NIH NATIONAL CANCER INSTITUTE | PURDUE UNIVERSITY CENTER FOR CANCER RESEARCH | INDIANA UNIVERSITY MELVIN AND BREN SIMON COMPREHENSIVE CANCER CENTER | BIG DATA GRAND CHALLENGE 2022

Big Data Training for Cancer Research

Special Lecture Series

Predictability, stability, and causality with a case study in biomedicine

Dr. Bin Yu

May 25, 2022, 1:00 – 2:30 PM (EDT)

Abstract:

"A.I. is like nuclear energy -- both promising and dangerous" -- Bill Gates, 2019.

Data Science is a pillar of A.I. and has driven most of recent cutting-edge discoveries in biomedical research and beyond. Human judgement calls are ubiquitous at every step of a data science life cycle, e.g., in choosing data and formulating the data science problem, choosing data cleaning methods, predictive algorithms and data perturbations. Such judgment calls are often responsible for the "dangers" of A.I. To maximally mitigate these dangers, we developed a framework based on three core principles: Predictability, Computability and Stability (PCS). The PCS framework (with PCS documentation) unifies and expands on the best practices of machine learning and statistics. It consists of a workflow and documentation and is supported by our software package v-flow.

In this talk, we first illustrate the PCS framework through the development of iterative random forests (iRF) for predictable and stable non-linear interaction discovery (in collaboration with the Brown Lab at LBNL and Berkeley Statistics). In pursuit of genetic drivers of a heart disease called hypertrophic cardiomyopathy (HCM) as a CZ Biohub project (in collaboration with the Ashley Lab at Stanford Medical School and others), we use iRF and UK Biobank data to recommend gene-gene interaction targets for knock-down experiments. We then analyze the experimental data to show promising findings about genetic drivers of HCM.

Speaker Bio: Bin Yu is Chancellor's Distinguished Professor and Class of 1936 Second Chair in the Departments of Statistics and EECS, and Center for Computational Biology at UC Berkeley. She has published more than 170 publications in premier venues and these papers not only investigate a wide range of research topics from practice to algorithms and to theory, but also seek deep insights. The breadth and depth of her research experience enabled unique and novel solutions to interdisciplinary data problems in audio and image compression, network tomography, remote sensing, neuroscience, genomics, and precision medicine.

She champions collaborative research with experts in the subject knowledge and leads research in statistical machine learning (e.g. boosting, sparse modeling, kernel methods, and spectral clustering, decision trees, and deep learning) and causal inference to design algorithms such as iterative random forests (iRF), agglomerative contextual decomposition (ACD) and adaptive wavelet distillation (AWD) for interpreting deep neural networks and X-learner for heterogeneous treatment effect estimation in causal inference.

She is a Member of the U.S. National Academy of Sciences and of the American Academy of Arts and Sciences. She is Past President of the Institute of Mathematical Statistics (IMS), Guggenheim Fellow, Tukey Memorial Lecturer of the Bernoulli Society, Rietz Lecturer of IMS, and a COPSS E. L. Scott Prize winner. She has been selected to deliver the Wald Memorial Lectures of IMS at JSM in 2023. She recently served on the scientific advisory committee for the IAS Special Year on optimization, statistics and theoretical machine learning, and the inaugural scientific advisory committees of the UK Turing Institute for Data Science and AI. She is serving on the editorial board of PNAS, the Scientific Advisory Boards of Canadian Statistical Sciences Institute (CANSSI), of the AI Policy Hub at UC Berkeley, and the Scientific Advisory Committee of the Department of Quantitative and Computational Biology at USC.

Register at: https://purdue-edu.zoom.us/webinar/register/WN_ZXBAz2R3QzO53ZyTQkpg1w