

**Title: “Explanatory vs. Predictive Modeling in Scientific Research”**

**Speaker:** Galit Shmueli / University of Maryland

**Abstract:** Explanatory models are designed for testing hypotheses that specify how and why certain empirical phenomena occur. Predictive models are aimed at predicting new observations with high accuracy. An age-old debate in philosophy of science deals with the difference between predictive and explanatory goals. In mainstream statistical research, however, the distinction between explanatory and predictive modeling is mostly overlooked, and there is a near-exclusive focus on explanatory methodology. This focus has permeated into empirical research in many fields such as information systems, economics and in general, the social sciences. We investigate the issue from a statistical modeling perspective. Our premise is that (1) both explanatory and predictive statistical models are essential for advancing scientific research; and (2) the different goals lead to key differences at each step of the modeling process. In this talk we discuss the statistical divergences between modeling for an explanatory goal and modeling for a predictive goal. In particular, we analyze each step of the statistical modeling process (from data collection to model use) and describe the different statistical components and issues that arise in explanatory modeling vs. predictive modeling. We close with a discussion of implications of this work to the general scientific community and to the field of statistics.

\*\*\*\*\*

**Title: “ARO Probability and Statistics Program”**

**Speaker:** Dr. Harry Chang, U.S. Army Research Office

**Abstract:** In this 30-minute talk, we will discuss the vision, philosophy/strategy, current research thrust areas, and dynamics of the ARO Probability and Statistics Program. A sample of current statistics projects supported by ARO will also be discussed briefly.

\*\*\*\*\*

**Title: “Data Mining with Decision Trees”**

**Speaker:** Wei-Yin Loh, University of Wisconsin, Madison  
(<http://www.stat.wisc.edu/~loh/guide.html>)

**Abstract:** Decision trees are particularly suited for data mining applications, because they produce intuitively interpretable prediction models. This talk will briefly introduce the GUIDE algorithm, which quickly and automatically constructs accurate decision tree models for classification or regression problems. An application to the analysis of some Army courts martial data will be used for illustration.

\*\*\*\*\*

**Title: “Reliability Case Study”**

**Speaker:** Alyson Wilson

**Abstract:** In this case study, I am developing methodology to assign an uncertainty to a reliability estimate, predict reliability into the future, and use these estimates to allocate

resources for future testing. This talk discusses the diverse data available for the assessment, with particular attention to multilevel data. Multilevel data occurs when data is simultaneously available at both the component and subsystem level. In addition, I outline the approach for addressing the resource allocation problem when there are many choices for potential data sources.

\*\*\*\*\*

**Title: "I could really do great statistics if my data wasn't so lousy"**

**Speaker:** Dr. Dave Olwell, NPS

**Abstract:** I am going to discuss lessons learned from looking at OIF combat data, Navy missile data, and Army CROWS data. The common theme: if you don't design and include a robust data collection system when fielding a system, it is almost impossible to do meaningful analysis later. Statisticians must insist upon participating in the systems engineering.

\*\*\*\*\*

**Title: "An Experimental Investigation of Robotic Technology"**

**Speaker:** Barry A. Bodt, U.S. Army Research Laboratory

**Abstract:** In keeping with the theme of this workshop, I will discuss a current program seeking to develop pedestrian detection algorithms for use on unmanned ground vehicles. The focus is on a fall 2007 experiment. In concept, the factorial design was simple enough. The role of ANOVA was clear – but conducting the actual experiment and performing related analysis was not as simple or clear. I will describe the purpose, planning and conduct of the experiment and analysis. Throughout, I will attempt to tap into some underlying themes emphasized by statisticians that have come before us, for example, the need to be immersed in context, the need for evolutionary experimentation, and the vital contribution statisticians can make in scientific investigation beyond confirmatory statistics.

\*\*\*\*\*

**Title: "Assessing the Effects of Individual Augmentation on the Retention of Navy Personnel"**

**Speaker:** Dr. Ron Fricker (NPS)

**Abstract:** In this talk we will describe an effort to assess the effects of Individual Augmentation (IA) on the retention of Navy enlisted and officer personnel. IAs are a unique form of deployment for Navy personnel in which personnel are individually deployed, often in direct support of OIF, to augment deployed Army or Marine Corps units. The talk will describe the data and models we used, focusing on their strengths and weaknesses, as well as the various challenges we faced, and then discuss our findings.

\*\*\*\*\*

**Title: "Drowning in Data, Starved for Knowledge: Survival in the World of the Future"**

**Speaker:** Dr. Ben Cole (NSA)

**Abstract:** In this talk we will address a profound problem that is already upon us as the Network comes of age, giving everyone access to more information than they possibly comprehend. Search engines don't assign real "meaning" to all they access. Where will knowledge arise in world with information exploding at an exponential rate? We'll explore some possible answers and point out some difficult challenges.

\*\*\*\*\*

**Title: "Recommender Systems for Fun and Profit"**

**Speaker:** Dr. Chris Volinsky (AT&T)

**Abstract:** In October 2006, Netflix released 100 million movie ratings as part of a \$1M prize for any team that could improve their movie recommendation system by more than 10%. This landmark data set generated intense interest from the statistics and machine learning communities, and attracted entries from over 3000 teams from academia and industry. In this talk, I will review AT&T's experience analyzing this data using collaborative filtering techniques, leading to our winning a \$50,000 progress prize, as well as a subsequent project applying the methodology to television viewing data.

\*\*\*\*\*

**Title: "Statistical Challenges with Massive Data Sets"**

**Speaker:** Dr. Karen Kafadar (Rudy Professor of Statistics and Physics Indiana University)

**Abstract:** Historically, advances in statistical research have arisen out of a need to analyze new data types; e.g., sequential analysis (interim monitoring of clinical trials), design of highly-fractionated experiments (many factors with limited numbers of animals), nonparametrics (non-Gaussian data). Today's challenges include identification of "interesting" objects from a population containing millions of candidates; e.g., selection of most promising chemical compounds for further experiments; identification of "unusual" patterns in travel routes or process operations; etc.

Massive data require new approaches to experimental design, analysis, and inference. We review some successes in the visualization and analysis of large data sets, with particular attention to considerations of multiple testing, robustness, and computation, and describe these challenges for data sets on Internet traffic and in high energy physics experiments.