for other crimes with similar properties within a 1 block radius of the incident location and a 14-day time period
- Top 10 incidents with the most number of related incidents in this space-time window are highlighted



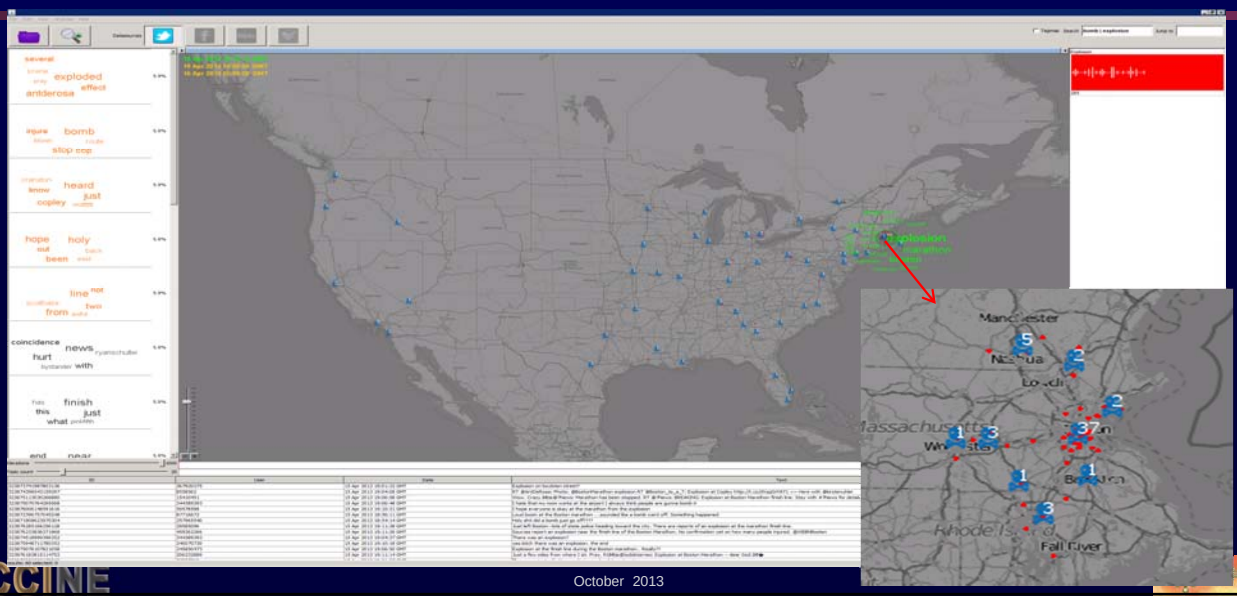
Social Media: Real-time Twitter Monitoring and Integration into Tools (Purdue, Stuttgart, Penn St.)

- Topic extraction using novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud





Detection using the Explosion Classifier



First Response (Tweet & Picture)



Visual Analytics of Activity During Hurricane Sandy

Two weeks before Sandy
10/14 (Sunday), 12:00 ~ 16:00

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00





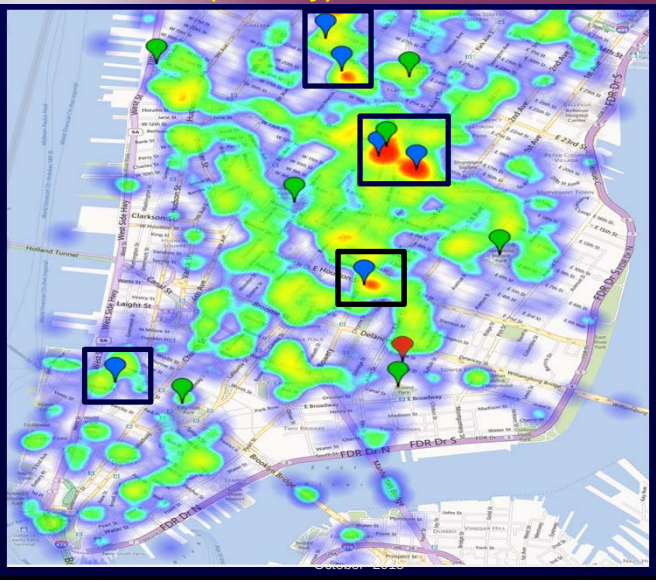
Supermarket Park Shelter

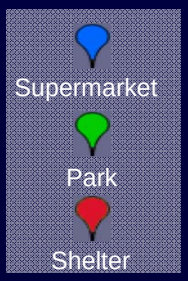
Visual Analytics of Activity During Hurricane Sandy

After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00

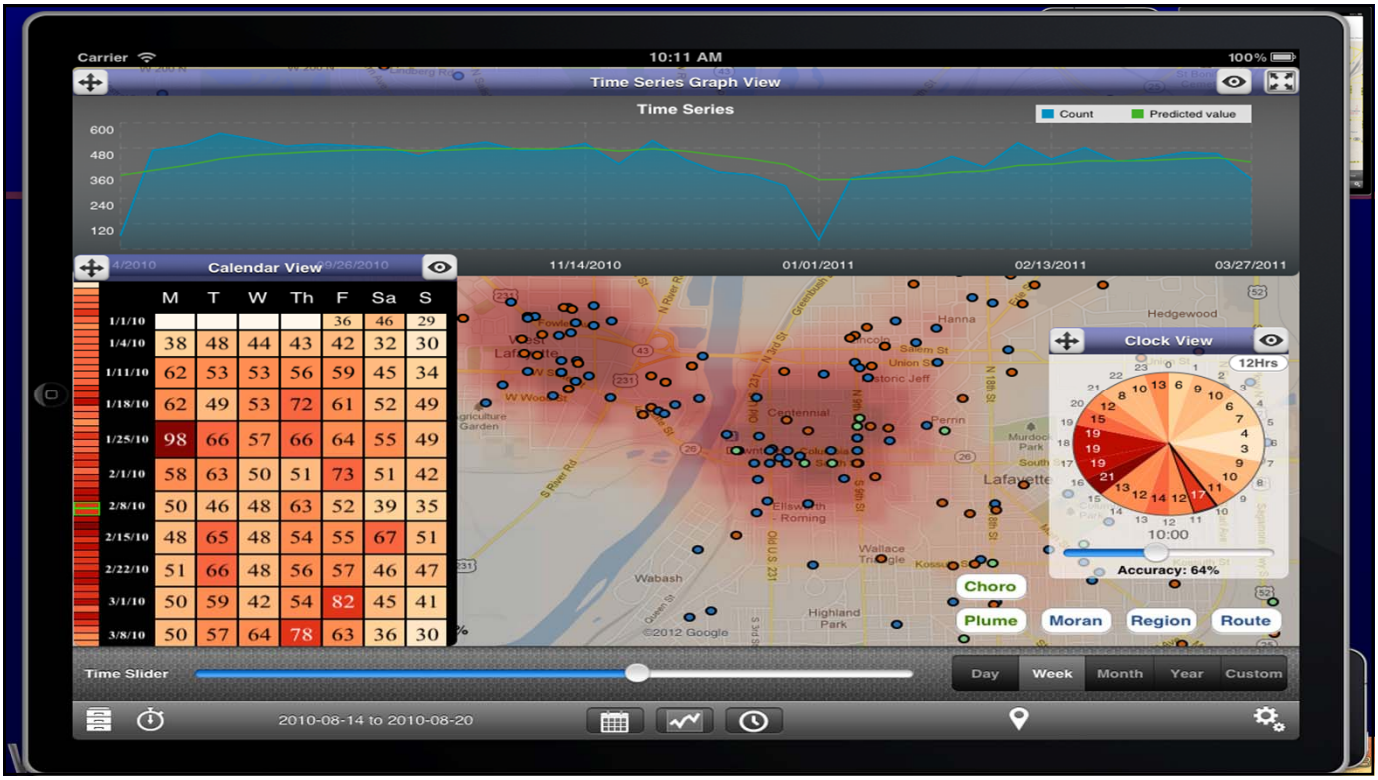
Evacuation order:
10/28, 10:30 AM

Hurricane Sandy's
Arrival at NYC:
10/29, 8:00 PM





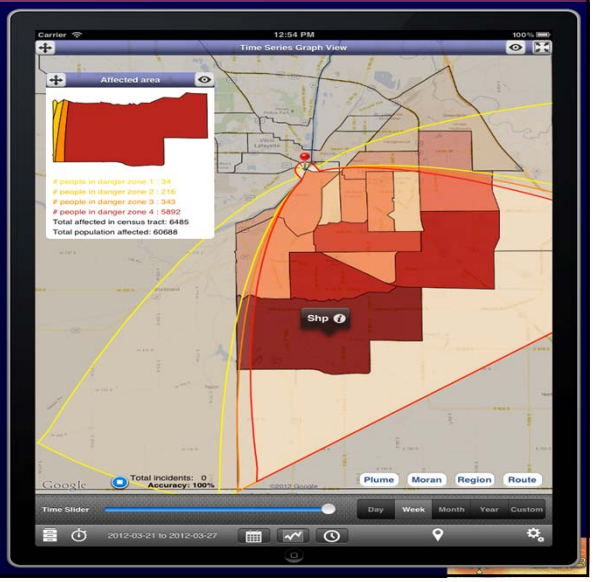
VACCINE



MERGE – iVALET Interactive Plume Visualization and Evacuation Planning

- Chemical release plume modeling identifies census tracts with the highest number of expected people affected





VACCINE

October 2013

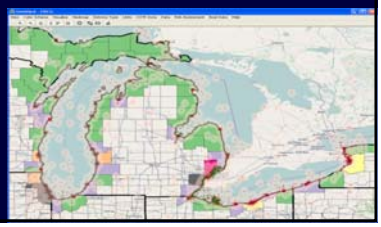
The Next Bend: US Coast Guard



Visual Analytics Uses for Risk-Based Decision Making



- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Interactive, multi-day, month, type investigation
- Multisource, multimedia data integration & analysis

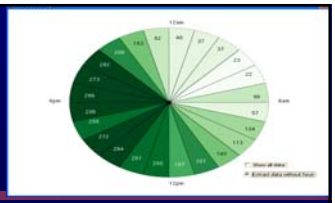


USCG: Effective Risk-based Decision Making and Resource Allocation Visual Analytics

- Evaluate current and historical mission area:
  - Demands
  - Risks (total, mitigated, residual)
  - Resource allocation
  - Return on investment
- Evaluate courses of action
- Evaluate above at both Strategic and Tactical/Operational level

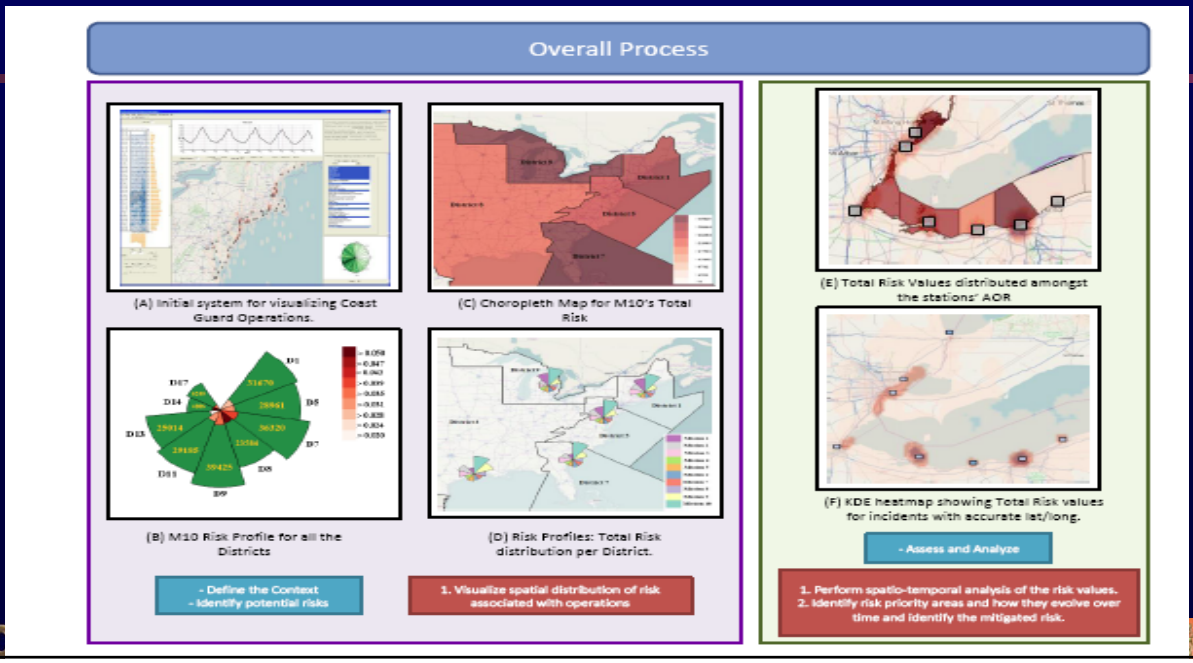


Risk-Based Allocations



- Comparative visual analysis of mission cases/hours vs. staffing hours
- Comparative visualization of resources vs. risk
- Trend visual analytics
  - Increase/decrease in resource allocation
  - Increase/decrease in risk (total, mitigated, residual)
  - Increase/decrease in incidents
- Exploration of alternatives and effect on risk
- Predictive analytics based on historical data (STL and EWMA)

VA For Risk-Based Decision Making Process

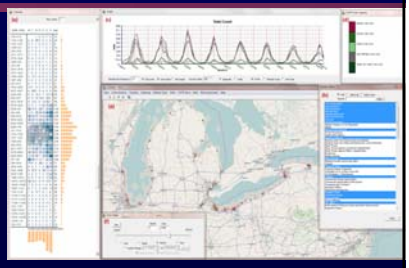


U.S. Coast Guard Search and Rescue VA (cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771

IMPACTS:

- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Hurricanes Sandy and Irene resource allocation decisions based on cgSARVA analysis and visualization
- Informed Commandant's budget testimony to Congress
- Key component of USCG D9 reallocation plan for 2011-12
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ



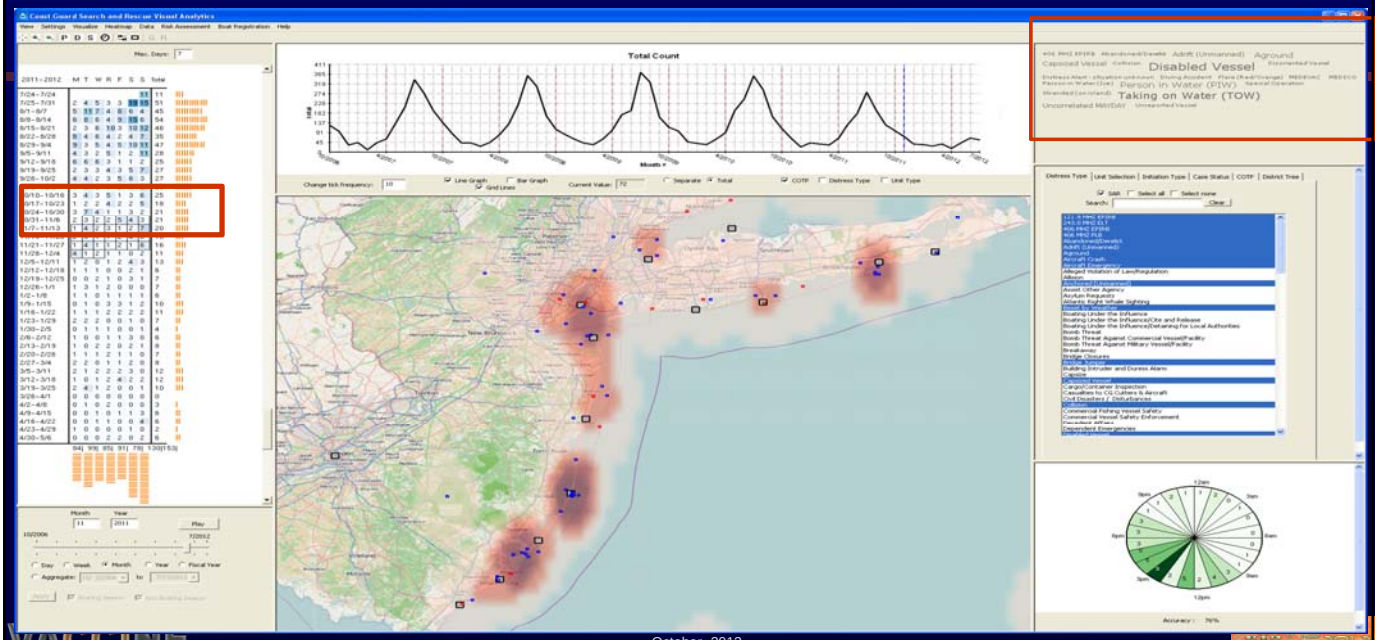


VACCINE

October 2013

Example: Risks and Consequences From Sandy:

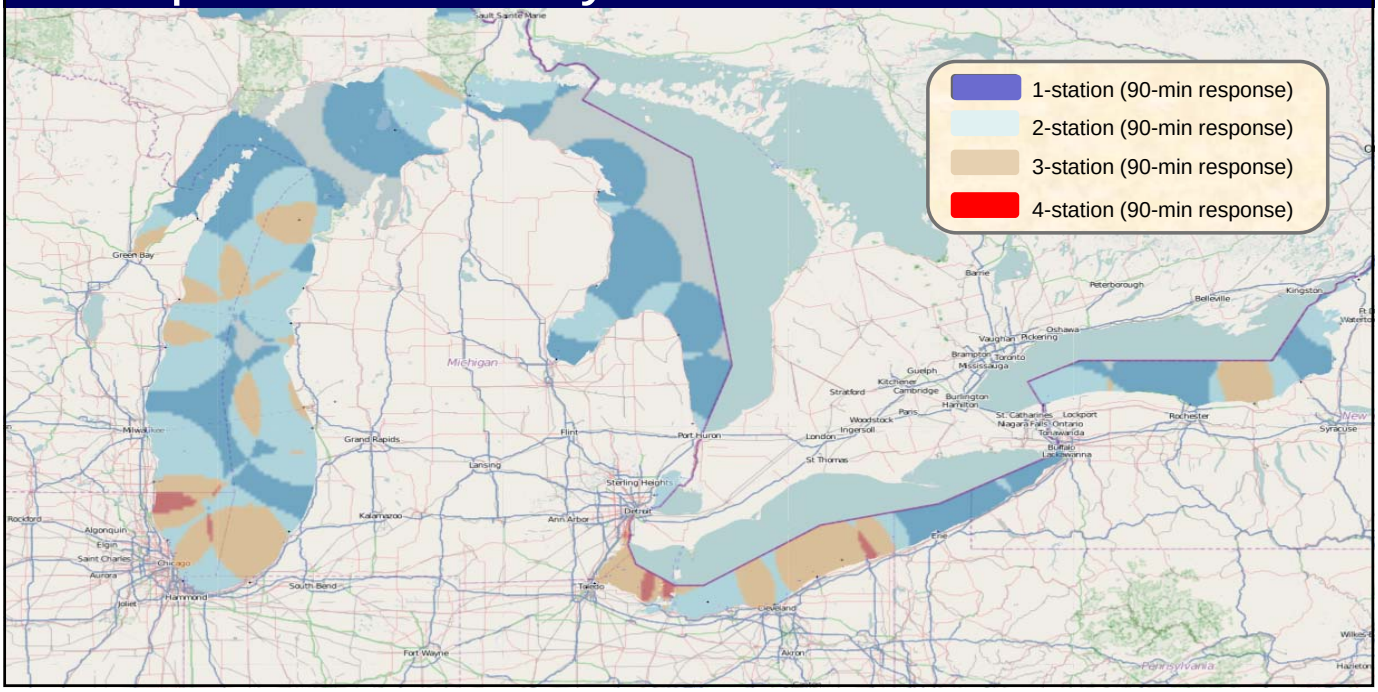
SAR Cases November 2011 NJ/NYC Area



VACCINE

October 2013

Response Efficiency – Possible Asset Allocation



Software Accredited for Decision Making

- April 22, 2013 cgSARVA VV&A'd for US Coast Guard system-wide use



VACCINE

Octo



Chasm Update – Crossed And Survived

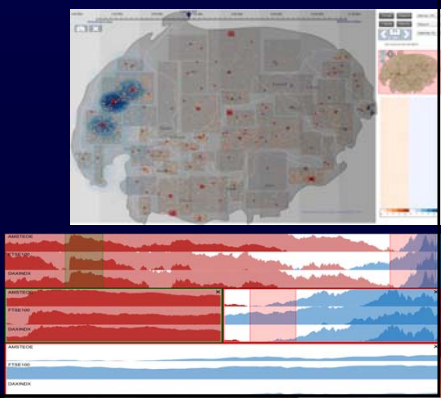
-

Lessons Learned

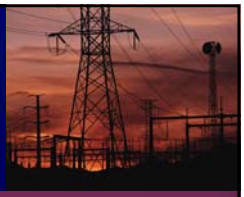
- Extremely worthwhile
- Communication and interaction are key
- Continually ask questions
- Many surprises around each turn (e.g., we need to VV&A the software)
- A growth and learning experience for everyone – a lot of acquired wisdom

Example VACCINE Team Work in Cybersecurity

- Visual Analytics for Security Application (VASA)
- Corporate Insider Threat Detection (Oxford, Leicester, Cardiff)
- Sensor Forensics (Purdue)
- SemanticPrism (Purdue)
- Multiscreen, Multiview, Interactive Cyber Investigation (VaTech, PNNL)
- Log Visualization (Purdue)



Cascading Critical Infrastructure Resiliency Modeling and Analytics (VASA)



- Purpose: Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.
- Collaborating Institution(s): Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities
- End-User(s): Power Suppliers (e.g., Duke Energy), Cyber Community (e.g., Cisco), Quick Service Restaurants and suppliers, food supply

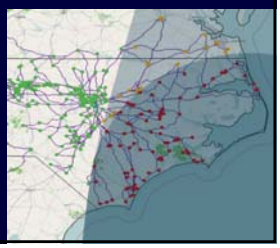
VASA: Visual Analytics for Security Applications

Collaborating Institution(s): Purdue, Minnesota, UTexas, UNCC + German universities
End-User(s): Fast-food restaurant chain, emergency management and planning personnel

Impacts and Accomplishments:



- Support decision-making for extreme weather and disaster (natural, man-made) scenarios
  - Combine real and simulation data
  - Allow "what-if" exploration
- **System of systems**: binds together multiple simulations models from collaborators into coherent whole
  - **Minnesota**: food distribution model
  - **Texas**: simulated and historical weather (hurricanes, storms)
  - **UNCC**: critical infrastructure
  - **Purdue**: roads + interaction visual analytics tool
- **Challenge**: combine interactive VA with complex simulation models for effective decision making



Corporate Insider Threat Detection: Cyber Security Inside and Out

(Universities of Oxford, Leicester, and Cardiff)

- **Sponsor:** Centre for the Protection of National Infrastructure
- **Academics:** Sadie Creese (PI), Min Chen, Michael Goldsmith, Michael Levi, David Upton and Monica Whitty
- **Combined Expertise** in cyber security, psychology, criminology, visual analytics, enterprise operations management and executive education
- **Objectives:**
  - Develop a model,
  - Understand psychological indicators
  - Identify the most effective algorithms
  - Understand enterprise culture and common practices
  - Provide a **visual analytical** interface
  - Develop an understanding of both the various organisational roles and awareness raising and educational methods
- **URL:** http://www.cs.ox.ac.uk/projects/CITD/index.html
- **Oxford Cybersecurity Centre:** http://www.cybersecurity.ox.ac.uk/index.html

Sensor Forensics
(Purdue – Delp)

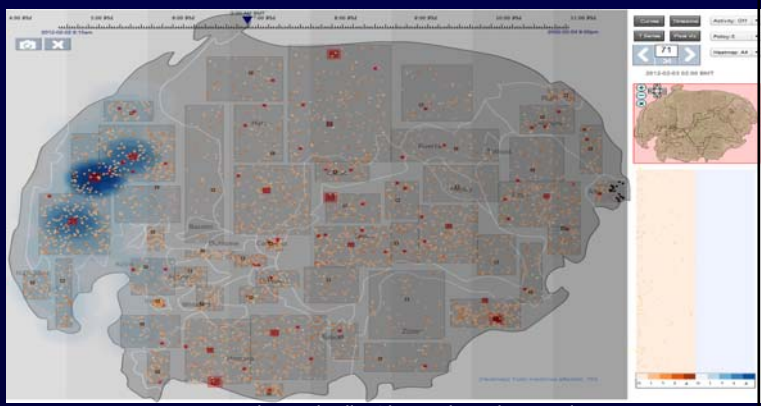
- Forensic characterization
  - Observe device output ➔ which device produced it?
  - Exploit how the device "makes" its output
- Device authentication
  - Performed using forensic characterization
  - Identify device type, make, model, configuration
  - Can the sensor be trusted?
- Detection of data forgery or alteration
- Fingerprint and trace

SemanticPrism: A Multi-aspect View of Large High-dimensional Data (Purdue University)

- **VAST 2012 Mini Challenge 1 Award: Outstanding Integrated Analysis and Visualization**

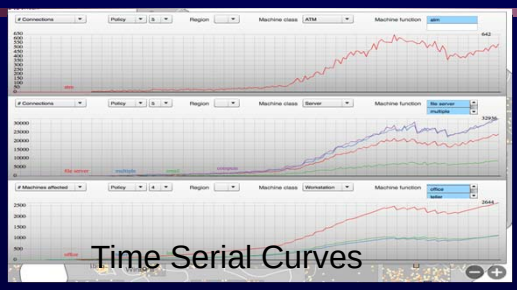
- **Geo-Temporal**
- **Time-serial**
- **Pixel-based**

- **Semantic Zoom**

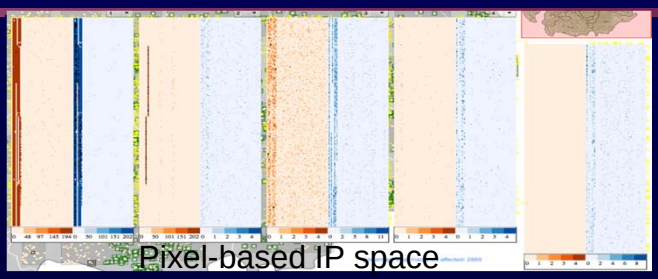


Victor Yingjie Chen, Ahmad M Razip, Sungahn Ko, Cheryl Zhenyu Qian, David S.Ebert

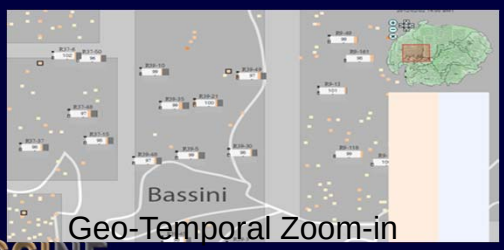
SemanticPrism



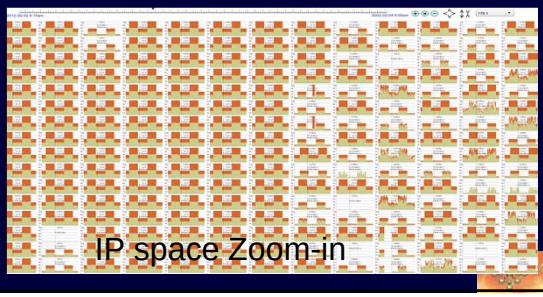
Time Serial Curves



Pixel-based IP space



Geo-Temporal Zoom-in



IP space Zoom-in

VACCINE

October 2013

VA for Cybersecurity Analysts

(VaTech – North, Endert)

Large High-Res Workspaces for Analysts



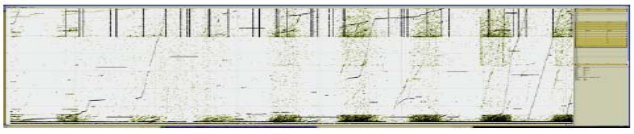


History and Traceability

Multiple, Simultaneous Investigation cases

Large, High-Resolution Visualizations

De-Aggregate Vital Information

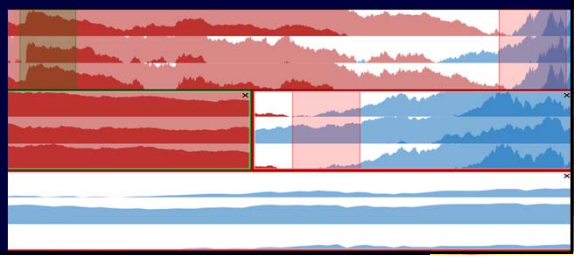




Server Cluster Log File Visualization

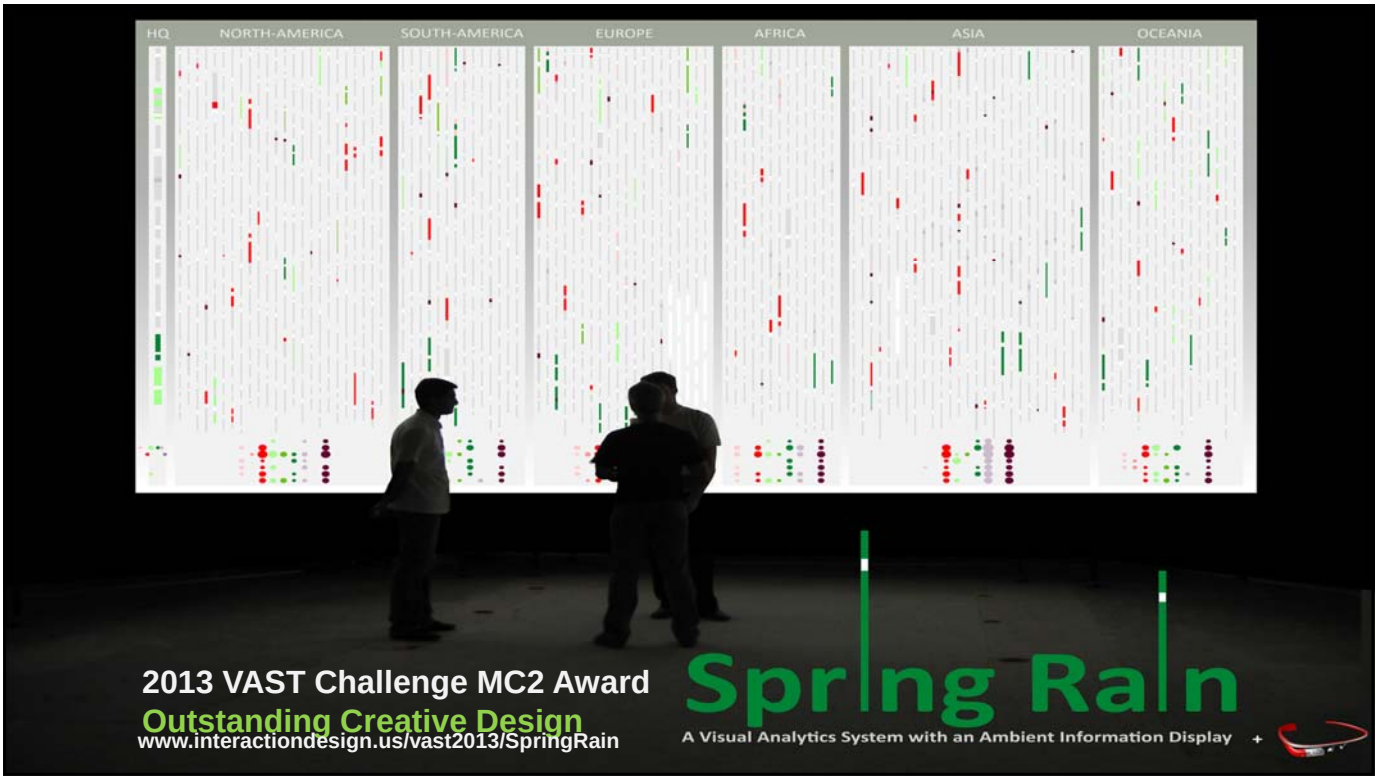
(Elmqvist, Purdue)

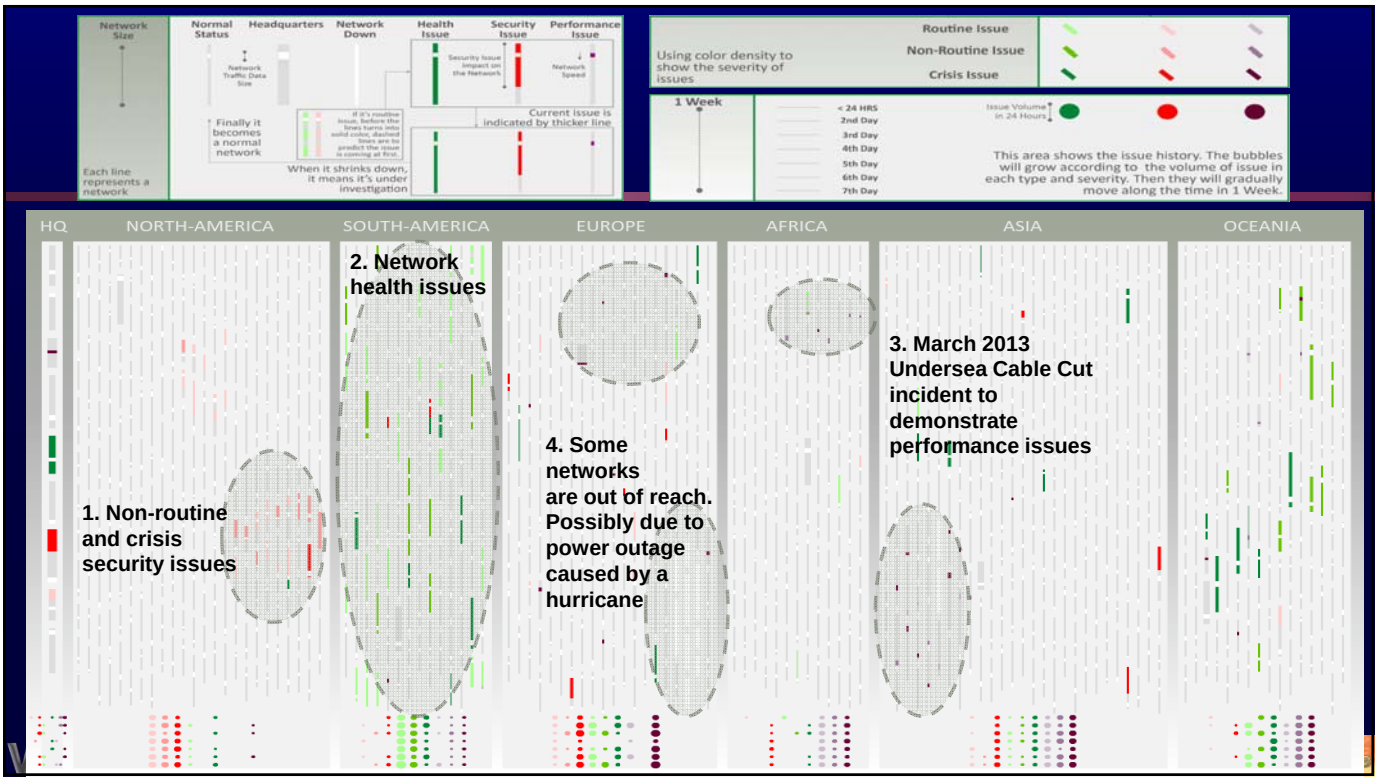
- Log file visualization for Purdue's ECN group
  - 200+ servers, 30 TB of storage, 6 million hits per month
  - User-centered, interview-based design
- Applied **stack zooming** to quantitative log data
  - CPU load, network hits, storage usage,
  - Users navigate in data
  - Very long time periods
- Limited deployment in Fall 2010
  - Very positive, powerful



VACCINE

October 2013





Application-Driven Visualiza Cybersecurity

- Should we be driving the research different users' goals are?
- Interesting survey article
  - Taxonomy of use-case classes:
    - Host-server monitoring
    - Internal/external monitoring
    - Port activity
    - Attack patterns
    - Routing behavior

Shiravi, et al., " A survey of visualization systems for networ security, *IEEE TVCG* 2012.

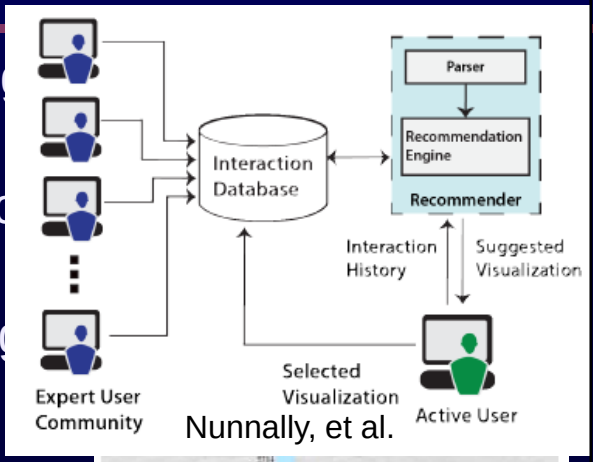
Visualization System	Visualization Technique(s)	Data Source(s)	Number of Citations
Host / Server Monitoring			
Erbacher et al. [4][5]	Glyph	Server Logs	106 | 7
Tudumi [6]	3D Node Link	Server Logs	38
NVisionIP [7,8]	Scatter Plot	NetFlows	145 | 20
Portall [9]	Node Link	Packet Traces	21
HoNe [10]	Node Link	Packet Traces	8
Perlman et al. [11]	Node Link | Glyph	Packet Traces	5
Radial Traffic [12]	Radial Panel	Packet Traces	23
Mansmann et al. [13]	Node Link	Packet Traces	2

Do citation counts show real-world value?

Cube of Doom [19]	3D Scatter Plot	Packet Traces	99
PortVis [20]	Scatter Plot	NetFlows	112
NetBytes Viewer [21]	3D Scatter Plot	NetFlows	7
Existence Plots [22]	Scatter Plot	Packet Traces	3
Attack Patterns			
Giardin [29]	Color Map	Packet Traces	60
NIVA [30]	Node Link | Glyph	Intrusion Alerts	51
Snort View [31]	Scatter Plot | Glyph	Intrusion Alerts	67
IDGraphs [32]	Scatter Plot	NetFlows	29
IP Matrix [33]	Scatter Plot | Color	Intrusion Alerts	21
Visual Firewall [34]	Scatter Plot	Packet Traces	24
IDS Rainstorm [35]	Scatter Plot	Intrusion Alerts	60
Vizalert [36][37][38]	Radial Panel	Intrusion Alerts	38 | 35 | 29
Rumint [39][40]	Parallel Coordinates	Packet Traces	15 | 35
Ren et al. [41]	Flying Term	DNS Traces	10
Xiao et al. [42]	Scatter Plot	Packet Traces	23
Svision [43]	3D Scatter Plot	Packet Traces	9
Mansmann et al. [44]	Treemap	Packet Traces	20
SpiralView [45]	Radial Panel	Intrusion Alerts	5
NFlowVis [46]	Treemap	NetFlows	17
Avisa [49]	Radial Panel	Intrusion Alerts	2
Routing Behavior			
BGPlay [50]	Node Link	BGP Traces	22
Wong et al. [51]	Node Link	BGP Traces	9
LinkRank [52]	Node Link	BGP Traces	16
Teoh et al. [53][54][55]	Histogram | Node Link	BGP Traces	54 | 28 | 35
BGP Eye [56]	Color Map	BGP Traces	8

VizSec Papers - My Analysis

- Expert users engaged: 3 of 9 (e.g. network analysts)
- Evaluation – performance most c informal user studies
- Data – 4 with actual data, 1 using
- Users involved from the start – 1 paper
- Training of Novices based on experts



Nunnally, et al.



Hao, et al.

Directions Forward, Keys to Success & Challenges



Cybersecurity Education

- How do we train practitioners in the field?
  - Varied backgrounds, varied tasks, communication to public
  - Short time to learn
  - Visualization is key
- Approaches:
  - How people learn framework
  - Personalizing learning - Community of practice train
- How do we educate the public?
  - Again, visualization is key



Some Challenges for Cybersecurity

1. Understanding the task and workflow, access to expert users, actual problems, environments
2. Creating decision making environments for analysts with realtime data and decision making constraints at real-world scale (computer-human visual cognition environments)
3. Solving specific scale issues (scalability) and cross-scale issues (machine, intranet, internet)
4. Managing uncertainty and time
5. Enabling risk-based decision making environments

Keys for Success

- User and problem driven
- Balance human cognition and automated analysis and modeling
- Interactivity and easy interaction
  - Intuitive and scalable solutions vital
- Understandability
- Intuitive visual cognition
- Not overloaded with features

For Further Information

www.VisualAnalytics-CCI.org



vaccine@purdue.edu
ebertd@purdue.edu

CVADA

Center for Visual Analytics and Data Analytics
A U.S. Department of Homeland Security Center of Excellence

Visual Analytics For Shared, Synchronized Situational Awareness

David Ebert
Purdue University

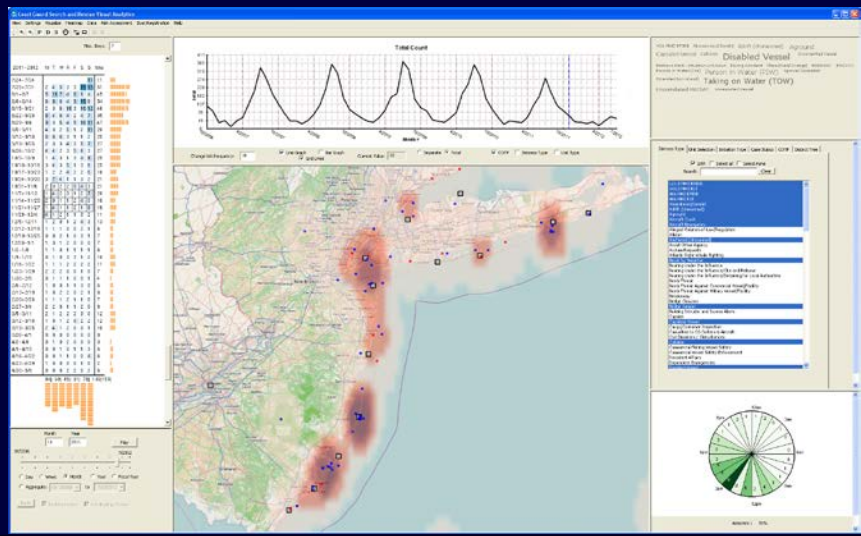
VACCINE

Visual Analytics for Command, Control, and Interoperability Environments
A U.S. Department of Homeland Security Center of Excellence

February 2014

Overview

- What we do
- S^3A and Team Situational Awareness (TSA)
- VACCINE work
  - Real-time Social media Analysis
  - Correlation with multiple social media
  - Predictive visual analytics
  - Mobile visual analytics



VACCINE: What We Do – Visual Analytic Solutions

We enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable user to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

Visual Analytic Solutions: What We Offer

- **Improved Effectiveness**: We enable users to be more effective through innovative, interactive visualization, analysis, and decision making tools
  - Provide the right information, in the right format, within the right time to solve the problem
  - Enable user to be more effective from planning to detection to response to recovery
  - Enable effective communication of information
- **Innovative Fielded Solutions:** We provide innovative visual analytic and scalable solutions to the extended homeland security community
- **People and Partnerships**
  - Interdisciplinary world-leading team of researchers and students

"cgSARVA has proven its worth time and again, providing key analytic information for decision makers for large scale projects…"

VADM Robert Parker, 2012 MRS Keynote Address

VADM Robert Parker with VACCINE student researchers (cgSARVA, COAST, iOPAR)



Why VA or What Big Data Analytics Can't Do

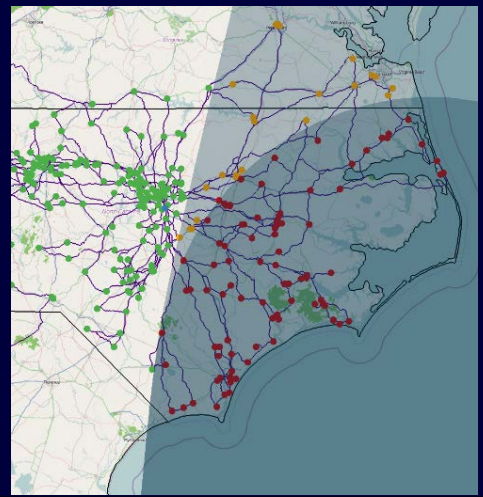
(inspired and adapted from David Brooks, New York Times, 2/18/2013)

- Qualitative, fuzzy, and social data
  - Preferences, significance of one relationship over another; Trust
- Context
  - Data is rarely complete, nor incorporates all relevant information Humans have extensive information and experience that never make it into the collected data
- Spurious vs. Significant
  - Big data means more statistically significant events and correlations, but they may not have any relevance
  - Increases noise to signal ratio
- Big problems
  - Complex, multifaceted, multiparameter big challenges with unquantified dependencies

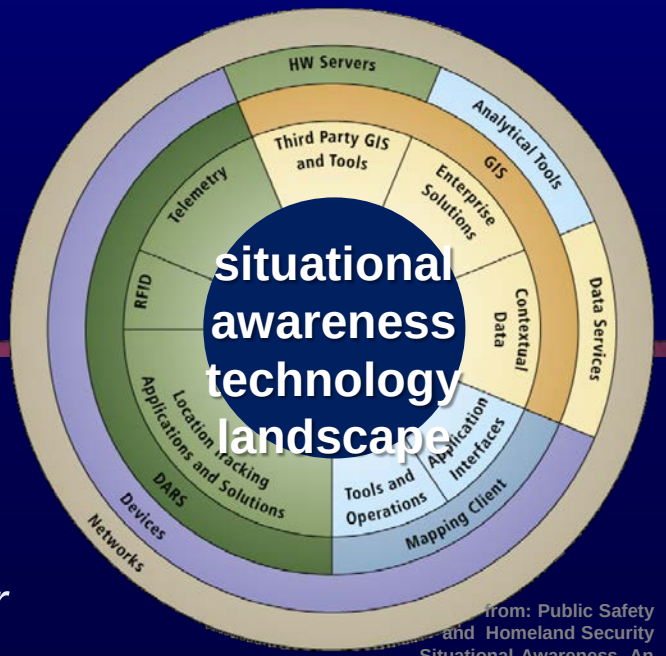
Visual Analytics Uses for Public Safety



- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Interactive, multi-day, month, type investigation
- Multisource, multimedia massive data integration & analysis



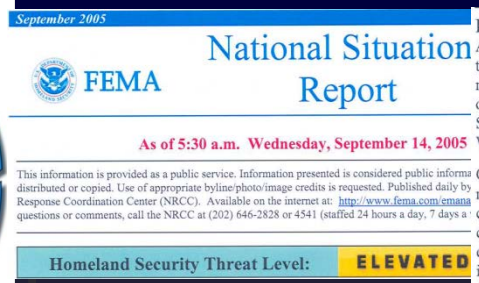
Situational Awareness & Assessment

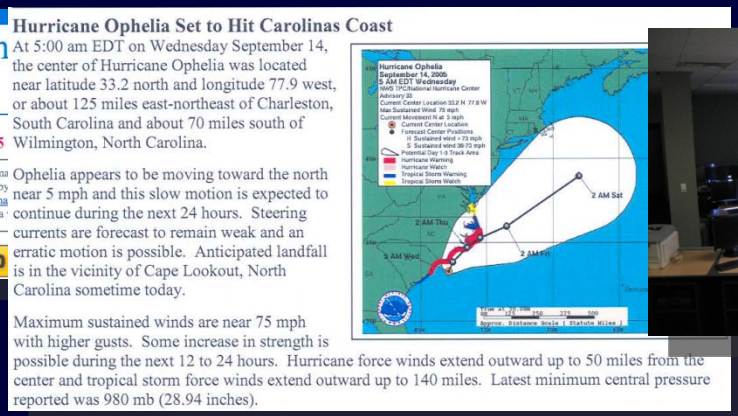


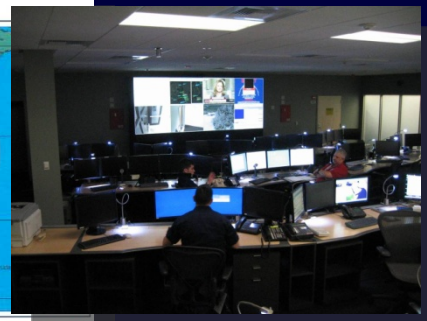
from: Public Safety and Homeland Security Situational Awareness, An ESRI ® White Paper • February 2008

- Three levels of SA (Endsley):
  - Perception, Comprehension, Projection
- *Team SA: "the degree to which every team member possesses the SA required for his or her responsibilities"*
- Problem addressed:
  - Develop, implement, test, and deploy methods that achieve (and continually update) the context for collection and integration of relevant information and support subsequent decision-making.









Team Situational Awareness/Assessment

- Outcomes:
  - A formal model of place-time context to support team situation awareness/ assessment
  - Analytic environments that enable information from multiple sources to be retrieved, summarized, fused and contextualized, allowing the end-users to navigate through the information space quickly and seamlessly to make real-time judgments and decisions.

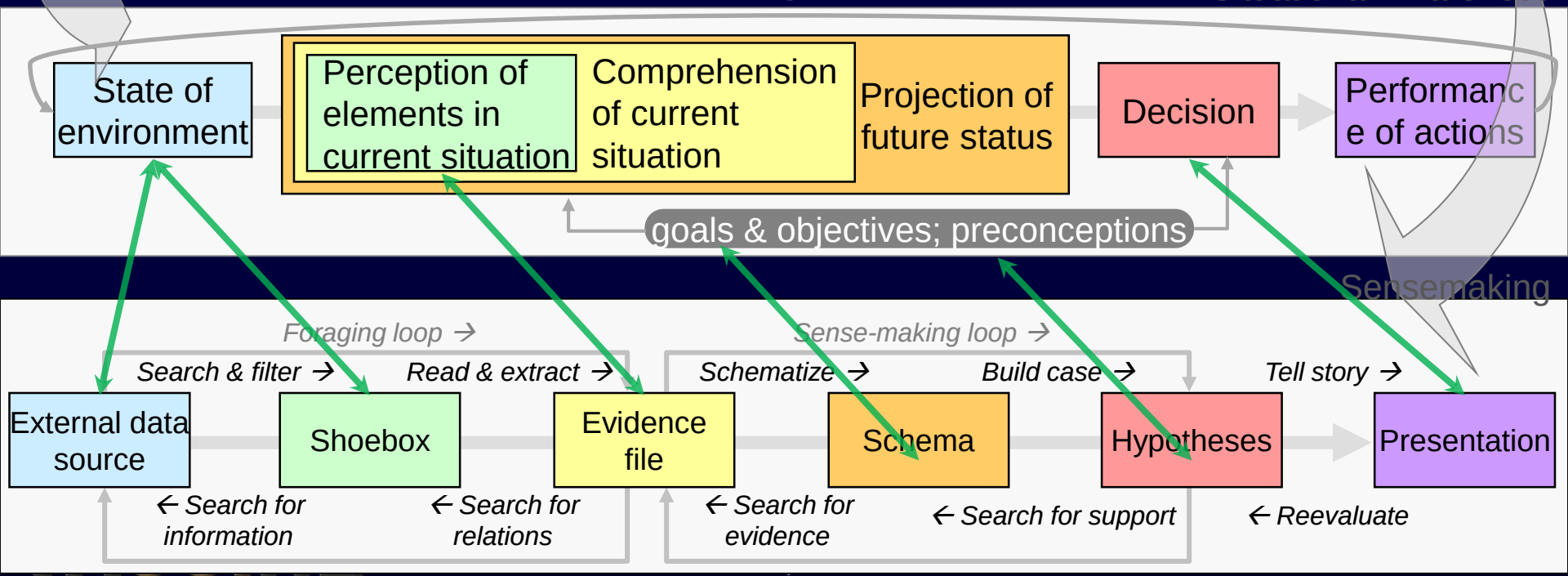


Visual Analytics vs. Situational Awareness

Situational Awareness ←→ Sensemaking

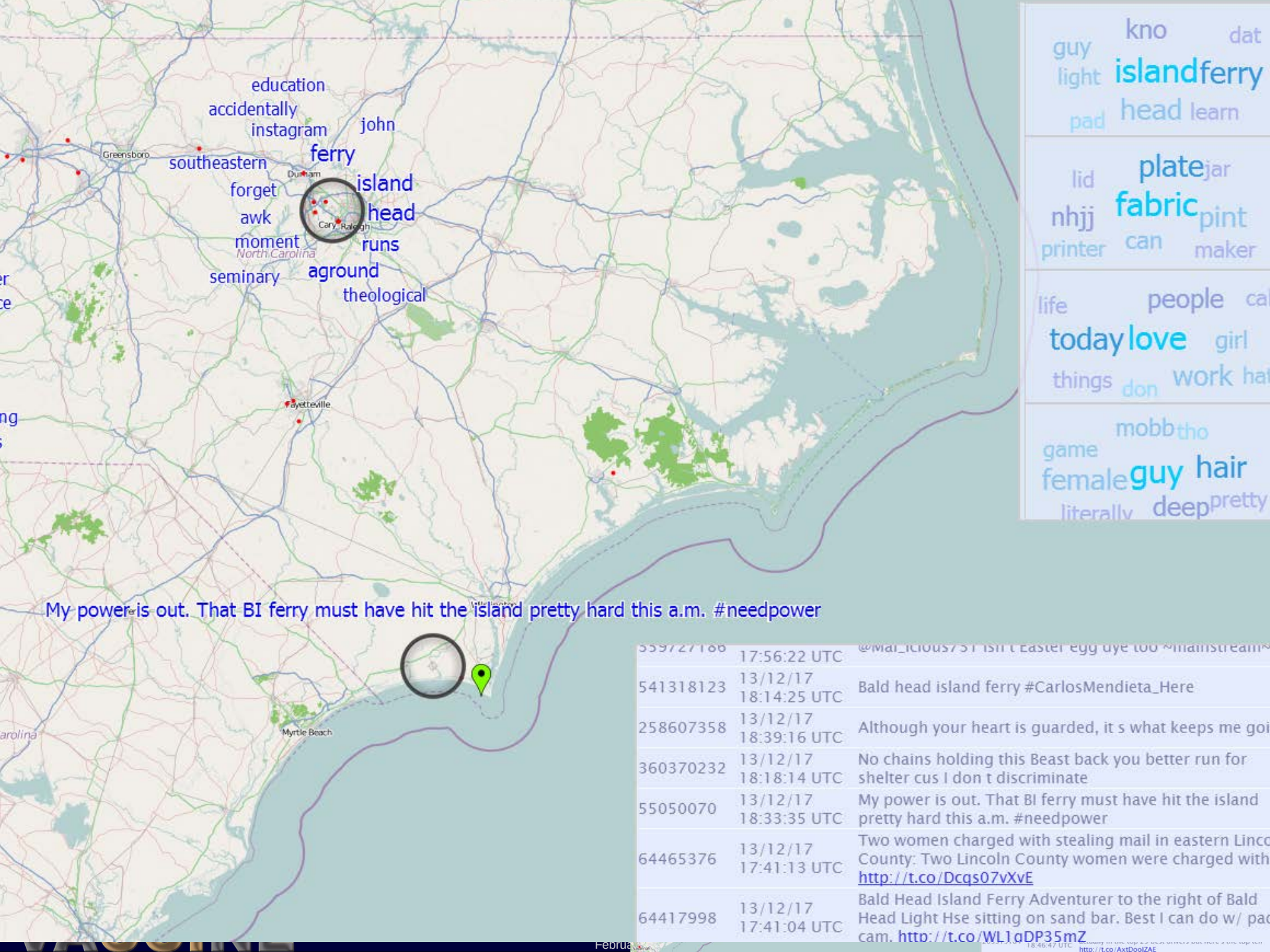
situational awareness as a sensemaking process (courtesy Alan MacEachren)

Situational awareness model (after Endsley, 1995; top) compared to model of sense making (after Pirolli & Card, 2005; bottom)



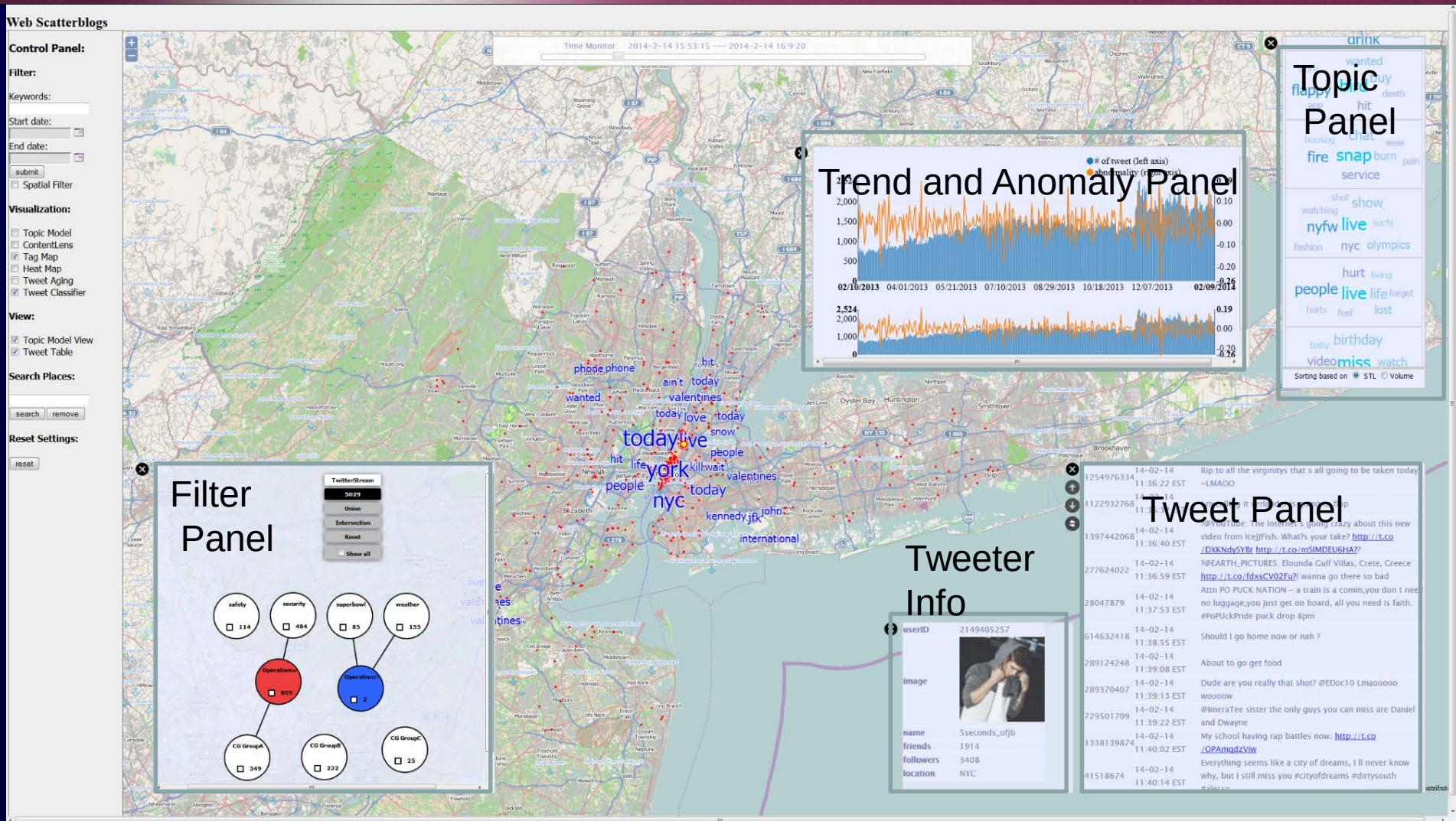
Team Situational Awareness, Planning, Investigation, Response Visual Analytics

- TSA from Big Data
  - In-field and desktop historical and current synchronized incident information from all relevant sources
  - Innovative spatiotemporal visualization, analysis, prediction
  - Image analysis tools
  - Integrated statistical forecasting and simulation models
  - Advanced social media analytics
- Tools for intelligence analysis
- Planning and response support
  - Integrated simulation, resource allocation, census, social media, weather data



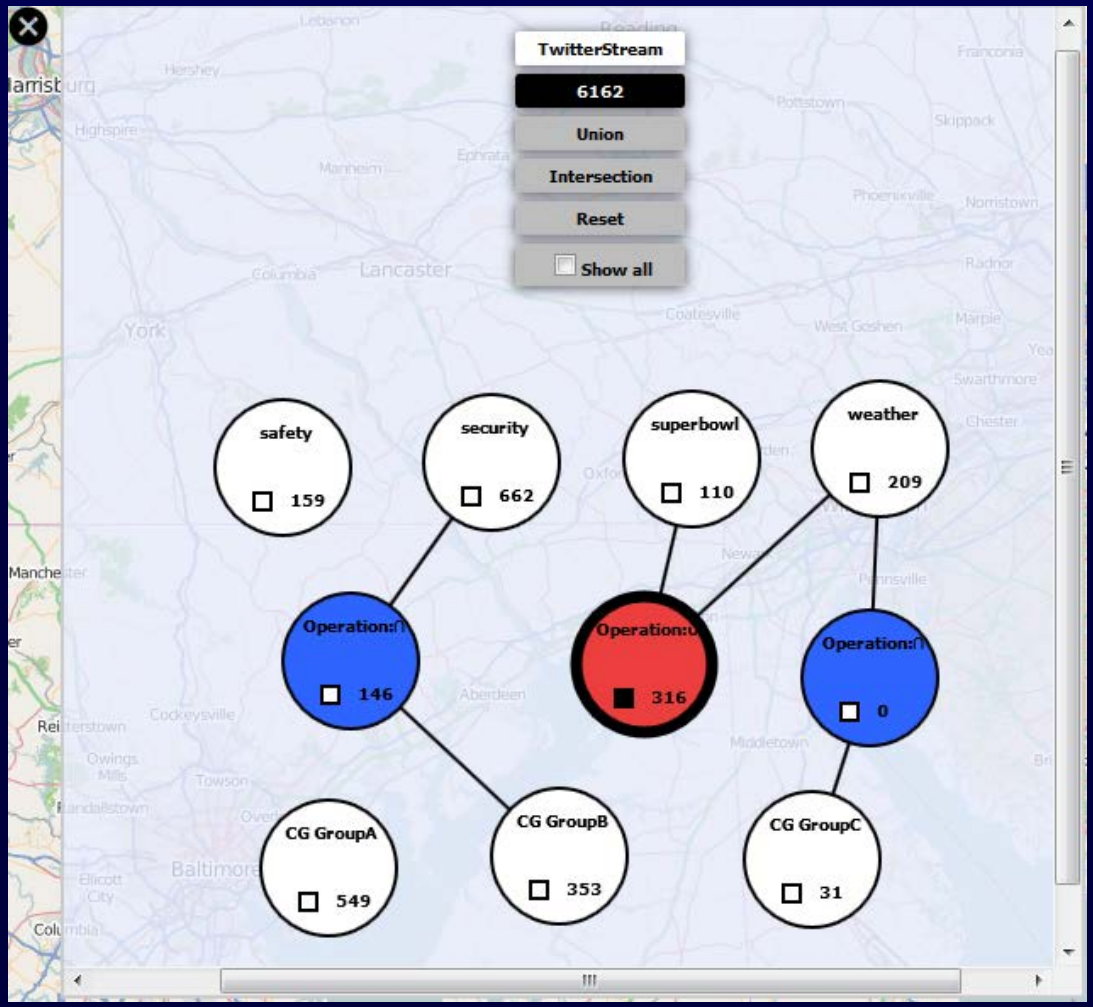
559727186	17:56:22 UTC	@Mal_icious731 isn't Easter egg dye too *mainstream*
541318123	13/12/17 18:14:25 UTC	Bald head island ferry #CarlosMendieta_Here
258607358	13/12/17 18:39:16 UTC	Although your heart is guarded, it s what keeps me goi
360370232	13/12/17 18:18:14 UTC	No chains holding this Beast back you better run for shelter cus I don t discriminate
55050070	13/12/17 18:33:35 UTC	My power is out. That BI ferry must have hit the island pretty hard this a.m. #needpower
64465376	13/12/17 17:41:13 UTC	Two women charged with stealing mail in eastern Linco County: Two Lincoln County women were charged with http://t.co/Dcqs07vXvE
64417998	13/12/17 17:41:04 UTC	Bald Head Island Ferry Adventurer to the right of Bald Head Light Hse sitting on sand bar. Best I can do w/ pac cam. http://t.co/WL1qDP35mZ

Scatterblogs – Web Version Overview

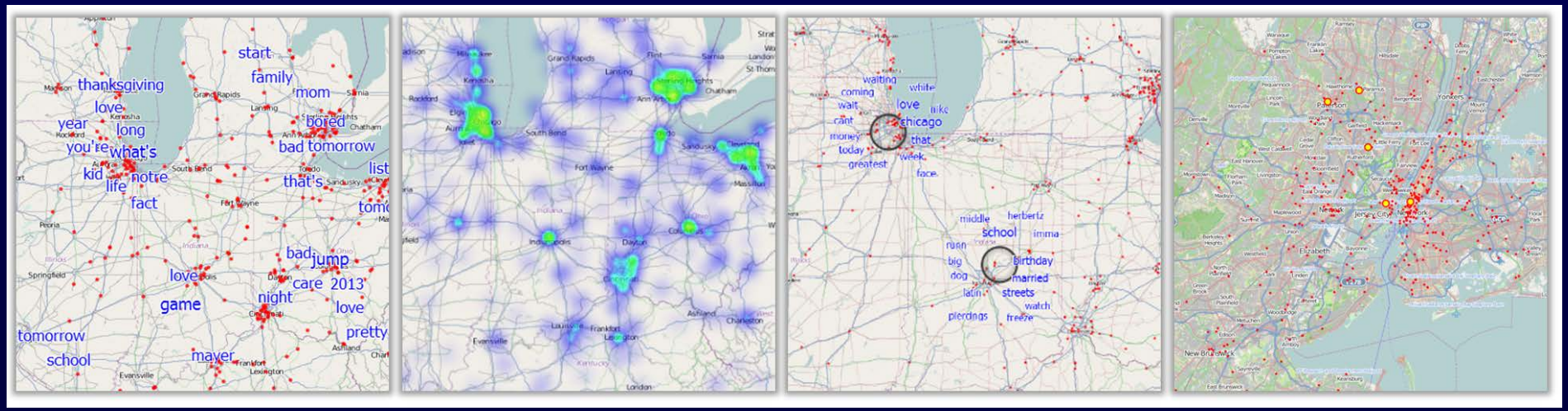


Scatterblogs – Filters

- Keywords to define various filters – can be an early warning system
- Can combine filters to specify overlap
- Can view tweets only associated with these filters to reduce information overload

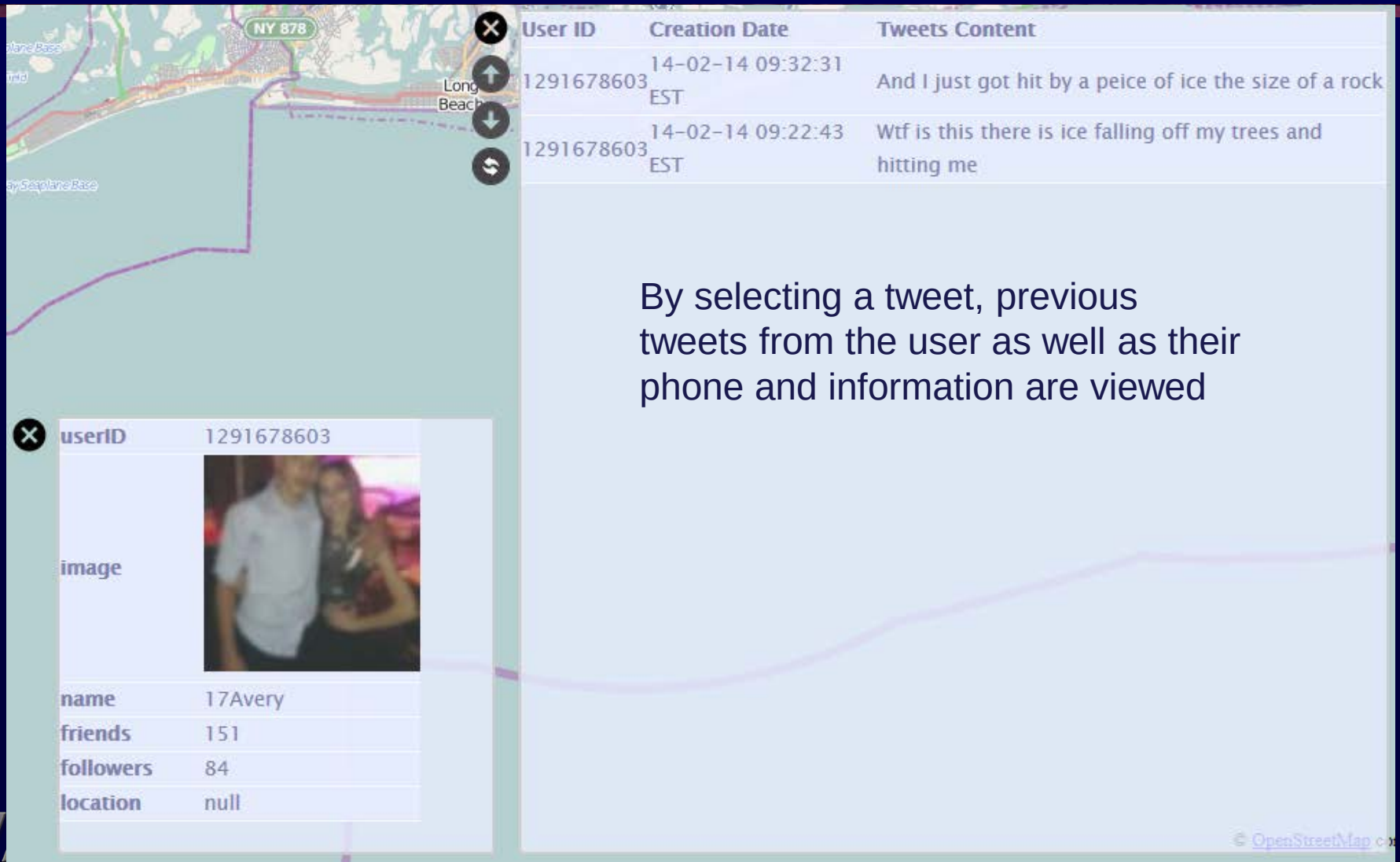


Scatterblogs – Screen Options



Tag Map | Heat Map | Content Lens | Realtime tweet aging

Scatterblogs – Tweeter information



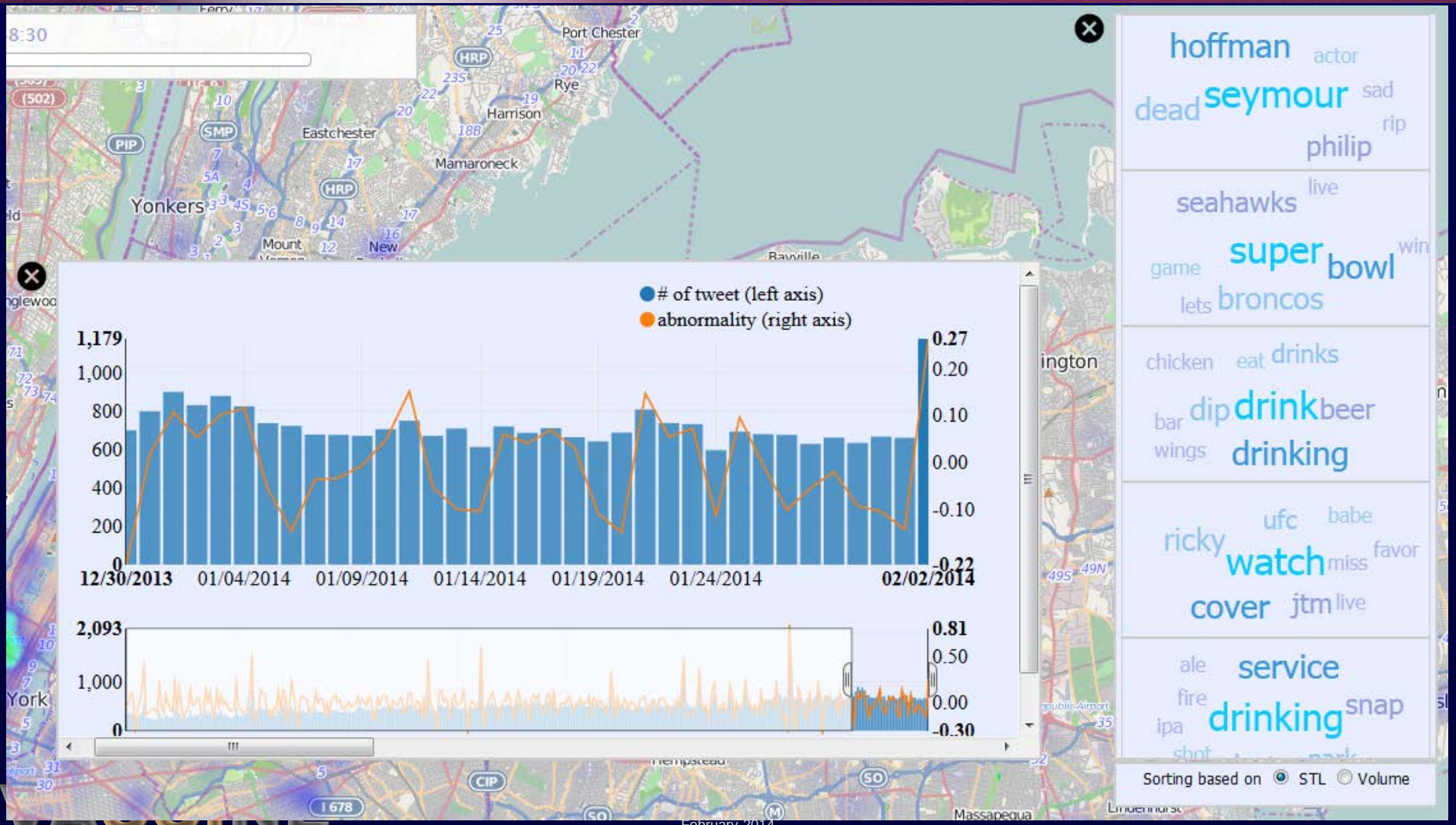
By selecting a tweet, previous tweets from the user as well as their phone and information are viewed

Scatterblogs – Tweet Panel

User ID	Creation Date	Tweets Content
480897943	14-02-14 09:51:50 EST	You dont deserve the restraint I show by not wiling out and smashing things that could be considered Vase Art in front of you
38029183	14-02-14 09:57:41 EST	YOU JUST SURVIVED 2 FEET OF SNOW, POWER LINES! WHY DO YOU CHOOSE 3 MINUTES BEFORE SKATING TO GO OUT!!!!
507982726	14-02-14 10:04:52 EST	I swear whenever there is a little accumulation of snow Mr. Sullivan looks outside and says Eh not today then goes and watches PBS
253765655	14-02-14 10:22:23 EST	@TheBeerWolf good to know lol @ESPNNYYankees: More spring training news: J.R. Murphy is now going by John Ryan Murphy
1058977081	14-02-14 10:34:20 EST	Can t wait to go snowboarding the next three days
76140820	14-02-14 10:36:20 EST	Part of me is secretly happy when we have snow days then I realize I m not in High school and I actually need to go to have class.
328196190	14-02-14 10:53:25 EST	My sis is going to Puerto Rico tomorrow and I ll be hibernating through another snow storm so dats cool
93842621	14-02-14 11:17:15 EST	I spit in the floor and the wind pushed it back and fell on the back of this girl s UGGS ??
28047879	14-02-14 11:37:53 EST	Attn PO PUCK NATION – a train is a comin,you don t need no luggage,you just get on board, all you need is faith. #PoPUckPride puck drop 8pm

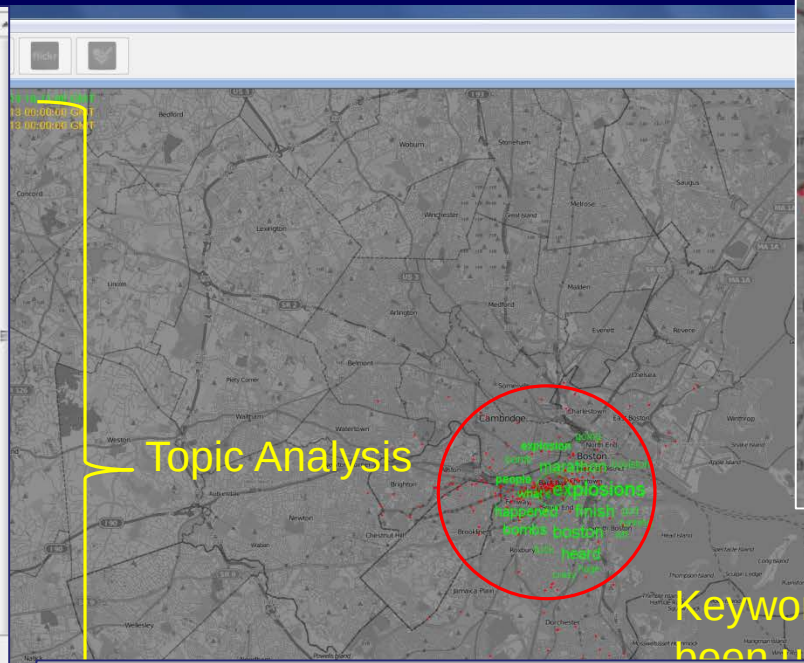
VACCINE

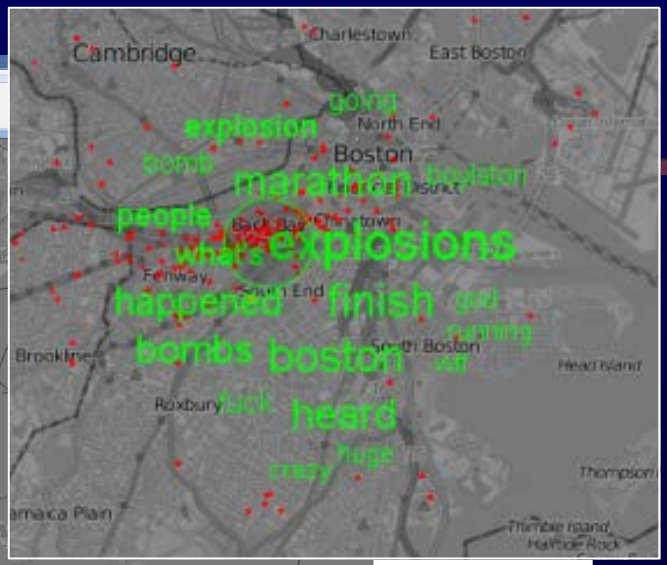
Anomaly Detection



Explosion Area in Boston







Keywords that have been used most often

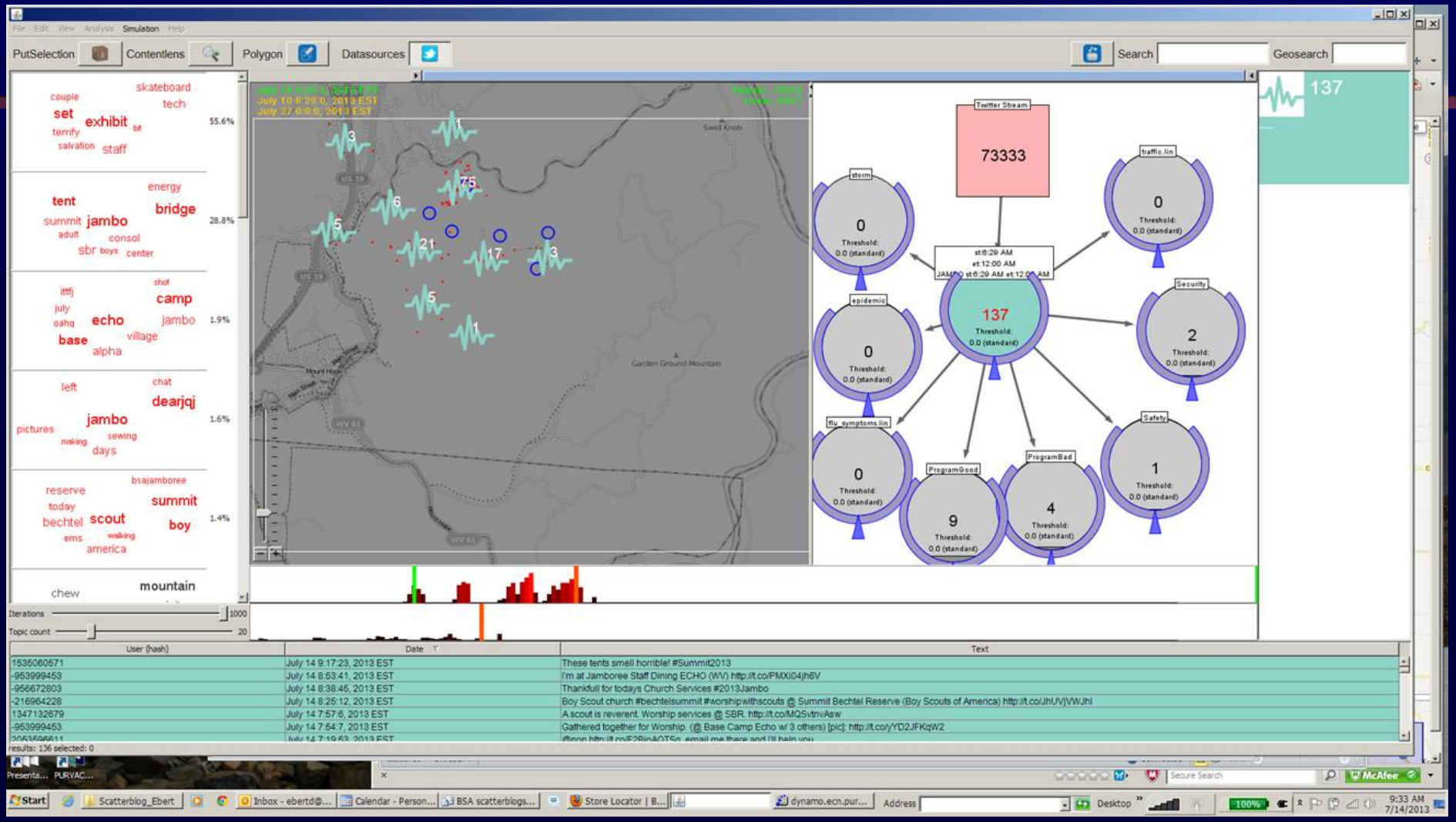
Tweets

Text
Oh my god what just happened
Something happened at the end of the Boston Marathon. Something bad and there is a lot of chatter on Twitter. What's going on?
Multiple people are injured near the Boston Marathon finish line after two explosions. The #BostonMarathon has been stopped.
two bombs just went off on boylston
BREAKING NEWS: Two powerful explosions detonated in quick succession right next to the Boston Marathon finsh line this afternoon.
What the FUCK was that
Back in Sept, @croon1 solicited me for $2000. He now has a music video with William Shatner. If you watch it (god forbid) keep that on mind.
Literally what the fuck get me out of here
@DTenenbaum my office right next to it
@FRANCESCalciO I figured 3 people would get the joke
Two explosions just rocked the finish line of the Boston Marathon. Sirens galore. People running in fear. Wonder what happened.
This is crazy i seen that blow up #bostonmarathon
can someone tell me what that explosion was!? #boston #bostonmarathon

First Response (Tweet & Picture)



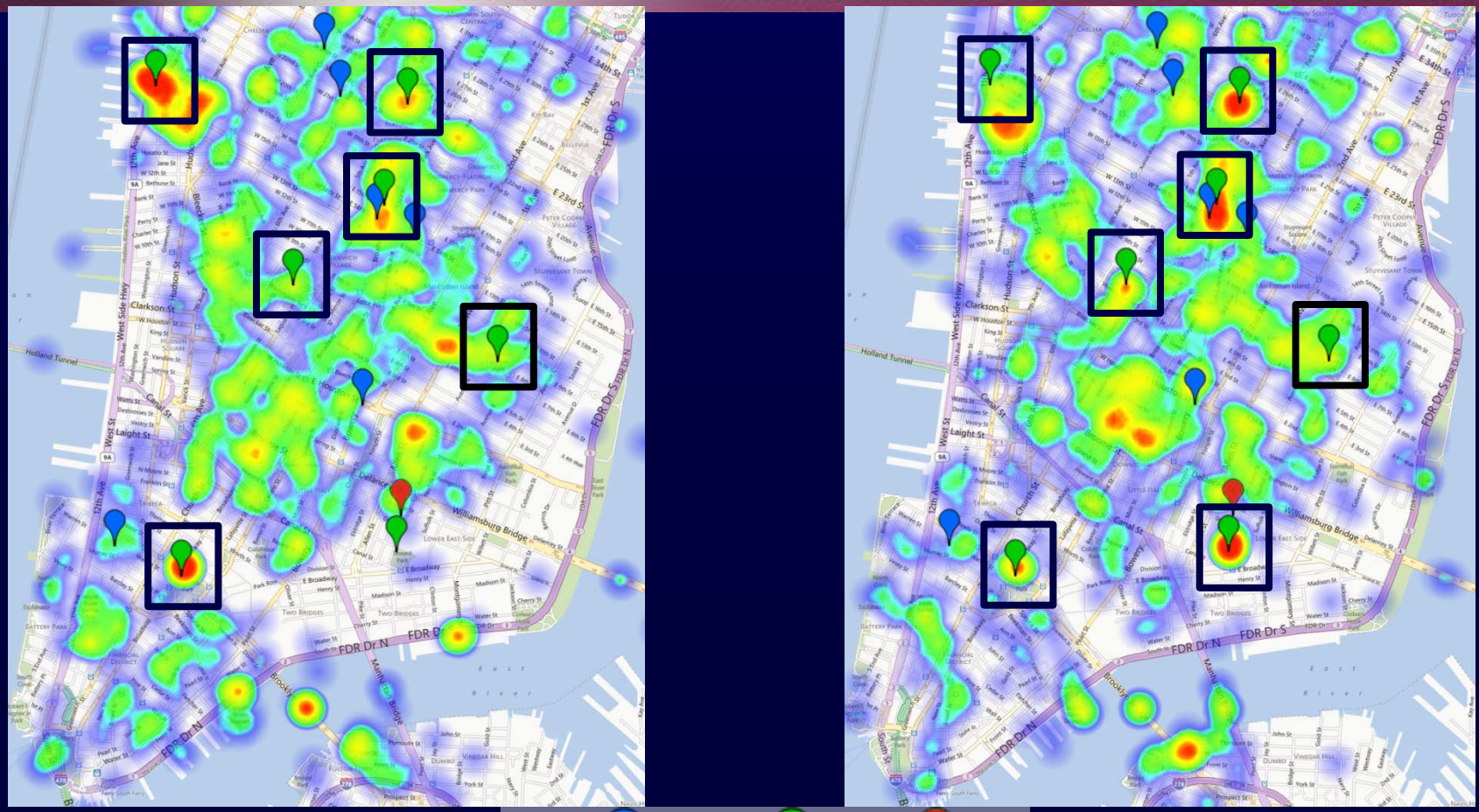
BSA Jamboree



Visual Analytics of Activity During Hurricane Sandy

Two weeks before Sandy
10/14 (Sunday), 12:00 ~ 16:00

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00

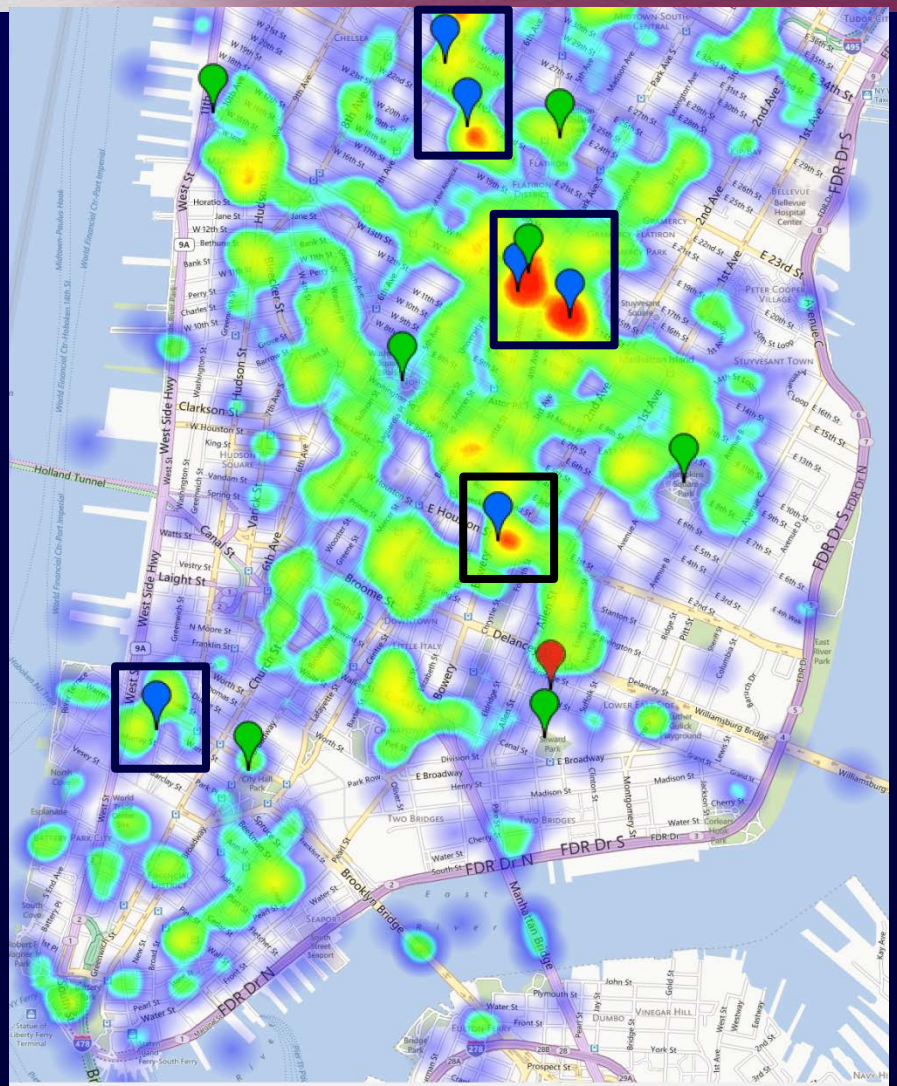


Visual Analytics of Activity During Hurricane Sandy

After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00

Evacuation order:
10/28, 10:30 AM

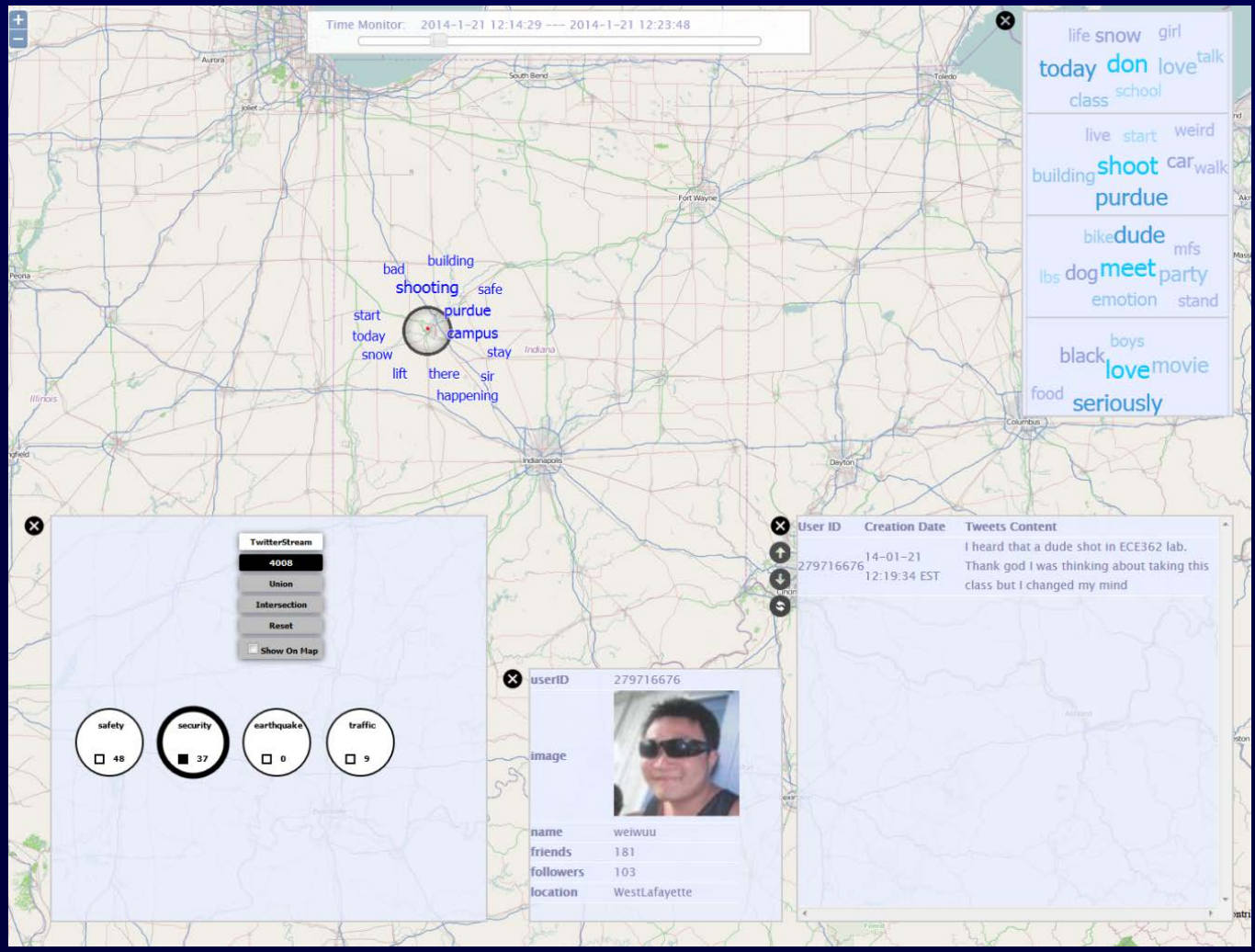
Hurricane Sandy's
Arrival at NYC:
10/29, 8:00 PM



