Interface 2011 Program Book

Symposium on the Interface of Computing Science and Statistics

Theme: Statistical, Machine Learning,
and Data Visualization Algorithms

Program Book
Schedule for Interface 2011

Wednesday, June 1, 2011

7:30 am – 5:00 pm **Registration** ***Bldg. C Lobby***

8:00 am – 9:45 am

Welcoming: **John Sall, Edward Wegman**

Keynote Address:

21st Century Analysis, Biomedical, Collaboration and Determinism Dogma: Emerging Challenges and Guidance for Interface of Computing and Statistics ***Auditorium***
Arnold Goodman, Collaborative Data Solutions

9:45 am – 10:15 am ***Morning Break*** **Prefunction Hall**

10:15 am – 12:00 noon ***Technical Sessions***

Statistical Analysis of Attributed Graphs and Networks. ***Room 1012***
Organizer: David Marchette, Naval Surface Warfare Center, Dahlgren Division

1. ***Tweets to Hypergraphs*****, Elizabeth Hohman, Naval Surface Warfare Center, Dahlgren Division**
2. ***Vertex Nomination: Finding Similar Vertices in Communications Graphs Attributed with Human Language*****, Glen Coppersmith, Johns Hopkins University**
3. ***Combining Geographic Information in Random Dot Product Graphs,*** **David Marchette, Naval Surface Warfare Center, Dahlgren Division**

Predictive Modeling**. Organizer: Georgiy Bobashev,** ***Room 1014***
Research Triangle Institute

1. ***Using Random Forest Models to Predict Organizational Violence*****, Burton Levine, Research Triangle Institute**
2. ***Patterns of Daily Alcohol Consumption*****, Dan Liao, Research Triangle Institute**
3. ***Predictive Modeling Processes and Heavy Alcohol Use*****, Barry Eggleston, Research Triangle Institute**
4. ***Using Tree-Based Prediction Models for Missing Data in the Complex Survey Setting*****, Daryl Creel, Research Triangle Institute**

12:00 noon – 1:30 pm **Lunch Break** ***Bldg. C Café***

1:30 pm – 3:15 pm **Technical Sessions**

Case Studies in Visualization. ***Room 1012***
Organizer: David W. Scott, Rice University

1. ***Response Surface Methodology*, John Sall, JMP/SAS Institute**
2. ***Exact and Approximate Area-proportional Circular Venn and Euler Diagrams*, Leland Wilkinson, SYSTAT Software, Inc.**
3. ***Two Micromap Extensions: 3-Way Hypothesis Test Conditioned Micromaps and More Options for Dynamic Comparative Micromaps*, Daniel Carr and Chunling Zhang, George Mason University**

Tree Based Machine Learning. ***Room 1014***
Organizer: Adele Cutler, Utah State University

1. ***Random Forests Variable Selection*, Hemant Ishwaran, Cleveland Clinic**
2. ***Finding Structural Variants in Individual Human Genomes with Random Forests*, Jacob Michaelson, Psychiatric Genomics - Sebat Group and University of California, San Diego**
3. ***Probability Machines*, James D. Malley, Center for Information Technology, National Institutes of Health**

3:15 pm – 3:45 pm **Afternoon Break** ***Prefunction Hall***

3:45 pm – 5:30 pm **Technical Sessions**

A Few Challenges on the Road toward Robot-Assisted ISR. ***Room 1012***
Organizer: Barry Bodt, Army Research Laboratory

1. ***Cross-country Autonomous Navigation: Trade-offs and Challenges*, Juan Pablo Gonzalez, General Dynamics Robotic Systems**
2. ***Using Hybrid Maps to Understand Space and Promote Common Ground*, MaryAnne Fields and Jason Owens, U.S. Army Research Laboratory and Kostas Daniilidis, University of Pennsylvania**
3. ***Similarity-Based Information Fusion*, Eric Heilman,Timothy P. Hanratty, U.S. Army Research Laboratory**

Reproducible Research. ***Room 1014***
Organizer: Jürgen Symanzik, Utah State University

1. ***Approaches and Barriers to Reproducible Practices in Biostatistics*, Matt Shotwell, Vanderbilt University**
2. ***Practical Implementation of Reproducible Research in the Madagascar Software Project*, Sergey Fomel, University of Texas**
3. ***Archiving Computational Research in Virtual Machines*, Sorin Mitran, University of North Carolina**

5:30 pm – 7:30 pm Happy Hour at Embassy Suites

6:00 pm – 8:00 pm Board of Directors Meeting and Dinner

7:30 pm – 9:30 pm Conference Mixer at Embassy Suites

Thursday, June 2, 2011

7:30 am – 5:00 pm Registration ***Bldg. C Lobby***

8:00 am – 9:45 am Technical Sessions

Three Approaches to Large p Inference*.** ***Room 1012
Organizer: David Banks, Duke University

1. ***Spatial Varying Coefficient Models for Neuroimaging Data*, Hongtu Zhu, University of North Carolina, Chapel Hill**
2. ***An Ordinary Differential Equation-Based Solution Path Algorithm*, Yichao Wu, North Carolina State University**
3. ***Cherry-Picking in Cluster Analysis and Regression*, David Banks, Duke University**

Business Knowledge and Networks in SAS Data Mining. ***Room 1014***
Organizer: David Duling, SAS Institute

1. ***Combining Knowledge-based and Statistical Approaches to Extract Sentiment from Text*, Russell Albright, SAS Institute**
2. ***GLM models for Insurance Rate Making*, Billie Anderson, SAS Institute**
3. ***Network Visualizations in Common Data Mining Analysis*, David Duling, SAS Institute**

9:45 am – 10:15 am Morning Break ***Prefunction Hall***

10:15 am – 12:00 noon Technical Sessions

Visual Inference and Uncertainty. ***Room 1012***
Organizer: Roy Welsch, MIT

1. ***Data Visualization by Navigating High Dimensional Space with Low Dimensional Trajectories*, Wayne Oldford, University of Waterloo (Canada)**
2. ***Visualizing the Variability of Plots*, Rajiv S. Menjoge, Google and Roy E. Welsch, MIT**
3. ***Visual Statistical Inference for Regression Parameters*, Mahbubul Majumder, Heike Hofmann, and Dianne Cook, Iowa State University**

Statistical Issues in Weather and Climate Research. ***Room 1014***
Organizer: Richard Smith, SAMSI and UNC, Chapel Hill

1. ***Nonparametric Spatial Models for Extreme Temperature Data*, Montse Fuentes, North Carolina State University**
2. ***Space-time Modeling of Climatic Trends*, Peter Craigmile, Ohio State University**
3. ***The Reliability of Millennial Multi-proxy Temperature Reconstructions*, Blake McShane, Northwestern University**

12:00 noon – 1:30 pm **Lunch Break** ***Bldg. C Café***

1:30 pm – 3:15 pm **Technical Sessions**

Celebrating the 20th Anniversary of JCGS: ***Room 1012***
Highlights at the Interface.
Organizer: Richard Levine, San Diego State University

1. ***Cross-section and Longitudinal Penalized Functional Regression*, Jeff Goldsmith, Johns Hopkins University**
2. ***Dissimilarity Plots: A Visual Exploration Tool for Partitional Clustering*, Michael Hahsler, Southern Methodist University**
3. ***Counting Contingency Tables via Multistage Markov Chain Monte Carlo*, George Fishman, University of North Carolina, Chapel Hill.**

Maps, Nets, Optimization and Profiling: ***Room 1014***
Data Visualization, Modeling and Interpretation in JMP.
Organizer: Bradley Jones, JMP/SAS Institute

1. ***Using Maps as a Background for Displaying Information that is Geographically Distributed*, Xan Gregg, JMP**
2. ***Neural Nets – Boosted and Cross-Validated*, Christopher Gotwalt, JMP**
3. ***Interactive Graphics for Exploring Multi-Dimensional and Multivariate Functions*, Bradley Jones, JMP**

