

**STANFORD UNIVERSITY**  
**DEPARTMENT OF STATISTICS**  
**DEPARTMENTAL SEMINAR**

4:15 p.m., Tuesday, November 27, 2007  
Sequoia Hall Room 200  
(Cookies at 3:45 in 1st Floor Lounge)

*Regina Liu*  
Department of Statistics  
Rutgers University

**Mining Massive Text Data: Classification, Construction of Tracking Statistics  
and Inference under Misclassification**

We present a systematic data mining procedure for exploring large free-style text datasets to discover useful features and develop tracking statistics (often referred to as performance measures or risk indicators). The procedure includes text classification, construction of tracking statistics, inference under error measurements and risk management. The main difficulty in deriving this inference scheme is the accounting for misclassification errors, for which we propose two types of approaches: "plug-in" and "projection" methods. We also consider the bootstrap calibration for fine tuning. Finally, as an illustrative example, the proposed data mining procedure is applied to analyzing an FAA aviation safety report repository to show its utility in aviation risk management or general decision-support systems.

Although most illustrations here are drawn from aviation safety data, the proposed data mining procedure applies to many other domains, including, for example, mining free-style medical reports for tracking medical errors or possible disease outbreaks.

This is joint work with Daniel Jeske, Department of Statistics, UC Riverside.