Stanford University
Department of Statistics

DEPARTMENTAL SEMINAR

4:30pm, Tuesday, June 6, 2017
Sloan Mathematics Center Room 380X
Refreshments served at 4pm in Sequoia Lounge.

Speaker: Jonathan Taylor
Department of Statistics,
Stanford University

Title: Interactive Data Analysis

Abstract:
We describe interactive data analysis, so-named to denote an interactive approach to data analysis with an emphasis on inference after data analysis. Our approach is a compromise between Tukey’s exploratory (roughly speaking “model free”) and confirmatory data analysis (roughly speaking classical and “model based”), also allowing for Bayesian data analysis. We view this approach as close in spirit to current practice of applied statisticians and data scientists while allowing frequentist guarantees for results to be reported in the scientific literature, or Bayesian results where the data scientist may choose the statistical model (and hence the prior) after some initial exploratory analysis. While this approach to data analysis does not cover every scenario, and every possible algorithm data scientists may use, we see this as a useful step in concrete providing tools (with frequentist statistical guarantees) for current data scientists.

The basis of inference we use is selective inference, in particular, its randomized form. The randomized framework, besides providing additional power and shorter confidence intervals, also provides explicit forms for relevant reference distributions (up to normalization) through the *selective sampler*. The reference distributions are constructed from a particular conditional distribution formed from what we call a DAG-DAG, a Data Analysis Generative DAG. As sampling conditional distributions in DAGs is generally complex, the selective sampler is crucial to any practical implementation of interactive data analysis.