3:15 pm – 3:45 pm **Afternoon Break** ***Prefunction Hall***

3:45 pm – 5:30 pm **Technical Sessions**

Algorithms for Data and Information Visualization*.** ***Room 1012
Organizer: Adalbert Wilhelm, Jacobs University, Germany

1. ***3-d Stereoscopic Plots: From History to R*, Juergen Symanzik, Utah State University**
2. ***Interactive Visualization of Multiple Time Series,* Anushka Anand, National Center for Data Mining, University of Illinois at Chicago**
3. ***Singular Spectrum Analysis Algorithm for Decomposing and Visualizing Time Series Data*, Michael Leonard, SAS Institute Inc.**

Statistics of Function and Shape. ***Room 1014***

Organizer: Michael Schimek, Medical University of Graz, Austria

1. ***Functional Varying Coefficient Models*****, Damla Senturk, UCLA**
2. ***Smoothing Dynamic Positron Emission Tomography Time Courses Using Functional Principal Components*****, Ci-Ren Jiang – UC Davis**
3. **Detection of Functional Abnormalities in Brain Using Shape Analysis of Subcortical Structures, Sebastian Kurtek, Eric Klassen, Florida State University, Zhaohua Ding, Malcolm J. Avison, Vanderbilt University, Anuj Srivastava, Florida State University**

Computing and Statistics. ***Auditorium***

Organizer: Edward Wegman, George Mason University

1. ***Finite User Pool Effect in Two Sample t-test of Controlled Experiments on the Web*** **(refereed), Shaojie Deng, Microsoft**
2. ***Visualizing Association Rules in Hierarchical Groups*** **(refereed), Michael Hahsler and Sudheer Chelluboina, Souther Methodist University**
3. **The Architecture of Rc2, E. James Harner and Mark Lilback, West Virginia University**

5:30 pm – 7:30 pm **Happy Hour at Embassy Suites**

7:00 pm – 9:00 pm **Conference Banquet at Embassy Suites**

Friday, June 3, 2010

7:30 am – 9:00 am **Registration** ***Bldg. C Lobby***

8:00 am – 9:45 am **Technical Sessions**

Business Analytics and Algorithms in Data Mining. ***Room 1012***

Organizer: Simon Sheather. Texas A & M University

1. ***Computing for Robust Process Engineering*****, John Sall, Bradley Jones and Christopher Gotwalt, JMP/SAS Institute**
2. ***Two Entities and Beyond: Challenges and Examples in Social Network Analysis*****, Jin-Whan Jung, Dominic Jann, Dan Kelly and John Brocklebank, SAS Institute**
3. ***Choosing between Logistic Regression and Classification Trees in Data Mining*****, Mike Speed and Simon Sheather, Texas A&M University**

Best of Statistical Analysis and Data Mining. ***Room 1014***
Organizer: Joe Verducci, Ohio State University

1. ***Symbolic Data Analysis and Statistical Analysis and Data Mining*****, Lynne Billard, University of Georgia**
2. ***Learning Representations of Language for Greater Generalization Capacity*****, Alexander Yates, Temple University**
3. ***Coupling Optional Polya Trees – A Bayesian Nonparametric Approach to Case-Control Studies*****, Li Ma, Stanford University**

9:45 am – 10:15 am **Morning Break** ***Prefunction Hall***

10:15 am – 12:00 noon **Technical Sessions**

Best of Wiley Interdisciplinary Reviews: ***Room 1012***
Computational Statistics**.**
Organizer: Yasmin Said, George Mason University

1. ***Streaming Data,*** **William F. Szewczyk, NSA**
2. ***Record Linkage,*** **William E. Winkler, Census Bureau**
3. ***Some Interdisciplinary Topics on: i) Game-Theoretical Solutions for Quantitative Cyber-Risk Estimation/ Management ii) CLOUD Computing Implementations to Compute Operational Risk using Java Programming,*** **Mehmet Sahinoglu, Auburn University**

Data Mining Social Data. ***Room 1014***
Organizer: Rida E. A. Moustafa, dMining Technology

1. ***Analysis on Racial Rates and the Effective Elements of Home Mortgage Lending Acceptance in California*****, Chong Zhang, George Washington University**
2. ***Predicting and Explaining Potential Caravan Insurance Policy Ownership*****, Haiyun Zheng, George Washington University**
3. ***Red Zone, Blue Zone: Discovering Parking Ticket Trends in New York City*****, Samuel S. Ackerman, George Washington University**
4. ***Clustering TB Microarray Data for Potential Pathway Discovery and Disease Discrimination*****, Laura Tipton, George Washington University**

Abstract Book
Interface 2011

Red Zone, Blue Zone: Discovering Parking Ticket Trends in New York City
Samuel S. Ackerman and Rida E. Moustafa, George Washington University

Abstract: This paper reviews data on parking ticket violations in New York City in the month of March 2010 to reveal basic trends relating to the types of vehicles receiving tickets in various parts of the city. Parking tickets are a frequent nuisance in the daily lives of New York City residents, who must move their cars often to comply with the city's alternate side parking rules, and for the commercial vehicles that must park illegally to conduct deliveries. The characteristics of the various boroughs—Staten Island's mostly residential character versus heavily commercial Manhattan, for instance—can strongly influence the differing traffic patterns and thus the types of violations issued there; tickets for double parking, for example, are less common in Staten Island than in the bustle of Manhattan. Using a decision tree classifier based on attributes such as vehicle type, the violation, and fine amount, we can generally separate between Manhattan and non-Manhattan tickets.

Combining Knowledge-based and Statistical Approaches to Extract Sentiment from Subjective Text
Russell Albright, SAS Institute

Abstract: An important application of text mining is to automatically characterize the sentiment, whether it is positive, negative, or neither, of documents in a variety of domains. In this presentation we explore the benefits and disadvantages of using domain-specific linguistic rules, statistically-based methods, and a combination of both to identify the sentiment of documents.

Interactive Visualization of Multiple Time Series
Anushka Anand, National Center for Data Mining, University of Illinois at Chicago

Abstract: The recent prevalence of large volumes of time series data has generated an explosion of interest in mining this data and a flurry of information visualization tools for visual exploration. Most of these tools focus on particular interaction abilities, data domains or are limited by the size of data they can handle. We discuss a methodology for Exploratory Data Analysis (EDA) of large scale, massively concurrent time series data from any domain leveraging summarization and multi-scale views. We describe an approach of observing clusters over time and analyzing similar behavior in an interactive, general framework that facilitates guided exploration. We implement and demo this as an extension to the iPlots eXtreme package in R for interactive visualization of large data.

GLM models for Insurance Rate Making
Billie Anderson, SAS Institute

Abstract: Insurance companies gain competitive advantage by offering better rates and services to attract and retain the best customers. Generalized linear models (GLMs) have become popular and proven techniques for ratemaking and actuarial work over the past decade. Claim frequency is typically modeled using a Poisson distribution; severity is modeled using a Gamma distribution; and pure premium is modeled using a Tweedie distribution. Insurance analysts build these models on large data sets and require specialized transformations and reporting. In this presentation, we introduce the methods of GLM models for Insurance and demonstrate use of SAS® Enterprise Miner™.

Cherry-Picking in Cluster Analysis and Regression
David Banks, Duke University

Abstract: In complex, high-dimensional models one often sees superpositions of simpler structures. In some situations, these can be identified by mixture models, but in other cases, exploratory methods perform better. This talk describes a greedy structure extraction procedure that successively finds and removes structure in regression and cluster analysis.

Symbolic Data Analysis and *Statistical Analysis and Data Mining*
Lynne Billard, University of Georgia

Abstract: In an earlier edition of *Statistical Analysis and Data Mining* (*SAM*), Goodman (2010) suggested that one of the emerging fields in our discipline was symbolic data and their analysis (SDA). Recently, the March 2011 Issue of *SAM* was dedicated to this area. In this talk, we describe SDA and review some of the contributions presented by papers appearing in this Issue. This includes ideas of principal component analysis for missing observations and for intervals. There is an article breaking new ground by subdividing modal multi-valued data and histogram data into quantiles which can then be used to find principal components for such data. Other papers deal with clustering for observations that are discrete distributions; and with forecasting for histogram time series values.

