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

# Special Colloquium

*Too much data!*

John Stufken (Arizona State University)

**Abstract:** The enormous amounts of data that are collected in applications in a wide variety of fields create challenges and opportunities for statisticians. One of the challenges is that traditional statistical methods for data of smaller size may no longer be applicable in the new "big data" environment, for computational reasons or otherwise. The corresponding opportunity lies in the need to develop methods that are applicable for big data. The simplest such methods, and often the most elegant ones, are based on innovations that allow familiar techniques to be applied in this new environment in a computationally feasible way. Adapting existing methods for this new environment can typically not be accomplished by putting "old wine in new bottles", but requires clever innovations.

Traditionally, it goes against a statistician's core principles to "discard" some of the data. Yet, some data sets are so large that exploration and analysis must proceed by using only some of the data. This leads to the idea of selecting subdata from big data and drawing conclusions from an analysis of the subdata. While this idea brings traditional statistical analysis methods potentially back into the picture, there are the immediate questions of how to select the subdata and, if needed, how to adjust analysis methods. Innovations to accomplish this are the focus of this presentation. We discuss subdata selection methods, with special emphasis on information-based subdata selection, as well as challenges and shortcomings

Wednesday, January 23 at 3:00 PM in 636 SEO

associated with these methods.

Wednesday, January 23 at 3:00 PM in 636 SEO