Big Data Analytics: Program and Abstracts

Friday, October 18

7:00-9:00 PM Reception in Bailey Hall 204

Saturday, October 19

8:00-9:00 Registration and Coffee in Bailey Hall 204

9:00-9:30 Richard Messmer (GE Global Research) *Big Data: Data Driven vs Domain Driven Analytics*

9:40-10:10 Jin Xia (GE Global Research), *Divide and Recombine (D&R): An Approach for Statistical Modeling of Big and Complex Data*

10:20-10:50 Jeremy Kepner (MIT Computer Science and AI Laboratory), *The Abstract Algebra of Big Data*

10:50-11:15 Coffee in Bailey Hall 204

11:15-12:15 Gunnar Carlsson (Stanford University), *The Shape of Data*

12:15-2:00 Lunch Break

2:00-2:30 Bouchra Bouqata (GE Global Research), *Big Data, Analytics, and Cloud Computing; the New Paradigm*

2:40-3:10 Mark Embrechts (Industrial and Systems Engineering, RPI), *Big Data: Hype or Haleluja*

3:30-4:30 Ron Snee (Temple University and Snee Associcates), *Enhancing Big Data Projects through Statistical Engineering*

4:30-5:00 Coffee in Bailey Hall 204

6:30 Banquet in Old Chapel

Featured Speakers

Ronald Snee, Temple University and Snee Associcates
Enhancing Big Data Projects through Statistical Engineering

Massive data sets or Big Data have become more common recently, due to improved technology for data acquisition, storage, and processing of data. New tools have been developed to analyze such data, including classification and regression trees (CART), neural nets, and methods based on bootstrapping. These tools make high-powered statistical methods available to not only professional statisticians, but also to casual users. As with any tools, the results to be expected are proportional to the knowledge and skill of the user, as well as the quality of the data. Unfortunately, much of the professional literature may give casual users the impression that if one has powerful enough algorithms and a lot of data, good models and good results are guaranteed at the push of a button.

This presentation focuses on the application of principles of statistical engineering (Anderson-Cook and Lu, Quality Engineering, 2012) to the Big Data problem. Viewed through the statistical engineering lens several potential pitfalls of commonly used approaches to Big Data projects become apparent. The consequences of four major issues are addressed: 1) Lack of a disciplined approach to modeling, 2) Use of "one shot studies" versus sequential approaches, 3) Assuming all data are high quality data, and 4) Ignoring subject matter knowledge.

Gunnar Carlsson, Stanford University
The Shape of Data

In recent years, a recognition has been developing that notions of shape can serve as a valuable paradigm for organizing and understanding large and complex data sets. Topology, which is the mathematical discipline which concerns itself with the study of shape, can therefore be adapted to the study of such data sets. We will talk about a number of such methods, with examples.

Contributed Talks

Bouchra Bouqata, GE Global Research
Big Data, Analytics, and Cloud Computing; the New Paradigm

Today, it has been estimated that data is created at a rate of 2.5 quintillion bytes/day. We are living an explosive growth of data that comes from everywhere: sensors, social media, images and video, transaction records, etc. The ubiquitous availability of this digital information has created a paradigm shift from information-poor to information-rich, and impacted our modern life. This talk's focus will be on discussing data-driven innovation through new technologies and infrastructure around Big Data and Analytics. Furthermore, I will give an overview of representative applications of big data analytics, their deployment solutions, current challenges, and open research problems.

Mark Embrechts, Industrial and Systems Engineering, RPI
Big Data: Hype or Haleluja

The phenomena of Big Data and Analytics bring a new life to the discipline of data mining. This talk will define and trace the origin of big data and answer the where, when, and why of Big Data Analytics. Big Data is more than data mining on steroids. The vast amount of data mandates novel algorithmic approaches to Big Data Analytics. But there is more to come: Big Data often has a significant crowdsourcing aspect and now places a heavy emphasis on data cleansing and outlier detection. Because of the nature of the data (often text and images) the first emphasis for Big Data Analytics is now on structuring the data.

This talk will highlight the differences between data mining and Big Data Analytics and why an engineering approach is necessary for data-driven science and engineering applications of Big Data Analytics.

Jeremy Kepner, MIT Computer Science and AI Laboratory
The Abstract Algebra of Big Data

Spreadsheets, databases, hash tables and dictionaries; these are the fundamental building blocks of big data storage, retrieval and processing. They are ubiquitous: triple store databases are the backbone of tech companies such as Amazon (Dynamo) and Google (Big Table). These databases also play a prominent role in other industries that utilize big data such as healthcare, finance and network security. Likewise, spreadsheets are used by $\sim 100,000,000$ people each day. D4M (Dynamic Distributed Dimensional Data Model) is an interface to the triple store databases and spreadsheets that allow developers to write analytics in the language of linear algebra, significantly reducing the time and effort required. The core data structure of D4M is the associative array. In practice, associative arrays allow big data to be represented as large sparse matrices. This sparse matrix D4M schema has been widely adopted in a number of domains. We explore the mathematics behind the D4M system by formulating an axiomatic definition for associative arrays in terms of semi-modules over semi-rings, then exploring the algebraic properties of the latter. We are primarily concerned with the linear systems theory and spectral theory of a three particular families of semi-rings. We present a structure theorem for the solutions of linear systems of an important family of semi-rings.

Richard Messmer, GE Global Research

Big Data: Data-Driven vs Domain-Driven Analytics

BIG DATA challenges have always existed in the computational/analytics world. A Big Data challenge exists at any point in time where the largest data sets exceed the then current capabilities and require technology innovations to make progress. Big Data is the latest name for this recurring challenge but it is also somewhat different now because the rate of growth of data sizes puts a strain on the innovation cycle necessary to keep up with the increase in data sizes.

As in any rapidly developing technology area, there is sometimes rather more hype than is needed and potentially dead-ends might be encountered.

In the talk, we discuss approaches to big data analytics from the open literature, where data-driven analytics are very effective and some cases where they are not.

Also the more traditional domain-driven analytics (that depend on domain knowledge) can have their limitations as well. There have been various vocal proponents on each side of these discussions but history tells us that a hybrid approach will likely evolve. A brief discussion of some hybrid candidates that have been proposed will be presented.

Jin Xia, Applied Statistics Lab, GE Global Research

Divide and Recombine (D&R): An Approach for Statistical Modeling of Big and Complex Data

With the advance of technology in recent years, large amounts of data are collected ubiquitously in academia, industry and government. They often contain complex information about machines, human behaviors, biological experiments, etc. These large and complex data sets create substantial challenges to computing and modeling in data analysis. In this talk, we aim to discuss what challenges big data bring to statistical analysis, approach computation through an R and Hadoop combined platform, and propose a statistical modeling framework called Divide and Recombine (D&R). R and Hadoop are combined to form a distributed computing environment for statistical computing through R package RHIPE (www.rhipe.org). Based on the highly scalable distributed computing environment provided by RHIPE, we can divide the data into subsets, model subsets in parallel, and recombine subset models to achieve statistical modeling on big and complex data sets.