Two Micromap Extensions: 3-Way Hypothesis Test Conditioned Micromaps and More Options for Dynamic Comparative Micromaps.
Daniel B. Carr and Chunling Zhang, George Mason University

Abstract: Micromaps are graphics that link statistical information to an organized set of small maps in order to explore and communication patterns related to associations among variables, spatial indices and sometimes temporal indices. This scope leaves much room for extending the existing micromap designs and implementations. The National Cancer Institute has used hypothesis tests about US state status and trends to partition states into a 3 x 3 table. This effective table design was previously re-expressed as a two-way conditioned micromap that reveals spatial patterns more readily. When a third set of hypothesis tests produces a 3-class categorical variable for the states, the color of the states can show the class memberships as will be illustrate in several examples. On a second micromap front, implementation has continued on Java software called TCmaps for temporal change micromaps. TCmaps includes one change blindness addressing comparative design that have not been previously shown and as well as other variations previous designs.

Vertex Nomination: Finding Similar Vertices in Communications Graphs Attributed with Human Language.
Glen Coppersmith, Johns Hopkins University

Abstract: If I know of a few persons of interest, how can human language technology and graph theory help me find other people similarly interesting? If I know of a few people committing a crime, how can I determine their co-conspirators? Given a set of actors deemed interesting, we seek other actors who are similarly interesting. We use a collection of communications encoded as an attributed graph, where vertices represents actors and edges connect pairs of actors that communicate. Attached to each edge is the set of documents wherein that pair of actors communicate, providing content in context – the communication topic in the context of who communicates with whom. In these documents, our identified interesting actors communicate amongst each other and with other actors whose interestingness is unknown. Our objective is to nominate the most likely interesting vertex from all vertices with unknown interestingness. As an illustrative example, the Enron email corpus consists of communications between actors, some of which are committed fraud. Some of their fraudulent activity is captured in emails, along with many innocuous emails (both between the fraudsters and between the other employees of Enron). We are given the identities of a few fraudster vertices and asked to nominate one other vertex in the graph as likely representing another actor committing fraud. This talk will exposit some of the foundational theory and initial experimental results indicating that approaching this task with a joint model of content and context improves the performance (as measured by standard information retrieval measures) over either content or context alone.

Space-time Modelling of Climatic Trends
Peter F. Craigmile, The Ohio State University

Abstract: Classical assessments of climatic trends are based on the analysis of a small number of time series. Considering trend to be only smooth changes of the mean value of a stochastic process through time is limiting, because it does not provide a mechanism to study changes of the mean that could also occur over space. Thus, in studies of climate there is a substantial interest in being able to jointly characterize trends over time and space. In this talk we discuss the salient features of climate data that must incorporated in statistical models that

characterize trend. We build wavelet-based space-time hierarchical Bayesian models that can be used to simultaneously model trend, seasonality, and error, allowing for the possibility that the error process may exhibit space-time long-range dependence. We demonstrate how these statistical models can be used to assess the significance of trend over time and space. We motivate and apply our methods to the analysis of space-time temperature trends.

Using Tree-Based Prediction Models for Missing Data in the Complex Survey Setting
Darryl Creel, RTI

Abstract: In the survey data context, the use of tree-based prediction models is slightly different than the traditional use of these models. If we define the traditional approach as producing a reasonable value for an observation missing the data, the focus is on the specific value given to the observation for which the data is missing. We can contrast this with the survey data approach where the focus is on identifying homogenous groups of observations, based on the target variable, so that these groups can be used by other algorithms to produce a reasonable value for an observation for which the data is missing. Surveys have two types of missing data: unit level missingness, e.g., someone did not respond, or item level missingness, e.g., respondent failed to provide information for certain questions. For either type of missing data in the survey context, tree-based prediction models can play an important role. The most common approach to account for unit nonresponse is to adjust the weights. Tree-based prediction models can identify weighting classes within which the weight adjustment will be made. The most common approach to account for item nonresponse is imputation. Tree-based methods can identify imputation classes within which imputation will be conducted. This paper describes examples of the use of tree-based prediction models for both types of missing data in the survey context.

Finite User Pool Effect in Two Sample t-test of Controlled Experiments on the Web
Shaojie (Alex) Deng, Microsoft

Abstract: The i.i.d. assumption is a basic assumption made in the application of two sample t-test in practice. In this paper, we show that when this assumption is not true and the sampling scheme is replaced by sample without replacement from a finite pool, the variance estimator assuming i.i.d. samples overestimates the variance. The overestimating effect is small or negligible when the sample size is small relative to the finite pool size. However, the results in this paper are relevant to controlled experiment on the web because the capability of conducting controlled experiments using a large proportion of the whole web users.

Network Analytics at SAS
Barry deVille and David Duling, SAS Institute

Abstract: Network analytics and visualization have been a significant component of a number of SAS R&D activities over the last decade, particularly in the areas of Data Mining, Text Mining and Operations Research. The development has intensified in recent years as computing resource growth, “active” web applications, and the growth of social media such as Twitter and Facebook have given rise to substantially new customer models based on network relations and associated real time web displays and animations. This presentation provides an overview of network analytics applications at SAS that ranges from a discussion of (1) link analysis of relational and transactional data and text, (2) social network analysis in telecommunications marketing, and (3) social media analytics.

Predictive Modeling Processes and Heavy Alcohol Use
Barry Eggleston, RTI

Abstract: Hypothesis generation and testing, estimation of population risk, and forecast of future outcomes are four distinct goals in clinical research including the studies of alcohol dependence and other alcohol related issues. Accomplishing these goals may require different analysis methodologies and processes, explanatory modeling for hypothesis testing and population risk estimation and predictive modeling for future outcome forecasting. This presentation will illustrate a predictive modeling process used to construct and select a predictive model of heavy alcohol use. The process will involve an assessment of predictive model strength relative to other potential models, followed by proper validation of candidate predictive models using an independent sample. The data used in the illustration are from two projects that studied association between

heavy alcohol use and factors such as novelty seeking trait, and sweet liking characteristics. The paper will conclude with a discussion about limitations to applicability of the resulting predictive model.

Using Hybrid Maps to Understand Space and Promote Common Ground
MaryAnne Fields, Jason Owens, U.S. Army Research Laboratory and Kostas Daniilidis, (University of Pennsylvania

Abstract: As robotic systems mature, researchers are investigating techniques to establish common ground between the robots and humans, enabling them to use the same terminology to refer to objects and actions in the environment. Effective spatial representation and reasoning requires the robot subdivide its environment into regions such as rooms, halls, plazas, or paths. In this paper, we present a method to incrementally partition an indoor environment into human-understandable regions such as rooms and halls. The method uses point-to-point visibility and estimates for the location of structural elements such as walls and doorways to find the regions. Supervised learning techniques allow region labels to be assigned from a number of region characteristics including size, connectedness and adjacency.

Counting Contingency Tables via Multistage Markov Chain Monte Carlo
George Fishman, University of North Carolina, Chapel Hill

Abstract: We describe a multistage Markov chain Monte Carlo (MSMCMC) procedure for estimating the count of a collection of contingency tables. For each stage, Hastings-Metropolis (HM) sampling generates states with equilibrium distribution that facilitates sampling by relaxing columns sums while maintaining row sums. Our choice of penalty function and novel algorithms developed to this end eliminate local minima, thus precluding the problem of sample paths becoming trapped at those minima. The proposed procedure cycles through stages in the reverse order of a proposed inverse-temperature schedule (heating), inducing faster convergence to equilibrium than cycling through a cooling schedule, as in simulated annealing. Examples demonstrate the efficacy of the procedure for tables of varying sizes. They also show that switching between two proposed algorithms on an approximately optimally chosen stage induces a notable improvement in statistical efficiency.

