Title: STATISTICAL DATA COMPRESSION WITH DISTORTION

Speaker: Mokshay Madiman, Department of Statistics, Yale University

Abstract:

Motivated by the powerful and fruitful connection between information-theoretic ideas and statistical model selection, we consider the problem of "lossy" data compression ("lossy" meaning that a certain amount of distortion is allowed in the decompressed data) as a statistical problem. After recalling the classical information-theoretic development
of Rissanen's celebrated Minimum Description Length (MDL) principle for model selection, we introduce and develop a new theoretical framework for _code selection_ in data compression. First we describe a precise correspondence between compression algorithms (or codes) and probability distributions, and use it to interpret arbitrary families of codes as statistical models. We then introduce "lossy" versions of several familiar statistical notions (such as maximum likelihood estimation and MDL model selection criteria), and we propose new principles for building good codes. Specifically, we show that in particular cases, our "lossy MDL estimator'" has the following optimality property: Not only it converges to the best available code (as the amount of data grows), but it also identifies the right class of codes in finite time with probability one. In contrast, the "lossy maximum likelihood estimator" we define fluctuates outside the right class of codes infinitely often.

[Joint work with Ioannis Kontoyiannis and Matthew Harrison.]