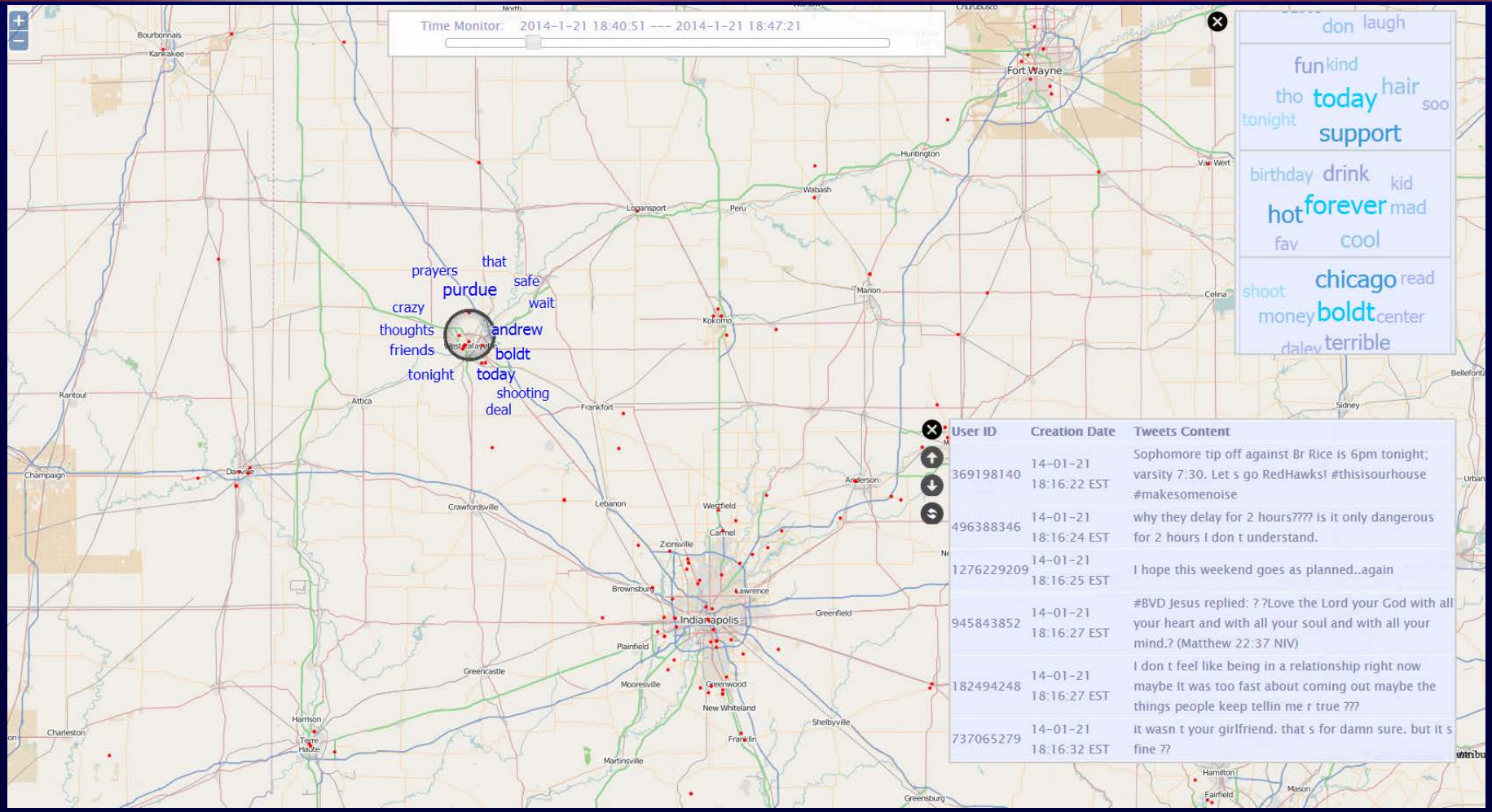
VACCINE

Purdue Shooting

- Specific individual's tweets 20 minutes after the shooting
- The content lense highlights the word "shooting"

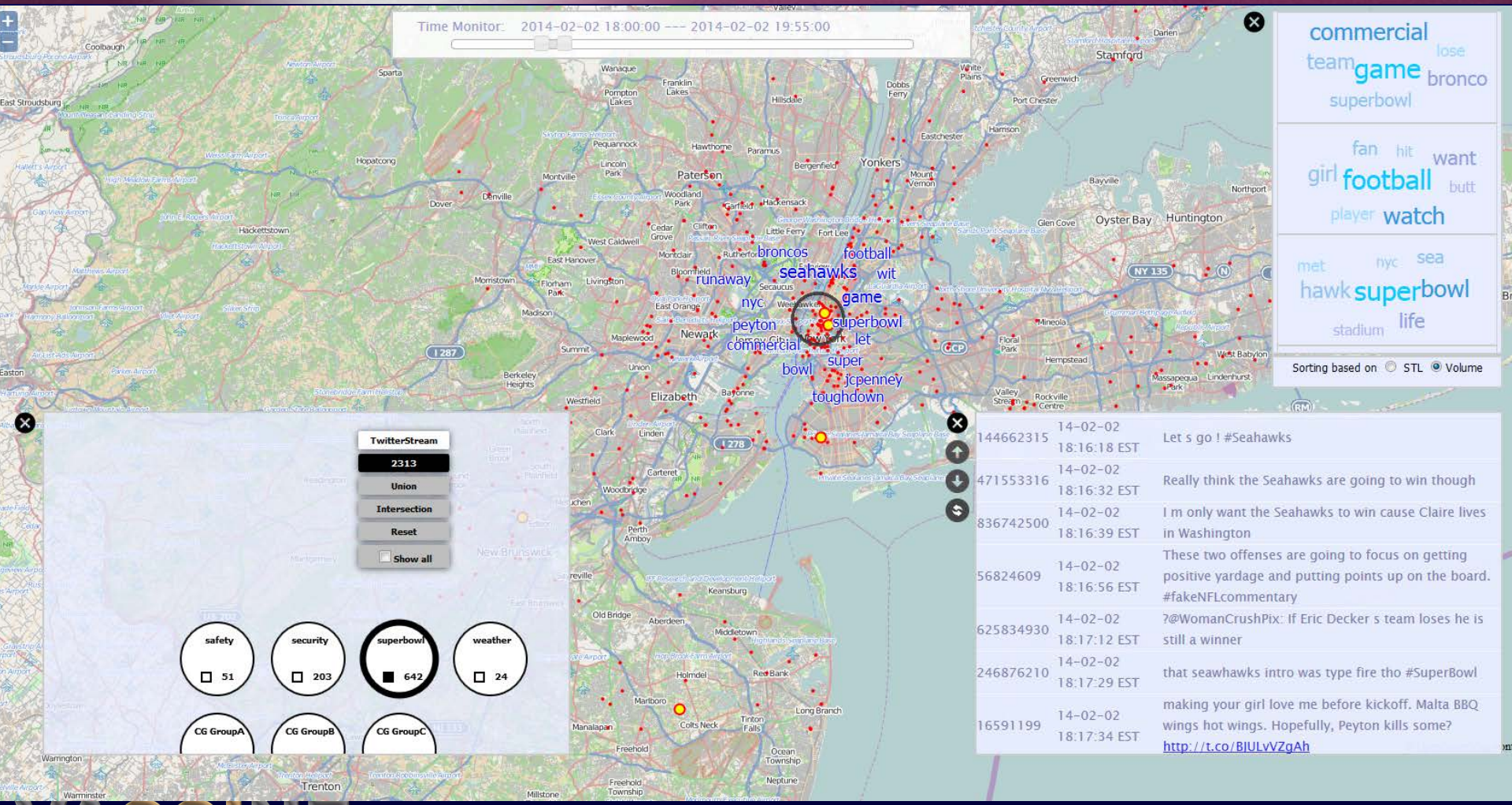


Purdue Shooting



"Andrew Boldt"'s name is highlighted in the content lens

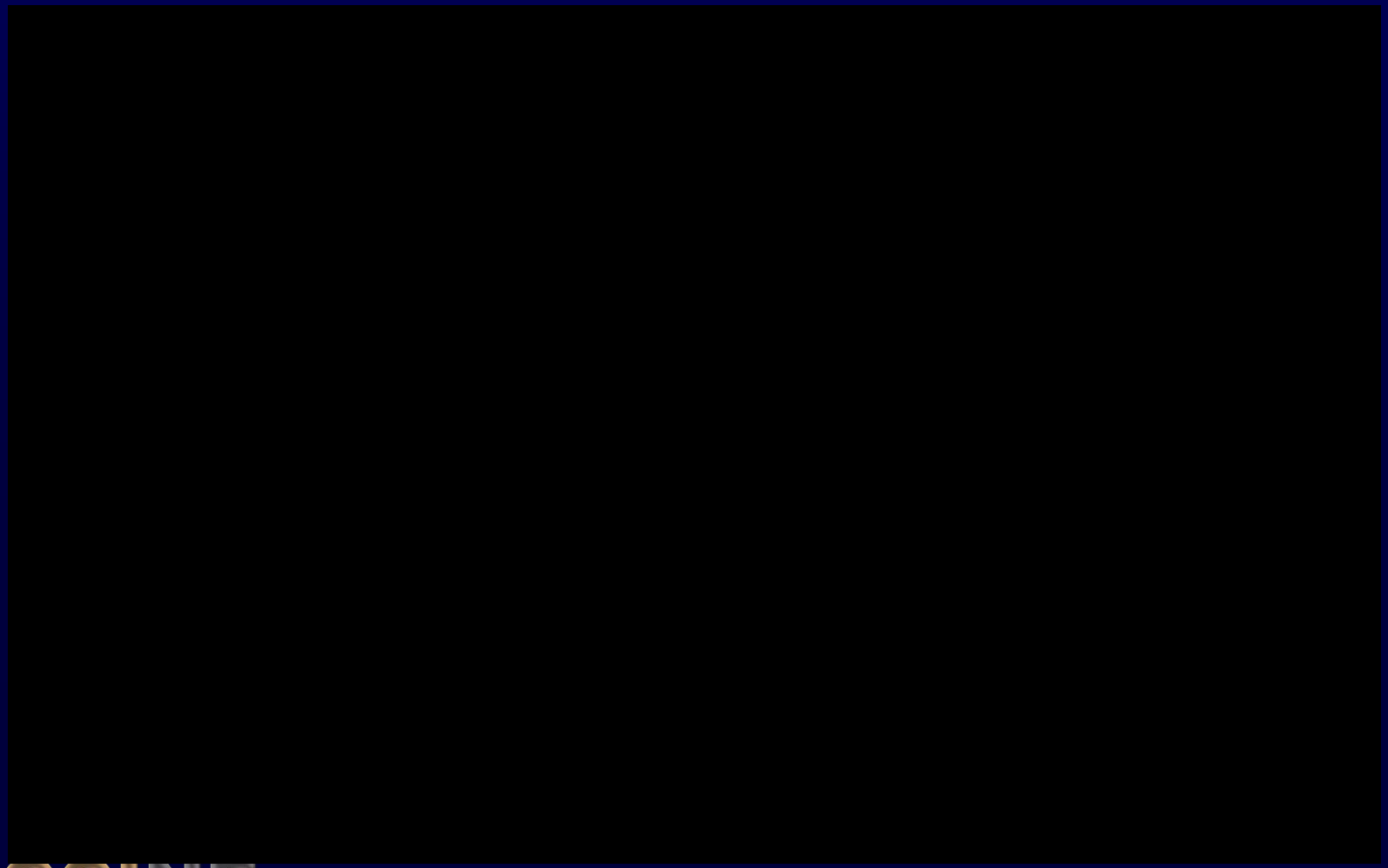
Super Bowl Examination



New York Harbor – 2014 Snapshot

- Data from Jan 1 ~ Feb 14
- At two points in the video (00:29 and 00:52) the words, 'help' appears in conjunction with 'children', 'life', 'future' and 'financial', indicating some tweet ads for children's financial future
- At 00:55, the words 'seymour' and 'hoffman' appear with 'body' and 'lost‘
- In the topic view, 'lost' goes with 'arizona' several times, indicating Arizona Wildcats lost their basketball game

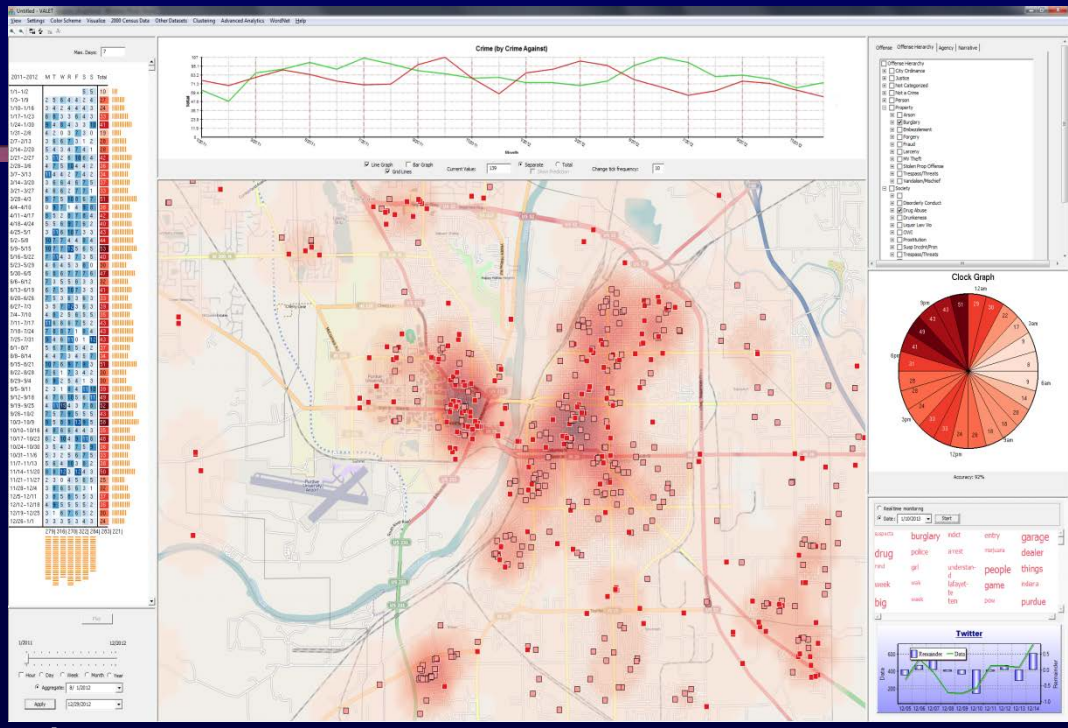
New York Harbor – 2014 Snapshot



Future Directions

- Integration of advanced text semantics (CMU)
- Text message alerts
- Multi-language
- Credibility/reliability of source and/or information
- Comparison of accuracy and speed of social media versus traditional news media

Integration with Other Tools: Visual Analytics Law Enforcement Toolkit (VALET, iVALET)



Impacts:

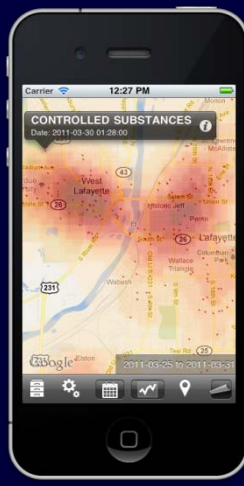
- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public (February 2013) for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

VALET delivered:

- Spring 2011: WL, Lafayette Police

iVALET delivered:

- October 2011: Purdue, WL Police

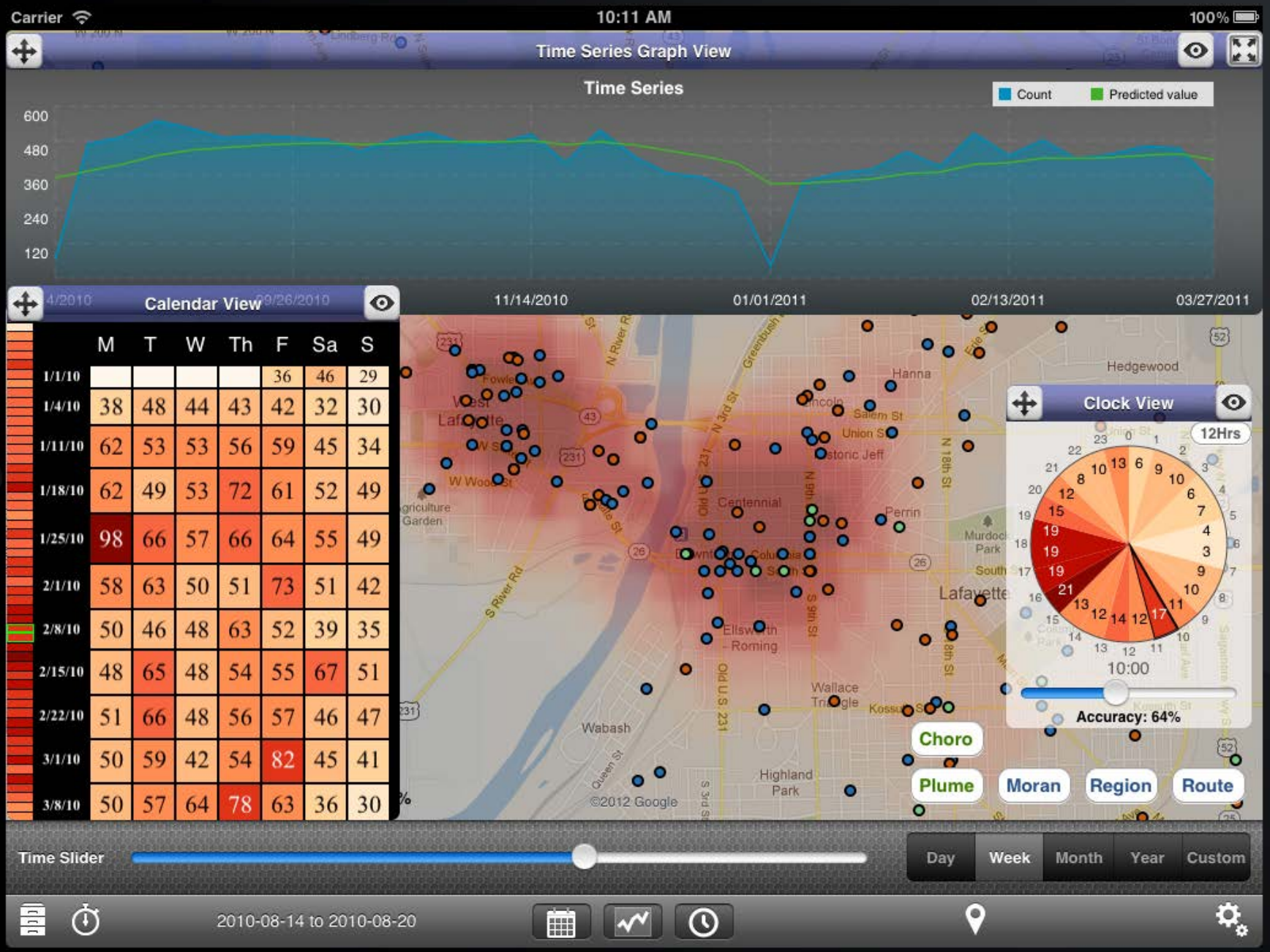


VALET Feature:
Real-time Twitter Monitoring and Integration

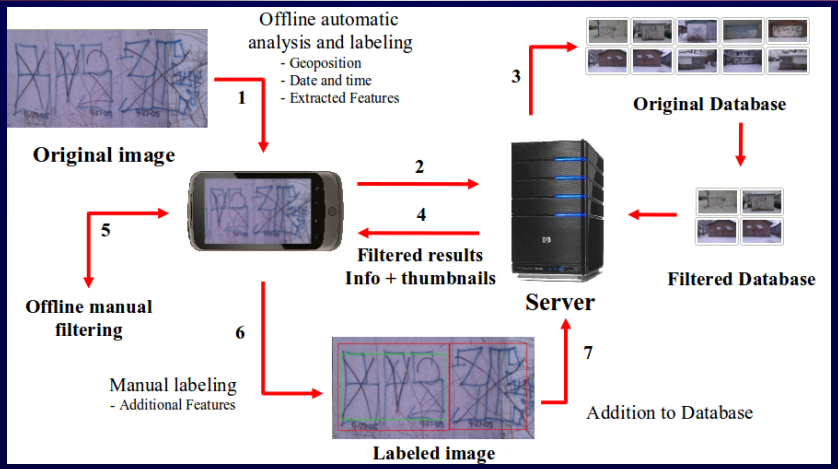
- Topic extraction using novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud



Grand Prix Weekend, Purdue University


Carrier
10:11 AM
100%
Time Series Graph View
Time Series
Count
Predicted value
600
480
360
240
120
11/14/2010
01/01/2011
02/13/2011
03/27/2011
Calendar View
M T W Th F Sa S
1/1/10 36 46 29
1/4/10 38 48 44 43 42 32 30
1/11/10 62 53 53 56 59 45 34
1/18/10 62 49 53 72 61 52 49
1/25/10 98 66 57 66 64 55 49
2/1/10 58 63 50 51 73 51 42
2/8/10 50 46 48 63 52 39 35
2/15/10 48 65 48 54 55 67 51
2/22/10 51 66 48 56 57 46 47
3/1/10 50 59 42 54 82 45 41
3/8/10 50 57 64 78 63 36 30
Clock View
12Hrs
10:00
Accuracy: 64%
Choro
Plume
Moran
Region
Route
Time Slider
Day
Week
Month
Year
Custom
2010-08-14 to 2010-08-20
©2012 Google

Gang Graffiti Analysis and Recognition Using a Mobile Telephone (GARI)





IMPACT:

- Allows police to catalog and analyze gang graffiti images into a database system to better track and determine gang activity throughout a region
- Will allow the graffiti images to be "interpreted"
- More than 140 users and 1300 graffiti images acquired

GARI delivered:

- Summer/Fall 2011:
  - IPD gang detectives
  - IPD at large
  - Ind Fusion Gang Task Force
  - Gang detectives across Indiana
- Spring/Summer 2013:
  - Cook County, IL
  - INGang ownership
  - Tattoo database added

VACCINE

Jigsaw

Visualization for Investigative Analysis across Document Collections

- Law enforcement & intelligence community
- Fraud (finance, accounting, banking)
- Academic research
- Journalism & reporting
- Consumer research

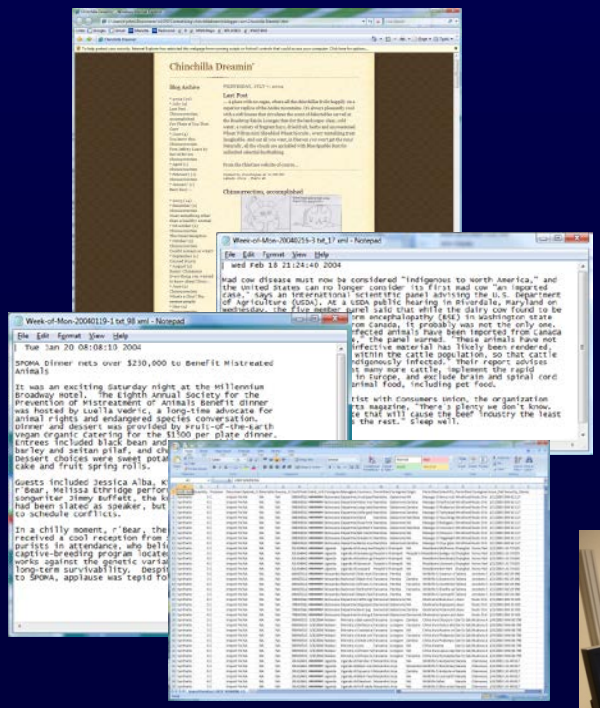
- Help “investigators” explore, analyze and understand large document collections



“Putting the pieces together”

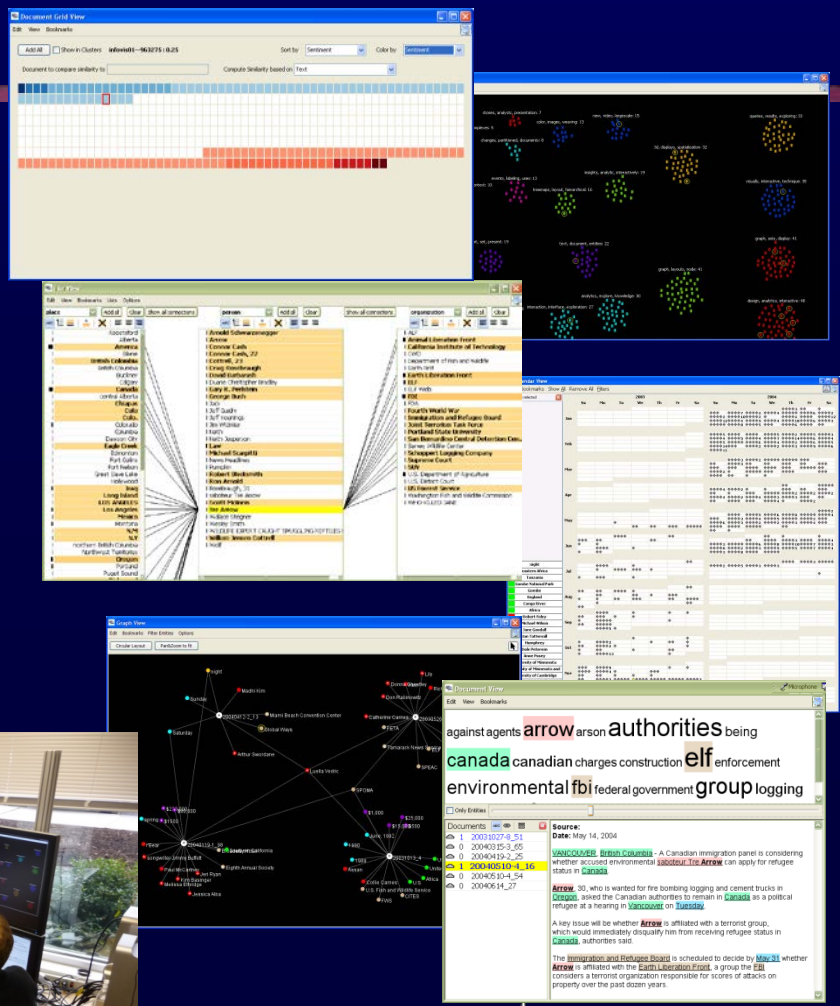
Jigsaw: Visual Analytics for Investigative Analysis

John Stasko



Documents in

Knowledge out





VACCINE

Georgia Tech College of Computing
School of Interactive Computing

For Further Information

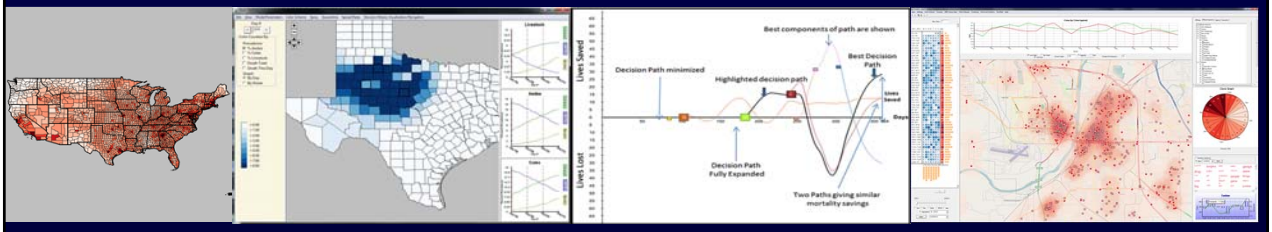
www.VisualAnalytics-CCI.org



VACCINE

Visual Analytics for Command, Control, and Interoperability Environments
A U.S. Department of Homeland Security Center of Excellence

Visual Analytics: Overcoming Challenges to Create Tools for Decision Making



David S. Ebert
Purdue University

October 2013

VACCINE

Outline

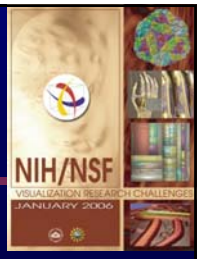
- NIH/NSF Visualization Research Challenges Report revisited
- Visual analytics
- Challenges and opportunities
- Example tools for decision makers
- Directions

VACCINE

October 2013

Vis 2013

NIH/NSF Visualization Research Challenges Report[1]: Key Points



- *"Continuing and accelerating these advancements will require people to comprehend vast amounts of data and information being produced from a multitude of sources. … Visualization will be critical in achieving this goal."*
- *"Further research in visualization, and the technology transfer of effective visualization methodologies into the working practice of medicine, science, engineering, and business, will be critical in handling the ongoing information explosion."*
- Policy changes should encourage collaboration between visualization and domain researchers and emphasize the importance of evaluation
- Promote foundational, transitional, and applied research
- Collaboration to solve real world problems

VACCINE

1) Johnson, C., Moorhead, R., Munzner, T., Pfister, H, Rheingans, P., Yoo, T., "NIH/NSF Visualization Research Challenges," January 2006

October 2013



NIH/NSF Visualization Research Challenges Report[1]: Key Points

- Characterize how and why visualization works
- Systematically explore the design space of visual representations
- Develop new interaction approaches
- Exploit the possibilities of novel display hardware
- Integrate visualization with other techniques and tools from other disciplines
- Create open data and task repositories

VACCINE

1)Johnson, C., Moorhead, R., Munzner, T., Pfister, H, Rheingans, P., Yoo, T., "NIH/NSF Visualization Research Challenges," January 2006

October 2013



Where Are We Today?

- Most points are still very relevant
- Federal policies and funding have not taken this advice
- Visualization and analytics advances have not kept up with the rate of data increase!
- More collaboration does exist and is producing some of the most impressive results!



Where Are We Today?

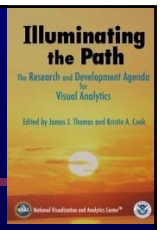
- Visual analytics has sparked integration of techniques from other disciplines – but it is just a start!
- Policy makers are starting to use visualization more effectively with the efforts in data driven policy development
- Visualization still isn't as heavily and effectively integrated in research, practice, and policy development as it should be
  - E.g., FDA impediments to interactive visualization for drug discovery
- Data is in some ways more available, but still less available



Where Are We Today? Data

- More
  - Electronic health records
  - Electronic syndromic surveillance data
  - Sources of public data (web searches, social media)
- Less
  - Most data sources not integrated
  - Privacy concerns/policies limit multiple uses
  - Techniques for appropriate comparable feature extraction not integrated
  - Correct aggregation and disambiguation techniques lacking

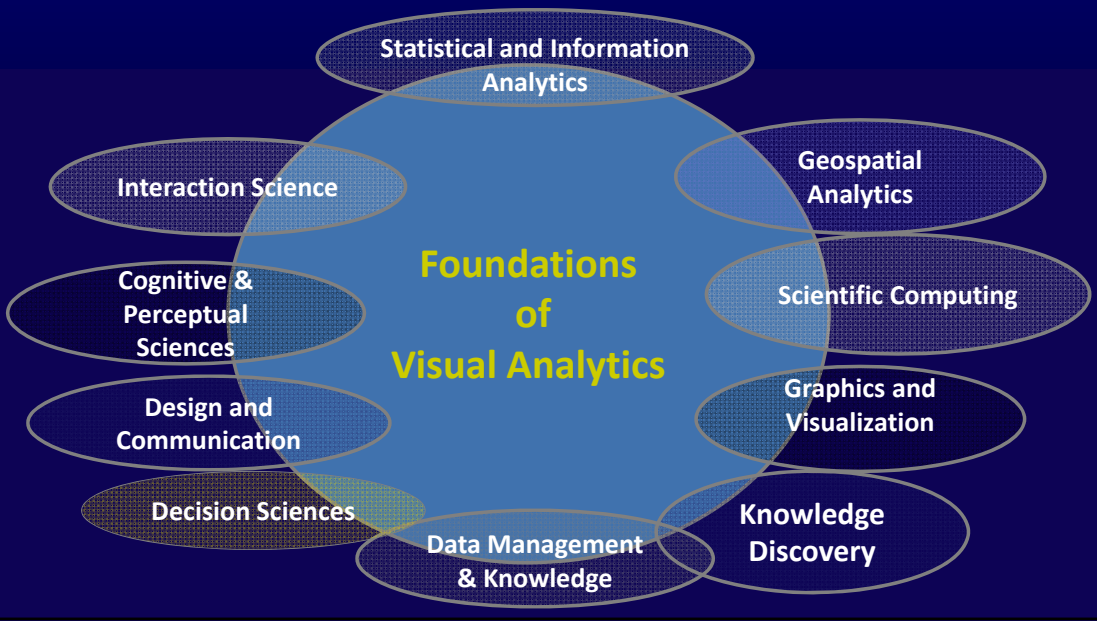
Role of Visual Analytics as a Solution



- Visual Analytics[1] is the science of analytical reasoning facilitated by interactive visual interfaces
- People use visual analytics tools and techniques to
  - Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
  - Detect the expected and discover the unexpected
  - Provide timely, defensible, understandable assessments
  - Communicate assessment effectively for action

1. *Illuminating the Path: The R&D Agenda for Visual Analytics, Editors: Thomas and Cook*

Role of Visual Analytics as a Solution



Role of Visual Analytics as a Solution

- Research agenda and funding has emphasized application-driven research to accelerate research and solve real world problems
- New interaction research in interactive decision making environments
- Evaluation of effectiveness of complex tasks a key research area

Example: VACCINE

- What we do
  - Help people make effective decisions from massive data
  - **Big Data Visual Analytics**
- Who we help
  - Law enforcement (NYPD, ISP, WLPD)
  - Public safety and health (USCG, TSA, ISDH)
  - Scientific discovery
  - Business production, resource allocation, risk-based decision making

Types of technologies

Novel visualization techniques and analytics

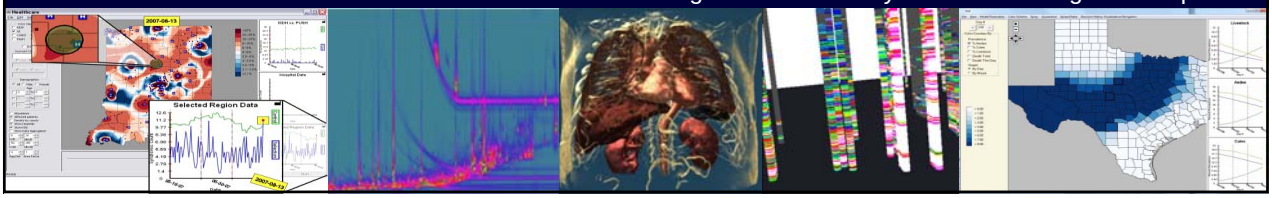
Mobile graphics, volume graphics, procedural abstraction of complex and massive data

Research Motivation:

- Solving these real-world problems requires
  - Novel theories, techniques, approaches, and adaptations of algorithms
  - Integration of cross-disciplinary expertise
- Solving these real-world problems provides
  - Compelling, publicly understandable value for your research
  - Advances in CS and in other disciplines
  - New publication opportunities
  - Great collaboration partners and proponents
  - Opportunities for new adventures

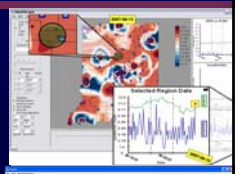
Some Examples: Health Visual Analytics

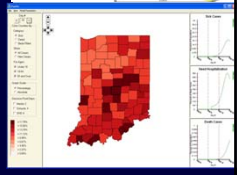
Health Surveillance | Cancer Detection | Treatment Planning | Biomarker Discovery | Zoonoses Planning and Response

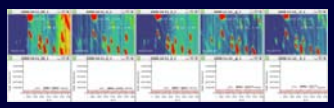


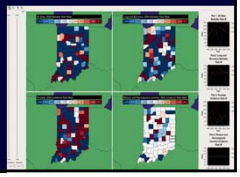
Range of Our VA Health Applications

- Public Health Surveillance
  - Indiana and Pakistan
- Pandemic Preparedness and Planning
- Cancer Care Engineering
  - Biomarker discovery
  - Regional cancer analysis







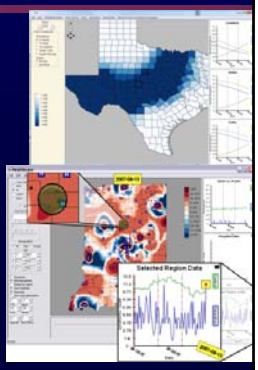


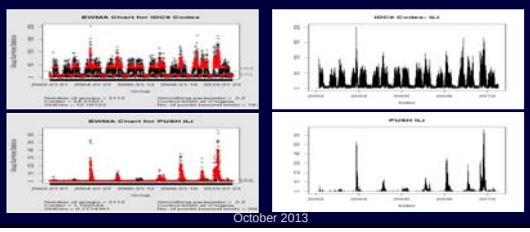
VACCINE

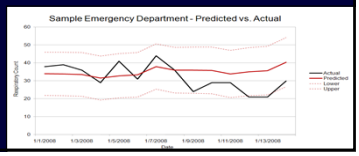
October 2013

Range of our VA Health Applications

- OneHealth – Linking Animal and Human Health
- Public Health Information Exchanges
- Practice Management





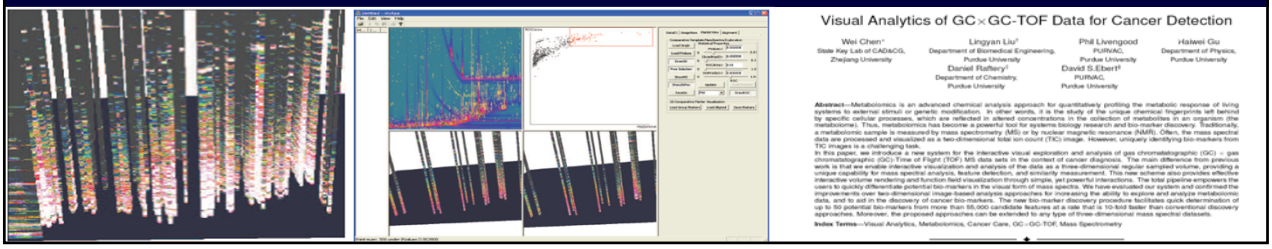


VACCINE

October 2013

Omics Visual Analytics

- 100 times speed-up in biomarker discovery process of metabolomics data
- Potential cancer biomarker found using tools
- Deployed on CCEhub
- Publication in *IEEE Transactions on Visualization and Computer Graphics*



Visual Analytics of GC×GC-TOF Data for Cancer Detection

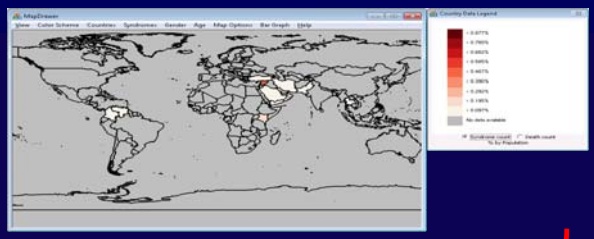
Visual Analytics For Syndromic Surveillance

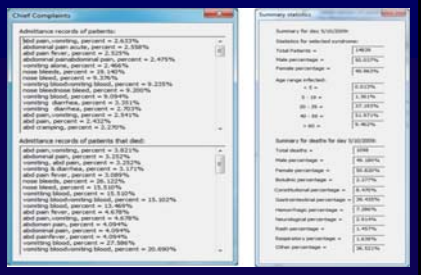
- Syndromic surveillance is the detection of adverse health events focusing on pre-diagnosis information to improve response time
- Pre-diagnosis information can consist of multiple data sources:
  - Over the counter medicine sales
  - News reports on emerging diseases
  - Promed news feeds
  - Emergency department chief complaints

Improving Syndromic Surveillance

- Benefits/ impact
  - Enhanced hypothesis testing capabilities
  - Linked views allow quicker cross validation of hypothesis
  - Less time investigating false positives
  - Systemic biological pandemic, syndromic, chem/bio/nuclear surveillance, management, and response

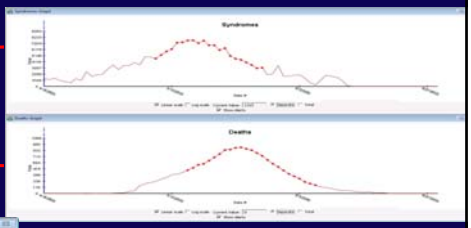
Syndromic Surveillance





Patient ID	Date	Chief Complaints	Location
9398	4/16/09	Ear pain	Karachi
10816	4/16/09	Stuffy nose	Lebanon
1491	4/16/09	Fever	Allepo
16237	4/16/09	Head bleed	Yemen

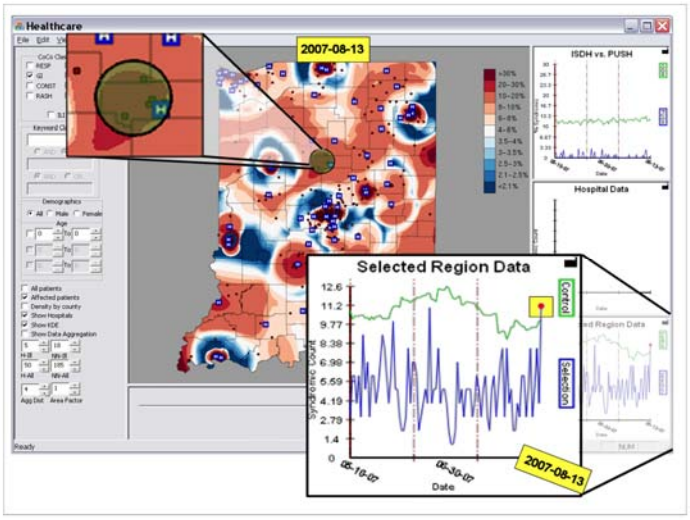
CoCo Classifier





VAST 2010 Challenge Award: Support for Future Detection, IEEE Conference on Visual Analytics Science and Technology, October 2010 for "VACCINATED," (Malik, A., Afzal, S., Hodgess, E., Ebert, D., Maciejewski, R.).

Visual Analytics for Syndromic Surveillance: Hypothesis Generation and Exploration

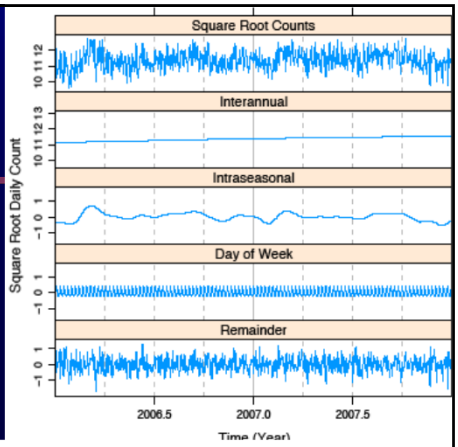


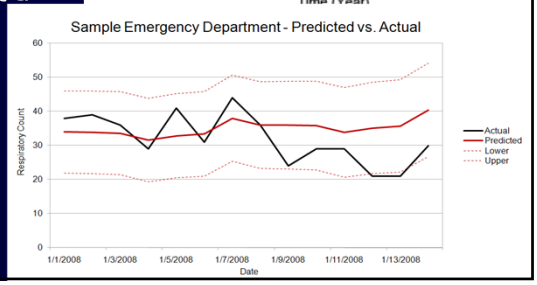
Best Paper Nominee, IEEE Symposium on Visual Analytics Science and Technology (VAST), October 2008, for "Understanding Syndromic Hotspots – A Visual Analytics Approach," (Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., Cleveland, W., Grannis, S., Wade, M., Ebert, D.).

20

Predictive Visual Analytics

- Time Series Modeling:
  - Seasonal-Trend Decomposition Based on Loess
    - Time series ~ sum of multiple trend components
    - For each data signal, components are extracted
    - Can then analyze correlation between components
- **Predictive Visual Analytics using STL**

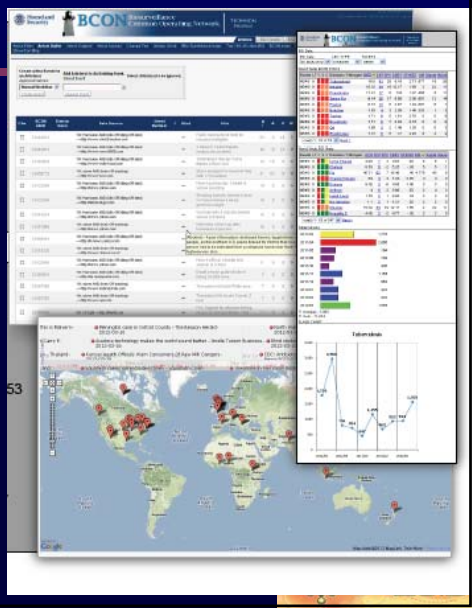




Hafen, R., Anderson, D., Cleveland, W., Maciejewski, R., Ebert, D., Abusalah, A., Yakout, M., Ouzzani, M., Grannis, S., "STL for Modeling, Visualizing, and Monitoring Disease Counts," *BMC Medical Informatics and Decision Making*, 9(21), 2009.

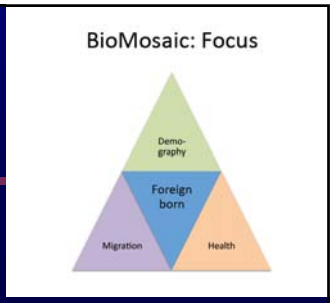
NBIC BCOP & BCON

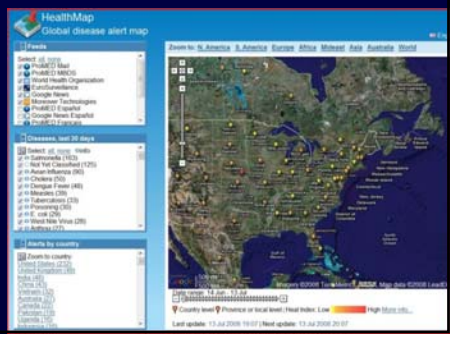
- Surveillance and information sharing



BioMosaic

- Healthmap incorporates infectious disease events reported by:
  - WHO, ProMed, Geosentinel, newspapers, OIE, FAO, WDIN, users, …
- BioDiaspora
  - International air transport association data
  - Use and distribution limited by license



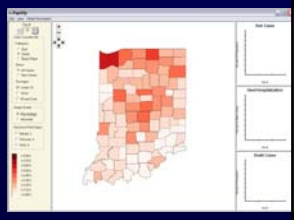


VACCINE

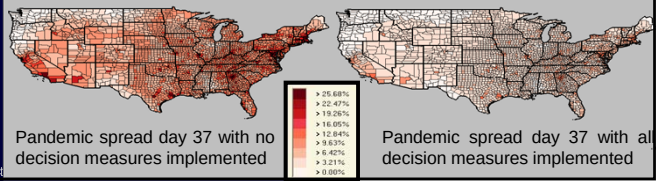
October 2013

Modeling a Pandemic

- Pandemic Influenza Planning Tool
- Models user specified:
  - Pandemic influenza characteristics
  - County population, demographics, hospital beds
- Decision measures
  - Strategic National Stockpile deployment
  - School Closures
  - Media Alerts



Map of the United States showing the estimated percent of the population ill based on a simulated pandemic influenza model originating in Chicago, IL.

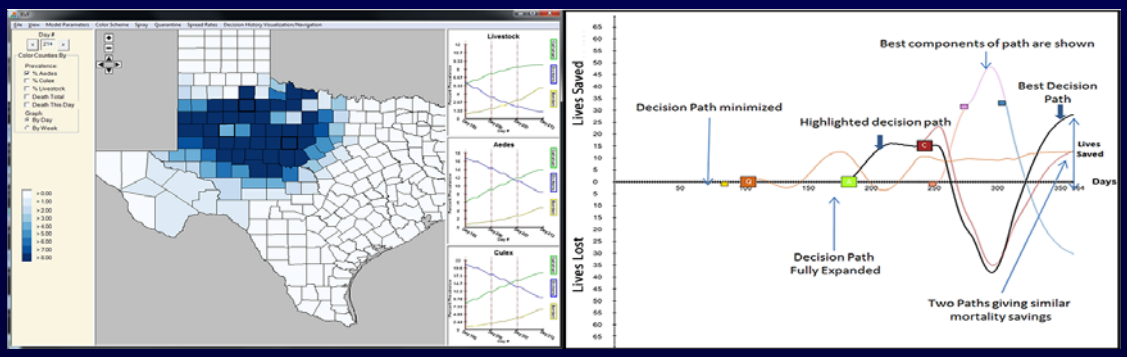


Pandemic spread day 37 with no decision measures implemented

Pandemic spread day 37 with all decision measures implemented

VACCINE

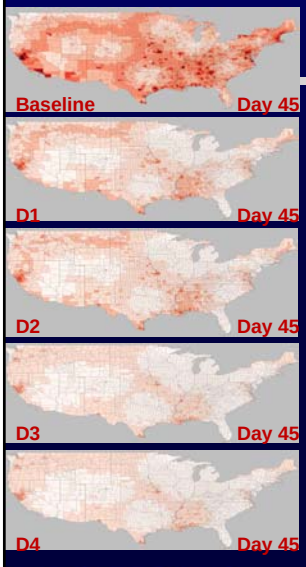
Advanced Decision Support Tools: Rift Valley Fever

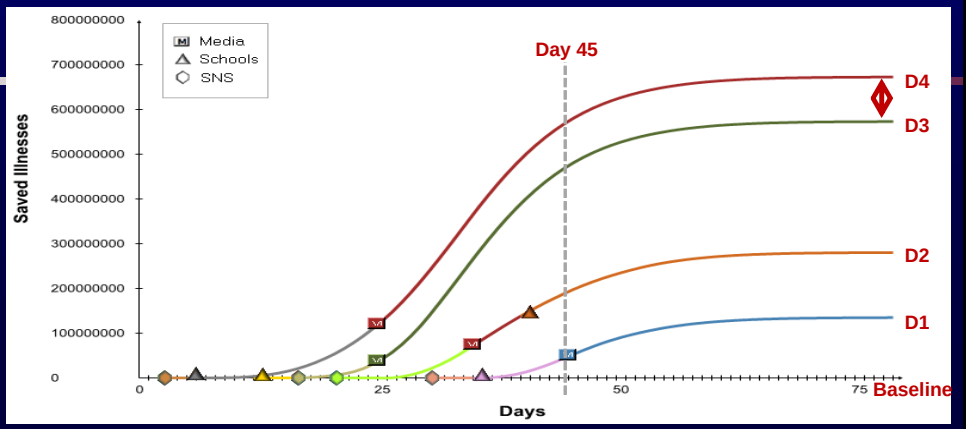


VACCINE

October 2013

Case Study – Pandemic Influenza





Decision Path D4

SNS Day 2 **School Closure Day 6** **Media Day 25**

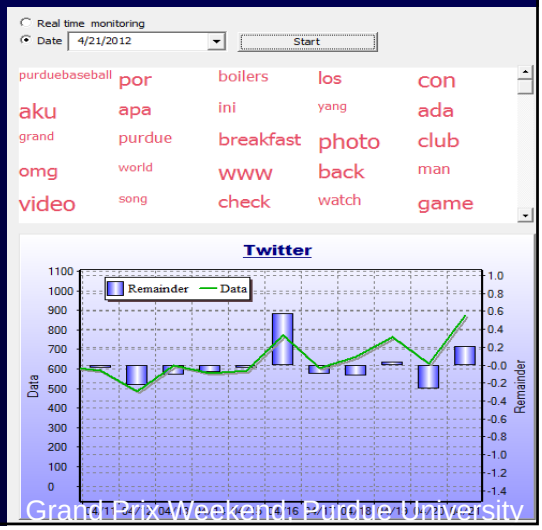
VACCINE

October 2013

Social Media: Real-time Twitter Monitoring and Integration into Tools

(Purdue, Stuttgart, Penn St.)

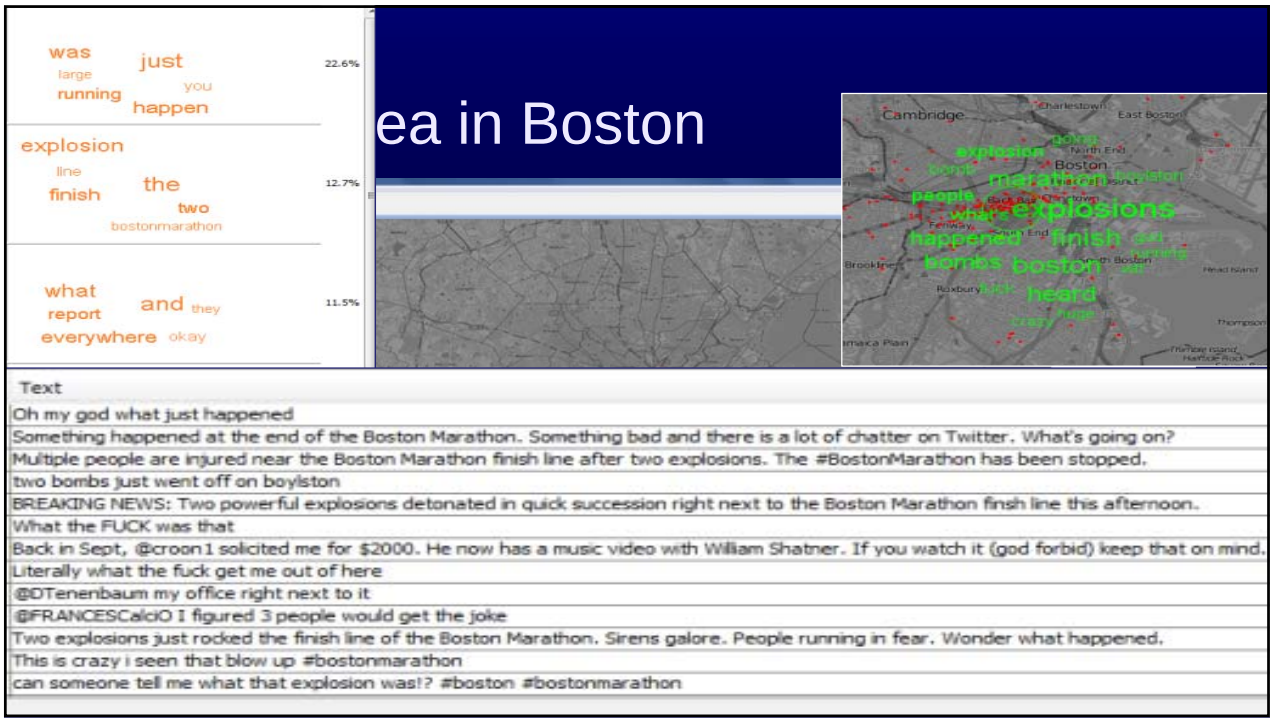
- Anomalous topic extraction using LDA and novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud



VACCINE

October 2013

Grand Prix Weekend, Purdue University



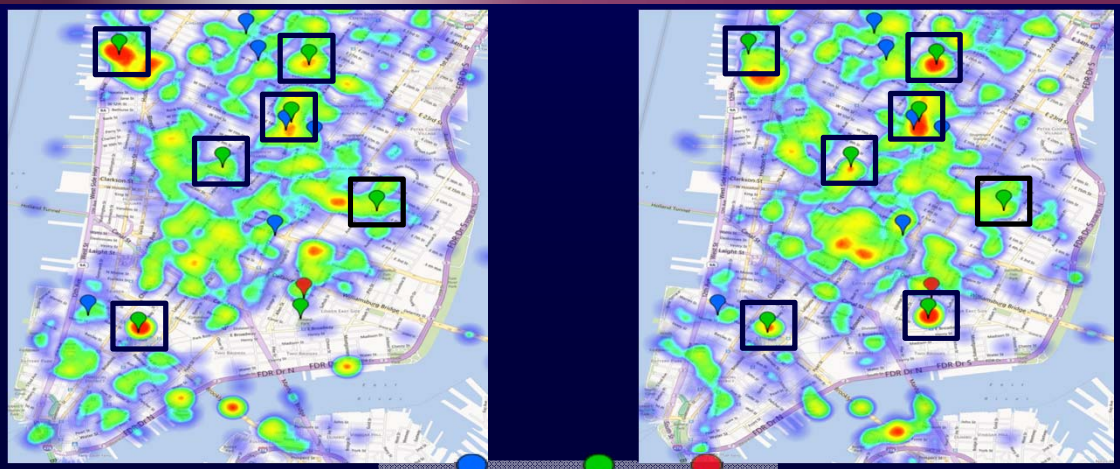
First Response (Tweet & Picture)



Visual Analytics of Activity During Hurricane Sandy

Two weeks before Sandy
10/14 (Sunday), 12:00 ~ 16:00

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00



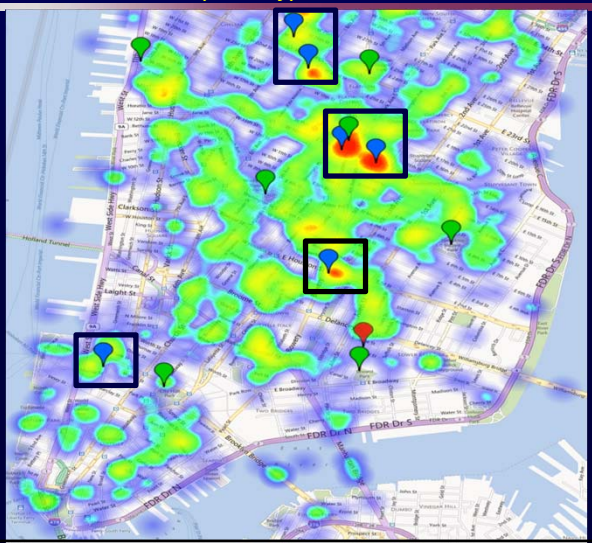
Supermarket Park Shelter

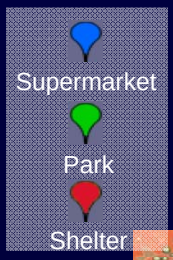
Visual Analytics of Activity During Hurricane Sandy

After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00

Evacuation order:
10/28, 10:30 AM

Hurricane Sandy's
Arrival at NYC:
10/29, 8:00 PM



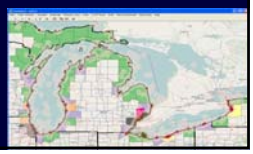


VACCINE

Visual Analytics Uses for Public Safety



- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Interactive, multi-day, month, type investigation
- Multisource, multimedia data integration & analysis



VACCINE

October 2013

Epidemiological Criminology (Akers et al.)

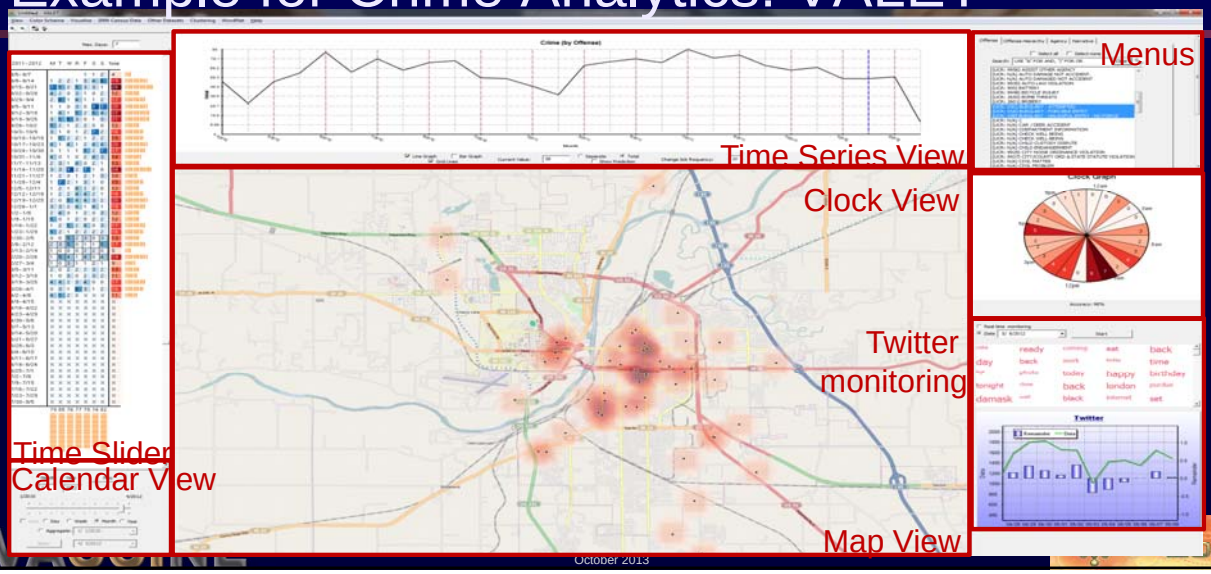
- Study criminal behavior using public health approach (epidemiology)





VACCINE

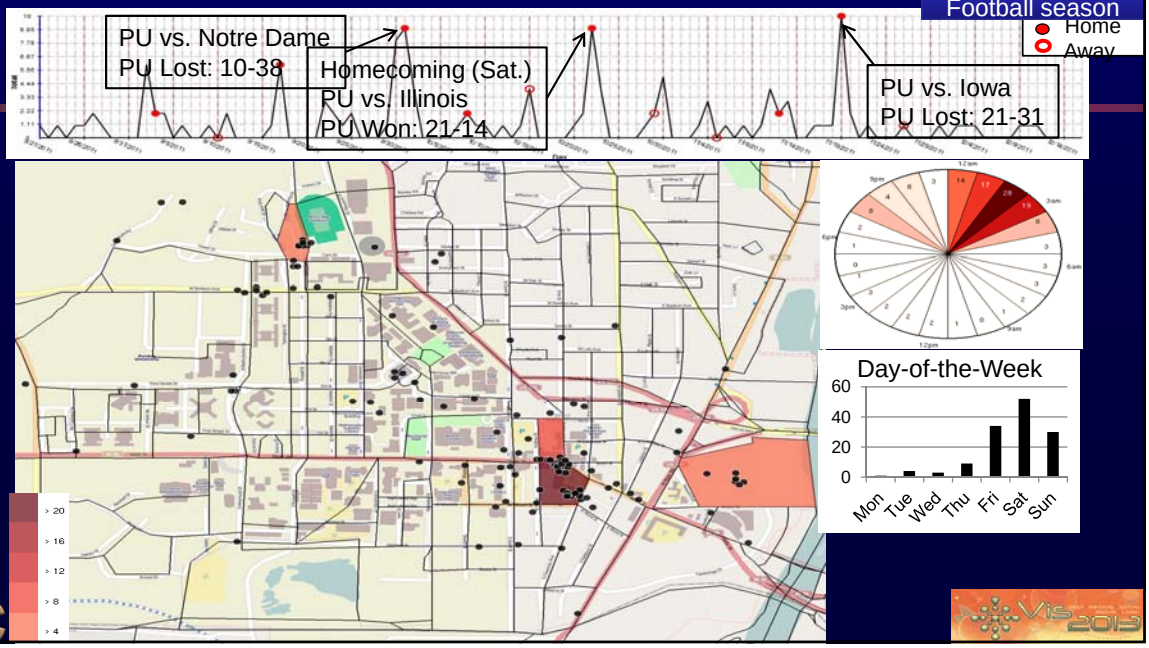
Example for Crime Analytics: VALET



VACCINE

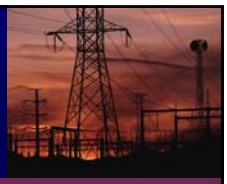
October 2013

Wicked Problem: High-Risk Alcohol Behavior: Drunkenness /Public Intoxication



VAC

Wicked Problem:
Cascading Critical Infrastructure Resiliency Modeling and Analytics (VASA)



- **Purpose:** Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.
- **Collaborating Institution(s):**
  Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities
- **End-User(s):** Power Suppliers (e.g., Duke Energy), Cyber Community (e.g., Cisco), Quick Service Restaurants and suppliers, food supply

VACCINE

October 2013

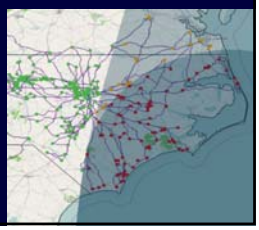
VASA: Visual Analytics for Security Applications

Ebert, Elmqvist, Gaither, Kennedy, Ribarsky, Wang
Collaborating Institution(s): Purdue, Minnesota, UTexas, UNCC + German universities
End-User(s): Fast-food restaurant chain, emergency management and planning personnel

Impacts and Accomplishments:



- Support decision-making for extreme weather and natural disaster scenarios
  - Combine real and simulation data
  - Allow "what-if" exploration
- **System of systems**: binds together multiple simulations models from collaborators into coherent whole
  - **Minnesota**: food distribution model
  - **Texas**: simulated and historical weather (hurricanes, storms)
  - **UNCC**: critical infrastructure
  - **Purdue**: roads + interaction visual analytics tool
- **Challenge**: combine interactive VA with complex simulation models



VACCINE

October 2013

Directions

- We still have a long way to go!
- Still a chasm between main visualization research community and other domains – but we are dramatically narrowing this.
- More acceptance of cross-disciplinary research and application results
- Still difficult for visualization researchers to publish application results

VACCINE

October 2013

Directions: Data

Huge amount of data available electronically

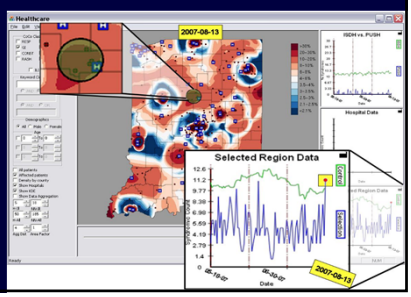
Challenges:

- Policies
- Privacy
- Security
- Integration to useful knowledge (smart aggregation)

Privacy

- Visualization can provide data privacy while still enabling understanding
- Aggregation and visualization are key assets in privacy preserving data sharing
- Normally only need to examine PII of a few affected individuals in taking action

Can you tell if I am sick?



Directions – We Still Need:

- Recognition of value of integrated visualization and analysis into science, health, and engineering research, policy, and decision making
  - Collaboration, integration, and interdisciplinary work
- Funding to support this integrated approach
- Research on key fundamental visualization and analytic problems (e.g., uncertainty, time, integrated analytics, science of interaction for decision making, evaluation)

VACCINE

For Further Information

www.VisualAnalytics-CCI.org



vaccine@purdue.edu
ebertd@purdue.edu

VACCINE

CVADA
Center for Visual Analytics and Data Analytics
A U.S. Department of Homeland Security Center of Excellence

Visual Analytics Solutions for Data-Driven Decision Making

David Ebert
Purdue University

VACCINE
Visual Analytics for Command, Control, and Interoperability Environments
A U.S. Department of Homeland Security Center of Excellence

July 2014

Who We Are:
International Team of Experts
75+ Faculty, 26 Institutions



- Purdue University
- Georgia Institute of Technology
- Pennsylvania State University
- Stanford University
- University of North Carolina at Charlotte
- University of Washington
- Arizona State University
- Simon Fraser University
- University of British Columbia
- Justice Institute of British Columbia
- Ontario Institute of Technology
- Dalhousie University
- University of Houston, Downtown
- Virginia Tech
- Indiana University
- Florida International University
- University of Texas at Austin
- Morgan State University
- Navajo Technical College
- University of Stuttgart
- University of Swansea
- Oxford University
- University of Calgary
- University of Manitoba
- Carleton University
- University of Victoria

VACCINE

July 2014

Some of Our Engaged End-Users

- Federal Operating Components:
  - US Coast Guard
  - US Transportation Security Agency
  - US Citizenship and Immigration Service
  - US Federal Emergency Management Agency
  - US Customs and Border Protection
  - US CERT
- Law Enforcement
  - Over 40 local and state agencies (IN, IL, OH, PA, NC, NY)
- State Fusion Centers
  - Ohio (SAIC), Indiana (IIFC)
- Companies
  - Motorola, Kimberly Clark, Banfield, Bank of America





VACCINE

July 2014

VACCINE: What We Do – Visual Analytic Solutions

We enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable user to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

VACCINE

July 2014

Why VA For Big Data?

(Or What Big Data Analytics Can't Do)

(inspired and adapted from David Brooks, New York Times, 2/18/2013)

- Qualitative, fuzzy, and social data
  - Preferences, significance of one relationship over another; Trust
- Context
  - Data is rarely complete, nor incorporates all relevant information Humans have extensive information and experience that never make it into the collected data
- Spurious vs. Significant
  - Big data means more statistically significant events and correlations, but they may not have any relevance
  - Increases noise to signal ratio
- Big problems
  - Complex, multifaceted, multiparameter big challenges with unquantified dependencies

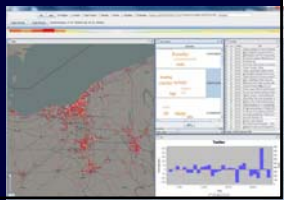
VACCINE

July 2014

Visual Analytics Uses for Public Safety



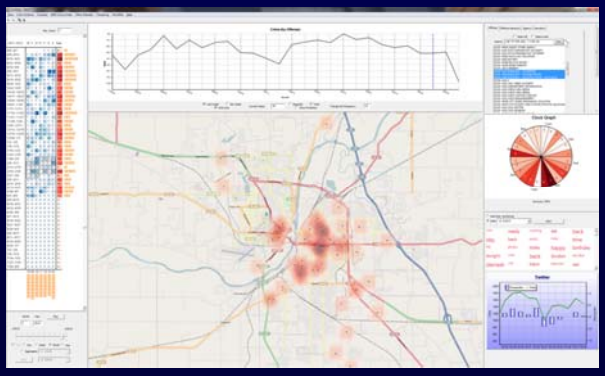
- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Data-driven policy analysis
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Multisource, multimedia massive data integration & analysis
- Purpose: Planning for resiliency, training, detection, investigation, response, recovery, remediation

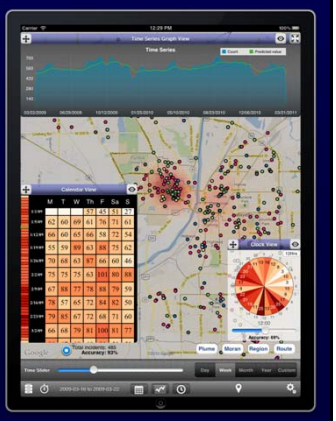


VACCINE

July 2014

Visual Analytics Law Enforcement Toolkit (i)VALET

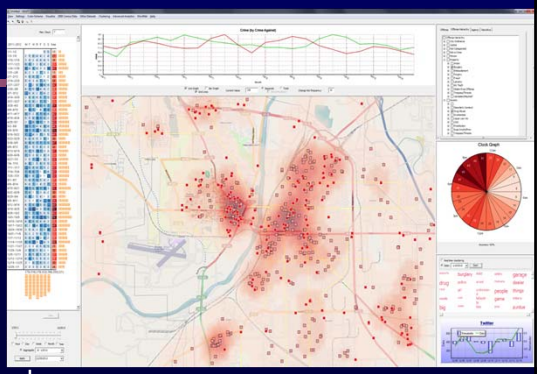




VACCINE

July 2014

Visual Analytics Law Enforcement Toolkit (VALET, iVALET)



Impacts:

- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public (February 2013) for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

VALET delivered:

- Spring 2011: WL, Lafayette Police

iVALET delivered:

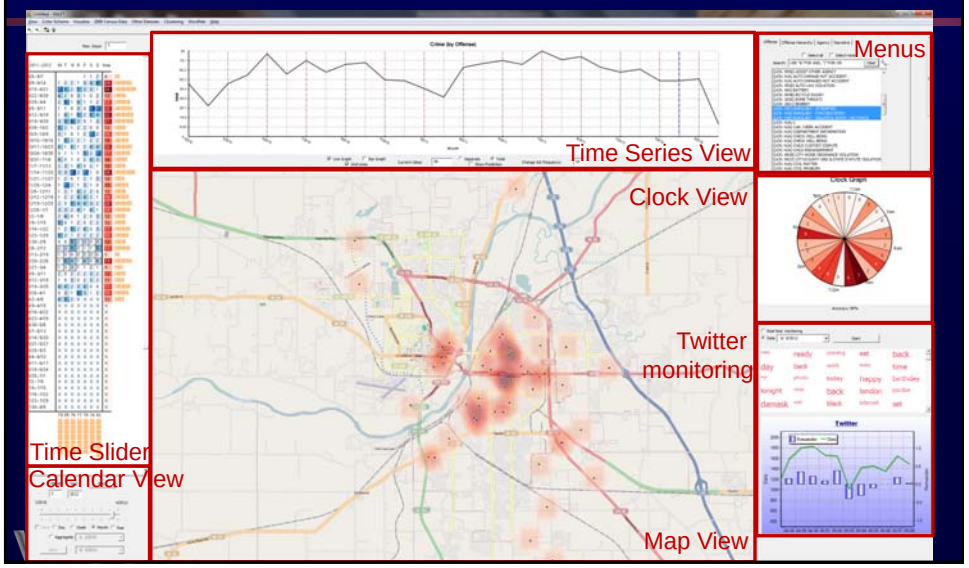
- October 2011: Purdue, WL Police



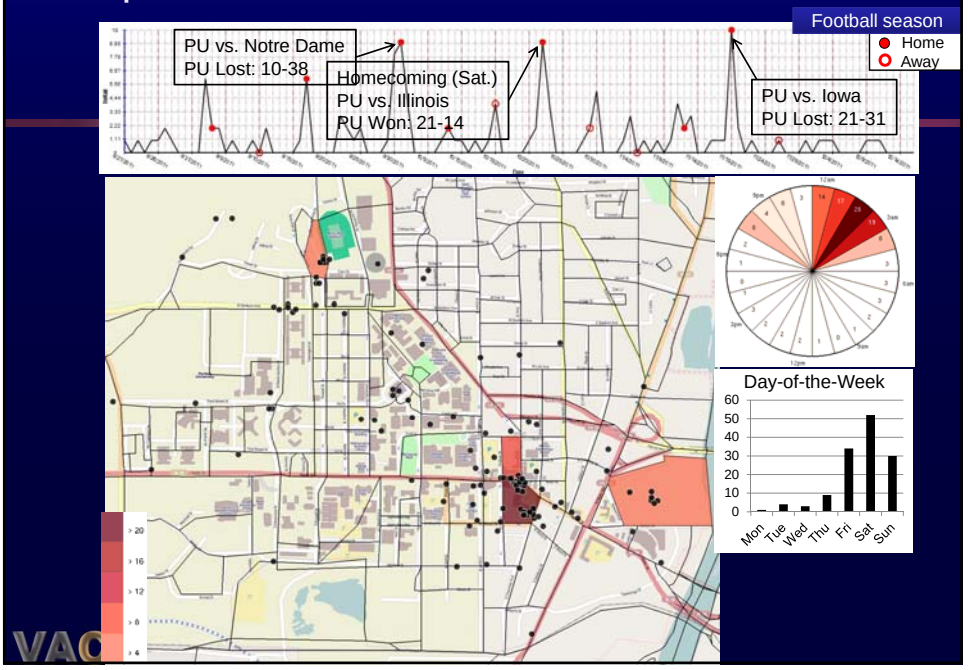
VACCINE

July 2014

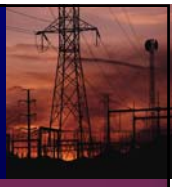
Example for Crime Analytics: VALET



Example: Drunkenness / Public Intoxication



Cascading Critical Infrastructure Resiliency Modeling and Analytics (VASA)



- Purpose: Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.
- Collaborating Institution(s): Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Texas (TACC), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities
- End-User(s): Power Suppliers (e.g., Duke Energy), Cyber Community (e.g., Cisco), Quick Service Restaurants and suppliers, food supply

Purdue

UNCC



U Minn

TACC

VACCINE

July 2014

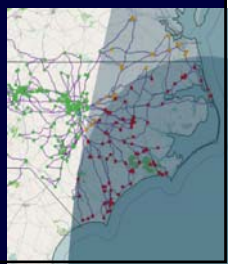
VASA: Visual Analytics for Simulation-Based Action

Collaborating Institution(s): Purdue, Minnesota, UTexas, UNCC + German universities
End-User(s): Fast-food restaurant chain, emergency management and planning personnel

Impacts and Accomplishments:



- Support decision-making for extreme weather and natural disaster scenarios
  - Combine real and simulation data
  - Allow "what-if" exploration
- **System of systems**: binds together multiple simulations models from collaborators into coherent whole
  - **Minnesota**: food distribution model
  - **Texas**: simulated and historical weather (hurricanes, storms)
  - **UNCC**: critical infrastructure
  - **Purdue**: roads + interaction visual analytics tool
- **Challenge**: combine interactive VA with complex simulation models

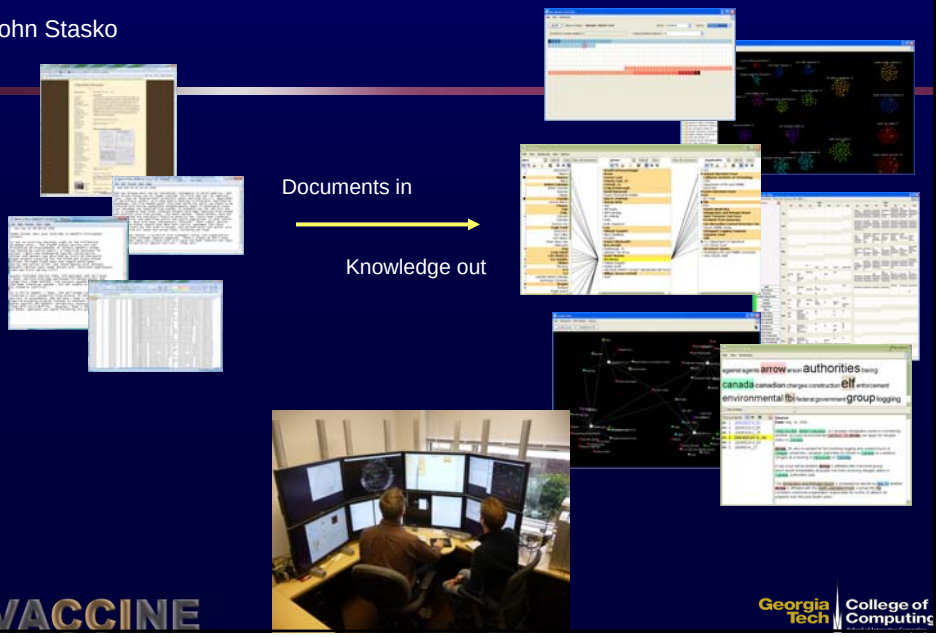


VACCINE

July 2014

Jigsaw: Visual Analytics for Investigative Analysis

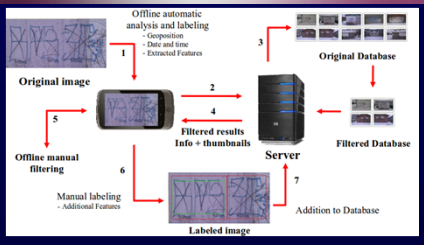
John Stasko



VACCINE

Georgia Tech College of Computing

Gang Graffiti Analysis and Recognition Using a Mobile Telephone (GARI)





IMPACT:

- Allows police to catalog and analyze gang graffiti images into a database system to better track and determine gang activity throughout a region
- Will allow the graffiti images to be "interpreted"
- More than 140 users and 1300 graffiti images acquired

GARI delivered:

- Summer/Fall 2011:
  - IPD gang detectives
  - IPD at large
  - Ind Fusion Gang Task Force
  - Gang detectives across Indiana
- Spring/Summer 2013:
  - Cook County, IL
  - INGang ownership
  - Tattoo database added

VACCINE

July 2014

Effective Risk-based Decision Making and Resource Allocation Visual Analytics

- Evaluate current and historical mission area:
  - Demands
  - Risks (total, mitigated, residual)
  - Resource allocation
  - Return on investment
- Evaluate courses of action
- Evaluate above at both Strategic and Tactical/Operational level

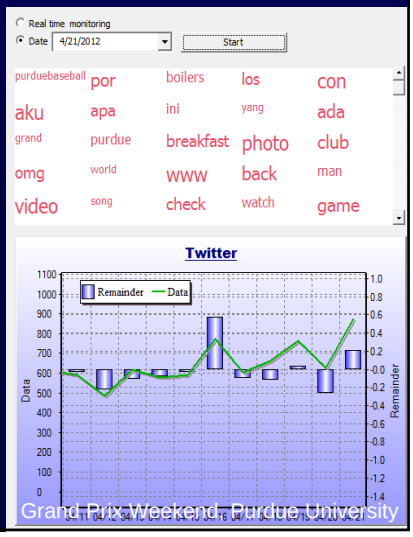
VACCINE

July 2014

Social Media: Real-time Twitter Visual Analytics

(Purdue, Stuttgart, Penn St.)

- Anomalous topic extraction using LDA and novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud

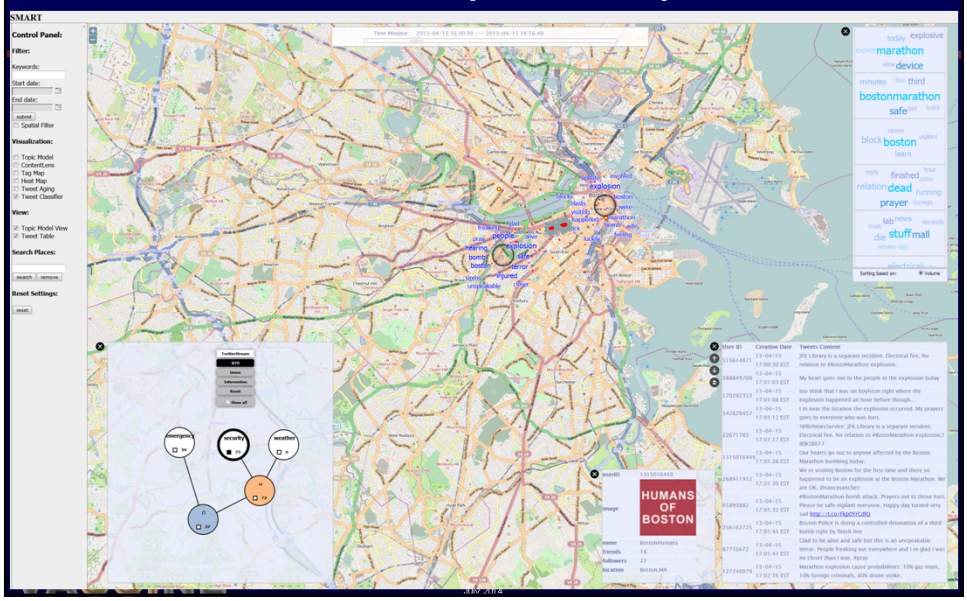


VACCINE

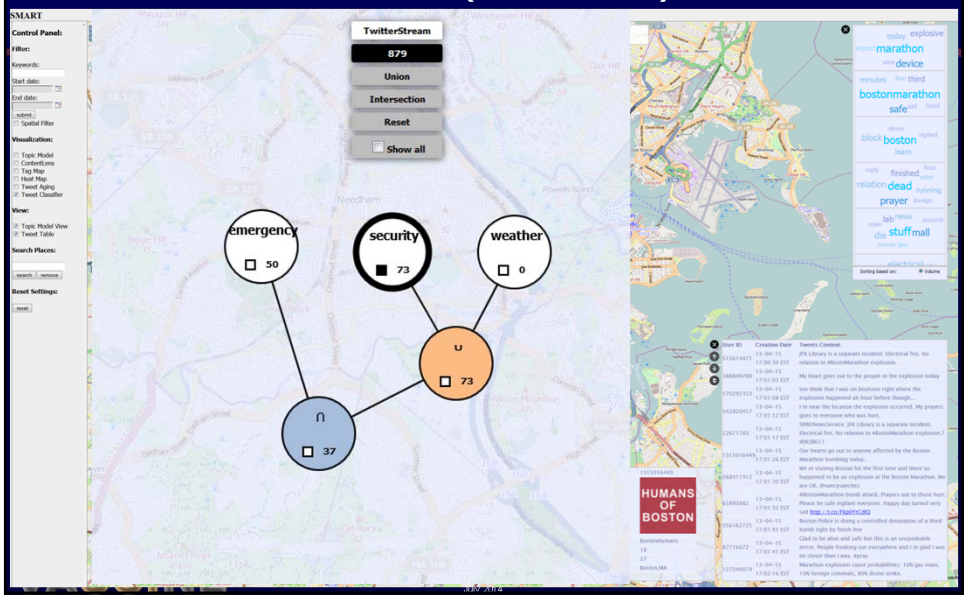
July 2014

Grand Prix Weekend - Purdue University

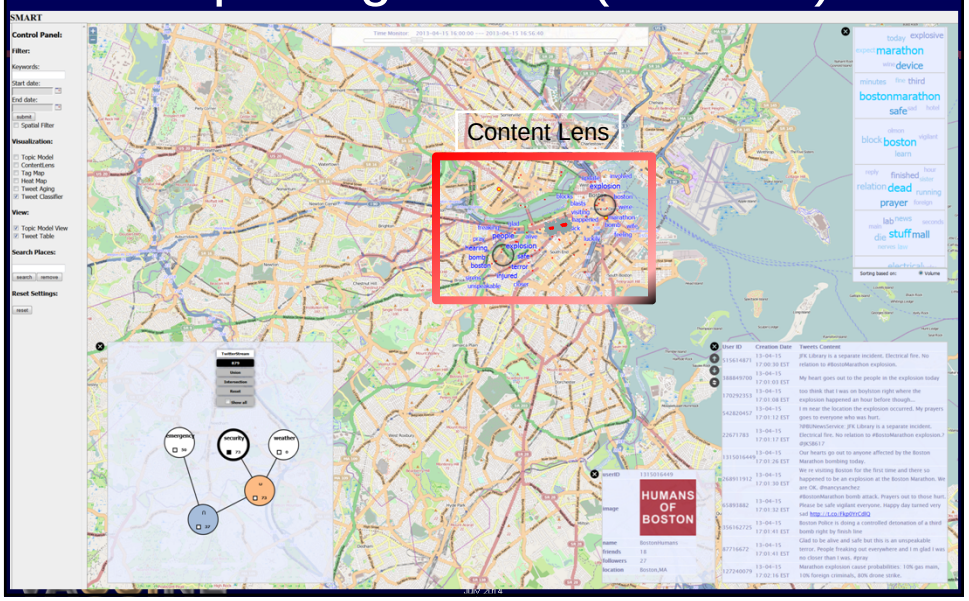
Social Media Analysis and Reporting Toolkit (SMART)



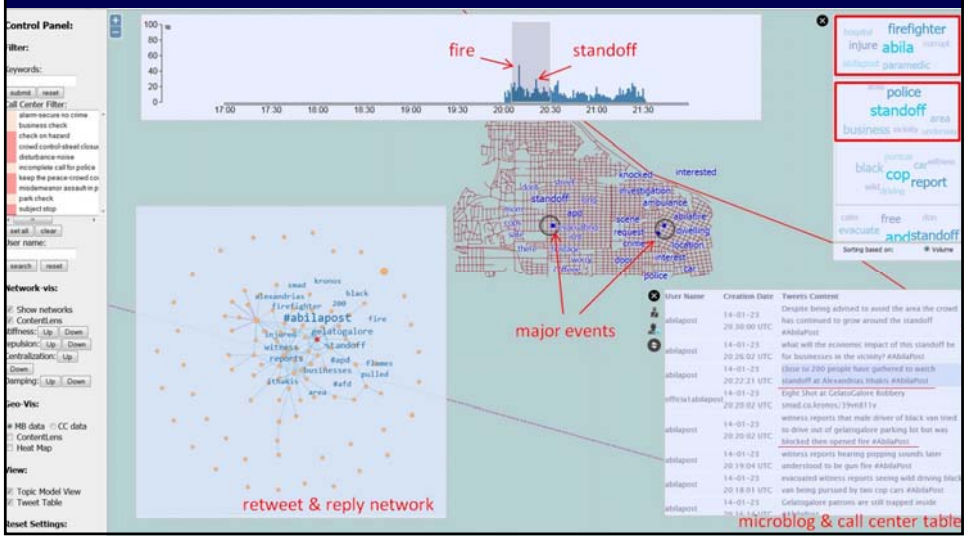
Social Media Analysis and Reporting Toolkit (SMART)



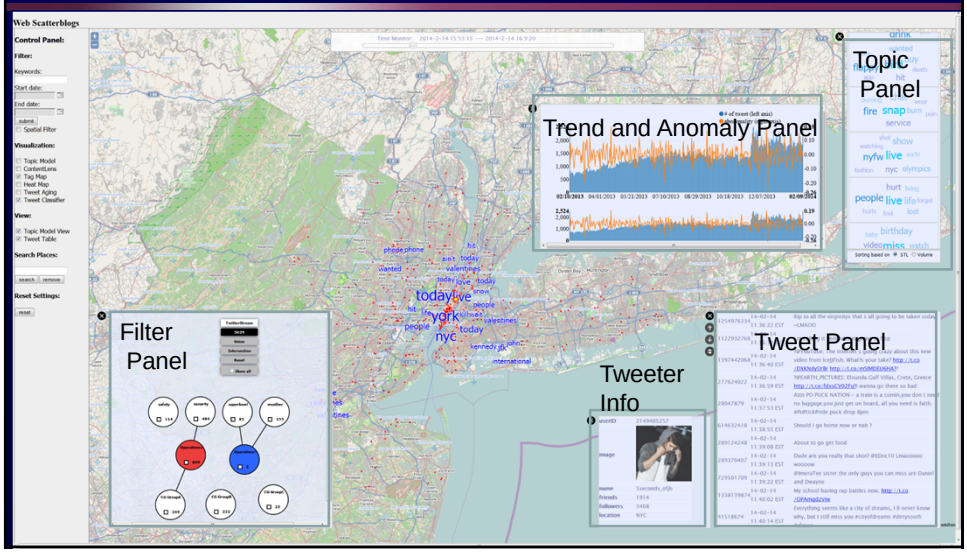
Social Media Analysis and Reporting Toolkit (SMART)



SMART: Network Linkage



SMART Anomaly Detection



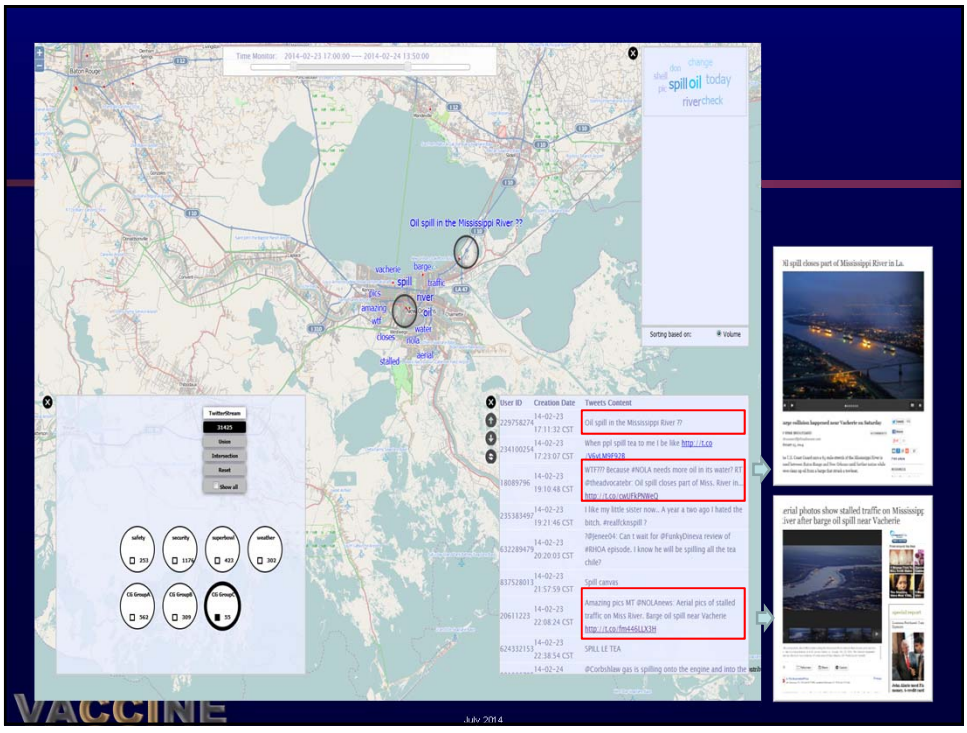
Anomaly Detection



Oil Spill in Mississippi River

2014/02/23

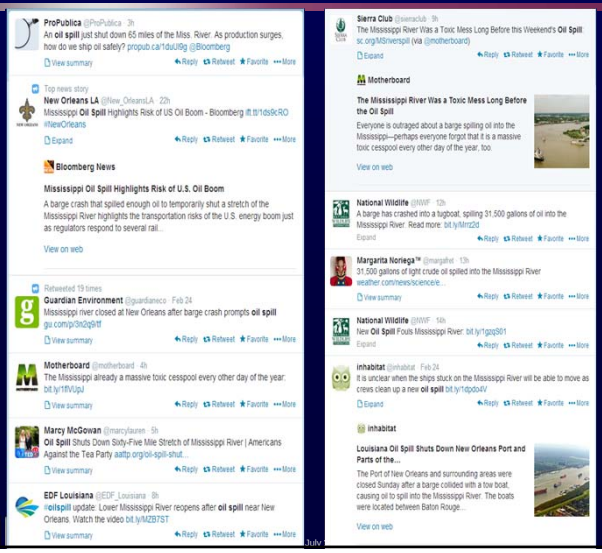
July 2014



VACCINE

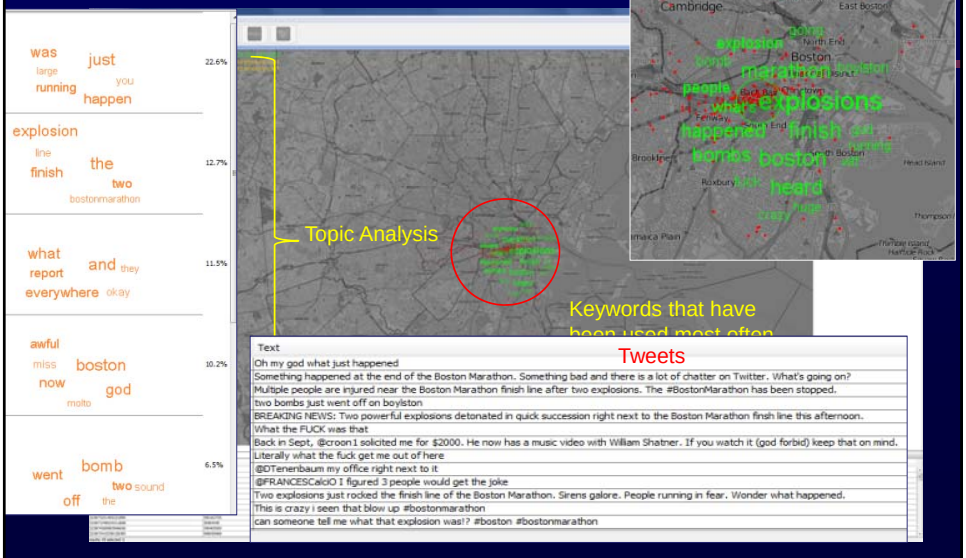
July 2014

Non-geo Tweets On Twitter About Oil Spill Event



VACCINE

Explosion Area in Boston



VACCINE

July 2014

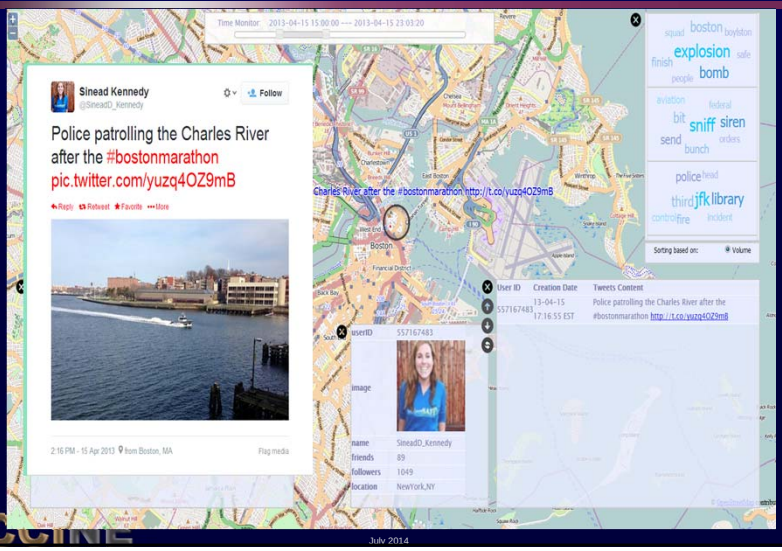
First Response (Tweet & Picture)



VACCINE

July 2014

Police Patrolling The River



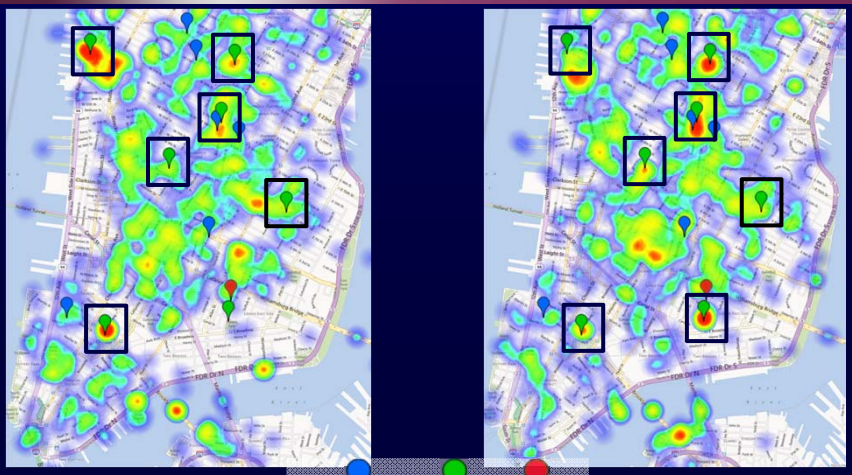
VACCINE

July 2014

Visual Analytics of Activity During Hurricane Sandy

Two weeks before Sandy
10/14 (Sunday), 12:00 ~ 16:00

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00



Supermarket Park Shelter

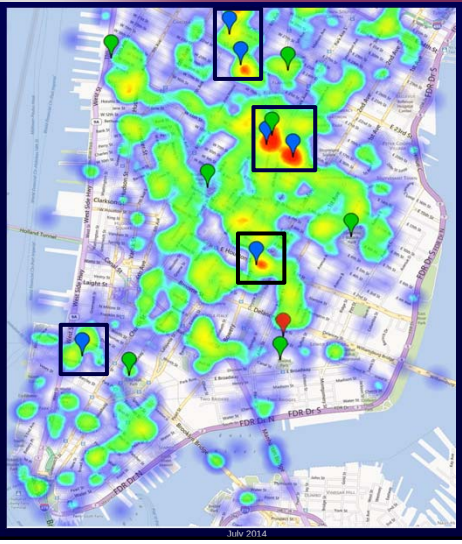
VACCINE

Visual Analytics of Activity During Hurricane Sandy

After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00

Evacuation order:
10/28, 10:30 AM

Hurricane Sandy's
Arrival at NYC:
10/29, 8:00 PM



Supermarket
Park
Shelter

VACCINE

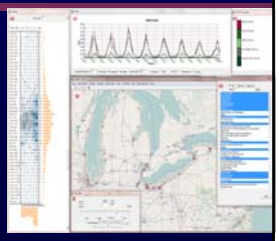
U.S. Coast Guard Search and Rescue VA

(cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771

IMPACTS:

- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Hurricanes Sandy and Irene resource allocation decisions based on cgSARVA analysis and visualization
- Informed Commandant's budget testimony to Congress
- Key component of USCG D9 reallocation plan for 2011-12
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ



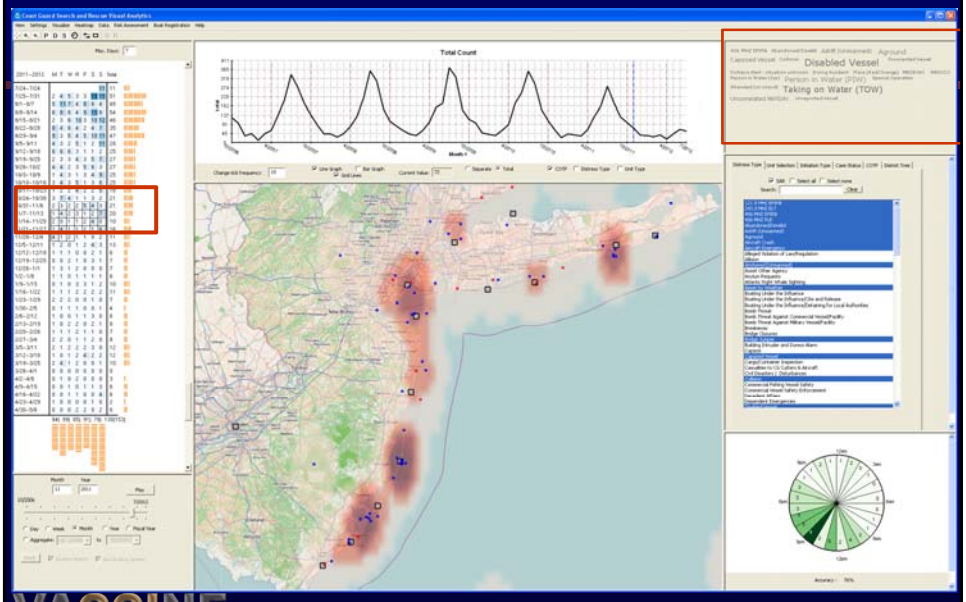


VACCINE

July 2014

Example: Risks and Consequences From Sandy:

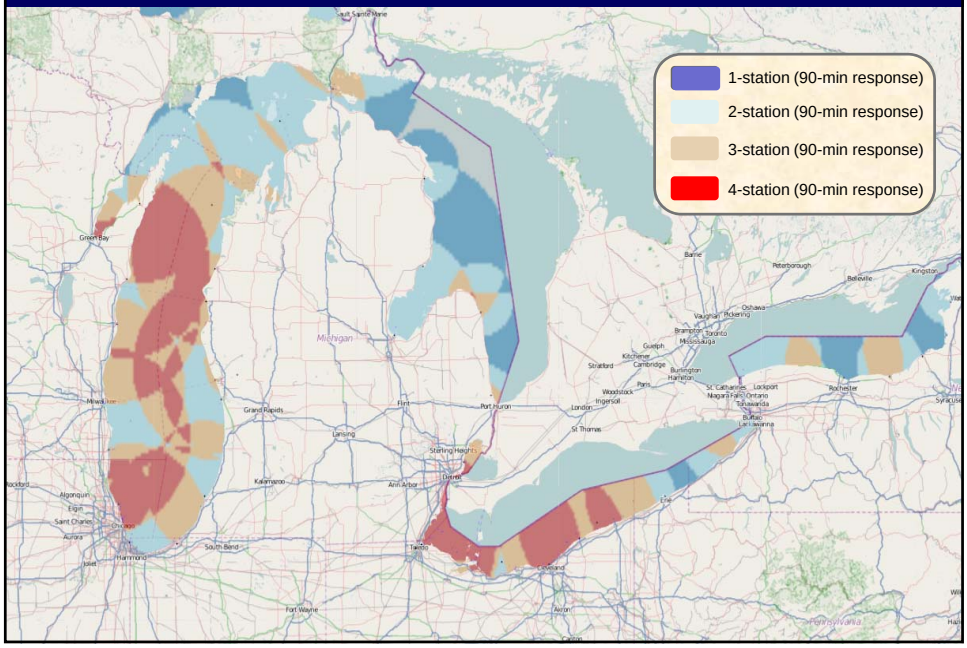
SAR Cases November 2011 NJ/NYC Area



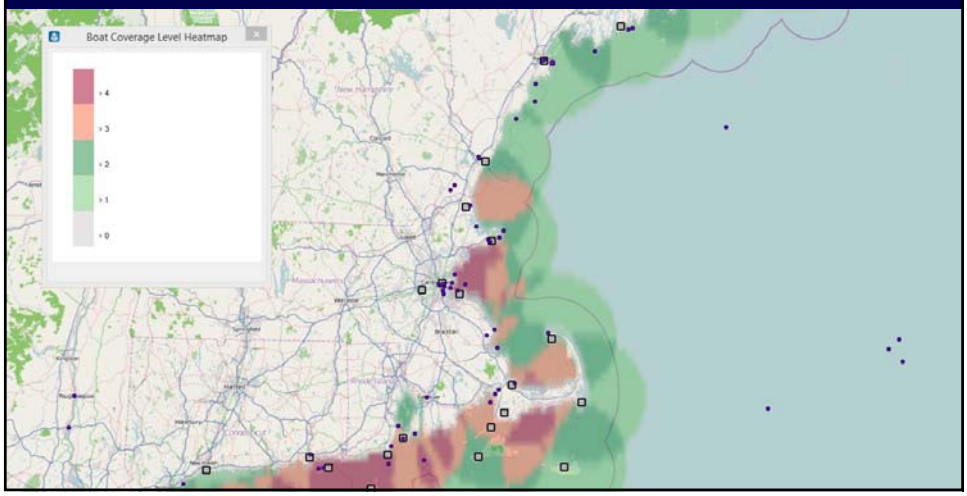
VACCINE

July 2014

Response Efficiency – Potential Future Assets



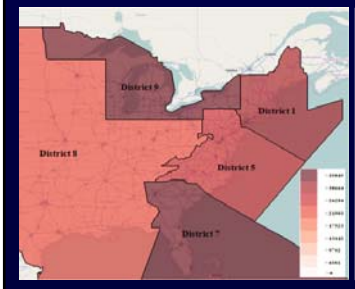
Coverage vs Lives Lost Aggregated Since Fiscal Year 2007



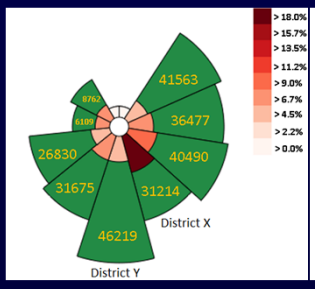
ORAM Visualization

Component Overview: Identification of Potential Risks

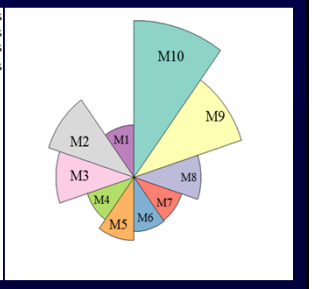
General understanding on how the risk is distributed between the different areas and missions.