Practical Implementation of Reproducible Research in the Madagascar Software Project
Sergey Fomel, University of Texas at Austin

Abstract: The Madagascar open-source software project is a community effort, which implements reproducible research practices as envisioned by Jon Claerbout. More than 100 geophysical data analysis papers have been published, together with open software code and data, and are maintained by the community. We have learned several important principles in our implementation of reproducible research: (1) The principal beneficiary is the author. (2) Each computational experiment is a test. (3) Computational reproducibility requires continuous maintenance and an open community. In the presentation, I will share these and other lessons. Reproducible research is important for scientific integrity and scientific collaboration. It is possible to implement it with the tools that exist today.

Spatial and Temporal Trends in Extreme Temperatures
Montserrat Fuentes*, North Carolina State University

Abstract: Estimating the probability of extreme temperature events is difficult because of limited records across time and the need to extrapolate the distributions of these events, as opposed to just the mean, to locations where observations are not available. Another related issue is the need to characterize the uncertainty in the estimated probability of extreme events at different locations. Although the tools for statistical modeling of univariate extremes are well-developed, extending these tools to model spatial extreme data is an active area of research. In this work, in order to make inference about spatial extreme events, we introduce a new nonparametric model for extremes. We present a Dirichlet-based copula model that is a flexible alternative to parametric copula models such as the normal and t-copula. This presents the most flexible multivariate copula approach in the literature, and allows for nonstationarity in the spatial dependence of the extremes. The proposed modeling approach is fitted using a Bayesian framework that allows us to take into account different sources of uncertainty in the data and models. We apply our methods to annual maximum temperature values in the east-south-central United States.*In collaboration with J. Henry and B. Reich (NCSU)

Cross-section and Longitudinal Penalized Functional Regression
Jeff Goldsmith, Jennifer Bobb, Ciprian Crainiceanu, Brian Caffo, Daniel Reich, Johns Hopkins University

Abstract: We develop fast-fitting methods for generalized functional linear models. The functional predictor is projected onto a large number of smooth eigenvectors and the coefficient function is estimated using penalized spline regression. Our method can be applied to many functional data designs, including functions measured with and without error, sparsely or densely sampled over regular or irregular grids. The methods are also extend to the recent but increasingly relevant longitudinal case, in which functional predictors and scalar outcomes are recorded over multiple visits. The approach can be implemented using standard mixed effects software or in a Bayesian framework. We are motivated by a study of white matter demyelination via diffusion tensor imaging in which various cerebral white matter tract property are used to predict cognitive and motor decline in multiple sclerosis patients. All methods are implemented in the 'refund' package available on CRAN.

Cross-country Autonomous Navigation: Trade-offs and Challenges
Juan Pablo Gonzalez, General Dynamics Robotic Systems

Abstract: Unmanned ground vehicles (UGVs) navigating in cross-country environments face significant challenges because of the size of the environments and the large number of parameters required to compute viable paths. Hierarchical planners decompose the search space such that long traverses can be completed in a timely fashion and with a limited amount of resources. While this approach has been widely used and provides satisfactory results in many environments, it usually underperforms or fails in very complex environments. In this talk, we will discuss some of the trade-offs and challenges involved in cross-country autonomous navigation, as well as some of the most recent approaches to tackle them.

21st Century Analysis, Biomedical, Collaboration and Determinism Dogma: Emerging Challenges and Guidance for Interface of Computing and Statistics
Arnold Goodman, Collaborative Data Solutions

Abstract: The 20th Century witnessed processing advance from desk calculators and mainframes, through timesharing and PCs, to supercomputers and cloud computing. It has evolved from too little data into almost too much data, and from theory dominating data into data beginning to dominate theory while needing new theory. In addition, it has witnessed problems advancing from simple in a lone discipline into becoming almost too complex in multiple disciplines, as well as from analysis driving solutions into solutions by data mining beginning to drive the analysis itself. How we accomplish this has transitioned from competition overcoming collaboration into collaboration starting to overcome competition. What is done being more important than how it is done is changing into how it is done becoming as important as what is done. In addition, what or how we do it being more important than what or how we should really do it is shifting into what or how we should do it becoming just as important as what or how we do it, if not more. Although we have come a long way in both our methodology and technology, are they sufficient for our complex and multidisciplinary problems and their massive databases? Since the apparent answer is not yet a resounding yes, we have inherited many challenges and opportunities. We now also need to explore the variation and likely uncertainty from our selection of assumptions and their models or our analysis and its software. To see emerging trends in statistical analysis and data mining, I sampled comments by experts at 2010 Joint Statistical Meetings, Symposium on Interface of Computing Science and Statistics, SIAM Data Mining Conference, plus Workshop IV on Data Hierarchies for Climate Modeling in UCLA Program on Model and Data Hierarchies for Simulating and Understanding Climate -- at its Institute for Pure and Applied Mathematics. My findings were summarized in “Emerging Topics and Challenges for Statistical Analysis and Data Mining”[1] and in “Analysis, Biology, Collaboration and Determinism Challenges and Guidance: A Wish List for Biopharmaceuticals on Interface of Computing and Statistics”[2]. This personal perspective goes well beyond them by expansion and extension. In times of exploding change, preview perspectives on what should be explored may be even more valuable than the far more typical reviews of what has already been accomplished in both conferences and journals. Of particular significance is the guidance for two novel challenges: my training of collaborative cultures that integrate over our disciplinary and organizational cultures which tend to behave much as tribes behave, and my replacing genetic determinism dogma with an uncertainty. This places uncertainty and statisticians at the very heart of molecular biology and genetics, now where the action and funding are. We are in a position of being capable for contributing to not only icing on the cake, but also the cake itself. The preview perspective shares my 57 years traveling over the interface of computing and statistics, from 8x8 matrix inversion

on desk calculators through digitizing of analogue data to programming in SAS. Information technology adventures span information systems from user needs to their evaluation, as well as data center management from computer capacity planning to performance evaluation. They also span from software quality and IT value into pioneering the new interface of data mining and statistics as well as an uncertainty in the cell. Experience has progressed through equal tours in aerospace, petroleum, government and university. I try to focus upon the right challenges and guidance, with absolutely no intention of being righteous.

Neural Nets – Boosted and Cross-Validated
Christopher Gotwalt, SAS/JMP

Abstract: Neural networks are a very flexible class of models that can be employed in a wide variety of contexts. In this talk we will illustrate their use as surrogate functions for finite element solvers with an emphasis on automated model selection of the number of hidden layer nodes via boosting.

Using Maps as a Background for Displaying Information that is Geographically Distributed
Xan Gregg, SAS/JMP

Abstract: We will discuss the advantages and the challenges associated with displaying geographic information in an exploratory data analysis context. Specifically, we must balance accuracy, including scale and projection, with responsive interaction. Examples will be demonstrated with JMP.

Dissimilarity Plots: A Visual Exploration Tool for Partitional Clustering
Michael Hahsler and Kurt Hornik, Southern Methodist University

Abstract: For hierarchical clustering, dendrograms are a convenient and powerful visualization technique. Although many visualization methods have been suggested for partitional clustering, their usefulness deteriorates quickly with increasing dimensionality of the data and/or they fail to represent structure between and within clusters simultaneously. In this paper we extend (dissimilarity) matrix shading with several reordering steps based on seriation techniques. Both ideas, matrix shading and reordering, have been well-known for a long time. However, only recent algorithmic improvements allow us to solve or approximately solve the seriation problem efficiently for larger problems. Furthermore, seriation techniques are used in a novel stepwise process (within each cluster and between clusters) which leads to a visualization technique that is able to present the structure between clusters and the micro-structure within clusters in one concise plot. This not only allows us to judge cluster quality but also makes misspecification of the number of clusters apparent. We give a detailed discussion of the construction of dissimilarity plots and demonstrate their usefulness with several examples. Experiments show that dissimilarity plots scale very well with increasing data dimensionality.

Visualizing Association Rules in Hierarchical Groups
Michael Hahsler and Sudheer Chelluboina, Southern Methodist University

Abstract: Association rule mining is one of the most popular data mining methods. However, mining association rules often results in a very large number of found rules, leaving the analyst with the task to go through all the rules and discover interesting ones. Sifting manually through large sets of rules is time consuming and strenuous. Visualization has a long history of making large amounts of data better accessible using techniques like selecting and zooming. However, most association rule visualization techniques are still falling short when it comes to a large number of rules. In this paper we present a new interactive visualization technique which lets the user navigate through a hierarchy of groups of association rules. We demonstrate how this new visualization technique can be used to analyze a large sets of association rules with examples from our implementation in the R-package arulesViz.

