**Conference proposal for**

**"Algorithmic, Mathematical, and Statistical Foundations of Data Science and Applications"**

*Petros Drineas and David F. Gleich*, Department of Computer Science, Purdue University

pdrineas@purdue.edu, dgleich@purdue.edu

## 1 Introduction

The amount of data in our world has exploded, and access to data and data science are increasingly the heart of modern economic activity, innovation, and growth.

Using data to achieve the potential economic and societal benefits is the goal of the growing field of data science. Doing so poses a unique set of multi-disciplinary and inter-disciplinary challenges across many disciplines within the college of science at Purdue.[1] As a taste, consider that a modern biology experiment will produce data that is too large to be analyzed in previously conventional tools such as Excel and SAS. This then requires new types of computing architectures (MapReduce, Spark, RHIPE, etc.) that enable scalable data computations. However, using these platforms may also require theoretical innovations in computer science to redesign fundamental analysis algorithms (randomized methods, communication-optimal, etc.). Now, the size, scale, and noise of these datasets requires new types of statistical modeling and analysis, generating further new algorithms.

### 1.1 Purdue's position in data science

As evidenced by the data science forums across campus, Purdue has a unique set of expertise in data science. Examples include excellence in algorithms for data science (computer science department), excellence in studying the privacy and security of data (computer science department and CERIAS), excellence in statistics and mathematics algorithms for data (stats and math department), novel applications (data science in climate, ITAP investigation of educational markers, uncertainty quantification in science and engineering).

[1] Of course, the field of data science extends more broadly than the areas inside the college of science and is also deeply involved with the areas in the college of engineering. We are just providing a single example.

### 1.2 Objectives

We propose to organize a two day workshop in the fall of 2018. The goal of the workshop is to drive the positive feedback cycle that exists in data science and to seek to open new, unique, Purdue specific and centric directions.

More specifically, we will bring together six high profile external speakers with Purdue faculty and researchers, with a particular emphasis on faculty and researchers from Computer Science, Statistics, Mathematics and applications in the Purdue College of Science. We will also advertise and invite to the workshop Purdue faculty and researchers that are interesting in learning about Data Science foundations in order to apply such foundational tools to their own research. The objectives of the workshop are three-fold:

(i) Identify fundamental foundational areas in the emerging discipline of Data Science where Purdue can transform it's current excellence into a global leadership via collaborations between computer scientists, mathematicians, statisticians, and applications.

(ii) Identify application domains that are of high priority to Purdue's mission and could be transformed by research in foundational areas of Data Science;

(iii) Assess how collaboration between computer scientists, mathematicians, statisticians, and applied scientists could potentially contribute to advancing and transforming the Data Science training at all educational levels: undergraduates, MS and Ph.D. students, and postdocs.

A ten-member committee (see Section 2.1) was formed in the College of Science in order to delineate the objectives of the workshop, prepare the proposal, and (if the proposal is funded) plan and organize the workshop. The team is led by the two PIs, who are ideally situated in order to organize this workshop. Their background in in Computer Science and Applied Mathematics; they have organized and participated in multiple Data Science workshops; they both have a significant amount of prior work and research on Data Science-related fields.

## 2 Details of the Proposed Workshop: Meeting Logistics

### 2.1 Steering committee

We have already formed a steering committee comprised of Purdue researchers from multiple relevant departments within the College of Science. All committee members have significant prior work in aspects of Data Science and will actively participate in the workshop organization. The members of the steering committee, in alphabetical order, are:

Bhadra, Anindya; Statistics · Cheng, Guang; Statistics ·
Cleveland, William S; Statistics · Drineas, Petros; Computer Science ·
Gleich, David; Computer Science · Lin, Guang; Mathematics ·
Neville, Jennifer; Computer Science · Taparowsky, Elizabeth J; Bio. Sci. ·
Xia, Jianlin; Mathematics · Zhang, Min; Statistics

### 2.2 Meeting format

We envision a two-day workshop, to be held on Friday and Saturday in October 2018. We intend to structure the workshop as six two-hour sessions, with each session hosting a talk from one of the six invited speakers and a few shorter presentations from Purdue researchers and faculty. Each session will be coherent and have a specific theme: for example, one session could focus on the impact of Numerical Linear Algebra and Matrix Computations in Data Science, with the external speaker giving a 50 minute presentation on the topic and three internal speakers giving shorter, 20-minute presentations on the topic, with 10 minutes left for Q&A sessions. Each session will also mention applications of the techniques.

In more detail, on Friday, after appropriate introductory comments from the organizers and the Dean and/or the Associate Deans of the College of Science, a potential format could include three two-hour sessions (two in the morning and one in the afternoon). The lunch break on Friday would be combined with a poster session, where anyone would be invited to present posters on their Data Science relevant work (with a focus on graduate students, but open to all!). On Friday evening, we will have a reception to facilitate interactions among the participants in a more relaxed environment. We intend to hold the reception within Purdue's facilities to further showcase our university to the external participants. On Saturday, we expect the same format, but instead of a

poster session during the lunch break, we intend to have a panel discussion in order to foster more lively interactions between the participants, the organizers, and the external speakers.

The exact format of the workshop will be determined after the list of external speakers have been finalized and in consultation with the steering committee.

### 2.3 List of potential external speakers

We expect to invite six external speakers. The following list of ten speakers serves as a starting point for invitations to be issued; the proposed speakers have collaborated with the organizers in the past or have spoken in meetings organized by members of our steering committee. Our list includes many members of the National Academies, foreign academies, officials of prestigious professional organizations (such as SIAM), Sloan fellows, Packard winners, etc. The list is in alphabetical order.