(A) District Choropleth Map for Total Risk across all missions



(B) Risk profiles across all districts and mission areas



(C) District X Risk Profiles across all mission area

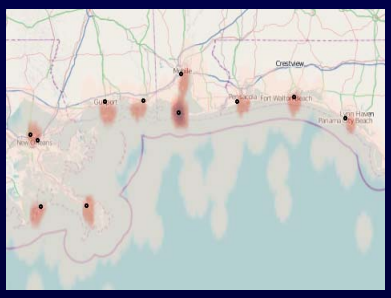
VACCINE

Component Overview: Analysis and Assessment

Specific areas can be analyzed by time and details on demand.



(D) Total M10 Risk Values distributed among District X's station AOR



(E) KDE heatmap showing Total M10 Risk values

VACCINE

July 2014

Interactive Operational Performance Assessment Report (iOPAR)

- OPAR
  - Standard report within the Coast Guard Business Intelligence (CGBI) system
  - Displays resource use and performance by core CG mission areas
- iOPAR – interactive iPad version to allow interactive visualization and analysis and inform decision making
  - Anaylze/visualize performance, targets, seasonal trends, predictions by boat, aircraft, cutter type, region levels

VACCINE

July 2014

USCG Port Closure Economic Impact VA

Partners: USC CREATE, USCG RDC, USCG D7, USCG LANT

IMPACT:

- Provided tool for use analysis and planning for impact of port closure in Port Arthur, Tx
- Economic sector impact, local and national impact
- Impact and effectiveness of alternative mitigation strategies



VACCINE

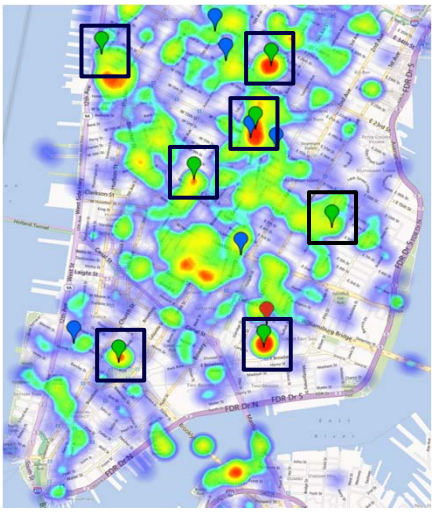
For Further Information

www.VisualAnalytics-CCI.org

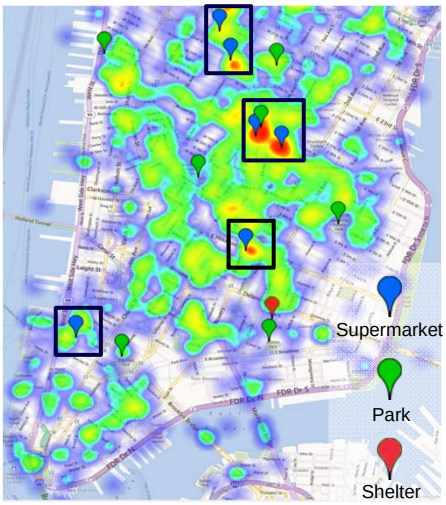


VAC

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00



After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00



VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

Visual Analytics: The Right Solution for Big Data and Global Challenges

David S. Ebert

Purdue University

ebertd@purdue.edu

CVADA

Center for Visual Analytics and Data Analytics

A U.S. Department of Homeland Security Center of Excellence

March 2014

Overview

- Big Data challenges
- Visual analytics
- VACCINE center
- Research area overview and example projects
- Path forward



Big Data: Solution to Global Challenges?



- **Big Data Definition:** A phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex and diverse types of data. (TechAmerica Foundation)
  Big Data is often defined along three dimensions:
  - **Volume** – size
  - **Velocity** – rate of input, update, change
  - **Variety** – different types, sources, variables
- Need:
  - Advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of information (TechAmerica Foundation)
  - **Enable effective, efficient analysis, decision making, planning, and action**

Big Data Challenges

- Heterogeneity
- Unprecedented volume and/or complexity
- Multivariate, Multidimensional
- Streaming
- Qualitative, Quantitative, Categorical, Temporal,…
- Distributed Sources: sensors, text, images, maps, …
- Inconsistent and Uncertain
- Different Scales and Cross Scale

Big Data Analytics – The Solution?

- Advanced data analytics for big data can help manage big data
  - Machine learning
  - Advanced statistical analysis
  - Hadoop-deployed analytics
- However, it falls short in several aspects

Why VA or What Big Data Analytics Can't Do

(inspired and adapted from David Brooks, New York Times, 2/18/2013)

- Qualitative, fuzzy, and social data
  - Preferences, significance of one relationship over another; Trust
- Context
  - Data is rarely complete, nor incorporates all relevant information
    Humans have extensive information and experience that never make it into the collected data
- Spurious vs. Significant
  - Big data means more statistically significant events and correlations, but they may not have any relevance
  - Increases noise to signal ratio
- Big problems
  - Complex, multifaceted, multiparameter big challenges with unquantified dependencies

Beyond Visualization to Visual Analytics

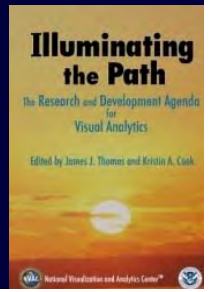
Visual Analytics[1] is the science of analytical reasoning facilitated by interactive visual interfaces

Interactive visualization, data analysis, exploration, and decision making with human in the loop!

People use visual analytics tools and techniques to

- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
- Detect the expected and discover the unexpected
- Provide timely, defensible, understandable assessments
- Communicate assessment effectively for action

1. *Illuminating the Path: The R&D Agenda for Visual Analytics, Editors: Thomas and Cook*



VACCINE

VA Approaches to Big Data

- Don’t display all the data!
- Extract relevant information at the appropriate natural scales using analytics and statistical models to reduce data space so that it is mappable to the visual space
- User interaction to guide above process
- Adapt to available hardware for the problem
  - Should we use big iron or cell phones?

Overview

- Big Data challenges
- Visual analytics
- VACCINE center
- Research area overview and example projects
- Path forward




Innovation with Impact
SFU, JIBC
UBC
UW
Purdue
VACCINE
Penn St.
Ind U
VaTech
NC A&T
UNCC
Stanford
Navajo Tech
JSU
GaTech
U Stuttgart
Swansea U
UT Austin
UHD
FIU
Selected Region Data
GeoDiscoverer

Who We Are:
International Team of Experts
75+ Faculty, 24 Institutions



- **Purdue University**
- **Georgia Institute of Technology**
- **Pennsylvania State University**
- **Stanford University**
- **University of North Carolina at Charlotte**
- **University of Washington**
- **Simon Fraser University**
- **University of British Columbia**
- **Ontario Institute of Technology**
- **Dalhousie University**
- **University of Manitoba**
- **Carleton University**
- **University of Houston, Downtown**
- **Virginia Tech**
- **Indiana University**
- **Florida International University**
- **University of Texas at Austin**
- **Morgan State University**
- **Arizona State University**
- **University of Stuttgart**
- **University of Swansea**
- **Oxford University**
- **University of Calgary**
- **University of Victoria**

VACCINE

Some of Our Engaged End-Users

- Federal Operating Components:
  - US Coast Guard
  - US Transportation Security Agency
  - US Citizenship and Immigration Service
  - US Federal Emergency Management Agency
  - US Customs and Border Protection
  - US CERT
- Law Enforcement
  - Over 40 local and state agencies (IN, IL, OH, PA, NC, NY)
- State Fusion Centers
  - Ohio (SAIC), Indiana (IIFC)
- Companies
  - Motorola, Kimberly Clark, Banfield, Bank of America





VACCINE: What We Do – Visual Analytic Solutions

We enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable user to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

Visual Analytic Solutions: What We Offer

- **Improved Effectiveness**: We enable users to be more effective through innovative, interactive visualization, analysis, and decision making tools
  - Provide the right information, in the right format, within the right time to solve the problem
  - Enable user to be more effective from planning to detection to response to recovery
  - Enable effective communication of information
- **Innovative Fielded Solutions:** We provide innovative visual analytic and scalable solutions to the extended homeland security community
- **People and Partnerships**
  - Interdisciplinary world-leading team of researchers and students

"cgSARVA has proven its worth time and again, providing key analytic information for decision makers for large scale projects…"

VADM Robert Parker, 2012 MRS Keynote Address

VADM Robert Parker with VACCINE student researchers (cgSARVA, COAST, iOPAR)



Our Research Themes To Solve Real World Problems

- **Visual Analytics** – integrated, interactive visual exploration, analysis, and decision making environments to enable effective decisions from massive, time-evolving multimedia data
- **Interactive scalable analytics techniques** – statistical, spatiotemporal, image, video, signal, machine learning
- **Science of Interaction for visual analytics**
- **Evaluation** in real world environments

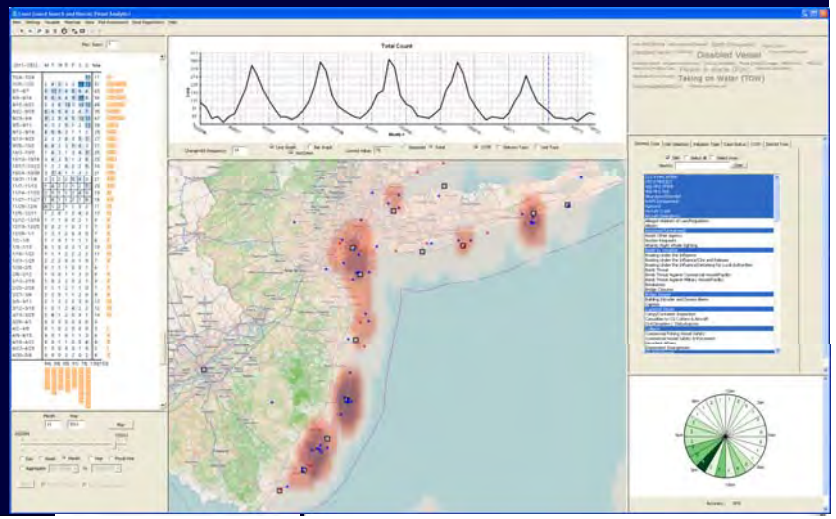
Overview

- Big Data challenges
- Visual analytics
- VACCINE center
- Research area overview and example projects
- Path forward



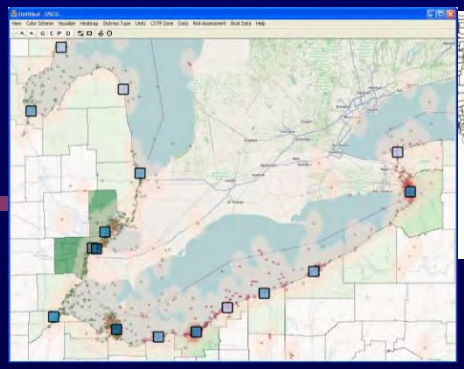
Research Area Overview and Example Projects

- Research applications areas
- Example projects
  - VASA
  - TSA
  - Public safety and law enforcement
  - Social media visual analytics
  - US Coast Guard
  - Additional projects

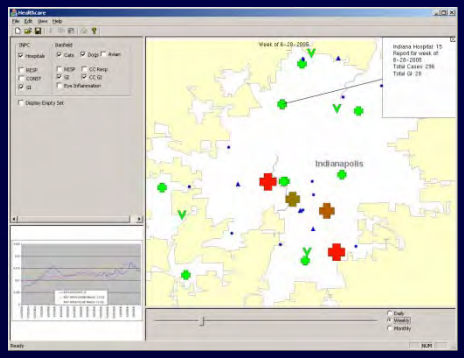


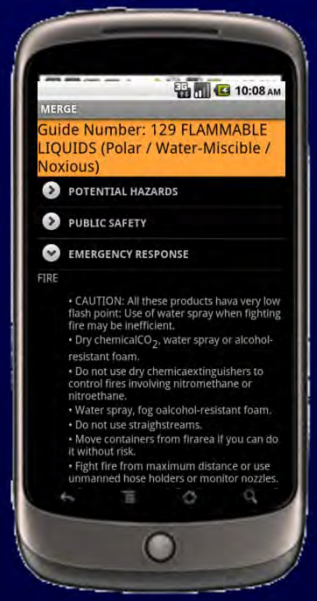
Public Safety & Health

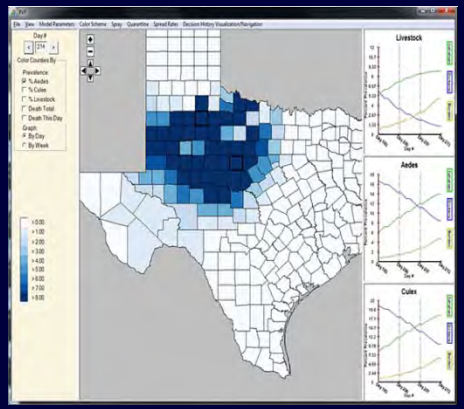
- Public safety visual analytics
  - cgSARVA
    - Coast Guard Search And Rescue Visual Analytics
  - MERGE
    - Mobile Emergency Response Guide
- Public health visual analytics
  - LAHVA
    - Linked Animal-Human Health Visual Analytics
  - RVF
    - Rift Valley Fever
    - Decision support environment for epidemic modeling and responses
  - PanVis
    - Pandemic influenza modeling and visualization tool
  - Cancer Care Engineering

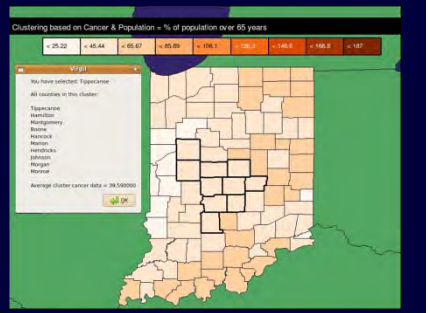










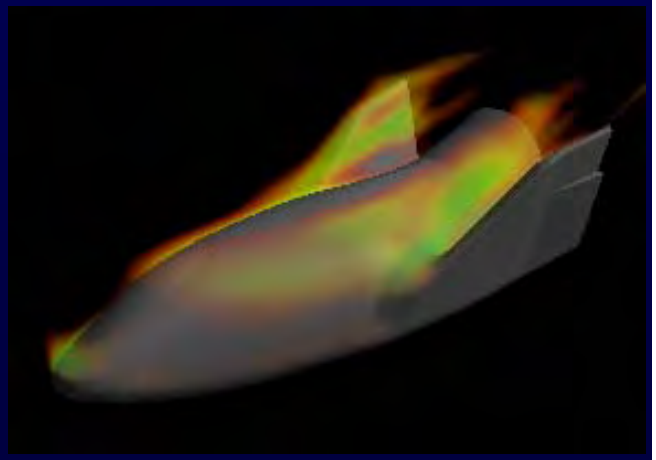


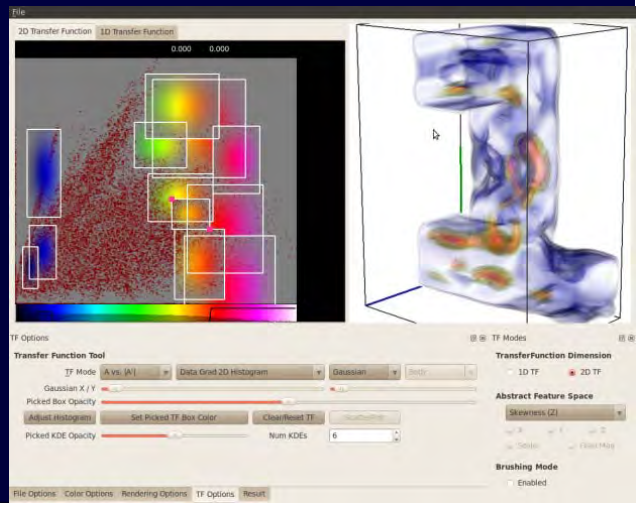
VACCINE

Scientific Discovery



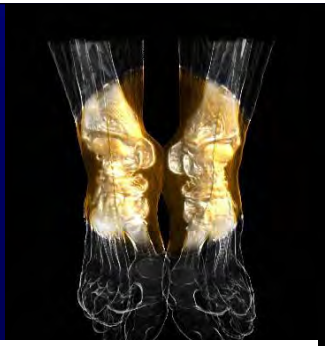
- Flow dynamics visualization
  - Providing insights on large flow data
  - Visualization linked with simulations
  - Innovative feature visualization
- Nanohub
  - Information-assisted data analysis and visualization of nanoelectronic models



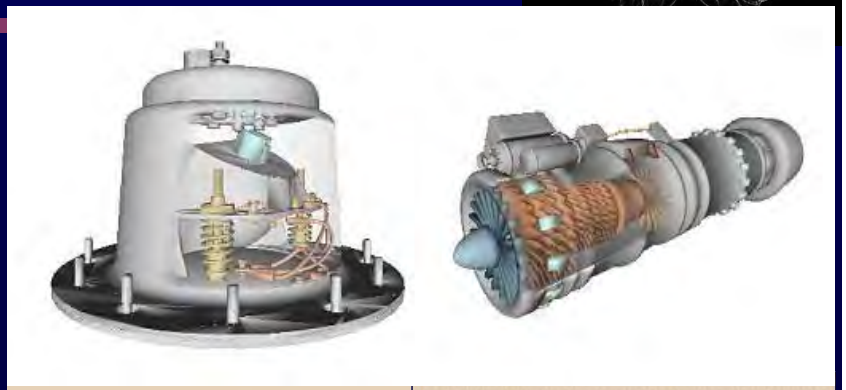


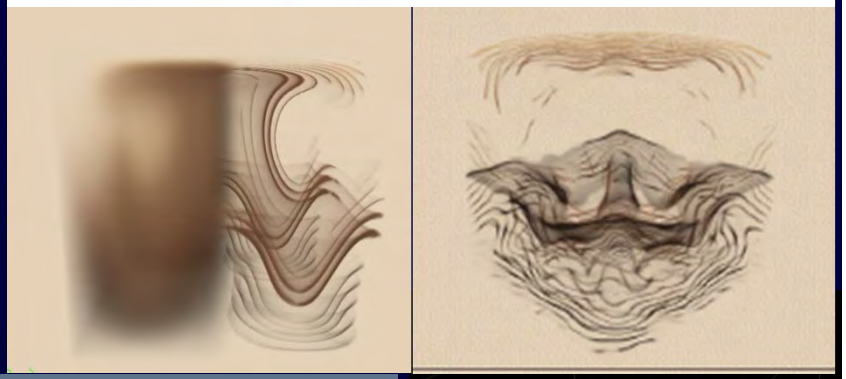
VACCINE

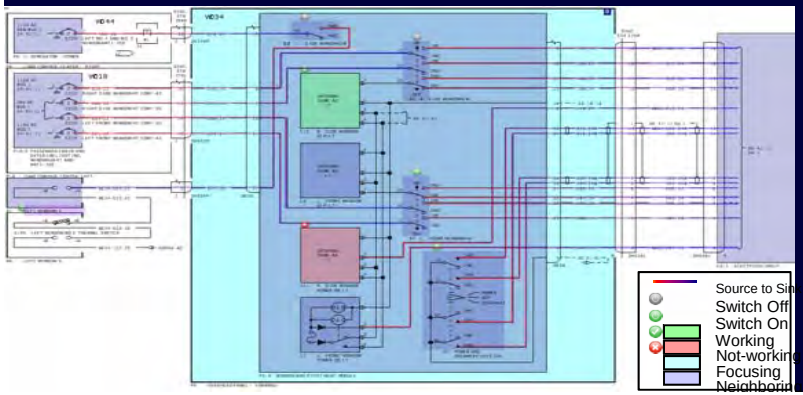
Scientific Discovery



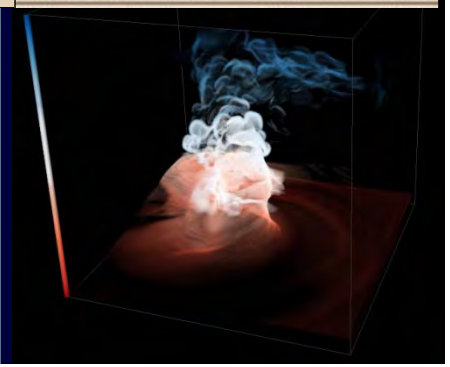
- Illustrative Visualization
  - Mechanical assemblies
  - Schematics and wiring diagrams
  - Illustrative flow visualization
  - Illustrative medical visualization
- Interactive multivariate atmospheric science data visual analanalytics







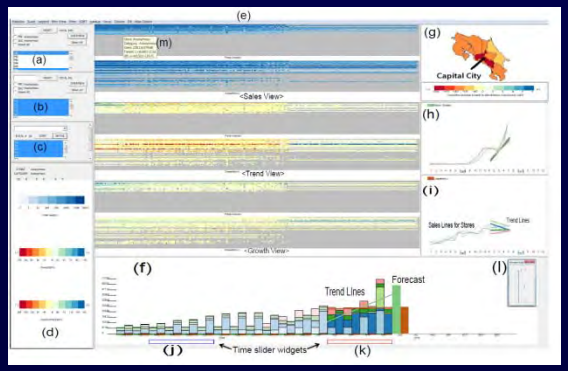




Business Visual Analytics

- Risk-based decision making and resource allocation
  - Coast guard operational risk assessment model
  - Helping to prioritize efforts to minimize risk
- Competitive Intelligence
  - Visual analytics system for business intelligence
- EconVIS
  - Visual analytics in various economic problems
  - Improving decision making and identifying key motivations in knowledge creation

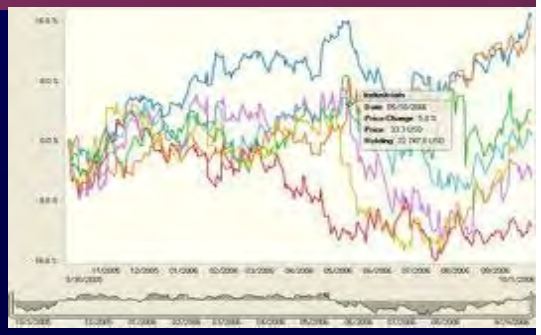






VACCINE

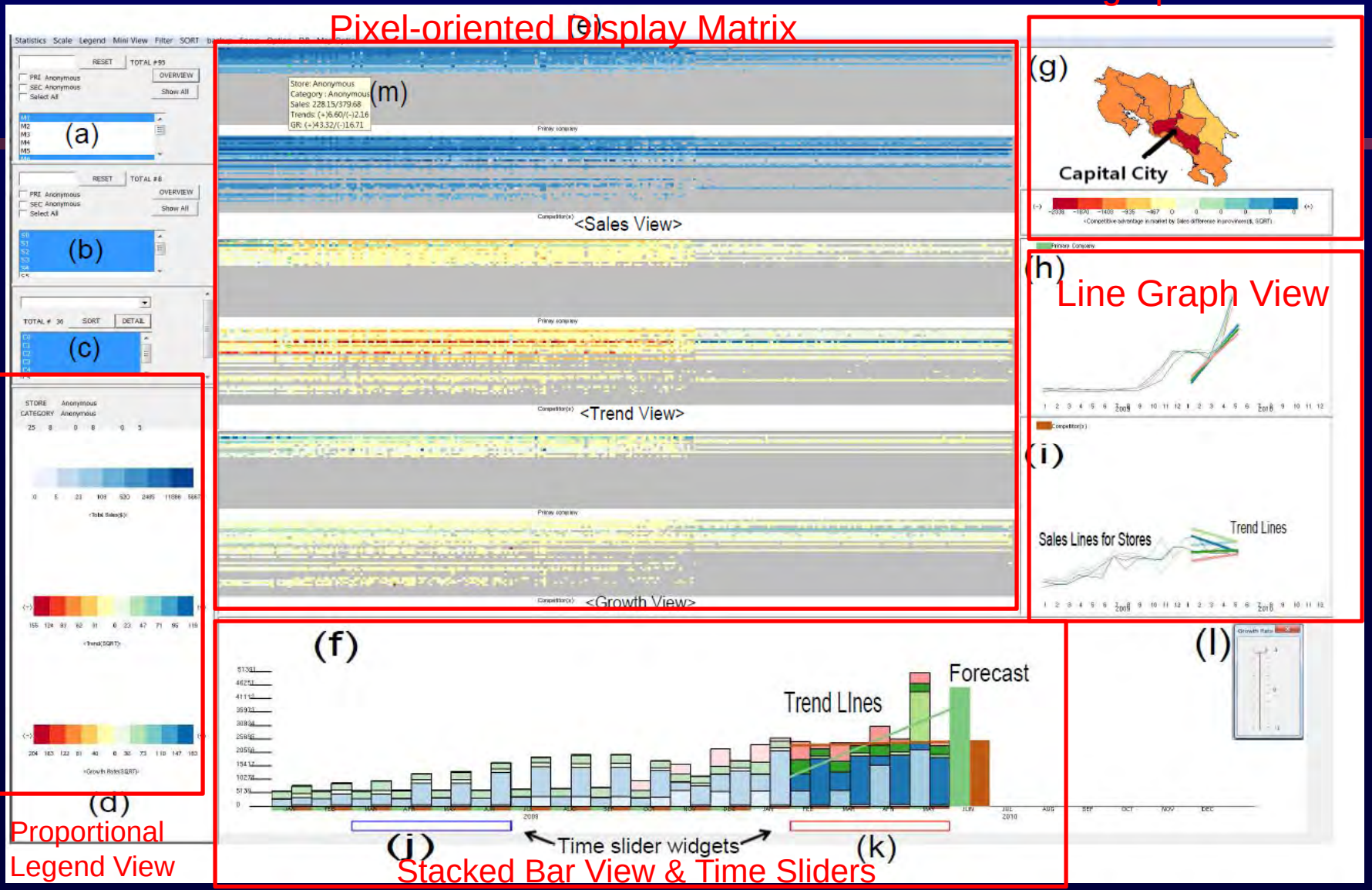
Competitive Business Intelligence Based On Point of Sales Data



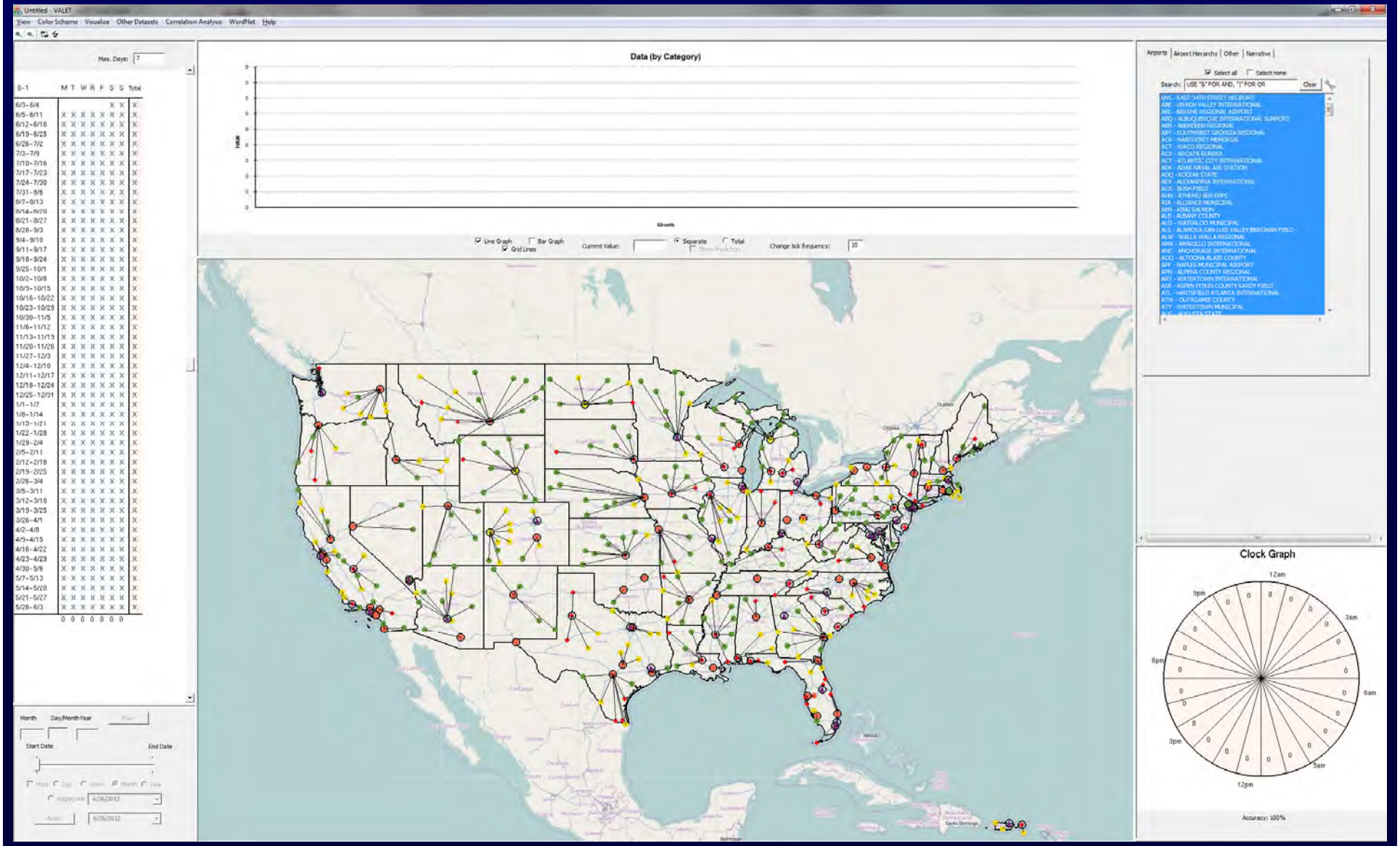
- Characteristics of point of sale data
  - Multivariate
    - Large # of dimensions
      - E.g., 38 categories(products) in 288 stores
  - Temporal : 18 months
  - Spatial
    - Stores located all over Costa Rica
  - Requirements
    - Support easy comparison among companies
      - E.g., visualize all data at once, sorting by importance
      - Enable geographical comparison
    - Enable easy recognition of any change in sales
      - E.g., proportional legends
    - Forecast for decision-making



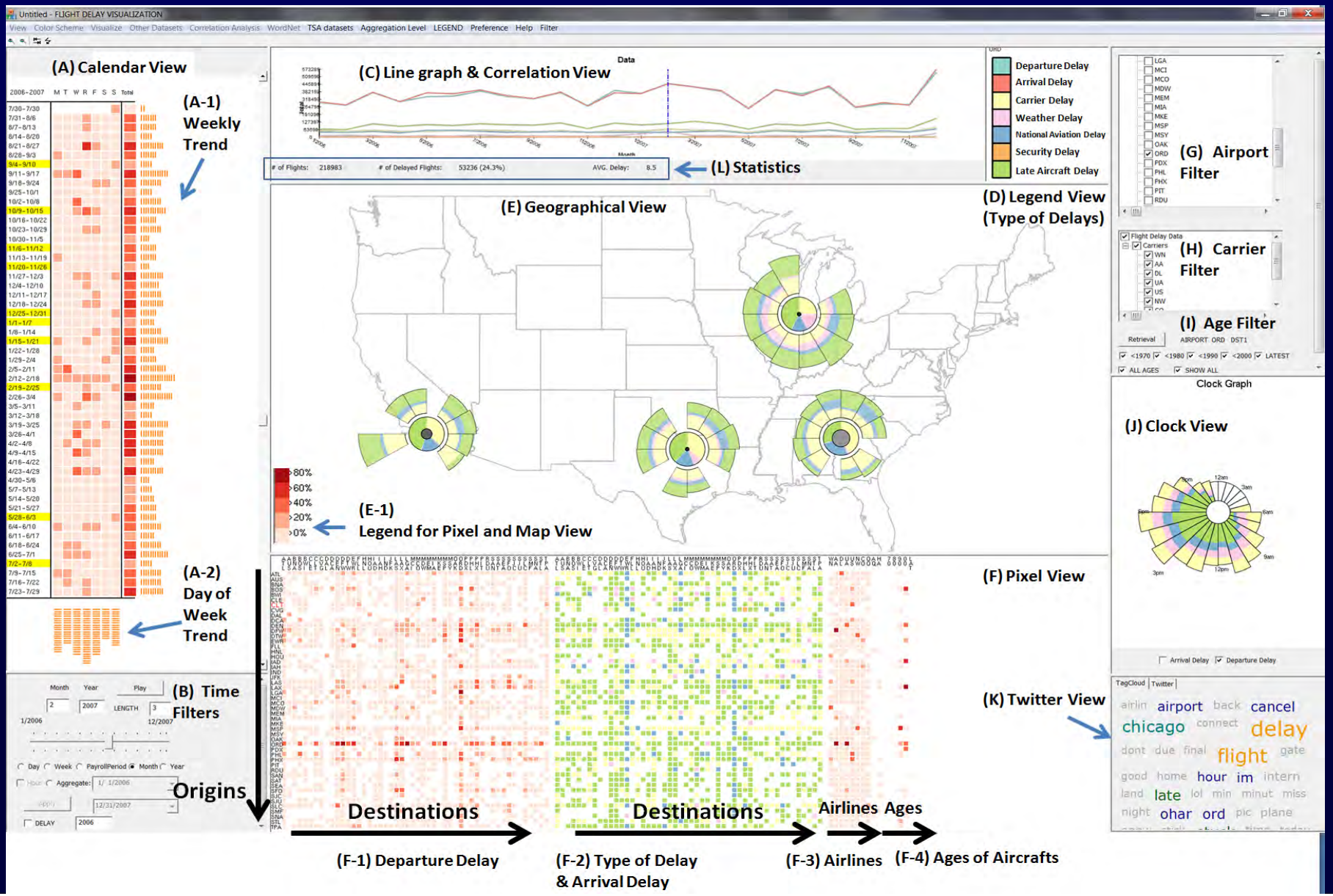
Market Analyzer



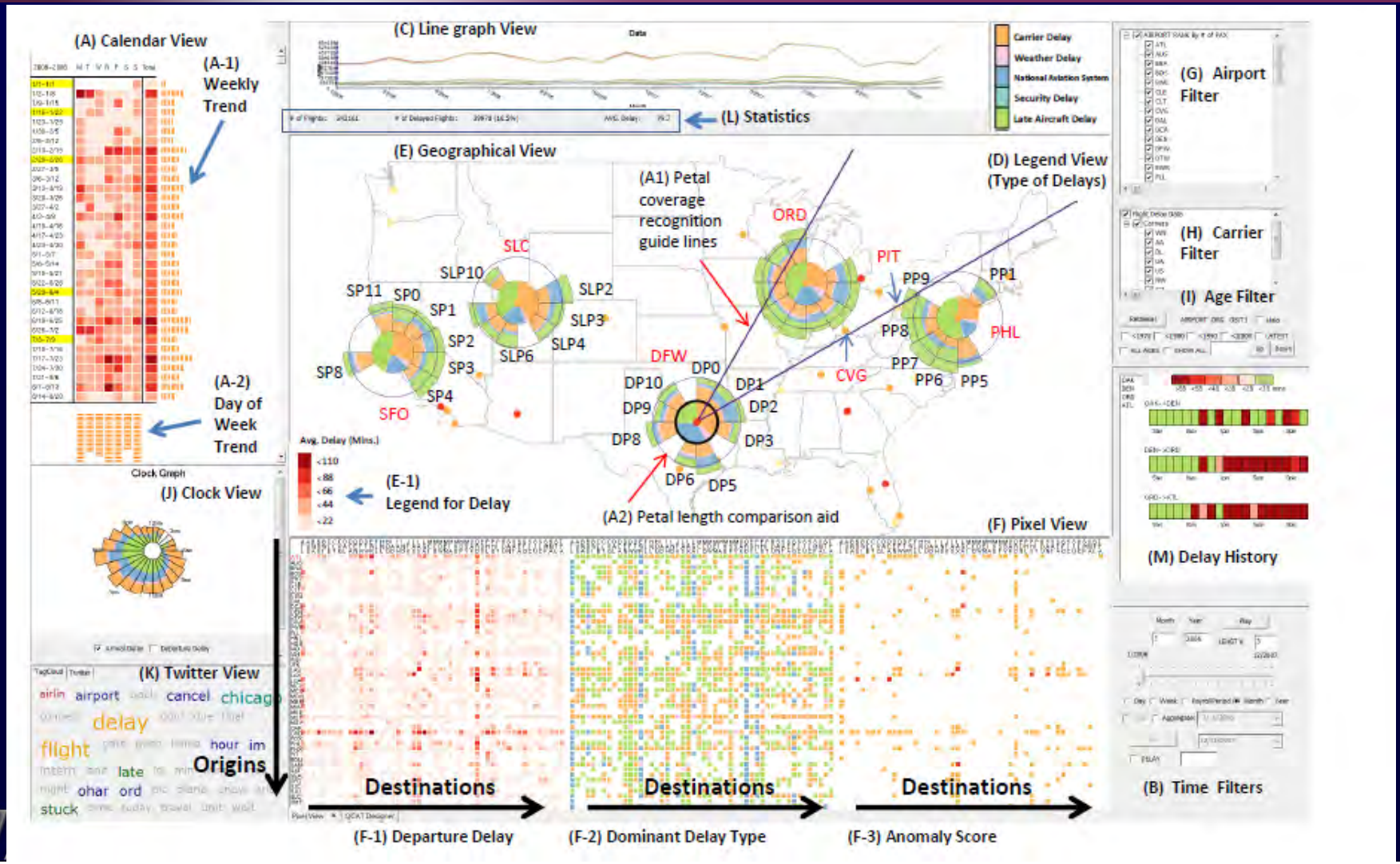
Example Projects: TSA Visual Analytics



Flight Delay Visual Analysis Prototype

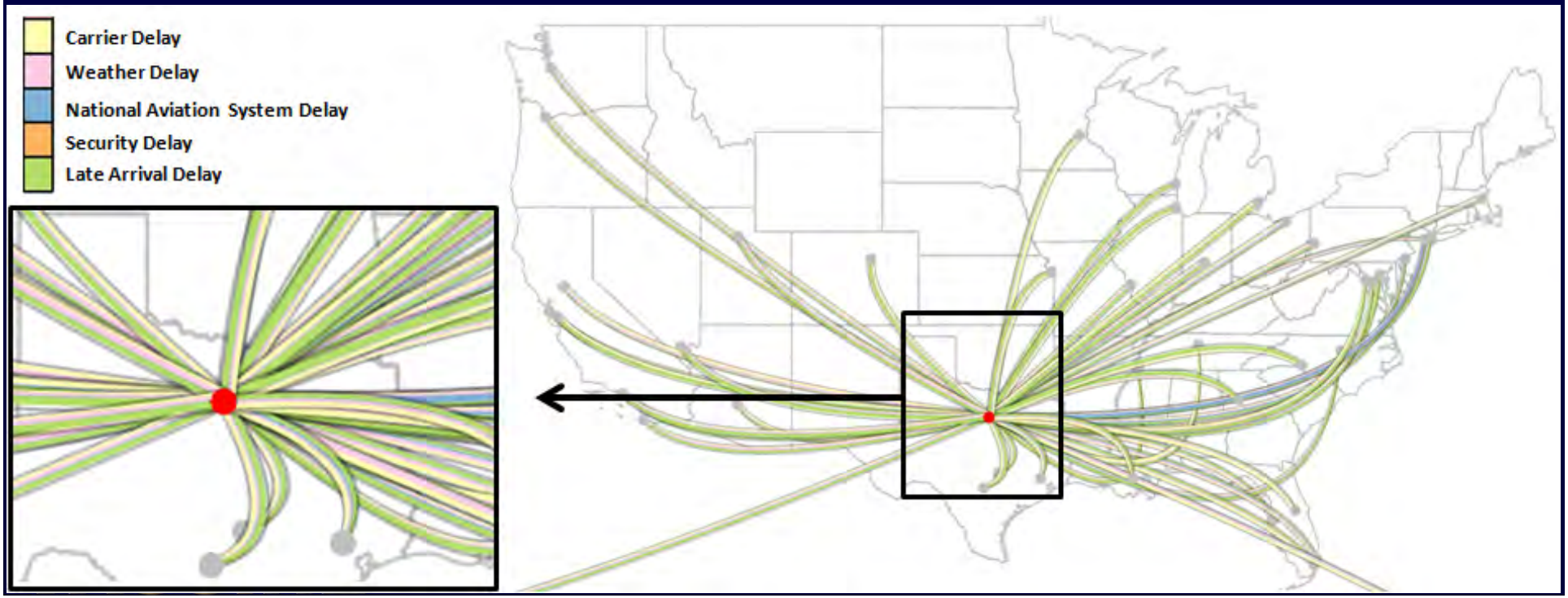


Multivariate Network Analytics with Information Theory Anomaly Detection

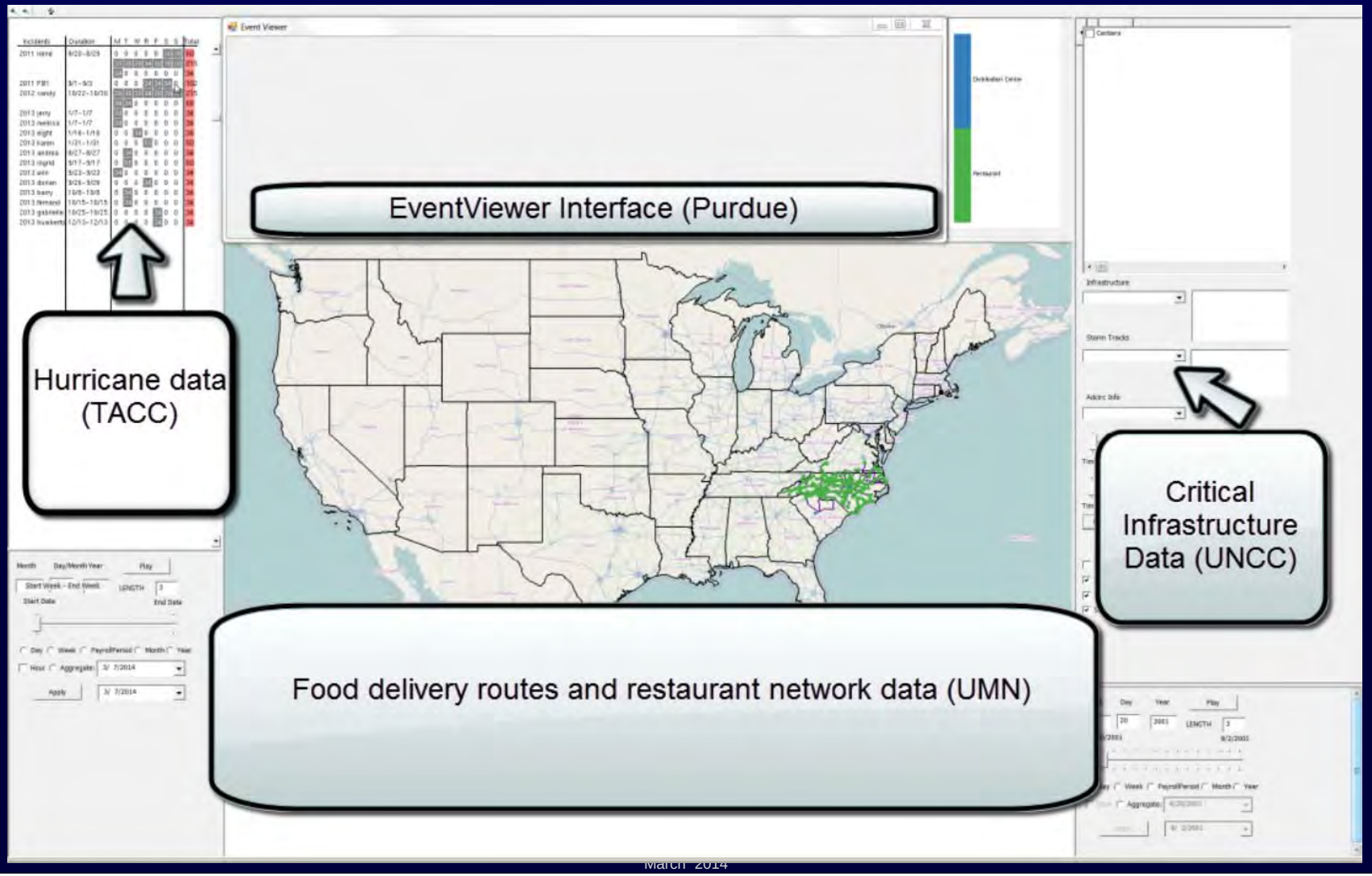


Flight Delay Visual Analysis Prototype

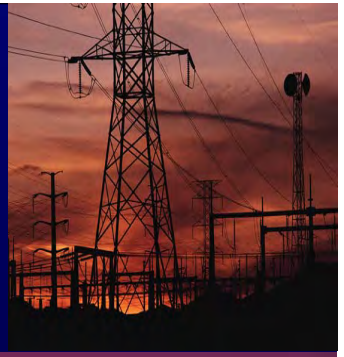
- Delay at DFW causing delays at destinations by delay type



Example Projects:
Systems of Systems Visual Analytics



Cascading Critical Infrastructure Resiliency Modeling and Analytics (VASA)

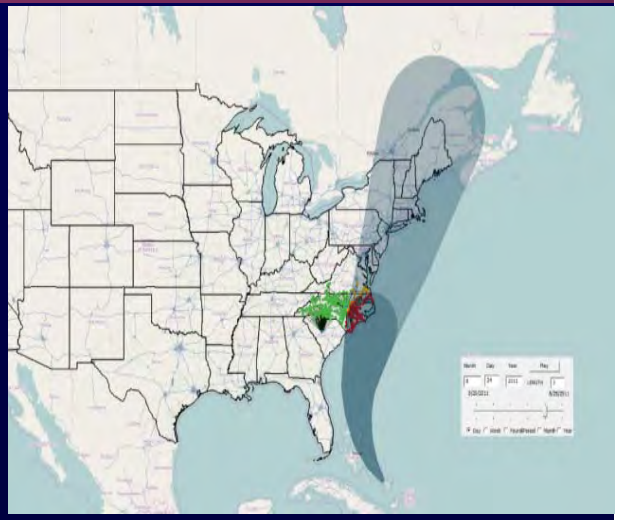


- Purpose: Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.
- Collaborating Institution(s):
  Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities
- End-User(s): Power Suppliers (e.g., Duke Energy), Cyber Community (e.g., Cisco), Quick Service Restaurants and suppliers, food supply

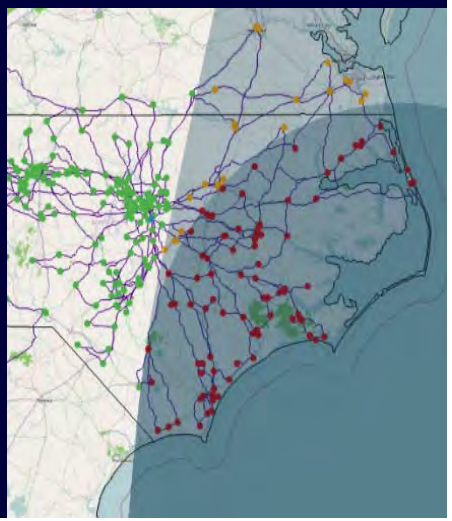
VASA: Visual Analytics for Security Applications

Ebert, Elmqvist, Gaither, Kennedy, Ribarsky, Wang
Collaborating Institution(s): Purdue, Minnesota, UTexas, UNCC + German universities
End-User(s): Fast-food restaurant chain, emergency management and planning personnel

Impacts and Accomplishments:

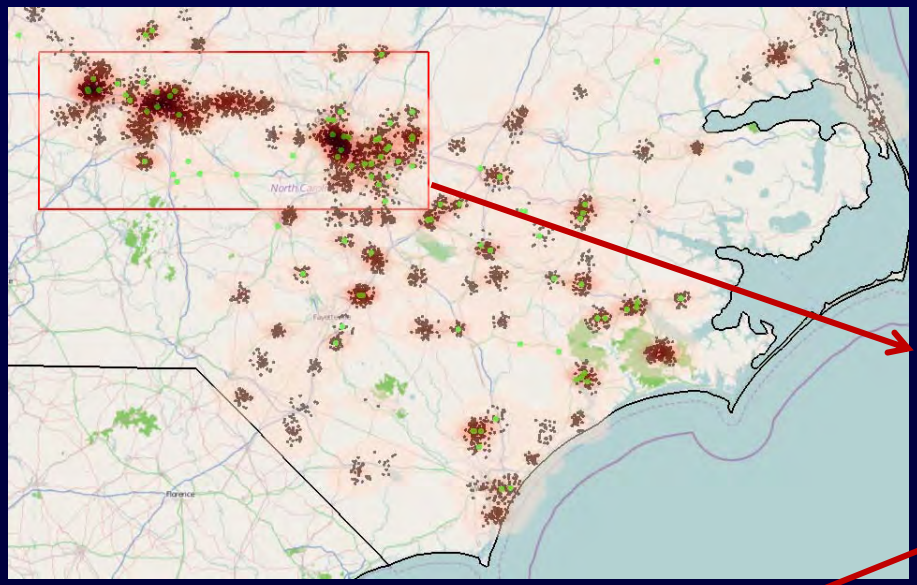


- Support decision-making for extreme weather and natural disaster scenarios
  - Combine real and simulation data
  - Allow “what-if” exploration
- **System of systems**: binds together multiple simulations models from collaborators into coherent whole
  - **Minnesota**: food distribution model
  - **Texas**: simulated and historical weather (hurricanes, storms)
  - **UNCC**: critical infrastructure
  - **Purdue**: roads + interaction visual analytics tool
- **Challenge**:
  combine interactive VA with complex simulation models



Example Foodborne Illness Scenario (Day 5, 8/31/11)

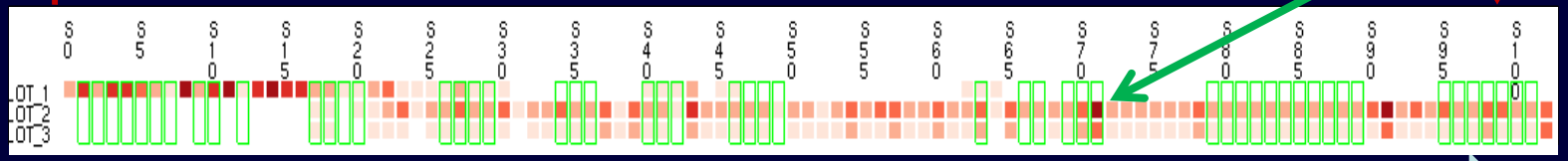
Day 5: Contaminated product consumption begins



- Production Time:
  - LOT_1, LOT_2, LOT_3
- People ill as brown dots
- Heatmap of people ill in red
- Stores in green

Product information for stores in selected region (where the most people are ill)

Squares: Darker red means larger quantity of that lot sold that day in that store



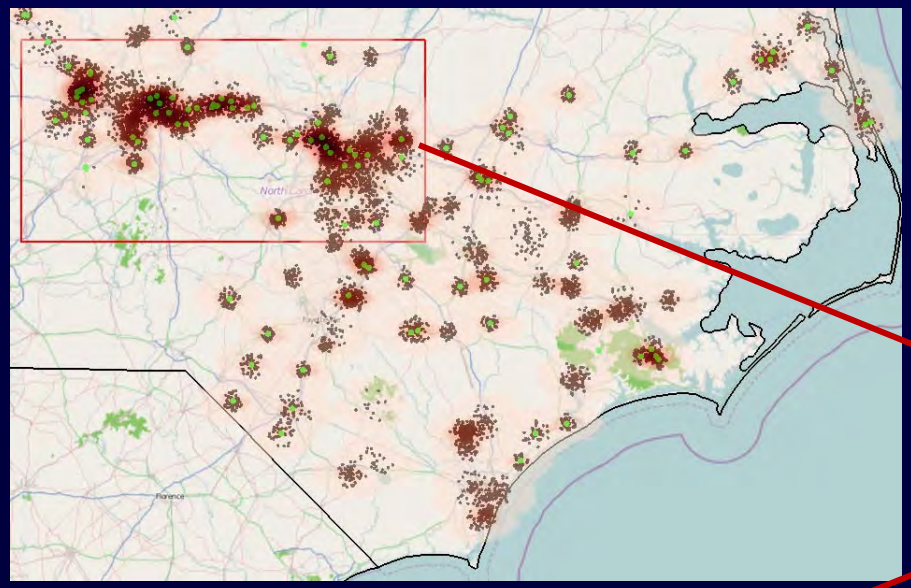
Y-axis: LOTs (production time)

X-axis: different stores

VACCINE

Example Foodborne Illness Scenario (Day 6, 9/1/11)

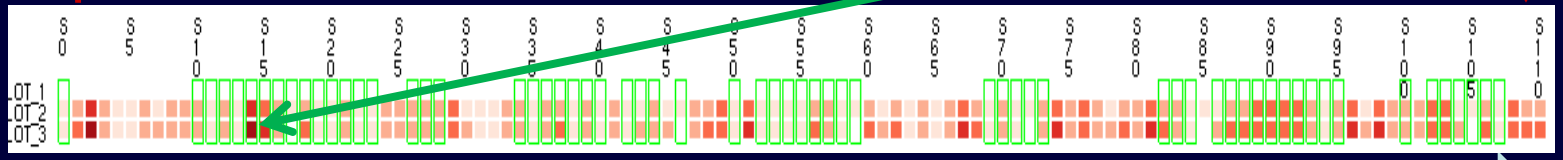
Day 6: Contaminated product consumption continues



- Production Time:
  - LOT_1, LOT_2, LOT_3
- People ill as brown dots
- Heatmap of people ill in red
- Stores in green

Product information for stores in selected region (where the most people are ill)

Squares: Darker red means larger quantity of that lot sold that day in that store



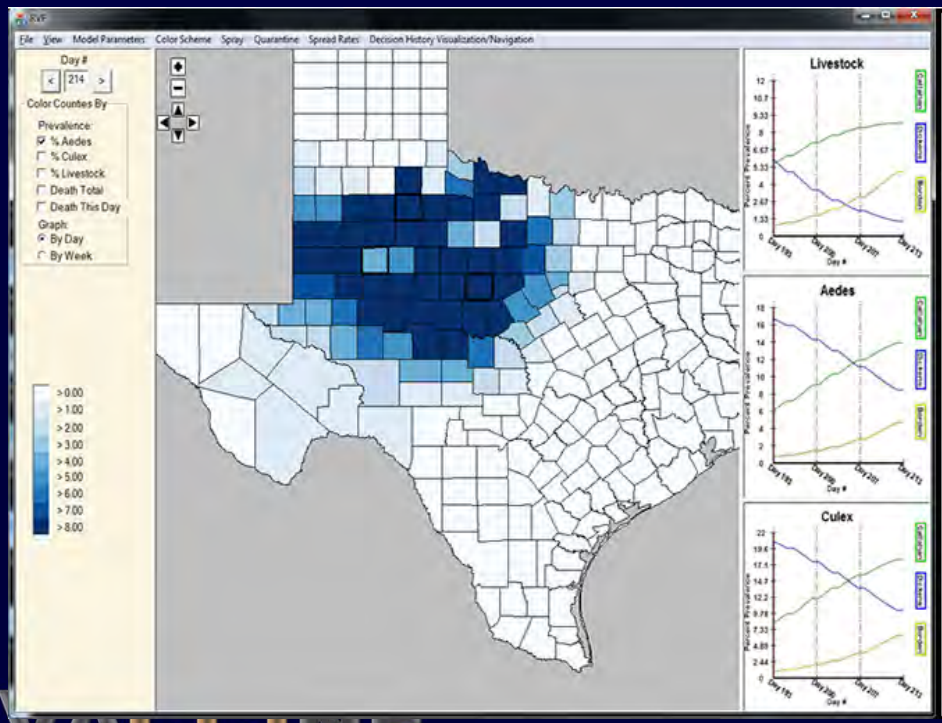
Y-axis: LOTs (production time)

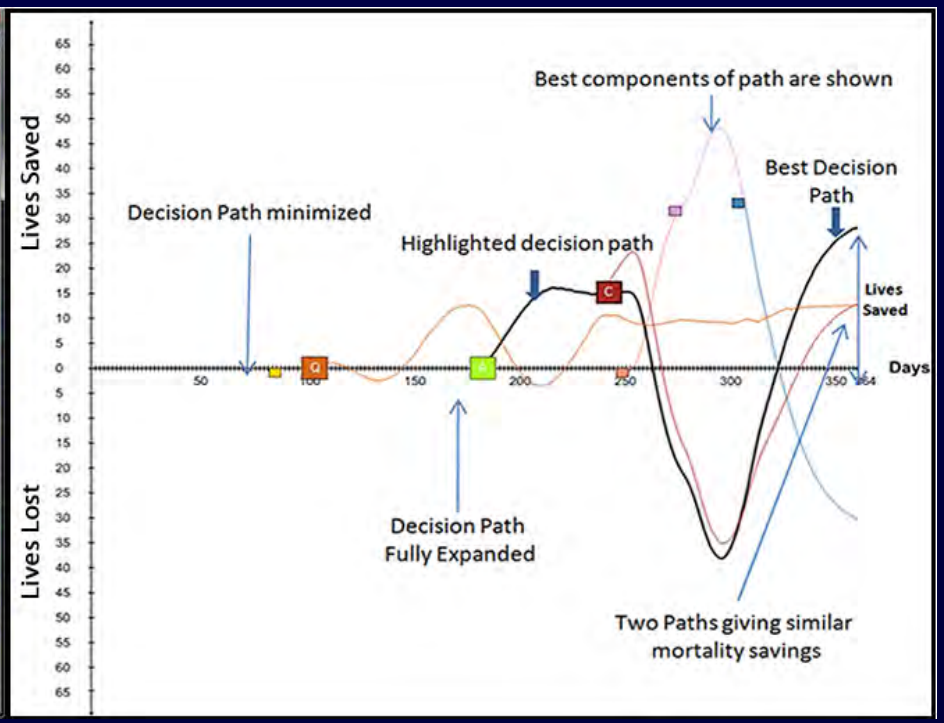
X-axis: different stores

VACCINE

Complex Decision Making: Advanced Decision Support Tools (e.g., Rift Valley Fever)

- Integrated simulation models
- Explore alternative courses of action in decision space and real space





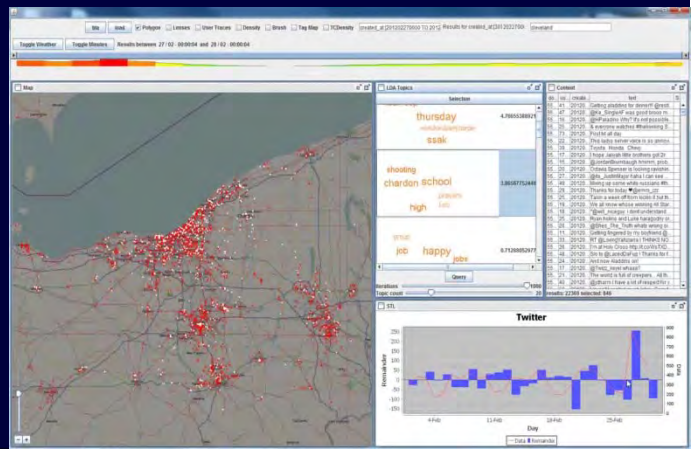
Example Projects: Public Safety and Law Enforcement Applications



Visual Analytics Uses for Public Safety



- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Multisource, multimedia massive data integration & analysis
- Purpose: Planning for resiliency, training, detection, investigation, response, recovery, remediation



VACCINE

Basis For Work: Novel Spatiotemporal VA

- Uses:
  - Anomaly detection
  - Situational awareness
  - Predictive policing
  - Coast Guard visual analytics
    - Search and rescue
    - Risk-based resource allocation
    - Risk-based operational performance

Team Situational Awareness, Planning, Investigation, Response Visual Analytics

- TSA from Big Data
  - In-field and desktop historical and current synchronized incident information from all relevant sources
  - Innovative spatiotemporal visualization, analysis, prediction
  - Image analysis tools
  - Integrated statistical forecasting and simulation models
  - Advanced social media analytics
- Tools for intelligence analysis
- Planning and response support
  - Integrated simulation, resource allocation, census, social media, weather data

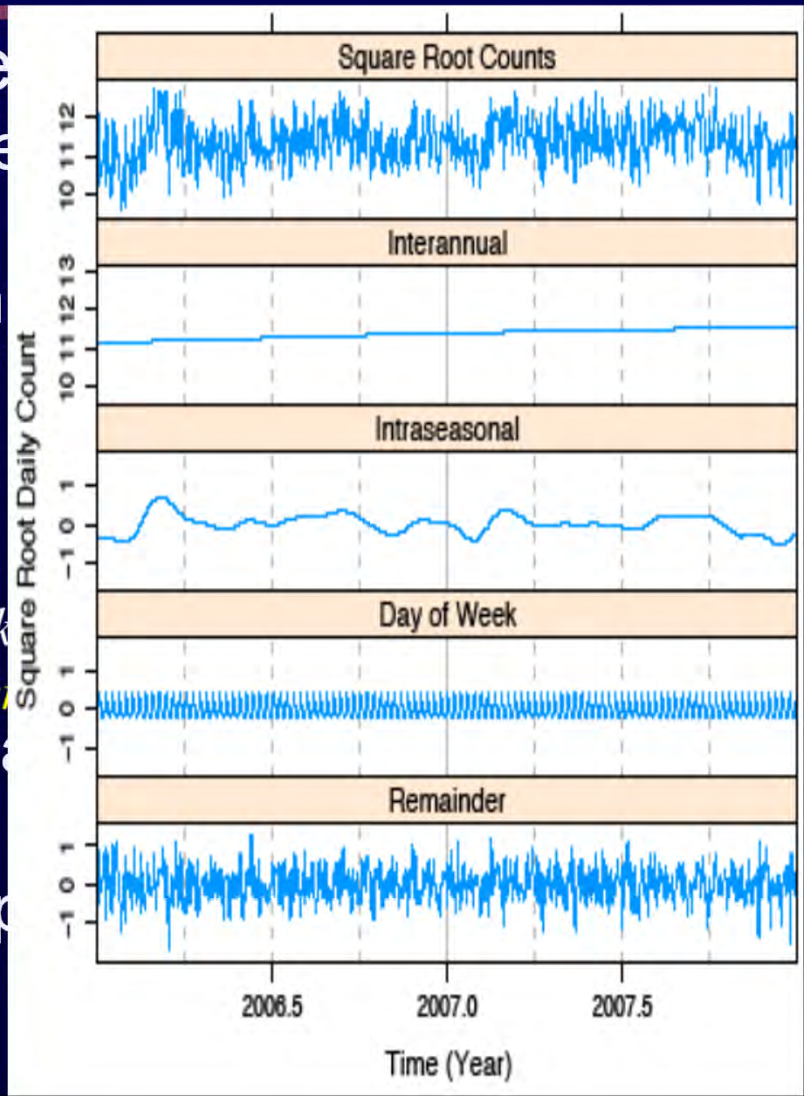
Event/Anomaly Detection - STL

- Temporal trend analysis and pre
  - Seasonal Trend decomposition base
    weighted regression) (STL)
  - Time series data viewed as the sum

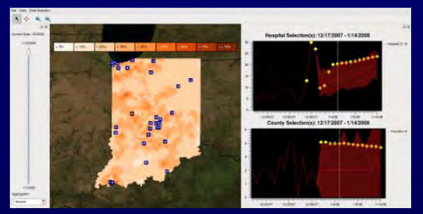
$$Y = T$$

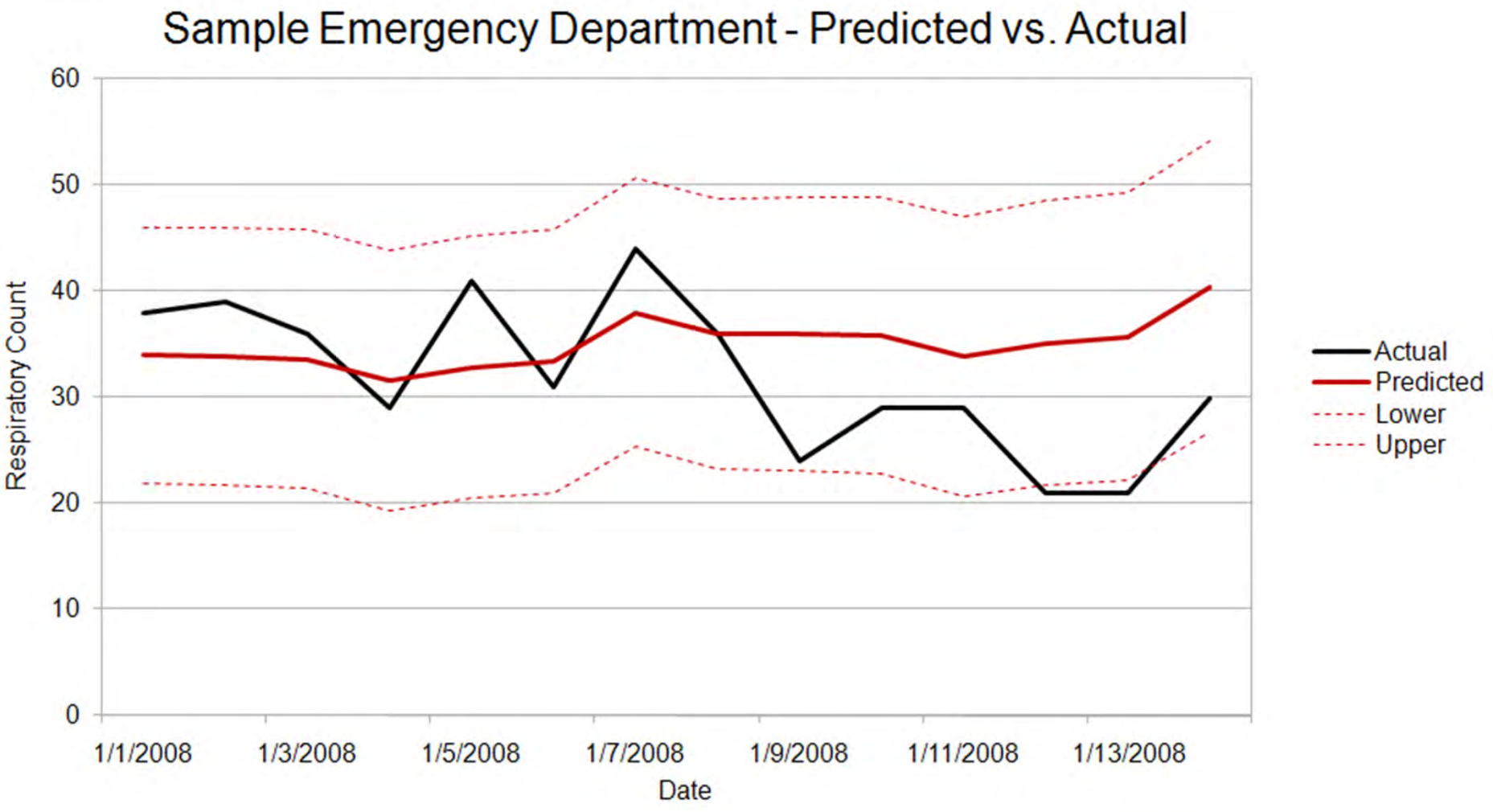
Y : original time series
T : trend component
S : seasonal (daily/week
R : remainder compone

  - Large value of *R* indicates substanti
  - Filter out the noise component
  - If value varies by 2 SD from STL exp

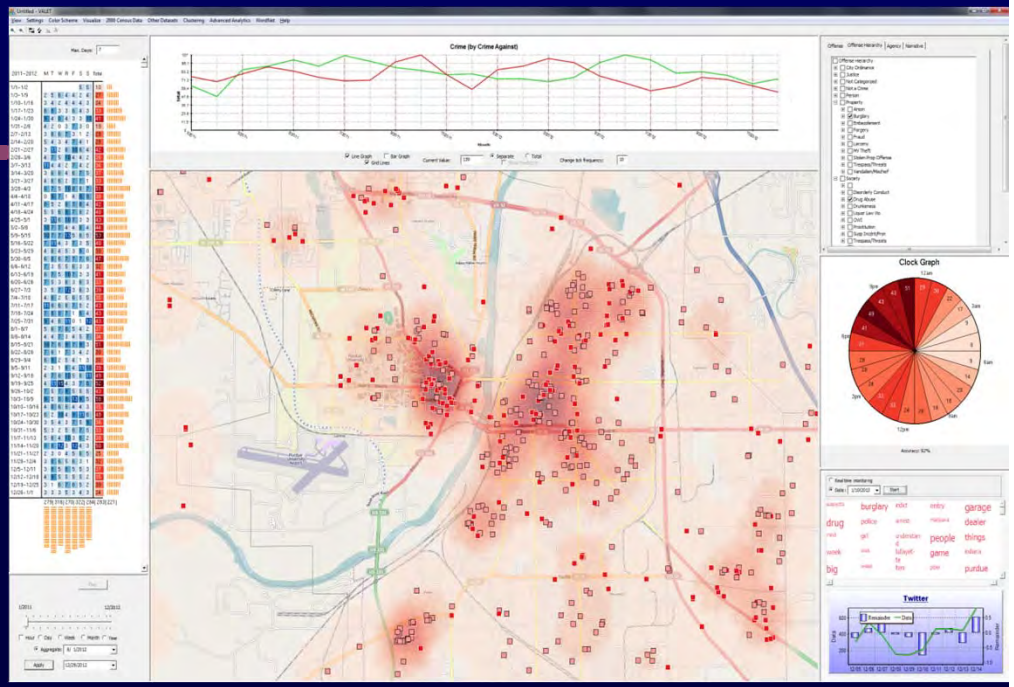


Predictive Visual Analytics





Visual Analytics Law Enforcement Toolkit (VALET, iVALET)



Impacts:

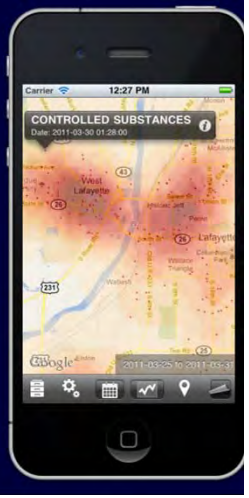
- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public (June 2014) for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

VALET delivered:

- Spring 2011: WL, Lafayette Police
- Fall 2013: NYPD

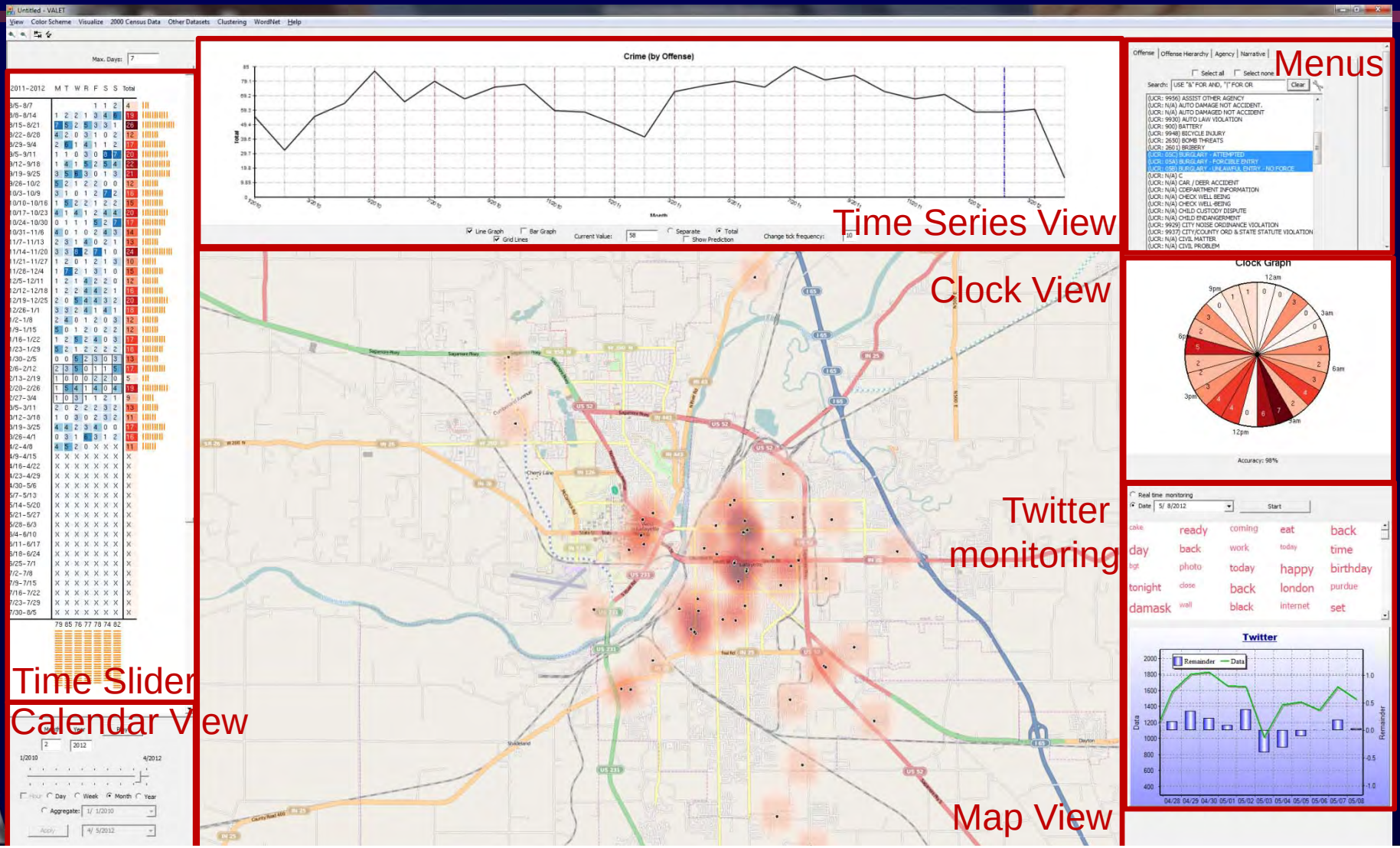
iVALET delivered:

- October 2011: Purdue, WL Police

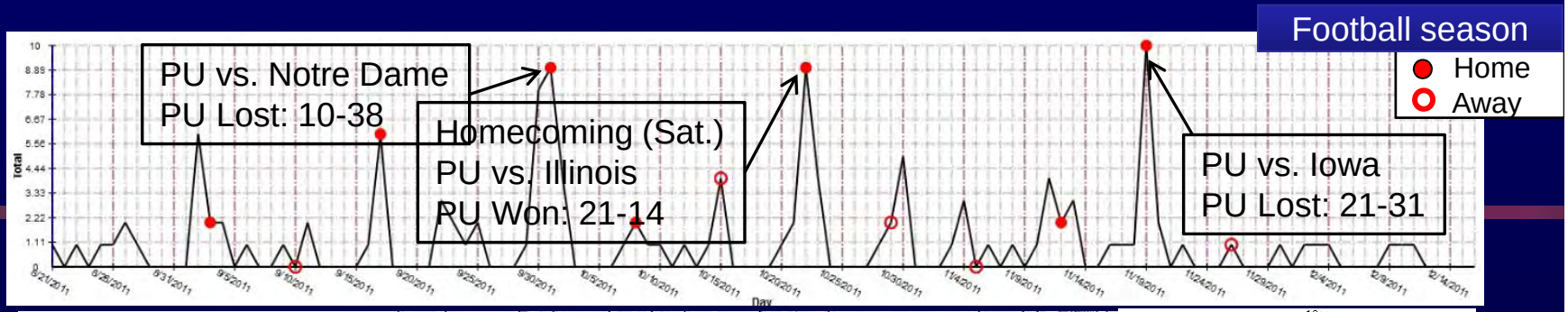


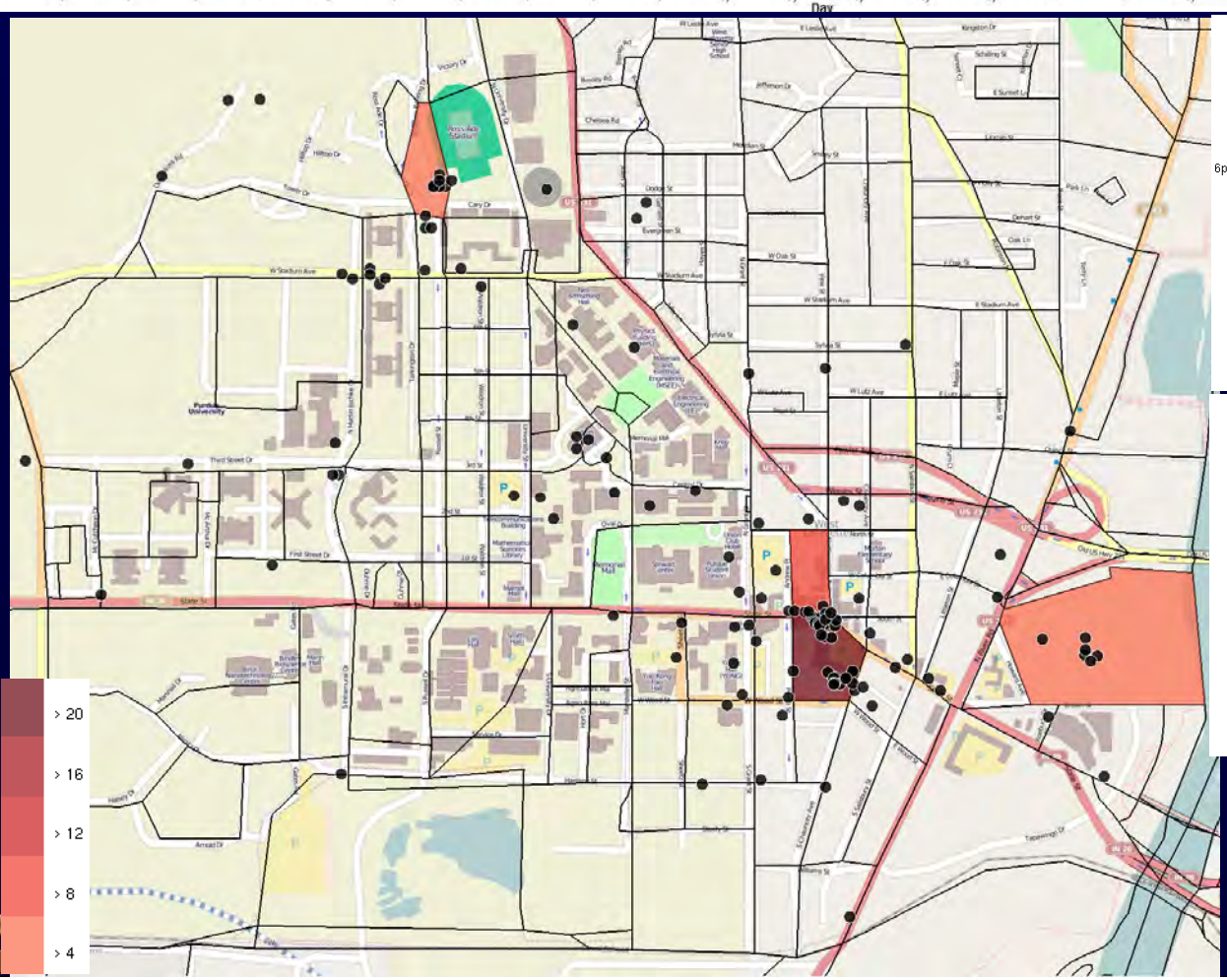
VACCINE

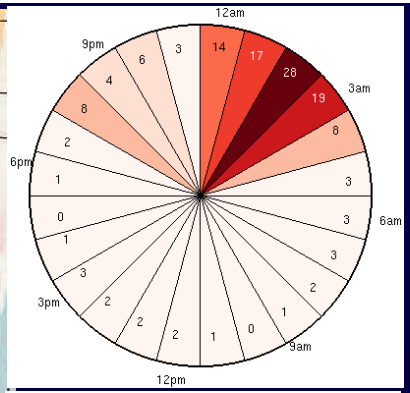
VALET Overview

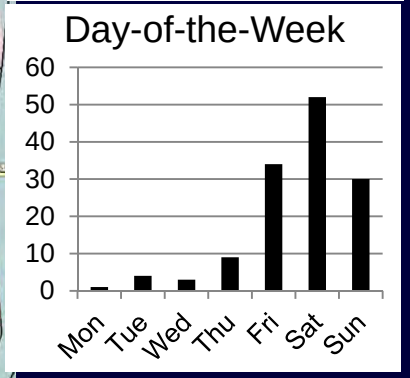


Example: Drunkenness / Public Intoxication



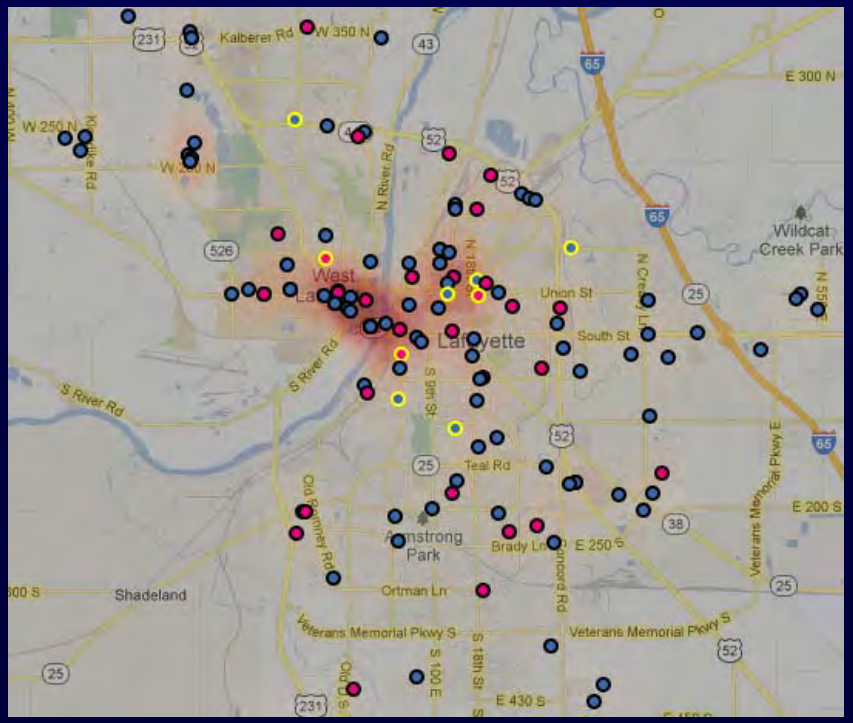






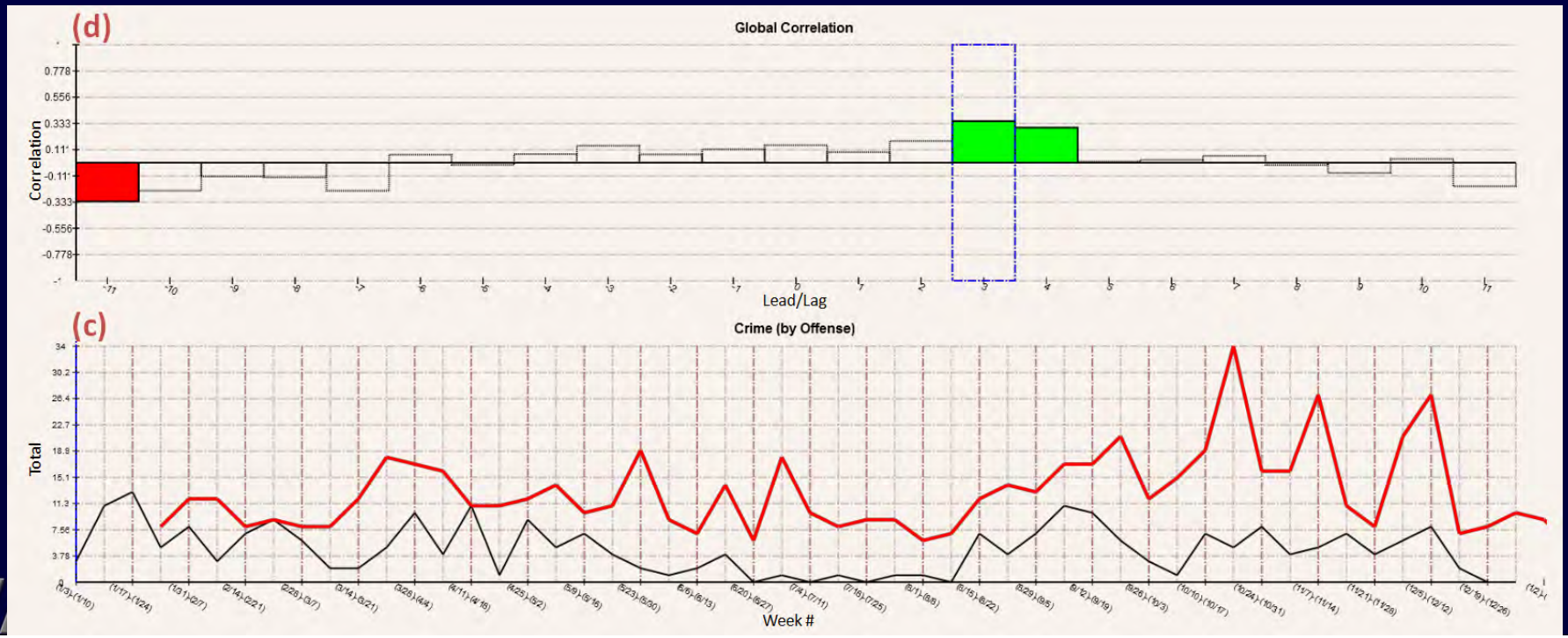
Top 10 Hot Incidents

- Identify unusual localized high-frequency patterns of crimes (near repeats)
- Each data entry is checked for other crimes with similar properties within a 1 block radius of the incident location and a 14-day time period;
- Top 10 incidents with the most number of related incidents in this space-time window are highlighted



Multivariate Correlative Predictive Analytics: Three Approaches

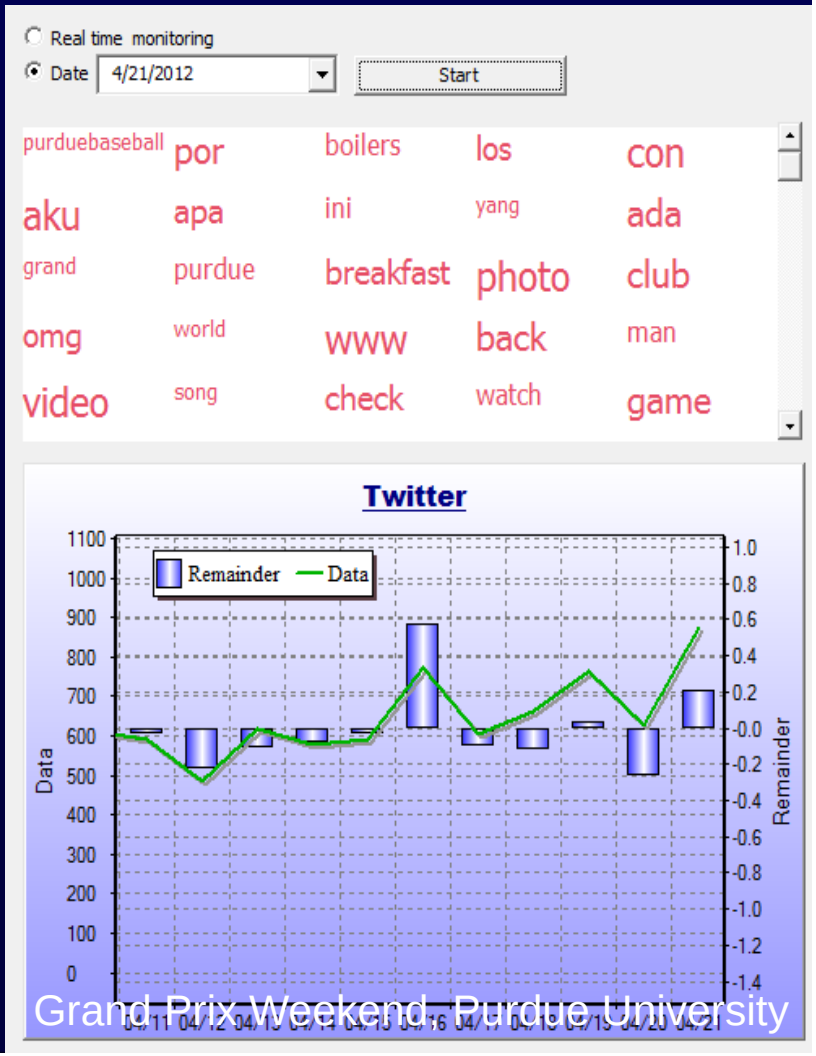
- Automatic correlation computation against lead/lags
- Temporal and spatial windowing
- Data category parameter space



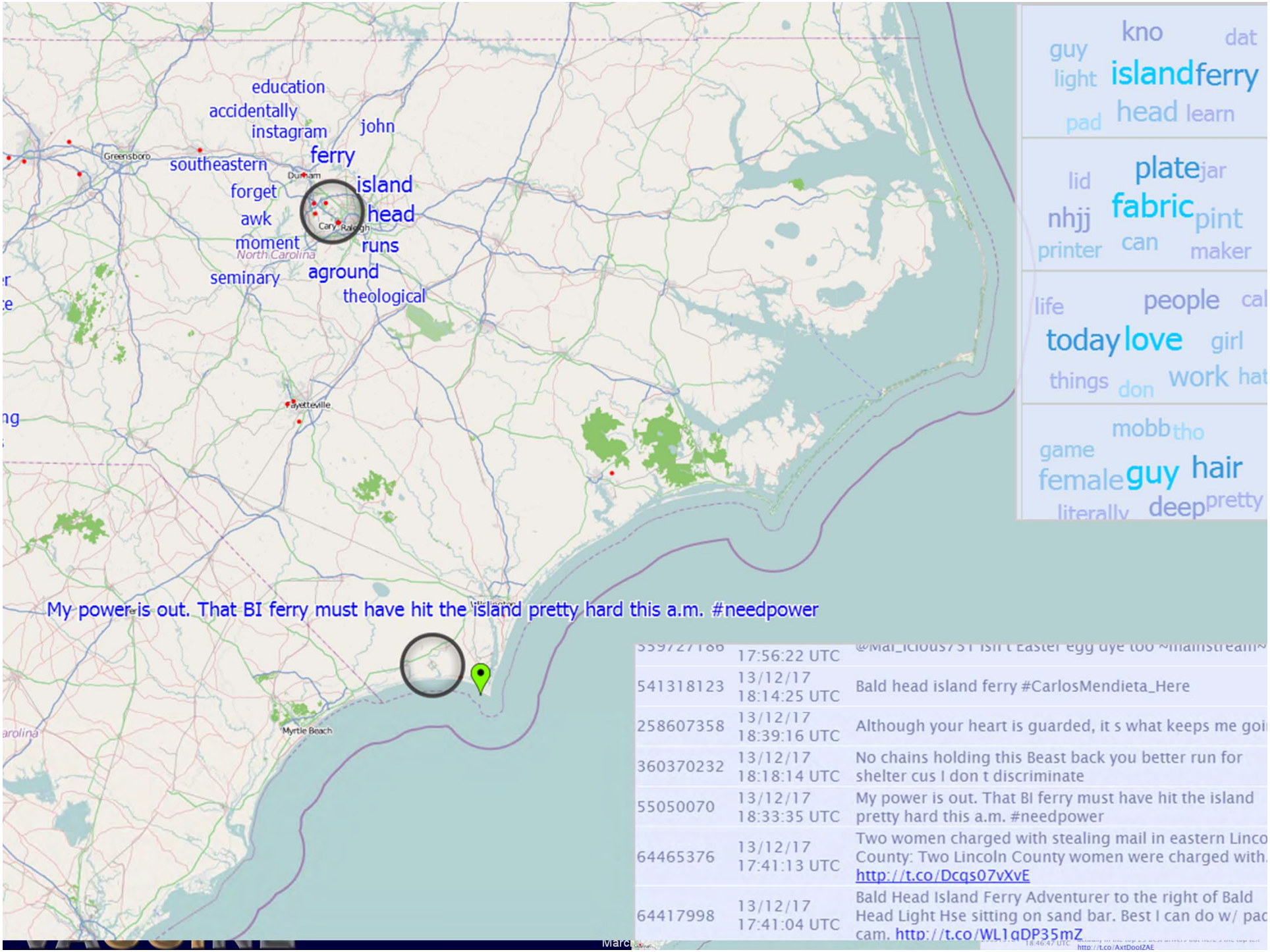
Social Media: Real-time Twitter Monitoring and Integration into Tools

(Purdue, Stuttgart, Penn St.)