Similarity-Based Information Fusion
Timothy P. Hanratty, Eric Heilman, John Richardson, Andrew Neiderer, U.S. Army Research Laboratory

Abstract: The US Army Research Laboratory (ARL) has developed a data fusion methodology and a software application to enable near-real time intelligence analysis of individuals or entities based on their resemblance or similarity to key reference population groups. The software implementation of the methodology, called the Visual

Multidimensional Scaling (VMDS), is presently a prototype software tool. VMDS provides an estimation of an entity's probable threat orientation based on resemblance to members of a known set of enemy and friendly entities. The embedded algorithm is tolerant of sparse or incomplete inputs and able to make reasonable comparisons with partial data. By using information parameters typically generated by robotic, automated, and manual sensors, the authors have adapted VMDS for use in the classification of terrain features such as buildings and specific terrain areas of interest such as parks and squares. Entry of a new set of sensed data describing specific terrain into the VMDS tool will result in the creation of a similarity graphic useful in making an estimation

The Architecture of Rc2
E. James Harner and Mark Lilback, West Virginia University

Abstract: Rc2 is a cloud-based, collaborative, web 2.0 interface to R accessed through WebKit-based browsers (e.g., Safari and Chrome). Client-specific style sheets and scripting provide a touch-optimized interface for mobile devices such as the iPad. R sessions are not tied to a specific computer or user, and collaborators can share the same R workspace and environment. For example, instructors can schedule interactive classroom R sessions where students can watch the instructor who can optionally turn over control to a student. These features allow researchers to collaborate over the Internet without concern for data becoming out of sync. Users can start long-running computations and Rc2 will notify the user(s) when the process is complete. Rc2 is implemented as a 4-tier system: a web 2.0 client, app server components, an R environment, and persistent storage. Client interfaces include a text editor for R scripts or R-embedded LaTeX code, a command line, full R output, a workspace display, and an interactive graphics display via our custom R graphics package. The R environment is a load-balanced set of Rserve instances. Persistent storage is split, with app server data stored in a SQL database and R-related data in Apache Cassandra, a distributed, write-optimized NoSQL database. This allows sessions to be reconstructed on another app server worker in the event of failure. App server components are highly scalable (both vertically and horizontally) and manage user authentication, dynamic auto-configuration, fast client-server communications using WebSockets, and load-balancing, among other tasks.

Random Forests Variable Selection
Hemant Ishwaran, Dept Quantitative Health Sciences at Cleveland Clinic

Abstract: It is now well known that the simple act of combining an elementary base learner such as a CART tree can yield a predictor with superior prediction performance. Learners that aggregate base learners are often referred to as ensembles. Of the many ensemble techniques introduced, one of the most successful is Random Forests (RF). It is an all-purpose tree-based method that can be used for survival, regression, and classification settings. The issue of variable selection with RF is however not straightforward. I discuss the widely used "permutation approach" and introduce a new method based on tree depth. I illustrate advantages of the new approach and discuss some of its properties including the issue of properly calibrating key RF tuning parameters in high dimensions.

Smoothing Dynamic Positron Emission Tomography Time Courses Using Functional Principal Components
Ci-Ren Jiang, Statistical and Applied Mathematical Sciences Institute

Abstract: A functional smoothing approach to the analysis of PET time course data is presented. By borrowing information across space and accounting for this pooling through the use of a nonparametric covariate adjustment, it is possible to smooth the PET time course data thus reducing the noise. A new model for functional data analysis, the Multiplicative Nonparametric Random Effects Model, is introduced to more accurately account for the variation in the data. A locally adaptive bandwidth choice helps to determine the correct amount of smoothing at each time point. This preprocessing step to smooth the data then allows subsequent analysis by methods such as Spectral Analysis to be substantially improved in terms of their mean squared error.

An Interactive Graph for Exploring Multi-Dimensional and Multivariate Functions
Bradley Jones, SAS/JMP

Abstract: Twenty years ago the main way experimenters visualized a multivariate response surface was by using contour plots with overlaid responses. This method was useful as long as there were only a few responses and two quantitative factors. However, our minds are limited by our three dimensional world when it comes to being able to visualize multi-factor multiple response functions. The graph described in this talk was invented to address this problem in visualization. It involves using a matrix of plots where each row of the matrix describes the conditional relationship of one response with multiple factors. Each column of the matrix shows the conditional relationship of one factor on multiple responses. The power of the graph comes from its interactivity. Through real-time updating of the entire matrix of graphs as the user drags a line controlling the value of any factor, the user can develop a powerful (tactile) understanding of the underlying function. This talk introduces the graph and shows several examples showing its usefulness from optimization using empirical models to illuminating concepts in statistics courses.

Hypergraphs from Twitter Data
Elizabeth Hohman, Naval Surface Warfare Center, Dahlgren Division

Abstract: We use data from the micro-blogging service Twitter. Public tweets from March, 2011 are used to form hypergraphs, where hyperedges result from words, hashtags, or topics. We explore the use of topic models on the hashtags and associated tweets in order to answer questions about the hashtags, such as how they change in time and whether they have multiple meanings. This is preliminary work and the majority of the presentation will focus on the problem definition and data processing.

Two Entities and Beyond: Challenges and Examples in Social Network Analysis (SNA)
Jin-Whan Jung, Dominic Jann, Dan Kelly, John Brocklebank, SAS Institute

Abstract: Network analysis seeks to combine entity-based measurements (an individual's behavior) with information about their peers (a group's behavior). This raises many questions that standard statistical techniques may ignore, including questions of matching entities. Out of deterministic and probabilistic record matching approaches, deterministic record matching is well-established and has been used widely in many industries. On the other hand, probabilistic record matching has been relatively underutilized due to the complexity of its assumptions. In this presentation, the matching problem in SNA will be reviewed and the probabilistic matching will be discussed as a challenging point. Fraud Detection example in Welfare and Contagious Churn example in Telco will be demonstrated.

Detection of Functional Abnormalities in Brain Using Shape Analysis of Subcortical Structures
Sebastian Kurtek, Eric Klassen, Florida State University, Zhaohua Ding, Malcolm J. Avison, Vanderbilt University, Anuj Srivastava, Florida State University

Abstract: We consider the task of computing shape statistics and classification of 3D anatomical structures in the brain (as continuous, parameterized surfaces) under a Riemannian framework. This task requires a Riemannian metric that allows: (1) re-parameterizations of surfaces by isometries, and (2) efficient computations of geodesic paths between surfaces. These tools allow for computing Karcher means and covariances (using tangent principal component analysis) for shape classes, and a probabilistic classification of surfaces into disease and control groups. We develop a path-straightening algorithm for computing geodesic paths. This process requires optimal re-parameterizations (deformations of grids) of surfaces and achieves a superior alignment of geometric features across surfaces. The resulting means and covariances are better representatives of the original data and lead to parsimonious shape models. These two moments specify a normal probability model on shape classes, which are then used for classifying test shapes. Through improved random sampling and a higher classification performance, we demonstrate the success of this model over some past methods. We use the Detroit Fetal Alcohol and Drug Exposure Cohort data to study brain structures and present classification results for the Attention Deficit Hyperactivity Disorder cases and controls in this study. We find that using the mean and covariance structure of the given data, we are able to attain an 88% classification rate, which is an improvement over a previously reported result of 82% on the same data.

Singular Spectrum Analysis Algorithm for decomposing and visualizing time series data
Michael Leonard, SAS Institute Inc.