- Lana Adamic, Facebook
- Albert-Laszlo Barabasi, Northeastern
- David Donoho, Stanford
- Anna Gilbert, University of Michigan Ann Arbor
- Piotr Indyk, MIT
- Ilse Ipsen, NCSU
- Michael Jordan, UC Berkeley
- Alex Smola, Amazon
- Steven Strogatz, Cornell
- Hal Varian, UC Berkeley

### 2.4 Budget

We are requesting **$25,000** for the special meeting. About half of the funds will be used towards reimbursing the travel and lodging expenses for the non-local speakers; a small honorarium will

be provided as well. The remaining funds will be used to *(i)* pay for local expenses (meeting venue, coffee breaks, reception for the participants, etc.) and *(ii)* provide a limited number of student fellowships to graduate and undergraduate students from nearby institutions to attend the conference, in order to further publicize Purdue's contributions in the field of Data Science.

## 3 Addressing the evaluation criteria

- **Linkage to the mission of Purdue.**

  The proposed conference will cover and publicize world changing research that is performed at Purdue University. Our format (sessions where external speakers present a topic and internal speakers highlight the contribution of Purdue researchers on this topic) will facilitate the exposition of our contributions in Data Science. Additionally, we expect that this format will motivate further research on Data Science topics. Since many of the involved disciplines in Data Science are fundamental for STEM education, we also expect that this conference will promote our leaderships in STEM. Finally, transformative educational ideas could be a potential output of the conference, since many of the external speakers come from colleges and universities that are also considering the incorporation of Data Science in the undergraduate curriculum, by even making mandatory elementary Data Science courses in the freshman year (e.g., UC Berkeley). We intend to ask our external speakers to discuss potentially transformative teaching practices that are employed in their colleges and universities in their presentations.

- **Reputation of potential invited speakers/lecturers.**

  As we already mentioned in Section 2.3, our speakers are world famous statisticians, mathematicians, computer scientists, and have had major contributions in Data Science. Many of the proposed speakers are members of the National Academies, foreign academies, officials of prestigious professional organizations (such as SIAM), Sloan fellows, Packard winners, etc.

- **Critical mass of Purdue scholars likely to participate.**

  Data science is literally everywhere within Purdue. Most certainly, within the College of Science, we expect significant participation from Computer Science, Mathematics, and Statis-

tics, but also from Biology, Chemistry, and Physics, which traditionally use data analysis tools and techniques to analyze their experimental data. Beyond the College of Science, we do intend to publicize our conference to the Agriculture, Engineering, and the Business School, since all these entities also leverage Data Science tools in their research. As a matter of fact, we intend to invite speakers from the aforementioned departments in sessions if they fit the topic of the session.

- **Likelihood to attract leading scholars from elsewhere.**

  We will publicize the conference to all nearby institutions (Indiana University, University of Chicago, Northwestern, University of Illinois Urbana Champaign, University of Michigan Ann Arbor, etc.). All these schools already have strong Data Science clusters and will be interested in participating in a local conference on this topic. Recall that we will allocate part of our budget to support students from nearby institutions to participate to our conference and present their work in the poster session.

- **Likelihood to showcase Purdue.**

  Our faculty (see, for example, the work of the members of the steering committee, Section 2.1) has had major contributions in Data Science. The presentations from the internal speakers will highlight those contributions by our faculty. Additionally, holding the conference and all related events within Purdue will highlight the university facilities and world-class research more generally.

- **Integration of networking opportunities (e.g., participants as well as Purdue stakeholders).**

  The meeting has two main networking opportunities. The first is the poster session, which is designed to enable interaction between individuals at Purdue and the external participants by showcasing what we expect to be mainly Purdue work. The goal here is really, though, to connect Purdue people with other Purdue people towards the goal of finding new directions to pursue and overcoming challenges. The second opportunity is the more relaxed environment of the reception, where the goal is to create a forum for open-ended discussions that will result in possible future investigations and grant proposals by the more established par-

ticipants (e.g. faculty). In addition, we will have the standard coffee breaks that facilitate these interactions as well as dedicated Q&A periods during the two-hour sessions.

- **Linkage to at least one professional association, if possible.**
  We will try to get SIAM (the Society for Industrial and Applied Mathematics) to endorse our conference. One of the proposed invited speakers (Ilse Ipsen) was SIAM's Vice President at large and we will coordinate with her (if she accepts our invitation) towards that end.

- **Value to Purdue.**
  Beyond making Purdue even more visible in the area of Data Science, the proposed conference will announce to the whole campus the existence of a very strong cluster at Purdue's College of Science in Data Science. Given that many researchers at Purdue use Data Science tools in their work, we believe that the conference will help foster significant further collaborations within Purdue.

- **Value to external audiences and the media (e.g., newsworthiness or pioneering work).**
  We will coordinate with Ms. Kristyn R. Childres, the communications specialist of the Computer Science department, to prepare press releases and ensure media coverage of the conference. Much of the research that will be presented in the conference has significant media value and we are confident that we will be able to appropriately showcase the conference in news outlets.

- **Proposed post-event impacts (e.g., future partnerships, publications, follow-up events).**
  We expect that the conference will result in and help foster collaborations between researchers working in foundations of Data Science and more applied researchers, who primarily use data science tools to analyze their data. We do expect joint proposals and publications to be the eventual outcome of this conference.

- **Proposed extramural sponsorships (e.g., joint funding, conference scholarships).**
  We will explore whether industrial partners or national labs (e.g., Sandia) are interested in providing sponsorship for the proposed event.