- Anomalous topic extraction using LDA and novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud

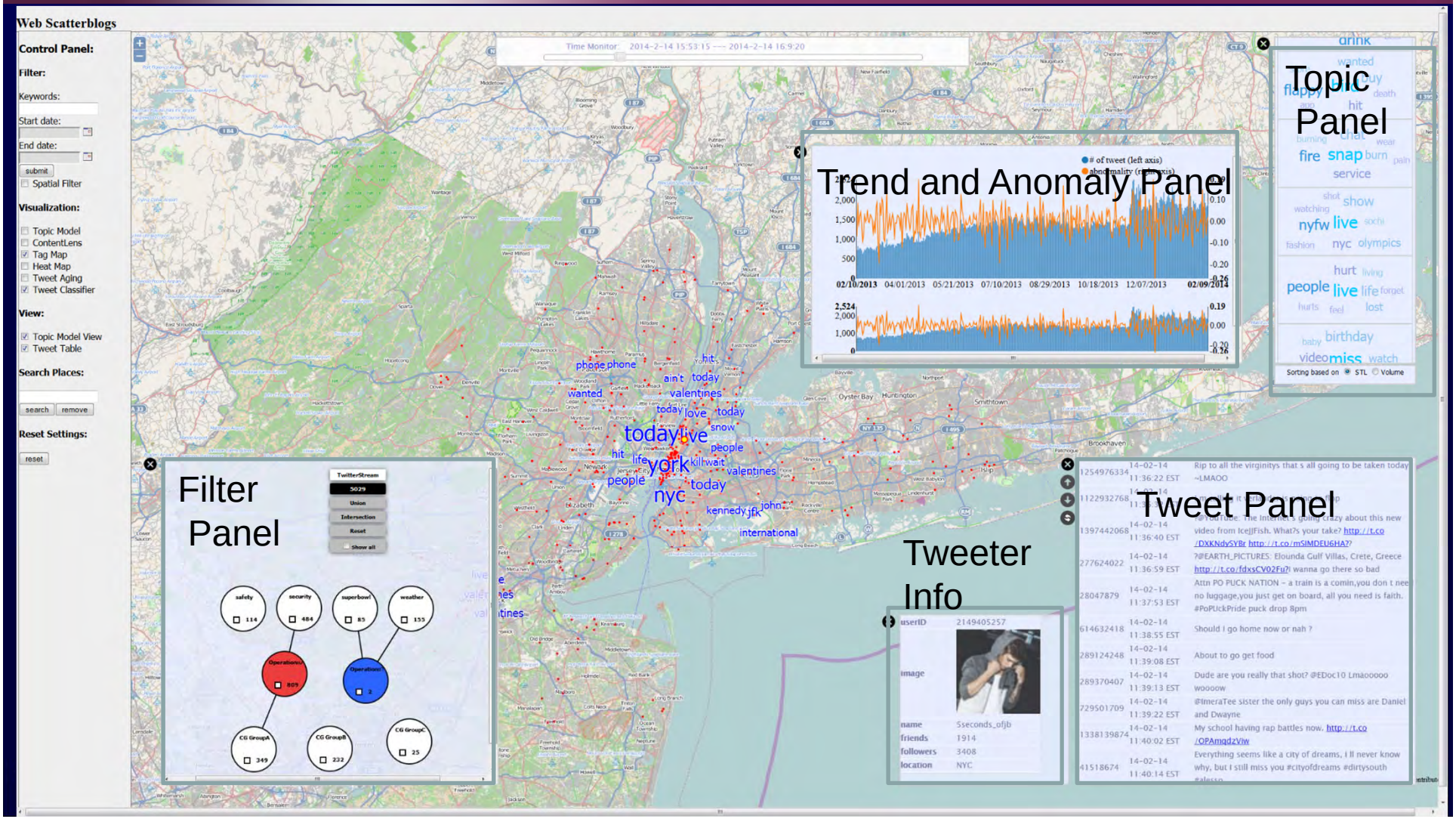


Example Projects:
Social Media Visual Analytics



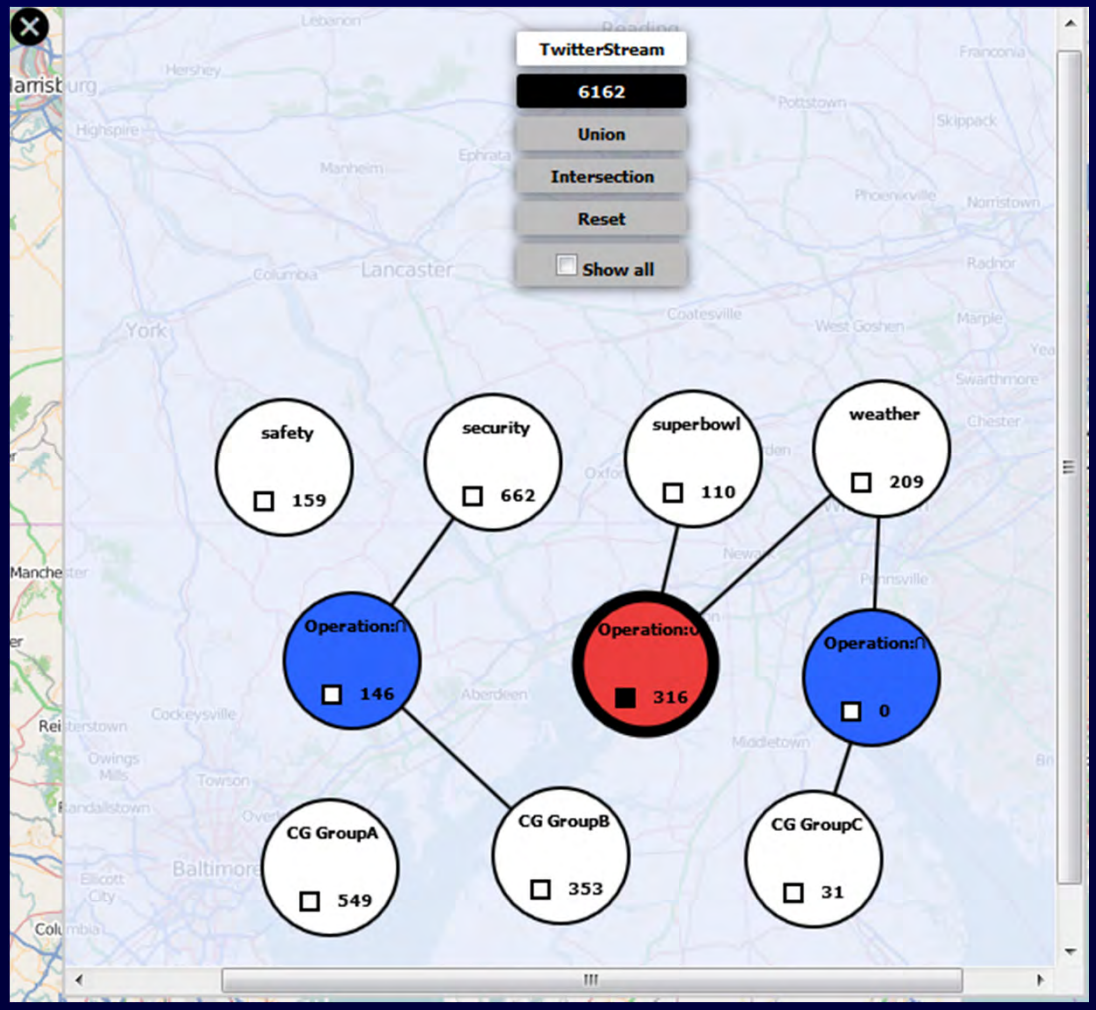
359727186	17:56:22 UTC	@Mal_icious731 isn't Easter egg dye too ~mainstream~
541318123	13/12/17 18:14:25 UTC	Bald head island ferry #CarlosMendieta_Here
258607358	13/12/17 18:39:16 UTC	Although your heart is guarded, it s what keeps me goi
360370232	13/12/17 18:18:14 UTC	No chains holding this Beast back you better run for shelter cus I don t discriminate
55050070	13/12/17 18:33:35 UTC	My power is out. That BI ferry must have hit the island pretty hard this a.m. #needpower
64465376	13/12/17 17:41:13 UTC	Two women charged with stealing mail in eastern Linco County: Two Lincoln County women were charged with. http://t.co/Dcqs07vXvE
64417998	13/12/17 17:41:04 UTC	Bald Head Island Ferry Adventurer to the right of Bald Head Light Hse sitting on sand bar. Best I can do w/ pac cam. http://t.co/WL1qDP35mZ

Scatterblogs – Web Version Overview

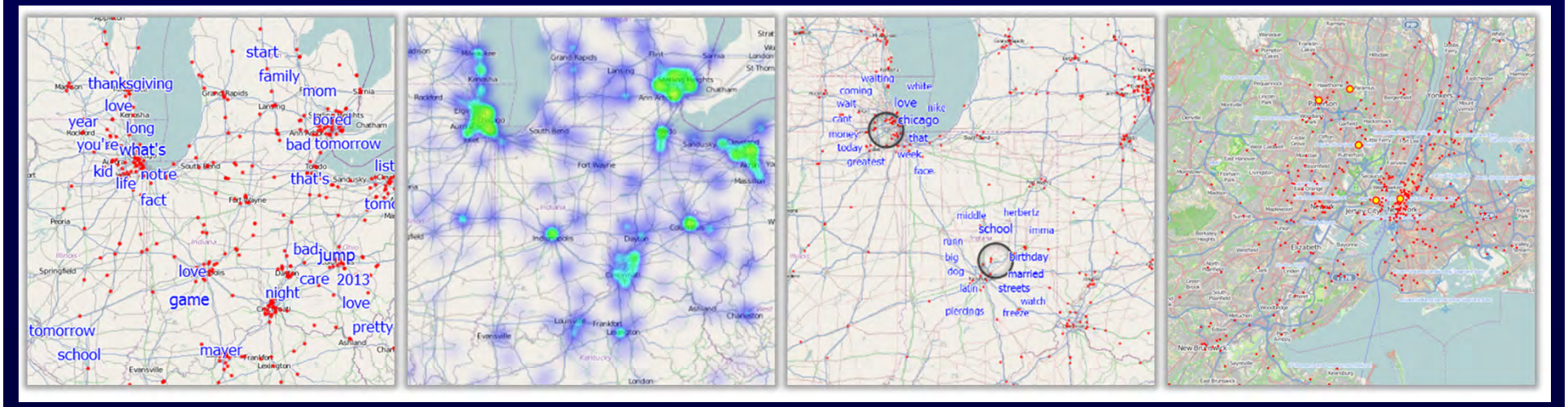


Scatterblogs – Filters

- Keywords to define various filters – can be an early warning system
- Can combine filters to specify overlap
- Can view tweets only associated with these filters to reduce information overload

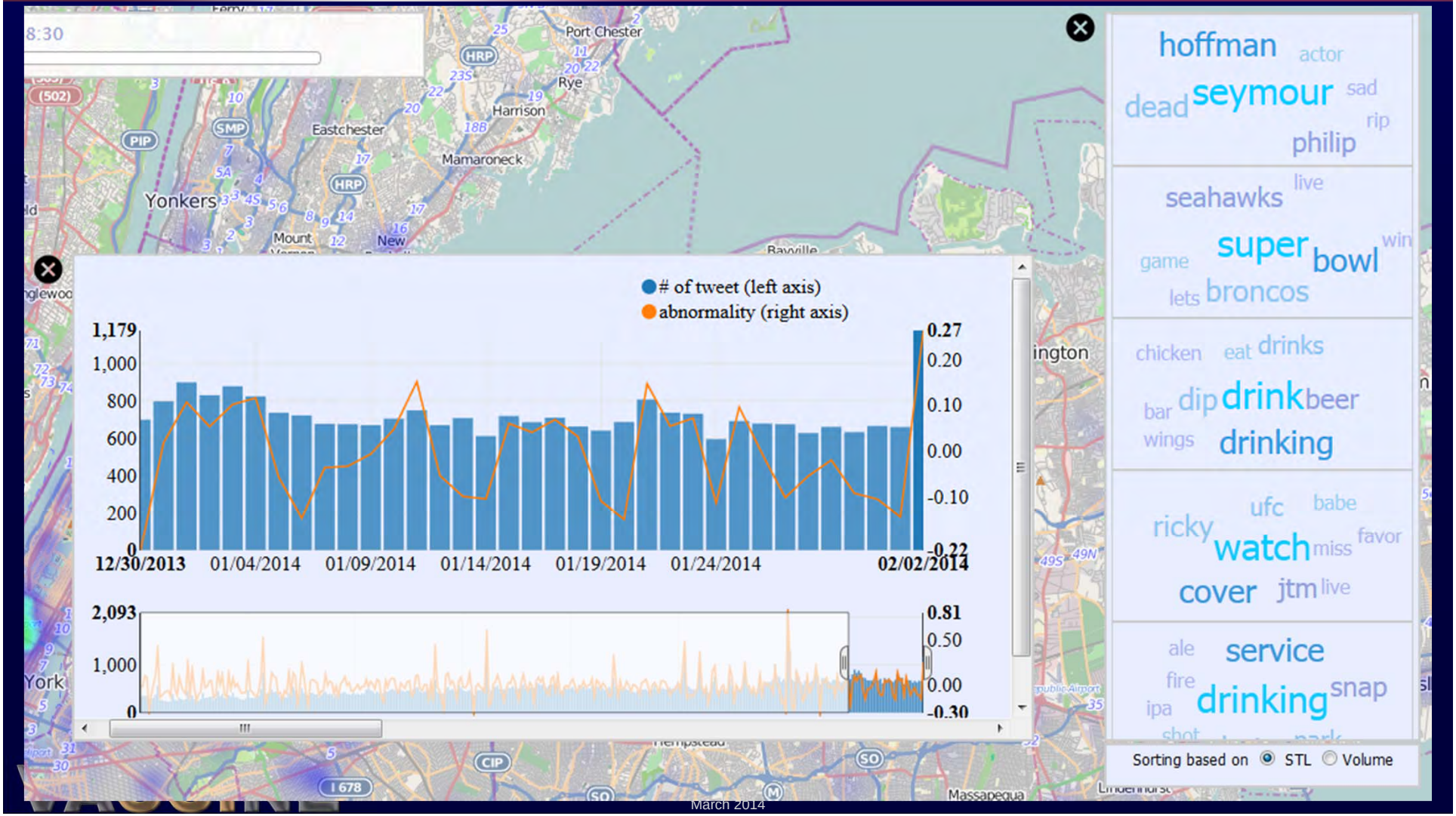


Scatterblogs – Screen Options



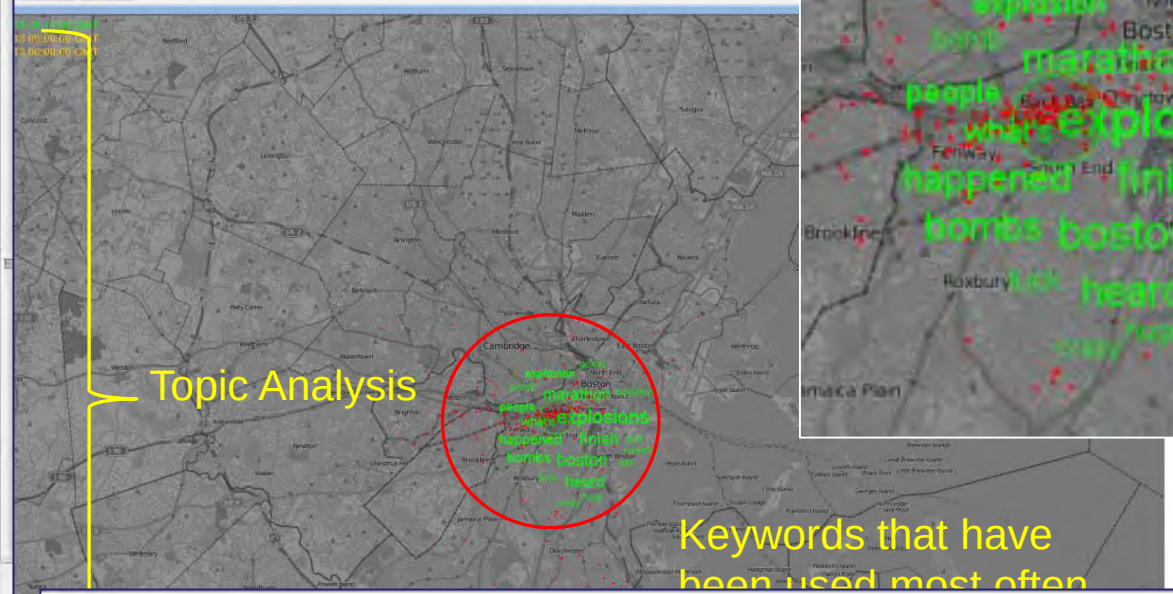
Tag Map | Heat Map | Content Lens | Realtime tweet aging

Anomaly Detection



Explosion Area in Boston







Keywords that have been used most often

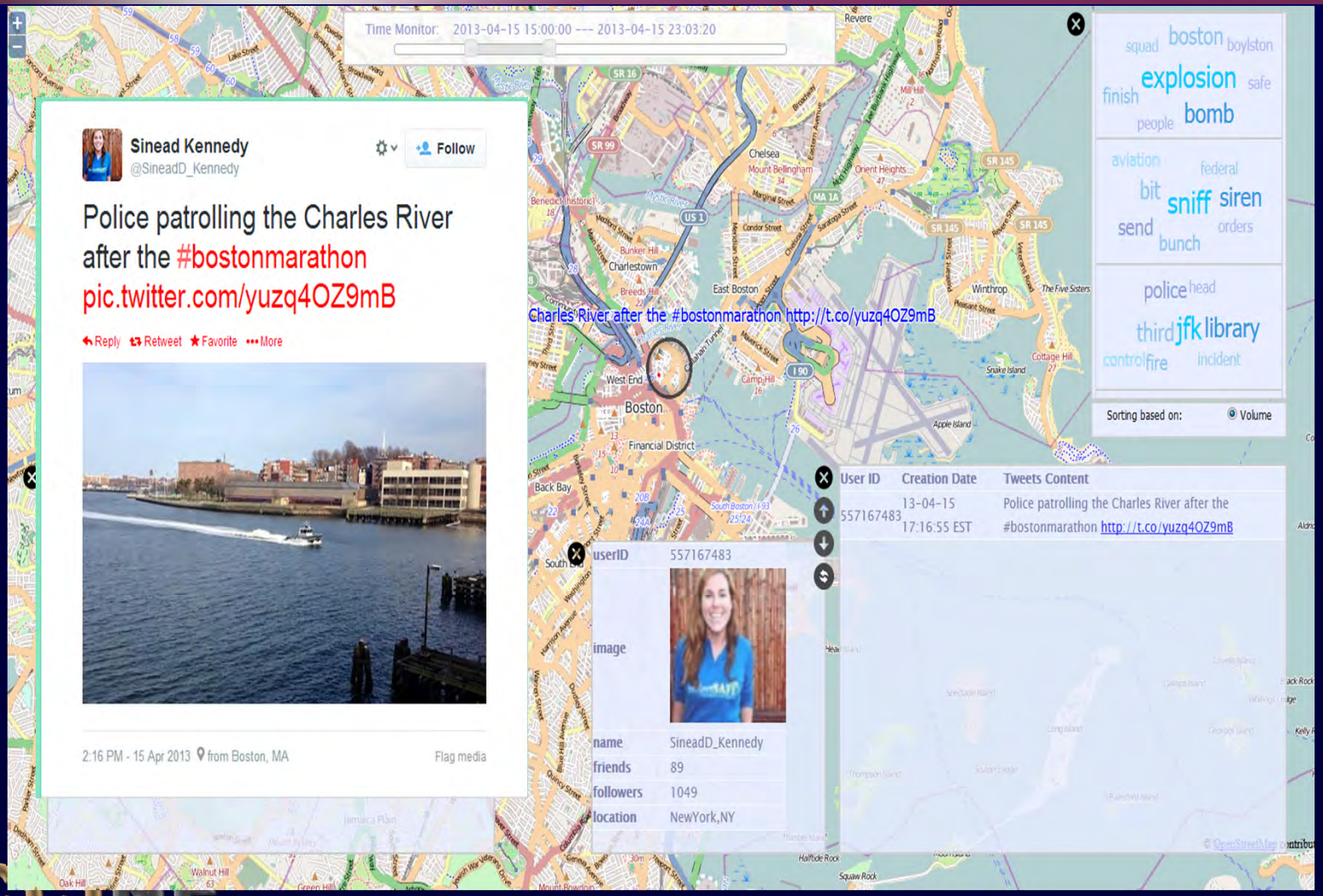
Tweets

Text
Oh my god what just happened
Something happened at the end of the Boston Marathon. Something bad and there is a lot of chatter on Twitter. What's going on?
Multiple people are injured near the Boston Marathon finish line after two explosions. The #BostonMarathon has been stopped.
two bombs just went off on boylston
BREAKING NEWS: Two powerful explosions detonated in quick succession right next to the Boston Marathon finish line this afternoon.
What the FUCK was that
Back in Sept, @croon1 solicited me for $2000. He now has a music video with William Shatner. If you watch it (god forbid) keep that on mind.
Literally what the fuck get me out of here
@DTenenbaum my office right next to it
@FRANCESCalcio I figured 3 people would get the joke
Two explosions just rocked the finish line of the Boston Marathon. Sirens galore. People running in fear. Wonder what happened.
This is crazy i seen that blow up #bostonmarathon
can someone tell me what that explosion was!? #boston #bostonmarathon

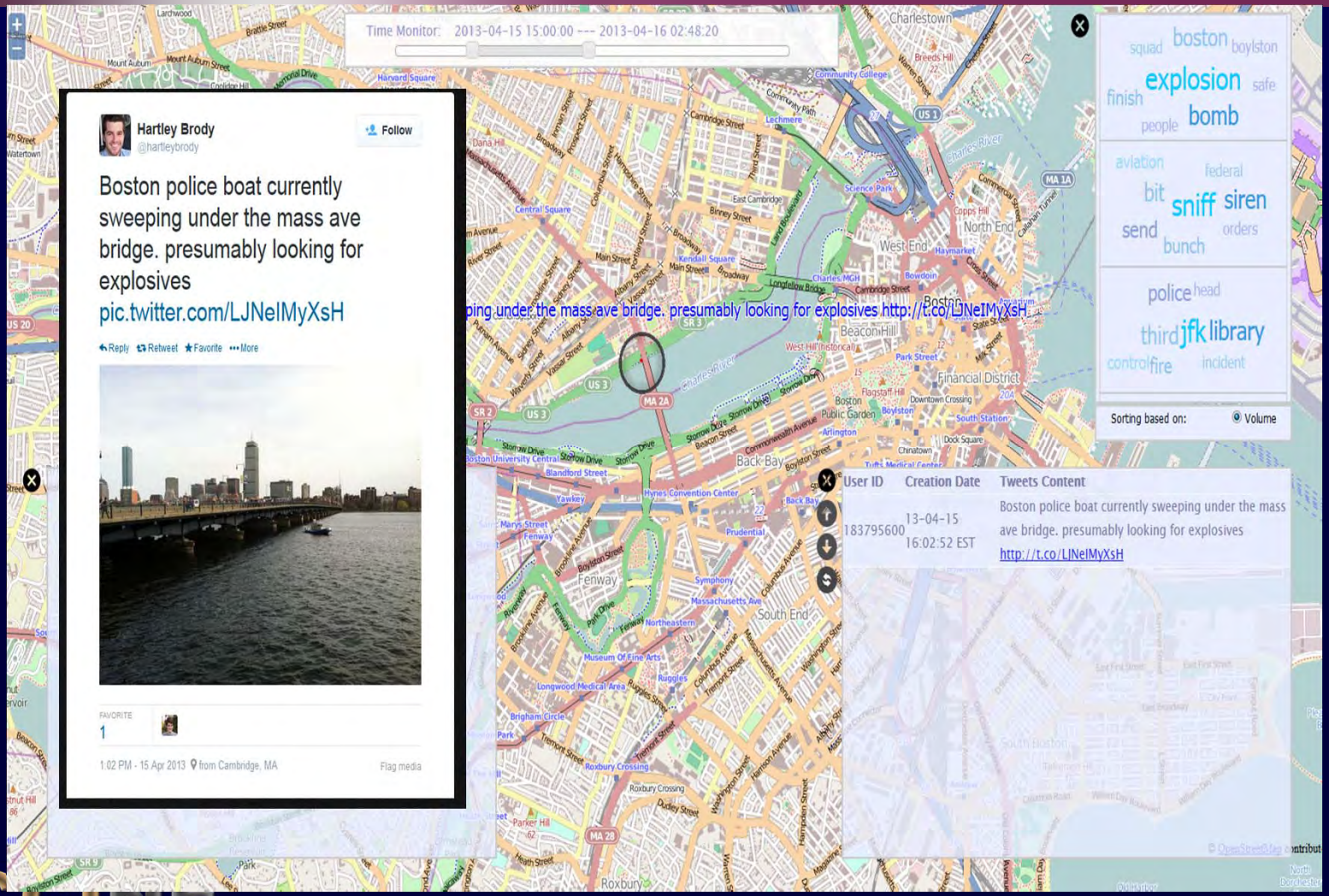
First Response (Tweet & Picture)



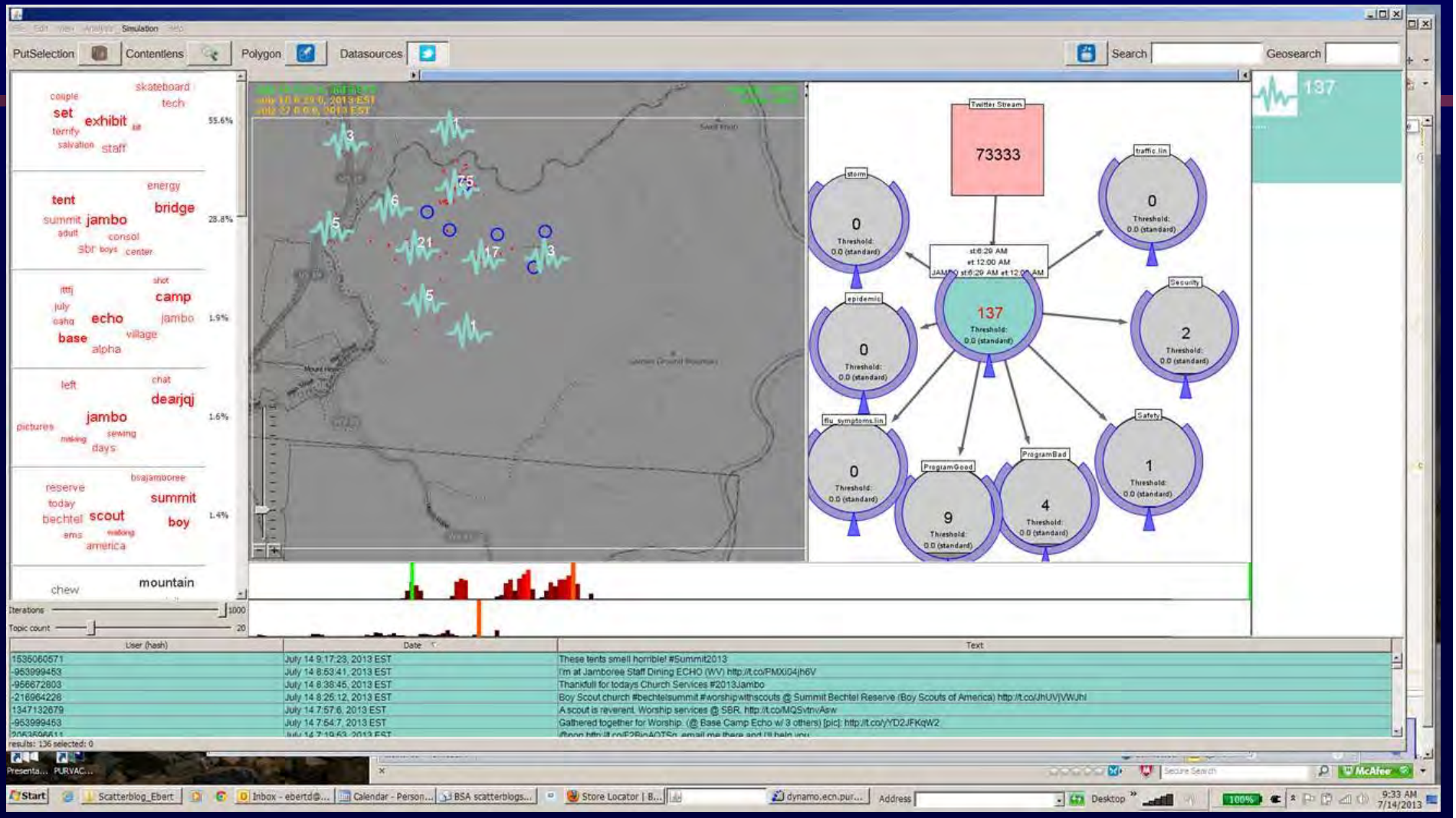
Police Patrolling The River



Boston Police Boat



BSA Jamboree

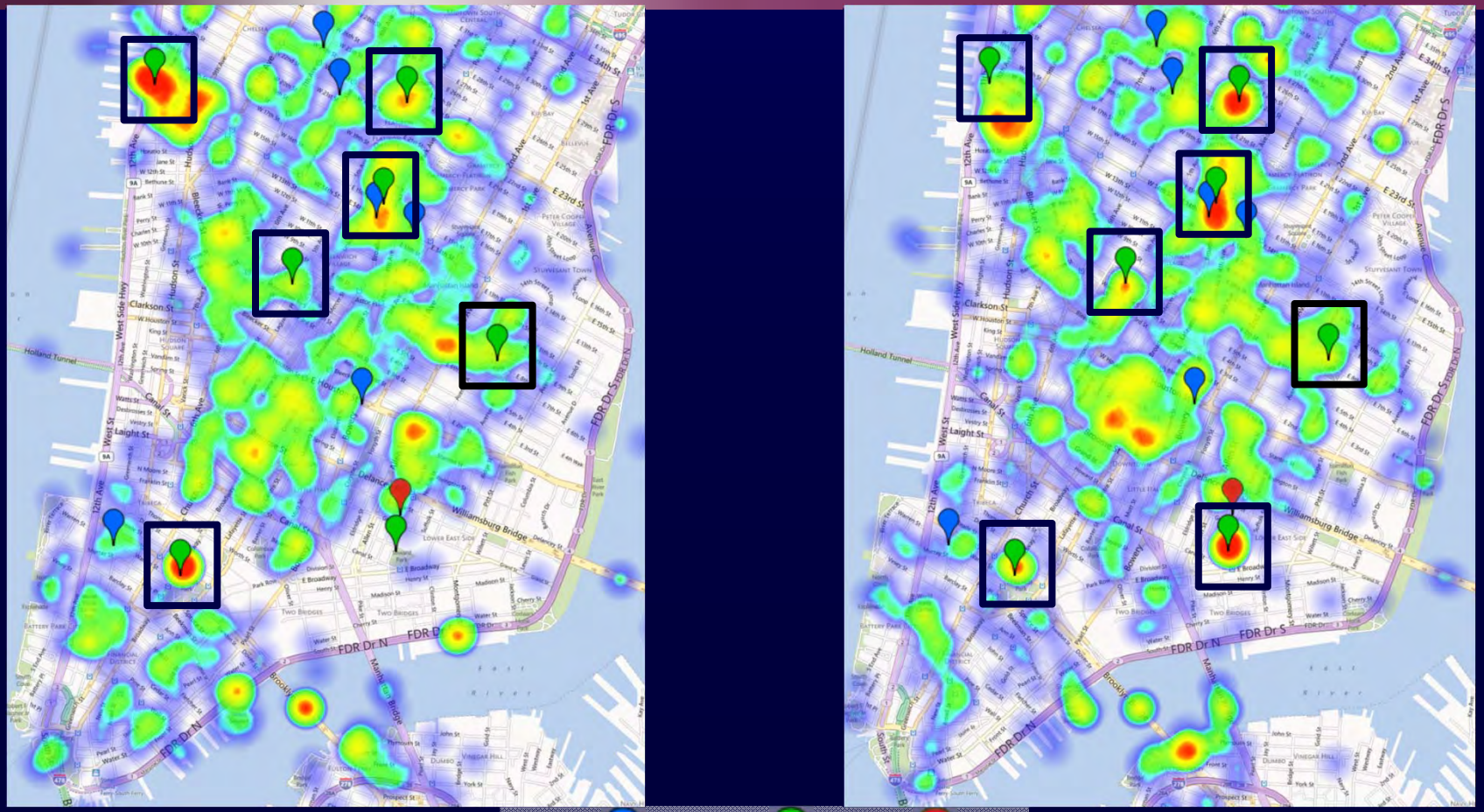


VACCINE

Visual Analytics of Activity During Hurricane Sandy

Two weeks before Sandy
10/14 (Sunday), 12:00 ~ 16:00

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00



VACCINE

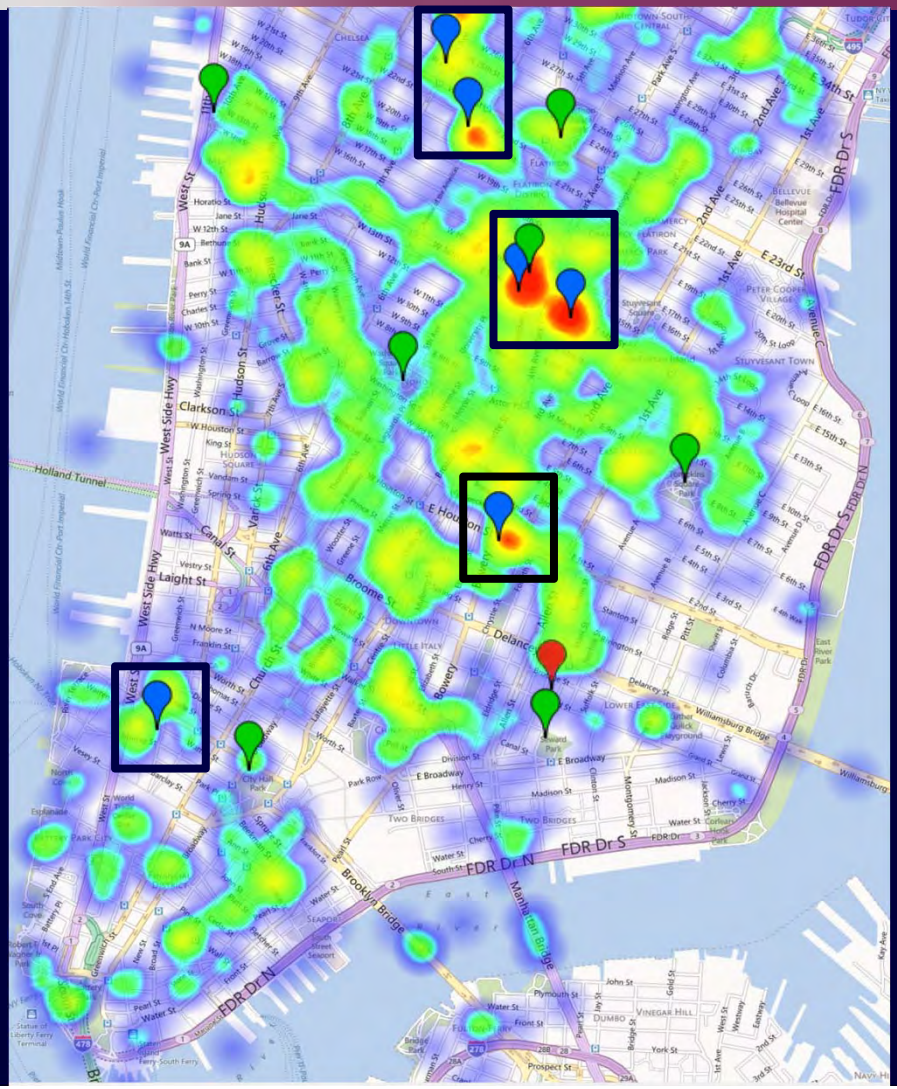
March 2014

Visual Analytics of Activity During Hurricane Sandy

After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00

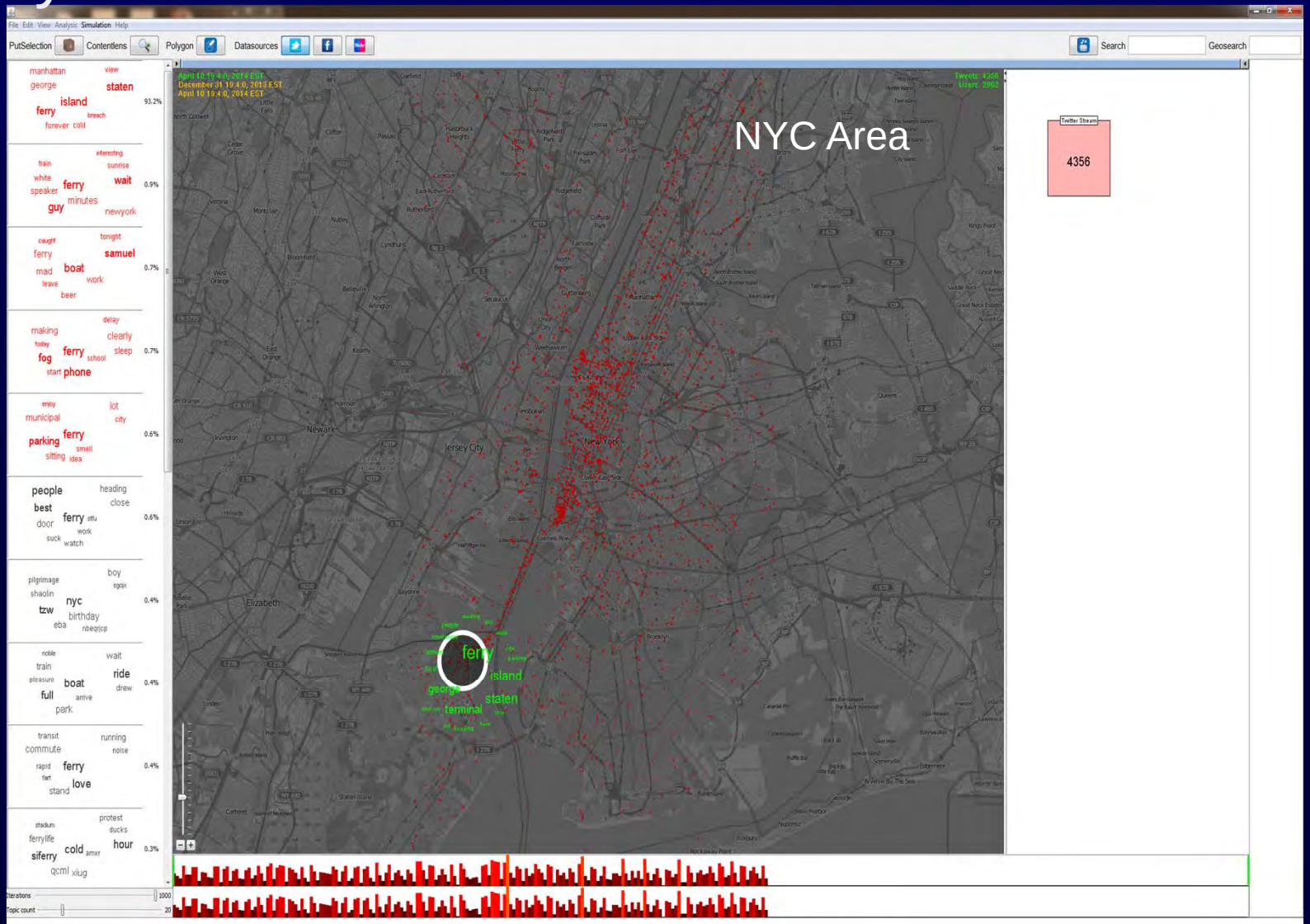
Evacuation order:
10/28, 10:30 AM

Hurricane Sandy's
Arrival at NYC:
10/29, 8:00 PM

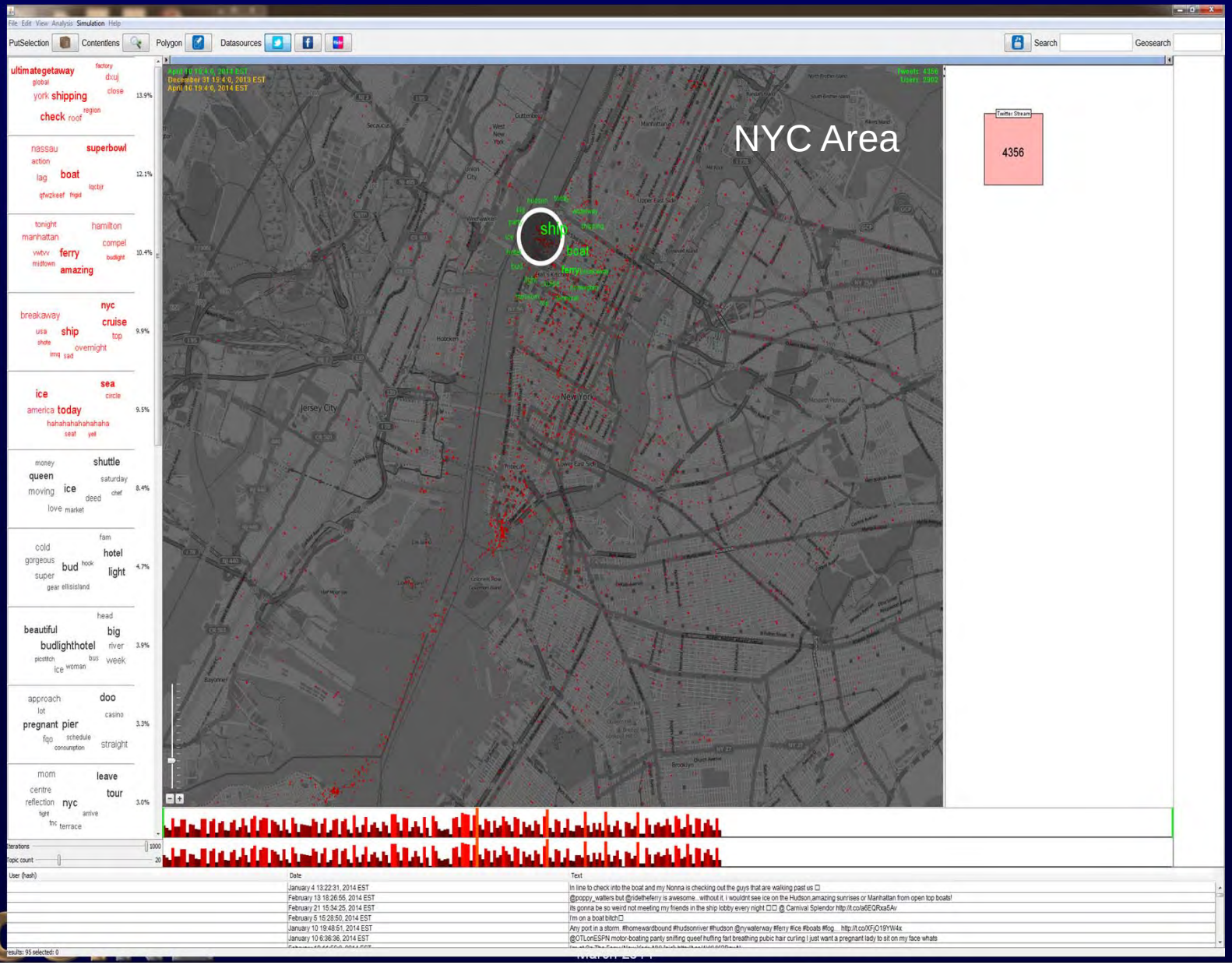


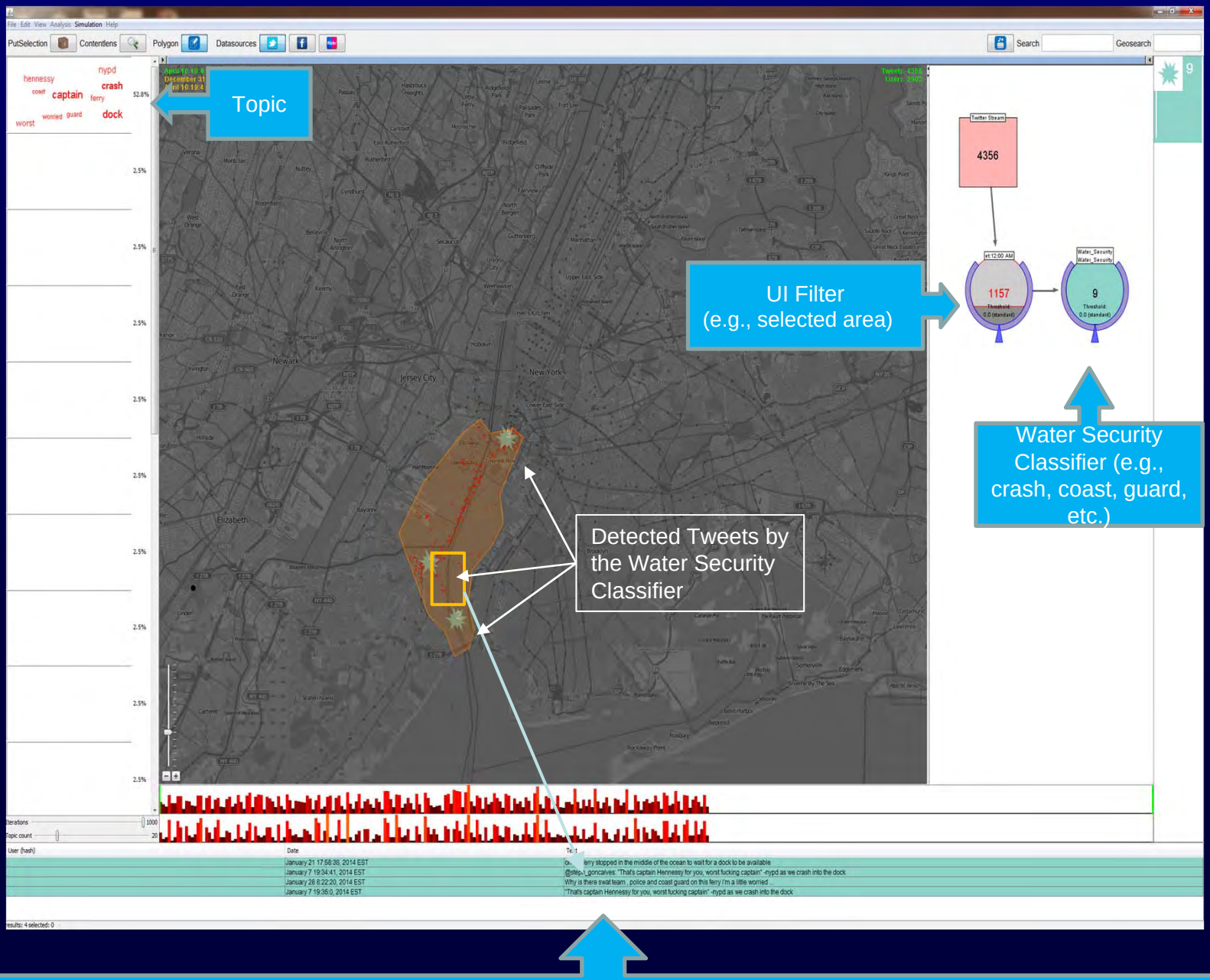
VACCINE

Twitter VA:
Ferry Traffic Between Staten Island And Manhattan



Two month Tweets containing water related keywords (e.g., ferry, ship, dock)


NYC Area
4356
PutSelection
Contentlens
Polygon
Datasources
Search
Geosearch
ultimategetaway
shipping
check
nassau
superbowl
boat
tonight
manhattan
hamilton
ferry
amazing
breakaway
nyc
cruise
ship
ice
sea
today
shuttle
queen
ice
cold
hotel
bud
light
beautiful
big
budlighthotel
approach
doo
pregnant
pier
mom
leave
tour
nyc
Jersey City
New York
Date
Text
January 4 13:22:31, 2014 EST
February 13 18:26:55, 2014 EST
February 21 15:34:25, 2014 EST
February 5 15:28:50, 2014 EST
January 10 19:48:51, 2014 EST
January 10 6:36:36, 2014 EST



oh the ferry stopped in the middle of the ocean to wait for a dock to be available
@steph_goncalves: "That's captain Hennessy for you, worst fucking captain" -nypd as we crash into the dock
Why is there swat team , police and coast guard on this ferry I'm a little worried ...
"That's captain Hennessy for you, worst fucking captain" -nypd as we crash into the dock

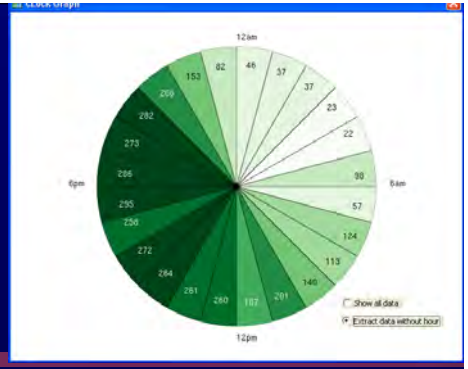
Example Projects: U.S. Coast Guard Visual Analytics

Effective Risk-based Decision Making and Resource Allocation Visual Analytics

- Evaluate current and historical mission area:
  - Demands
  - Risks (total, mitigated, residual)
  - Resource allocation
  - Return on investment
- Evaluate courses of action
- Evaluate above at both Strategic and Tactical/Operational level



Risk-Based Allocations



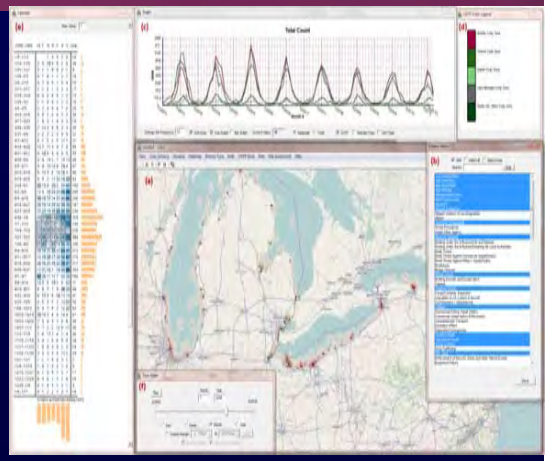
- Comparative visual analysis of mission cases/hours vs. staffing hours
- Comparative visualization of resources vs. risk
- Trend visual analytics
  - Increase/decrease in resource allocation
  - Increase/decrease in risk (total, mitigated, residual)
  - Increase/decrease in incidents
- Exploration of alternatives and effect on risk
- Predictive analytics based on historical data (STL and EWMA)

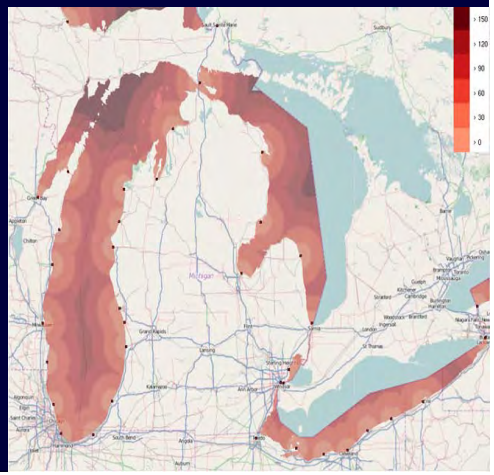
U.S. Coast Guard Search and Rescue VA (cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771

IMPACTS:

- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Hurricanes Sandy and Irene resource allocation decisions based on cgSARVA analysis and visualization
- Informed Commandant's budget testimony to Congress
- Key component of USCG D9 reallocation plan for 2011-12
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ

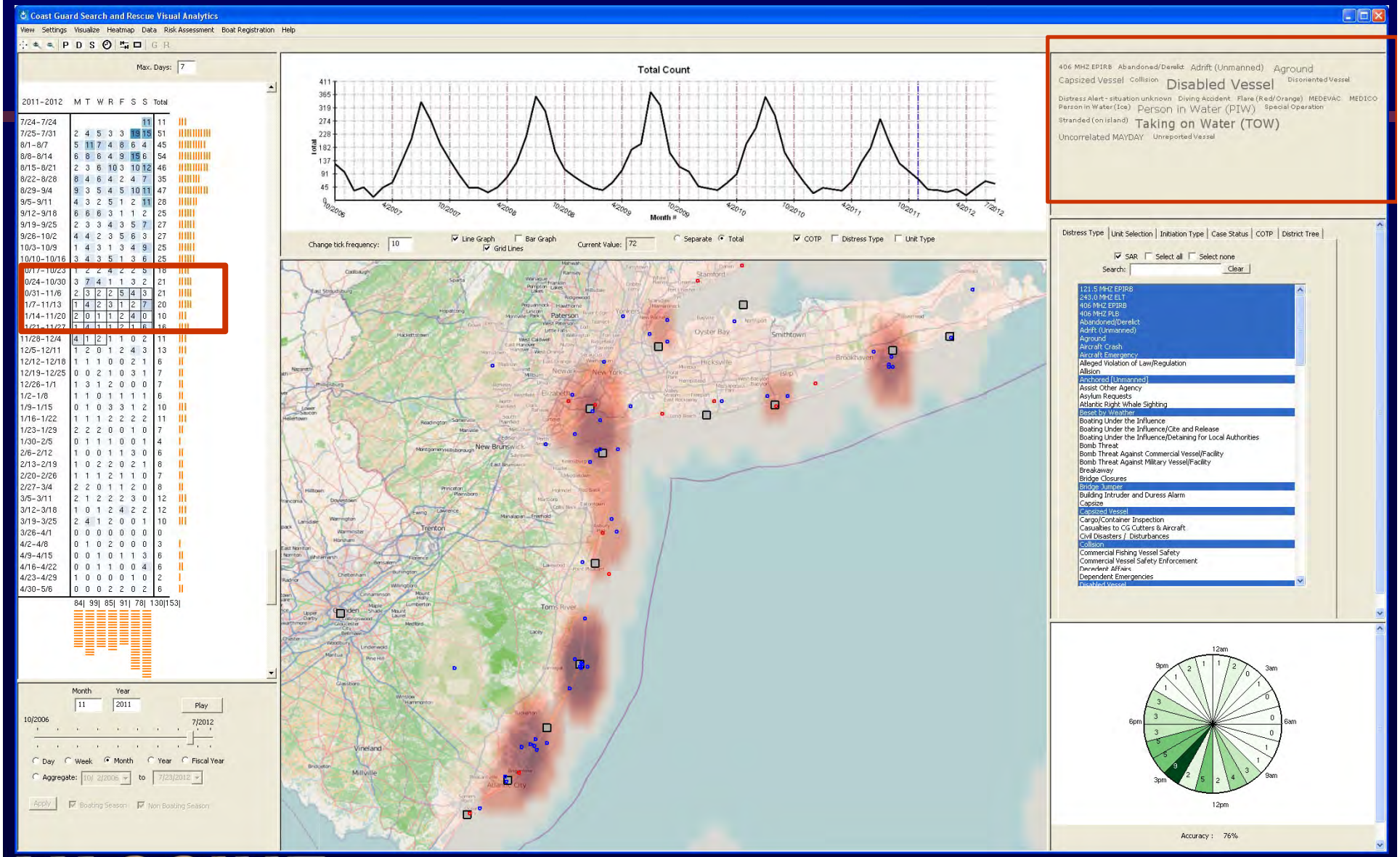




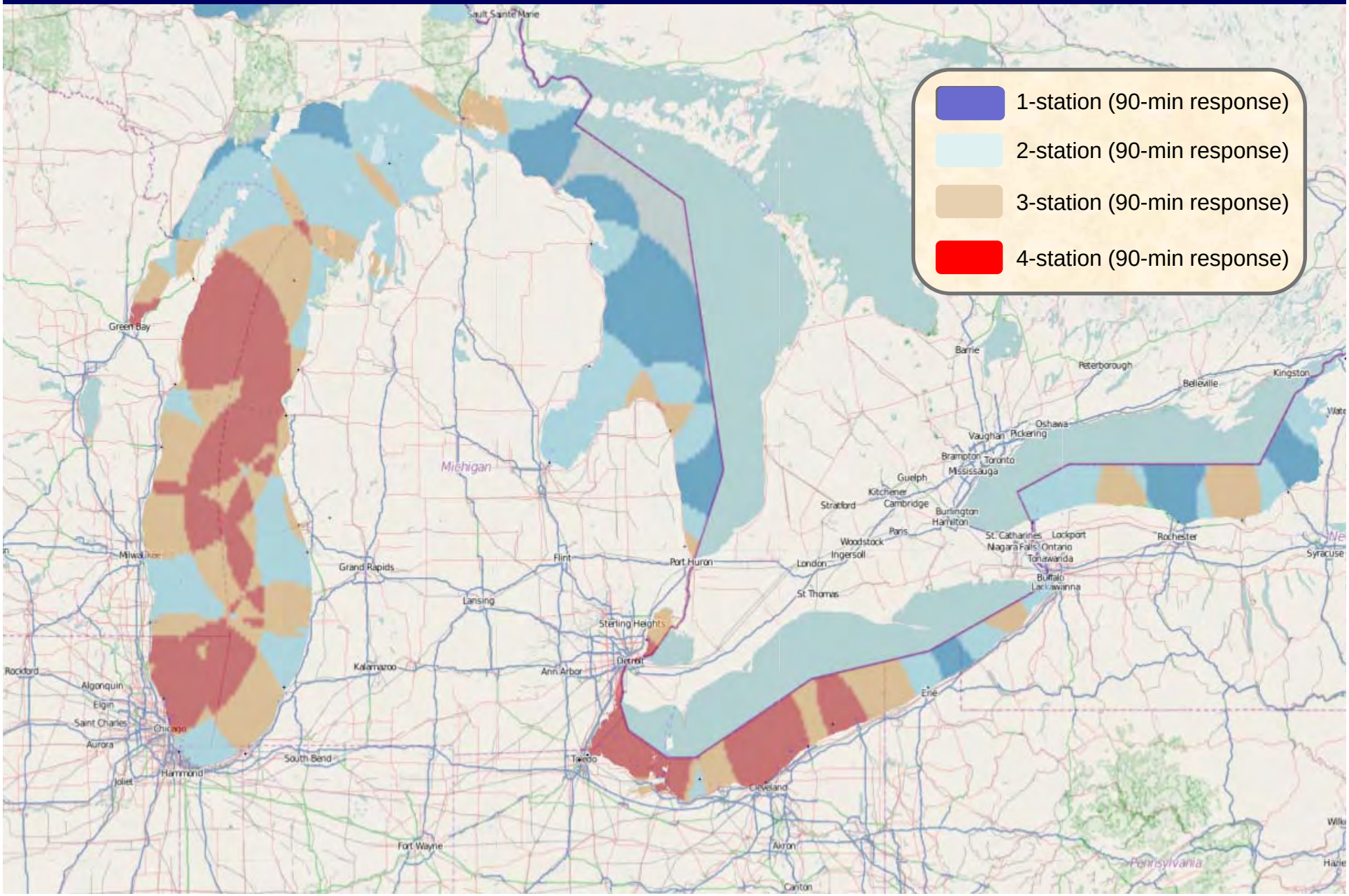
VACCINE

Example: Risks and Consequences From Sandy:

SAR Cases November 2011 NJ/NYC Area



Response Efficiency – Potential Future Assets



Software Accredited for Decision Making

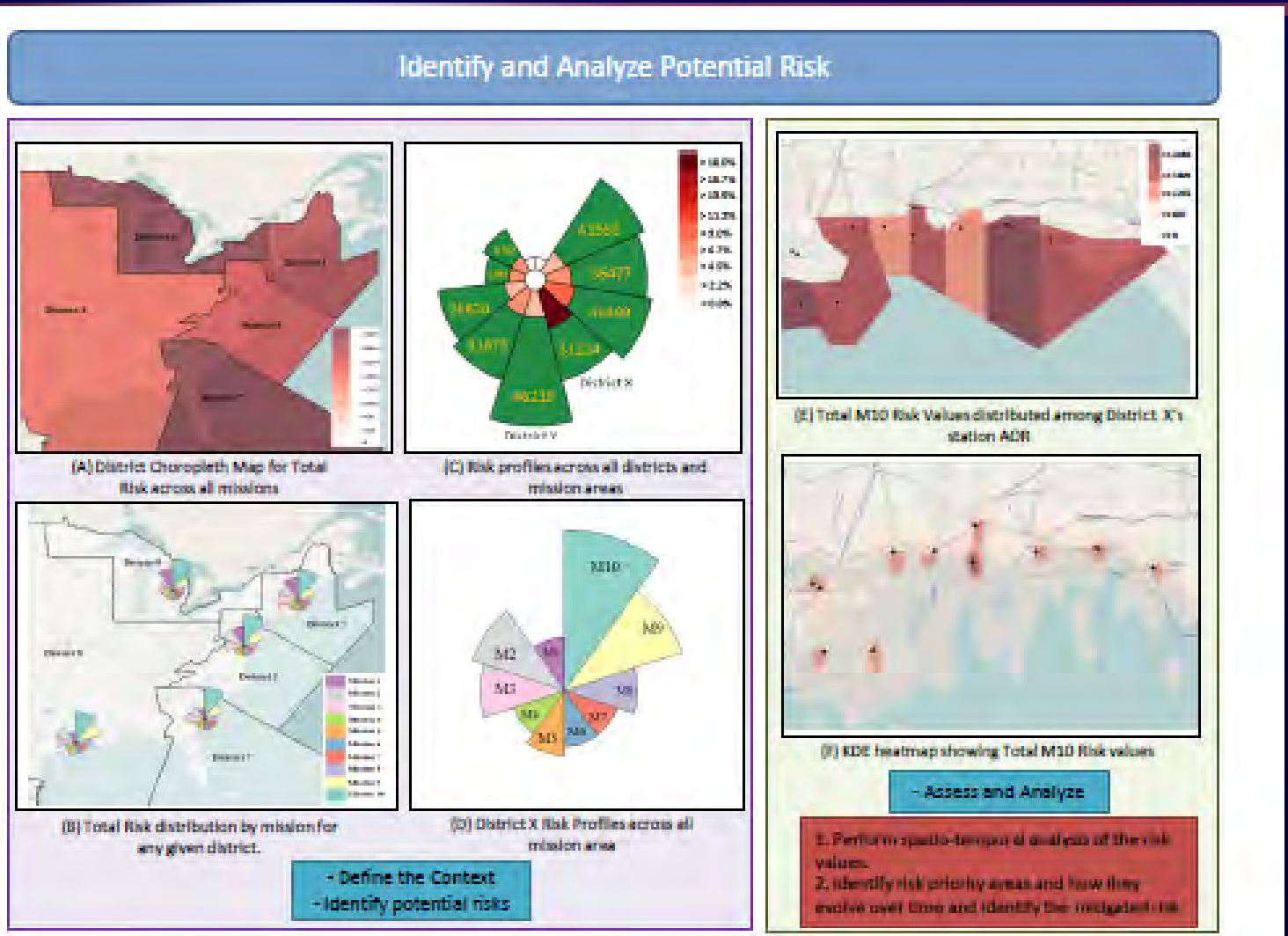
- April 22, 2013 cgSARVA VV&A'd for US Coast Guard system-wide use



Interactive Operational Performance Assessment Report (iOPAR)

- OPAR
  - Standard report within the Coast Guard Business Intelligence (CGBI) system
  - Displays resource use and performance by core CG mission areas
- iOPAR – interactive iPad version to allow interactive visualization and analysis and inform decision making
  - Anaylze/visualize performance, targets, seasonal trends, predictions by boat, aircraft, cutter type, region levels

Multi-Visualization, Multivariate VA for Risk-Based Decision Making



USCG Port Closure Economic Impact VA

Partners: USC CREATE, USCG RDC, USCG D7, USCG LANT

IMPACT:

- Provided tool for use analysis and planning for impact of port closure in Port Arthur, Tx
- Economic sector impact, local and national impact
- Impact and effectiveness of alternative mitigation strategies



VACCINE

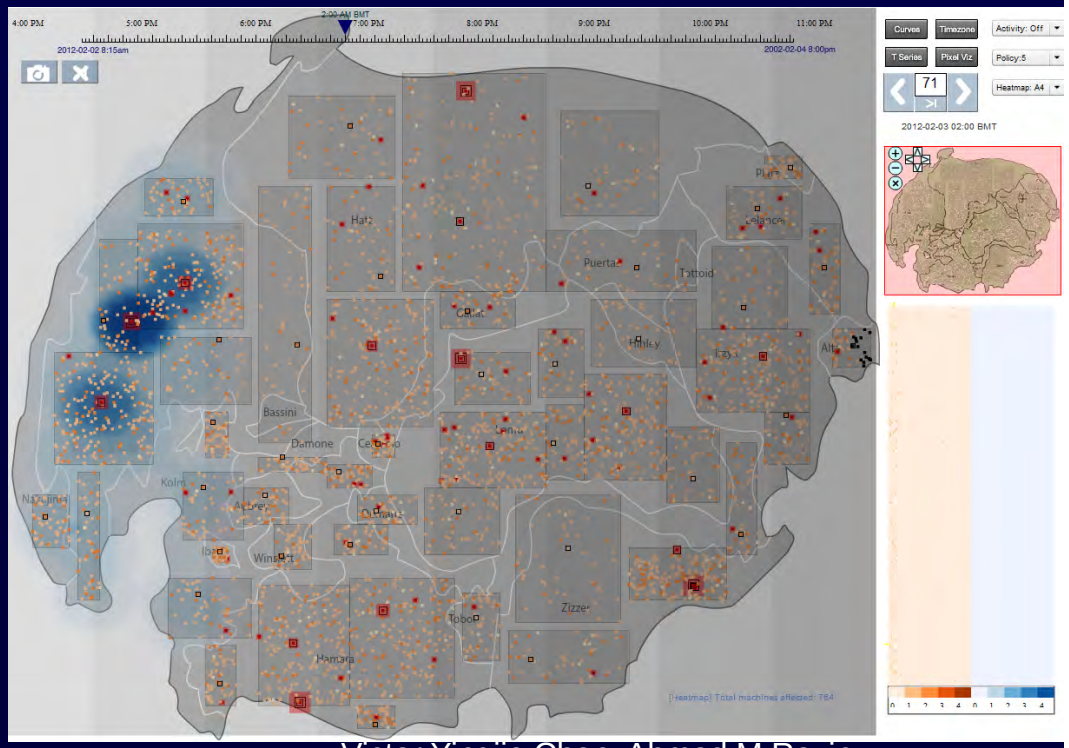
Example Projects:
Cybersecurity Visual Analytics

SemanticPrism: A Multi-aspect View of Large High-dimensional Data (Purdue University)

- **VAST 2012 Mini Challenge 1 Award: Outstanding Integrated Analysis and Visualization**

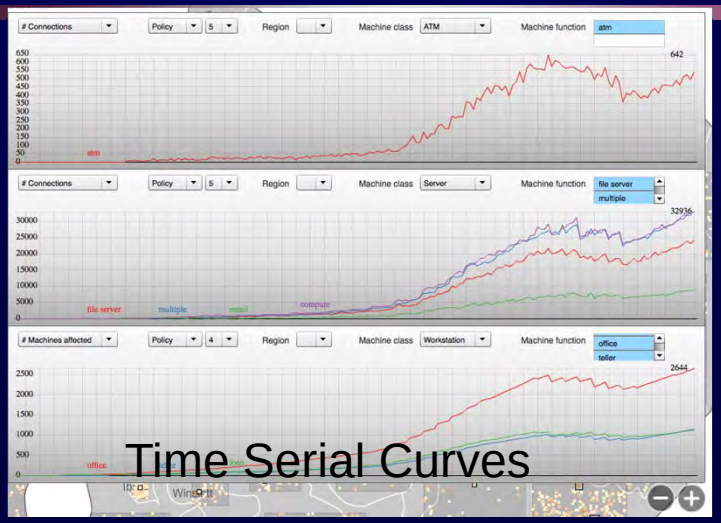
- **Geo-Temporal**
- **Time-serial**
- **Pixel-based**

- **Semantic Zoom**

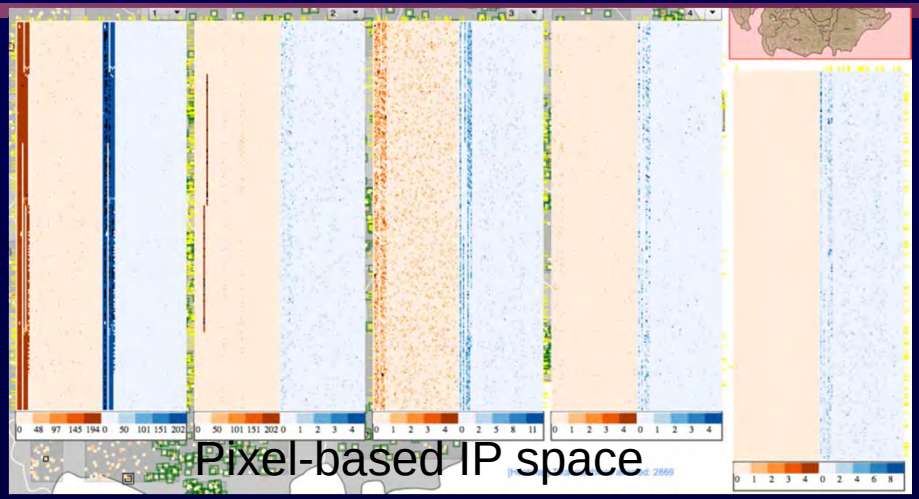


Victor Yingjie Chen, Ahmad M Razip, Sungahn Ko, Cheryl Zhenyu Qian, David S.Ebert

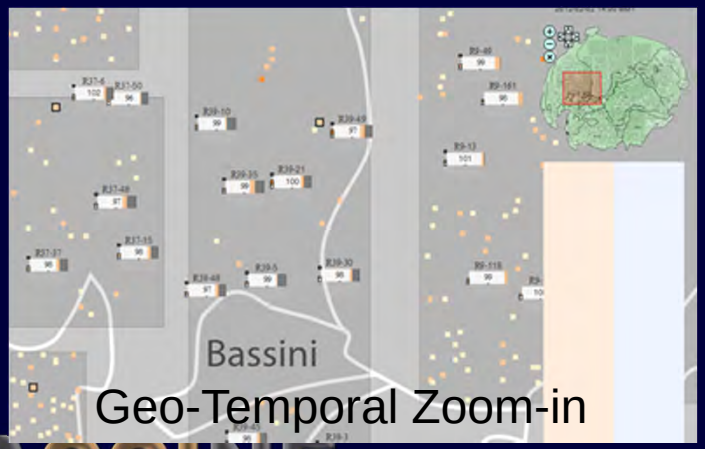
SemanticPrism



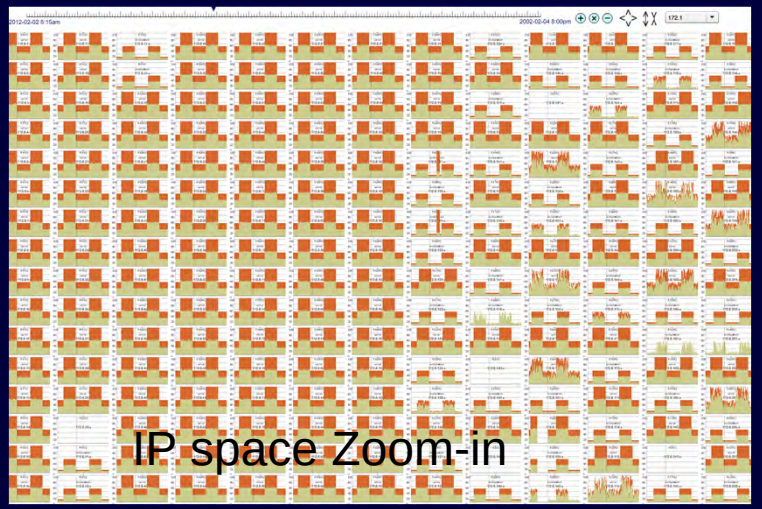
Time Serial Curves



Pixel-based IP space

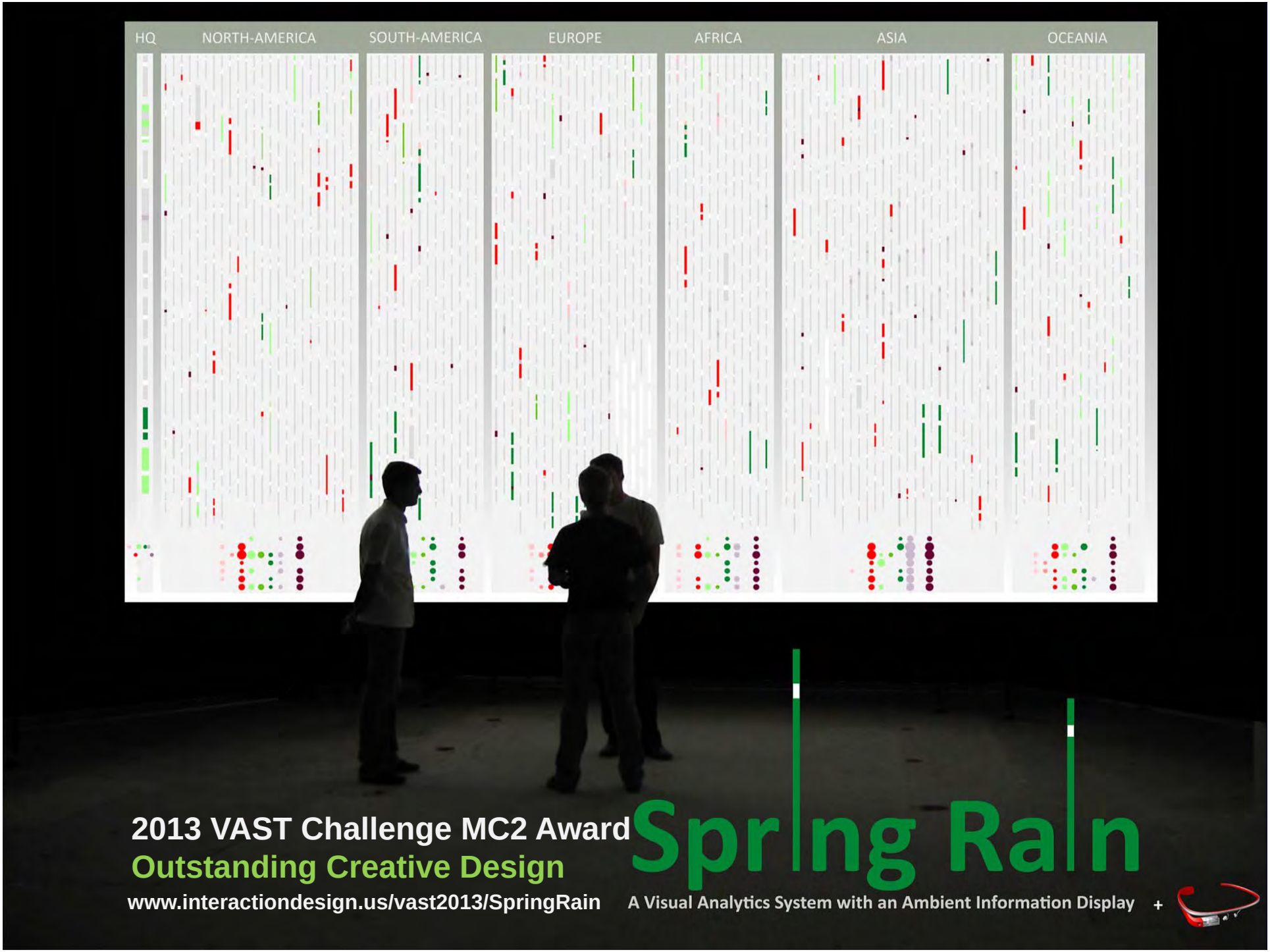


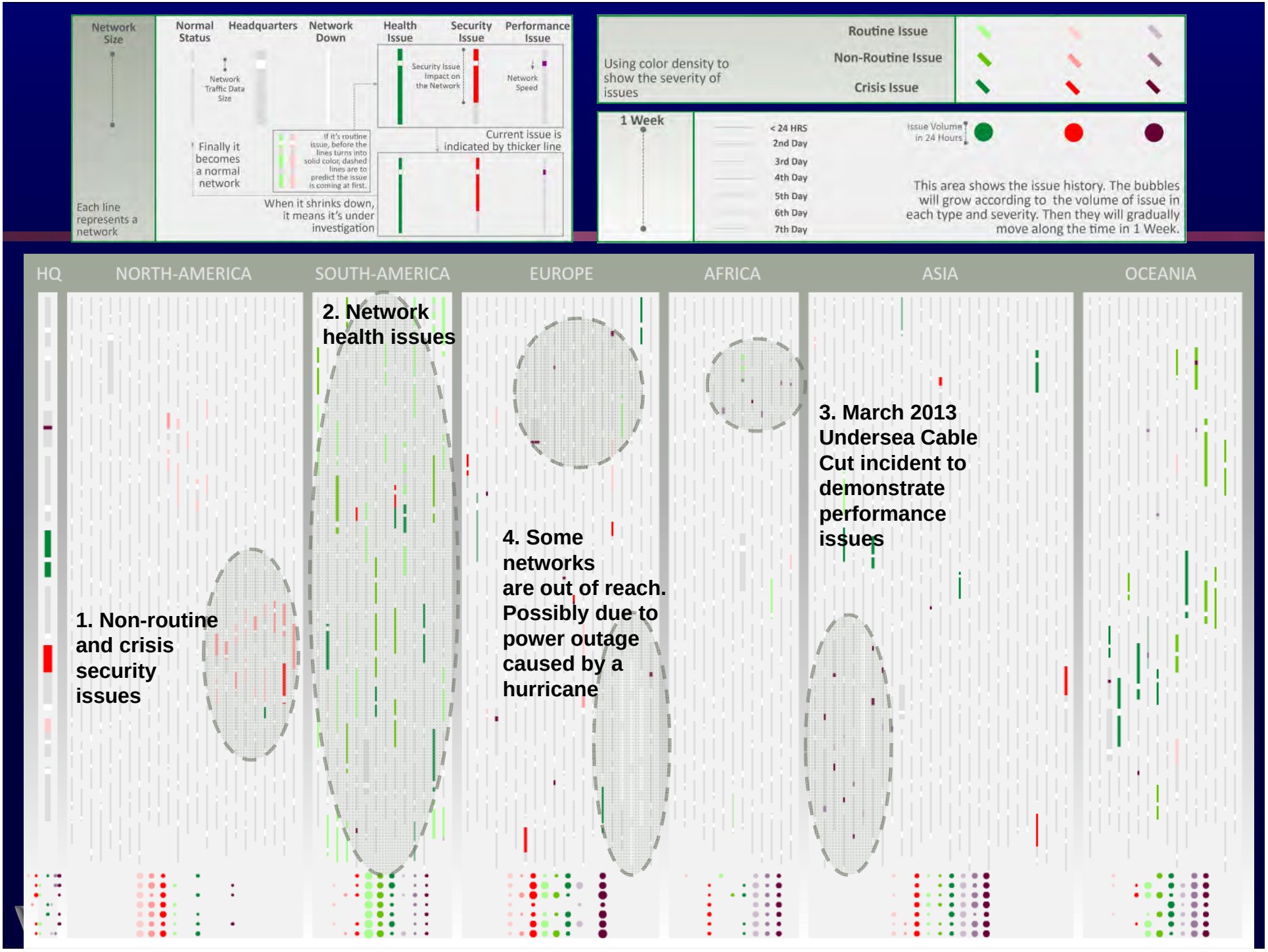
Geo-Temporal Zoom-in



IP space Zoom-in

VACCINE


HQ
NORTH-AMERICA
SOUTH-AMERICA
EUROPE
AFRICA
ASIA
OCEANIA
2013 VAST Challenge MC2 Award
Outstanding Creative Design
www.interactiondesign.us/vast2013/SpringRain
Spring Rain
A Visual Analytics System with an Ambient Information Display +


Network Size
Each line represents a network
Normal Status
Headquarters
Network Down
Health Issue
Security Issue
Performance Issue
Network Traffic Data Size
Security Issue Impact on the Network
Network Speed
Finally it becomes a normal network
If it's routine issue, before the lines turns into solid color, dashed lines are to predict the issue is coming at first.
When it shrinks down, it means it's under investigation
Current issue is indicated by thicker line
Using color density to show the severity of issues
Routine Issue
Non-Routine Issue
Crisis Issue
1 Week
< 24 HRS
2nd Day
3rd Day
4th Day
5th Day
6th Day
7th Day
Issue Volume in 24 Hours
This area shows the issue history. The bubbles will grow according to the volume of issue in each type and severity. Then they will gradually move along the time in 1 Week.
HQ
NORTH-AMERICA
SOUTH-AMERICA
EUROPE
AFRICA
ASIA
OCEANIA
1. Non-routine and crisis security issues
2. Network health issues
3. March 2013 Undersea Cable Cut incident to demonstrate performance issues
4. Some networks are out of reach. Possibly due to power outage caused by a hurricane

Directions Forward, Keys to Success & Challenges



Six Challenges for Proactive & Predictive Visual Analytics

1. Creating computer-human visual cognition environments
2. Integrating interactive simulation and analytics
3. Solving specific scale issues and cross-scale issues
4. Managing uncertainty and time
5. Enabling risk-based decision making environments
6. Developing the Science of Interaction for Visual Analytics

Visual Analytics: Remember…

- We need to be cognizant of parameters for visual representations
- Appropriate analysis can guide users to interesting features in the data
- Refined analysis through user interaction and their domain knowledge can help discover hidden problems
- There is no single catch-all visual representation or analysis

Keys for Success

- User and problem driven
- Balance human cognition and automated analysis and modeling
  - Often applied on-the-fly for specific components identified by the user
- Interactivity and easy interaction
  - Utilizing HPC and novel analysis approaches
- Understandability of why predicted value is what it is
- Intuitive visual cognition
- Not overloaded with features

For Further Information

www.VisualAnalytics-CCI.org



VACCINE

Visual Analytics for Command, Control, and Interoperability Environments

A U.S. Department of Homeland Security Center of Excellence

Visual Analytics: The Right Solution for Big Data and Global Challenges

David S. Ebert

CVADA

Center for Visual Analytics and Data Analytics

A U.S. Department of Homeland Security Center of Excellence

September 2013

Overview

- Big Data challenges
- Visual analytics
- VACCINE center
- Example projects
- Path forward



Big Data: Solution to Global Challenges?



- **Big Data Definition:** A phenomenon defined by the rapid acceleration in the expanding volume of high velocity, complex and diverse types of data. (TechAmerica Foundation)
  Big Data is often defined along three dimensions:
  - **Volume** – size
  - **Velocity** – rate of input, update, change
  - **Variety** – different types, sources, variables
- Need:
  - Advanced techniques and technologies to enable the capture, storage, distribution, management and analysis of information (TechAmerica Foundation)
  - **Enable effective, efficient analysis, decision making, planning, and action**

Big Data Challenges

- Heterogeneity
- Unprecedented volume and/or complexity
- Multivariate, Multidimensional
- Streaming
- Qualitative, Quantitative, Categorical, Temporal,…
- Distributed Sources: sensors, text, images, maps, …
- Inconsistent and Uncertain
- Different Scales and Cross Scale

Big Data Analytics – The Solution?

- Advanced data analytics for big data can help manage big data
  - Machine learning
  - Advanced statistical analysis
  - Hadoop-deployed analytics
- However, it falls short in several aspects

What Big Data Analytics Can't Do
(inspired and adapted from David Brooks, New York Times, 2/18/2013)

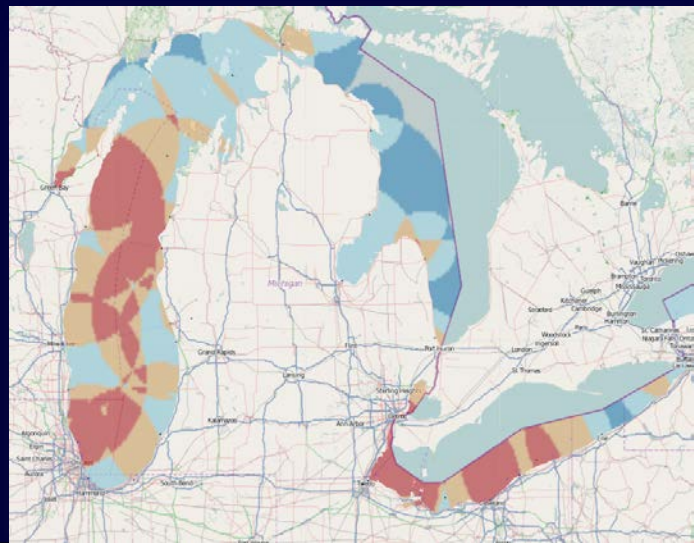
- Qualitative, fuzzy, and social data
  - Preferences, significance of one relationship over another
  - Trust
- Context
  - Data is rarely complete, nor does it incorporate all relevant information for decision making
  - Humans have extensive information and experience that never make it into the collected data

What Big Data Analytics Can't Do
(inspired and adapted from David Brooks, New York Times, 2/18/2013)

- Spurious vs. Significant
  - Big data means more statistically significant events and correlations, but they may not have any relevance
  - Increases noise to signal ratio
- Big problems
  - Complex, multifaceted, multiparameter big challenges with unquantified dependencies

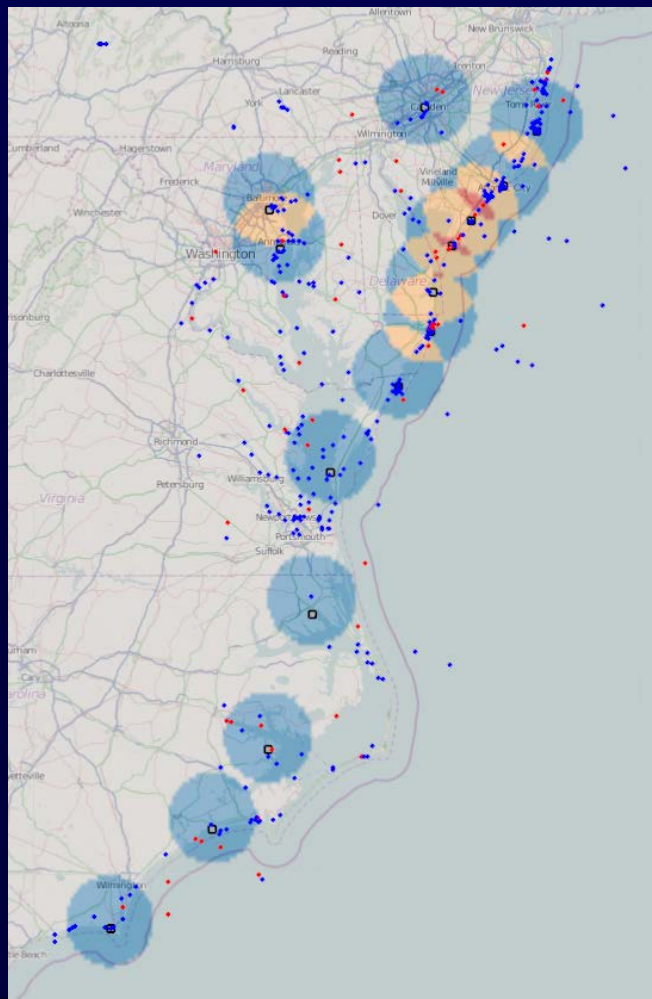
The Power of Visualization

- Turn the abstract into the concrete
- Enable rapid perception and visual cognition of information to increase efficient understanding and decision making
- Goal: Transform *data* into *information and knowledge*
- Data size doesn't matter,
  size of insights matters
- Visualization is a key method to increase understandability of data and analytical results



Visualization, Data, and Decision Making

- At some level most decision making is driven by data
- What role does visual representation play?
- Good or bad – people believe charts and graphics – “seeing is believing”



Beyond Visualization to Visual Analytics

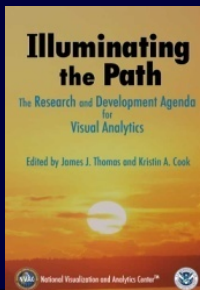
Visual Analytics[1] is the science of analytical reasoning facilitated by interactive visual interfaces

Interactive visualization, data analysis, exploration, and decision making with human in the loop!