Abstract: Singular spectrum analysis (SSA) is a relatively new approach to modeling time series data. Now supported in SAS/ETS software, the SSA method of time series analysis applies nonparametric techniques to decompose time series into principal components. SSA is particularly valuable for long time series, for which patterns (such as trends and cycles) are difficult to visualize and analyze. This paper provides an introduction to singular spectrum analysis and demonstrates how to use SAS/ETS software to perform SSA. As an illustration, monthly data on U.S. temperatures over the last century are analyzed to discover significant patterns.

Using Random Forest Models to Predict Organizational Violence
Burton Levine, Georgiy Bobashev, RTI International

Abstract: We use Random Forest modeling in the context of social theories to predict the probability that an organization will commit violence against non-government security forces or civilians. Effective implementation of these models could aid intelligence analysts in accessing terrorist threats. We leverage existing social theories and publically available data sources to build our predictive models. Our data source is longitudinal; consequently, implementation of this Random Forest model is complicated by the correlations within the data. We discuss the methodology used to account for this correlation. Finally, we present the results of the model fit and discuss model validation and the interpretation of the results.

Patterns of Daily Alcohol Consumption
D. Liao, J. Hampton., G. Bobashev, RTI

Abstract: Individual alcohol consumption varies dramatically between individuals even those who are considered “heavy users”. Understanding the drivers behind the patterns can help design prevention and treatment programs. We have analyzed time series of daily alcohol consumption over 6 months period for over 200 and classified then into 9 classes using innovative approach that considers sliding windows in three dimensions: mean, variance (also the coefficient of variance) and median. We also used the periodic (weekly) forcing and local patterns to identify sub-patterns. Our analysis have shown that binge drinkers are characterized by small median and variance, while for heavy alcohol users an assumption of scaled Poisson distribution can hold true. For many subjects however we observe a switch between pattern types. Specifically the most prevalent switch was between heavy frequent drinking and binge drinking. The data on individual alcohol consumption is very limited, and this is the first study aiming to build forecasting models using such data. Although the sample is not representative of the general or risk population our analysis is likely to capture most of the patterns. Validation on another study with 33 subjects providing their daily data up to 2 days have shown that the patterns fall into existing categories. This study offers the field a 'next step' that can possibly offer clinical implications for relapse prevention, increased treatment efficiency, and enhance understanding of the factors driving the variety of daily patterns of use.

Coupling Optional Polya Trees – A Bayesian Nonparametric Approach to Case-Control Studies
Li Ma, Stanford University

Abstract: Testing and characterizing the difference between two data samples (case versus control) is of fundamental interest in statistics. Parametric methods such as (logistic) regression-based approaches are often too restrictive in complex problems, while existing nonparametric methods do not scale well as the dimensionality increases and often provide no easy way to characterize the difference should it exist. In this talk, we introduce an inferential framework that addresses these challenges in the form of a prior for Bayesian nonparametric analysis. This prior, called the “coupling optional Polya tree” (co-OPT) distribution, is constructed based on a procedure of random recursive partitioning and probability assignment on the sample space. It has the ability to jointly generate multiple random distributions. These probability distributions are allowed to randomly "couple", that is to have the same conditional distribution, on subsets of the sample space. We show that posterior inference on the coupling state of the distributions provides an effective way both for testing the existence and for learning the structure of two sample difference, even in the presence of data sparsity. Several simulated and real

data analytical examples in genetic epidemiology and flow cytometry will be provided to illustrate the application of this method.

Visual Statistical Inference for Regression Parameters
Mahbubul Majumder, Heike Hofmann, Dianne Cook, Iowa State University

Abstract: Statistical graphics play a crucial role in exploratory data analysis, model checking and diagnosis. Until recently there were no formal visual methods in place for determining statistical significance of findings. This changed, when Buja et al.(2009) conceptually introduced two protocols for formal tests of visual findings. In this paper we take this a step further by comparing the lineup protocol (Buja et al.2009) against classical statistical testing of the significance of regression model parameters. A human subjects experiment is conducted using simulated data to provide controlled conditions. Results suggest that the lineup protocol provides results equivalent to the uniformly most powerful (UMP) test and for some scenarios yields better power than the UMP test.

Probability Machines
James D. Malley, Center for Information Technology, National Institutes of Health

Abstract: Many statistical learning machines can provide an optimal classification for binary outcomes. However, probabilities are required for risk estimation using individual patient characteristics for personalized medicine. This talk shows that statistical learning machines that are consistent for the nonparametric regression problem are also consistent for the probability estimation problem. These will be called probability machines. Probability machines discussed include classification and regression random forests and two nearest-neighbor machines, all of which use any collection of predictors with arbitrary statistical structure. Two simulated and two real data sets illustrate the use of these machines for probability estimation for an individual.

Combining Geographic Information in Random Dot Product Graphs
David Marchette, Naval Surface Warfare Center, Dahlgren Division

Abstract: We consider a random graph model that associates a collection of vectors (attributes) to the edge probabilities of the graph. This model, the random dot product graph (RDPG), has several nice properties for modeling social networks, and has efficient methods for fitting the vectors to a given graph. We will discuss this model within the context of Twitter graphs, in which we observe both the graph and the geographic location of the tweeter. We observe that the individuals tend to "tweet locally" -- most of their tweets are directed to other individuals geographically close to them. We discuss a method for fusing the information from the graph, encoded as the RDPG vectors, with geographic information, to produce an improved method for predicting future graphs.

The Reliability of Millennial Multi-proxy Temperature Reconstructions
Blake McShane, Northwestern University

Abstract: Predicting historic temperatures based on tree rings, ice cores, and other natural proxies is a difficult endeavor. The relationship between proxies and temperature is weak and the number of proxies is far larger than the number of target data points. Furthermore, the data contain complex spatial and temporal dependence structures which are not easily captured with simple models. In this paper, we assess the reliability of such reconstructions and their statistical significance against various null models. We find that the proxies do not predict temperature significantly better than random series generated independently of temperature. Furthermore, various model specifications that perform similarly at predicting temperature produce extremely different historical backcasts. Finally, the proxies seem unable to forecast the high levels of and sharp run-up in temperature in the 1990s either in-sample or from contiguous holdout blocks, thus casting doubt on their ability to predict such phenomena if in fact they occurred several hundred years ago. We propose our own reconstruction of Northern Hemisphere average annual land temperature over the last millennium, assess its reliability, and compare it to those from the climate science literature. Our model provides a similar reconstruction but has much wider standard errors, reflecting the weak signal and large uncertainty encountered in this setting.

Visualizing the Variability of Plots
Rajiv Menjoge and Roy Welsch, MIT

Abstract: Plots are an essential part of any statistical analysis, but are, in practice, typically reported as a single stand-alone summary of data. This can be limiting, since, like any other parameter of the data, they have variability which emerges from how data were drawn from the population, how the plot was chosen, and other assumptions that went into the generation of the data. As an example, if the data were resampled from the population, the plot could look entirely different. We present a method for displaying the variability of a plot, by using simulation to generate several plots that could have emerged under a different sample from the population or a different set of assumptions, and then using a distance metric between plots to methodically select a representative subset. We demonstrate this method on examples of simple plots, where we make no assumptions about the distribution of the data, to show the method's usefulness, in improving the validity of visualization in statistics. We then present an alternative application of this method to simplify the visualization of large scatterplot matrices.

Finding Structural Variants in Individual Human Genomes with Random Forests
Jacob Michaelson, University of California, San Diego

Abstract: Genomic structural variants are lesions in genetic material that are major drivers of individual and between-species diversity, as well as disease. The reliable identification of the locations of these lesions from high-throughput sequencing data is an open problem, and is becoming increasingly timely as sequencing costs fall and the technology becomes poised to revolutionize medicine. We have developed a method, based on Random Forests, to accurately identify the nature, location, and size of these structural variants. This problem represents an additional layer of complexity over traditional classification problems in that both the class and the event boundaries must be accurately predicted to be useful. We evaluated the method on data recently released by the 1000 Genomes Project, and performance is encouraging.

