University of California, Irvine
Statistics Department
Distinguished Seminar

Data Perturbation

Xiaotong Shen
John Black Johnston Distinguished Professor
School of Statistics
University of Minnesota

4 p.m., Thursday, Oct. 27, 2022
6011 Donald Bren Hall

Data perturbation is a technique for generating synthetic data by adding ``noise" to original data, which has a wide range of applications, primarily in data security. Yet, it has not received much attention within data science. In this presentation, I will describe a fundamental principle of data perturbation that preserves the distributional information, thus ascertaining the validity of the downstream analysis and a machine learning task while protecting data privacy. Applying this principle, we derive a scheme to allow a user to perturb data nonlinearly while meeting the requirements of differential privacy and statistical analysis. It yields credible statistical analysis and high predictive accuracy of a machine learning task. Finally, I will highlight multiple facets of data perturbation through examples.

This work is joint with B Xuan and R Shen.