People use visual analytics tools and techniques to

- Synthesize information and derive insight from massive, dynamic, ambiguous, and often conflicting data
- Detect the expected and discover the unexpected
- Provide timely, defensible, understandable assessments
- Communicate assessment effectively for action

1. *Illuminating the Path: The R&D Agenda for Visual Analytics, Editors: Thomas and Cook*



VA Approaches to Big Data

- Don’t display all the data!
- Extract relevant information at the appropriate natural scales using analytics and statistical models to reduce data space so that it is mappable to the visual space
- User interaction to guide above process
- Adapt to available hardware for the problem
  - Should we use big iron or cell phones?


Innovation with Impact
SFU, JIBC
UBC
UW
Purdue
VACCINE
Penn St.
Ind U
VaTech
Stanford
NC A&T
UNCC
Navajo Tech
JSU
GaTech
U Stuttgart
UT Austin
UHD
Swansea U
FIU
Selected Region Data
GeoDiscoverer
VACCINE
PURDUE

Who We Are:
International Team of Experts
75+ Faculty, 26 Institutions



- **Purdue University**
- **Georgia Institute of Technology**
- **Pennsylvania State University**
- **Stanford University**
- **University of North Carolina at Charlotte**
- **University of Washington**
- **Arizona State University**
- **Simon Fraser University**
- **University of British Columbia**
- **Justice Institute of British Columbia**
- **Ontario Institute of Technology**
- **Dalhousie University**
- **University of Houston, Downtown**
- **Virginia Tech**
- **Indiana University**
- **Florida International University**
- **University of Texas at Austin**
- **Morgan State University**
- **Navajo Technical College**
- **University of Stuttgart**
- **University of Swansea**
- **Oxford University**
- **University of Calgary**
- **University of Manitoba**
- **Carleton University**
- **University of Victoria**

VACCINE

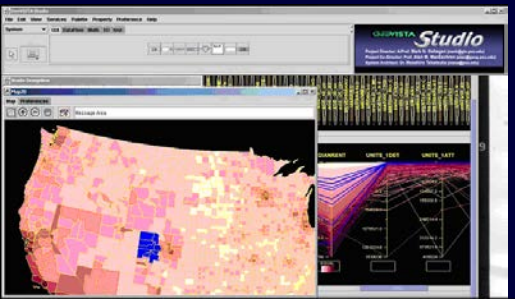
VACCINE's Role



Problem: To solve current and future homeland security problems requires exploring, analyzing, and reasoning with massive, multi-source, multi-scale, heterogeneous, streaming data –**BIG DATA**

- Cuts across entire spectrum of homeland security needs

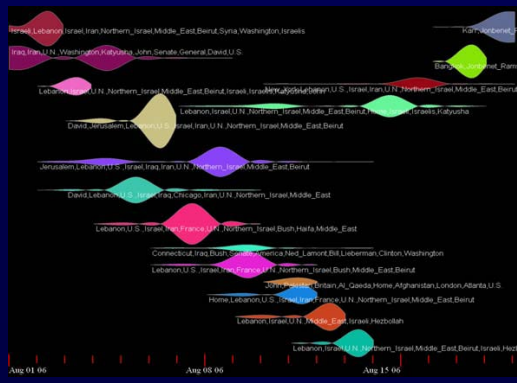
We provide tools to enable end users to get the relevant information they need during any situation to make a decision or take action





VACCINE Mission





- Provide visual analytic and scalable solutions to 2.3 million extended homeland security personnel
  - 185,000 DHS personnel, 350,000 law enforcement personnel, 750,000 homeland security practitioners
- Achieve excellence in visual analytics and visualization sciences
- Educate homeland security stakeholders and the next generation of talent



VACCINE Value

Our Value / Solution: Enable users to be more effective through innovative interactive visualization, analysis, and decision making tools

- Provide the right information, in the right format within the right time to solve the problem
- Turn data deluge into a pool of relevant, actionable knowledge
- Enable users to be more effective from planning to detection to response to recovery
- Enable effective communication of information

Approach: Partner-driven solutions and research

VACCINE Value Part II



Our people and partnerships

- Interdisciplinary world-leading team of researchers
- Defining and extending the new science of visual analytics driven by real-world, real-scale problems of engaged partners (local, state, federal)







VACCINE Research Themes To Solve Real World Problems

- **Visual Analytics** – integrated, interactive visual exploration, analysis, and decision making environments to enable effective decisions from massive, time-evolving multimedia data
- **Interactive scalable analytics techniques** – statistical, spatiotemporal, image, video, signal, machine learning
- **Science of Interaction for visual analytics**
- **Evaluation** in real world environments

Visual Analytic Solutions: What We Offer

- **Improved Effectiveness**: We enable users to be more effective through innovative, interactive visualization, analysis, and decision making tools
  - Provide the right information, in the right format, within the right time to solve the problem
  - Enable user to be more effective from planning to detection to response to recovery
  - Enable effective communication of information
- **Innovative Fielded Solutions:** We provide innovative visual analytic and scalable solutions to the extended homeland security community
- **People and Partnerships**
  - Interdisciplinary world-leading team of researchers and students
  - Actively Engaged Partners –
    We define and extend the new science of visual analytics driven by real-world, real-scale problems of engaged partners (local, state, federal)

"cgSARVA has proven its worth time and again, providing key analytic information for decision makers for large scale projects…"
VADM Robert Parker, 2012 MRS Keynote Address

VADM Robert Parker with VACCINE student researchers (cgSARVA, COAST, iOPAR)



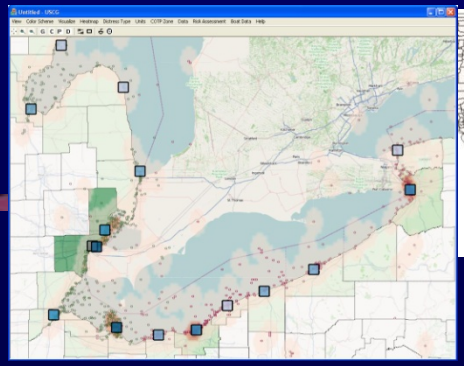
Engaged End-Users

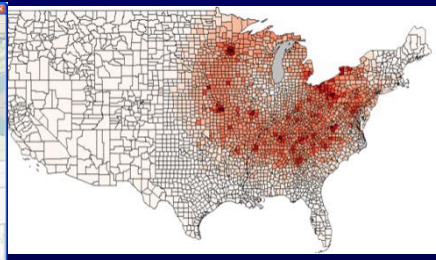
- Federal Operating Components:
  - US Coast Guard
  - US Transportation Security Agency
  - US Customs and Immigration Service
  - US Federal Emergency Management Agency
  - US Customs and Border Protection
  - US CERT
  - US ICE (in progress)
- Law Enforcement
  - Over 40 local and state agencies (IN, IL, OH, PA, NC, NY)
- Fusion Centers
  - Ohio (SAIC)
  - Indiana (IIFC)

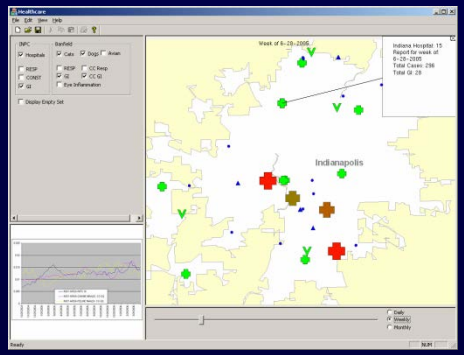


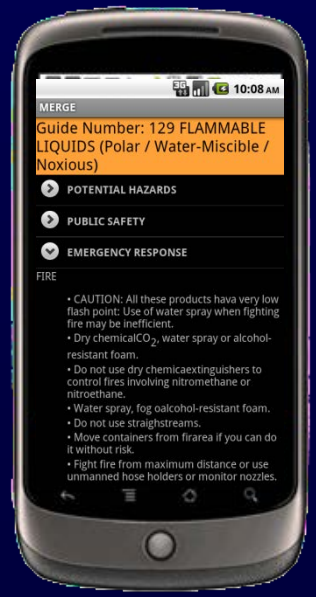
Public Safety & Health

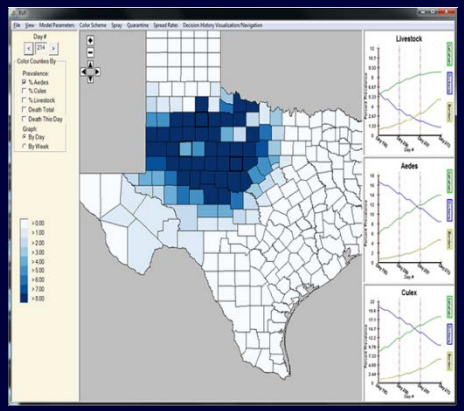
- Public safety visual analytics
  - cgSARVA
    - Coast Guard Search And Rescue Visual Analytics
  - MERGE
    - Mobile Emergency Response Guide
- Public health visual analytics
  - LAHVA
    - Linked Animal-Human Health Visual Analytics
  - RVF
    - Rift Valley Fever
    - Decision support environment for epidemic modeling and responses
  - PanVis
    - Pandemic influenza modeling and visualization tool
  - Cancer Care Engineering









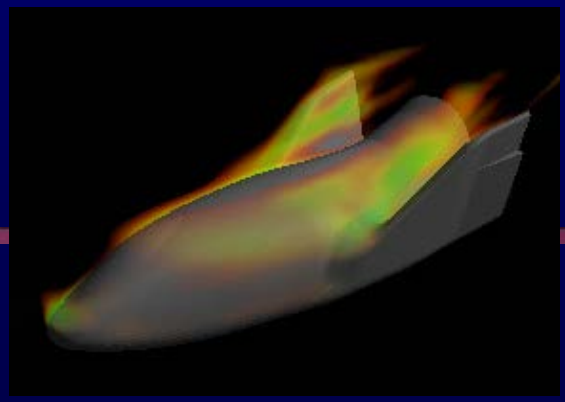


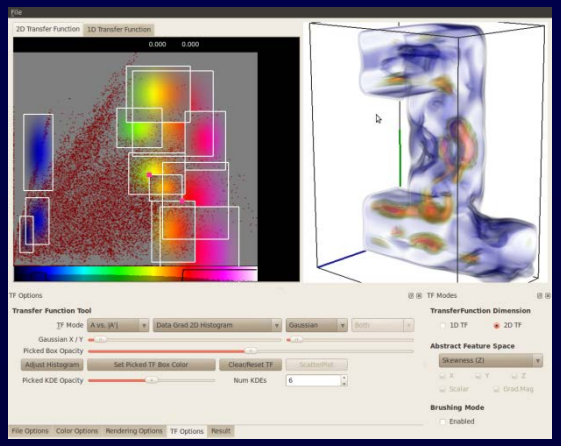


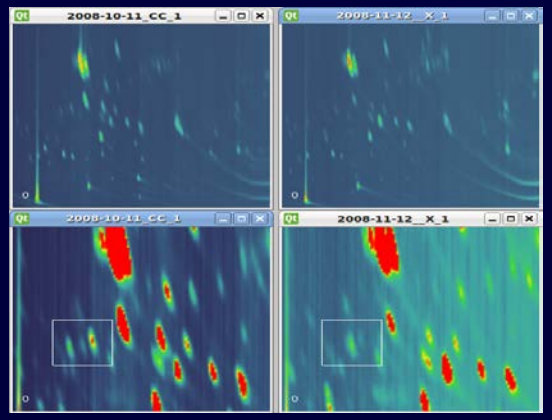
VACCINE

Scientific Discovery

- Flow dynamics visualization
  - Providing insights on large flow data
  - Visualization linked with simulations
- Nanohub
  - Information-assisted data analysis and visualization of nanoelectronic models
- Cancer biomarker
  - Visual analysis suite for exploring multiple samples of data from cancer care engineering



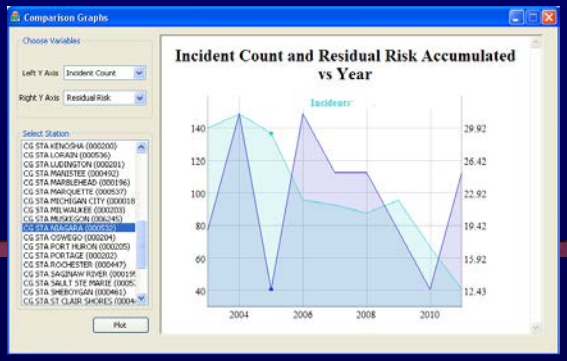


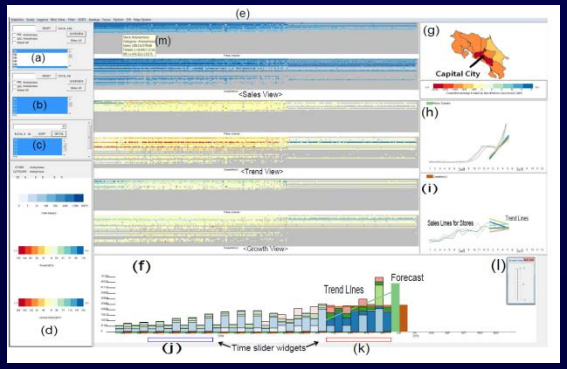


VACCINE

Business Visual Analytics

- Risk-based decision making and resource allocation
  - Coast guard operational risk assessment model
  - Helping to prioritize efforts to minimize risk
- Competitive Intelligence
  - Visual analytics system for business intelligence
- EconVIS
  - Visual analytics in various economic problems
  - Improving decision making and identifying key motivations in knowledge creation

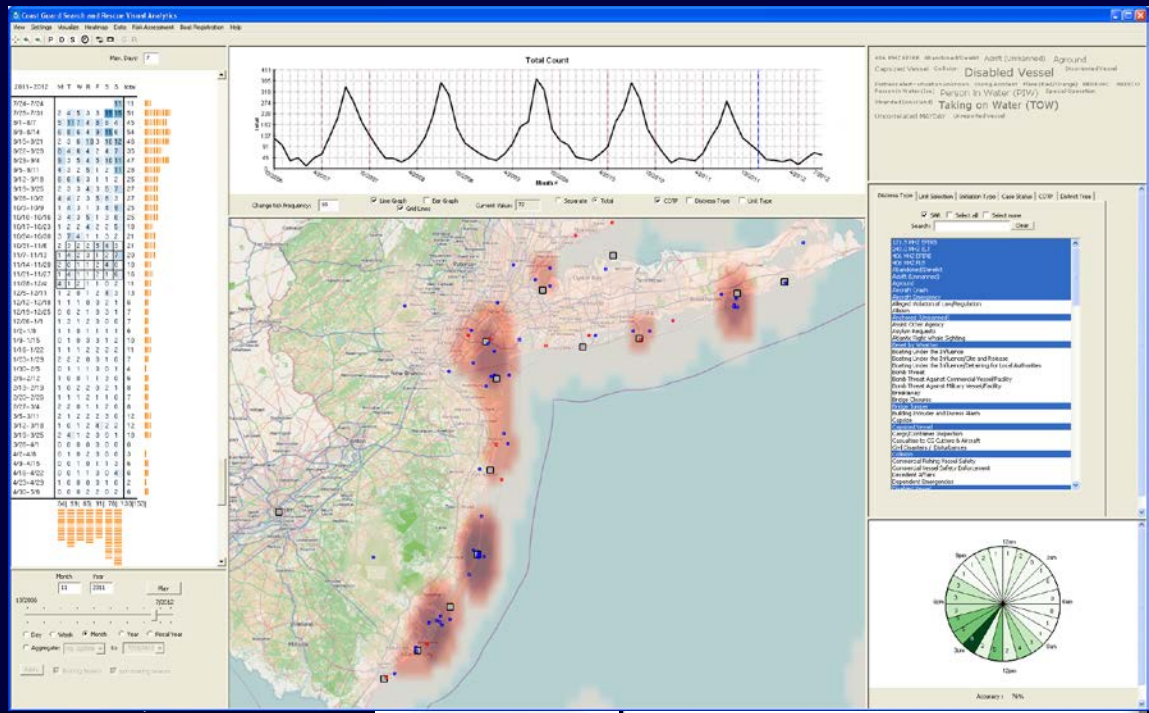






Example Projects

- VASA
- TSA
- Public Safety and Law Enforcement
- US Coast Guard



VACCINE

VASA: Visual Analytics for Security Applications

Cascading Critical Infrastructure Resiliency Modeling and Analytics



- **Purpose:** Apply visual analytics to the problem of monitoring and understanding cyber networks and critical infrastructures during detrimental cascading effects, and to the management of the ensuing crisis response.

- **Collaborating Institution(s):**
  Purdue, UNC Charlotte, U. Minn. (NCFPD), U. Konstanz, U. Stuttgart, Fraunhofer IGD, Siemens, German utilities

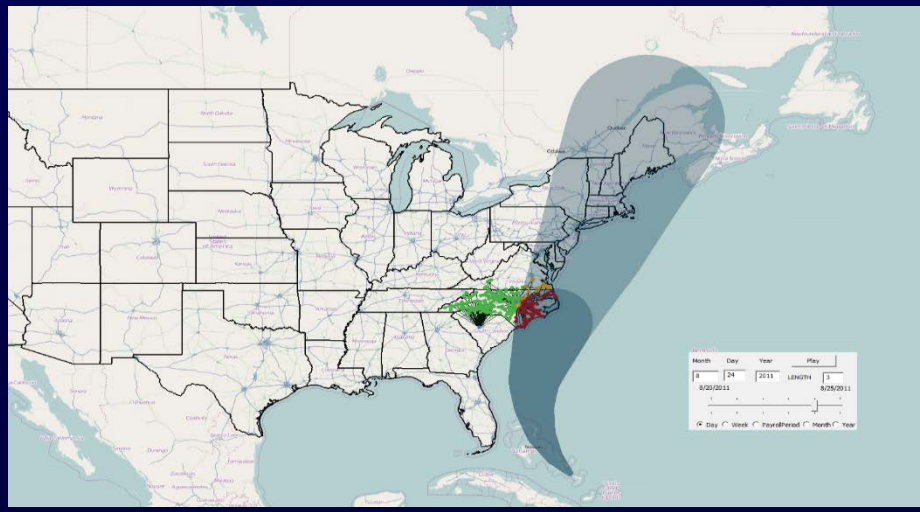
- **End-User(s):** Power Suppliers (e.g., Duke Energy),
  Cyber Community (e.g., Cisco),
  Quick Service Restaurants and suppliers

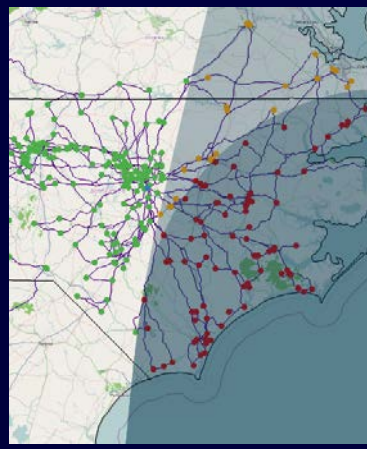
VASA: Visual Analytics for Security Applications

Ebert, Elmqvist, Gaither, Kennedy, Ribarsky, Wang
Collaborating Institution(s): Purdue, Minnesota, UTexas, UNCC + German universities
End-User(s): Fast-food restaurant chain, emergency management and planning personnel

Impacts and Accomplishments:

- Support decision-making for extreme weather and natural disaster scenarios
  - Combine real and simulation data
  - Allow “what-if” exploration
- **System of systems**: binds together multiple simulations models from collaborators into coherent whole
  - **Minnesota**: food distribution model
  - **Texas**: simulated and historical weather (hurricanes and storms)
  - **UNCC**: critical infrastructure
  - **Purdue**: roads + visual tool
- **Challenge**: combine interactive VA with heavy-weight simulation models





Transition updates:

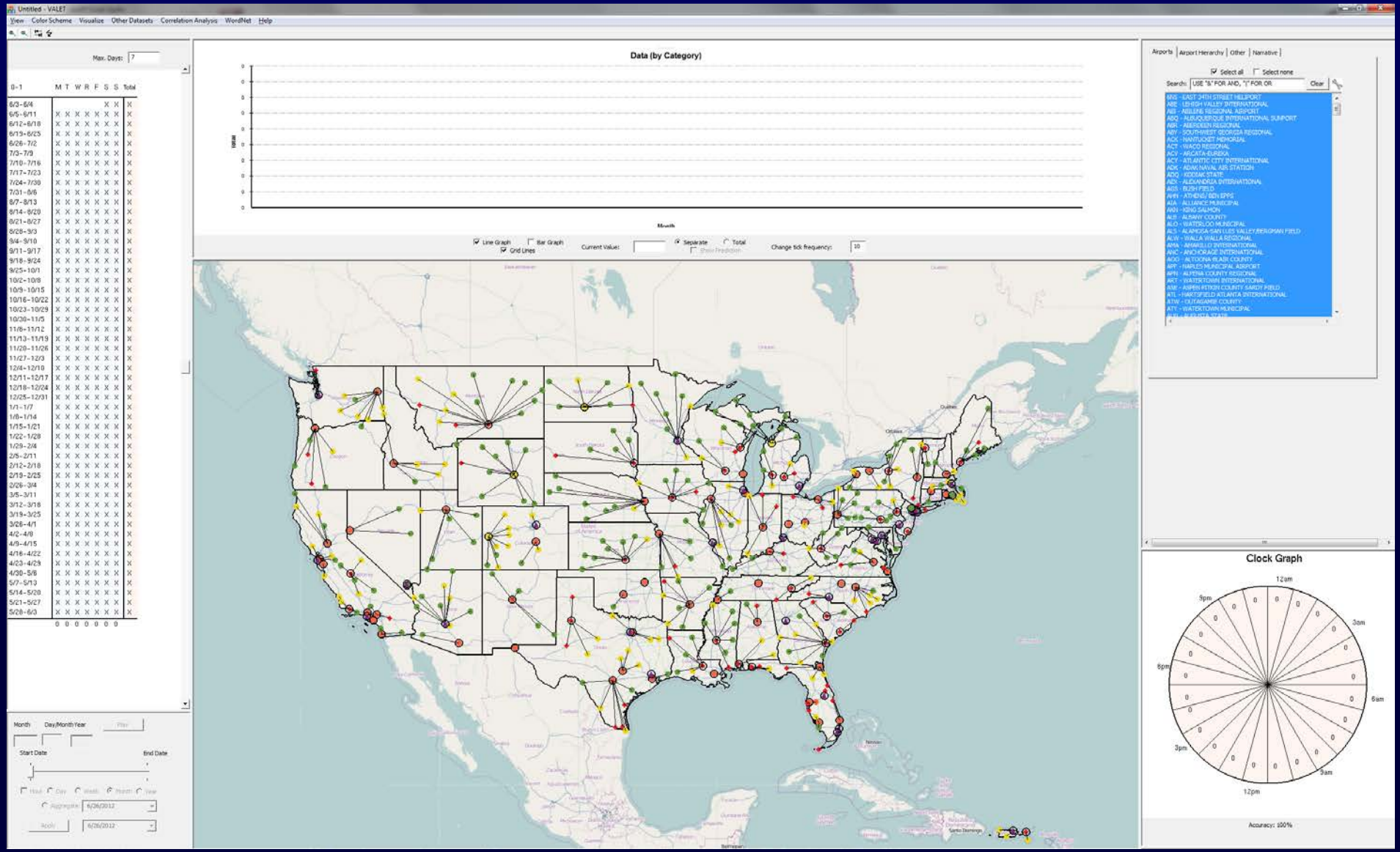
- June 2013: Meeting with German collaborators
- August 2013: US teams establish common vision
- September 2013: Architecture established

VACCINE

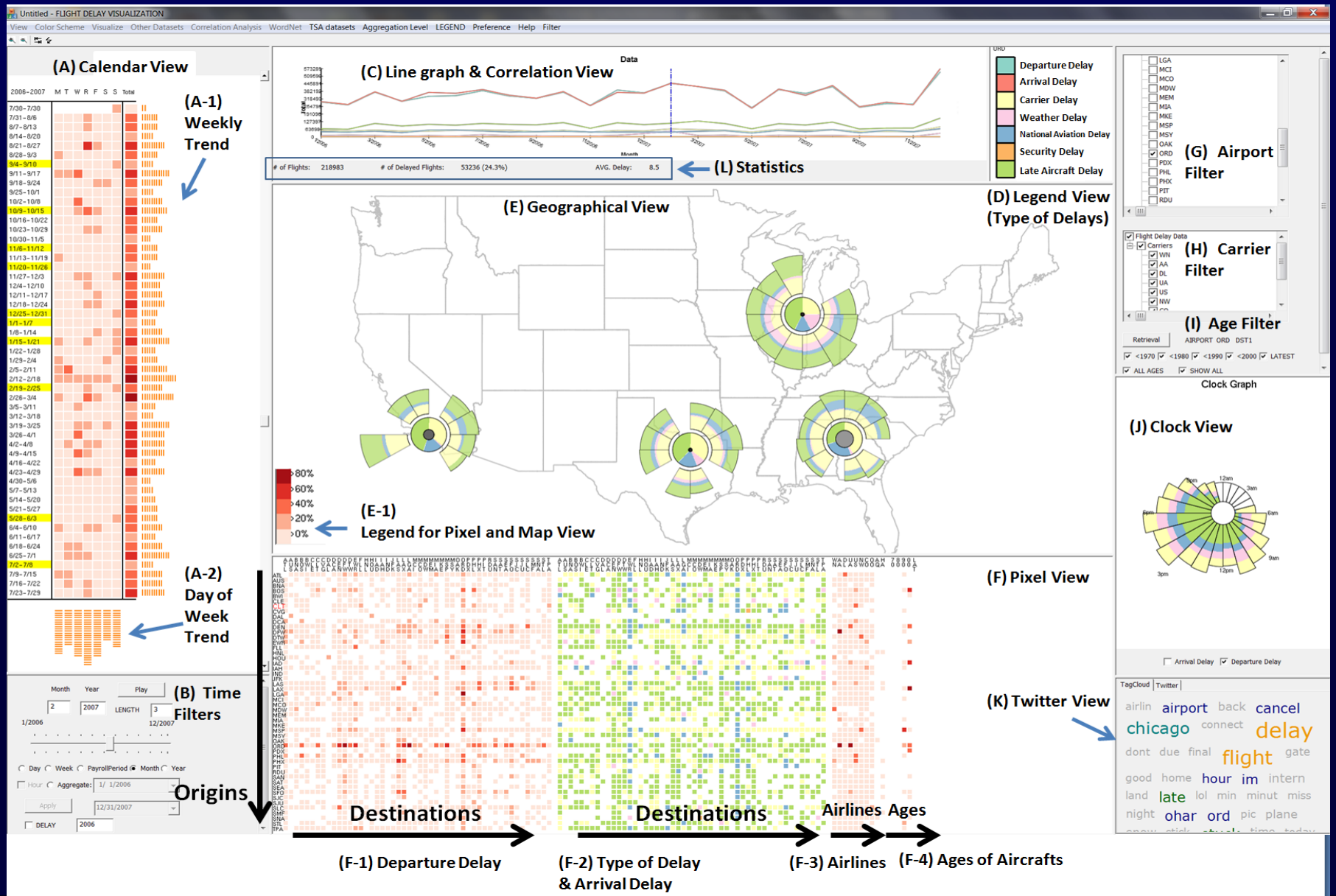
TSA Projects

- Visual Analytics for multifaceted operational analysis
  - Analysis and correlative analysis of
    - Operational performance (throughput per lane, use of AIT, targets)
    - Security performance (incidents)
    - Financial performance (expenses vs. performance metrics)
  - Spatial, temporal, spoke, hub, and regional analysis
- Customer satisfaction visual analytics
  - Jigsaw use to understand and find patterns in compliment/complaint database
  - Potential linkage to above data
- MERGE for rail inspectors

TSA Visual Analysis Prototype

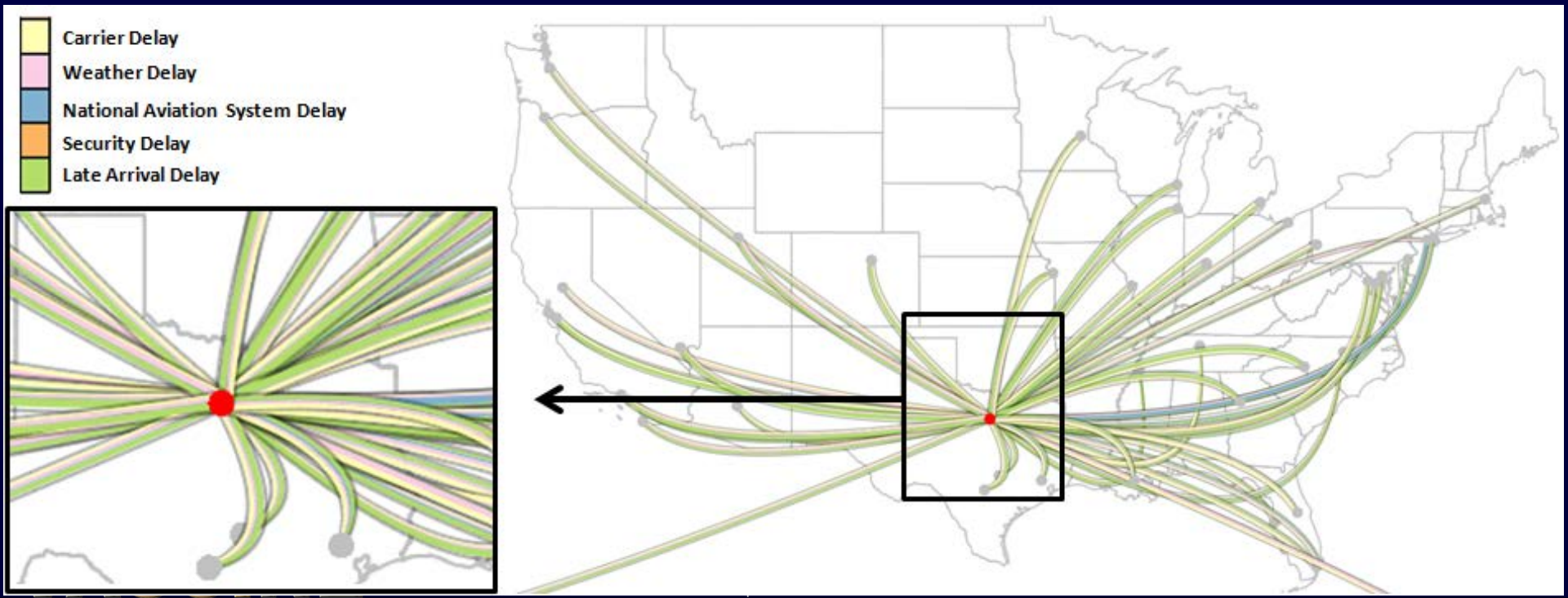


Flight Delay Visual Analysis Prototype



Flight Delay Visual Analysis Prototype

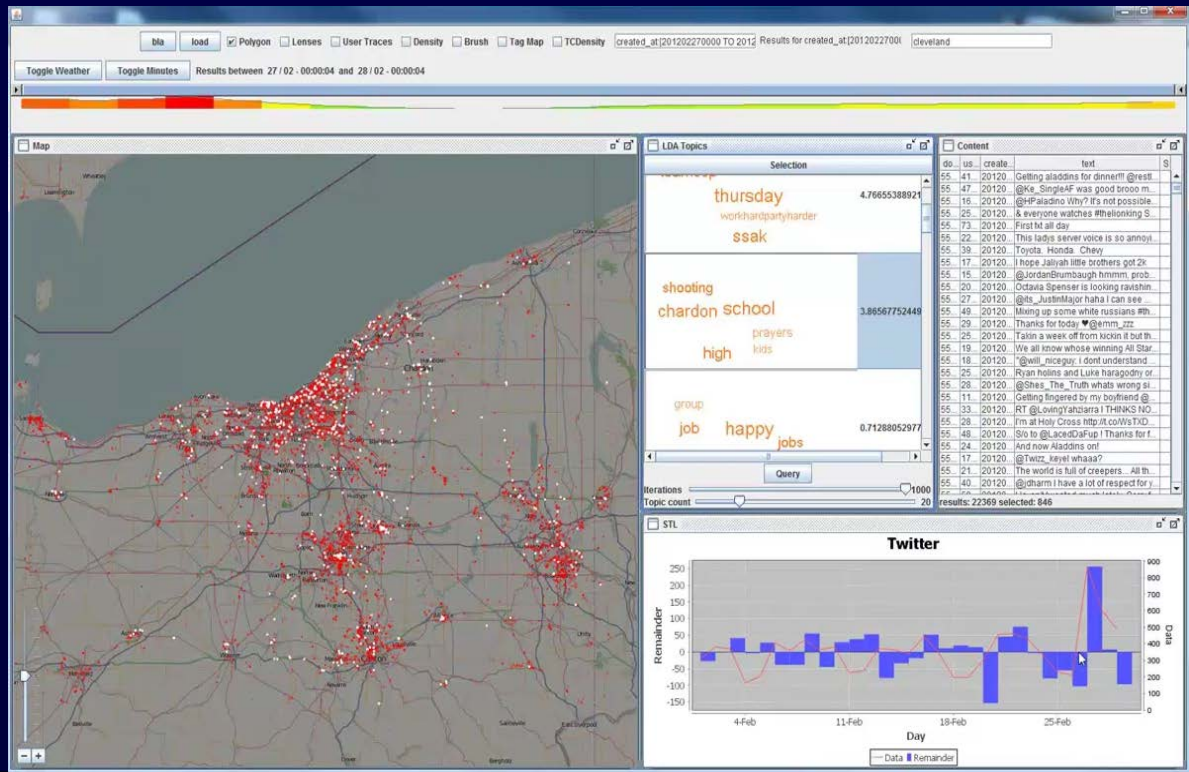
- Delay at DFW causing delays at destinations by delay type



Visual Analytics Uses for Crisis Management and Response



- Trend analysis, clustering, anomaly detection
- Interactive, multi-day, month, type investigation
- Multisource, multimedia data integration & analysis
- Purpose:
  - Planning for resiliency
  - Long-term analysis
  - Predictive analytics
  - Training
  - Detection
  - Investigation
  - Response
  - Recovery, remediation



VACCINE

Situational Awareness, Planning, Investigation, Response Visual Analytics

- Situational Awareness and Assessment from massive data
  - In-field and desktop historical and current incident/crime information
  - Innovative spatiotemporal visualization, analysis, prediction
  - Image analysis tools
  - Integrated statistical forecasting and simulation models
  - Advanced social media analytics
- Tools for Intelligence analysis
- Planning and response support
  - Integrated simulation, resource allocation, census, weather data

Current Public Safety Coalition Projects: Emergency Management

Desktop

- Miami-Dade EOC Image-based Disaster Assessment
  - Delivered to EOC Spring 2011, students deployed 2011, 2012
- Flooding damage assessment (cooperation with CHC COE)
  - Used by CHC for assessment and by FEMA/DHS
- Visual Analytics in EOCs (partnership with DRDC Canada)
  - Used in Richmond, BC EOC and for exercises

In-field

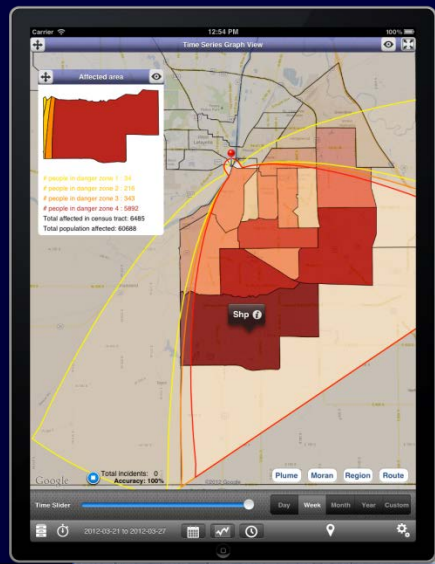
- HazMat Visual Analytic Application (MERGE)
  - Delivered to Carroll County EMA, Purdue Police, Ind. Fire, 08/2011
  - Delivered to TSA July 2012



Visual Analytics Uses for Public Safety



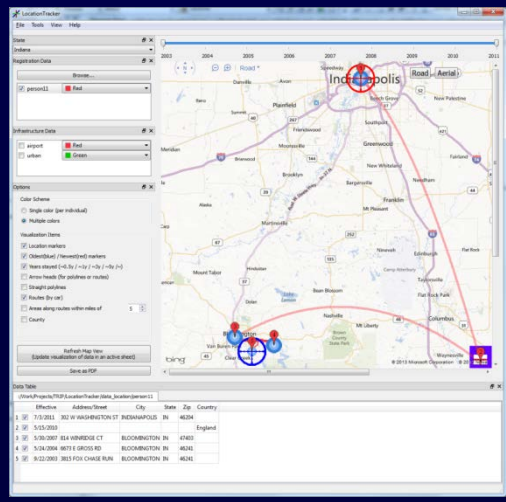
- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Interactive and multi-day investigations
- Multisource, multimedia massive data integration & analysis
- Visual standardization and interoperability



Current Public Safety Coalition Projects: Law Enforcement

Desktop crime analytics and intelligence analytics

- VALET & CrimeViz
  - Delivered to Lafayette PD 2011, WL PD Spring 2012
  - Delivered to Harrisburg Bureau of Police Spring 2011
- TRIP
  - Delivered to Indiana Intelligence Fusion Center May 2012
- Jigsaw
  - Delivered to West Lafayette PD, IMPD 09/2010, TSA 06/2012



In-field crime analytics

- Gang Graffiti Mobile Application (Gari)
  - Delivered to IMPD 09/2011, in use by 15 different police agencies in Indiana
- iValet
  - Delivered to WLPD, LPD, Purdue PD Fall 2011



Jigsaw

Visualization for Investigative Analysis across Document Collections

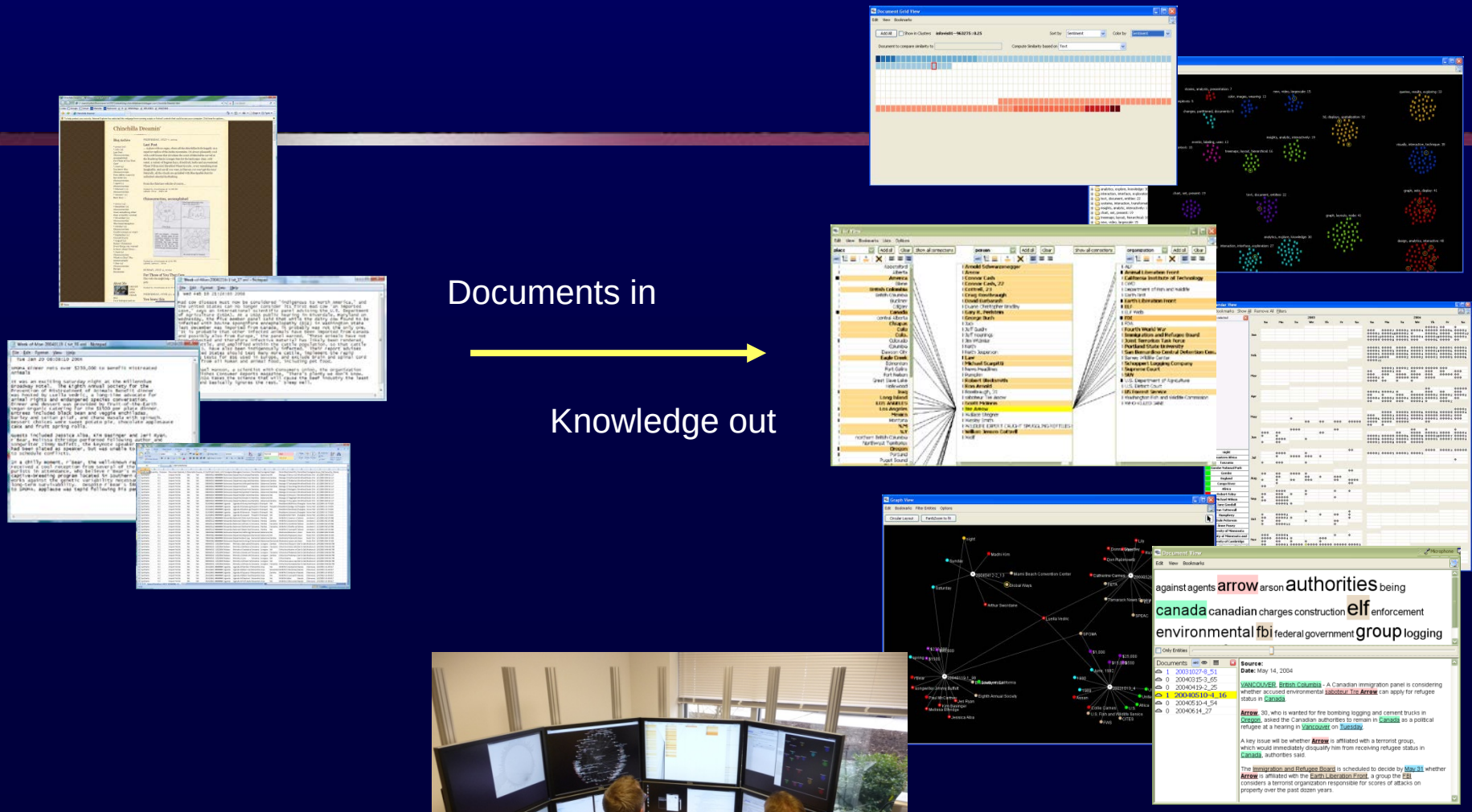
- Law enforcement & intelligence community
- Fraud (finance, accounting, banking)
- Academic research
- Journalism & reporting
- Consumer research

- Help “investigators” explore, analyze and understand large document collections



“Putting the pieces together”

Jigsaw: Visual Analytics for Investigative Analysis



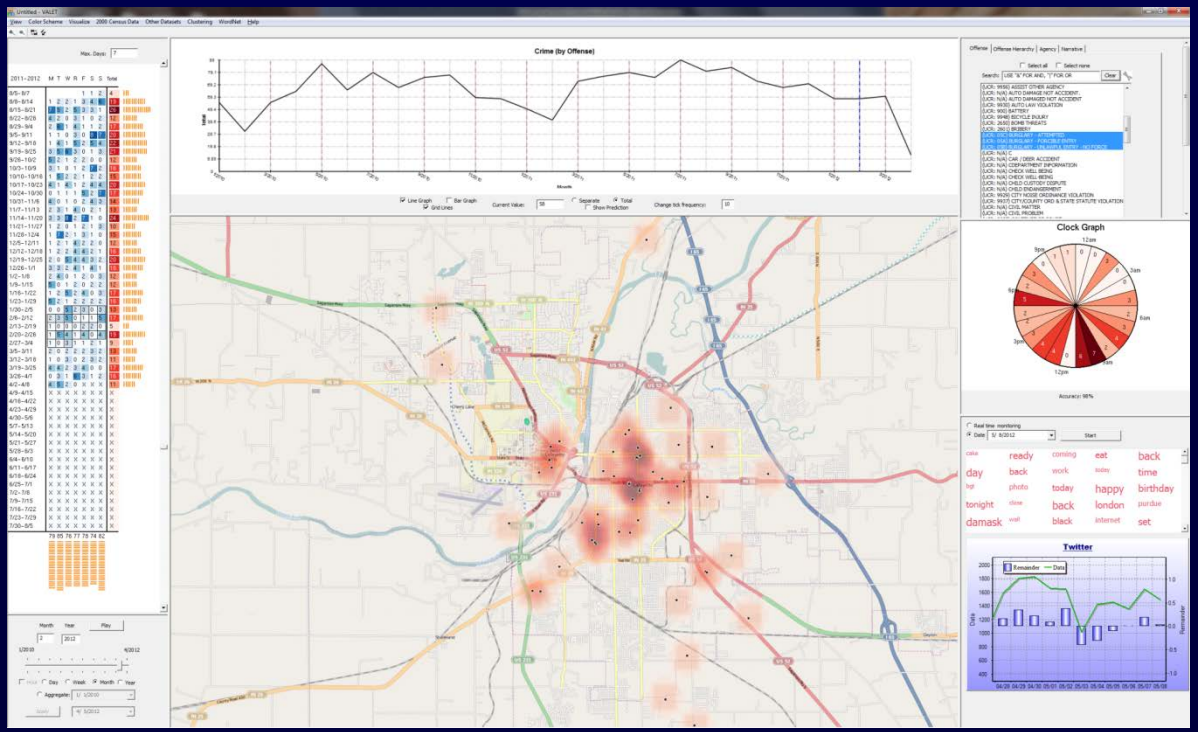


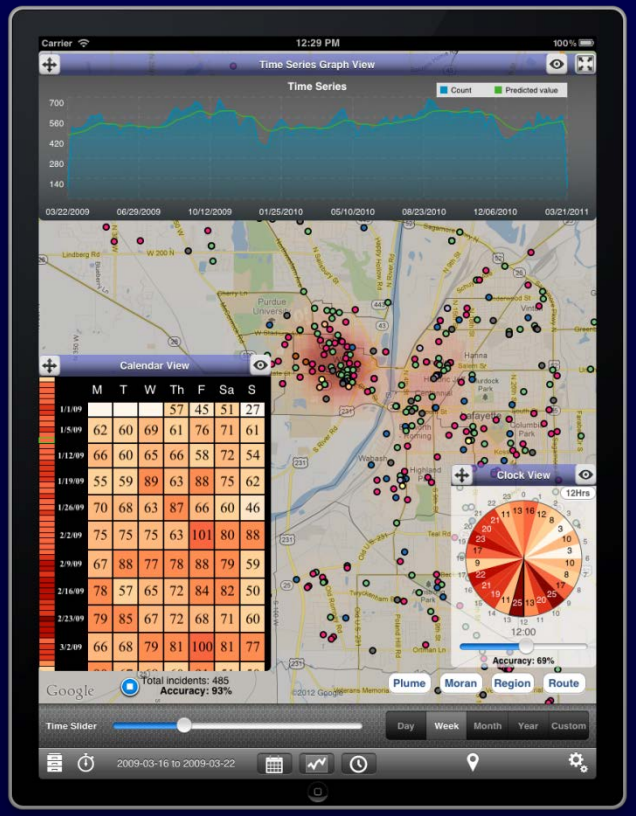
John Stasko

VACCINE

Georgia Tech College of Computing
School of Interactive Computing

Visual Analytics Law Enforcement Toolkit (i)VALET





VACCINE

VALET

Visual Analytics Law Enforcement Toolkit (VALET, iVALET)

Impacts:

- In use to analyze crime patterns in Lafayette, Indiana and to connect strings of activities
- Mobile version being released to public (February 2013) for community-based policing
- Investigating correlation factors
- Analyzing time of day problems and improving accuracy of police record management system
- Novel statistical predictive model incorporated for planning
- Incorporating predictive alerts

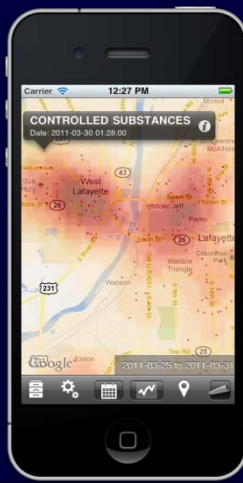
iVALET
Debriefing Example

VALET delivered:

- Spring 2011: WL, Lafayette Police

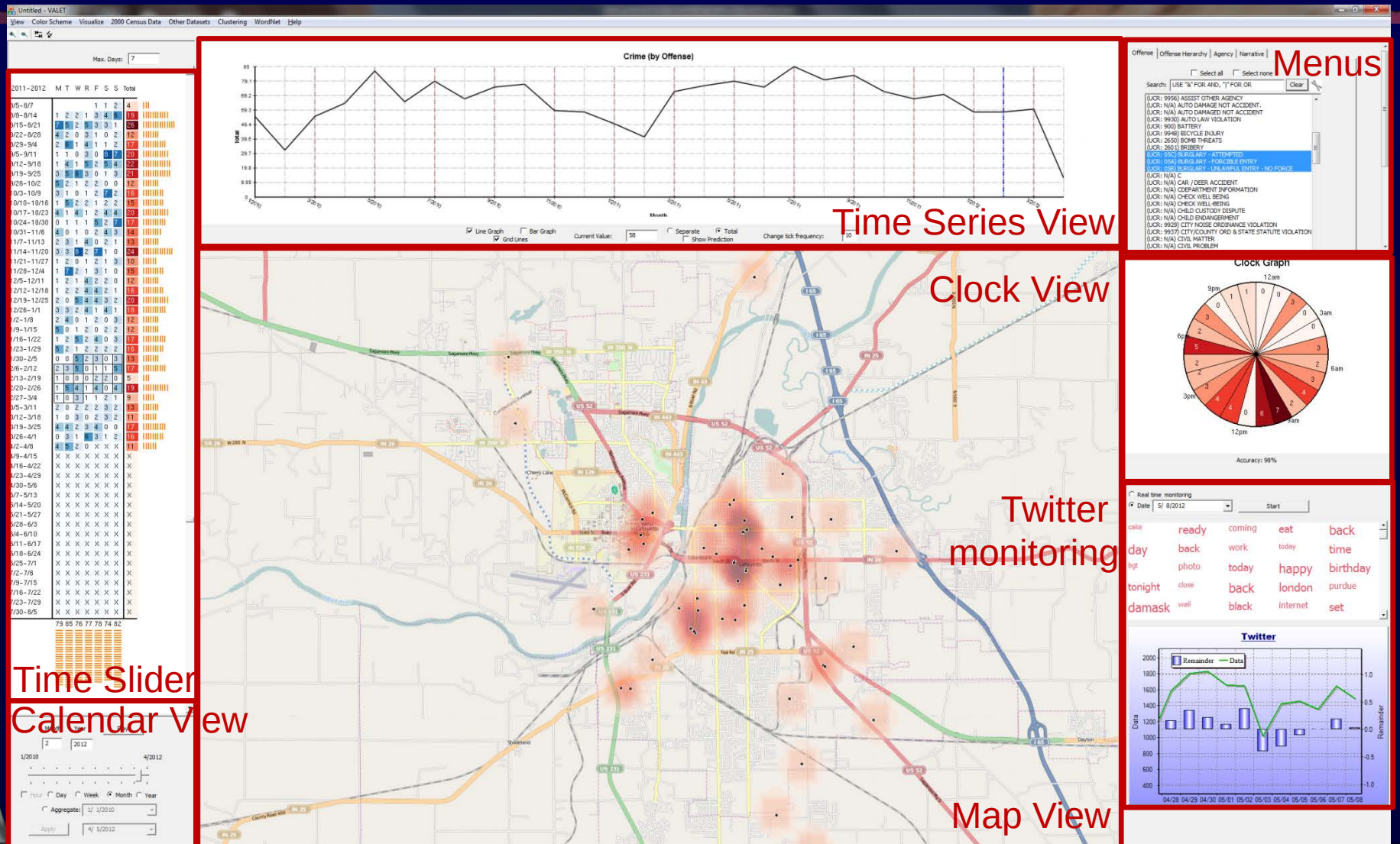
iVALET delivered:

- October 2011: Purdue, WL Police

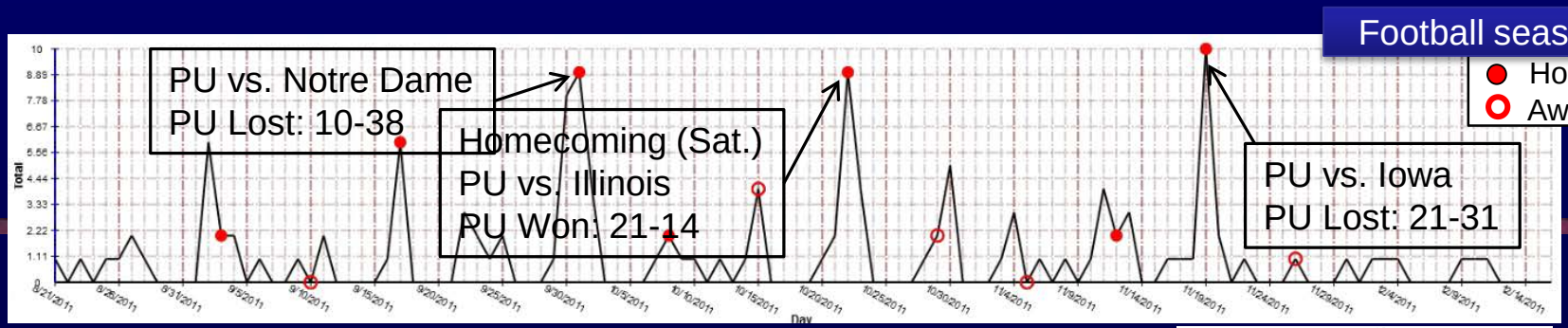


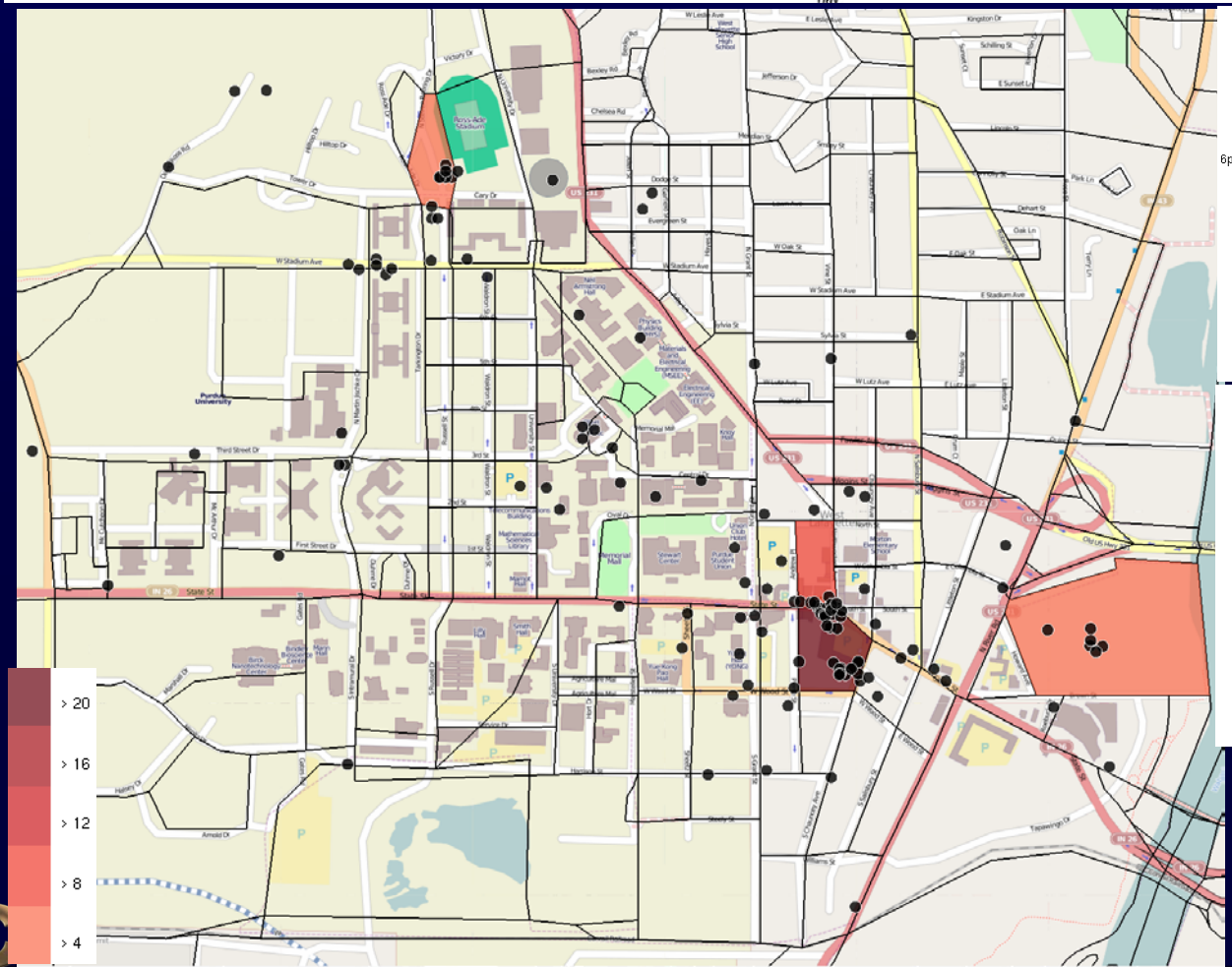
VACCINE

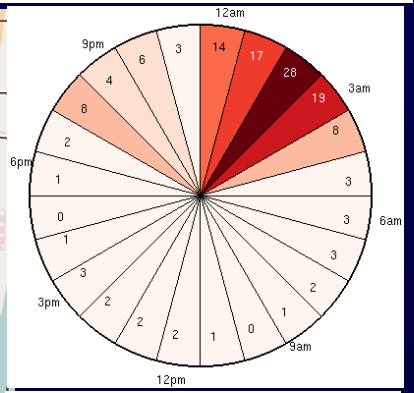
VALET Overview

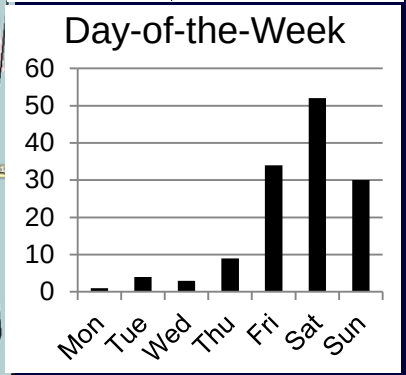


Example: Drunkenness / Public Intoxication



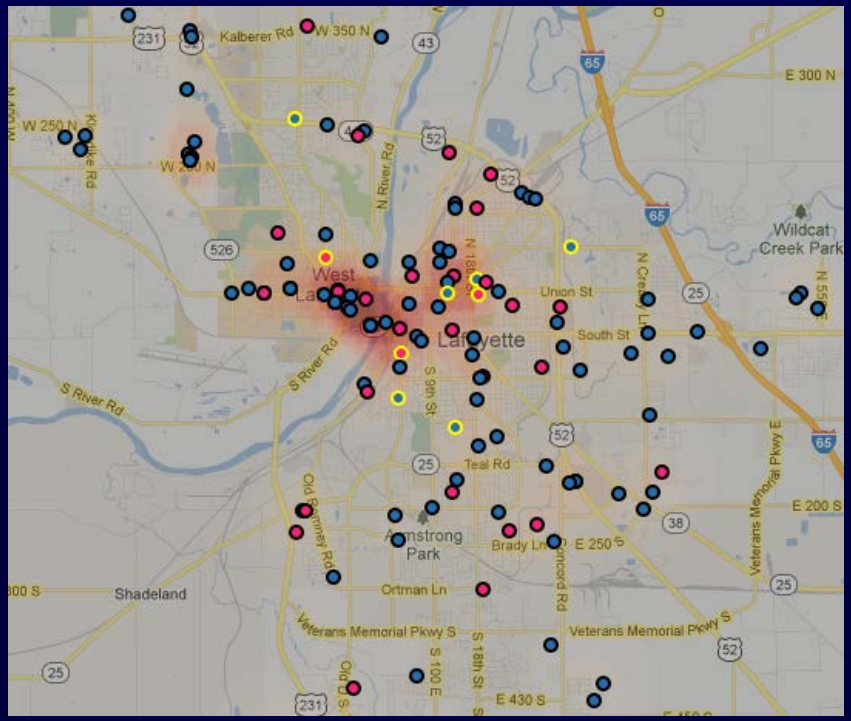






Top 10 Hot Incidents

- Identify unusual localized high-frequency patterns of crimes (near repeats)
- Each data entry is checked for other crimes with similar properties within a 1 block radius of the incident location and a 14-day time period;
- Top 10 incidents with the most number of related incidents in this space-time window are highlighted



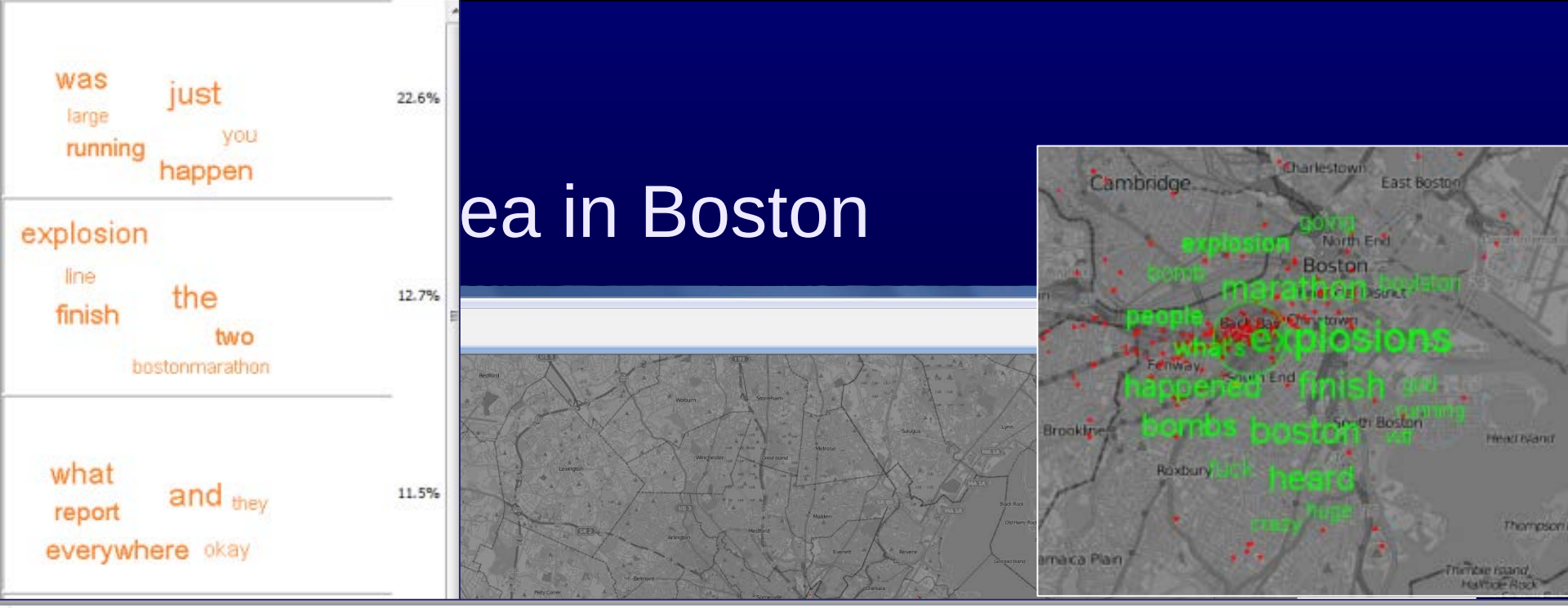
Social Media:Real-time Twitter Monitoring and Integration into Tools

(Purdue, Stuttgart, Penn St.)

- Topic extraction using novel STL based remainder estimation technique
- Dynamically linked views providing options to monitor emerging / emergent twitter feeds
- Topics extracted shown as a dynamic word cloud

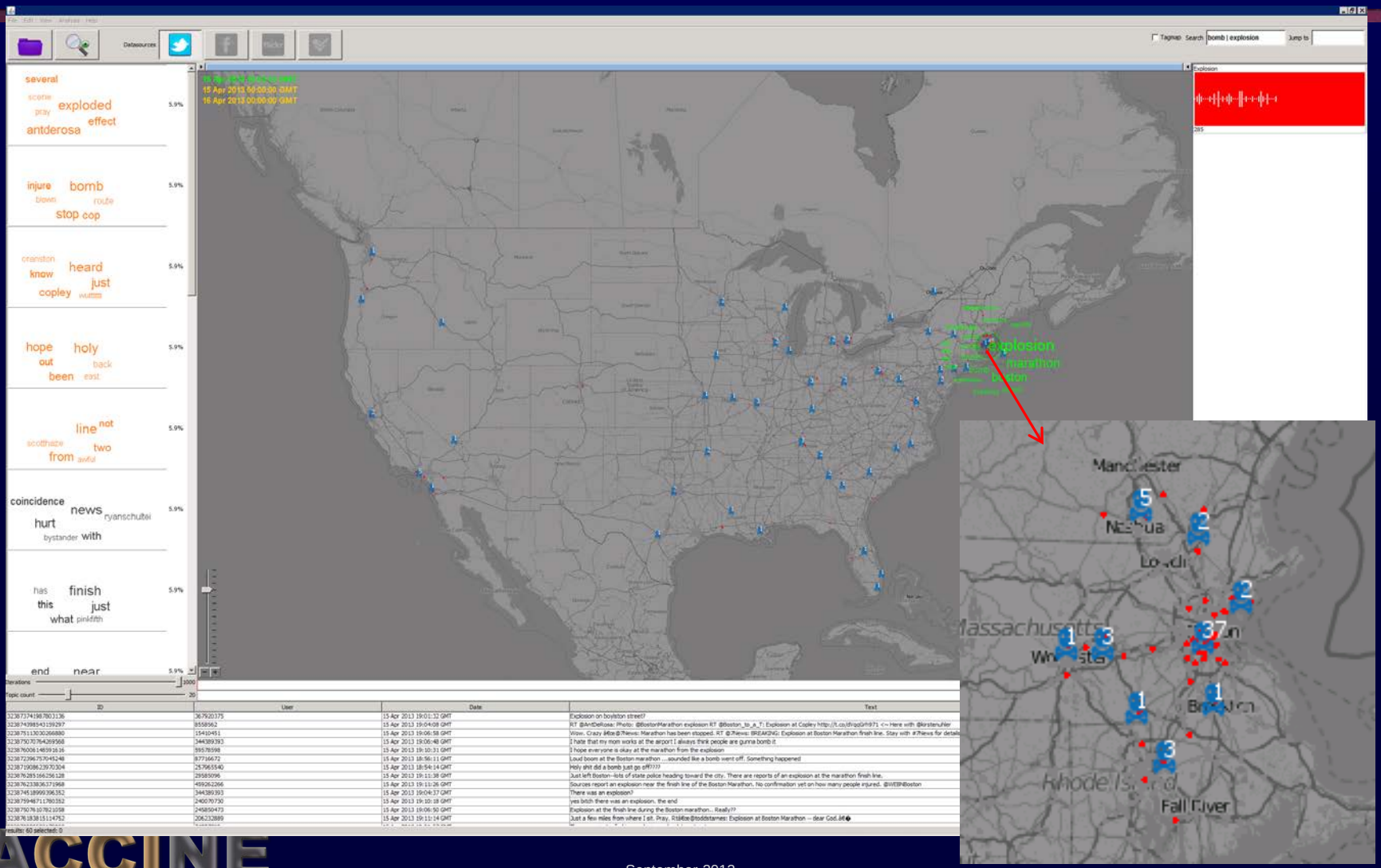


Grand Prix Weekend, Purdue University



Text
Oh my god what just happened
Something happened at the end of the Boston Marathon. Something bad and there is a lot of chatter on Twitter. What's going on?
Multiple people are injured near the Boston Marathon finish line after two explosions. The #BostonMarathon has been stopped.
two bombs just went off on boylston
BREAKING NEWS: Two powerful explosions detonated in quick succession right next to the Boston Marathon finsh line this afternoon.
What the FUCK was that
Back in Sept, @croon1 solicited me for $2000. He now has a music video with William Shatner. If you watch it (god forbid) keep that on mind.
Literally what the fuck get me out of here
@DTenenbaum my office right next to it
@FRANCESCalciO I figured 3 people would get the joke
Two explosions just rocked the finish line of the Boston Marathon. Sirens galore. People running in fear. Wonder what happened.
This is crazy i seen that blow up #bostonmarathon
can someone tell me what that explosion was!? #boston #bostonmarathon

Detection using the Explosion Classifier



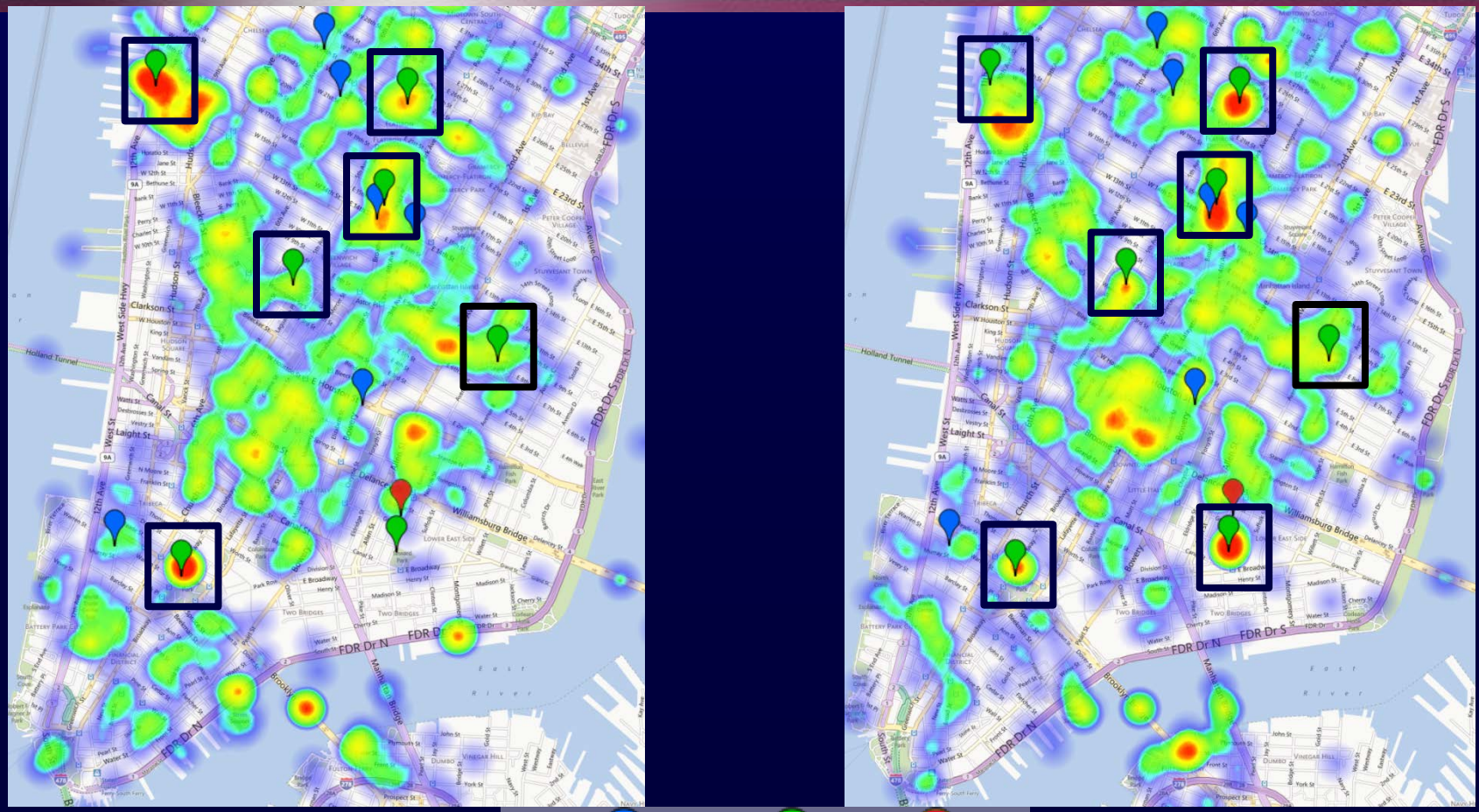
First Response (Tweet & Picture)



Visual Analytics of Activity During Hurricane Sandy

Two weeks before Sandy
10/14 (Sunday), 12:00 ~ 16:00

One week before Sandy
10/21 (Sunday), 12:00 ~ 16:00

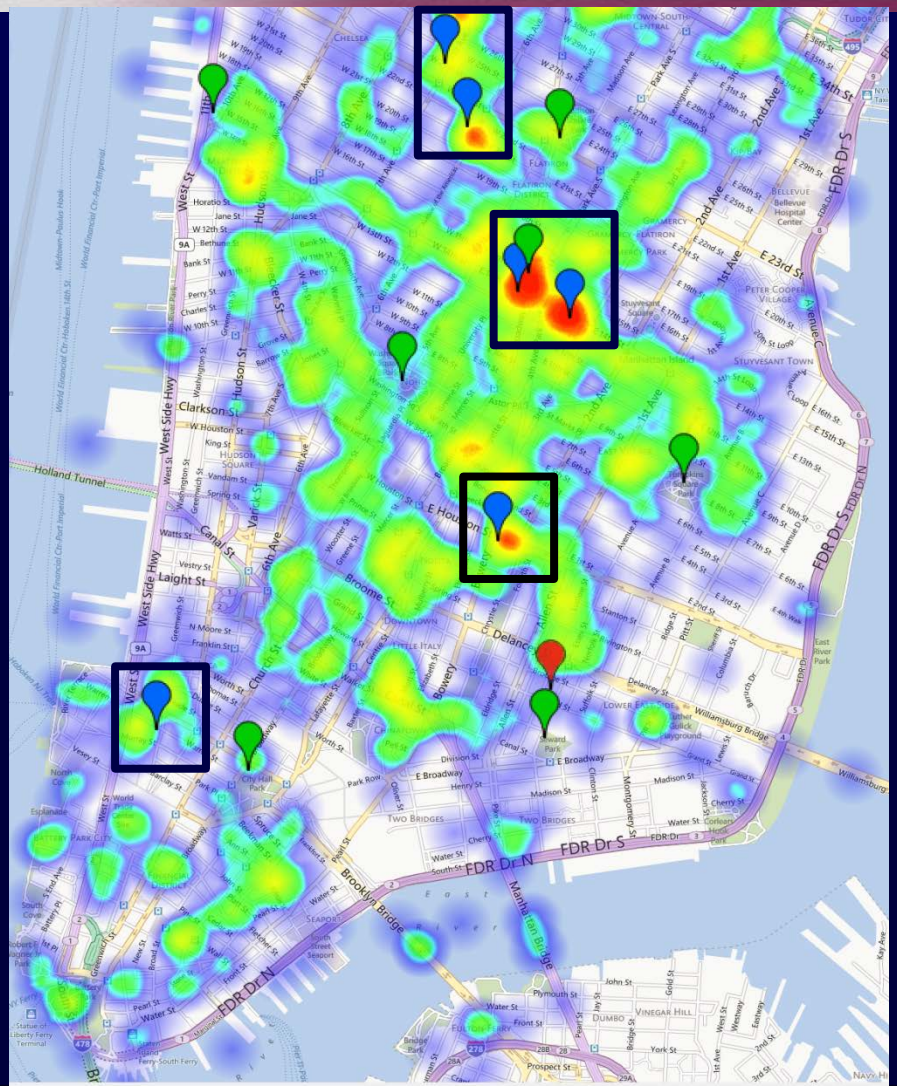


Visual Analytics of Activity During Hurricane Sandy

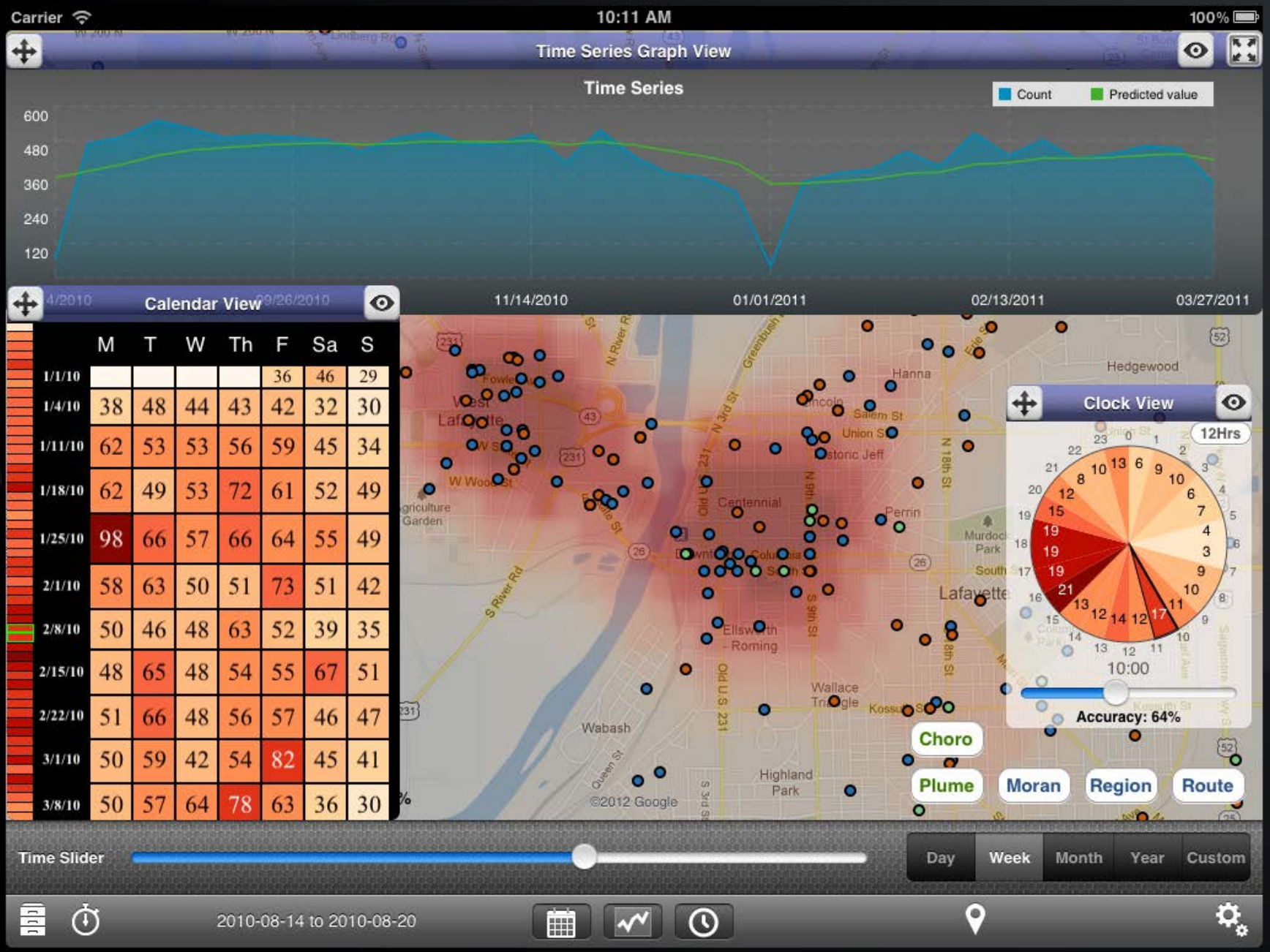
After the evacuation order
10/28 (Sunday), 12:00 ~ 16:00

Evacuation order:
10/28, 10:30 AM

Hurricane Sandy's
Arrival at NYC:
10/29, 8:00 PM

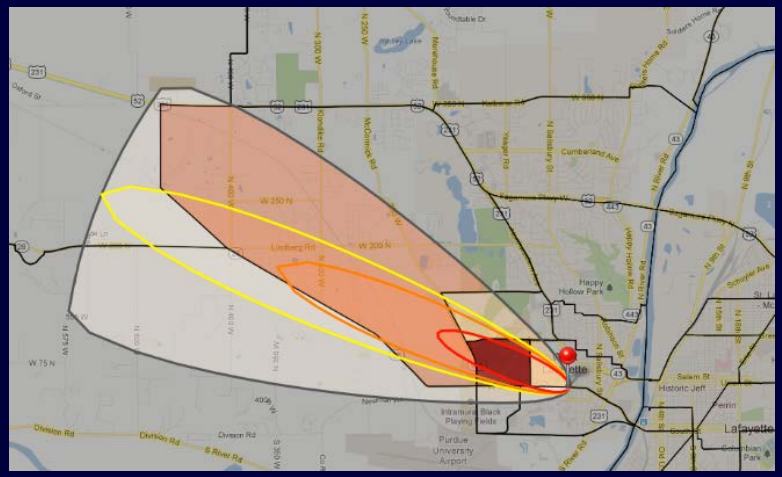


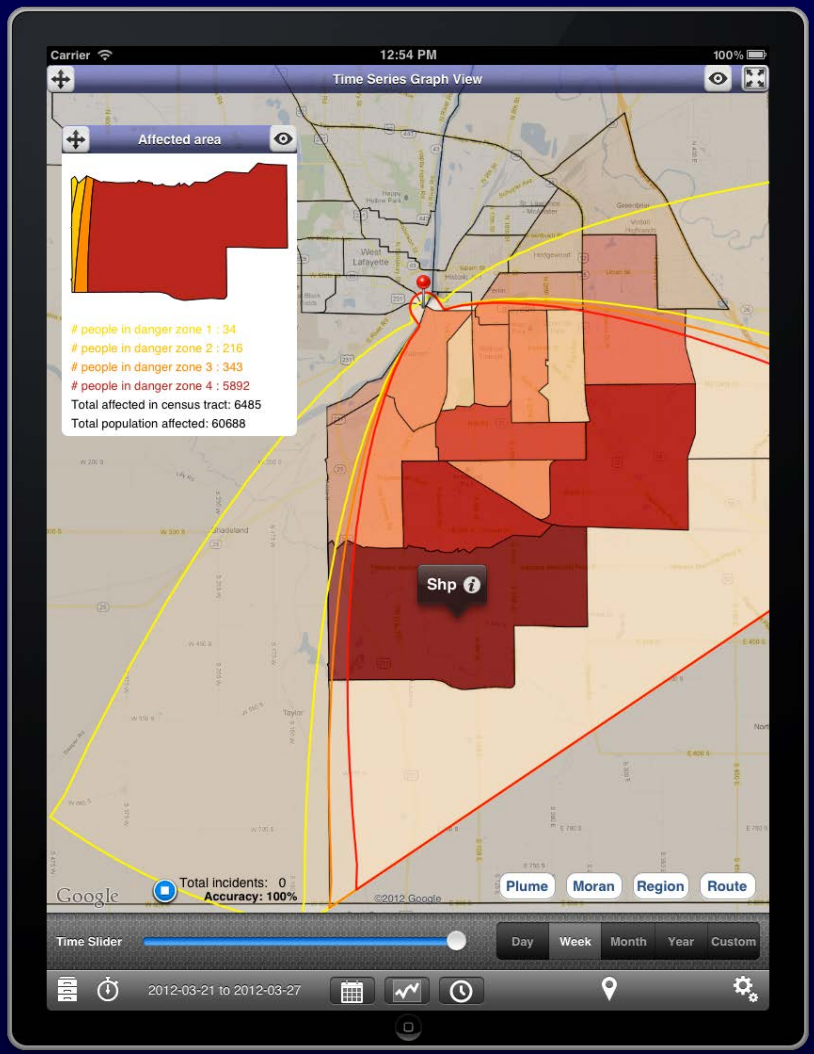
VACCINE


Carrier
10:11 AM
100%
Time Series Graph View
Time Series
Count
Predicted value
600
480
360
240
120
11/14/2010
01/01/2011
02/13/2011
03/27/2011
Calendar View
M T W Th F Sa S
1/1/10 36 46 29
1/4/10 38 48 44 43 42 32 30
1/11/10 62 53 53 56 59 45 34
1/18/10 62 49 53 72 61 52 49
1/25/10 98 66 57 66 64 55 49
2/1/10 58 63 50 51 73 51 42
2/8/10 50 46 48 63 52 39 35
2/15/10 48 65 48 54 55 67 51
2/22/10 51 66 48 56 57 46 47
3/1/10 50 59 42 54 82 45 41
3/8/10 50 57 64 78 63 36 30
Clock View
12Hrs
10:00
Accuracy: 64%
Choro
Plume
Moran
Region
Route
Time Slider
Day
Week
Month
Year
Custom
2010-08-14 to 2010-08-20
©2012 Google

MERGE – iVALET Interactive Plume Visualization and Evacuation Planning

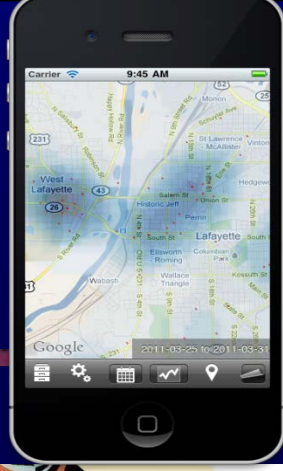
- Chemical release plume modeling identifies census tracts with the highest number of expected people affected





We Have an App For That!

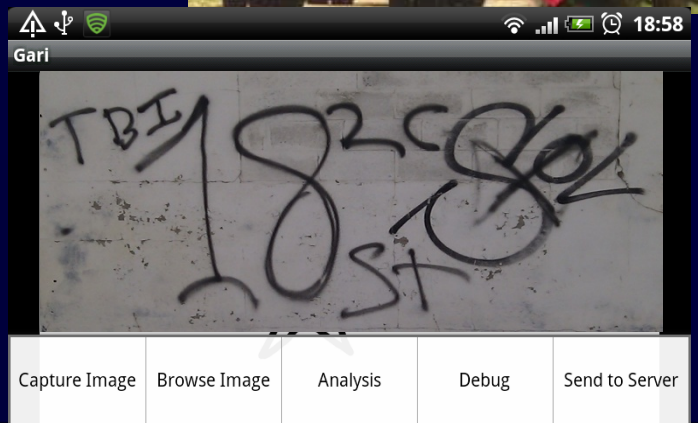
Our Mobile Tools



- VALET
- Evacuation Planning
- Rosetta Phone
- Hazmat app
- Gang graffiti app
- Tatoo app

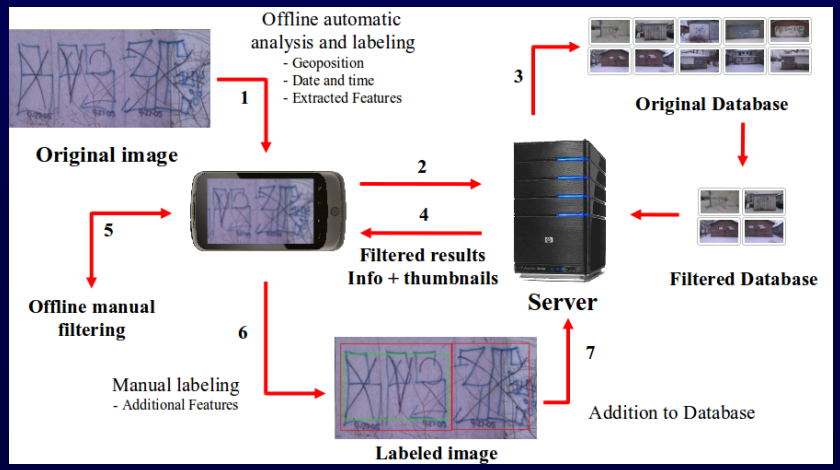






VACCINE

Gang Graffiti Analysis and Recognition Using a Mobile Telephone (GARI)





IMPACT:

- Allows police to catalog and analyze gang graffiti images into a database system to better track and determine gang activity throughout a region
- Will allow the graffiti images to be "interpreted"
- More than 75 users and 800 graffiti images acquired

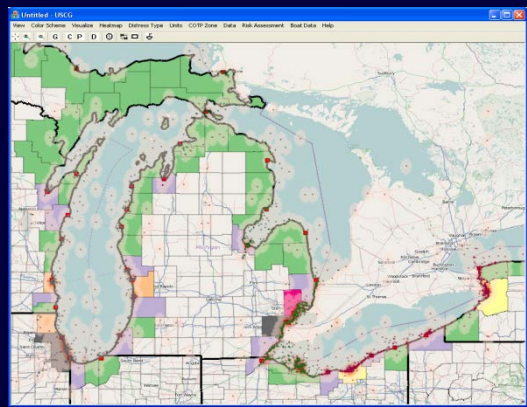
GARI delivered:

- Summer 2011:
  - IPD gang detectives
- August 2011:
  - IPD at large
  - Ind Fusion Gang Task Force
- September 2011:
  - Gang detectives across Indiana

Visual Analytics Uses for Risk-Based Decision Making



- Risk visualization and analysis
- Predictive analytics
- Uncertain decision making
- Alternative evaluation and consequence investigation
- Trend analysis, clustering, anomaly detection
- Interactive, multi-day, month, type investigation
- Multisource, multimedia data integration & analysis

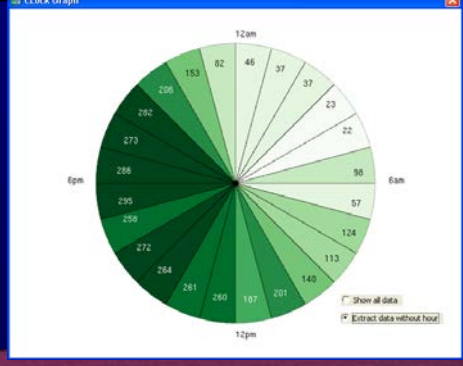


USCG: Effective Risk-based Decision Making and Resource Allocation Visual Analytics

- Evaluate current and historical mission area:
  - Demands
  - Risks (total, mitigated, residual)
  - Resource allocation
  - Return on investment
- Evaluate courses of action
- Evaluate above at both Strategic and Tactical/Operational level



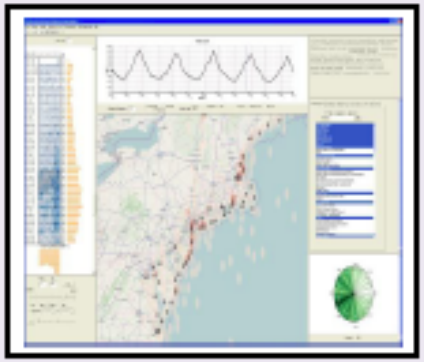
Risk-Based Allocations



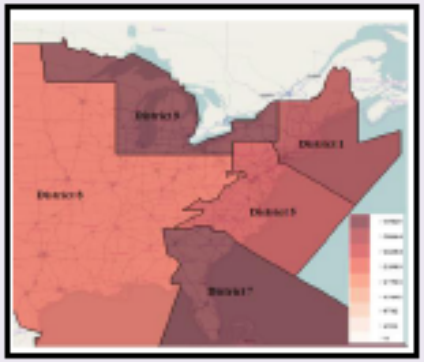
- Comparative visual analysis of mission cases/hours vs. staffing hours
- Comparative visualization of resources vs. risk
- Trend visual analytics
  - Increase/decrease in resource allocation
  - Increase/decrease in risk (total, mitigated, residual)
  - Increase/decrease in incidents
- Exploration of alternatives and effect on risk
- Predictive analytics based on historical data (STL and EWMA)

VA For Risk-Based Decision Making Process

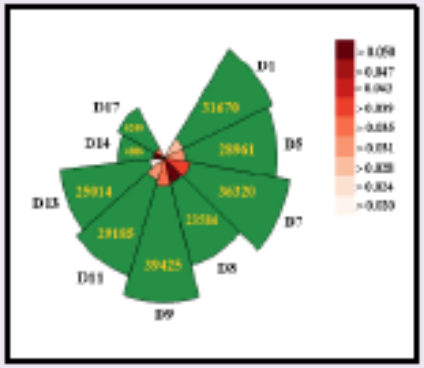
Overall Process



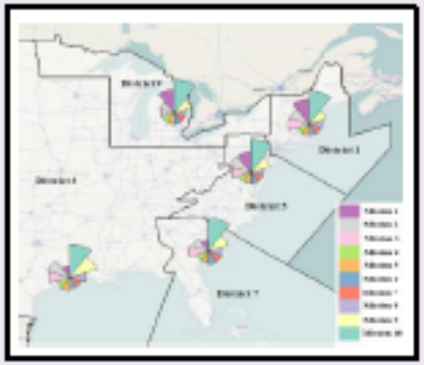
(A) Initial system for visualizing Coast Guard Operations.



(C) Choropleth Map for M10's Total Risk



(B) M10 Risk Profile for all the Districts



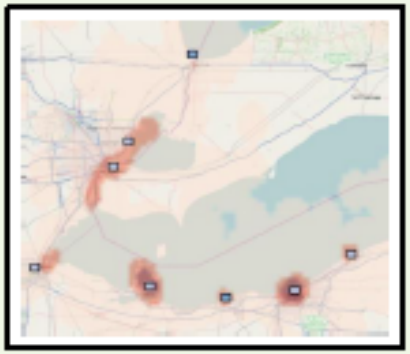
(D) Risk Profiles: Total Risk distribution per District.

- Define the Context
- Identify potential risks

1. Visualize spatial distribution of risk associated with operations



(E) Total Risk Values distributed amongst the stations' AOR



(F) KDE heatmap showing Total Risk values for incidents with accurate lat/long.

- Assess and Analyze

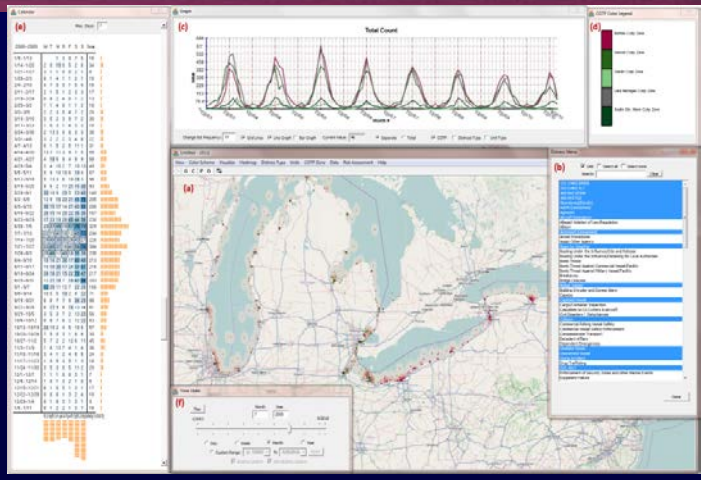
1. Perform spatio-temporal analysis of the risk values.
2. Identify risk priority areas and how they evolve over time and identify the mitigated risk.

