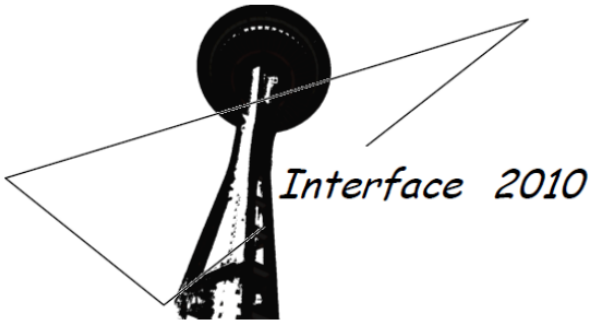
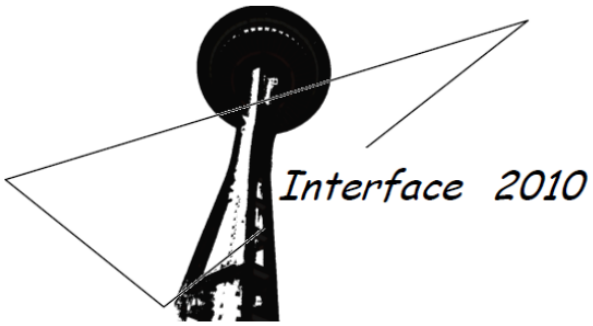


PROGRAM

41st SYMPOSIUM ON THE INTERFACE: COMPUTING SCIENCE AND STATISTICS

THEME: COMPUTATIONAL STATISTICS AND HUMAN BEHAVIOR

June 16-19, 2010
Seattle Westin
Seattle, WA



KEYNOTE SPEAKER

ADRIAN E. RAFTERY
UNIVERSITY OF WASHINGTON

Probabilistic Projections of HIV Prevalence Using Bayesian Melding with Incremental Mixture Importance Sampling (IMIS)

Sponsored by:
Interface Foundation of North America, Inc.

Financial Sponsors:
ASA Section on Statistical Computing
ASA Section on Statistical Graphics
Salford Systems

Cooperating Organizations:
ASA, CSNA, ENAR, IASC, IMS, INFORMS, SIAM, WNAR

This symposium is a long-standing forum focusing on the interface between computing science and statistics

http://www.interfacesymposia.org/Interface2010/

Interface 2010
Program Committee

Edward J. Wegman, George Mason University
Program Co-Chair

Yasmin H. Said, George Mason University
Program Co-Chair

Georgiy Bobashev, RTI International

Barry Bodt, Army Research Laboratory

Hamparsum Bozdogan, University of Tennessee

David van Dyk, University of California, Irvine

Arnold Goodman, Collaborative Data Solutions

Mark Handcock, University of California, Los Angeles

Tim Hesterberg, Google

David J. Marchette, Naval Surface Warfare Center, Dahlgren Division

Rida Moustafa, dMining Technology, LLC

Rebecca Nugent, Carnegie Mellon University

Adrian Raftery, University of Washington

Stephan Sain, National Center for Atmospheric Research

C. Shane Reese, Brigham Young University

Michael Schimek, Danube University, Austria

David Scott, Rice University

Simon Sheather, Texas A&M University

Jeffrey L. Solka, Naval Surface Warfare Center, Dahlgren Division

William Szewczyk, National Security Agency

Michael Trosset, Indiana University

Antony Unwin, University of Augsburg, Germany

Roy Welsch, Massachusetts Institute of Technology

Adalbert Wilhelm, Jacobs University, Bremen, Germany

Interface 2010
Miscellaneous Information

Administrative support for Interface 2010 is provided by Ms. Elizabeth Quigley and Dr. Don Faxon. Interface 2010 is a conference sponsored by the Interface Foundation of North America, Inc. which is a non-profit educational/scientific corporation. The Army Conference on Applied Statistics is also a conference of the Interface Foundation of North America, Inc.

The 2010 Symposium on the Interface is being held at the Westin Hotel in Seattle. Technical sessions including the Keynote address take place in the Cascade Ballroom and St. Helens, Olympic, Vashon and Adams breakout rooms. See the accompanying map of the Cascade Ballroom level. The Vashon and Adams breakout rooms are one level up from the Cascade Ballroom level and are easily reachable via the escalators.