Archiving Computational Research in Virtual Machines
Sorin Mitran, University of North Carolina

Abstract: Several approaches have been taken by computational scientists to ensure open access to their research codes: providing source codes, using a purpose-built archival system, literate programming tools. These procedures reflect standard practices in experimental sciences where laboratory techniques, supplies and equipment are documented in a research paper. Computational research has one advantage with respect to experimental science: our entire laboratory can be packaged and sent to independent parties for validation of research results. Virtualization has advanced to a stage in which direct access to graphics processing hardware and multiple CPU parallel processing can be included in virtual machines. The entire panoply of open-source tools for scripting and documentation can be included with the virtual machine. This talk will present experience with this approach in the context of interdisciplinary research that uses two of the author’s codes (BEARCLAW and Diapason). Particular attention is paid to documentation and use of the TeXmacs environment to present both theory and implementation of algorithmic ideas.

Visual Exploration of High-Dimensional Data by Interactive Navigation of Low-Dimensional Data Spaces
Wayne Oldford, University of Waterloo, Canada

Abstract: The structure of a set of high dimensional data objects (e.g. images, documents, molecules, genetic expressions, etc.) is notoriously difficult to visualize. In contrast, lower dimensional structure (esp. 3 or fewer dimensions) is natural to us and easy to visualize. A not unreasonable approach, then, might be to explore one low-dimensional visualization after another in the hope that, together, these will shed light on the higher dimensional structure. In this talk, I will introduce some graph-theoretic structures which have low dimensional spaces as nodes/vertices and transitions from one space to another as edges. To be concrete, suppose that each node is a 2-d scatterplot of the data and that an edge exists between nodes whose corresponding scatterplots share a variable. In this case, travel along an edge amounts to a 3d transition effected by rotating one 2d scatterplot into the next. More generally, imagine a user moving a “You are here” circle, or “bullet”, from one node to another along defined edges, causing one data visualization to be smoothly morphed into the other. A walk on the graph represents a low-dimensional trajectory through the higher dimensional space. Of interest,

are walks along these graphs that reveal meaningful structure in the displayed data. These ideas will be demonstrated on different data sets using an interactive R package called RnavGraph. Rnavgraph allows a user to visually explore any data set by dynamically walking the graph structure and interacting with the displayed data. Methods for constructing these graphs and for identifying interesting subgraphs will also be described and demonstrated. Some dimensionality reduction (manifold learning) methods will also be used to constrain the size of the graph.

Some Interdisciplinary Topics on: i) Game-Theoretical Solutions for Quantitative Cyber-Risk Estimation/Management; ii) CLOUD Computing Implementations to Compute Operational Risk using Java Programming
M. Sahinoglu, Auburn University

Abstract: i) Risk management with Mixed Strategy (Nash equilibrium) is examined as an alternative to Two-Player Zero-Sum as applied to the author's Security-Meter automated software tool for quantitative risk assessment. A *pure strategy* provides a complete definition of how a player will play a game as in a two-player zero-sum solution. In particular, it determines the move a player will make for any situation he or she could face. A player's strategy set is the set of pure strategies available to that player. A *mixed strategy* is an assignment of a probability to each pure strategy. This allows for a player to randomly select a pure strategy. Since probabilities are continuous, there are infinitely many mixed strategies available to a player, even if their strategy set is finite. Of course, one can regard a pure strategy as a degenerate case of a mixed strategy, in which that particular pure strategy is selected with probability 1 and every other strategy with probability 0. For example, in a 2x2 setting against any strategy of Person 2, this gives Person 1 an expected gain in a scenario where risks are chosen to assure a Nash equilibrium:

(3/8)(20) + (5/8)(-10) = 5/4 (against 1 of Person 2);
(3/8)(-30) + (5/8) (20) = 5/4 (against 2 of Person 2).

We plan to implement a varying range of mixed strategy solutions, including a Nash equilibrium as a steady state to see the impact as judged by the expert opinions and sound judgment calls from the practicing engineers and managerial cadre regarding the security-meter assessment for the risk management algorithm so as to compare with a two-player zero-sum solution. A mixed strategy equilibrium predicts that the outcome of a game is stochastic, so that for a single play its prediction is less precise than that of a pure strategy. The author will show as an original contribution on how the Nash equilibrium is derived for the security-meter algorithm as an alternative to the conventional two-player zero-sum solution. In summary:

1) Two-Person Zero-Sum with Minimax = Maximin solution with an existing saddle point; that is, the pure strategy by von Neumann (1928).
2) Not all two-person zero-sum games have saddle points. Such games employing Minimax ≥ Maximin using probability mixes of strategies will enable the game to have a saddle point in mixed strategies. Nash equilibrium (1953) is the optimal steady-state solution.

ii) CLOUD Computing implementations on how to estimate RoS (Risk of Service) using CLOURA (CLOUD Risk Assessor) will be demoed for small and large cyber-systems. What-if scenarios will be examined in the light of recently occurring reliability glitches in the commercial cloud arena regarding major Cloud services like Amazon, HP, or Google to name a few.

Response Surface Methodology
John Sall, JMP/SAS Institute

Abstract: When doing response surface modeling, visualization of many dimensions is a challenge. Good visualizations can be done by showing various cross-sections. However to better explore acceptable factor regions, it is often better to explore the space by interactive filtering. Visualizing the Pareto frontier adds further value. Brute-force Interactive filtering becomes a general purpose tool, which is useful in many other exploration contexts.

Computing for Robust Process Engineering
John Sall, Bradley Jones, and Christopher Gotwalt, JMP/SAS Institute

Abstract: When you design a manufacturing process, you want to make it robust with respect to variation in the inputs, including environmental inputs. What you want to model is the defect rate with respect to a set of output specification limits. Simulation is the tool for doing this. However exploring the defect rate for regions of very low probability in an age of six-sigma requires a sampler that aggressively probes these outlying regions, and using a Gaussian Process fit on the log-defect-rate surface becomes the surrogate model to optimize for robust factor settings.

Functional Varying Coefficient Models
Damla Senturk, UCLA School of Public Health and Hans Georg Mueller, University of California, Davis

Abstract: The proposed functional varying coefficient model provides a versatile and flexible analysis tool for relating longitudinal responses to longitudinal predictors. Specifically, this approach provides a novel representation of varying coefficient functions through suitable auto- and cross-covariances of the underlying stochastic processes, which is particularly advantageous for sparse and irregular designs, as often encountered in longitudinal studies. Existing methodology for varying coefficient models is not adapted to such data. The proposed approach extends the customary varying coefficient models to a more general setting, in which not only current but also recent past values of the predictor time course may have an impact on the current value of the response time course. The influence of past predictor values is modeled by a smooth history index function, while the effects on the response are described by smooth varying coefficient functions. The resulting estimators for varying coefficient and history index functions are shown to be asymptotically consistent for sparse designs. In addition, prediction of unobserved response trajectories from sparse measurements on a predictor trajectory is obtained, along with asymptotic pointwise confidence bands. The proposed methods perform well in simulations, especially when compared with commonly used local polynomial smoothing methods for varying coefficient models, and are illustrated with longitudinal primary biliary liver cirrhosis data.

Choosing between Logistic Regression and Classification Trees in Data Mining
Simon Sheather and Mike Speed, Texas A&M University

Abstract: Classification trees and logistic regression are key data mining tools used when the response variable is binary (e.g., responded to offer Yes/No). In this talk we describe situations in which logistic regression is expected to outperform classification trees and vice versa. We also provide graphical tools which enable the user to decide what terms to include in a logistic regression. A number of real data sets will be used to illustrate these points.

Approaches and Barriers to Reproducible Practices in Biostatistics
Matthew S. Shotwell and JoAnn M. Alvarez

Abstract: Reproducible research methods aim to establish a record of research activities, so that others can more easily replicate and evaluate scientific findings. In applied statistics, reproducible research means fully documenting or scripting all data analysis plans and procedures. Despite the appeal, reproducibility is not uniformly practiced in biostatistics. A survey conducted within the Department of Biostatistics at Vanderbilt University has identified several practical barriers to adopting reproducible practices. This work addresses some pragmatic measures to overcome these barriers using a reproducible research workflow and free software tools, including revision control systems and the R utility Sweave.