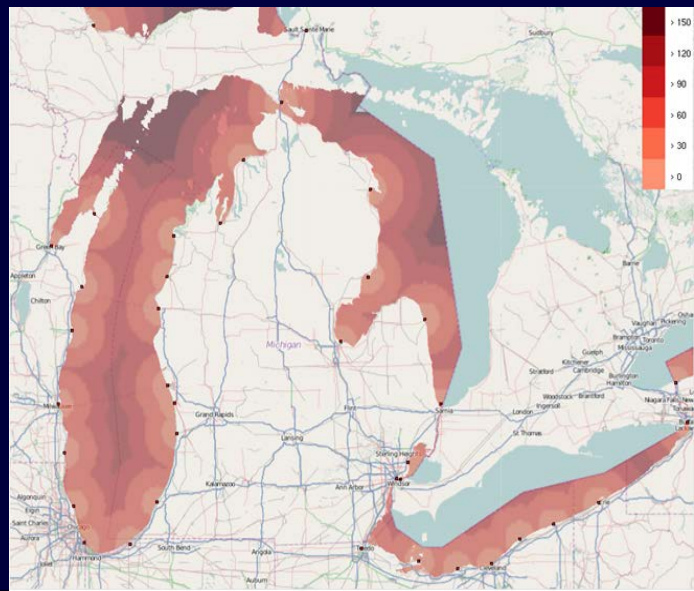
U.S. Coast Guard Search and Rescue VA (cgSARVA)

Partners: USCG LANT 7, USCG D9, USCG D5, USCG HQ 771

IMPACTS:

- Analyzed impact of CG auxiliary stations on search and rescue mission in Great Lakes
- Used for resource allocation for SAR
- Provided new insights to SAR mission
- Hurricanes Sandy and Irene resource allocation decisions based on cgSARVA analysis and visualization
- Informed Commandant's budget testimony and recommendations to Congress
- Key component of USCG D9 reallocation plan for 2011-2012 based on decreased budget
- Key component of Coastal Operations Allocation Suite of Tools (COAST) – USCG HQ

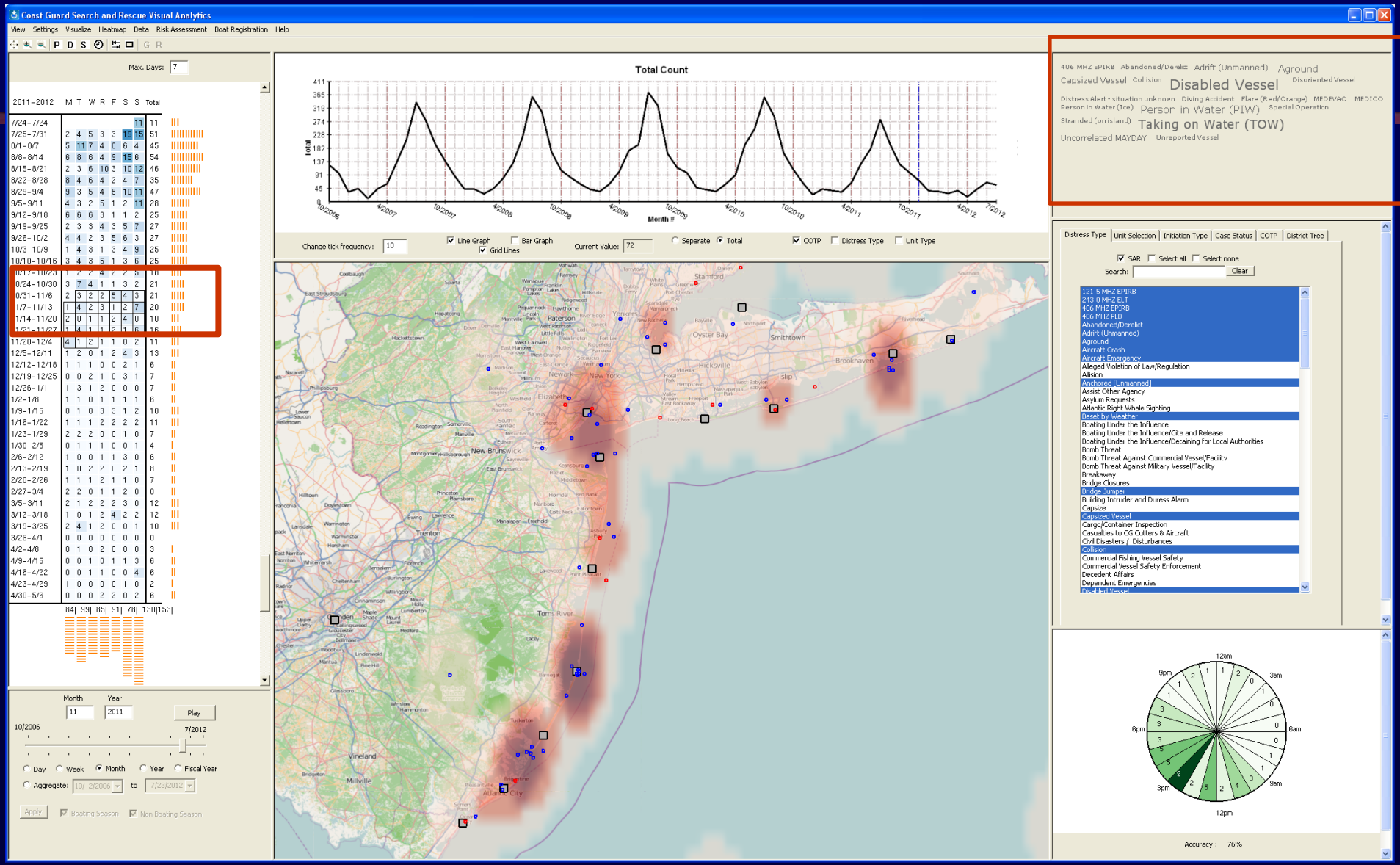




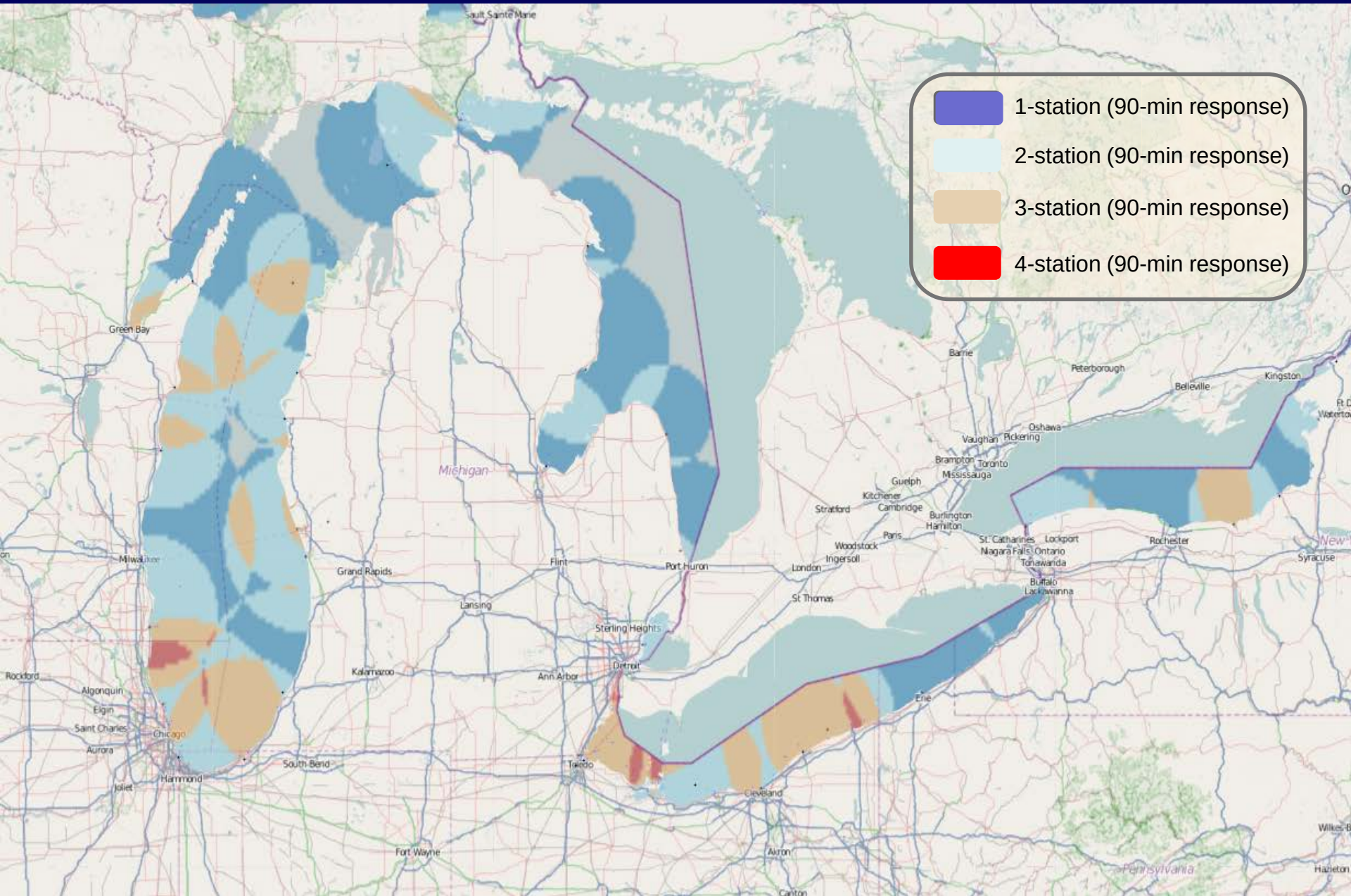
VACCINE

Example: Risks and Consequences From Sandy:

SAR Cases November 2011 NJ/NYC Area



Response Efficiency – Possible Asset Allocation



Software Accredited for Decision Making

- April 22, 2013 cgSARVA VV&A'd for US Coast Guard system-wide use

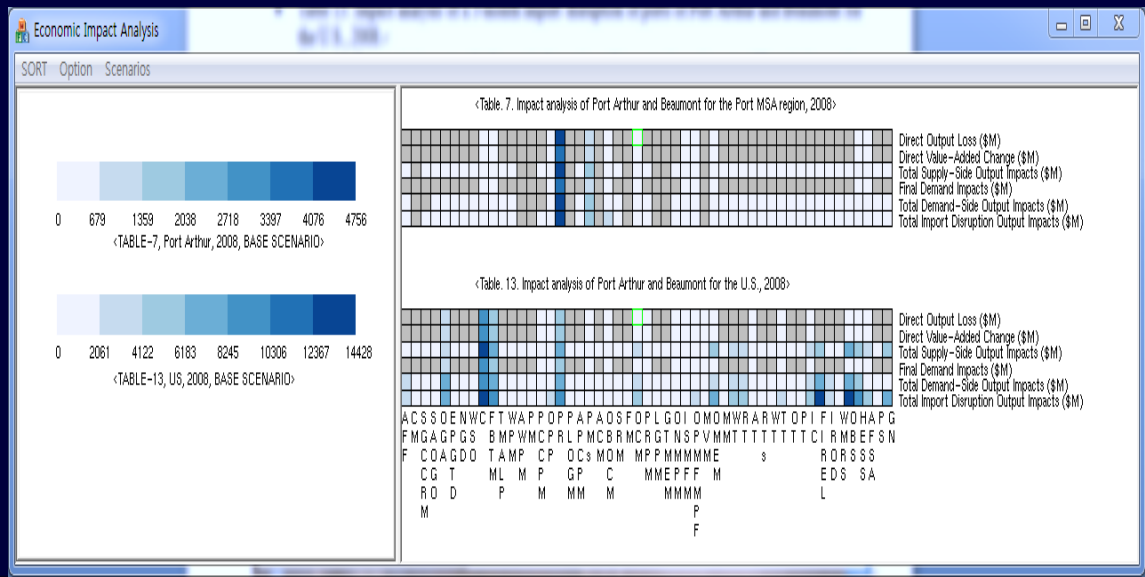


USCG Port Closure Economic Impact VA

Partners: USC CREATE, USCG RDC, USCG D7, USCG LANT

IMPACT:

- Provided tool for use analysis and planning for impact of port closure in Port Arthur, Tx
- Economic sector impact, local and national impact
- Impact and effectiveness of alternative mitigation strategies



VACCINE

Directions Forward, Keys to Success & Challenges



Six Challenges for Proactive & Predictive Visual Analytics

1. Creating computer-human visual cognition environments
2. Integrating interactive simulation and analytics
3. Solving specific scale issues and cross-scale issues
4. Managing uncertainty and time
5. Enabling risk-based decision making environments
6. Developing the Science of Interaction for Visual Analytics

Visual Analytics: Remember…

- We need to be cognizant of parameters for visual representations
- Appropriate analysis can guide users to interesting features in the data
- Refined analysis through user interaction and their domain knowledge can help discover hidden problems
- There is no single catch-all visual representation or analysis

Keys for Success

- User and problem driven
- Balance human cognition and automated analysis and modeling
  - Often applied on-the-fly for specific components identified by the user
- Interactivity and easy interaction
  - Utilizing HPC and novel analysis approaches
- Understandability of why predicted value is what it is
- Intuitive visual cognition
- Not overloaded with features

For Further Information

www.VisualAnalytics-CCI.org



vaccine@purdue.edu
ebertd@purdue.edu

VACCINE

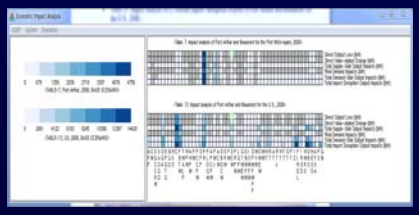
VACCINE

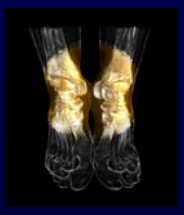
VACCINE

Visual Analytics for Command, Control, and Interoperability Environments
A U.S. Department of Homeland Security Center of Excellence

Visualization Literacy: Research Challenges

Do You See What I See?





David S. Ebert
Purdue University
ebertd@purdue.edu

Do I Really See a Bone?
IEEE Visualization panel

June 2014

Visualization Literacy Research: Foundations

- What is the current state of visualization literacy aptitude of students and the public?
- Can we create simple aptitude exercises?
- How do we tailor visualization to user's cognitive profile?

VACCINE

June 2014

Visualization Literacy Research: Interactive vs. Static

- How does interaction change literacy of individuals?
- How do we capture the interactive exercise for explanation, training, etc.?



Visualization Literacy Research: Evaluation

- How do we do evaluation studies across diverse populations?
  - Scientists/engineers vs. poets?
  - Fireman vs. statisticians?
  - How do we understand variation of the casual experts?
- Can we determine universality, category group norms, …



Visualization Literacy Research: Training

- How do we develop undergraduate training (e.g., this workshop)?
- When should training start (K-12)?
- How do we train professionals in various fields?



Visualization Literacy Research: Complex Analytics and Decision Making

- What techniques work in complex analytic and long-term decision making environments?
- Should visualization explain complex analysis techniques and evidence from automated processing for casual experts?
- How do capabilities change as problem complexity, size, and duration change (learning vs. overloading)



For Further Information

www.VisualAnalytics-CCI.org



AN INFORMATION THEORETIC APPROACH TO THE FLIGHT DATA PROBLEM

Min Chen and Simon Walton
Oxford University

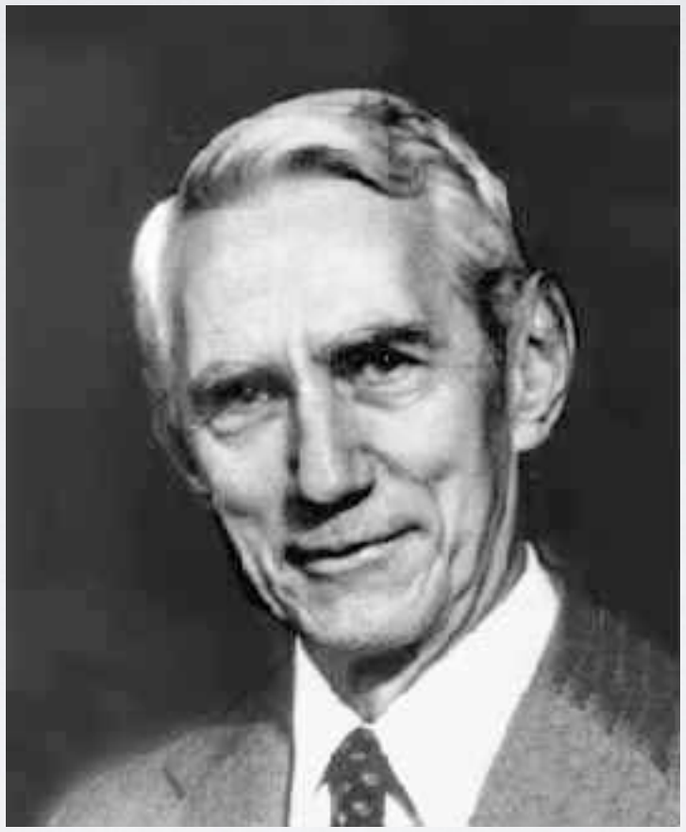






CONTENTS

- Overview of the information-theoretic approach
- Implementation using a query-driven database
- Designing queries using a GUI
- Visualising queries using a map & pixel-based visualisation



Claude Shannon

VISUAL ANALYTICS

An information-theoretic approach

VISUAL ANALYTICS TASKS

- Given a flight record, inform whether it's normal or abnormal.
- Show all `abnormal' records.
- What are the key indicators for a specific type of phenomenon, problem, or risk?

WHY NOT MACHINE LEARNING?

- Insufficient anomalies to train
- Uncertainty in annotation
- Importance of different dimensions unknown/uncertain
- Context-sensitive knowledge not taken into account

FLIGHT DATA PROBLEM

- 29 attributes in **delay** table
- 7 additional attributes in linked **airport** table
- Some attribute pairs have a high mutual information rating

INFORMATION THEORY

- Surprisal measure $-\log_2(p(R))$
- Entropy measure $\boldsymbol{H}(X) = -\sum_{i}^{m} p(x_i)\log_2 p(x_i)$
  - I.E. the weighted sum of each element's entropy

	field	distinct	min	max	avg	stddev	data type	example values
2	id	1111000	1	1111000	555500.5	320718.219	integer	
3	year	1	2008	2008	2008	0	smallint	2008
4	month	2	1	2	1.454756976	0.49794909	smallint	1; 2
5	day_of_mon	31	1	31	15.64767237	8.71671849	smallint	14; 25; 27; 8; 12; 17; 28; 15; 1; 10; 26; 11; 4; 18; 30; 16; 6; 19; 29; 21; 2; 3; 23; 31; 20; 5; 13; 22; 9; 7; 24
6	day_of_week	7	1	7	3.900422142	1.94035508	smallint	6; 4; 5; 1; 2; 7; 3
7	dep	70232	01/01/2008 00:00	29/02/2008 23:59	N/A	N/A	timestamp without time zone	
8	dep_crs	61016	01/01/2008 00:10	29/02/2008 23:59	N/A	N/A	timestamp without time zone	
9	arr	0	None	None	N/A	N/A	timestamp without time zone	None
10	arr_crs	71100	01/01/2008 00:00	29/02/2008 23:59	N/A	N/A	timestamp without time zone	
11	carrier	20	9E	YV	N/A	N/A	character	FL; OH; US; AS; OO; B6; 9E; CO; AQ; HA; XE; NW; DL; EV; MQ; YV; WN; UA; F9; AA
12	flight_num	7132	1	9740	2310.82087	1974.25791	smallint	
13	tail_num	4937		NHZOAL	N/A	N/A	character	
14	elapsed_time	635	-1	693	121.7318227	71.5420011	smallint	
15	elapsed_time	451	-9	1435	126.6794689	68.1159275	smallint	
16	air_time	610	-1	645	98.90645095	67.8274253	smallint	
17	delay_arr	784	-91	2461	11.20040144	40.546436	smallint	
18	delay_dep	763	-92	2457	12.16753825	37.3754576	smallint	
19	origin	286	ABE	YUM	N/A	N/A	character varying	
20	dest	287	ABE	YUM	N/A	N/A	character varying	
21	distance	1349	21	4962	703.3896382	544.819541	smallint	
22	taxi_in	143	-1	213	6.600512151	5.09845531	smallint	
23	taxi_out	257	-1	383	16.20731233	11.4317721	smallint	
24	cancelled	2	N/A	N/A	N/A	N/A	boolean	False; True
25	cancellation_	5		D	N/A	N/A	character	B; A; D; C;
26	diverted	2	N/A	N/A	N/A	N/A	boolean	False; True
27	delay_carrier	639	-1	1455	3.364172817	21.5347329	smallint	
28	delay_weath	431	-1	1148	0.107864986	10.6567594	smallint	
29	delay_nas	431	-1	1357	3.418187219	17.4503411	smallint	
30	delay_securi	94	-1	149	-0.71913411	0.98303751	smallint	
31	delay_late_a	438	-1	1143	4.960390639	22.5465128	smallint	

FOCUS & CONTEXT SUBSET

- Given a record R, with a set of attributes a_1, a_2, a_3,..., a_{29}
- Divide attributes into three groups:
  - Conditional
  - Variants of Normality

	A	B
1	**Attribute**	**Notes**
2	year	
3	month	
4	day_of_month	
5	day_of_week	
6	dep_time	Planned departure time
7	arr_time	Planned arrival time
8	dep_crs_offset	CRS departure offset from planned (mins)
9	arr_crs_offset	CRS arrival time offset from planned (mins)
10	carrier	Airline carrier code
11	flight_num	
12	tail_num	Unique aircraft identifier (around 800 total)
13	elapsed_time_actual	Actual elapsed time
14	elapsed_time_crs	Elapsed time according to CRS
15	air_time	Time in air
16	delay_arr	Delay arriving
17	delay_dep	Delay departing
18	origin	Origin airport code
19	dest	Dest airport code
20	distance	Miles (rounded)
21	taxi_in	Mins spent taxiing in
22	taxi_out	Mins spent taxiing out
23	cancelled	Was cancelled?
24	cancellation_code	Five possible codes
25	diverted	Was diverted?
26	delay_carrier	Delay caused by carrier (mins)
27	delay_weather	Delay caused by weather (mins)
28	delay_nas	Delay caused by NAS (mins)

QUERYING A FACAS

- *Bind* conditionals (origin, dest)
- Query data source for all records satisfying condition
- Each unique combination of variants in result set is considered a letter in the alphabet

Origin: LAX
Dest: JFK

Carrier, ElapsedTime, AirTime, TailNum, TaxiOut, Cancelled, CancellationCode, Diverted

An *instantiated* FaCAS

Query – flight_database on postgres@localhost:5432 *

flight_database on postgres@localhost:5

SQL Editor | Graphical Query Builder

Previous queries | Delete | Delete All

```
select * from delay where origin = 'LAX' and dest = 'JFK'
```

Scratch pad

Output pane

Data Output | Explain | Messages | History

	id integer	year smallint	month smallint	day_of_month smallint	day_of_week smallint	dep timestamp without time zone	dep_crs timestamp without time zone	arr timestamp without time zone	arr_crs timestamp without time
1	218650	2008	1	1	2	2008-01-01 06:13:00	2008-01-01 06:15:00	2008-01-01 14:07:00	2008-01-01 14:31:00
2	218651	2008	1	2	3	2008-01-02 06:15:00	2008-01-02 06:15:00	2008-01-02 14:35:00	2008-01-02 14:31:00
3	218652	2008	1	3	4	2008-01-03 06:07:00	2008-01-03 06:15:00	2008-01-03 14:54:00	2008-01-03 14:31:00
4	218653	2008	1	4	5	2008-01-04 06:18:00	2008-01-04 06:15:00	2008-01-04 15:23:00	2008-01-04 14:31:00
5	218654	2008	1	5	6	2008-01-05 06:15:00	2008-01-05 06:15:00	2008-01-05 14:16:00	2008-01-05 14:31:00
6	218655	2008	1	6	7	2008-01-06 06:08:00	2008-01-06 06:15:00	2008-01-06 14:22:00	2008-01-06 14:31:00
7	218662	2008	1	1	2	2008-01-01 08:40:00	2008-01-01 08:25:00	2008-01-01 17:05:00	2008-01-01 16:48:00
8	218663	2008	1	2	3	2008-01-02 08:21:00	2008-01-02 08:25:00	2008-01-02 16:34:00	2008-01-02 16:48:00
9	218664	2008	1	3	4	2008-01-03 08:16:00	2008-01-03 08:25:00	2008-01-03 16:26:00	2008-01-03 16:48:00
10	218665	2008	1	4	5	2008-01-04 08:29:00	2008-01-04 08:25:00	2008-01-04 16:38:00	2008-01-04 16:48:00
11	218666	2008	1	5	6	2008-01-05 08:25:00	2008-01-05 08:25:00	2008-01-05 16:53:00	2008-01-05 16:48:00
12	218667	2008	1	6	7	2008-01-06 08:23:00	2008-01-06 08:25:00	2008-01-06 16:44:00	2008-01-06 16:48:00
13	218668	2008	1	7	1	2008-01-07 08:14:00	2008-01-07 08:10:00	2008-01-07 16:06:00	2008-01-07 16:22:00
14	218669	2008	1	8	2	2008-01-08 08:05:00	2008-01-08 08:10:00	2008-01-08 16:20:00	2008-01-08 16:22:00

OK. Unix Ln 1, Col 33, Ch 33 940 rows. 80 ms

Rows: $R_1, R_2, R_3, \ldots, R_N$

Alphabet: $Z_1, Z_2, Z_3, \ldots, Z_M$

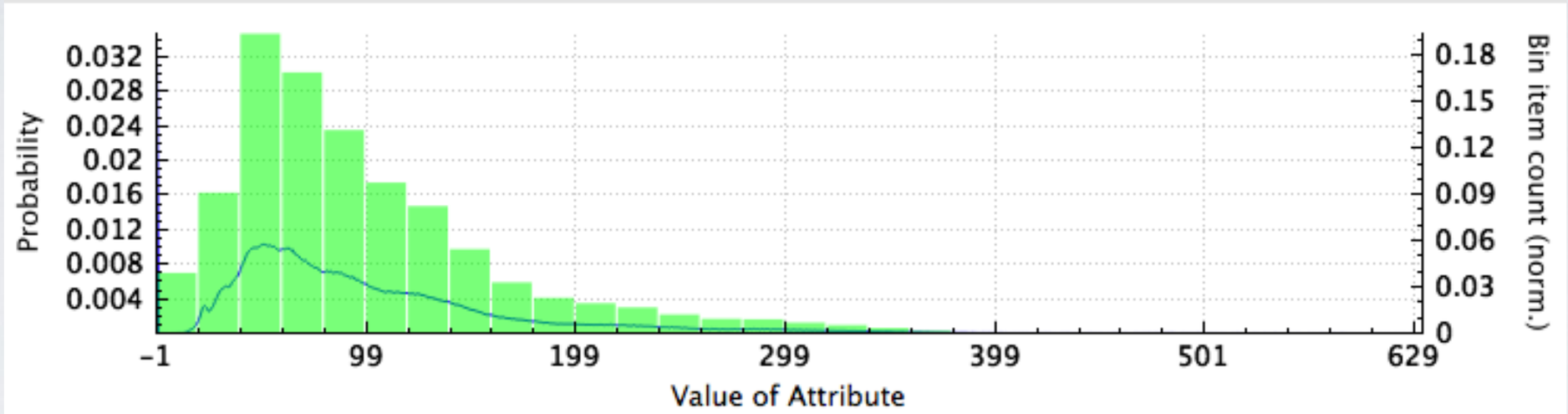
$(M \leq N)$

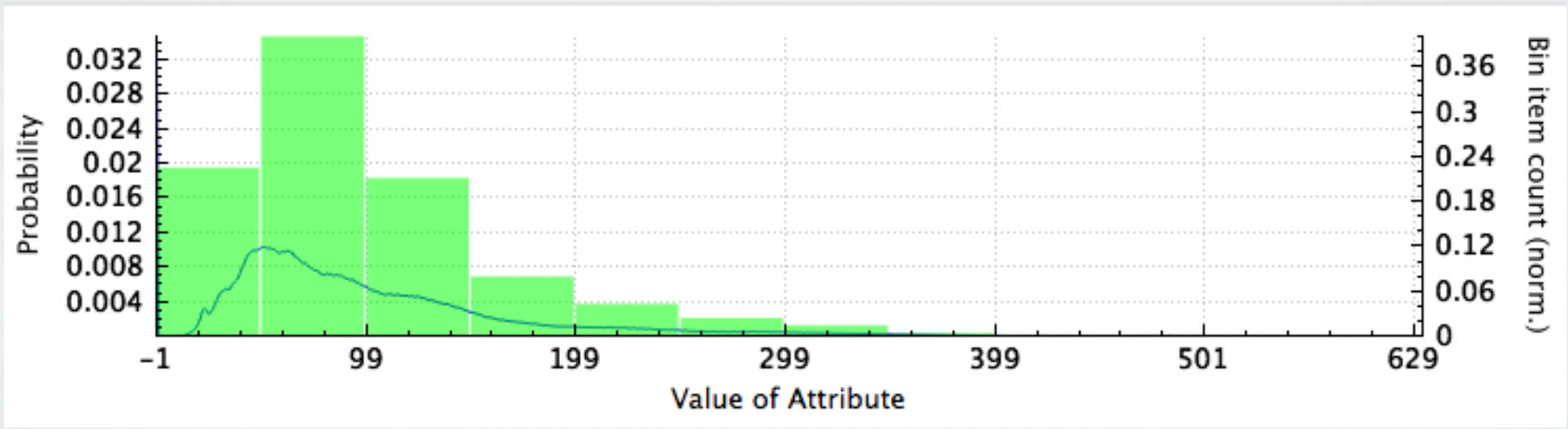
REDUCING THE VON SPACE: BINNING

- No 'continuous' value types in the flight database; however:
  - Obviously the VON space needs to be reduced otherwise there are far too many 'unique' combinations!
- How to choose sensible bin widths?
  - Should be a user-guided procedure as it relies upon specialist knowledge of the data's present and future ranges

REDUCING THE VON SPACE: VISUAL FEEDBACK

Smaller bins





Larger bins

BINNING FOR DATATYPES

- *Boolean*: uhh... let's not bother :)
- *Integer*: Divide into **N** bins? Or fix a width of **B**?
- *Datetime*: Semantic meaning; fixing a width seems logical
  - **B** seconds? **B** minutes? **B** hours? ...
- *String*: Depends on the semantics of the attribute

INITIAL GOALS

- Confirm the FaCAS maths / implementation
  - Initially in Python
  - Further implementation in C++
- Show surprisal in a basic visualization
  - Completed using Qt



IMPLEMENTATION

With Postgres / pqxx (Postgres C++ API)

SERVER OR CLIENT?

- Where should the bulk of the computation occur?
  - *Server side*: DB Server performs all necessary aggregation
  - *Client side:* Run individual queries through DB server, gather records on client and aggregate using results

SERVER OR CLIENT?

- Answer depends on:
  - Latency & bandwidth between client & server
  - Computational power / memory of server verses client
- In our tests with a localhost DB server, client-based method quicker than server-based method

CLIENT PSEUDOCODE

```
F = facas(conditionals={...},vons={...})
Z = hash<row_hash,int>
R = server.query(f.conditionals())

// compute Z
for (row in R)
  Z[hash(row.vons)].count++

// compute entropy
for(letter in Z)
  letter.prob = letter.count / |Z|
  letter.surp = -log2(letter.prob)
  HZ += letter.prob * -letter.surp
```



SERVER IMPLEMENTATION

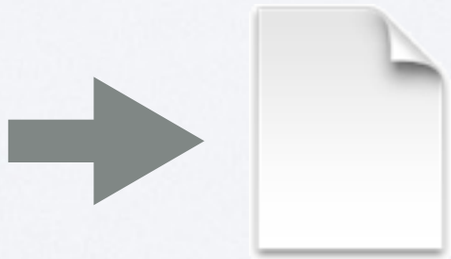
- Implemented as a Stored Procedure (SP) in PSQL
- Client calls SP through API
- API returns result when DB completes

```
-- Function: facas_server(text, text, text)

-- DROP FUNCTION facas_server(text, text, text);

CREATE OR REPLACE FUNCTION facas_server(table_name text, condition text, von_hash text)
  RETURNS record AS
$BODY$
declare
       sql text;
       total_recs integer;
       N real;
       rec record;
begin

sql = 'select count(id) FROM ' || table_name || ' WHERE ' || condition;
execute sql into total_recs;

raise info 'A) %', sql;

if total_recs = 0 THEN
       execute 'select cast(0 as bigint)  as totalRowCounts, cast(0 as bigint)  as Zcount, cast(0 as numeric) as HZ, cast(0 as numeric)
       return rec;
end if;

N = cast(total_recs as real);

sql = 'select cast(' || N || ' as bigint) as totalRowCount, count(t.cnt) as Zcount,
       cast(-SUM(_log2(cast(t.cnt/cast(' || N ||' as real) as real)) * (cast(t.cnt as real)/ ' || N || ')) as numeric) AS HZ,
       cast(SUM(-_log2(cast(t.cnt/cast(' || N ||' as real) as real))) as numeric) AS sum_surprise
from (
select count(id) as cnt
 from ' || table_name ||
 ' where  ' || condition ||
 ' group by ' || von_hash ||
') t';

raise info 'B) %', sql;

 execute sql into rec;
 return rec;

end
$BODY$
  LANGUAGE plpgsql VOLATILE
  COST 100;
ALTER FUNCTION facas_server(text, text, text)
  OWNER TO postgres;
```

SERVER PERFORMANCE

- Initial shock: server-side implementation ~5x *slower* than client-side implementation
- Culprit: built-in log() function too slow
- Luckily: can plug-in our own C log2 implementation from an object file and get Postgres to call it

```c
#include <math.h>

#include <postgres.h>
#include <fmgr.h>

#ifdef PG_MODULE_MAGIC
PG_MODULE_MAGIC;
#endif

PG_FUNCTION_INFO_V1(_log2);
Datum _log2(PG_FUNCTION_ARGS) {
    PG_RETURN_FLOAT4(log2(PG_GETARG_FLOAT4(0)));
}
```





SERVER PERFORMANCE

- Other performance consideration: choice of hash function for computing unique sets of variants of normality

- Hashing quick in C++ using std::map & std::string concat

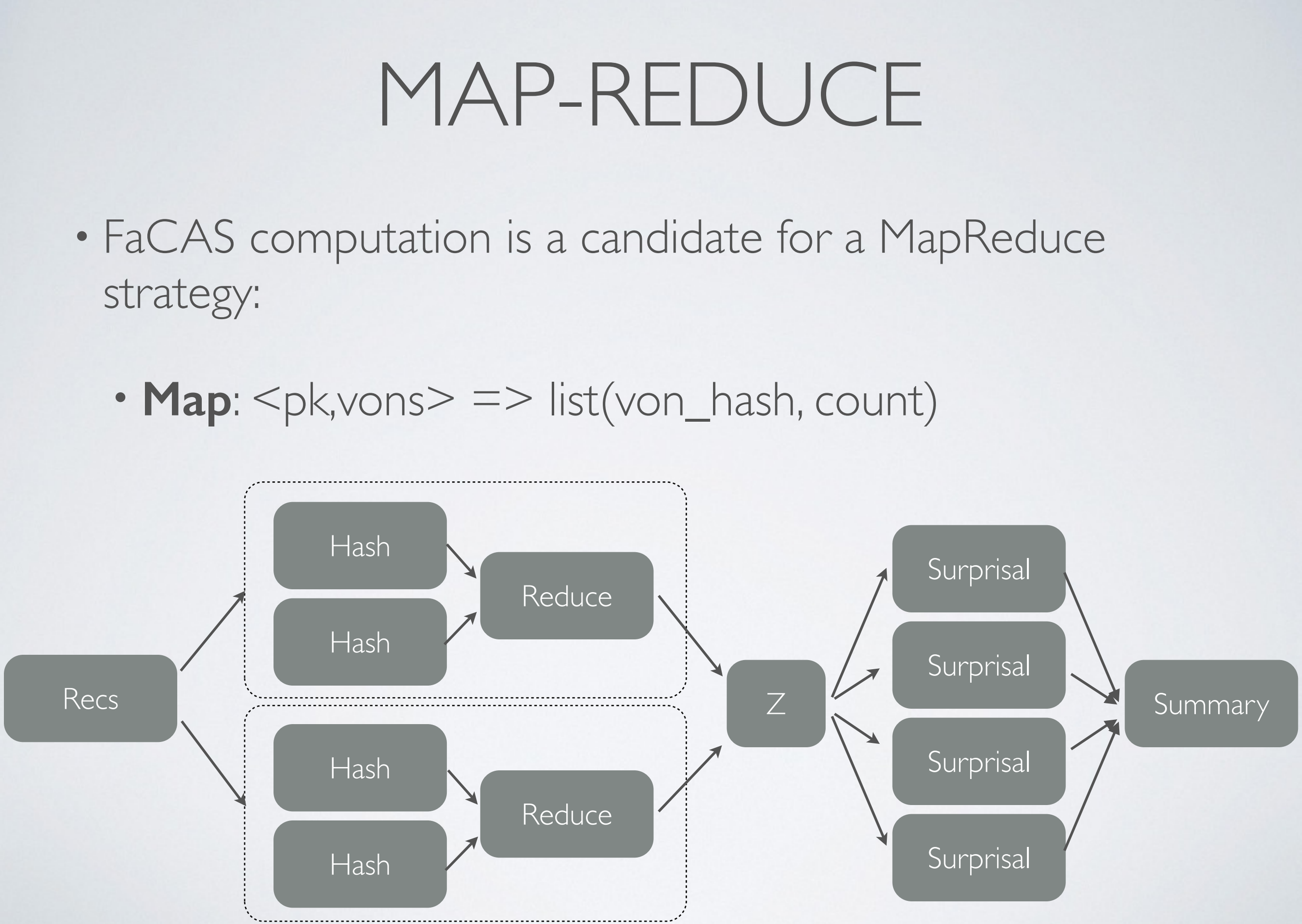
- Not so quick in Postgres!

Origin: LAX
Dest: JFK

Carrier, ElapsedTime, AirTime, TailNum, TaxiOut, Cancelled, CancellationCode, Diverted

MAP-REDUCE

- FaCAS computation is a candidate for a MapReduce strategy:
  - **Map**: <pk,vons> => list(von_hash, count)

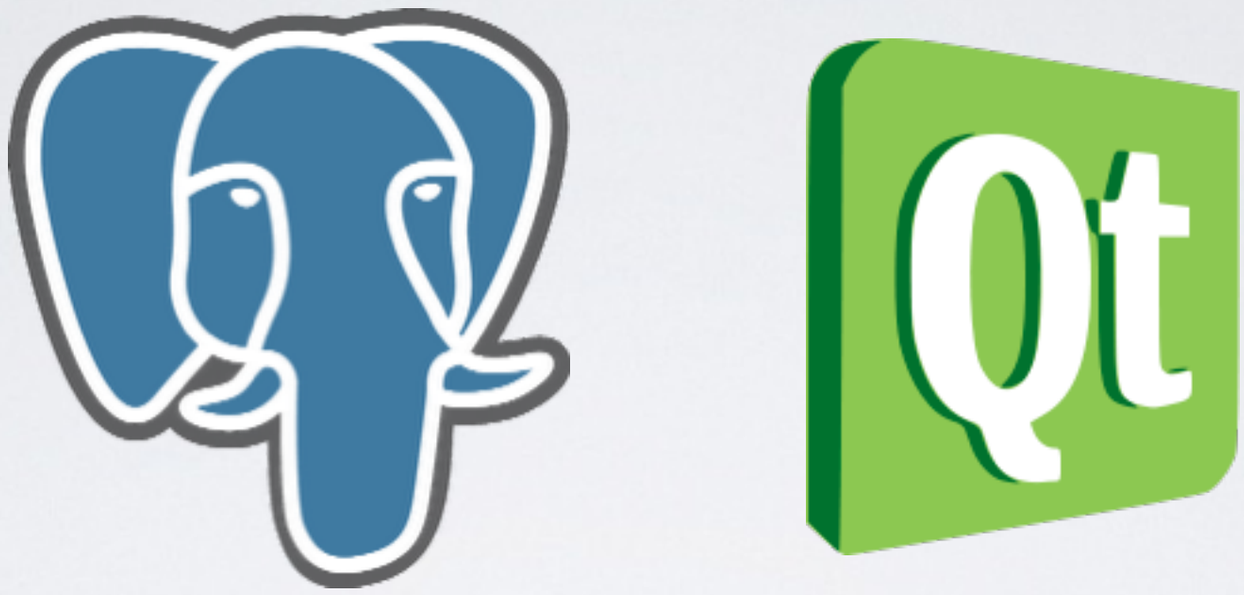


INDEXING

- Candidates for indexes:
  - All fields used by any FaCAS condition (inc. multicolumn combinations)
  - Bin formulae (i.e. `to_int(integer/20)`)
- All of these currently created automatically by system

PERFORMANCE, OVERALL

- 1,000,000 records surprisal computation: ~1.2 sec
- Performance largely dependent upon:
  - Initial conditional matching (quicker if indexed)
  - Number of VONS to hash together

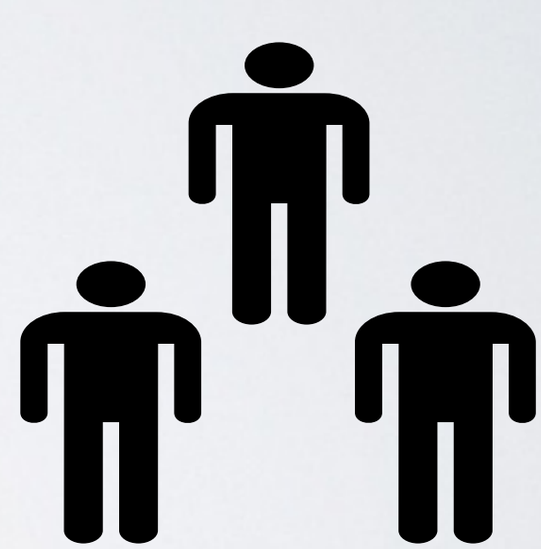


IMPLEMENTATION

Graphical User Interface

MOTIVATION

- In our system, we imagine three user roles:
  - *Observer*: observes incoming data
  - *Analyst*: analyses new and existing data
  - *Designer*: designs appropriate FaCASs



SYSTEM FLOW

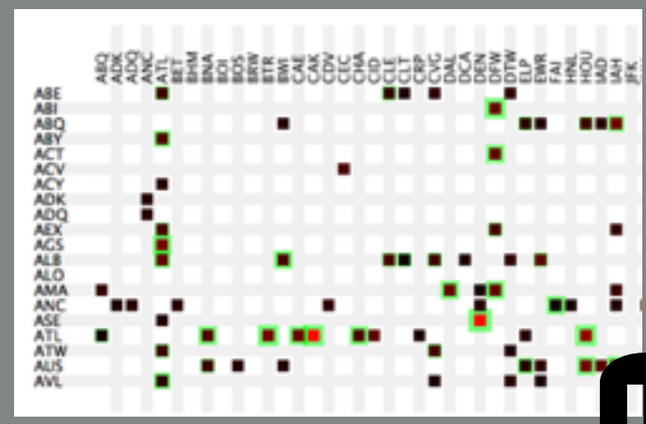
Designer Software

FaCAS Store

F1 F2 Fn

Analytical Engine

Vis



FaCAS Summaries

Origin: LAX
Dest: JFK

FACAS DESIGNER

- Should provide the user with useful statistics for each field in order to assist user decision in:
  - Denoting fields as conditionals or VONs
  - Choosing the bin widths for the conditionals and VONs
- Should allow test runs of FaCAS
- Save/load functionality

FaCAS Designer

Attribute List:

Attribute	Bin Spec	Role
id	10 integers	None
year	10 integers	None
month	10 integers	None
day_of_month	10 integers	None
day_of_week	10 integers	None
dep	10 minutes	None
dep_crs	10 minutes	None
arr	10 minutes	None
arr_crs	10 minutes	None
carrier	1 string	None
flight_num	10 integers	None
tail_num	1 string	None
elapsed_tim...	10 integers	None
elapsed_tim...	10 integers	None
air_time	10 integers	None
delay_arr	10 integers	None
delay_dep	10 integers	None
origin	1 string	None
dest	1 string	None
distance	10 integers	None
taxi_in	10 integers	None
taxi_out	10 integers	None
cancelled	1 boolean	None
cancellation...	1 string	None
diverted	1 boolean	None
delay_carrier	10 integers	None
delay_weat...	10 integers	None
delay_nas	10 integers	None
delay_security	10 integers	None
delay_late_...	10 integers	None

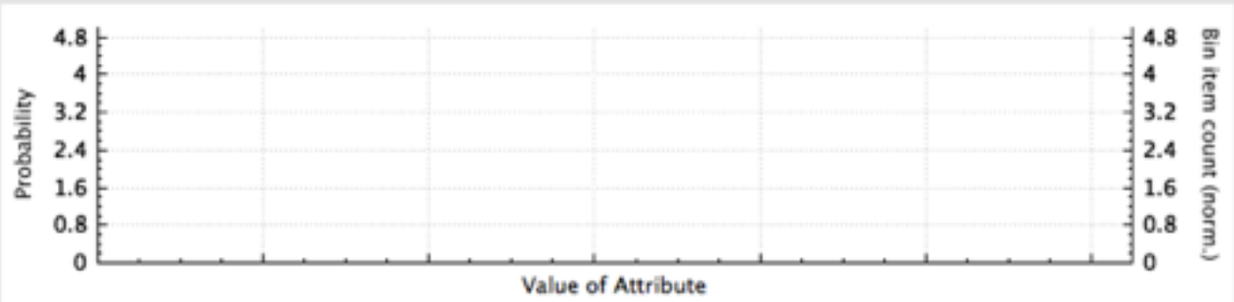
Conditionals:

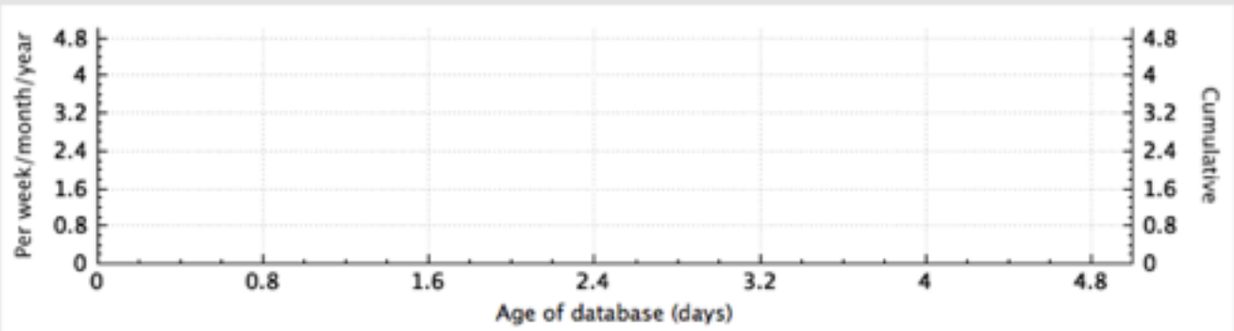
Attribute	Op	Value

Field statistics (no field selected)

Min: - Max: - Unique: - Avg: - StdDev: -

-





Bin Specification

Width: = # Bins: Maintain constant number of bins with incoming records

FaCAS Statistics

Message:

Entropy: Avg Surprise: Alphabet size: Total matches:

Wall time: Uncertainty:

FaCAS Designer

Attribute List:

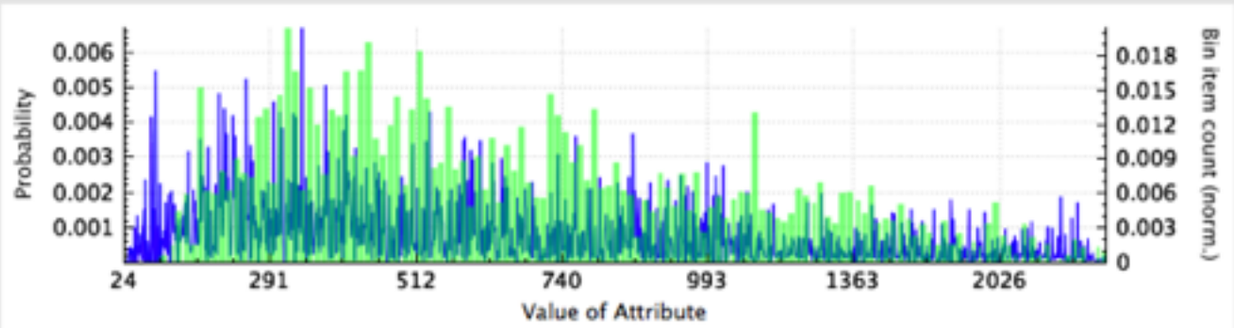
Attribute	Bin Spec	Role
id	10 integers	None
year	10 integers	None
month	10 integers	None
day_of_month	10 integers	None
day_of_week	10 integers	None
dep	10 minutes	None
dep_crs	10 minutes	None
arr	10 minutes	None
arr_crs	10 minutes	None
carrier	1 string	None
flight_num	10 integers	None
tail_num	1 string	None
elapsed_tim...	10 integers	None
elapsed_tim...	10 integers	None
air_time	10 integers	None
delay_arr	10 integers	None
delay_dep	10 integers	None
origin	1 string	None
dest	1 string	None
distance	10 integers	None
taxi_in	10 integers	None
taxi_out	10 integers	None
cancelled	1 boolean	None
cancellation...	1 string	None
diverted	1 boolean	None
delay_carrier	10 integers	None
delay_weat...	10 integers	None
delay_nas	10 integers	None
delay_security	10 integers	None
delay_late_...	10 integers	None

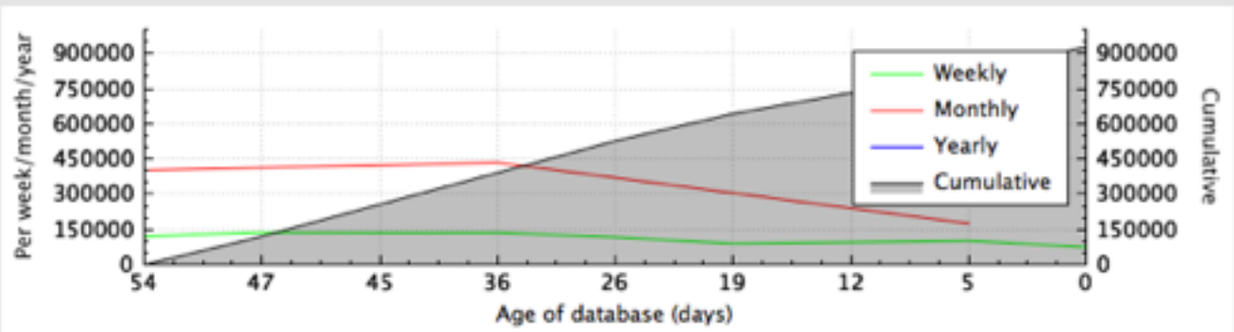
Conditionals:

Attribute	Op	Value

Stats for field: distance

Min: 24.00 Max: 4962.00 Unique: 1347 Avg: 689.36 StdDev: 539.06





Bin Specification

Width: 10 integers = # Bins: 134 Maintain constant number of bins with incoming records

FaCAS Statistics

Message: No VONs yet; FaCAS not run.

Entropy: 0 Avg Surprise: 0 Alphabet size: 0 Total matches: 0

Wall time: 0 Uncertainty: 0

FaCAS Designer

Attribute List:

Attribute	Bin Spec	Role
id	10 integers	None
year	10 integers	None
month	10 integers	None
day_of_month	10 integers	None
day_of_week	10 integers	None
dep	10 minutes	None
dep_crs	10 minutes	None
arr	10 minutes	None
arr_crs	10 minutes	None
carrier	1 string	None
flight_num	10 integers	None
tail_num	1 string	None
elapsed_tim...	10 integers	None
elapsed_tim...	10 integers	None
air_time	10 integers	None
delay_arr	10 integers	None
delay_dep	10 integers	None
origin	1 string	Conditional
dest	1 string	Conditional
distance	10 integers	VON
taxi_in	10 integers	VON
taxi_out	10 integers	None
cancelled	1 boolean	None
cancellation...	1 string	VON
diverted	1 boolean	None
delay_carrier	10 integers	None
delay_weat...	10 integers	None
delay_nas	10 integers	None
delay_security	10 integers	None
delay_late_...	10 integers	None

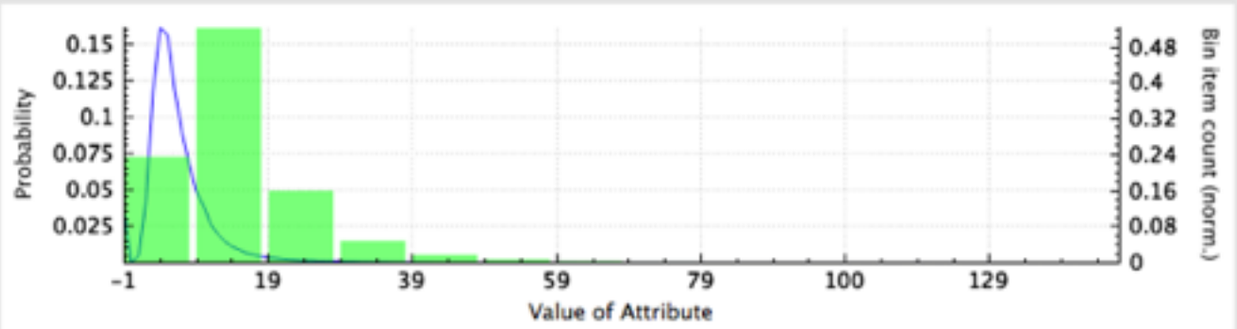
Conditionals:

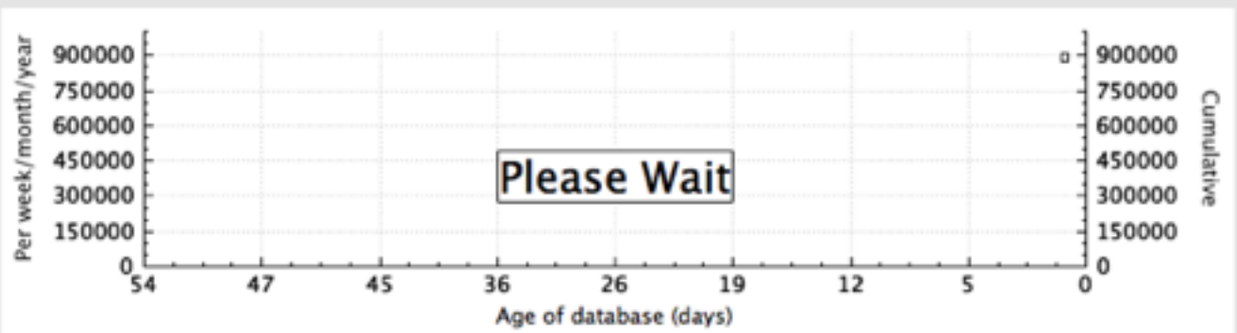
Attribute	Op	Value
dest	=	
origin	=	

Field statistics (no field selected)

Min: - Max: - Unique: - Avg: - StdDev: -

-





Bin Specification

Width: 1 string = # Bins: 286 Maintain constant number of bins with incoming records

FaCAS Statistics

Message: A conditional is not complete; FaCAS not run.

Entropy: 0 Avg Surprise: 0 Alphabet size: 0 Total matches: 0

Wall time: 0 Uncertainty: 0

FaCAS Designer

Attribute List:

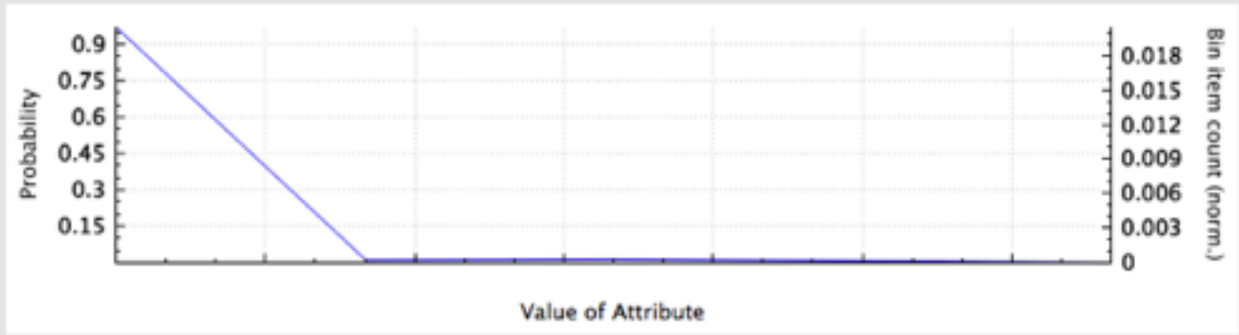
Attribute	Bin Spec	Role
id	10 integers	None
year	10 integers	None
month	10 integers	None
day_of_month	10 integers	None
day_of_week	10 integers	None
dep	10 minutes	None
dep_crs	10 minutes	None
arr	10 minutes	None
arr_crs	10 minutes	None
carrier	1 string	None
flight_num	10 integers	None
tail_num	1 string	None
elapsed_tim...	10 integers	None
elapsed_tim...	10 integers	None
air_time	10 integers	None
delay_arr	10 integers	None
delay_dep	10 integers	None
origin	1 string	Conditional
dest	1 string	Conditional
distance	10 integers	VON
taxi_in	10 integers	VON
taxi_out	10 integers	None
cancelled	1 boolean	None
cancellation...	1 string	VON
diverted	1 boolean	None
delay_carrier	10 integers	None
delay_weat...	10 integers	None
delay_nas	10 integers	None
delay_security	10 integers	None
delay_late_...	10 integers	None

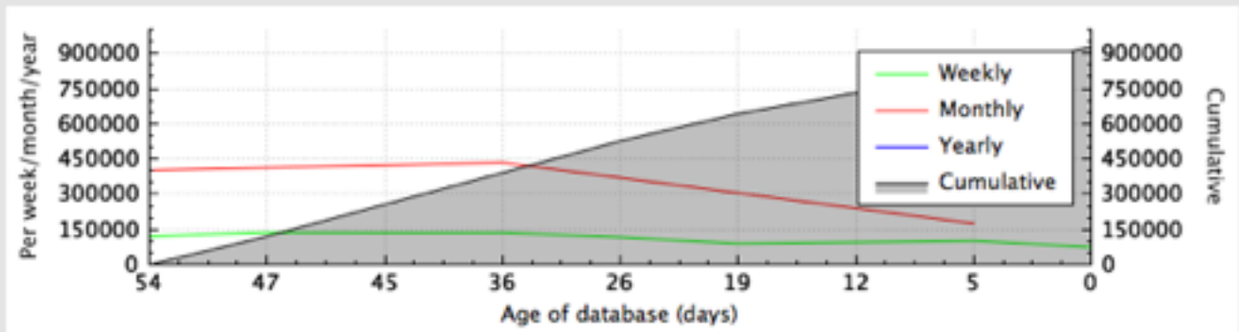
Conditionals:

Attribute	Op	Value
dest	=	LAX
origin	=	JFK

Stats for field: distance

Min: 24.00 Max: 4962.00 Unique: 1347 Avg: 689.36 StdDev: 539.06





Bin Specification

Width: 10 integers = # Bins: 134 Maintain constant number of bins with incoming records

FaCAS Statistics

Message: Successfully run FaCAS.

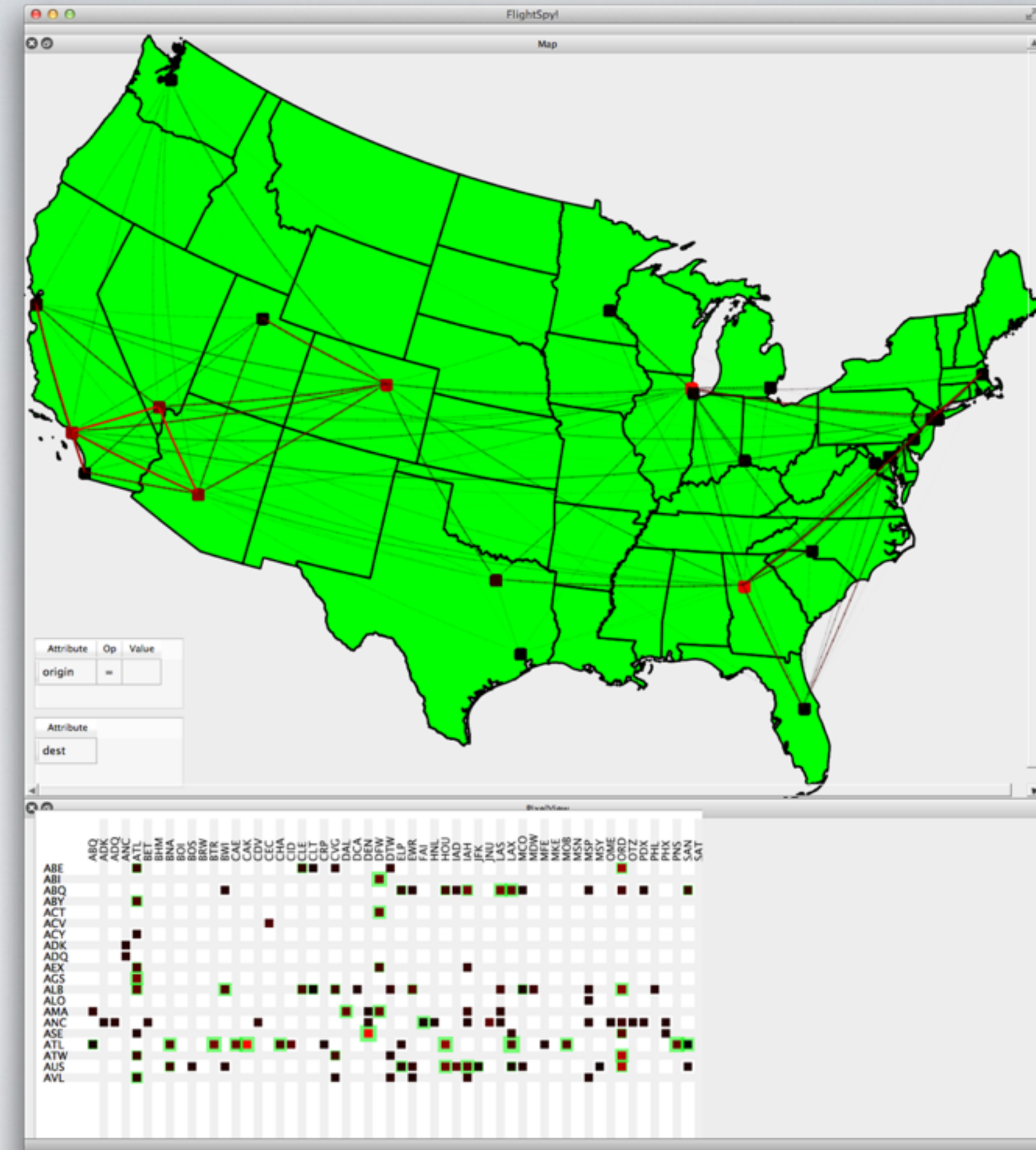
Entropy: 1.6493 Avg Surprise: 6.56186 Alphabet size: 11 Total matches: 940

Wall time: 0 Uncertainty: 0.166992



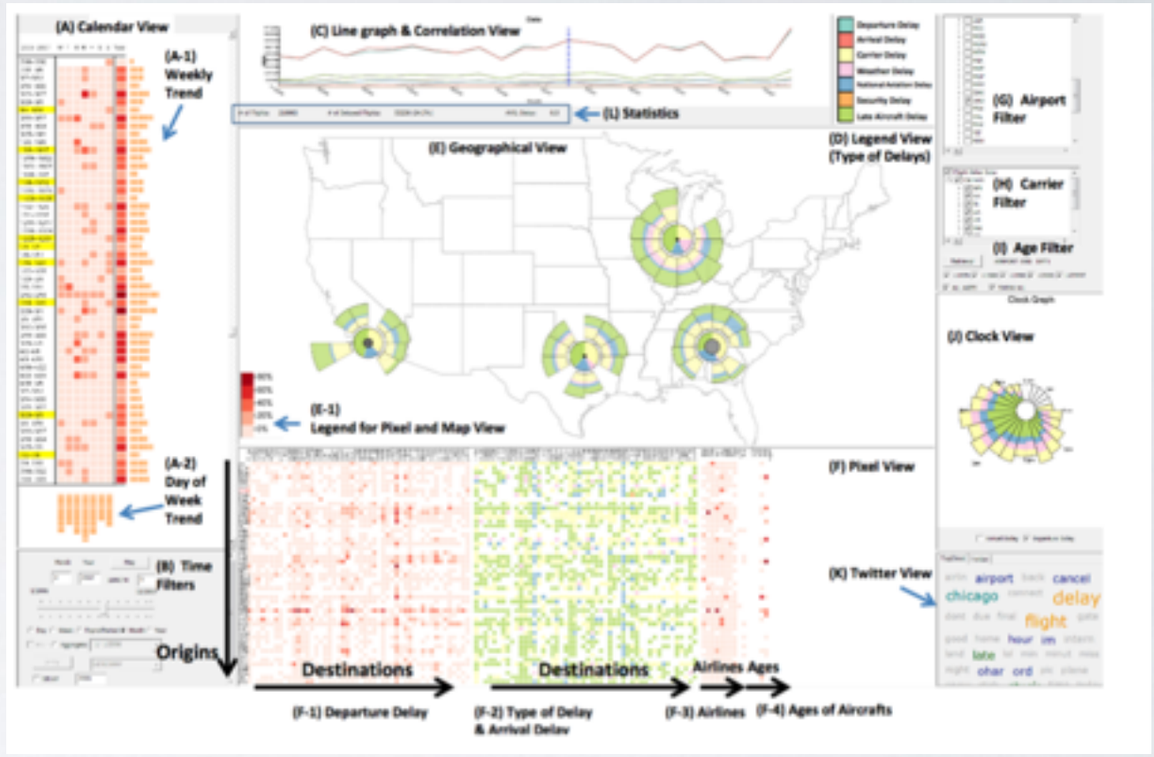
VISUALISATIONS

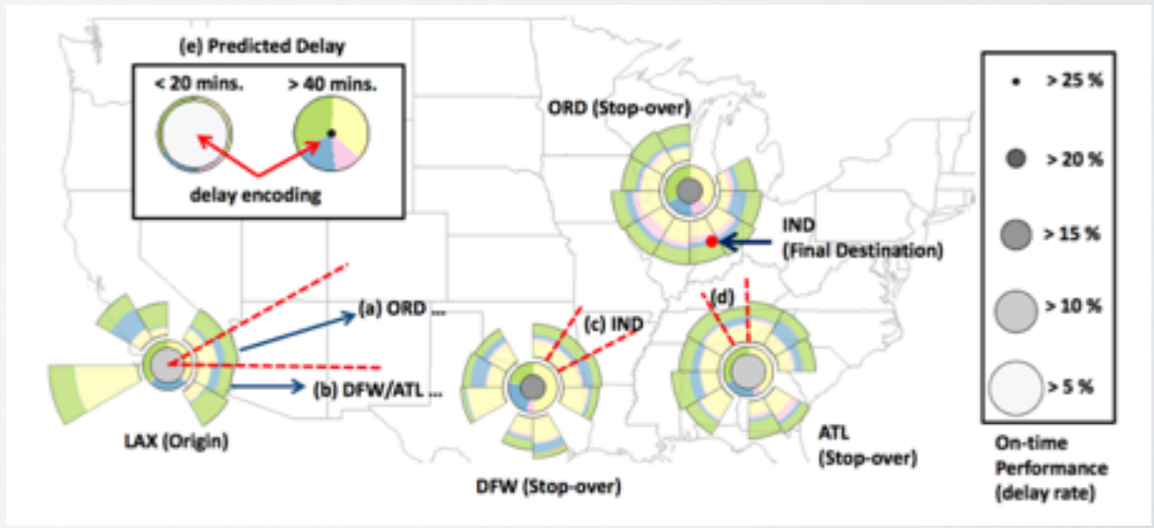
- Two initial visualisations using flight data and FaCAS surprisal highlighting strategy
  - Pixel grid of origin verses destination airport
  - US map view


FlightSpy!
Map
Attribute
Op
Value
origin
=
Attribute
dest
ABQ ADK ADQ ANC ATL BET BHM BNA BOI BOS BRW BTR BWI CAE CAK CDV CEC CHA CID CLE CLT CRP CVG DAL DCA DEN DFW DTW ELP EWR FAI HNL HOU IAD IAH JFK JNU LAS LAX MCO MDW MFE MKE MOB MSN MSP MSY OME ORD OTZ PDX PHL PHX PNS SAN SAT
ABE
ABI
ABQ
ABY
ACT
ACV
ACY
ADK
ADQ
AEX
AGS
ALB
ALO
AMA
ANC
ASE
ATL
ATW
AUS
AVL

CONCLUSION

- Two GUI prototypes developed to support the proposed FaCAS workflow
- Working towards integration with existing software developed by Purdue (Ko *et al.*),
- Speed of FaCAS aggregation not an issue, even on local machine





QUESTIONS?

AN INFORMATION THEORETIC APPROACH TO THE FLIGHT DATA PROBLEM

Min Chen and Simon Walton
Oxford University

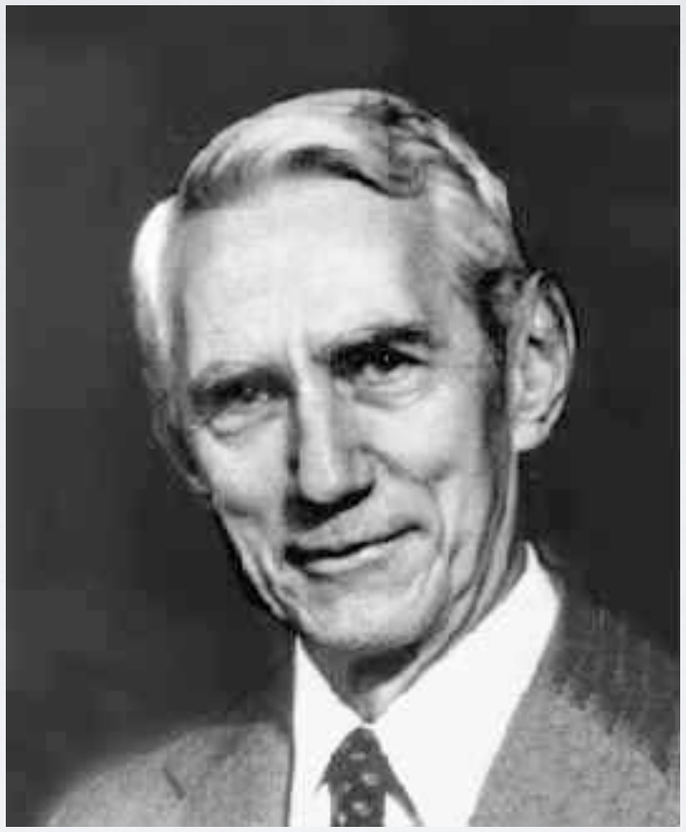






CONTENTS

- Information Theory: A Crash Course
- Our Information-Theoretic Model: QCATs
- Algorithms, Implementation and Scalability
- User Interaction



Claude Shannon

INFORMATION THEORY

Crash Course

INFORMATION THEORY

- The number of all possible answers is the amount of information
- Roll a fair dice: 6 outcomes, equiprobable
  - What if I roll it *n* times?
- We can make information additive:

$$log_b(r^n) = n \cdot log_b(r)$$

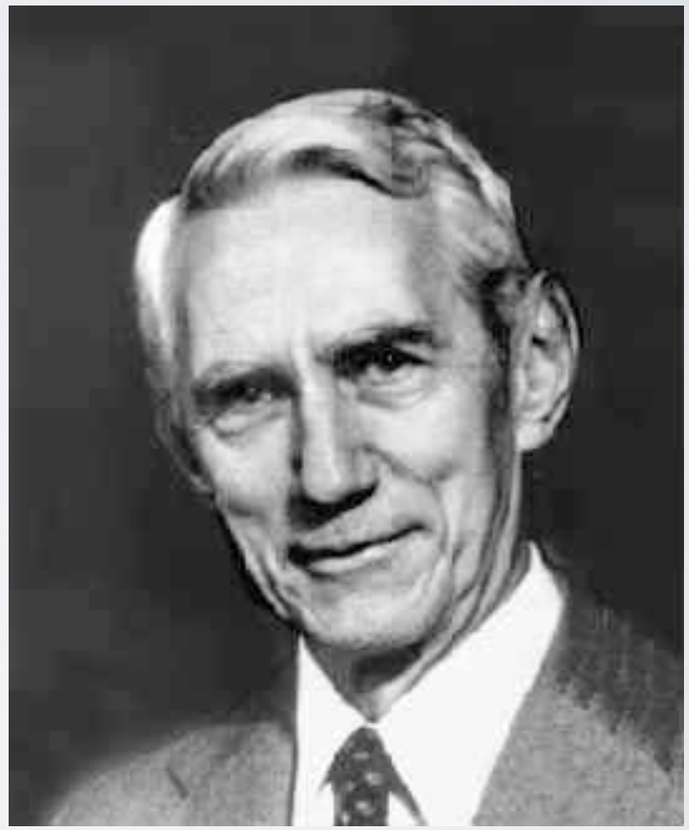




INFORMATION THEORY

- What if not all events are equiprobable?

- Intuitively we get more ‘information’ for the less probable outcomes since they are more surprising



CUPS AND BALLS

- Game: cups and balls
- 1/3 chance of getting the ball



  - ✔ $log_2(\frac{1}{\frac{1}{3}}) = log_2(\frac{3}{1}) = log_2(3) = \approx 1.58$
  - ✖ $log_2(\frac{3}{2}) = \approx 0.58$

- What is the amount of information then in the answer?

CUPS AND BALLS



- Average of all outcomes - i.e. weight according to their probabilities:

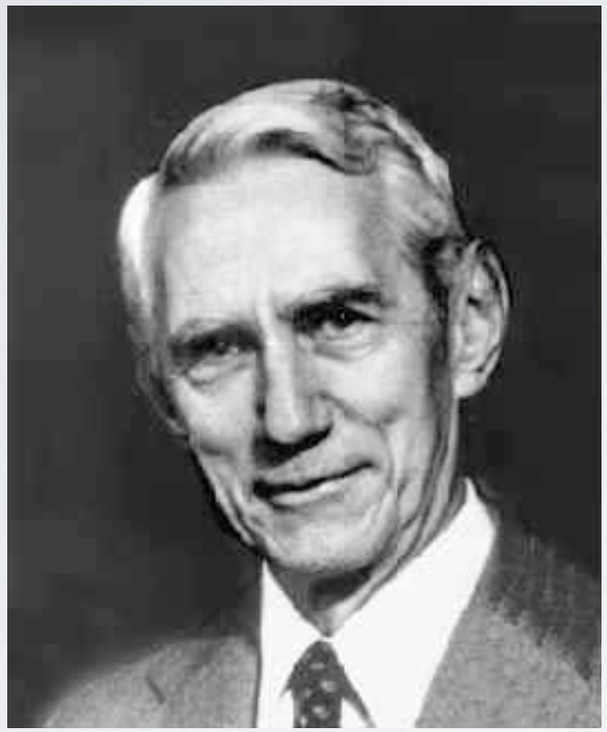
$$\frac{1}{3}log_2(3) + \frac{2}{3}log_2(\frac{3}{2}) =\approx 0.92$$

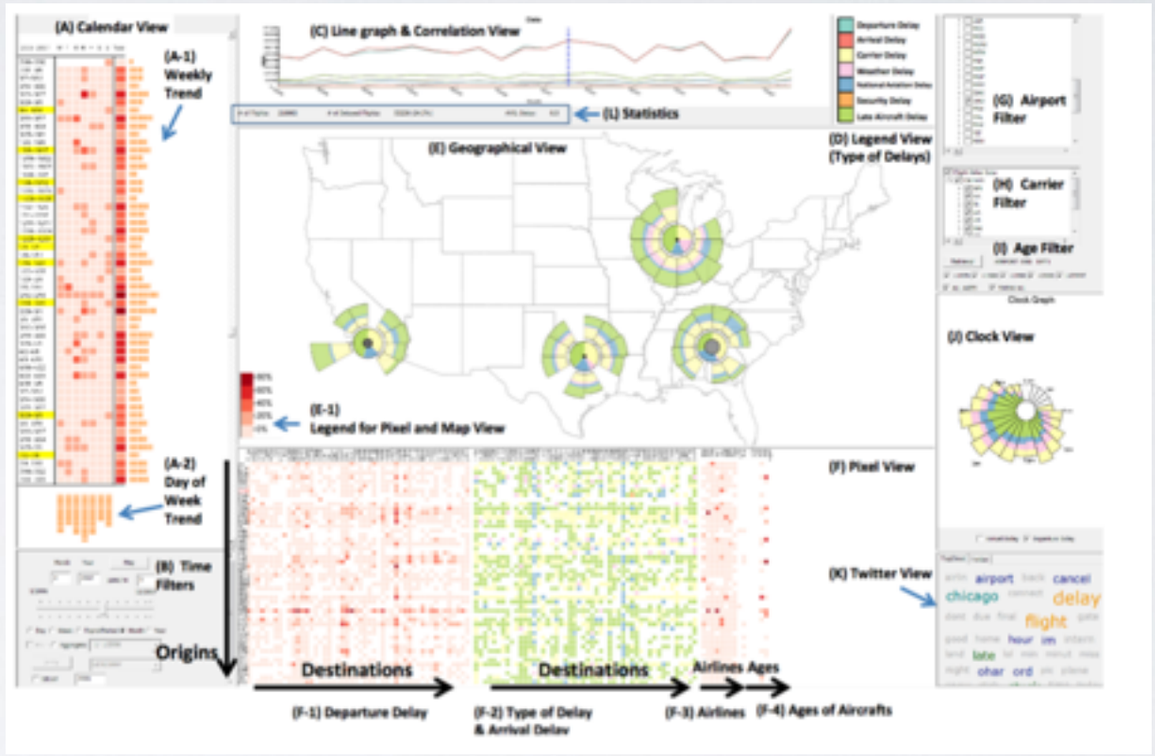
- More generally:

$$\sum_{i=1}^{r} p_i \cdot log_2(\frac{1}{p_i})$$

INFORMATION IS…

- The `surprise' that we get from learning an outcome
- Not all information is created equal
  - Not everything in the real world is equiprobable
  - Some pieces of information can be less than a bit
  - But even one bit of information can be huge





QCATS

Our Information Theoretic Model for Anomaly Detection

VISUAL ANALYTICS TASKS

- Given a flight record, inform whether it's normal or abnormal.
- Show all `abnormal' records.
- What are the key indicators for a specific type of phenomenon, problem, or risk?

WHY NOT MACHINE LEARNING?

- Insufficient anomalies to train
- Uncertainty in annotation
- Importance of different dimensions unknown/uncertain
- Context-sensitive knowledge not taken into account

FLIGHT DATA PROBLEM

- 29 attributes in **delay** table
- 7 additional attributes in linked **airport** table
- Some attribute pairs have a high mutual information rating

INFORMATION THEORY

- Surprisal measure $-\log_2(p(R))$
- Entropy measure $\boldsymbol{H}(X) = -\sum_{i}^{m} p(x_i)\log_2 p(x_i)$
  - I.E. the weighted sum of each element's entropy

	field	distinct	min	max	avg	stddev	data type	example values
2	id	1111000	1	1111000	555500.5	320718.219	integer	
3	year	1	2008	2008	2008	0	smallint	2008
4	month	2	1	2	1.454756976	0.49794909	smallint	1; 2
5	day_of_mon	31	1	31	15.64767237	8.71671849	smallint	14; 25; 27; 8; 12; 17; 28; 15; 1; 10; 26; 11; 4; 18; 30; 16; 6; 19; 29; 21; 2; 3; 23; 31; 20; 5; 13; 22; 9; 7; 24
6	day_of_week	7	1	7	3.900422142	1.94035508	smallint	6; 4; 5; 1; 2; 7; 3
7	dep	70232	01/01/2008 00:00	29/02/2008 23:59	N/A	N/A	timestamp without time zone	
8	dep_crs	61016	01/01/2008 00:10	29/02/2008 23:59	N/A	N/A	timestamp without time zone	
9	arr	0	None	None	N/A	N/A	timestamp without time zone	None
10	arr_crs	71100	01/01/2008 00:00	29/02/2008 23:59	N/A	N/A	timestamp without time zone	
11	carrier	20	9E	YV	N/A	N/A	character	FL; OH; US; AS; OO; B6; 9E; CO; AQ; HA; XE; NW; DL; EV; MQ; YV; WN; UA; F9; AA
12	flight_num	7132	1	9740	2310.82087	1974.25791	smallint	
13	tail_num	4937		NHZOAL	N/A	N/A	character	
14	elapsed_time	635	-1	693	121.7318227	71.5420011	smallint	
15	elapsed_time	451	-9	1435	126.6794689	68.1159275	smallint	
16	air_time	610	-1	645	98.90645095	67.8274253	smallint	
17	delay_arr	784	-91	2461	11.20040144	40.546436	smallint	
18	delay_dep	763	-92	2457	12.16753825	37.3754576	smallint	
19	origin	286	ABE	YUM	N/A	N/A	character varying	
20	dest	287	ABE	YUM	N/A	N/A	character varying	
21	distance	1349	21	4962	703.3896382	544.819541	smallint	
22	taxi_in	143	-1	213	6.600512151	5.09845531	smallint	
23	taxi_out	257	-1	383	16.20731233	11.4317721	smallint	
24	cancelled	2	N/A	N/A	N/A	N/A	boolean	False; True
25	cancellation_	5		D	N/A	N/A	character	B; A; D; C;
26	diverted	2	N/A	N/A	N/A	N/A	boolean	False; True
27	delay_carrier	639	-1	1455	3.364172817	21.5347329	smallint	
28	delay_weath	431	-1	1148	0.107864986	10.6567594	smallint	
29	delay_nas	431	-1	1357	3.418187219	17.4503411	smallint	
30	delay_securi	94	-1	149	-0.71913411	0.98303751	smallint	
31	delay_late_a	438	-1	1143	4.960390639	22.5465128	smallint	

QCAT - QUERY & CONDITIONAL ATTRIBUTES

- Given a record R, with a set of attributes a_1, a_2, a_3,..., a_{29}
- Divide attributes into three groups:
  - Conditional
  - Variants of Normality

	A	B
1	**Attribute**	**Notes**
2	year	
3	month	
4	day_of_month	
5	day_of_week	
6	dep_time	Planned departure time
7	arr_time	Planned arrival time
8	dep_crs_offset	CRS departure offset from planned (mins)
9	arr_crs_offset	CRS arrival time offset from planned (mins)
10	carrier	Airline carrier code
11	flight_num	
12	tail_num	Unique aircraft identifier (around 800 total)
13	elapsed_time_actual	Actual elapsed time
14	elapsed_time_crs	Elapsed time according to CRS
15	air_time	Time in air
16	delay_arr	Delay arriving
17	delay_dep	Delay departing
18	origin	Origin airport code
19	dest	Dest airport code
20	distance	Miles (rounded)
21	taxi_in	Mins spent taxiing in
22	taxi_out	Mins spent taxiing out
23	cancelled	Was cancelled?
24	cancellation_code	Five possible codes
25	diverted	Was diverted?
26	delay_carrier	Delay caused by carrier (mins)
27	delay_weather	Delay caused by weather (mins)
28	delay_nas	Delay caused by NAS (mins)

QUERYING A QCAT

- *Bind* conditionals (origin, dest)
- Query data source for all records satisfying condition
- Each unique combination of variants in result set is considered a letter in the alphabet

Origin: LAX
Dest: JFK

Carrier, ElapsedTime, AirTime, TailNum, TaxiOut, Cancelled, CancellationCode, Diverted

An *instantiated* QCAT

Query – flight_database on postgres@localhost:5432 *

flight_database on postgres@localhost:5

SQL Editor | Graphical Query Builder

Previous queries | Delete | Delete All

```
select * from delay where origin = 'LAX' and dest = 'JFK'
```

Scratch pad

Output pane

Data Output | Explain | Messages | History

	id integer	year smallint	month smallint	day_of_month smallint	day_of_week smallint	dep timestamp without time zone	dep_crs timestamp without time zone	arr timestamp without time zone	arr_crs timestamp without time
1	218650	2008	1	1	2	2008-01-01 06:13:00	2008-01-01 06:15:00	2008-01-01 14:07:00	2008-01-01 14:31:00
2	218651	2008	1	2	3	2008-01-02 06:15:00	2008-01-02 06:15:00	2008-01-02 14:35:00	2008-01-02 14:31:00
3	218652	2008	1	3	4	2008-01-03 06:07:00	2008-01-03 06:15:00	2008-01-03 14:54:00	2008-01-03 14:31:00
4	218653	2008	1	4	5	2008-01-04 06:18:00	2008-01-04 06:15:00	2008-01-04 15:23:00	2008-01-04 14:31:00
5	218654	2008	1	5	6	2008-01-05 06:15:00	2008-01-05 06:15:00	2008-01-05 14:16:00	2008-01-05 14:31:00
6	218655	2008	1	6	7	2008-01-06 06:08:00	2008-01-06 06:15:00	2008-01-06 14:22:00	2008-01-06 14:31:00
7	218662	2008	1	1	2	2008-01-01 08:40:00	2008-01-01 08:25:00	2008-01-01 17:05:00	2008-01-01 16:48:00
8	218663	2008	1	2	3	2008-01-02 08:21:00	2008-01-02 08:25:00	2008-01-02 16:34:00	2008-01-02 16:48:00
9	218664	2008	1	3	4	2008-01-03 08:16:00	2008-01-03 08:25:00	2008-01-03 16:26:00	2008-01-03 16:48:00
10	218665	2008	1	4	5	2008-01-04 08:29:00	2008-01-04 08:25:00	2008-01-04 16:38:00	2008-01-04 16:48:00
11	218666	2008	1	5	6	2008-01-05 08:25:00	2008-01-05 08:25:00	2008-01-05 16:53:00	2008-01-05 16:48:00
12	218667	2008	1	6	7	2008-01-06 08:23:00	2008-01-06 08:25:00	2008-01-06 16:44:00	2008-01-06 16:48:00
13	218668	2008	1	7	1	2008-01-07 08:14:00	2008-01-07 08:10:00	2008-01-07 16:06:00	2008-01-07 16:22:00
14	218669	2008	1	8	2	2008-01-08 08:05:00	2008-01-08 08:10:00	2008-01-08 16:20:00	2008-01-08 16:22:00

OK. Unix Ln 1, Col 33, Ch 33 940 rows. 80 ms

Rows: $R_1, R_2, R_3, \ldots, R_N$

Alphabet: $Z_1, Z_2, Z_3, \ldots, Z_M$

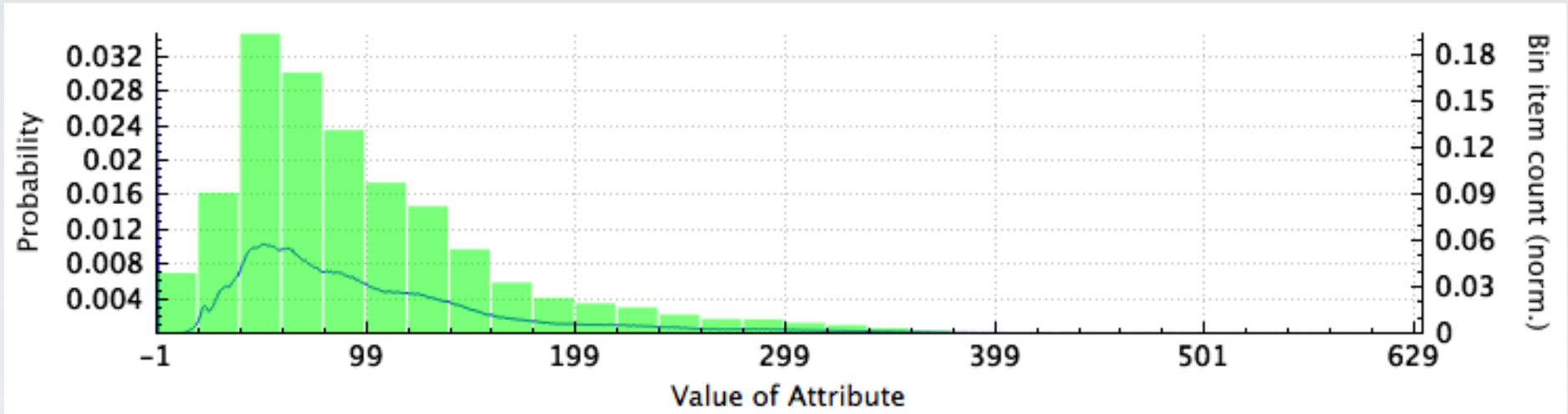
$(M \leq N)$

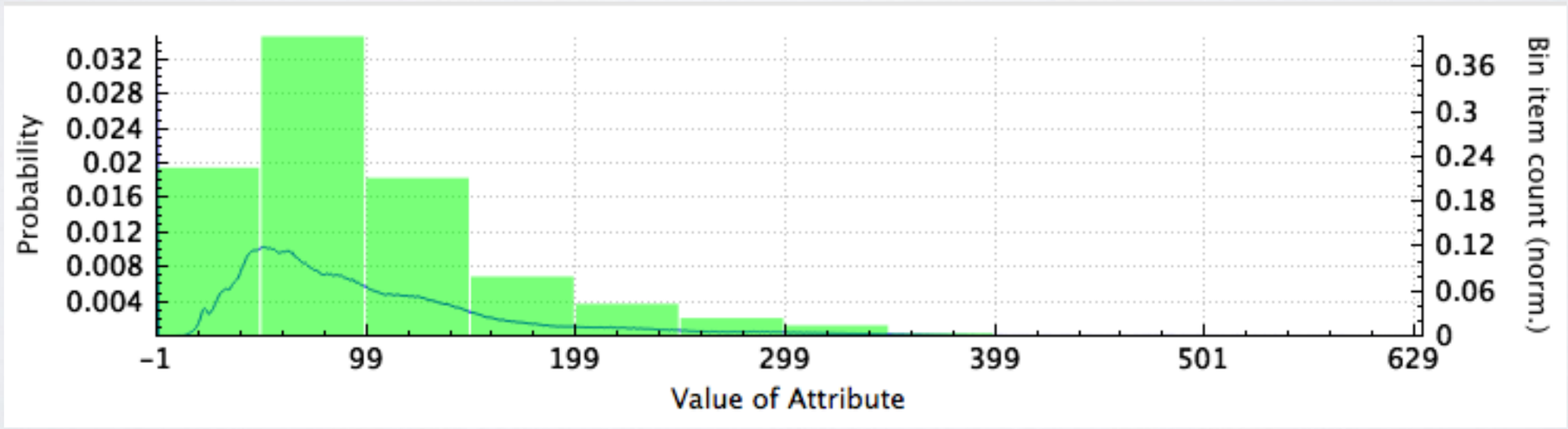
REDUCING THE VON SPACE: BINNING

- No 'continuous' value types in the flight database; however:
  - Obviously the VON space needs to be reduced otherwise there are far too many 'unique' combinations!
- How to choose sensible bin widths?
  - Should be a user-guided procedure as it relies upon specialist knowledge of the data's present and future ranges

REDUCING THE VON SPACE: VISUAL FEEDBACK

Smaller bins





Larger bins

BINNING FOR DATATYPES

- *Boolean*: uhh... let's not bother :)
- *Integer*: Divide into **N** bins? Or fix a width of **B**?
- *Datetime*: Semantic meaning; fixing a width seems logical
  - **B** seconds? **B** minutes? **B** hours? ...
- *String*: Depends on the semantics of the attribute





IMPLEMENTATION

Algorithms, Technical Details & Scalability

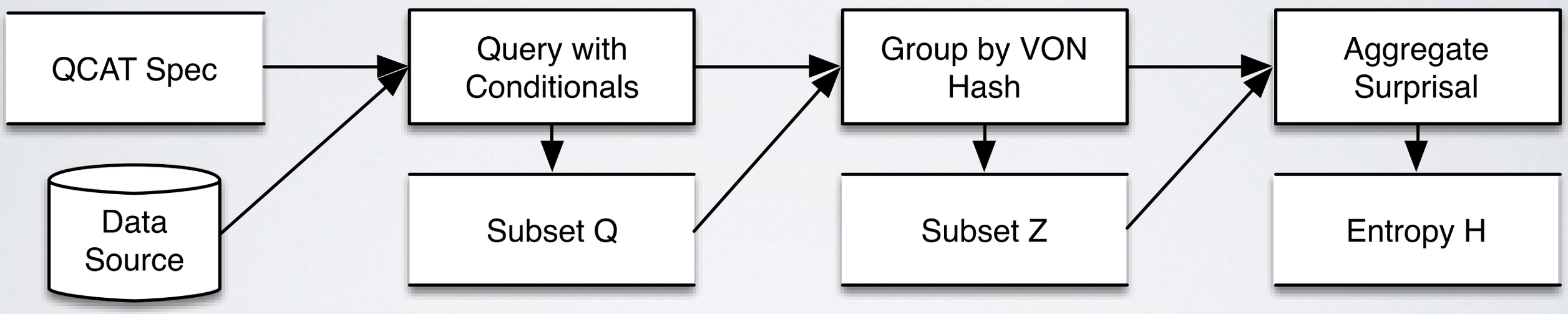
SERVER OR CLIENT?

- Where should the bulk of the computation occur?
  - *Server side*: DB Server performs all necessary aggregation
  - *Client side:* Run individual queries through DB server, gather records on client and aggregate using results

SERVER OR CLIENT?

