NYU Stern School of Business
Department of Information, Operations & Management Sciences
STATISTICS RESEARCH SEMINAR

TOPIC: **An introspection on using sparse regression techniques to analyze text**
SPEAKER: **Luke Miratrix (Harvard University)**
DATE: **Friday, April 11, 2014**
TIME: **11:30 AM - 12:30 PM**
PLACE: **KMC 5-75**

Abstract

In this talk, I propose a general framework for topic-specific summarization of large text corpora, and illustrate how it can be used for analysis in two quite different contexts: legal decisions on workers' compensation claims (to understand relevant case law) and an OSHA database of occupation-related accident reports (to search for high risk circumstances). Our summarization framework, built on sparse classification methods, is a lightweight and flexible tool that offers a compromise between simple word frequency based methods currently in wide use, and more heavyweight, model-intensive methods such as Latent Dirichlet Allocation (LDA). For a particular topic of interest (e.g., emotional disability, or chemical gas), we automatically labels documents as being either on- or off-topic, and then use sparse classification methods to predict these labels with the high-dimensional counts of all the other words and phrases in the documents. The resulting small set of phrases found as predictive are then harvested as the summary. Using a branch-and-bound approach, this method can be extended to allow for phrases of arbitrary length, which allows for potentially rich summarization. I further discuss how focus on specific aspects of the corpus and the purpose of the summaries can inform choices of regularization parameters and constraints on the model. I briefly present a new R package, ngram, that implements the methods discussed. Overall, I argue that sparse methods have much to offer text analysis, and hope that this work opens the door for a new branch of research in this important field.