3-d Stereoscopic Plots: From History to R
Juergen Symanzik, Utah State University

Abstract: Three-dimensional (3-d) stereoscopic plots allow human viewers to interpret printed plots, plots shown on a computer screen, or plots projected to a wall as realistic 3-d images. Our human perception of depth, i.e., the third dimension, is due to the fact that each of our eyes sees a slightly different image. When these images are combined in the human brain, we interpret the result as a third dimension that represents depth or distance. In 3-d stereoscopic plots, two slightly different images are created and are presented to the two eyes of the

viewer. When done well, a realistic 3-d image is created in our brain. Various techniques exist to create and present the two different images to the human viewer. In this talk, based on an article recently submitted to *Wiley Interdisciplinary Reviews - Computational Statistics (WIREs),* we will focus on techniques that have been used extensively in the field of statistics, i.e., freeviewing of side-by-side images and anaglyphs. We will start with the origins of these plots in the 19th century, look at their heydays in statistics in the mid 1980s to early 1990s, and finish with their revival in a variety of recent R packages.

Streaming Data
William F. Szewczyk, NSA

Abstract: As the ability to collect data continues to outstrip the ability to process and analyze it, the age-old paradigm of store-and-process is becoming untenable. Finding one or two interesting items in the midst of many possible signals depends on context which often changes over time. A new way to interact with data is needed to handle some of these challenges. The streaming data model is one such approach. In this article, we present the streaming data model as well as two approaches to designing algorithms to handle streaming data. The first, the single processor method, traces its heritage to database models. The second, the multi-processor method, is more aligned with signal processing algorithms. We will close with a critique of the current approaches and some of the statistical challenges that streaming data pose. *WIREs Comp Stat* 2011 3 22–29 DOI: 10.1002/wics.130. Reprinted with Permission of John Wiley and Sons, inc.

Clustering TB Microarray Data for Potential Pathway Discovery and Disease Discrimination
Laura Tipton, Rida Moustafa, George Washington University

Abstract: In 2010 Berry, et al. analyzed microarray data from several populations with tuberculosis presence to find a blood transcriptional signature to discriminate between disease states. Their approach to clustering by subjects has failed to find a way to distinguish healthy subjects from those with latent or active tuberculosis infections. In fact, the subjects do not even cluster by disease state and several subjects with active infections are clustered with the healthy subjects. If instead an older approach of clustering by genes is used then some of the resulting clusters show differentiation between disease states. These distinguishable clusters can then be analyzed for potential disease pathways and highly discriminate gene expression levels. Once highly discriminate pathways and genes are identified, they can be used to classify subjects by disease state. The ultimate goal of tuberculosis microarray analysis would be to predict those subjects with latent tuberculosis who will go on to develop an active infection and classifying subjects as having latent or active infections would be the first step to identifying that predictor.

Exact and Approximate Area-proportional Circular Venn and Euler Diagrams
Leland Wilkinson, SYSTAT

Abstract: Scientists conducting microarray and other experiments use circular Venn and Euler diagrams to analyze and illustrate their results. As one solution to this problem, we introduce a statistical model for fitting area-proportional Venn and Euler diagrams to observed data. The statistical model outlined in this report includes a statistical loss function and a minimization procedure that enables formal estimation of the Venn/Euler area-proportional model for the first time. A significance test of the null hypothesis is computed for the solution. Residuals from the model are available for inspection. As a result, this algorithm can be used for both exploration and inference on real datasets. A Java program implementing this algorithm is available under the Mozilla Public License. An R function {\tt venneuler()} is available as a package in CRAN and a plugin is available in Cytoscape.

Record Linkage
William Winkler, Bureau of the Census

Abstract: Record Linkage This article describes methods for matching duplicates within and across files using non-unique identifiers such as first name, last name, date-of-birth, address, and other characteristics. WIREs Comp Stat 2010 5 535-543 DOI: 10.1002/wics.108. Reprinted with Permission of John Wiley and Sons, Inc.

An Ordinary Differential eEquation-based Solution Path Algorithm
Wu, Yichao, North Carolina State University

Abstract: Efron, Hastie, Johnstone, and Tibshirani [(2004), Least Angle Regression (with discussions), The Annals of Statistics, 32, 409-499] proposed least angle regression (LAR), a solution path algorithm for the least squares regression. They pointed out that a slight modification of the LAR gives the LASSO [Tibshirani, R. (1996), Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society, Series B, 58, 267-288] solution path. However, it is largely unknown how to extend this solution path algorithm to models beyond the least squares regression. In this work, we propose an extension of the LAR for generalised linear models and the quasi-likelihood model by showing that the corresponding solution path is piecewise given by solutions of ordinary differential equation (ODE) systems. Our contribution is twofold. First, we provide a theoretical understanding on how the corresponding solution path propagates. Second, we propose an ODE-based algorithm to obtain the whole solution path.

Learning Representations of Language for Greater Generalization Capacity
Alexander Yates, Temple University

Abstract: Supervised Natural Language Processing (NLP) systems traditionally rely on carefully-engineered, manually-tuned features for accuracy. This paradigm has led to rapid progress in the field. However, recent experiments have shown that such systems have difficulty generalizing to language that differs from the small subset of language observed during training. This is especially true for systems that are trained on one domain, such as newswire text, and tested on another, like biomedical text. Both empirical evidence and theoretical argument point to traditional feature sets as the culprits. In this talk we describe our recent efforts at automatically learning features for supervised NLP tasks. We describe a new paradigm in which language models are trained on a large unlabeled corpus, and then used to embed sentences in a novel feature space. By carefully designing our language models, we can combat problems --- like sparsity and polysemy --- that plague traditional representations, and we can take advantage of linguistic intuitions to improve our representations. Experiments in a variety of domain adaptation settings and for a variety of NLP tasks show that this paradigm leads to state-of-the-art performance.

Analysis on Racial Rates and the Effective Elements of Home Mortgage Lending Acceptance in California
Chong Zhang and Rida Moustafa, George Washington University

Abstract: Efforts to promote equal access to mortgage capital by racial and ethnic minorities have historically been a key component of the civil rights agenda in the United States. Through years, U.S. governments have put lots of efforts to improve the race discrimination situation. However, this is still a big issue. In this paper, we will discuss the mortgage approval rates between each different minority racial group, and the factors to the mortgage acceptance of those racial groups. We used the mortgage lending acceptance as a target variable, and used data mining classification models, including Logistic Regression, Decision Tree, Neutral Network, Support Vector Machine, Naïve Bayes and K-Nearest Neighborhood to generate models and determine the influence variables to the target one, diagnose those models, and select the optimal one that fitted this database.

Predicting and Explaining Potential Caravan Insurance Policy Ownership
Haiyun Zheng and Rida Moustafa, George Washington University

Abstract: As a way to market a product or service, direct mailings to a company's potential customers - "junk mail" to many –is a money-consuming process. If we can accurately predict the potential customers, direct mailings to them will save the company a lot of money. In this report, an efficient prediction model by Naïve Bayes classification method has been built up. This model can predict the potential Caravan Insurance Policy customers out of a big population with high accuracy. The most important prediction variables have also been picked out by feature selection algorithm. From these important prediction variables, we can see that people who have more than one car and are wealthier than average, and who in general carry more insurance coverage than average are more likely to be Caravan Insurance Policy customers.

Smoothing Imaging Data in Population Studies
Hongtu Zhu, University of North Carolina-Chapel Hill

Abstract: Motivated by recent work studying massive imaging data in large neuroimaging studies, we propose various multiscale adaptive smoothing models (MARM) for spatially modeling the relation between high-dimensional imaging measures on a three-dimensional (3D) volume or a 2D surface with a set of covariates. Statistically, MARM can be regarded as a novel generalization of propagation-separation, local kernel smoothing, functional principal component analysis (fPCA) and varying coefficient models (VCM) in higher dimensional space. We develop novel estimation procedures for MARMs and systematically study their theoretical properties. We conduct Monte Carlo simulation and real data analyses to examine the finite-sample performance of the proposed procedures.

BUILDING C - FIRST FLOOR