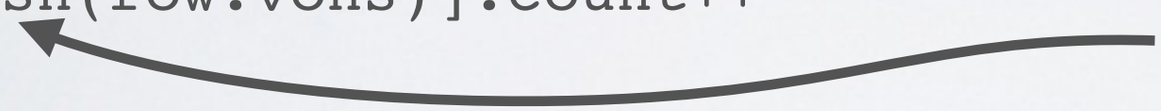
- Answer depends on:
  - Latency & bandwidth between client & server
  - Computational power / memory of server verses client
- In our tests with a localhost DB server, client-based method quicker than server-based method

GENERAL WORKFLOW



CLIENT PSEUDOCODE

```
F = qcat(conditionals={...},vons={...})
Z = hash<row_hash,int>
R = server.query(f.conditionals())

// compute Z
for (row in R)
  Z[hash(row.vons)].count++

// compute entropy
for(letter in Z)
  letter.prob = letter.count / |Z|
  letter.surp = -log2(letter.prob)
  HZ += letter.prob * -letter.surp
```



$str(v_1)+str(v_2)+...+str(v_{|V|})$

SERVER IMPLEMENTATION

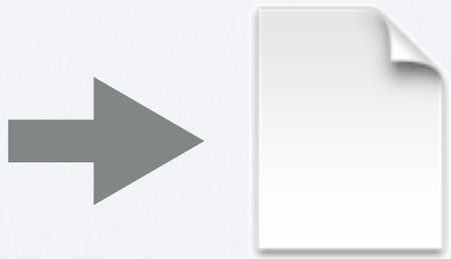
- Implemented as a Stored Procedure (SP) in PSQL
- Client calls SP through API
- API returns result when DB completes

```
-- Function: facas_server(text, text, text)

-- DROP FUNCTION facas_server(text, text, text);

CREATE OR REPLACE FUNCTION facas_server(table_name text, condition text, von_hash text)
  RETURNS record AS
$BODY$
declare
       sql text;
       total_recs integer;
       N real;
       rec record;
begin

sql = 'select count(id) FROM ' || table_name || ' WHERE ' || condition;
execute sql into total_recs;

raise info 'A) %', sql;

if total_recs = 0 THEN
       execute 'select cast(0 as bigint)  as totalRowCounts, cast(0 as bigint)  as Zcount, cast(0 as numeric) as HZ, cast(0 as numeric)
       return rec;
end if;

N = cast(total_recs as real);

sql = 'select cast(' || N || ' as bigint) as totalRowCount, count(t.cnt) as Zcount,
       cast(-SUM(_log2(cast(t.cnt/cast(' || N ||' as real) as real)) * (cast(t.cnt as real)/ ' || N || ')) as numeric) AS HZ,
       cast(SUM(-_log2(cast(t.cnt/cast(' || N ||' as real) as real))) as numeric) AS sum_surprise
from (
select count(id) as cnt
 from ' || table_name ||
 ' where  ' || condition ||
 ' group by ' || von_hash ||
') t';

raise info 'B) %', sql;

 execute sql into rec;
 return rec;

end
$BODY$
  LANGUAGE plpgsql VOLATILE
  COST 100;
ALTER FUNCTION facas_server(text, text, text)
  OWNER TO postgres;
```

SERVER PERFORMANCE

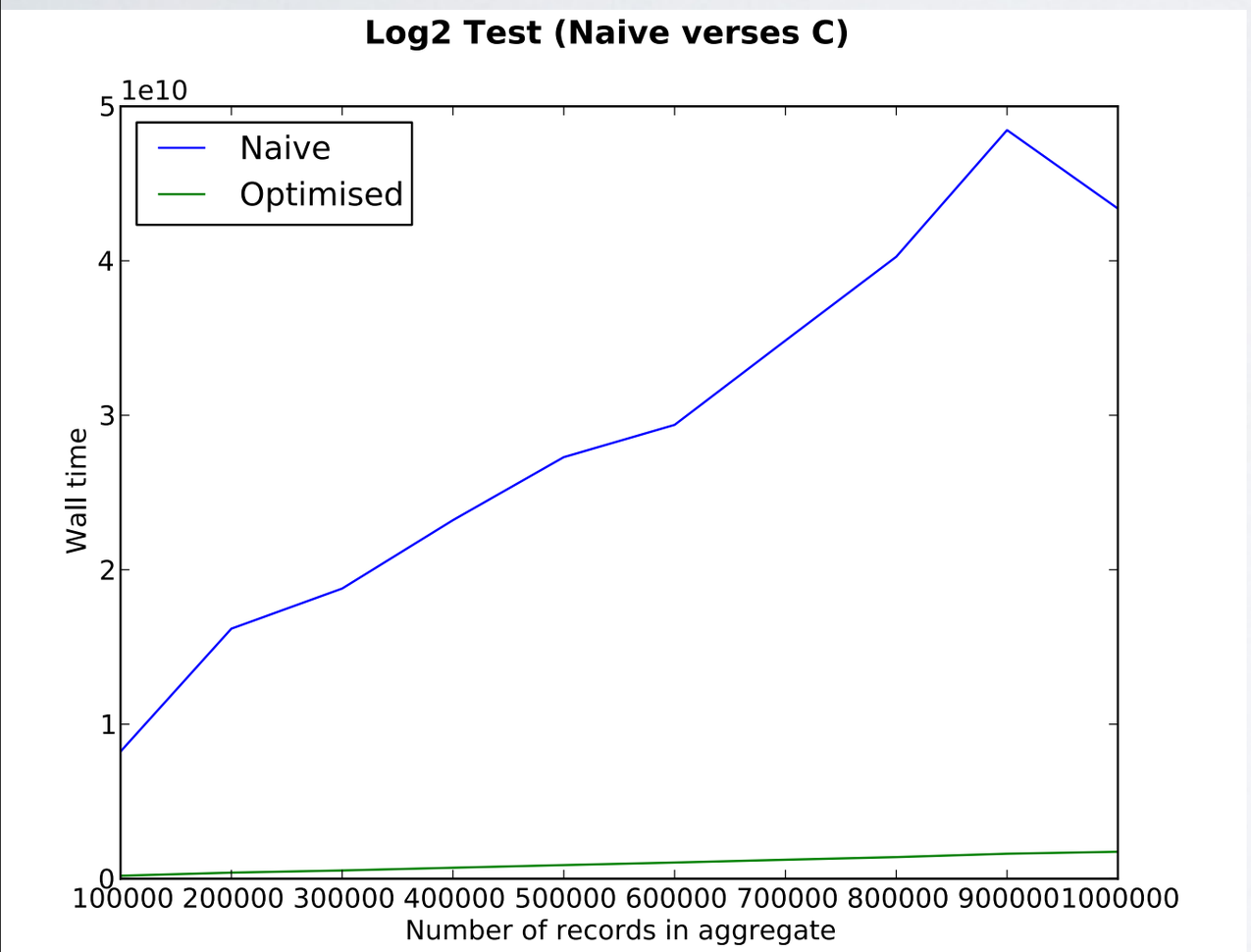
- Initial shock: server-side implementation ~5x *slower* than client-side implementation
- Culprit: built-in log() function too slow
- Luckily: can plug-in our own C log2 implementation from an object file and get Postgres to call it

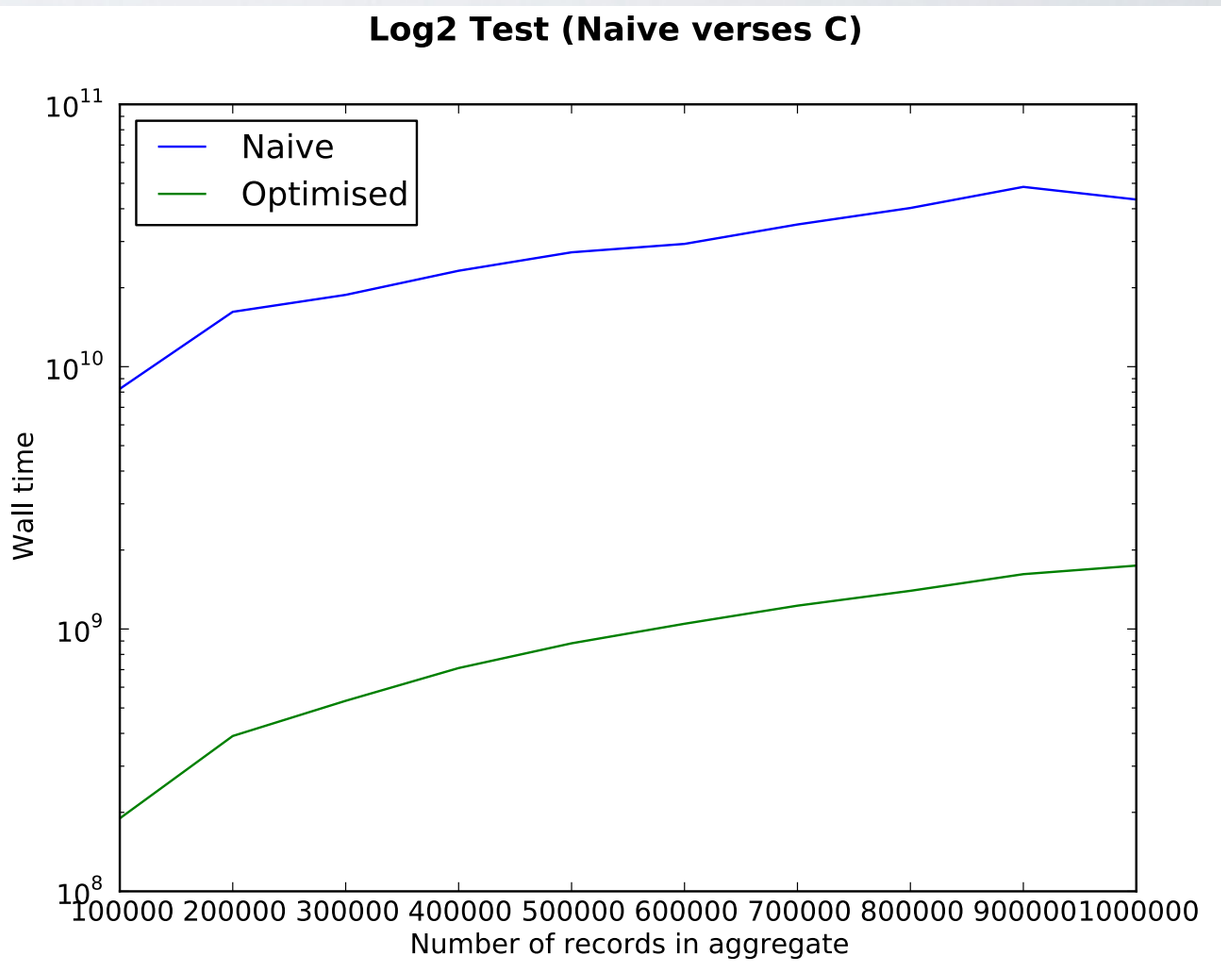
```c
#include <math.h>

#include <postgres.h>
#include <fmgr.h>

#ifdef PG_MODULE_MAGIC
PG_MODULE_MAGIC;
#endif

PG_FUNCTION_INFO_V1(_log2);
Datum _log2(PG_FUNCTION_ARGS) {
    PG_RETURN_FLOAT4(log2(PG_GETARG_FLOAT4(0)));
}
```





LOG2 COMPARISON





SERVER PERFORMANCE

- Other performance consideration: choice of hash function for computing unique sets of variants of normality

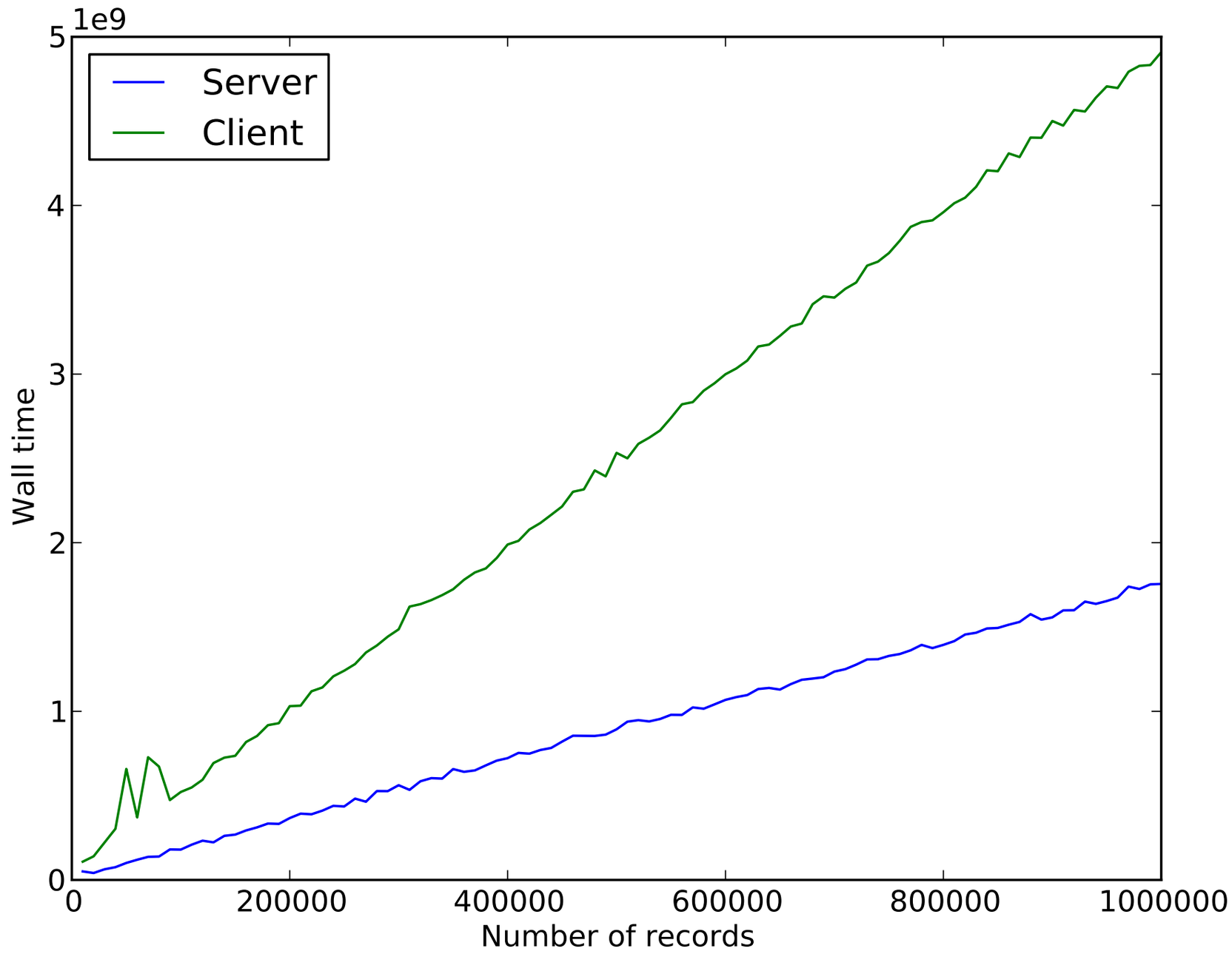
- Hashing quick in C++ using std::map & std::string concat

- Not so quick in Postgres!

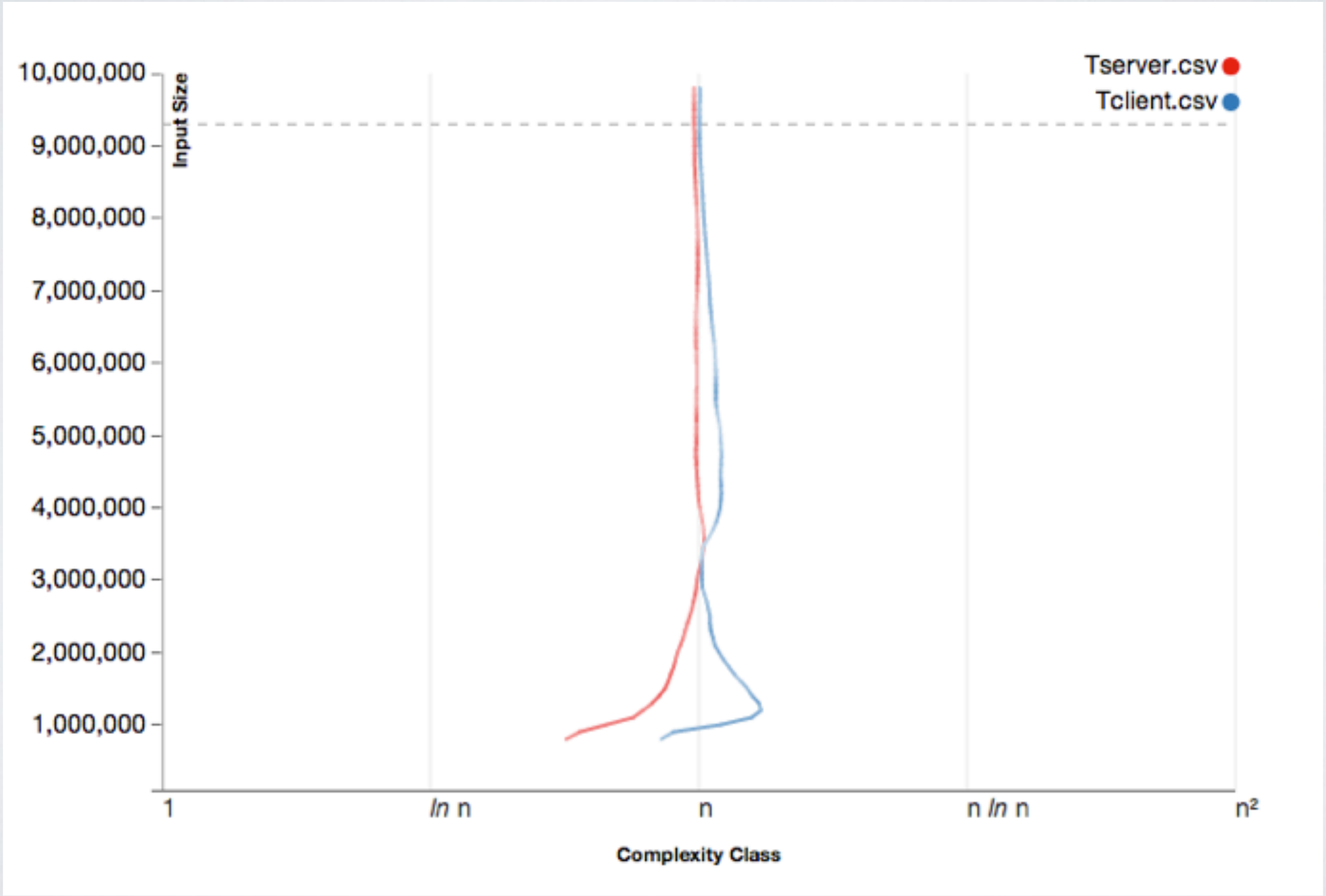
Origin: LAX
Dest: JFK

Carrier, ElapsedTime, AirTime, TailNum, TaxiOut, Cancelled, CancellationCode, Diverted

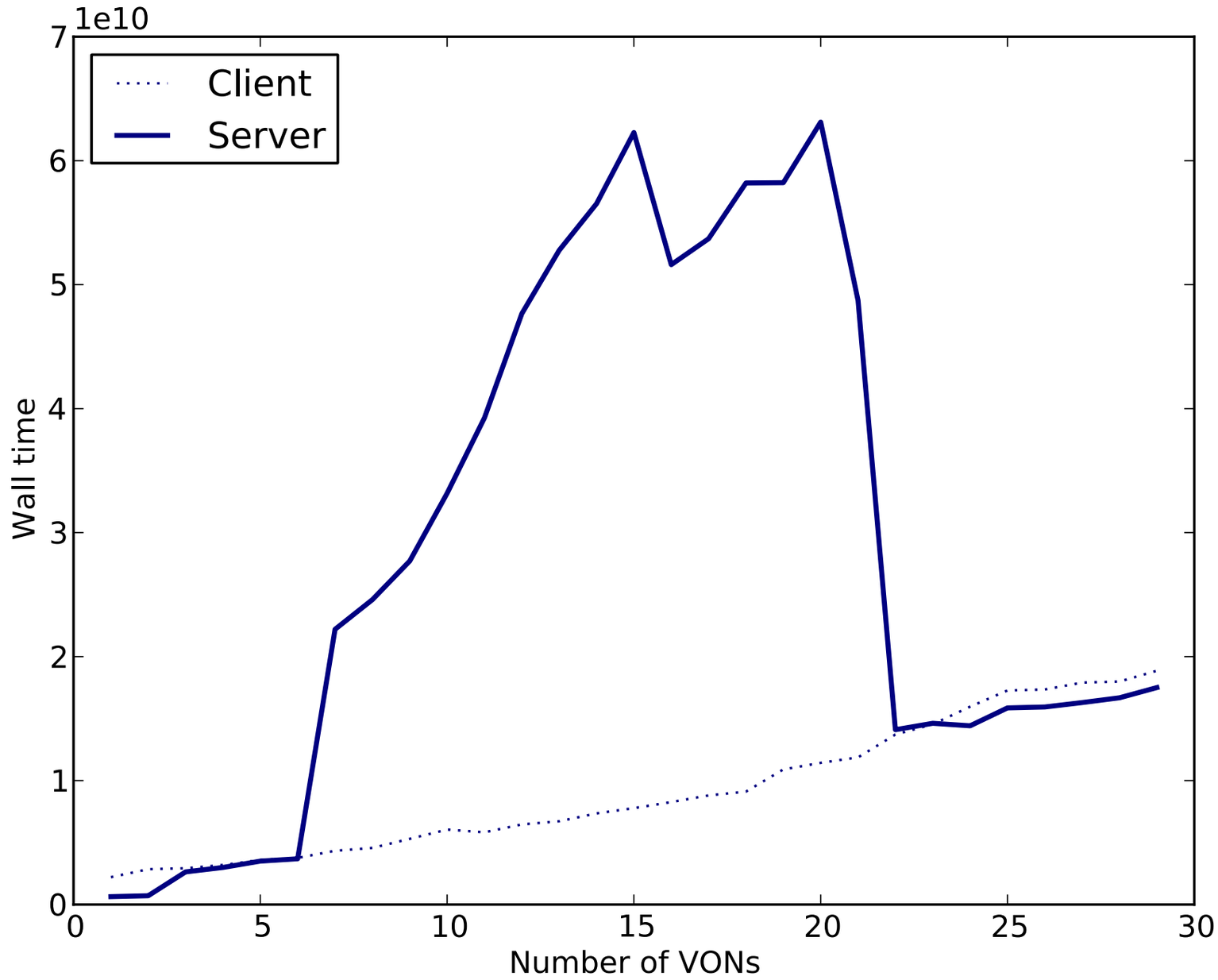
CLIENT VS. SERVER



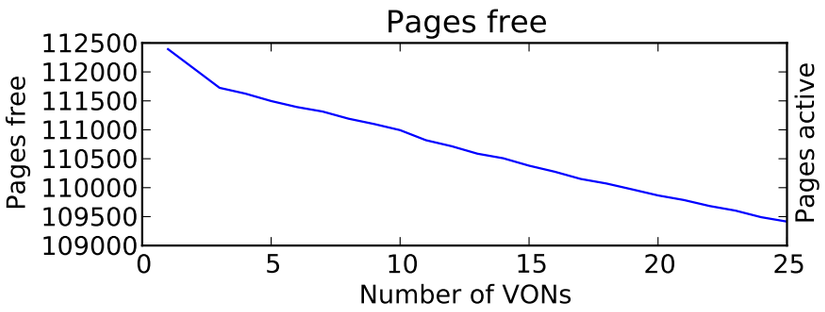
CLIENT VS. SERVER

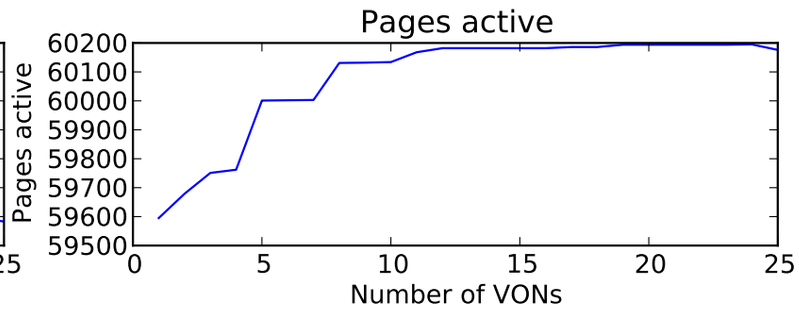


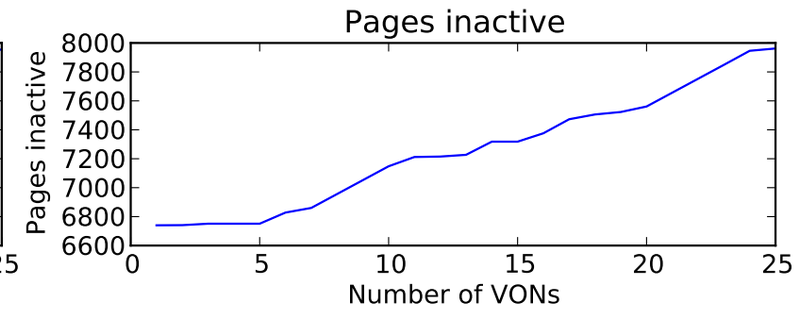
VON STRESS TEST

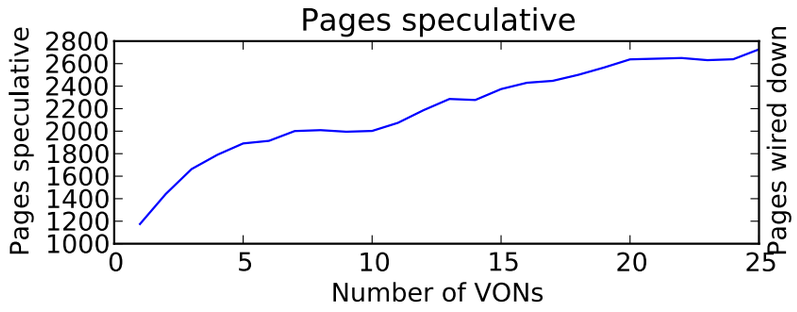


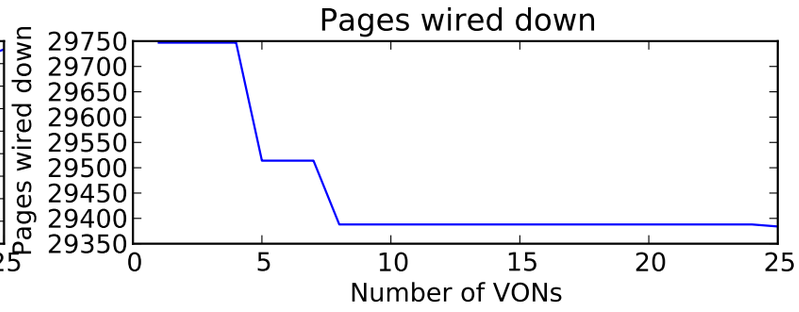
Memory Usage During FaCAS Postgres VON Test

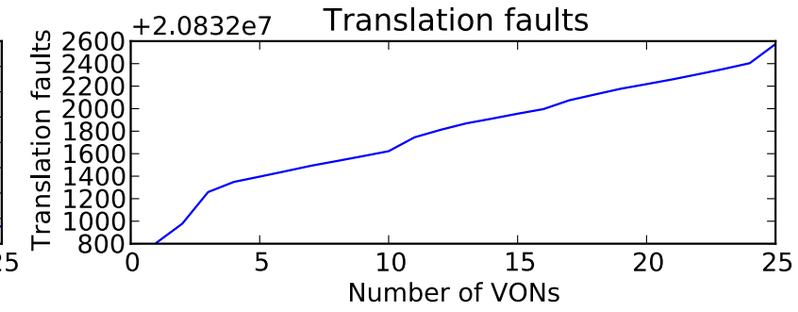


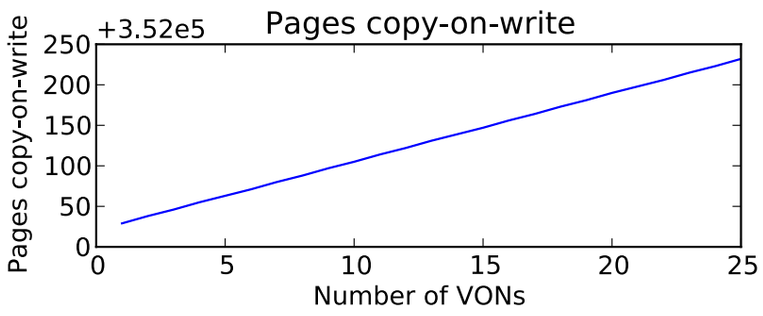


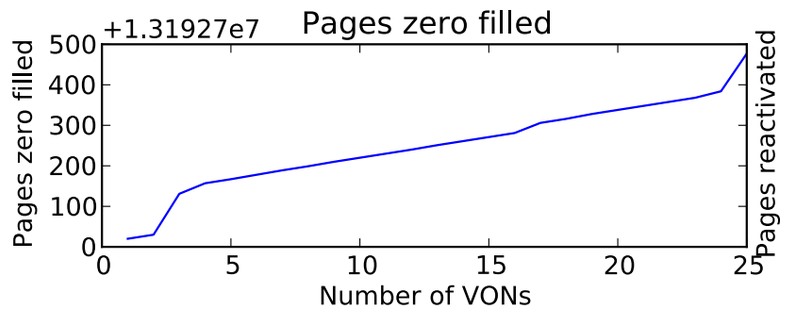


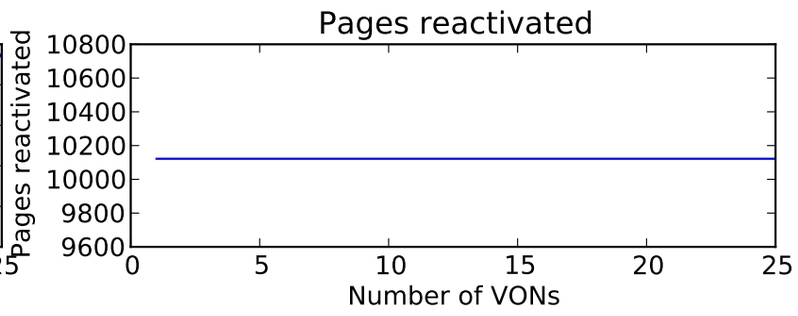


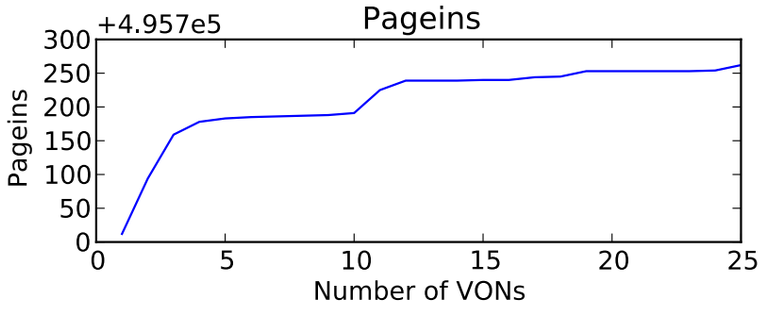


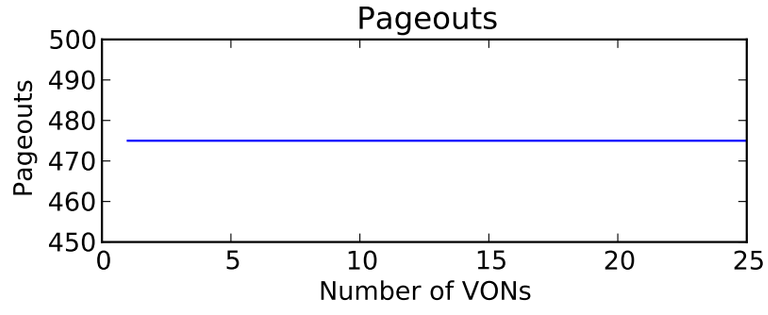




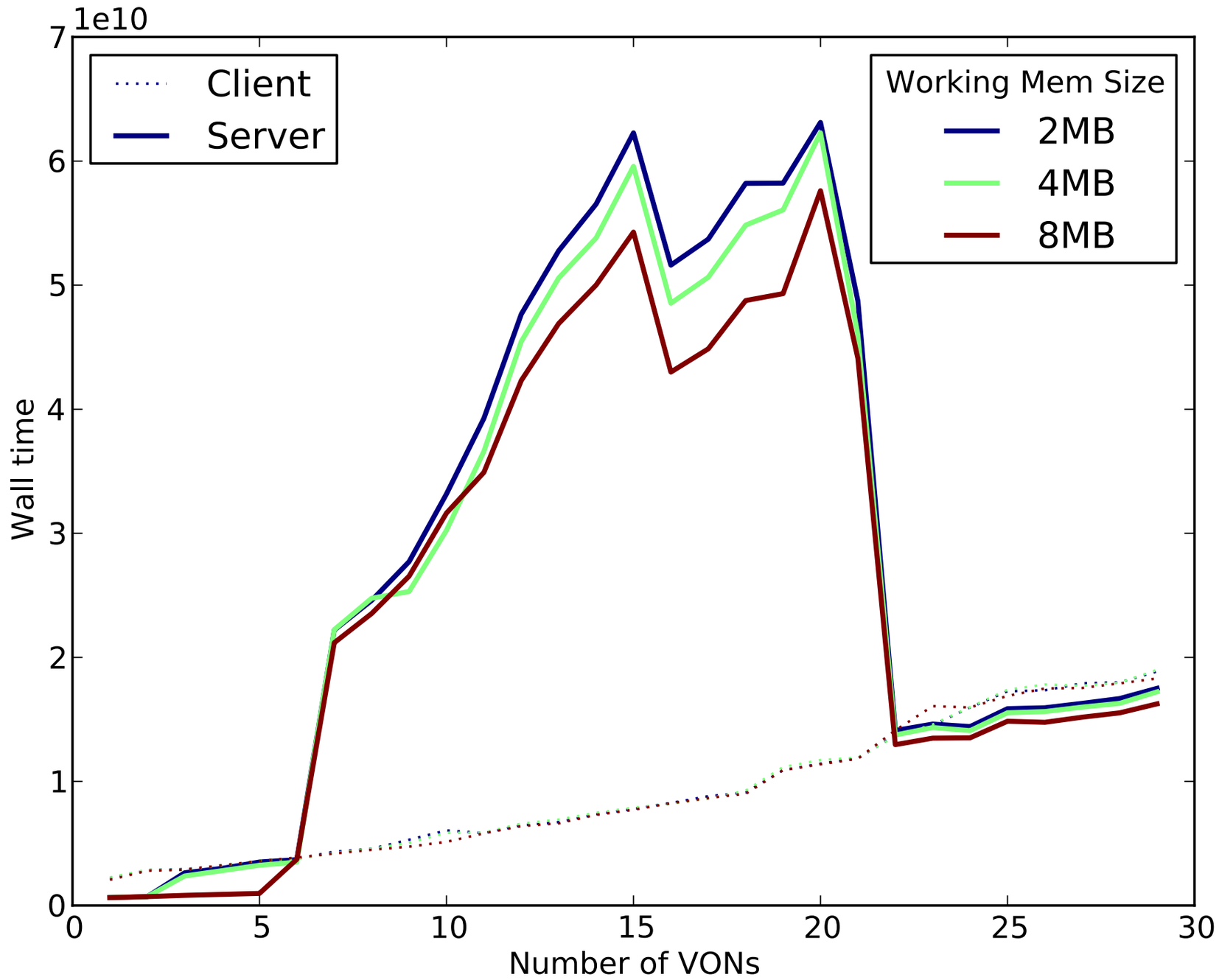




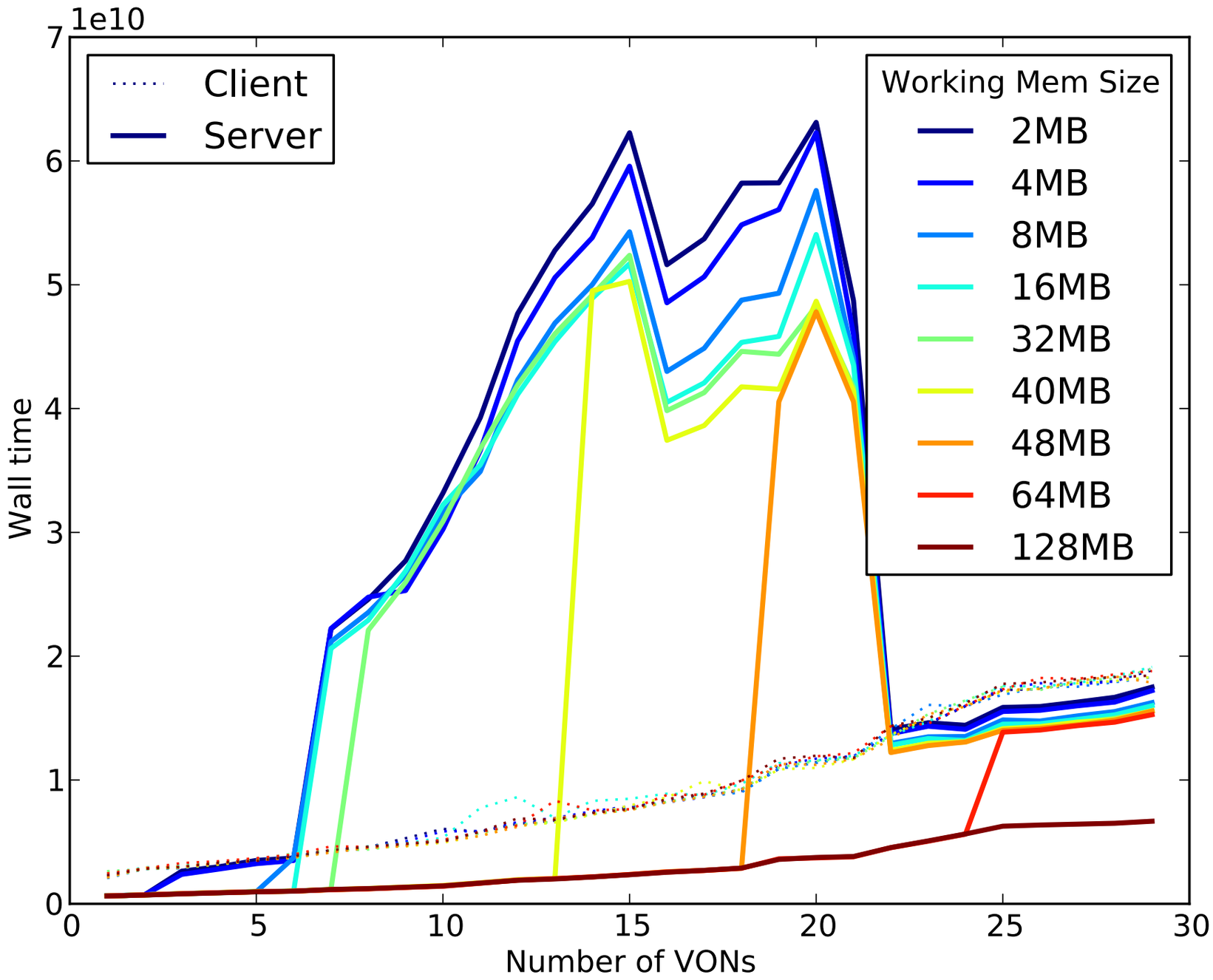




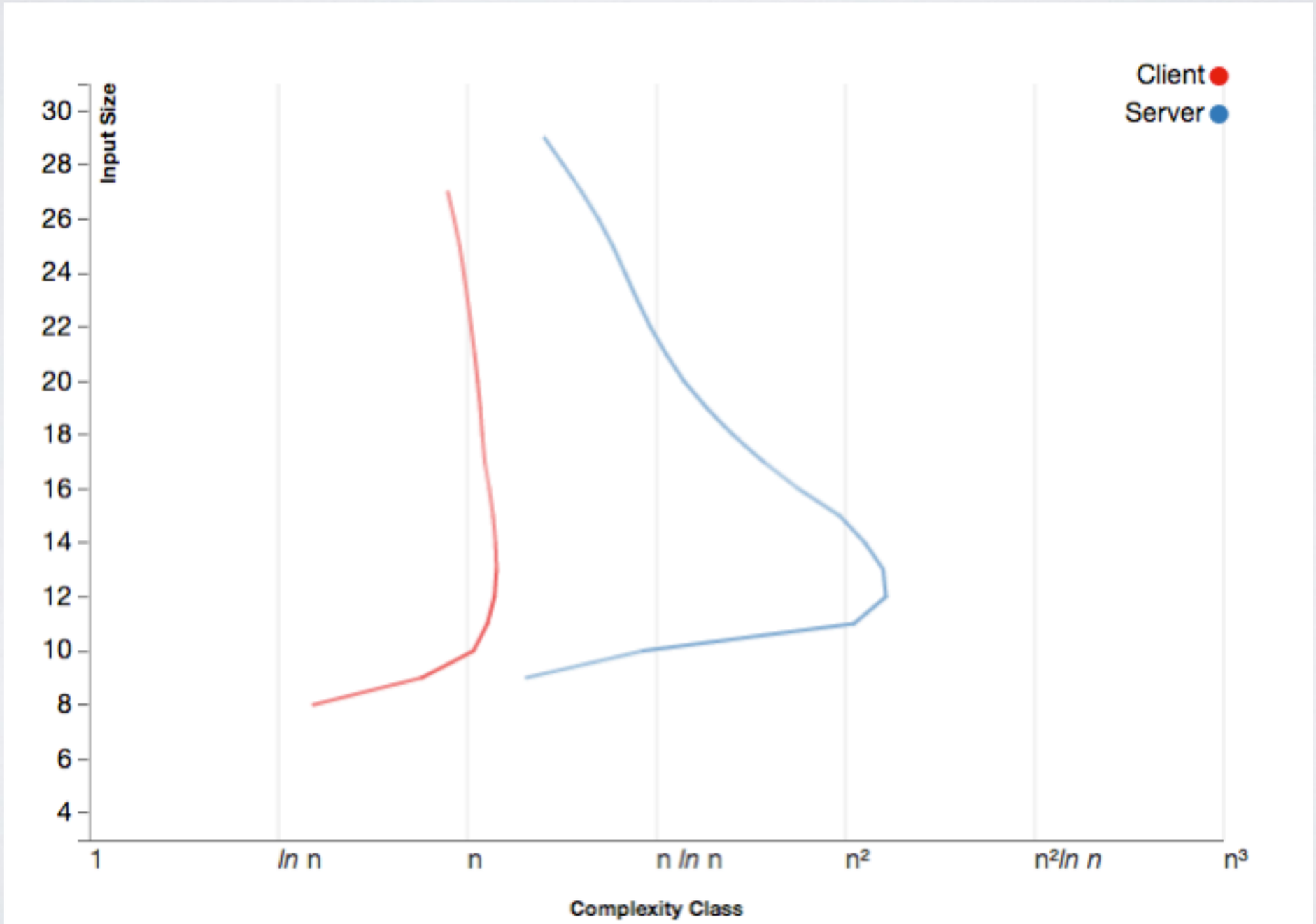
VON STRESS TEST



VON STRESS TEST



VON SCALABILITY: LINEAR



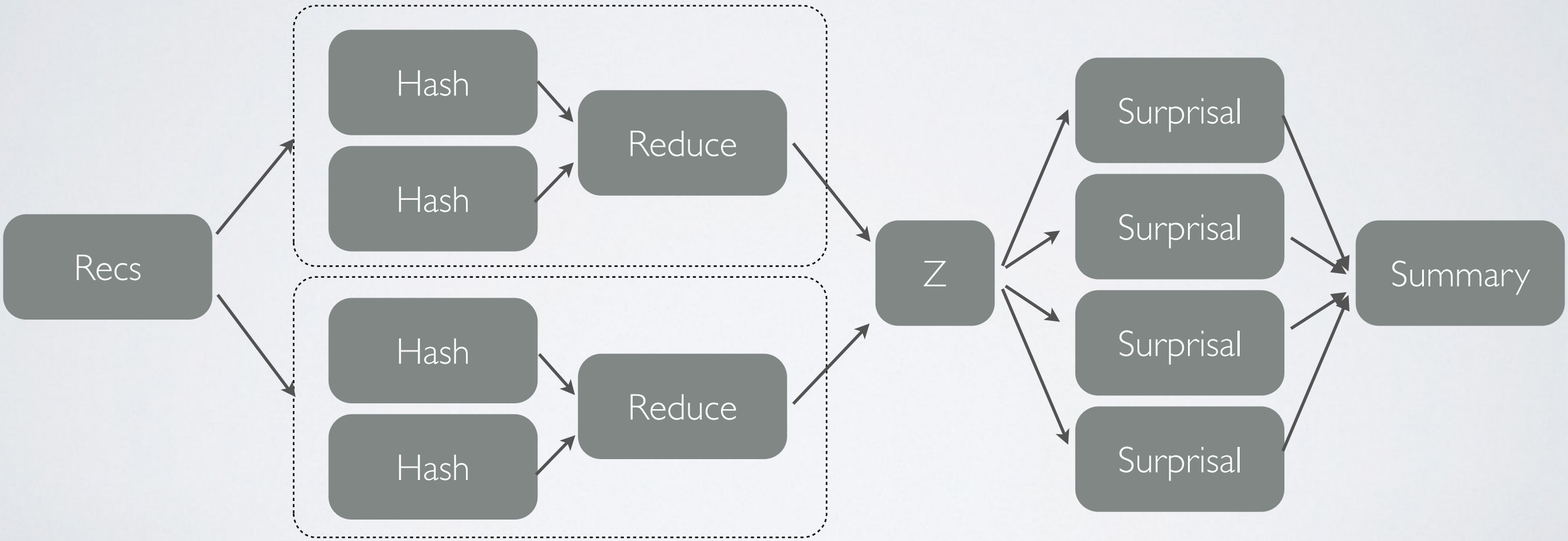
INDEXING

- Candidates for indexes:
  - All fields used by any QCAT condition (inc. multicolumn combinations)
  - Bin formulae (i.e. `to_int(integer/20)`)
- All of these currently created automatically by system

PERFORMANCE, OVERALL

- 1,000,000 records surprisal computation: ~1.2 sec
- Performance largely dependent upon:
  - Initial conditional matching (quicker if indexed)
  - Number of VONS to hash together

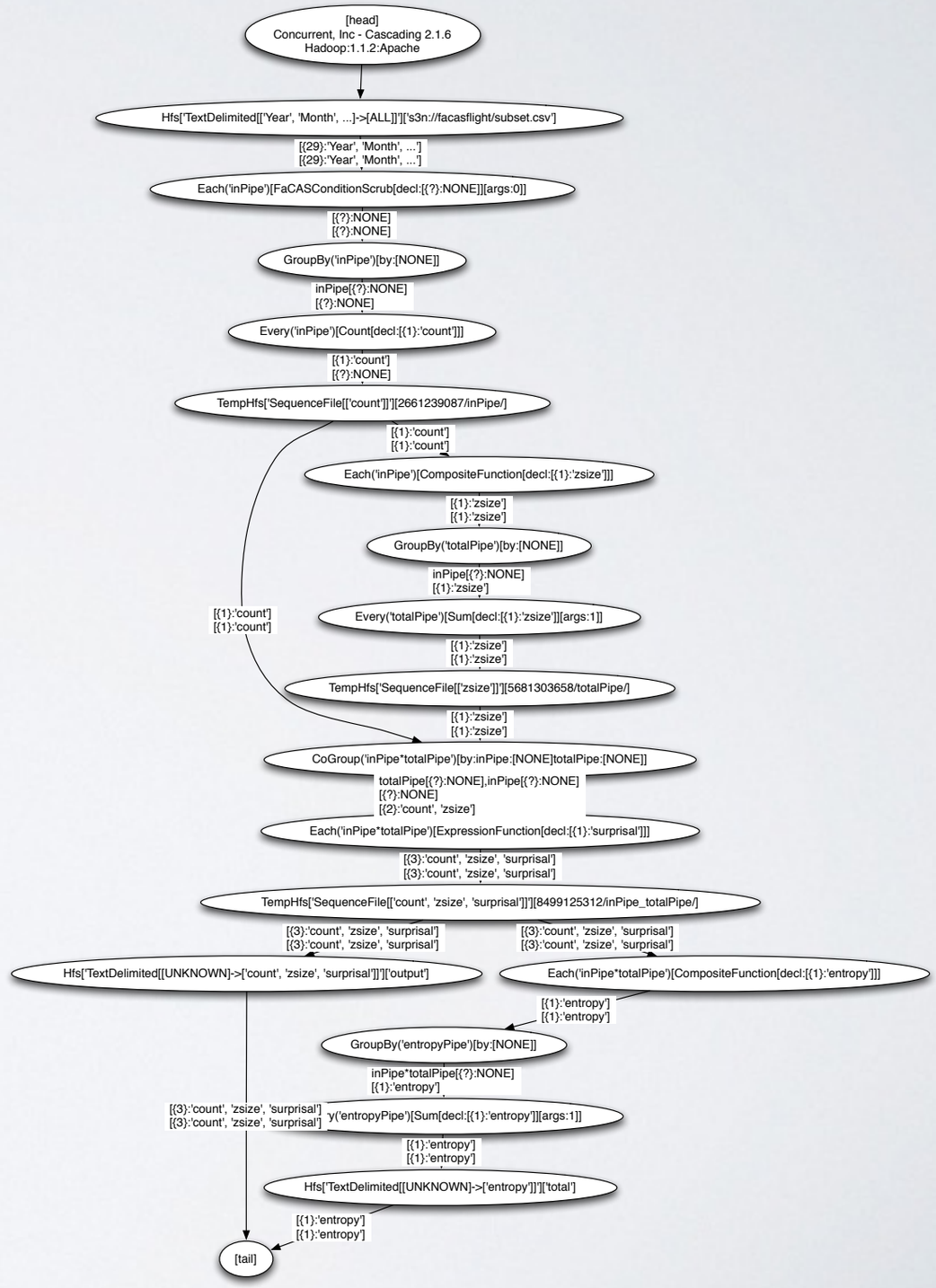
MAP/REDUCE

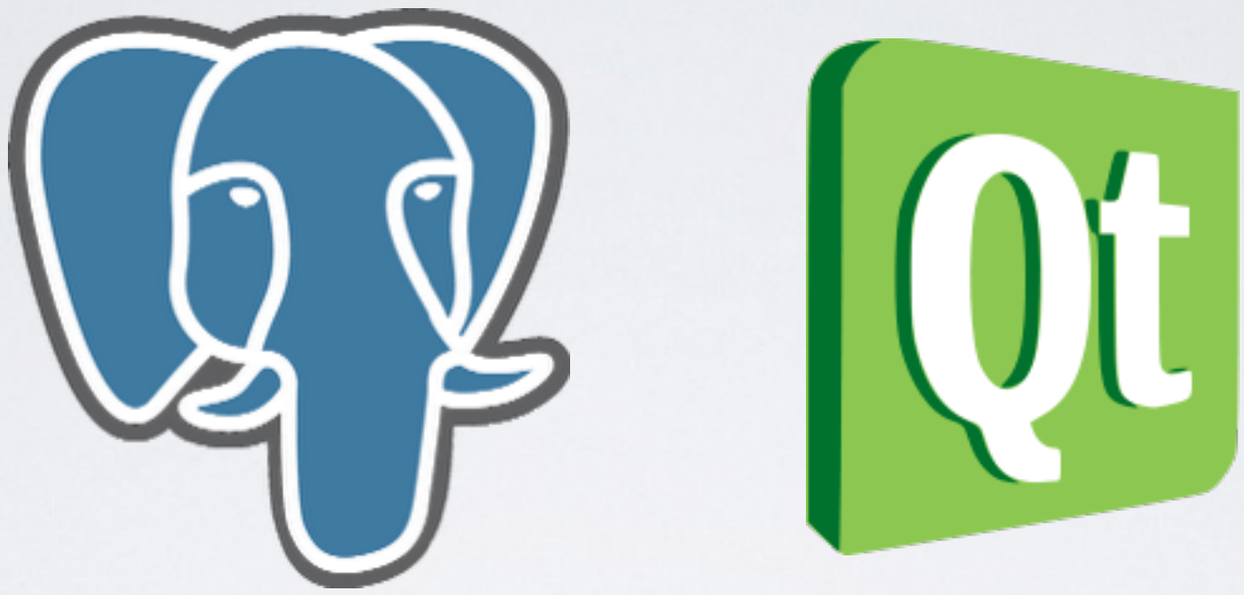

Recs
Hash
Hash
Reduce
Hash
Hash
Reduce
Z
Surprisal
Surprisal
Surprisal
Surprisal
Summary

HADOOP VERSION

- Java/Cascading* implementation
- Deployed as a .JAR
- Very slow
- But obviously scalable

* http://www.cascading.org/



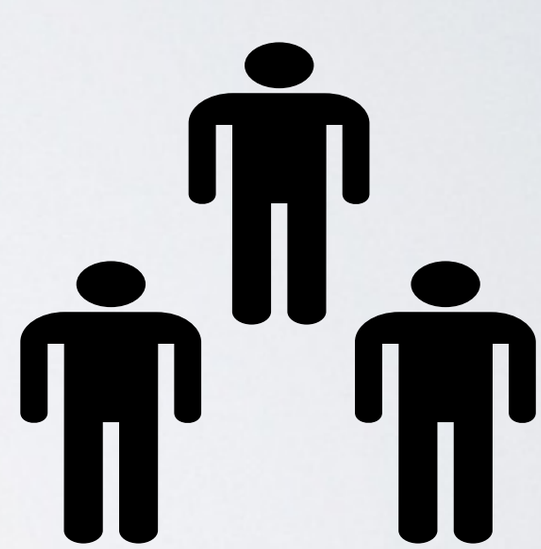


USER-INTERACTION

Graphical User Interfaces & Visualisations

MOTIVATION

- In our system, we imagine three user roles:
  - *Observer*: observes incoming data
  - *Analyst*: analyses new and existing data
  - *Designer*: designs appropriate QCATs



SYSTEM FLOW



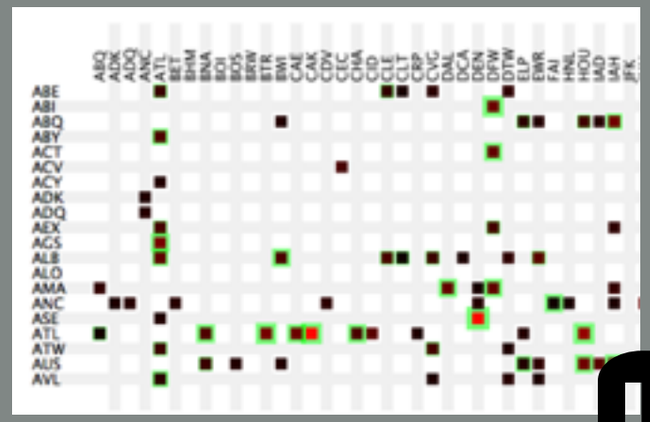
Designer Software

QCAT Store

F_1 F_2 F_n

Analytical Engine

Vis



QCAT Summaries

Origin: LAX
Dest: JFK



QCAT DESIGNER

- Should provide the user with useful statistics for each field in order to assist user decision in:
  - Denoting fields as conditionals or VONs
  - Choosing the bin widths for the conditionals and VONs
- Should allow test runs of QCAT
- Save/load functionality

FaCAS Designer

Attribute List:

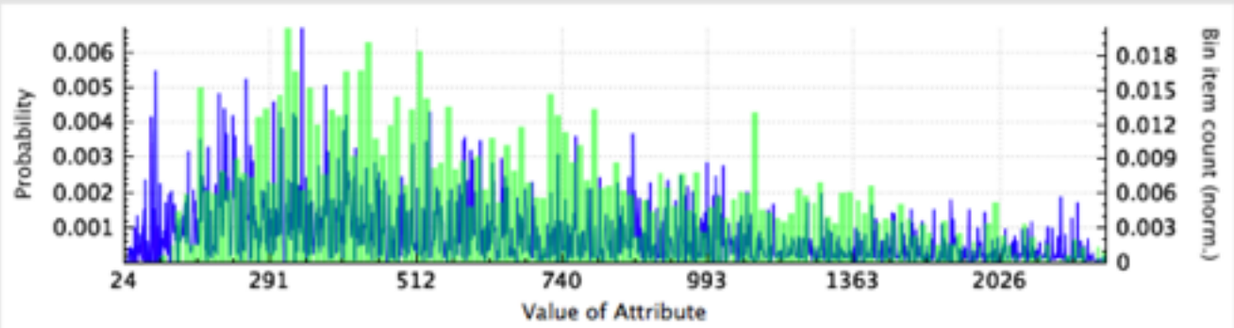
Attribute	Bin Spec	Role
id	10 integers	None
year	10 integers	None
month	10 integers	None
day_of_month	10 integers	None
day_of_week	10 integers	None
dep	10 minutes	None
dep_crs	10 minutes	None
arr	10 minutes	None
arr_crs	10 minutes	None
carrier	1 string	None
flight_num	10 integers	None
tail_num	1 string	None
elapsed_tim...	10 integers	None
elapsed_tim...	10 integers	None
air_time	10 integers	None
delay_arr	10 integers	None
delay_dep	10 integers	None
origin	1 string	None
dest	1 string	None
distance	10 integers	None
taxi_in	10 integers	None
taxi_out	10 integers	None
cancelled	1 boolean	None
cancellation...	1 string	None
diverted	1 boolean	None
delay_carrier	10 integers	None
delay_weat...	10 integers	None
delay_nas	10 integers	None
delay_security	10 integers	None
delay_late_...	10 integers	None

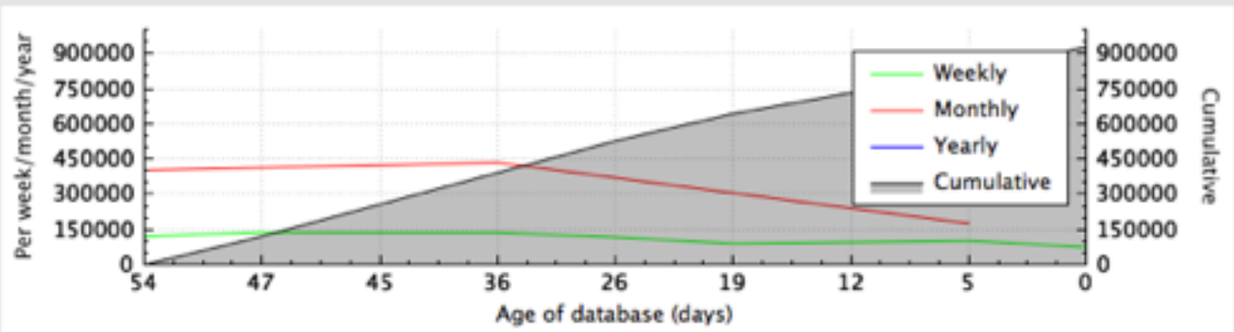
Conditionals:

Attribute	Op	Value

Stats for field: distance

Min: 24.00 Max: 4962.00 Unique: 1347 Avg: 689.36 StdDev: 539.06





Bin Specification

Width: 10 integers = # Bins: 134 Maintain constant number of bins with incoming records

FaCAS Statistics

Message: No VONs yet; FaCAS not run.

Entropy: 0 Avg Surprise: 0 Alphabet size: 0 Total matches: 0

Wall time: 0 Uncertainty: 0

FaCAS Designer

Attribute List:

Attribute	Bin Spec	Role
id	10 integers	None
year	10 integers	None
month	10 integers	None
day_of_month	10 integers	None
day_of_week	10 integers	None
dep	10 minutes	None
dep_crs	10 minutes	None
arr	10 minutes	None
arr_crs	10 minutes	None
carrier	1 string	None
flight_num	10 integers	None
tail_num	1 string	None
elapsed_tim...	10 integers	None
elapsed_tim...	10 integers	None
air_time	10 integers	None
delay_arr	10 integers	None
delay_dep	10 integers	None
origin	1 string	Conditional
dest	1 string	Conditional
distance	10 integers	VON
taxi_in	10 integers	VON
taxi_out	10 integers	None
cancelled	1 boolean	None
cancellation...	1 string	VON
diverted	1 boolean	None
delay_carrier	10 integers	None
delay_weat...	10 integers	None
delay_nas	10 integers	None
delay_security	10 integers	None
delay_late_...	10 integers	None

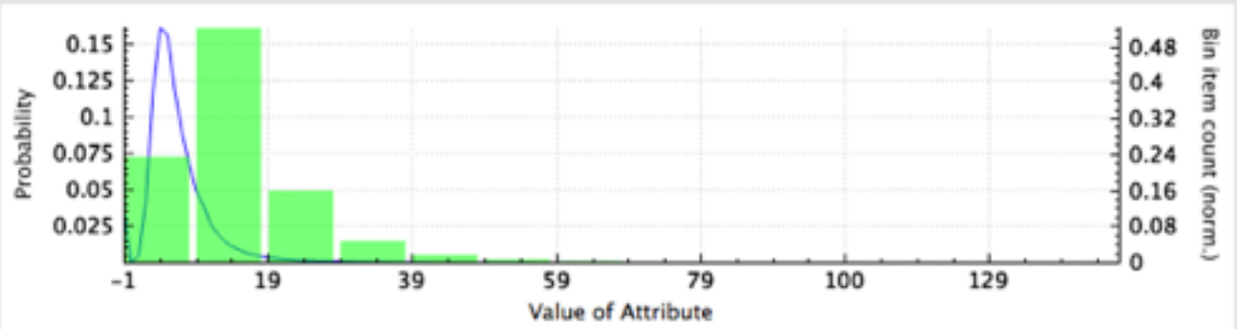
Conditionals:

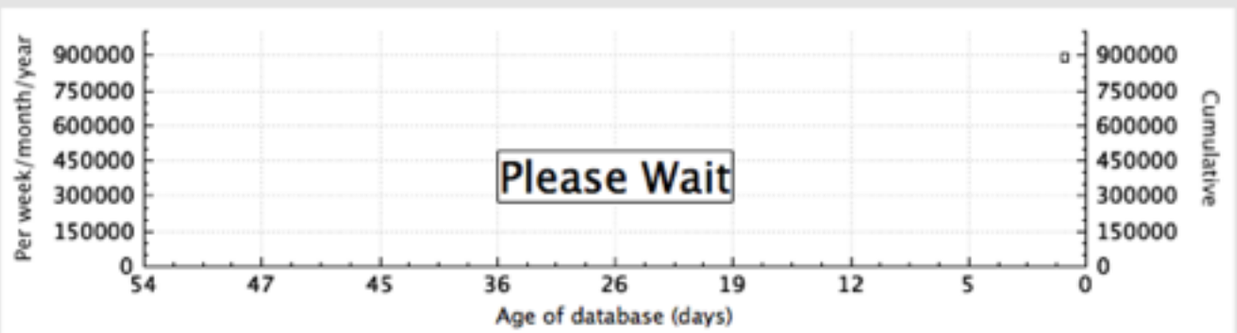
Attribute	Op	Value
dest	=	
origin	=	

Field statistics (no field selected)

Min: - Max: - Unique: - Avg: - StdDev: -

-





Bin Specification

Width: 1 string = # Bins: 286 Maintain constant number of bins with incoming records

FaCAS Statistics

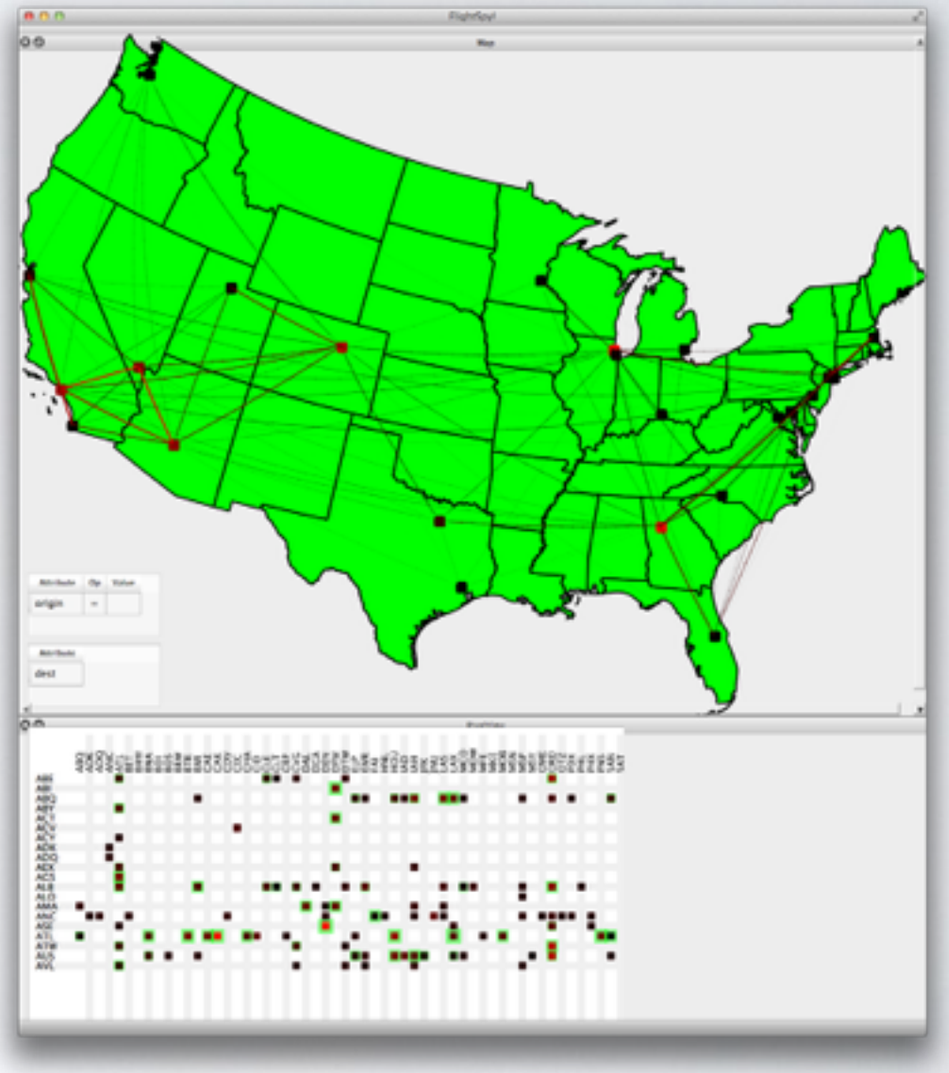
Message: A conditional is not complete; FaCAS not run.

Entropy: 0 Avg Surprise: 0 Alphabet size: 0 Total matches: 0

Wall time: 0 Uncertainty: 0

VISUALISATIONS

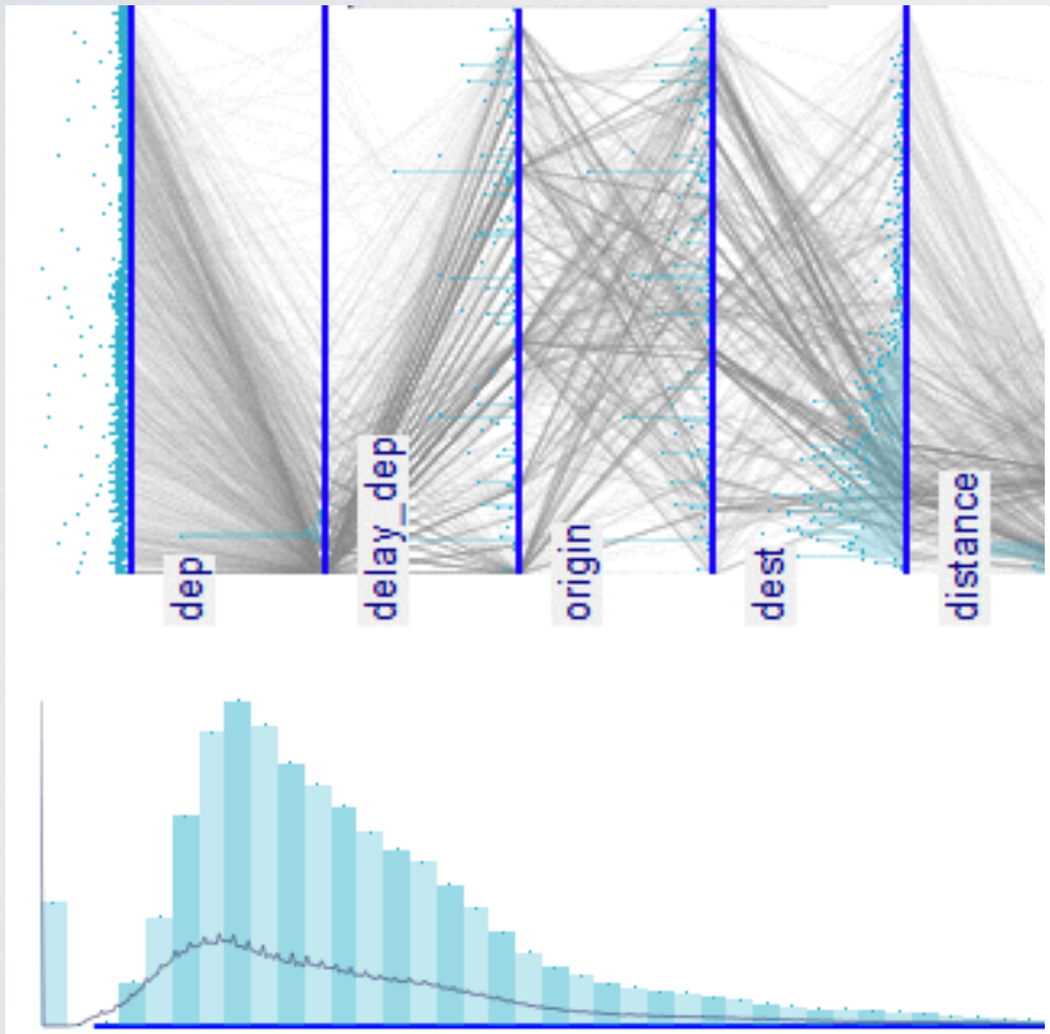
- During API testing stage: two initial visualisations using flight data and QCAT surprisal highlighting strategy
  - Pixel grid of origin verses destination airport
  - US map view

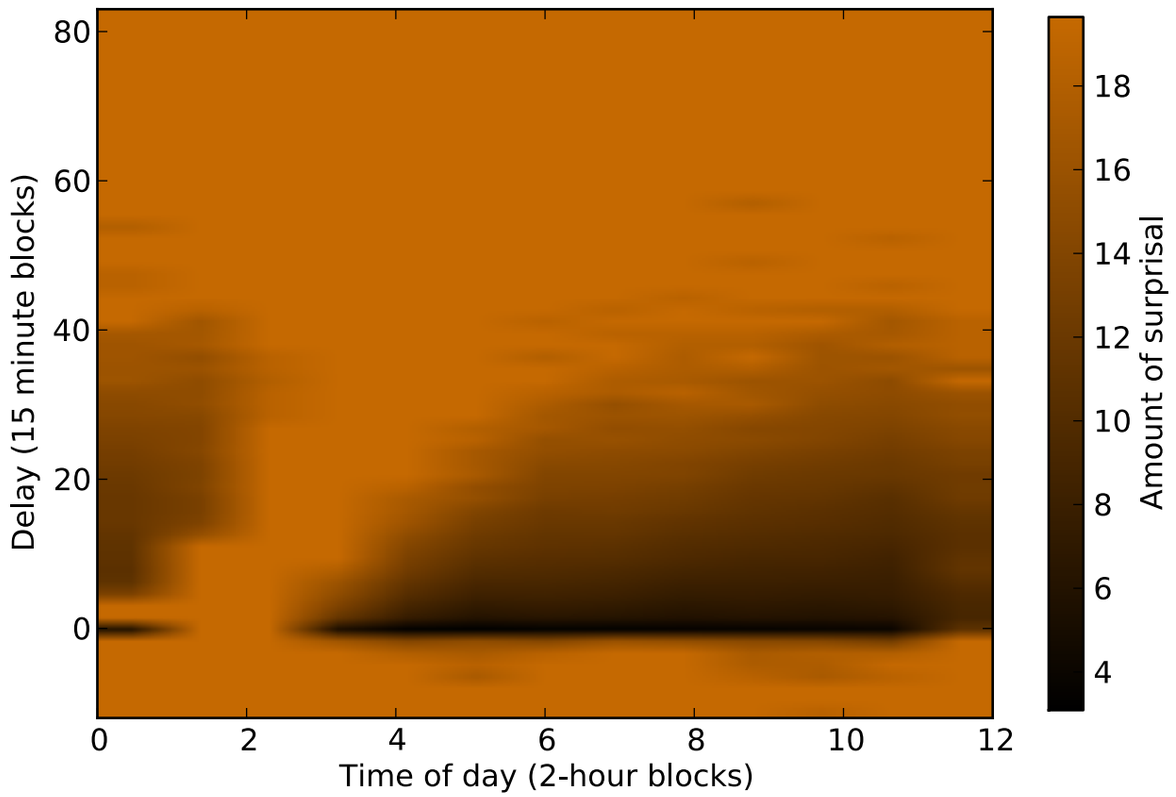


PURDUE COLLABORATION



PURDUE COLLABORATION

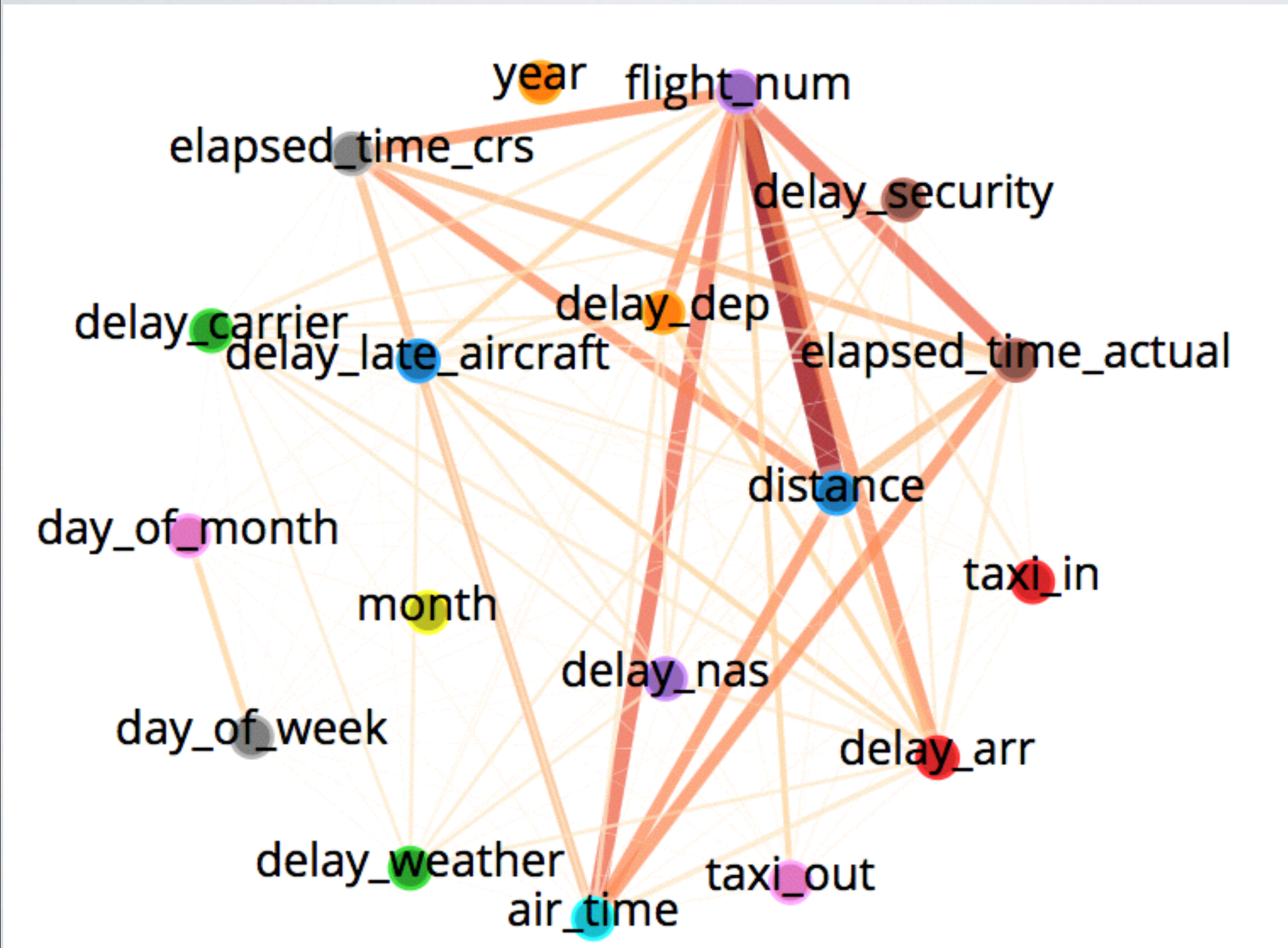




CONCLUSION

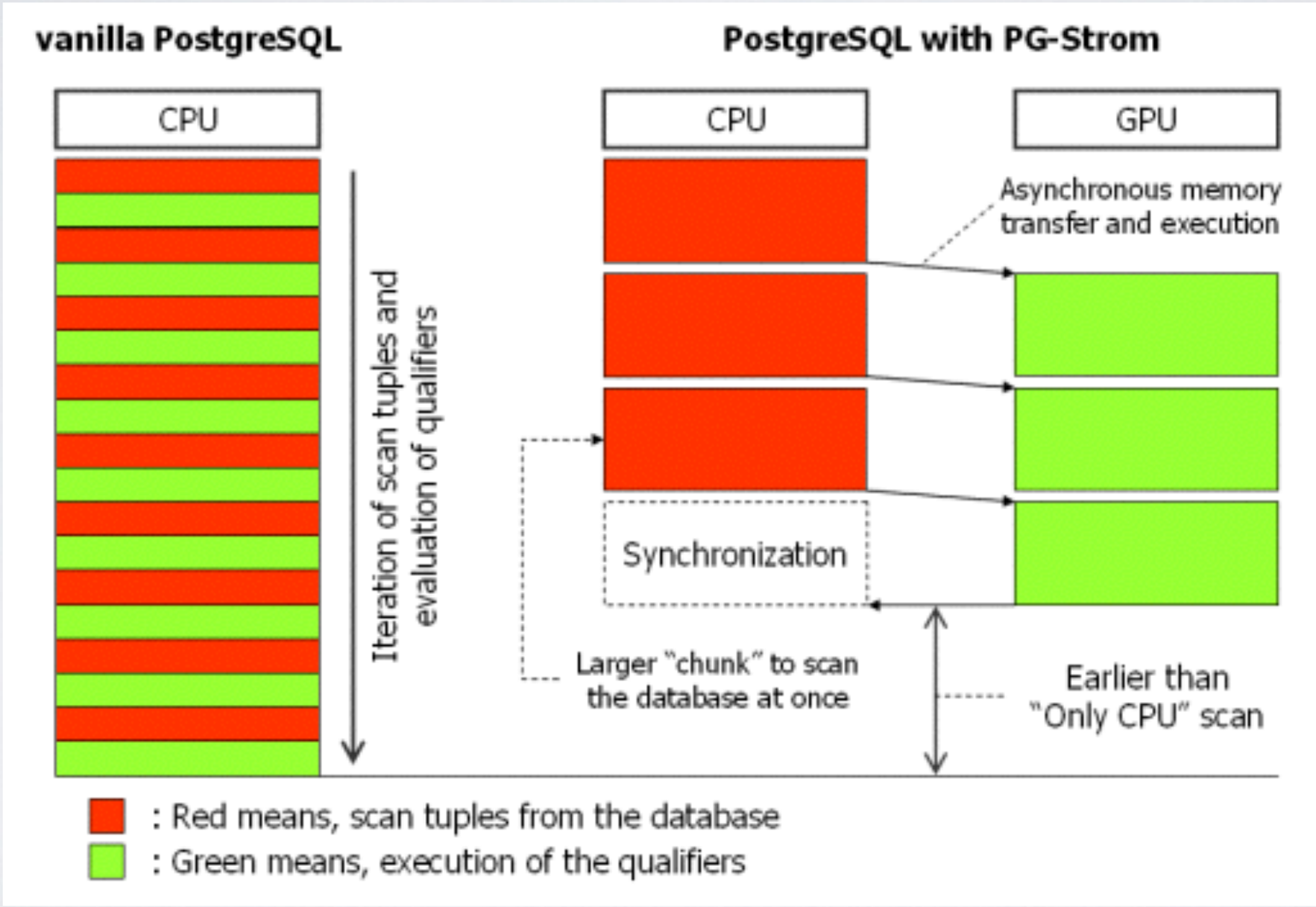
- Development and Implementation of a new Information-Theoretic approach to anomaly detection
  - Implemented in a documented C++ API
  - Scalability tested
- Collaboration with Purdue resulting in Eurovis paper submission

FUTURE WORK


year
flight_num
elapsed_time_crs
delay_security
delay_dep
delay_carrier
delay_late_aircraft
elapsed_time_actual
distance
day_of_month
taxi_in
month
delay_nas
day_of_week
delay_arr
delay_weather
taxi_out
air_time


distance
elapsed_time_actual

FUTURE WORK



http://wiki.postgresql.org/wiki/PGStrom

QUESTIONS?

UKVAC II JRP

WP1: Information-Theoretic Framework

Professor Min Chen
Dr. Simon Walton
Oxford e-Research Centre
University of Oxford















UKVAC II Meeting, 11 January 2013

Objectives of WP1

a) To formulate an **information-theoretic framework** for visual analytics processes in a context exemplified by these two problems.

b) To develop **visual representations** that depict information contained in a dataset as well as the uncertainty quantities and other probabilistically-derived qualities such as importance and value.

c) To develop a **prototype** demonstration system for illustrating the potential use of information theoretic quantities in the process of hypothesis generation and evaluation.

d) To compare the information-theoretic framework with captured user study data (Objective 4) and to offer an explanation about any correlation and inconsistency.

e) To explore the application of the information-theoretic framework in interaction exploration (Objective 2), data management and data analysis (Objective 3), hierarchical decision models (Objective 5).

The Flight Data (FD) Problem

- *The flight data problem is about wanting answers to questions that could be contained in or constructed from the interrogation of 120 million records covering all commercial flights in the US from October 1987 to April 2008, about 12Gb of uncompressed data. Questions that could be asked include; when is the best time to fly to minimise delays? Can you detect cascading failures as delays in one airport create delays in others? Are there critical links in the system?*

The Nobel Laureates (NL) Problem

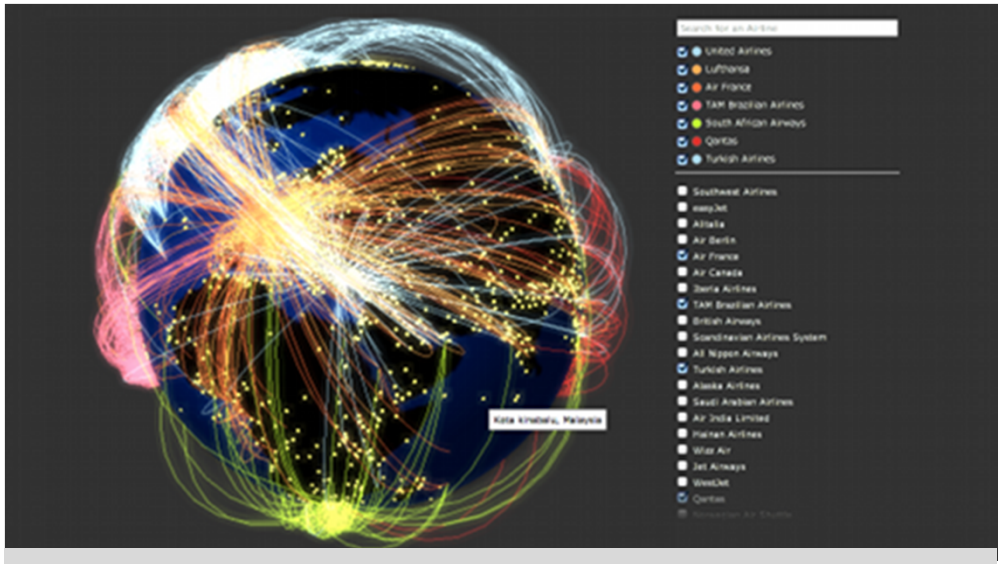
- *Who will be the short-listed candidates for the Nobel Laureates in 2012?*

State of the Art: Visualization

- Spatial visualization
  - Real-timing tracking
- Spatiotemporal visualization
  - Flight Route Visualization
- Information and geo-information visualization
  - Pixel-based visualization
  - Radial graph visualization
  - Tree-map
  - Simple glyph visualization



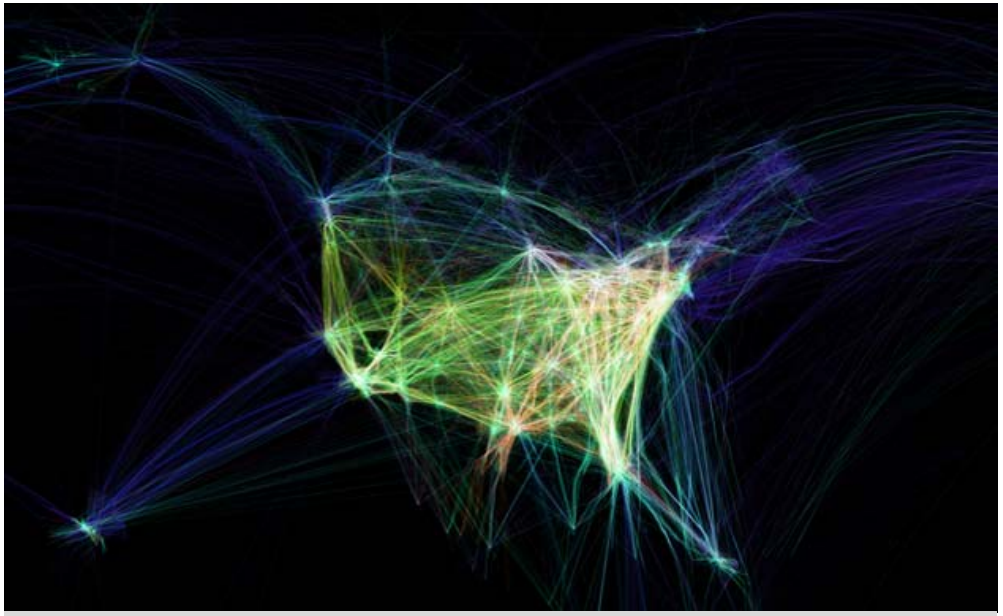
http://data-informed.com/flightradar24-real-time-visualization-of-air-traffic-uses-global-sensors-network/
Stephanie Overby



http://philogb.github.com/world-flights/
about 200 airlines, 57,000 flights and 6000 destinations



http://www.barnabu.co.uk/flight-route-maps-for-australia-and-new-zealand/



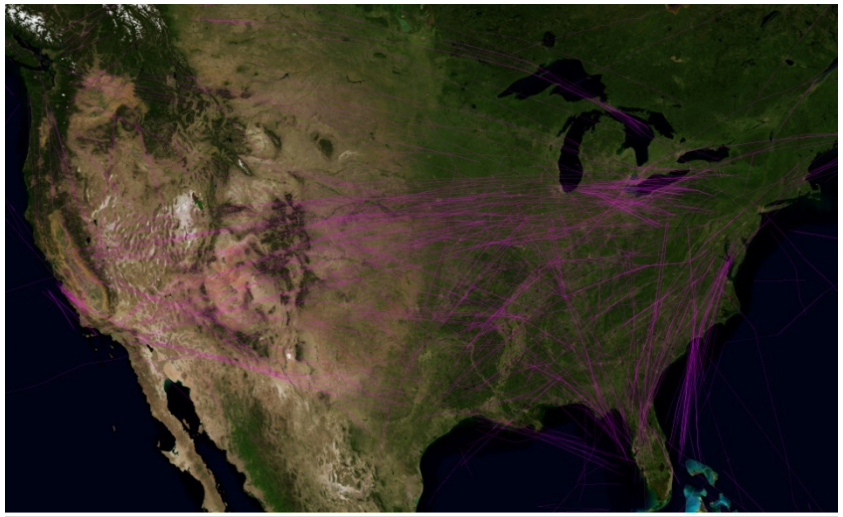
http://www.aaronkoblin.com/work/flightpatterns/
Aaron Koblin



http://www.navigadget.com/
Visualizing flight paths



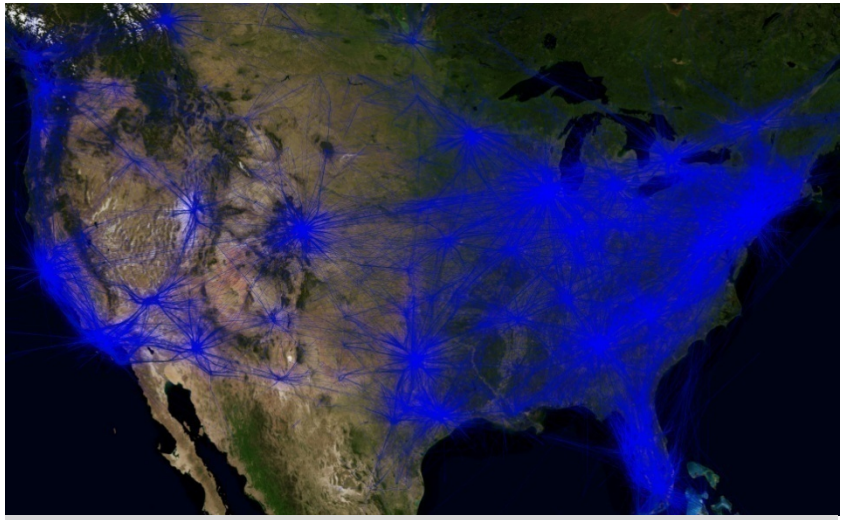
http://metabunk.org/threads/812-Visualizing-Flight-Paths-Above-30-000-Feet
Nick: FAA flight data for a 24 hour period



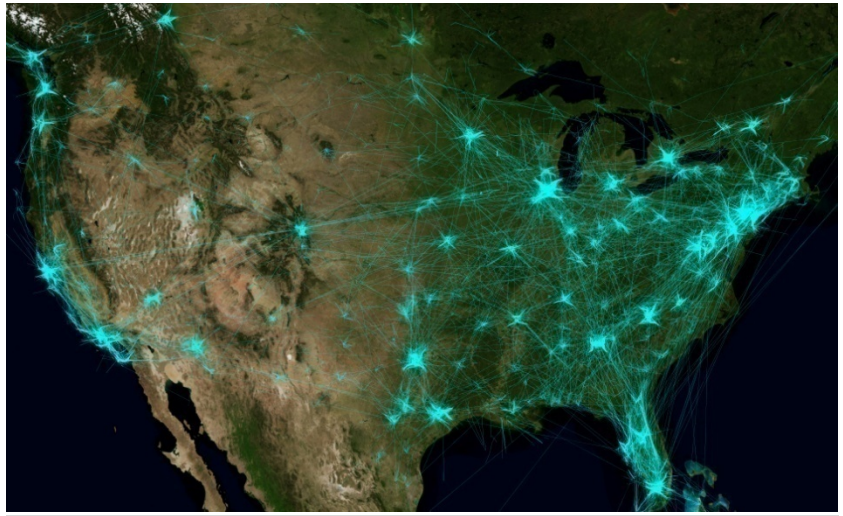
Above 40,000 feet



Above 30,000 feet



Below 30,000 feet

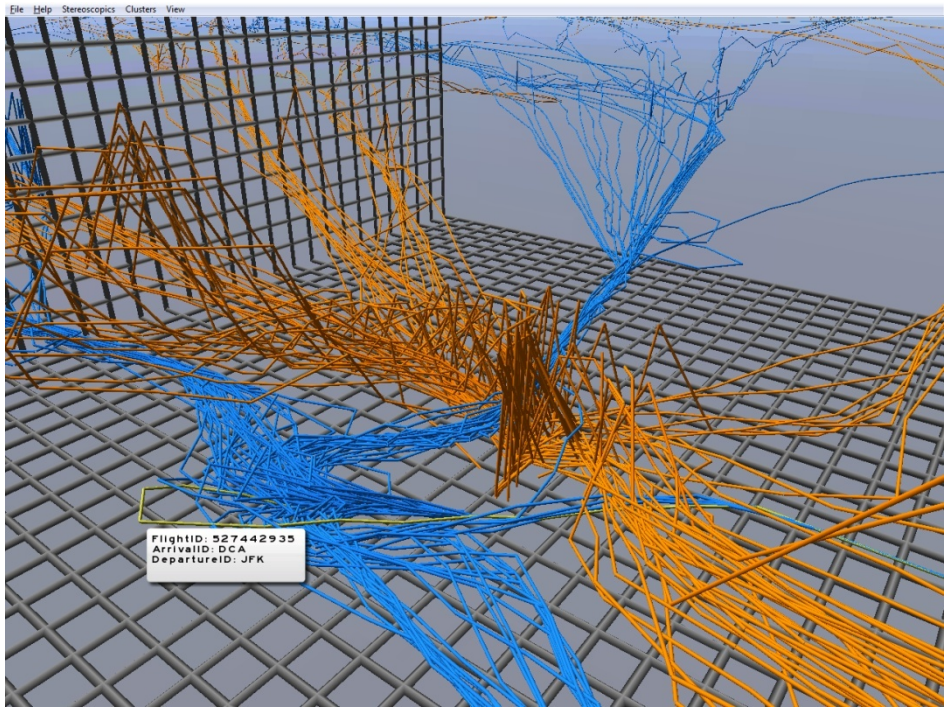


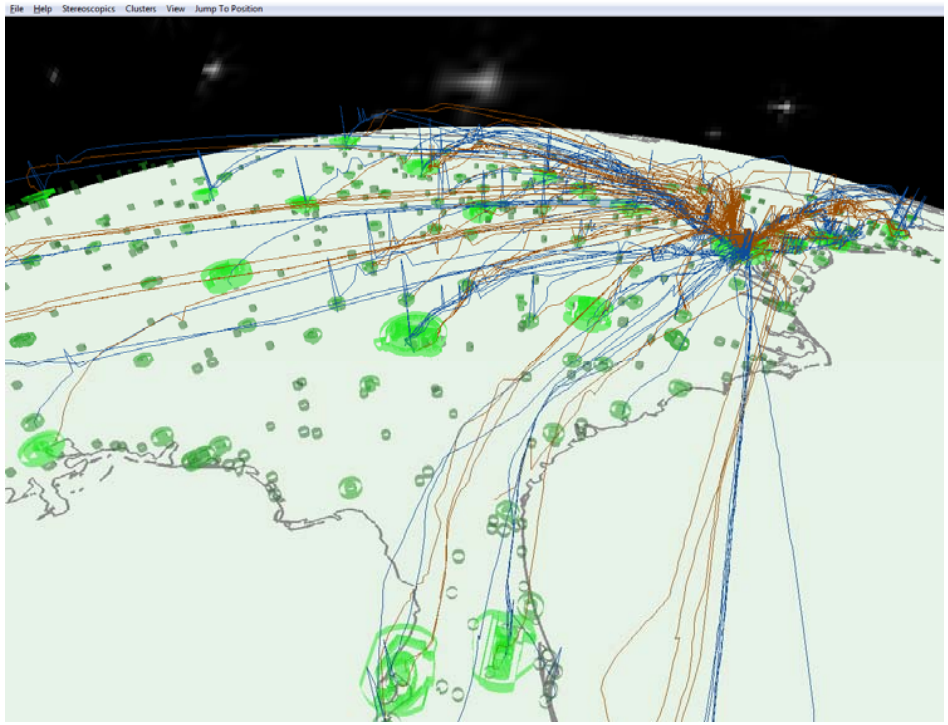
Below 10,000 feet

http://www.cattlab.umd.edu/?portfolio=sky-graph

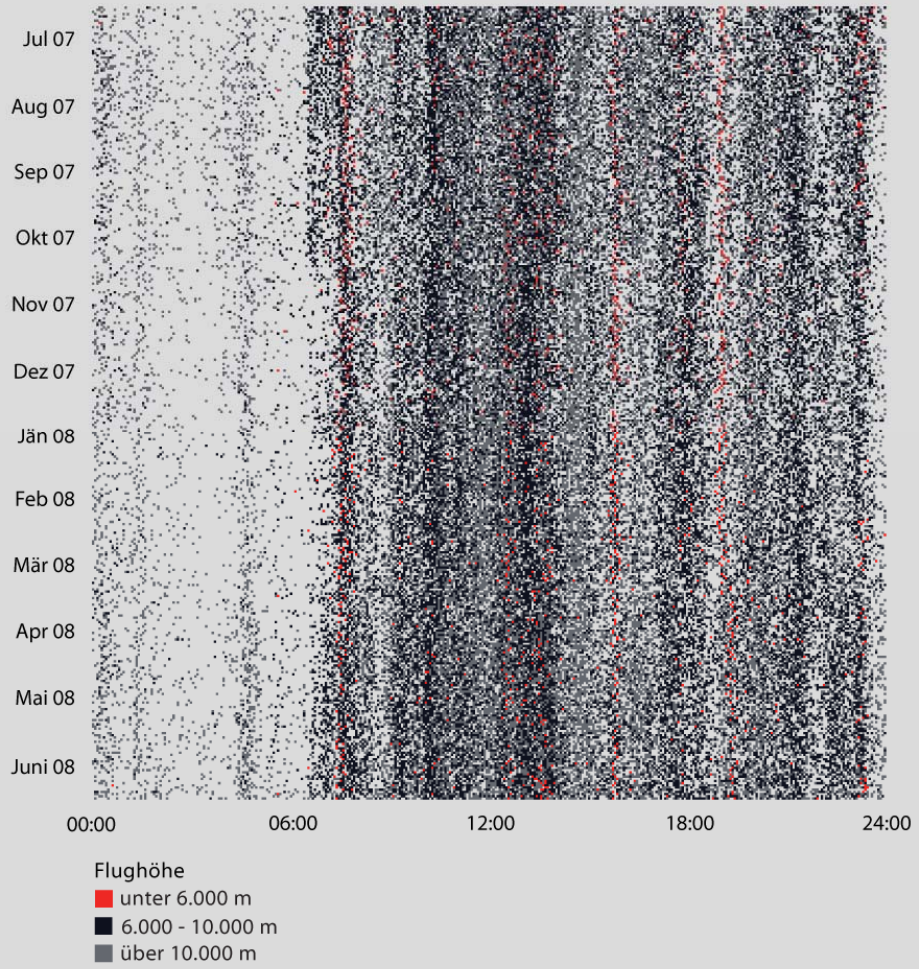
As air traffic rises at a rapid rate, congestion in National Airspace (NAS) becomes a primary concern of Air Traffic Control (ATC). This increase in traffic is pushing the limits of what controllers can handle in a given airspace. It is up to NAS designers to lay out optimal airspaces and routes that reduce the risk of incidents, allow for fuel efficient landings and limit the stress on air traffic controllers. **Sky Graph** provides a stereoscopic 3-dimensional environment where airspace designers can modify existing NAS elements as well as visually analyze historical flight and airspace data.

