Veridical Data Science and PCS Uncertainty Quantification

Bin Yu

(UC Berkeley)

11/06/2025, 4:15 pm

Department of Statistics, Ludwigstr. 33, Seminar Room 144,
and online via Zoom [Link]
(Meeting-ID: 631 1190 7291; Password: StatsCol)

Abstract

Data Science is central to Al and has driven most of the recent advances in biomedicine and beyond. Human judgment calls are ubiquitous at every step of the data science life cycle (DSLC): problem formulation, data cleaning, EDA, modeling, and reporting. Such judgment calls are often responsible for the "dangers" of Al by creating a universe of hidden uncertainties well beyond sample-to-sample uncertainty. To mitigate these dangers, veridical (truthful) data science is introduced based on three key principles: Predictability, Computability and Stability (PCS). The PCS framework (including documentation) unifies, streamlines, and expands on the ideas and best practices of statistics and machine learning. This talk showcases PCS through a collaborative project on finding genetic drivers of a heart disease and through uncertainty quantification (PCS-UQ) with applications to prediction including deep learning. It compares PCS-UQ and makes connections with Conformal Prediction (CP). Over a collection of 17 regression tabular datasets, 6 multi-class tabular datasets, and 4 deep learning datasets, PCS-UQ reduces the size of the prediction intervals or sets by around 20% on average when compared to the best CP method among the ones used by PCS-UQ, and has better subgroup coverages than CP overall.

About the Speaker:
Bin Yu is CDSS Chancellor's Distinguished Professor in Statistics, EECS, and Computational Biology, and Scientific Advisor at the Simons Institute for the Theory of Computing, all at UC Berkeley. Her research focuses on the practice and theory of statistical machine learning, veridical data science, responsible and safe AI, and solving interdisciplinary data problems in neuroscience, genomics, and precision medicine. She and her team have developed algorithms such as iterative random forests (iRF), stability-driven NMF, and adaptive wavelet distillation (AWD) from deep learning models. She is a member of the National Academy of Sciences and of the American Academy of Arts and Sciences. She holds an Honorary Doctorate from The University of Lausanne.