DEPARTMENT OF STATISTICS

COLLOQUIA SERIES

Monday October 17th, Talk: 4:15 PM — Science Center 300H
Reception: 3:50 PM

Guest Speaker: Xiao-Li Meng
Dean of the Graduate School of Arts and Sciences
Whipple V. N. Jones Professor
Department of Statistics
Harvard

Statistical Paradises and Paradoxes in Big Data

Abstract:

Statisticians are increasingly posed seemingly paradoxical questions, challenging our qualifications for entering the statistical paradises created by Big Data. Two such questions represent the use of Big Data for population inferences and individualized predictions: (1) "Which one should I trust: a 1% survey with 60% response rate or a self-reported administrative dataset covering 80% of the population?" and (2) "Personalized treatments -- that sounds heavenly, but where on earth did they find the right guinea pig for *me*?" Investigating the first question reveals a *Big Data Paradox*: the bigger the data, the more certain we will miss our target. We need *data-quality indexes*, not merely quantitative sizes, to answer the question: a seemingly tiny self-reporting bias will make an on-line database with 160,000,000 entries equivalent to a simple random sample of 400 for estimating population averages. The second question is fundamentally yoked to the familiar Simpson's Paradox: how do we ensure that the level of aggregation (i.e., data resolution) does not alter our (treatment) conclusions? A *multi-resolution framework*, inspired by wavelets, provides a theoretical platform for studying statistical evidence for predicting individual outcomes. In contrast to the first question, where the goal is to infer population quantities from samples, with the second question we seek a *primary inference resolution*, that is, a sensible bias-variance tradeoff to form populations for approximating individuals. Theoretical links between optimal resolution and sparsity will be discussed.