The goal of Sky Graph is to better visualize NAS, thus giving ATC personnel a better understanding of the structure of its elements and there interrelations. This visualization is also interactive allowing the user to select individual elements and modify them. Sky Graph can focus on different aspects of NAS to prevent an excess of data representation which can impede its exploration. All of these features combined with a stereoscopic viewing system offer an engaging and informative tool for ATC to use while planning NAS.

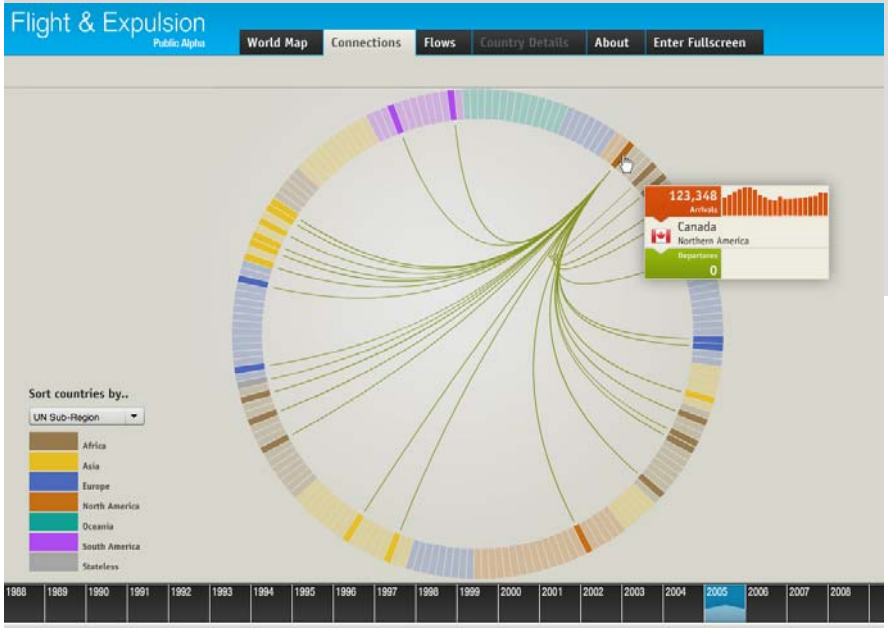




http://13bit.at/2009/01/31/airborne-radar/
Florian Waldner , All flights that passed Vorarlberg within a year



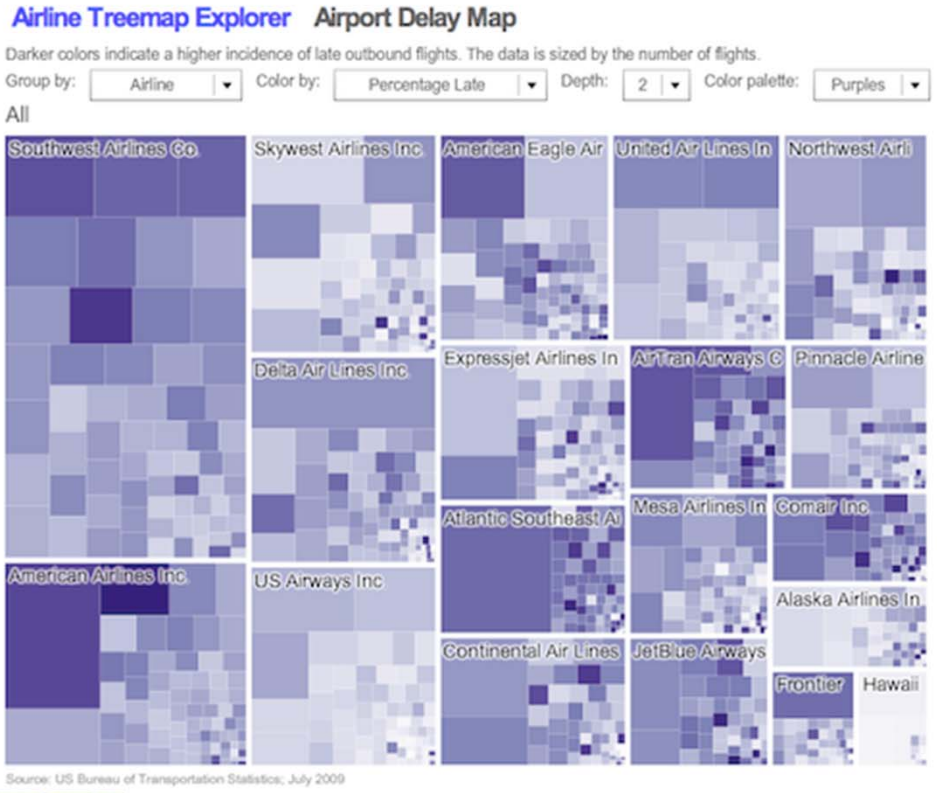
http://www.visualizing.org/stories/visualizing-human-migration
Christian Behrens , Visualizing Human Migration

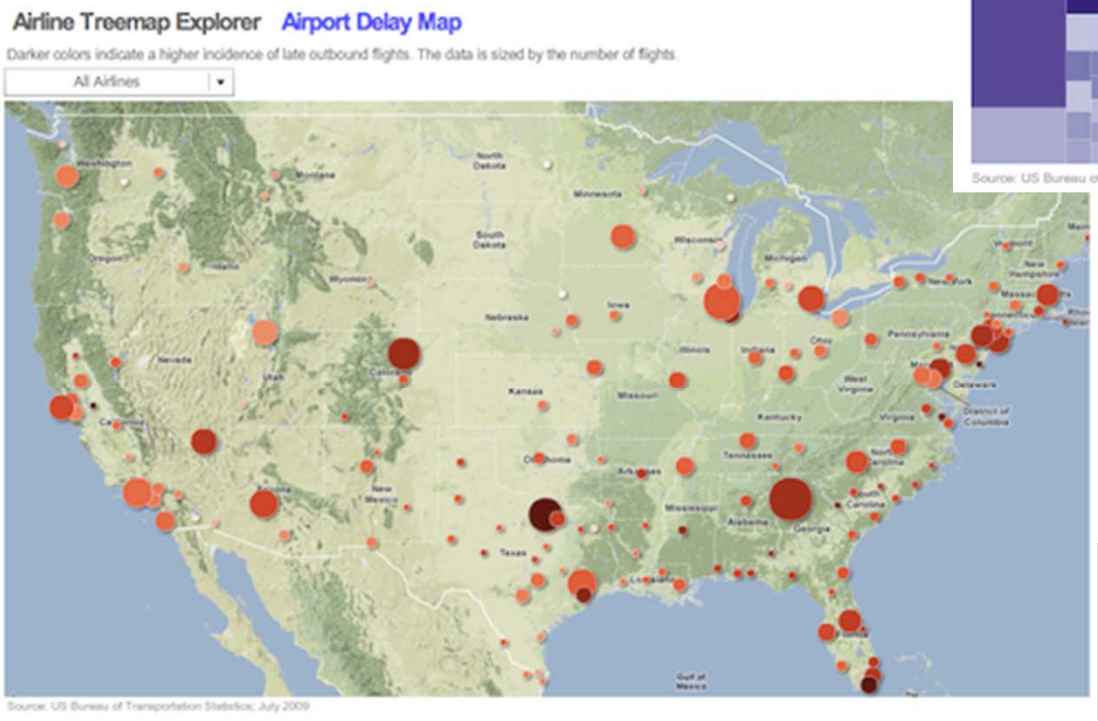




http://www.juiceanalytics.com/writing/airline-and-airport-traffic-and-delays/
Sal Uryasev, Airline and Airport Traffic and Delays: A JuiceKit Visualization Demo

Uses **size** to represent the number of flights by airline and by point of origin. The **color** is used to show delay time — we've got all sorts of delay metrics, each of which tells an interesting story.





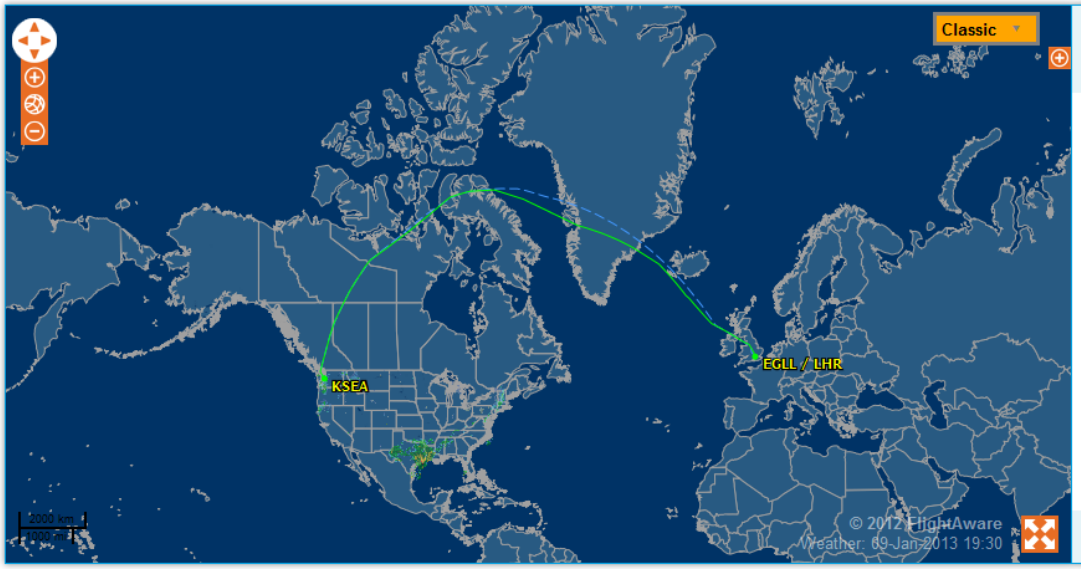
Uses **size** to represent the number of flights and the **color** to display delay time. Filtering by airline yields additional details.

Example Sources of Data

- http://stat-computing.org/dataexpo/2009/
- http://www.flightstats.com/
  - Real-time data sources include: FAA ASDI Data Feed, European Data Feed, GDS (Sabre, Amadeus, Apollo, Galileo), Direct Airport / Airline Data Feeds
  - Batch data sources include: Innovata Schedules, TSA Security Wait Times, Security Information, Health Information, Consular Information
- http://www.oagaviation.com/
  - Flight Status Data (dynamic, on demand, historical, alerts)
  - OAG Schedules (alerts, dynamic, SSIM, WDF, OnDemand)
  - OAG Reference data (DST/Time Zones, Locations, Carriers, etc.)
- http://www.faa.gov/
- http://www.caa.co.uk/
  - mostly statistical data, online data and CDs with a charge
- http://apps.bts.gov/
  - Research and Innovative Technology Administration, Bureau of Transportation Statistics
- http://uk.flightaware.com/
- http://www.flightglobal.com/

Live Flight Tracking
Aviation Photos
Squawks & Headlines
Discussions
Pilot Resources
Commercial Services
About FlightAware
Contact

FlightAware → Live Flight Tracker → British Airways (BA) #49 → 09-Jan-2013 → EGLL / LHR - KSEA Flight Tracker





British Airways 49 (Track inbound flight)

BAW49 · "Speedbird" (all flights) (website)

London Heathrow (EGLL / LHR) Terminal 5	Seattle-Tacoma Intl (KSEA)
14:46 GMT	15:49 PST
Scheduled: 14:30 GMT	Scheduled: 15:18 PST
7-day average: 14:43 GMT	7-day average: 15:26 PST

Other flights between these airports

Duration: 9 hours

Wednesday, 9 January 2013

Status	Arrived over 14 hours ago (track log & graph)
Aircraft	Boeing 747-400 (quad-jet) (B744 – photos)
Speed	Filed: 473 kts (graph)
Distance	Direct: 4,791 sm Planned: 4,793 sm Flown: 7,411 sm
Cabin	First / Business / Economy: Meal
Route	IWACK098002 JAKSN GLASR8

Google EARTH

LIVE FLIGHT TRACKER

PRIVATE FLIGHT TRACKER:
Flight/Tail# e.g. VIR1
TRACK FLIGHT

AIRLINE FLIGHT TRACKER:
Airline British Airways - BA
Flight # 49
TRACK FLIGHT

FORGOT THE FLIGHT NUMBER?

AIRPORT TRACKER/INFO
Airport Code KSEA
-or-
Airport City e.g. London
VIEW ACTIVITY
VIEW INFO



Up to 30% off weekend stays



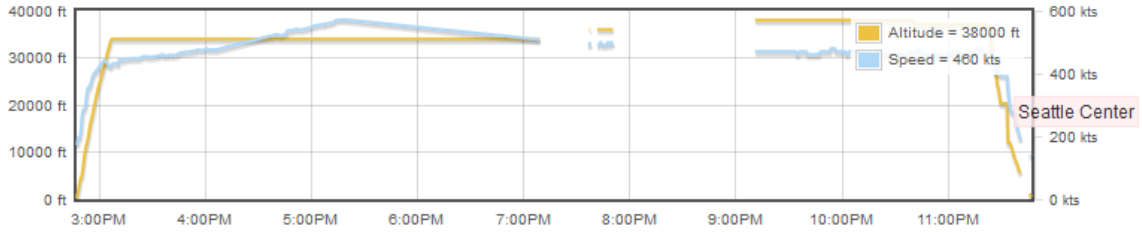
ACTIVITY LOG

Want a full history search for BAW49 dating back to 1998? Buy now. Get it within one hour.

Date	Aircraft	Origin	Destination	Departed	Arrival	Duration
11-Jan-2013	*B744*	London Heathrow (EGLL / LHR)	Seattle-Tacoma Intl (KSEA)	*14:40 GMT*	*15:55 PST*	Scheduled
10-Jan-2013	*B744*	London Heathrow (EGLL / LHR)	Seattle-Tacoma Intl (KSEA)	*14:40 GMT*	*15:36 PST*	Scheduled
09-Jan-2013	B744	London Heathrow (EGLL / LHR)	Seattle-Tacoma Intl (KSEA)	14:46 GMT	15:49 PST	9:03
08-Jan-2013	B744	London Heathrow (EGLL / LHR)	Seattle-Tacoma Intl (KSEA)	14:51 GMT	15:56 PST	9:04
07-Jan-2013	B744	Lond				
06-Jan-2013	B744	Lond				
05-Jan-2013	B744	Lond				
04-Jan-2013	B744	Lond				
03-Jan-2013	B744	Lond				
02-Jan-2013	B744	Lond				
01-Jan-2013	B744	Lond				
31-Dec-2012	B744	Lond				

easy!) view 4 months history. Join

(More)



Accident & Incident Data
Aviation Data & Statistics
Aviation Forecasts
Commercial Space Data
Funding & Grant Data
Passengers & Cargo
Research
Safety

Data & Research

Print

The FAA conducts research to ensure that commercial and general aviation is the safest in the world. Below you will find information about how the research is done, the resulting data and statistics, and information on funding and grant data.

Accident & Incident Reports

- Preliminary Data
- Final Data

More » Accident & Incident Data

Aviation Data & Statistics

- Airline On-Time Statistics & Delay Causes
- Airport Operations and Ranking

More » Aviation Data & Research

Commercial Space Data

- Upcoming & Recent Launch Data
- Historical Launch Data

More » Commercial Space Data

Forecasts

- FAA Aerospace Forecasts FY 2012-2032

More » Aviation Forecasts

Passengers & Cargo

- Unruly Passengers Statistics
- Passenger Boarding and All-Cargo Data

More » Passenger & Cargo Data

Safety

- Aviation Safety Information Analysis and Sharing (ASIAS)
- National Wildlife Aircraft Strike Database

More » Safety Data

Funding & Grant Data

- Airport Improvement Program (AIP)
- Procurement Forecast

More » Funding & Grant Data

Do You Want To... ?

Review preliminary accident data

Read final accident data

Look up airline on-time statistics

Search airport data

Research DOT airline data & statistics

Airport Cooperative Research Program
ACRP



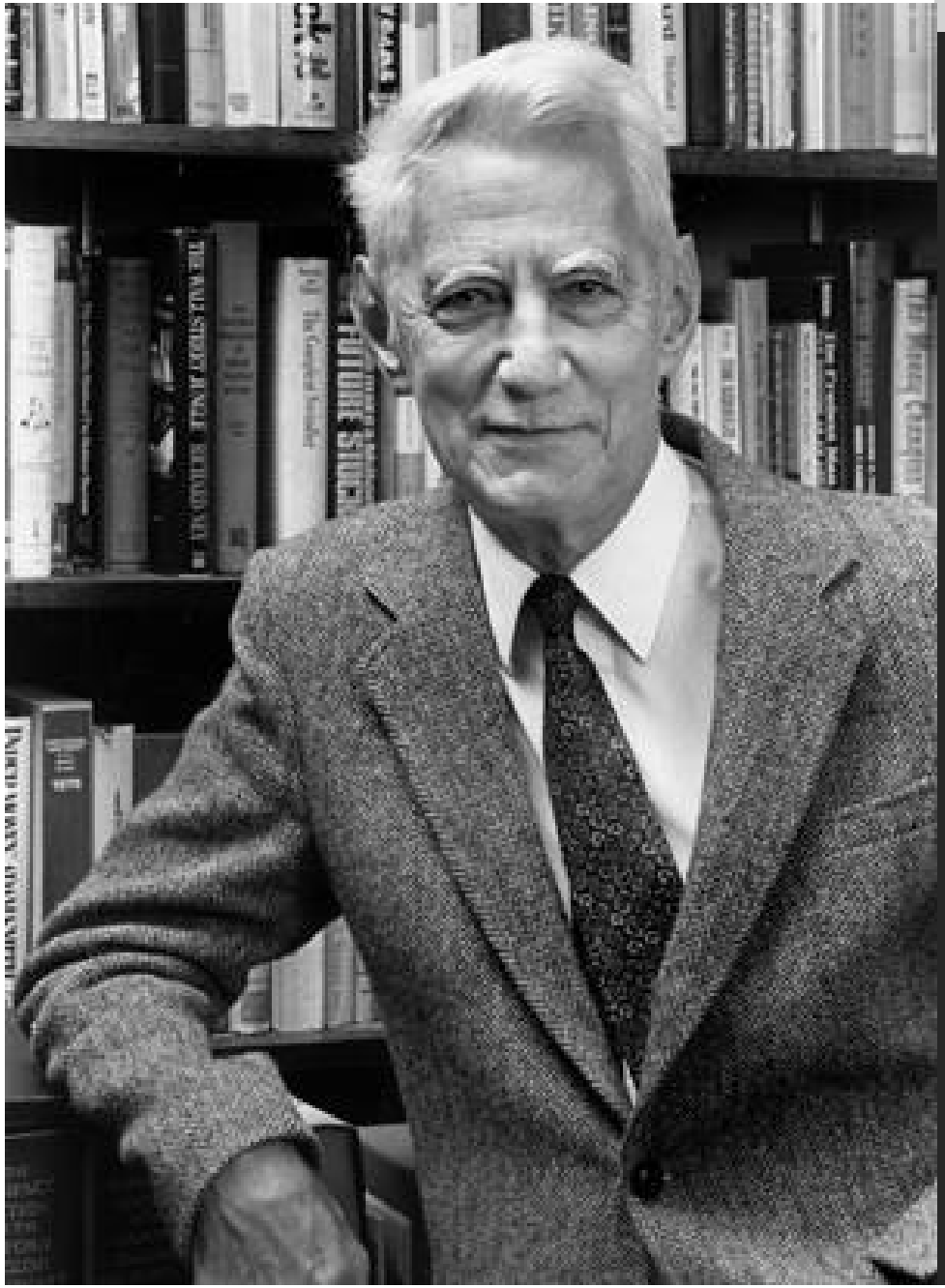
ASA Data

1. Year — 1987-2008
2. Month — 1-12
3. DayofMonth — 1-31
4. DayOfWeek — 1 (Monday) - 7 (Sunday)
5. DepTime — actual departure time (local, hhmm)
6. CRSDepTime — scheduled departure time (local, hhmm)
7. ArrTime — actual arrival time (local, hhmm)
8. CRSArrTime — scheduled arrival time (local, hhmm)
9. UniqueCarrier — unique carrier code
10. FlightNum — flight number
11. TailNum — plane tail number
12. ActualElapsedTime — in minutes
13. CRSElapsedTime — in minutes
14. AirTime — in minutes
15. ArrDelay — arrival delay, in minutes
16. DepDelay — departure delay, in minutes
17. Origin origin — IATA airport code
18. Dest destination — IATA airport code
19. Distance — in miles
20. TaxiIn — taxi in time, in minutes
21. TaxiOut — taxi out time in minutes
22. Cancelled was the flight cancelled?
23. CancellationCode — reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24. Diverted 1 = yes, 0 = no
25. CarrierDelay — in minutes
26. WeatherDelay — in minutes
27. NASDelay — in minutes
28. SecurityDelay — in minutes
29. LateAircraftDelay — in minutes

Our Main Technical Challenges

- How to define the probability distribution function?
  - simple statistics?
- How to relate entropy in different attribute dimensions?
  - simply mutual information?
- What is the visual representation?
  - spatial mapping
  - temporal mapping
  - visual mapping
- Dealing with incomplete and uncertain data
- Dealing with analytical algorithms (e.g., clustering, classifiers)

Claude E. Shannon (1916-2001)



Entropy

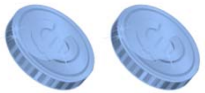
- Random variable
  - X
- It takes values
  - $x_1, x_2, \ldots, x_m$
- Probability mass function
  - $p(x_i)$
- Entropy

$$\boldsymbol{H(X) = -\sum_{i}^{m} p(x_i) \log_2 p(x_i)}$$

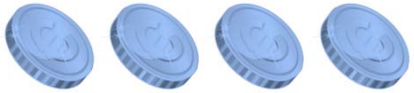
- Toss a bad coin once



  - x_1:0.1; x_2:0.9;
  - $\boldsymbol{H}(X)$ = 0.47 bits
- Toss a bad coin twice



  - x_1:0.01; x_2:0.09; x_3:0.09; x_4:0.81;
  - $\boldsymbol{H}(X)$ = 0.94 bits
- Toss a bad coin 4 times



  - $x_1 \sim x_{16}$;
  - $\boldsymbol{H}(X)$ = 1.88 bits

Entropy: Example

- Toss a good coin once



  - x_1:½; x_2:½;
  - **H**(X) = 1 bit
- Toss a good coin twice



  - x_1:¼; x_2:¼; x_3:¼; x_4:¼;
  - **H**(X) = 2 bits

An Information-theoretic Framework for Visualization

Min Chen and Heike Jänicke

Abstract— In this paper, we examine whether or not information theory can be one of the theoretic frameworks for visualization. We formulate new concepts and measurements for qualifying visual information. We illustrate these concepts with examples that manifest the intrinsic and implicit use of information theory in many existing visualization techniques. We outline the broad correlation between visualization and the major applications of information theory, while pointing out the difference in emphasis and some technical gaps. Our study provides compelling evidence that information theory can explain a significant number of phenomena or events in visualization, while no example has been found which is fundamentally in conflict with information theory. We also notice that the emphasis of some traditional applications of information theory, such as data compression or data communication, may not always suit visualization, as the former typically focuses on the efficient use of a communication channel, whilst the latter focuses on the effectiveness in aiding the perceptual and cognitive process for data understanding and knowledge discovery. These findings suggest that further theoretic developments are necessary for adopting and adapting information theory for visualization.

Index Terms—Information theory, theory of visualization, quantitative evaluation.

1 Introduction

Information theory is a branch of probability theory [27]. It is "the science of quantification, coding and communication of information" [44]. Visualization is concerned with visually coding and communicating information. Many aspects of a visualization pipeline feature events of a probabilistic nature, bearing a striking resemblance to a modern communication pipeline. For instance,

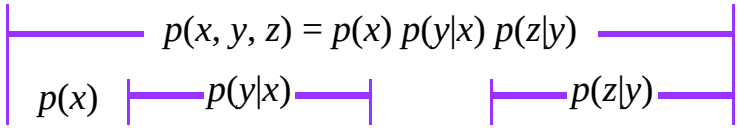
- data abstraction usually results in data compression;
- creating and viewing a visualization is usually an information discovery process;
- the messages in a visualization are not guaranteed to be received by a viewer;
- the quality of a visualization is often measured by probabilistic experiments; and so forth.

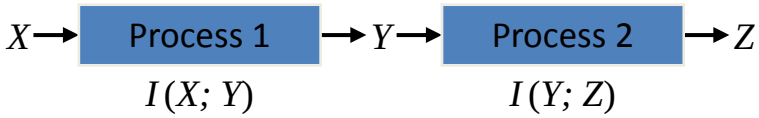
This suggests a strong connection between information theory and visualization. It is a reasonable assumption that the science of visual-

prove (a) is false. We thereby adopt an approach to examine the major concepts of information theory through its taxonomy, and for each major concept, we appraise its applicability to visualization. If a concept contradicts with some observations in visualization, we consider this as evidence indicating that information theory cannot explain one particular phenomenon. As the theoretical development of information theory is mature by nature, that of visualization is also makes sense to use the taxonomy of information theory as a basis and to view visualization from the perspective of information theory.

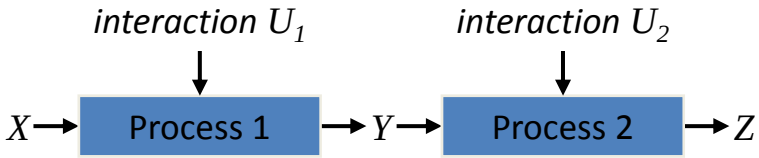
In the remainder of this paper, after a brief review of the literature, we examine the similarity and difference between a visualization system and a communication system in Section 3. In Section 4, we juxtapose the two subjects by outlining an information-theoretic taxonomy, and annotate the relevance in visualization. This is followed by a detailed examination of four major concepts of information theory in Sections 5-8. We then briefly discuss the role of user studies under an information theoretic framework in Section 9. We offer our

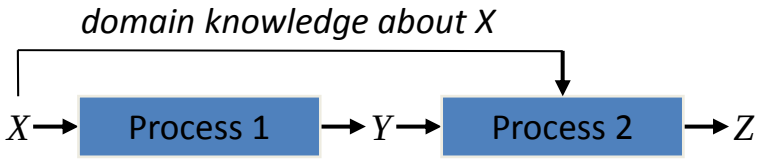
IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, 16(6):1206-1215, 2010





$$I(X;Y) \geq I(X;Z)$$

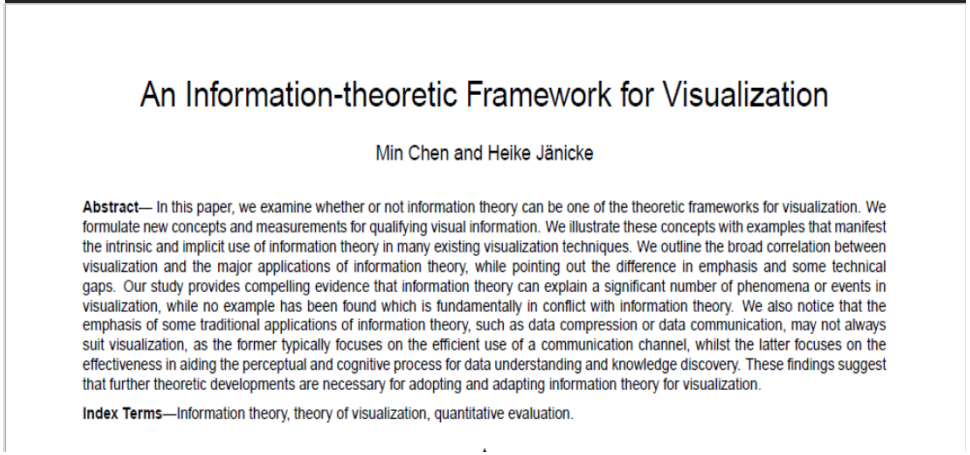




$$I(X;Y) \not\geq I(X;Z)$$

Data Processing Inequality

- "No clever manipulation of data can improve the inferences that can be made from the data"
  [Cover and Thomas, 2006]
- Markov chain conditions
  - Closed coupling: (X, Y), (Y,Z)
  - X and Z are conditionally independent
- What if the condition is broken?



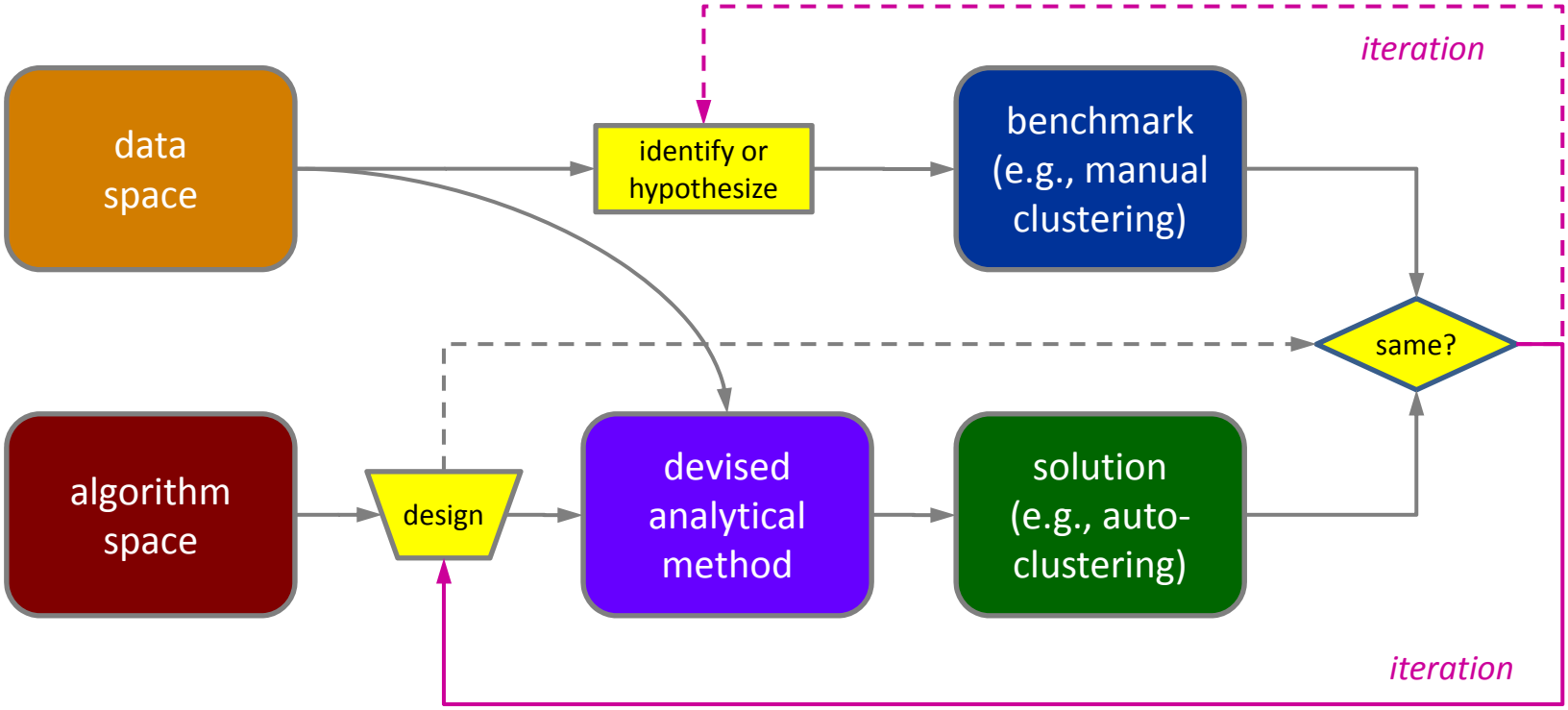
An Information-theoretic Framework for Visualization

Min Chen and Heike Jänicke

Abstract— In this paper, we examine whether or not information theory can be one of the theoretic frameworks for visualization. We formulate new concepts and measurements for qualifying visual information. We illustrate these concepts with examples that manifest the intrinsic and implicit use of information theory in many existing visualization techniques. We outline the broad correlation between visualization and the major applications of information theory, while pointing out the difference in emphasis and some technical gaps. Our study provides compelling evidence that information theory can explain a significant number of phenomena or events in visualization, while no example has been found which is fundamentally in conflict with information theory. We also notice that the emphasis of some traditional applications of information theory, such as data compression or data communication, may not always suit visualization, as the former typically focuses on the efficient use of a communication channel, whilst the latter focuses on the effectiveness in aiding the perceptual and cognitive process for data understanding and knowledge discovery. These findings suggest that further theoretic developments are necessary for adopting and adapting information theory for visualization.

Index Terms—Information theory, theory of visualization, quantitative evaluation.

A Typical Visual Analytics Pipeline

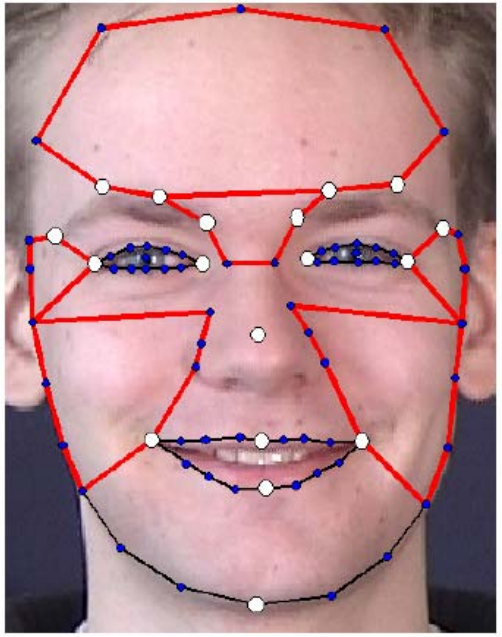




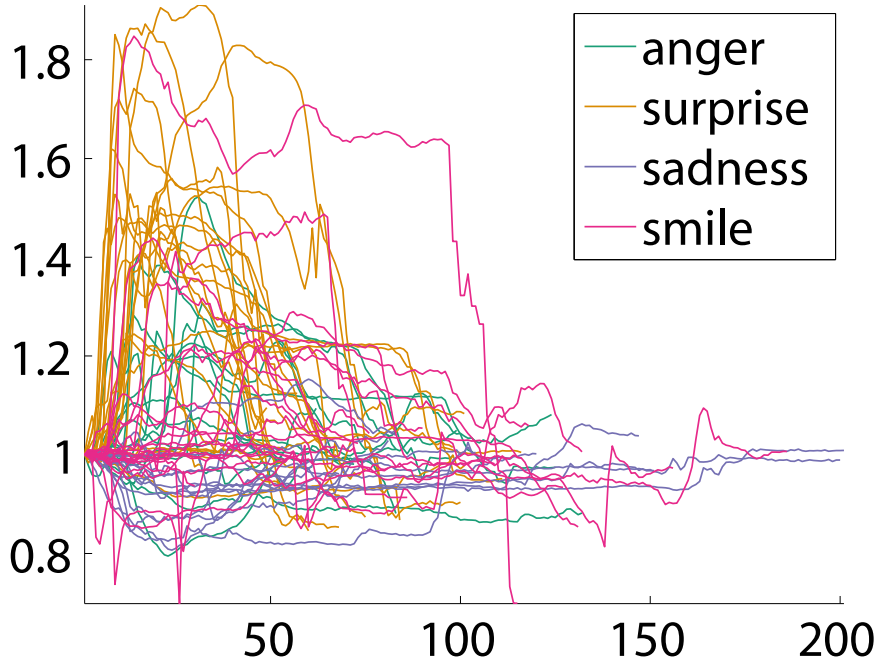
(a) smile (b) sadness



(c) surprise (d) anger



(e) feature points and regions



Example: Facial Dynamics

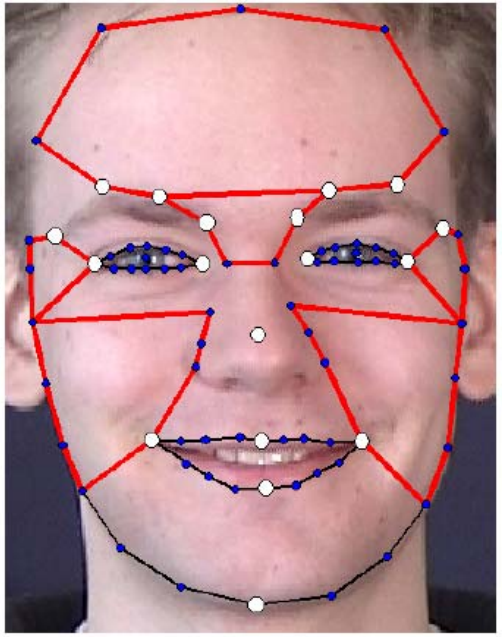
- Expression Recognition
  - Humans are very good at
  - Machine vision is far behind
  - Limited understanding
- Data
  - Video
  - Feature changes
  - Time series
- Challenges
  - A lot of features
  - A lot of ways of measuring features
  - Non-uniform temporal behavior



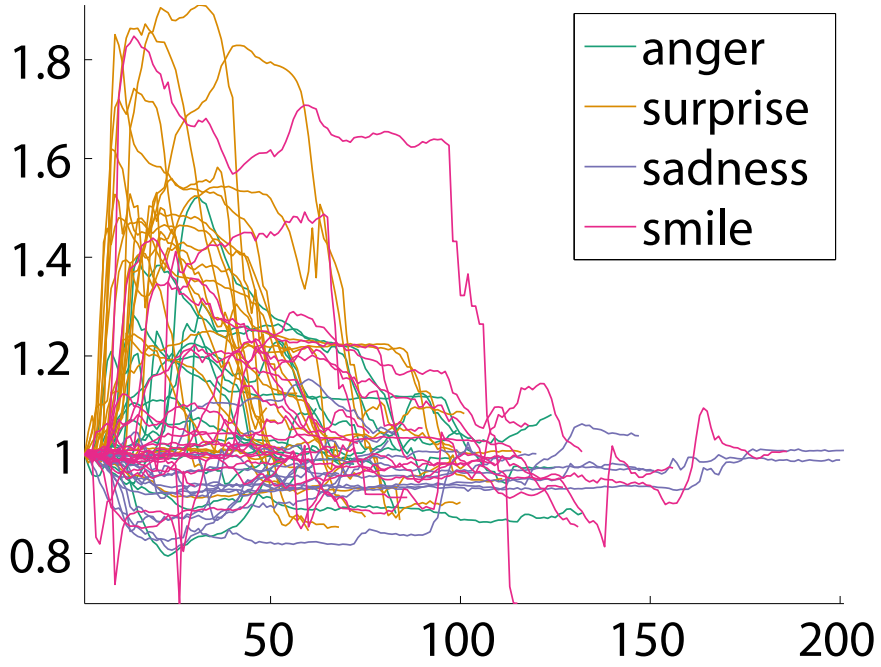
(a) smile (b) sadness



(c) surprise (d) anger



(e) feature points and regions



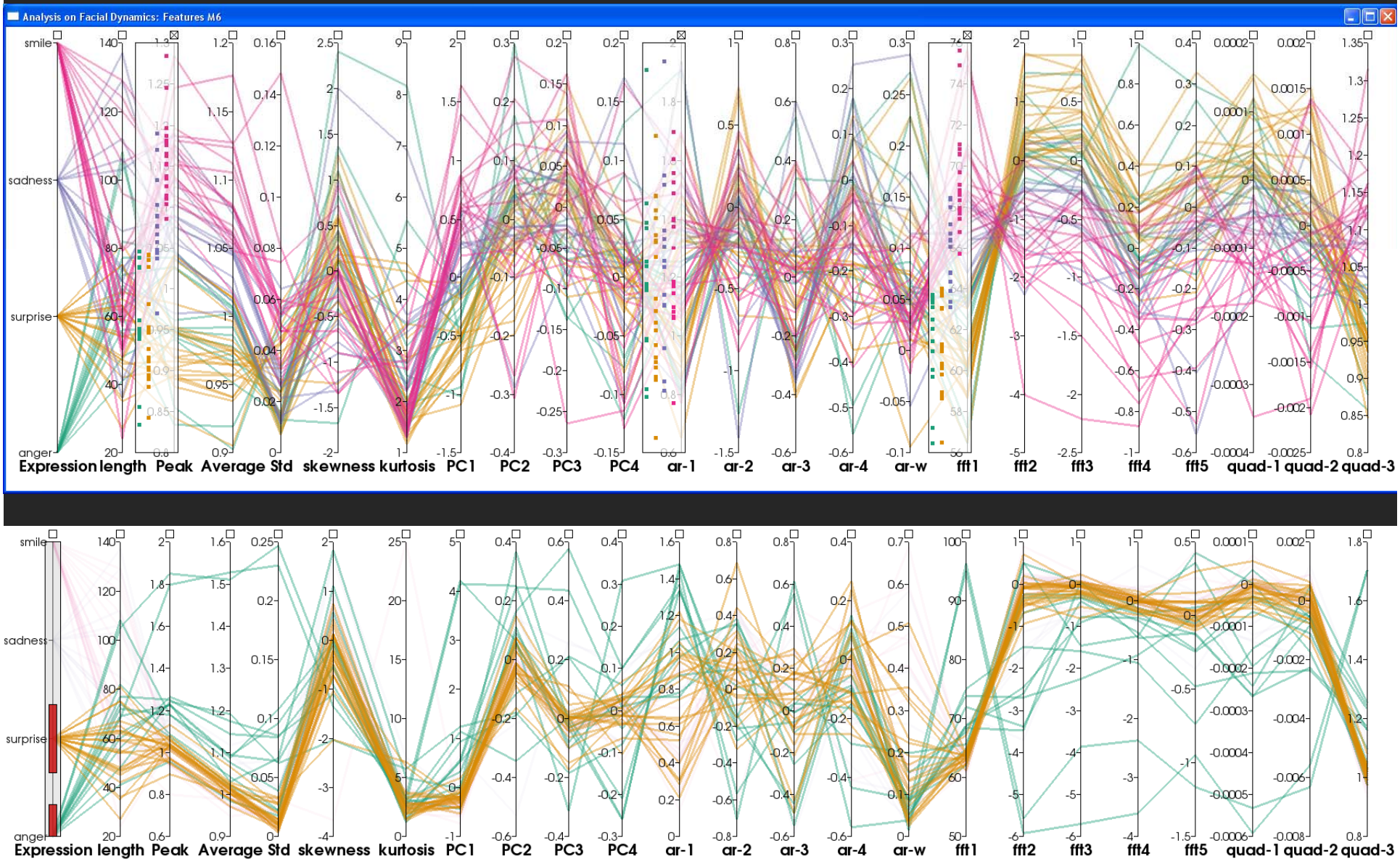
Auto-Analytical Methods

- Extract the Essences of Temporal Data

- Simple Descriptors
- Low-order Moments
- Principle Component Analysis (PCA)
- Fourier Transform
- Polynomial Fitting
- Auto-regressive Model
- … Many many more

- Algorithm Space

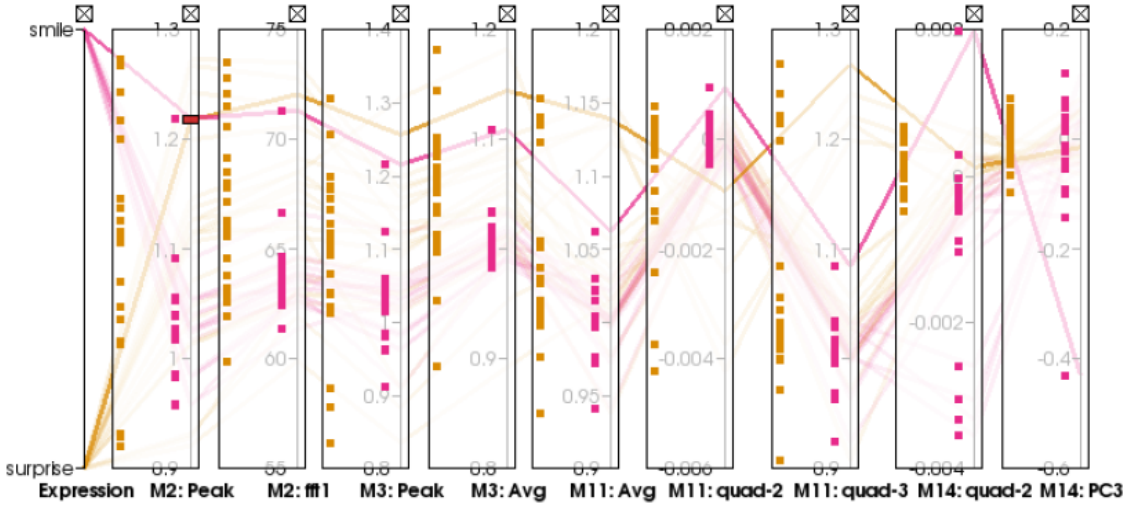
Visualization: Algorithm/Parameter Space



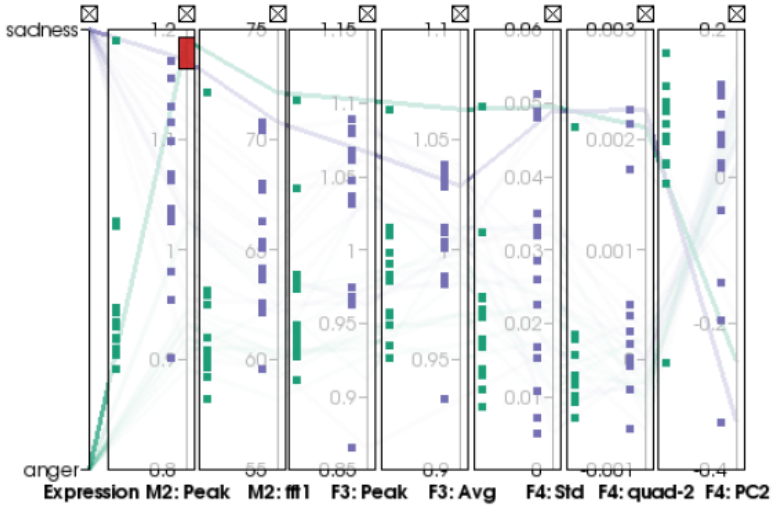
Interactive Visualization: Outliers



(a) 1905 (b) 1807 (c) 1931 (d) 1943

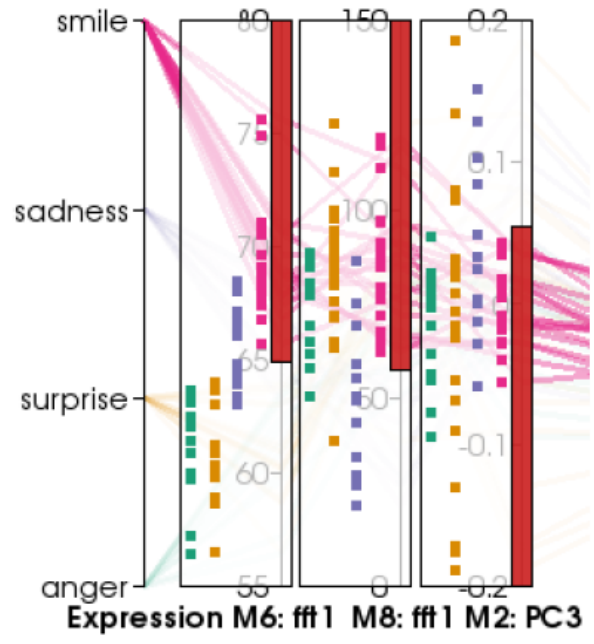


(a) Smile 1905 (outlier), Surprise 1807

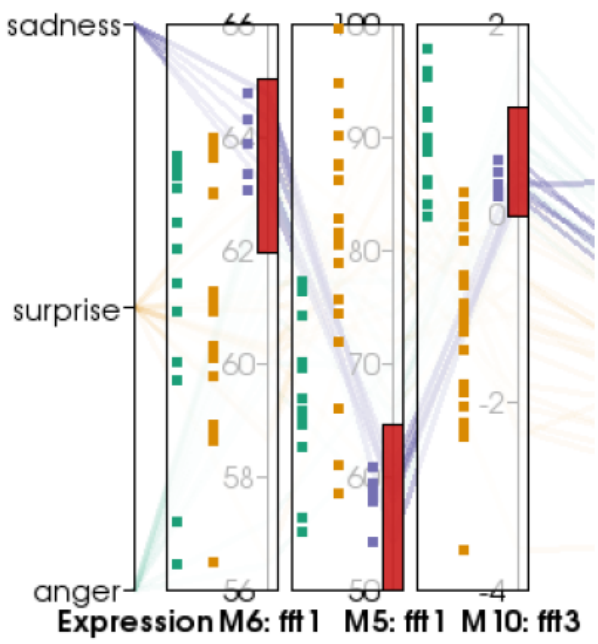


(b) Anger 1931 (outlier), Sadness 1943

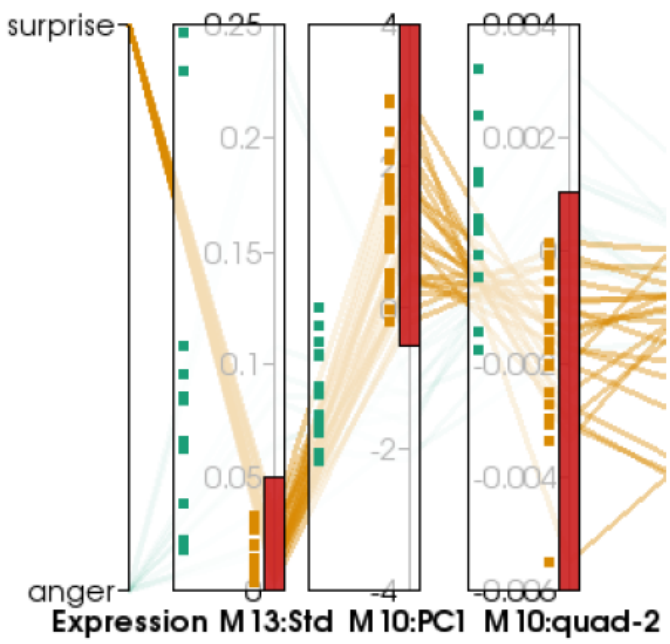
Interactive Visualization: Formulating Decisions



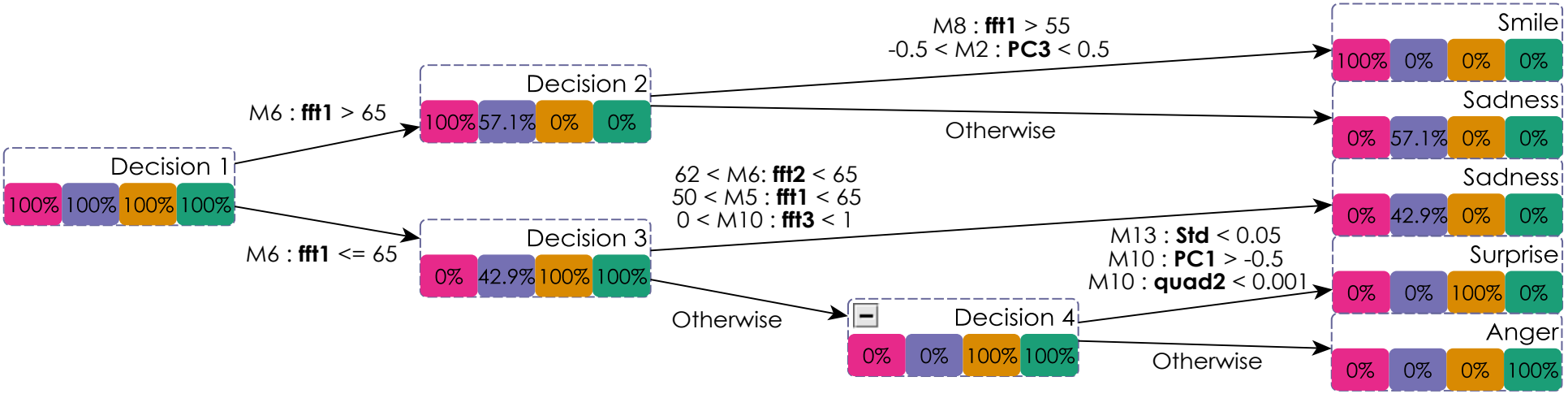
(a) Separate smile from all expressions

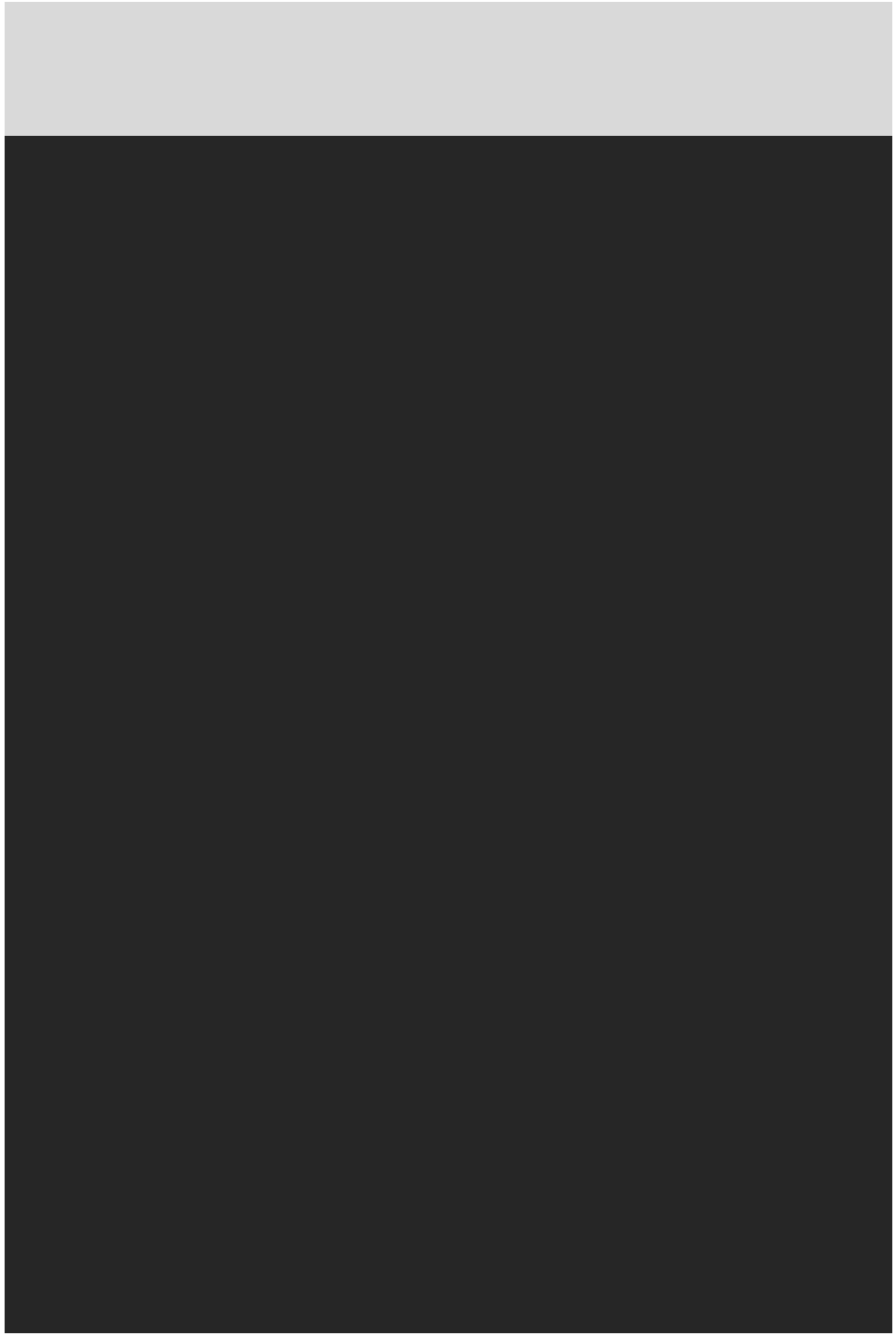


(b) Identifying sadness at Decision 3



(c) Identifying surprise at Decision 4





UKVAC II JRP

WP1: Information-Theoretic Framework

Part 2 (Confidential)

Professor Min Chen
Dr. Simon Walton
Oxford e-Research Centre
University of Oxford

UKVAC

PRIFYSGOL BANGOR UNIVERSITY 1884

Imperial College London

Middlesex University London

UNIVERSITY OF OXFORD

UCL

OXFORD e-Research CENTRE

UKVAC II Meeting, 31 May 2013

Objectives of WP1

a) To formulate an **information-theoretic framework** for visual analytics processes in a context exemplified by these two problems.

b) To develop **visual representations** that depict information contained in a dataset as well as the uncertainty quantities and other probabilistically-derived qualities such as importance and value.

c) To develop a **prototype** demonstration system for illustrating the potential use of information theoretic quantities in the process of hypothesis generation and evaluation.

d) To compare the information-theoretic framework with captured user study data (Objective 4) and to offer an explanation about any correlation and inconsistency.

e) To explore the application of the information-theoretic framework in interaction exploration (Objective 2), data management and data analysis (Objective 3), hierarchical decision models (Objective 5).

Accident & Incident Data
Aviation Data & Statistics
Aviation Forecasts
Commercial Space Data
Funding & Grant Data
Passengers & Cargo
Research
Safety

Data & Research

Print

The FAA conducts research to ensure that commercial and general aviation is the safest in the world. Below you will find information about how the research is done, the resulting data and statistics, and information on funding and grant data.

Accident & Incident Reports

- Preliminary Data
- Final Data

More » Accident & Incident Data

Aviation Data & Statistics

- Airline On-Time Statistics & Delay Causes
- Airport Operations and Ranking

More » Aviation Data & Research

Commercial Space Data

- Upcoming & Recent Launch Data
- Historical Launch Data

More » Commercial Space Data

Forecasts

- FAA Aerospace Forecasts FY 2012-2032

More » Aviation Forecasts

Passengers & Cargo

- Unruly Passengers Statistics
- Passenger Boarding and All-Cargo Data

More » Passenger & Cargo Data

Safety

- Aviation Safety Information Analysis and Sharing (ASIAS)
- National Wildlife Aircraft Strike Database

More » Safety Data

Funding & Grant Data

- Airport Improvement Program (AIP)
- Procurement Forecast

More » Funding & Grant Data

Do You Want To... ?

Review preliminary accident data

Read final accident data

Look up airline on-time statistics

Search airport data

Research DOT airline data & statistics

Airport Cooperative Research Program
ACRP



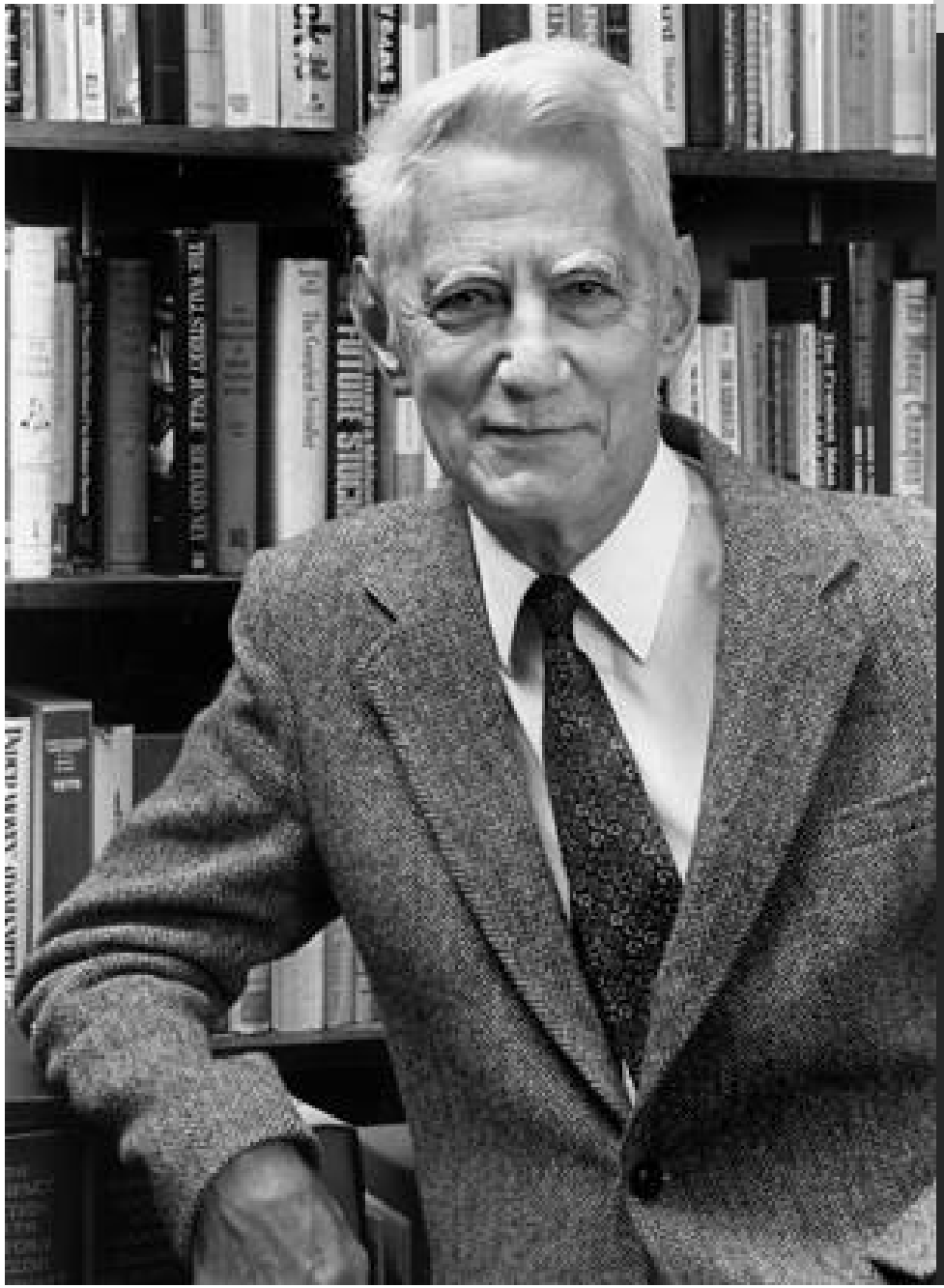
ASA Data

1. Year — 1987-2008
2. Month — 1-12
3. DayofMonth — 1-31
4. DayOfWeek — 1 (Monday) - 7 (Sunday)
5. DepTime — actual departure time (local, hhmm)
6. CRSDepTime — scheduled departure time (local, hhmm)
7. ArrTime — actual arrival time (local, hhmm)
8. CRSArrTime — scheduled arrival time (local, hhmm)
9. UniqueCarrier — unique carrier code
10. FlightNum — flight number
11. TailNum — plane tail number
12. ActualElapsedTime — in minutes
13. CRSElapsedTime — in minutes
14. AirTime — in minutes
15. ArrDelay — arrival delay, in minutes
16. DepDelay — departure delay, in minutes
17. Origin origin — IATA airport code
18. Dest destination — IATA airport code
19. Distance — in miles
20. TaxiIn — taxi in time, in minutes
21. TaxiOut — taxi out time in minutes
22. Cancelled was the flight cancelled?
23. CancellationCode — reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24. Diverted 1 = yes, 0 = no
25. CarrierDelay — in minutes
26. WeatherDelay — in minutes
27. NASDelay — in minutes
28. SecurityDelay — in minutes
29. LateAircraftDelay — in minutes

Visual Analytics Questions

- Given a flight record, *R*, is it normal or abnormal?
- Show me all abnormal records.
- What are the key indicator(s) for a specific type of phenomenon, or problem, or risk?
- Difficulties in Machine Learning
  - not enough anomalies to train.
  - uncertainty in annotation (assumed normality is not necessarily normal).
  - importance of different dimensions is unknown or uncertain.
  - context-sensitive knowledge (e.g., a VIP onboard or nearby)

Claude E. Shannon (1916-2001)



Entropy

- Consider one of the 29 attributes
  - X
- It takes values (e.g., letters in an alphabet)
  - $x_1, x_2, \ldots, x_m$
- Probability mass function
  - $p(x_i)$
- Entropy

$$\boldsymbol{H}(X) = -\sum_{i}^{m} p(x_i)\,\mathbf{log}_2\; p(x_i)$$

- For an attribute with continuous values, we can use differential entropy, with Δ and mean x_i

$$p(x_i)\Delta = \int_{x \in bin_i} p(x)\,dx$$

ASA Data

1. Year — 1987-2008
2. Month — 1-12
3. DayofMonth — 1-31
4. DayOfWeek — 1 (Monday) - 7 (Sunday)
5. DepTime — actual departure time (local, hhmm)
6. CRSDepTime — scheduled departure time (local, hhmm)
7. ArrTime — actual arrival time (local, hhmm)
8. CRSArrTime — scheduled arrival time (local, hhmm)
9. UniqueCarrier — unique carrier code
10. FlightNum — flight number
11. TailNum — plane tail number
12. ActualElapsedTime — in minutes
13. CRSElapsedTime — in minutes
14. AirTime — in minutes
15. ArrDelay — arrival delay, in minutes
16. DepDelay — departure delay, in minutes
17. Origin origin — IATA airport code
18. Dest destination — IATA airport code
19. Distance — in miles
20. TaxiIn — taxi in time, in minutes
21. TaxiOut — taxi out time in minutes
22. Cancelled was the flight cancelled?
23. CancellationCode — reason for cancellation (A = carrier, B = weather, C = NAS, D = security)
24. Diverted 1 = yes, 0 = no
25. CarrierDelay — in minutes
26. WeatherDelay — in minutes
27. NASDelay — in minutes
28. SecurityDelay — in minutes
29. LateAircraftDelay — in minutes

ignore condition normality variants

Conditional Probability

- Given a record R, with a set of known attributes
  - $a_1, a_2, \ldots, a_{29}$
  - It is always a low frequency event.
- Assuming Orthogonal Attribute Dimensions (to be discussed further)
- Divide attributes into three groups
  - (Possibly) Conditional Attributes
  - (Possibly) Insignificant Attributes
  - Variants of Normality (all remaining attributes)
- We call each of such configurations as a FaCAS (Focus and Context Attribute Subset).
  - In Portugueses, facas means knives

Conditional Probability (Assuming Orthogonal Attribute Dimensions)

- Given a FaCAS, and a record R, we can retrieve all records that have the same conditional attribute values as R.
- Let this collection of records are $R_1, R_2, \ldots, R_N$.
- Consider only the Variants of Normality, which defines an alphabet Z.
- Each unique combination (taking Δ whenever necessary) of variants is considered as a letter in the alphabet.
  - There are many letters normally.
  - Let the total number of letters to be $T_z \leq N$, and ideally, $T_z << N$.
- We can compute the probability of each letter and derive the entropy $H(Z)$.
- The level of uncertainty of this alphabet can be measured by $\frac{H(Z)}{\log_2 T_z}$
- Let the probability of R in the alphabet Z is $p(R)$. The level of self-information (also called surprisal) is $-\log_2(p(R))$
- For $R_1, R_2, \ldots, R_N$, We can compute the mean surprisal, standard deviation, etc. We can also visualize the surprisal value of R in conjunction with those of $R_1, R_2, \ldots, R_N$.

How to Develop a Heuristic System?

- Define a collection of “useful” FaCASs,
  - each of which has an operational role similar to a spam filter,
  - except in this case “spam” $\rightarrow$ “anomaly” and “filter” $\rightarrow$ “highlighter”
- One can interactively inspect a specific record by calling a set of FaCASs.
- One can also process a stream of incoming records dynamically and automatically
  - in a way similar to email spam filtering
- Visualization of Detected “Possible Anomalies”
  - An obvious approach is to colour-code records according to an aggregated anomaly level.
  - An alternative, or supplementary, approach is to use a glyph to display the multivariate detection results. A simple design of the glyph can just be a $k \times k$ pixel map.
  - Parallel coordinates can also be used to display a large set of incoming records.
- This approach may not sound ideal, as it makes use of human knowledge to identify useful FaCASs. Nevertheless, since the entire spam and virus defence industry is built on this methodology, it cannot be that “wrong”.

Some of the Challenges

- Understand the space of FaCASs
  - 29 attributes without considering sub-ranges
  - The number of combinations of these attributes in terms of the 3 attribute groups

$$3^{29} - 2^{29} = 68{,}629{,}840{,}493{,}971$$

- Understand the relationships between attributes, as they are clearly nor orthogonal.
  - Mutual information
  - Dimension reduction, e.g., MDS, PCA
  - But, some these techniques may remove anomalies.

Additional Notes

UKVAC II JRP

Problem 2: Flight Data (Overview, Part II)

Min Chen

Joint presentation with Jonathan C. Roberts
with Contributions from David Eberts,
Simon Walton and all other partners

Oxford e-Research Centre
University of Oxford















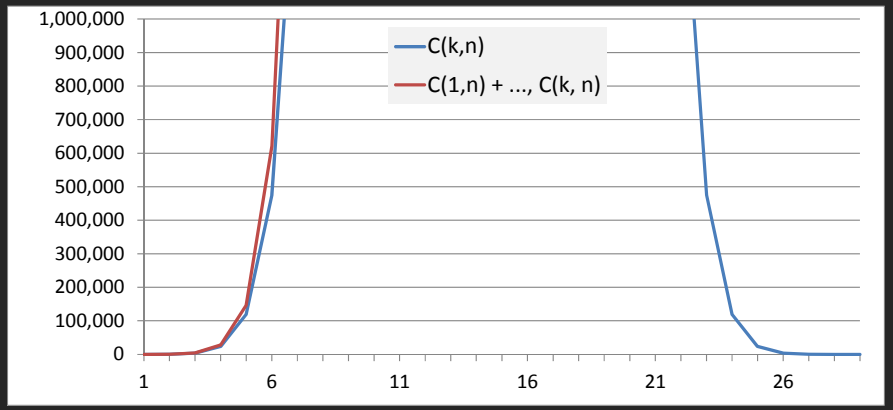
UKVAC II Meeting, 13 December 2013

Outline

1. Describe the problem
2. Describe how it represents some other big data problems
3. Existing visualization and visual analytics solutions
4. Quick point: Need an integrated software solution for VA
5. Software development environments for an integrated solution and a software development strategy
6. Why it is a hard problem and why existing solutions cannot address this fully
7. Three main categories of tasks for dealing with such data
8. How the work of WP1, WP2, WP3, WP4, WP5 fits with these three categories?
9. How to take WP1 and WP2 further?
10. How to transfer WP3, WP4, WP5 works into the FD problem?

Why it is a Hard Problem?

- 29 Attribute Dimension, plus many additional ones
- High range of data values for each most dimensions
- Seasonal effects
- Noise (disturbance) from infrequent events (e.g., airline merge, Olympics)
- Changing environmental variables (e.g., airlines merge, increasing fleet size)
- Are statistical indicators statistically significant (e.g., $p < 0.05$)?



- Anomaly is context-dependent
  - How many possible contexts?
  - Theoretical Total: 536,870,911
- Anomaly can be subjective or perceptual
- Paradox of Anomaly Annotation:
  - Anomaly may lead to negative or non-negative outcomes.
  - Do we annotate “Anomalies” with non-negative outcomes as “Anomalies”?

Further Statements about the Challenges

Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly Detection: A Survey", *ACM Computing Survey*, 14(3):a-15, 2009

- Encompasses every possible normal behaviour is very difficult. The boundary between normal and anomalous is often not precise.
- The malicious adversaries often adapt themselves to make the anomalous observations appear normal.
- In many domains normal behaviour keeps evolving and a current notion of normal behaviour might not be sufficiently representative in the future.
- The exact notion of an anomaly is different for different application domains.
- Availability of labelled data for training/validation of models used by anomaly detection techniques is usually a major issue.
- Often the data contains noise that tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.

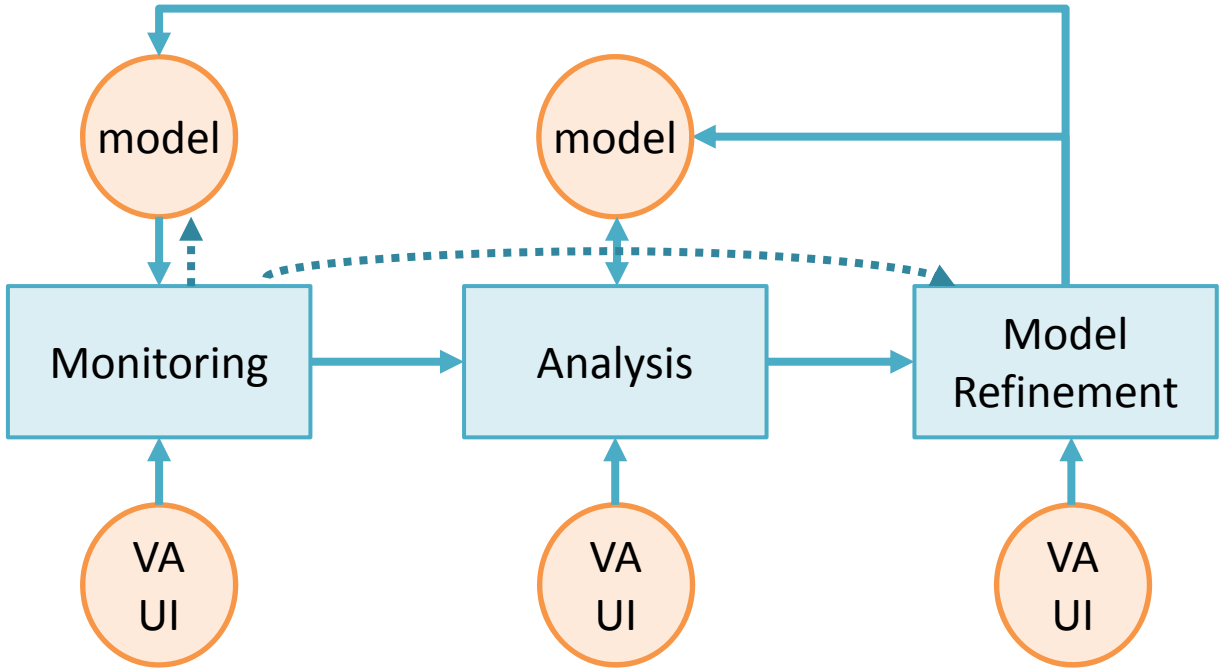
Existing Methods

Varun Chandola, Arindam Banerjee, and Vipin Kumar, "Anomaly Detection: A Survey", *ACM Computing Survey*, 14(3):a-15, 2009

- Classification Based
  - rely on the availability of accurate labels, threshold problem
- Clustering-based
  - distance measurement for multivariate data is a black art
  - cannot deal with sparsely distributed anomalies
- Nearest Neighbor Based
  - distance measurement for multivariate data is a black art
- Statistical
  - The assumption about data distribution may nor hold
  - Multivariate histogram is difficult to analyse
- Information-Theoretic
  - performance is dependent on the choice of the information theoretic measure
  - It is difficult to associate an anomaly score with a test instance
- Spectral
  - limited to situations where the normal and anomalous instances are separable in the lower dimensional embedding of the data

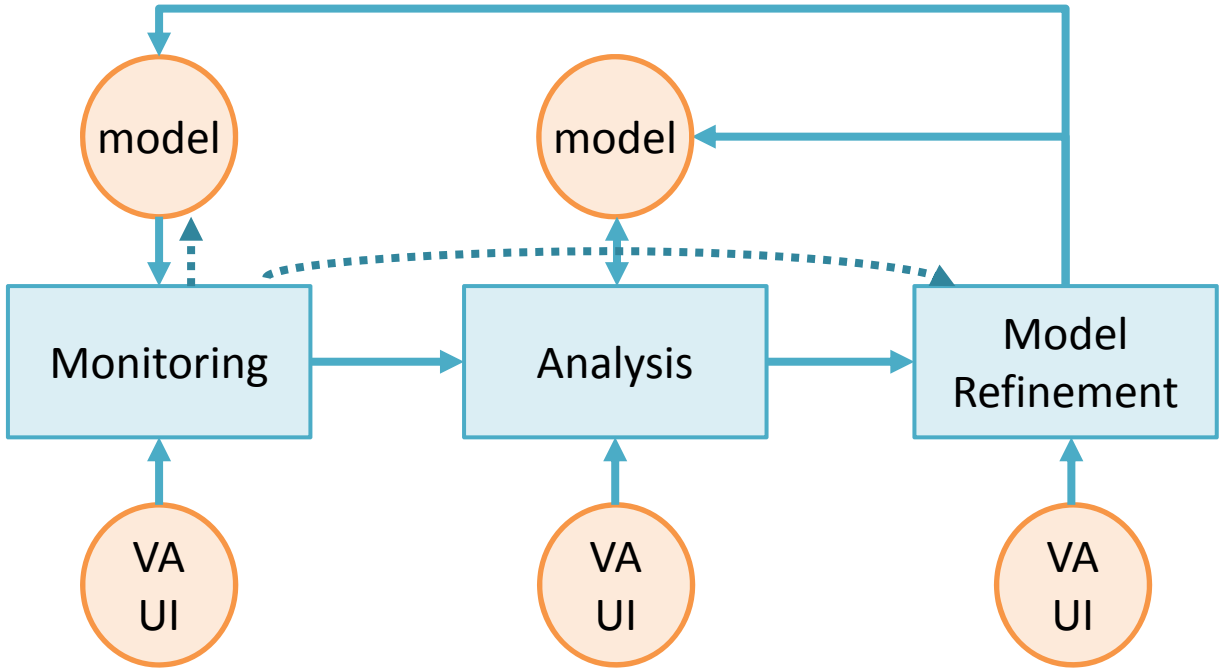
Three Categories of Tasks

- Monitoring: routine, low skill, quick RT, and more human resources
- Analysis: medium skill level, may take time, may need provenance
- Model Refinement: highly skilled, occasional operation (excluding active learning)



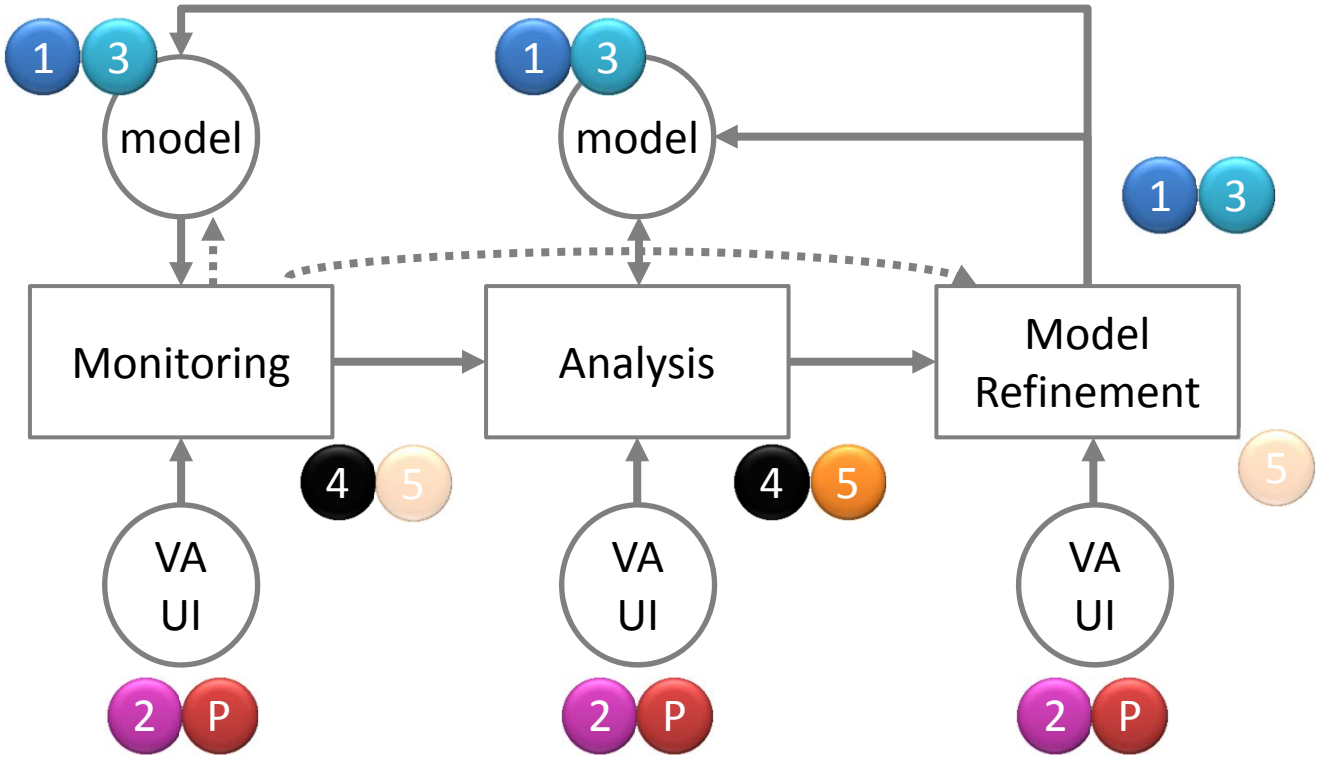
How the work of WP1-5 (and Purdue) Fits with These Three Categories?

- Monitoring: routine, low skill, quick RT, and more human resources
- Analysis: medium skill level, may take time, may need provenance
- Model Refinement: highly skilled, occasional operation (excluding active learning)



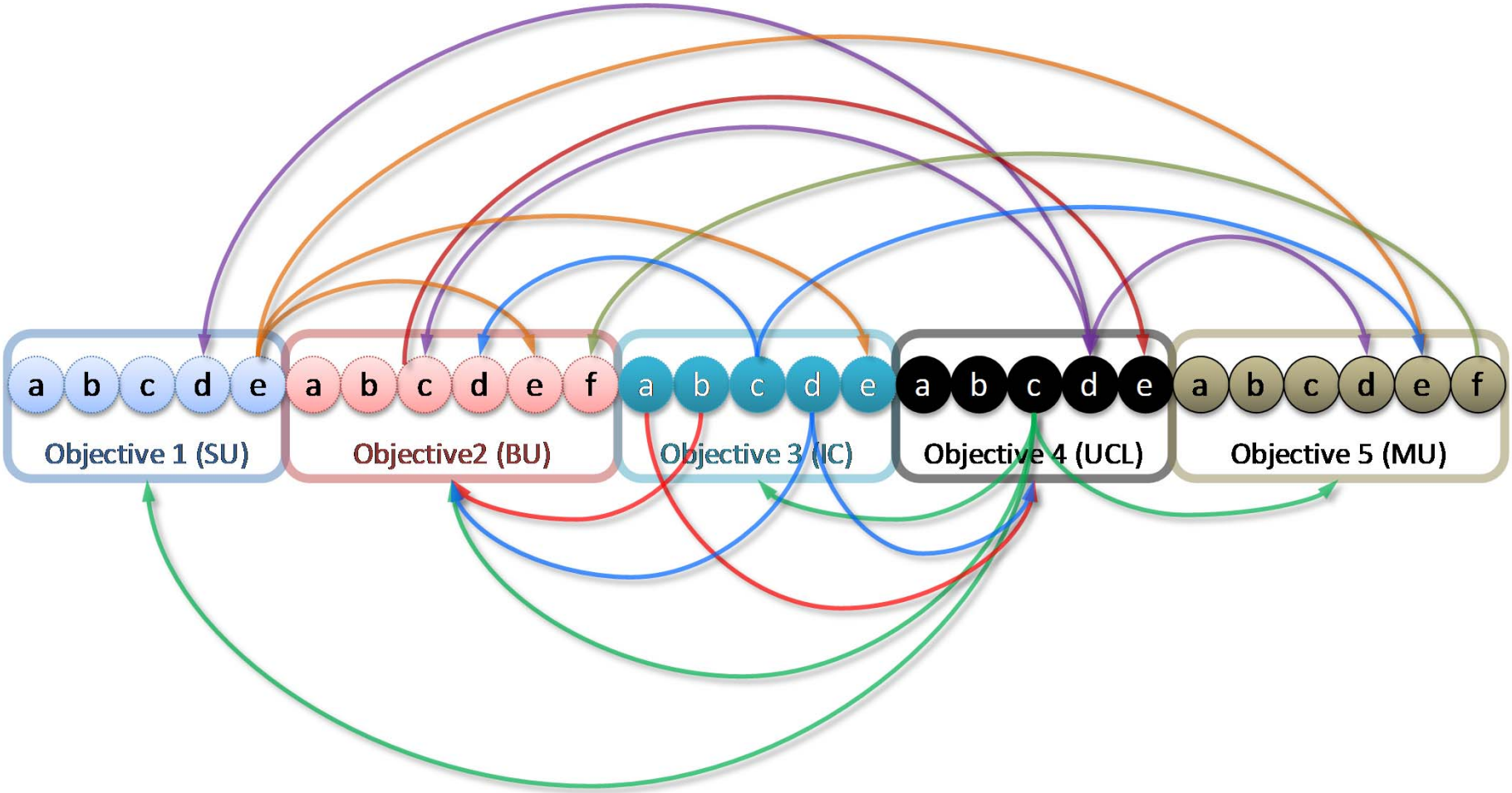
WPs (and Purdue)

- Monitoring: routine, low skill, quick RT, and more human resources
- Analysis: medium skill level, may take time, may need provenance
- Model Refinement: highly skilled, occasional operation (excluding active learning)



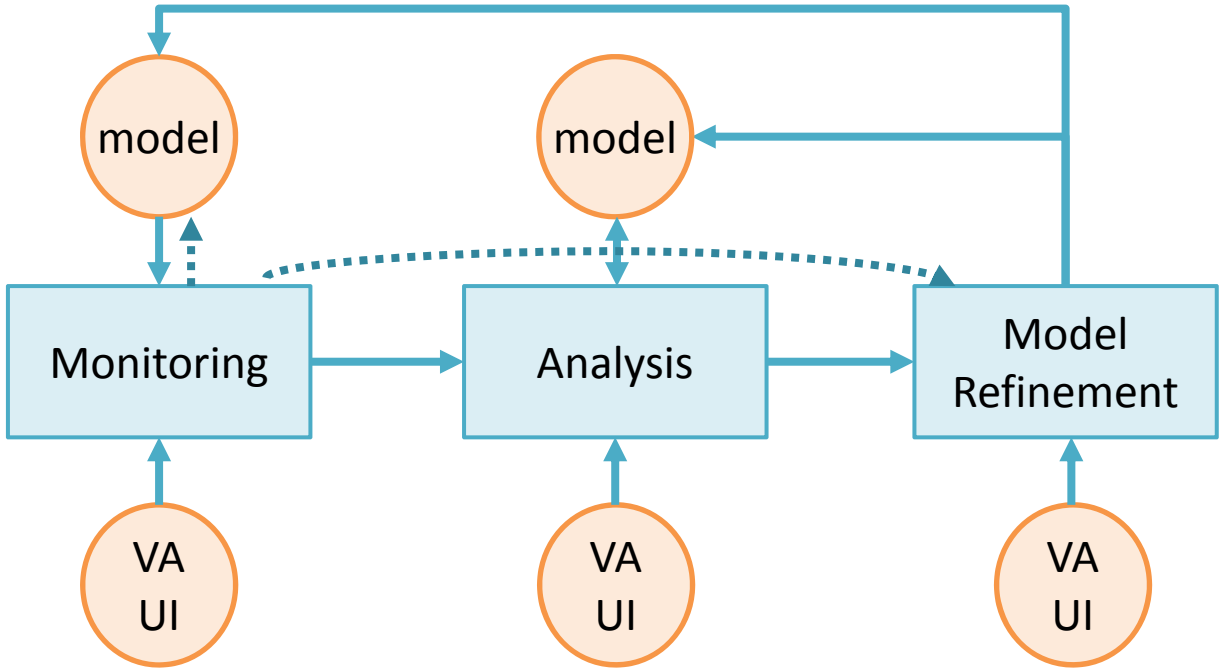
Future Work

- Information Theoretic Methods for Unstructured Data



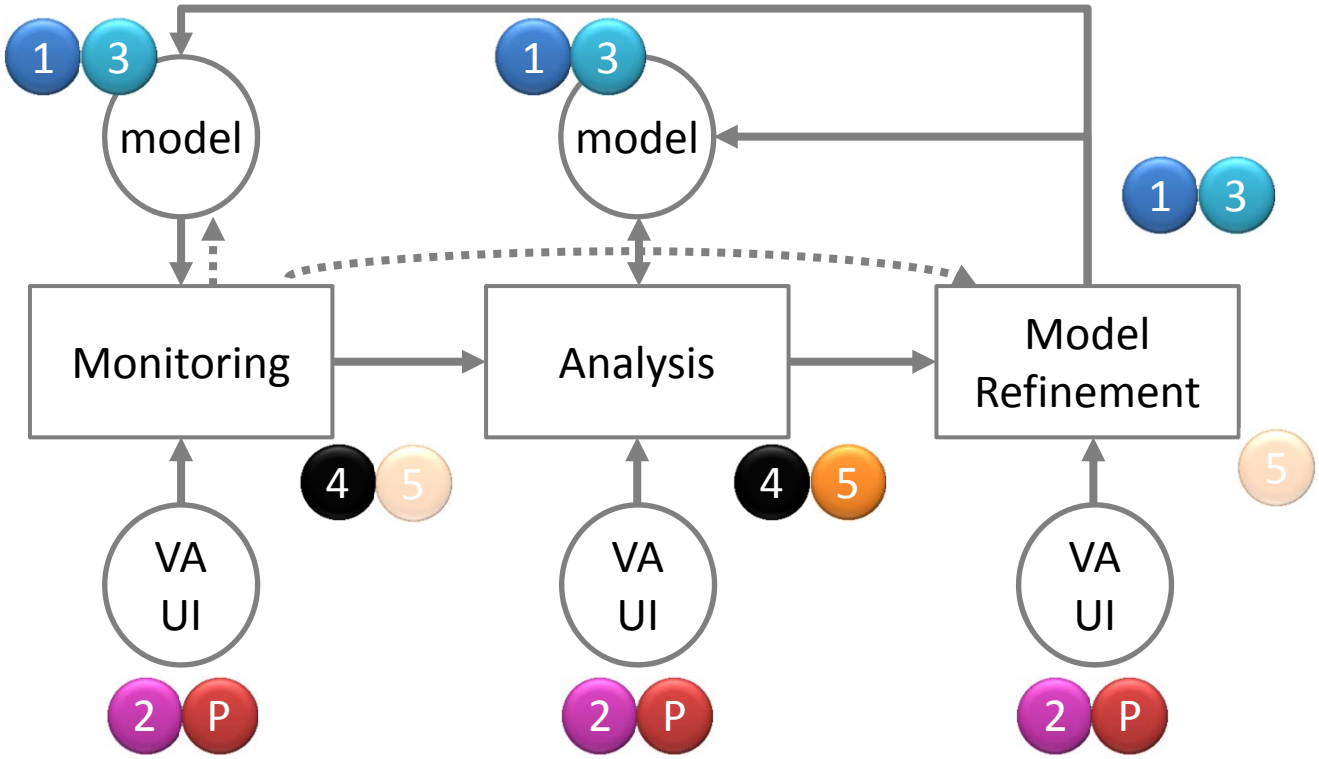
How to take WP1 and WP2 further?

- Mechanisms for context analysis and model building
- Model visualization, analysis and refinement
- Different visual and interaction designs



Transfer WP3, WP4, WP5 Results to the FD Problems

- With WP3: Hybrid Models, and Ensemble Models
- With WP4: Different Reasoning Strategies at Different Stages
- With WP5: Search in Anomaly detection



Future Work

- Many exciting directions that require in-depth research and collaboration

