Mathematics, Statistics, and Computer Science **@ UIC**

Departmental Colloquium

Data perturbation for data science

Richard Samworth (University of Cambridge)

Abstract: When faced with a dataset and a statistical problem of interest, should we propose a statistical model and use that to inform an appropriate algorithm, or dream up a potential algorithm and then seek to justify it? The former is the more traditional statistical approach, but the latter appears to be becoming more popular. I will present an example of a 20th century analysis that falls into the first category, and explain why it may not be as suitable for modern statistical challenges. I'll then discuss a class of algorithms that belong in the second category, namely those that involve data perturbation (e.g. subsampling, random projections, artificial noise, knockoffs,...). As an illustration, I will consider Complementary Pairs Stability Selection for variable selection.

Friday, April 26 at 3:00 PM in 636 SEO