The general website for the Interface Foundation is

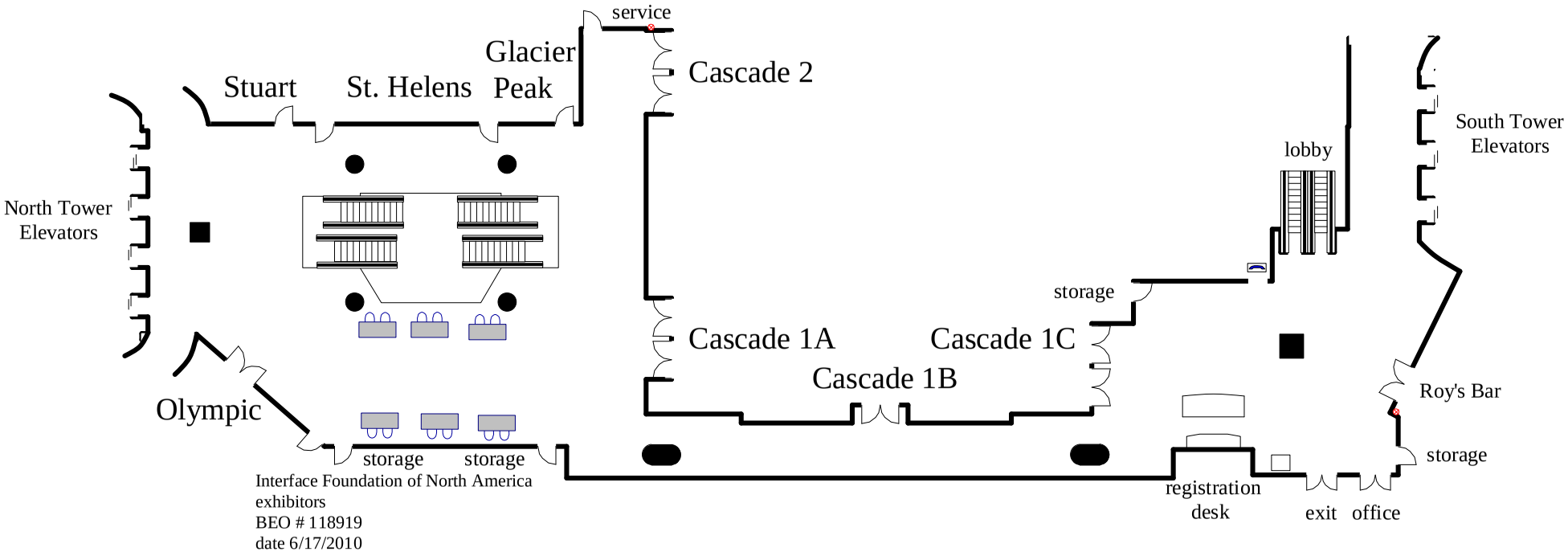
http://www.interfacesymposia.org.

Information about the upcoming Army Conference on Applied Statistics can be found at

http://www.armyconference.org.

Certified
Room Diagram

Cascade Foyer



Interface Foundation of North America
exhibitors
BEO # 118919
date 6/17/2010

Interface 2010 At a Glance

Wednesday - June 16

6:30 pm-8:00 pm	Board of Directors
8:00 pm-10:00 pm	Evening Mixer

Thursday - June 17

8:00 am-9:45 am	Key - 1

Breakouts

	Track 1	Track 2	Track 3	Track 4
10:15 am-12:00 pm	Inv - 1	Inv - 2	Inv - 3	Inv - 4
1:30 pm-3:15 pm	Inv - 5	Inv - 6	Inv - 7	Con - 1
3:45 pm-5:30 pm	Inv - 8	Inv - 9	Inv - 10	Inv - 11

7:00 pm-10:30 pm	Conference Banquet

Friday - June 18

8:00 am-9:45 am	Inv - 12	Inv - 13	Inv - 14	Inv - 15
10:15 am-12:00 pm	Inv - 16	Inv - 17	Inv - 18	Con - 2
1:30 pm-3:15 pm	Inv - 19	Inv - 20	Inv - 21	Work - 1
3:45 pm-5:30 pm	Inv - 22	Inv - 23	Ref - 1	Work - 2

Saturday - June 19

8:00 am-9:45 am	Inv - 24	Inv - 25	Inv - 26
10:05 am-11:50 am	Inv - 27	Inv - 28	Inv - 29

Program for Interface 2010

Wednesday, June 16, 2010

7:30 am – 5:00 pm **Registration**

8:00 am – 12:00 noon **Short Course: Agent Based Modeling and Simulation**

1:30 pm – 5:30 pm **Short Course: Statistical Natural Languages and Text Mining**

6:30 pm – 8:00 pm **Board of Directors Meeting (Closed)** **Baker**

8:00 pm – 10:00 pm **Evening Mixer** **Cascade Ballroom**

Thursday, June 17, 2010

7:30 am – 5:00 pm **Registration**

8:00 am – 9:45 am

Key – 1 ***Keynote Address: Adrian Raftery*** **Cascade Ballroom**
Probabilistic Projections of HIV Prevalence Using Bayesian Melding with Incremental Mixture Importance Sampling (IMIS)

9:45 am – 10:15 am **Morning Break**

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

Inv –1 ***Uncertainty and Simplifications in ABMs*** **St. Helens**
Organizer and Session Chair: Georgiy Bobashev
Using Simulation Models for Decision Making under Deep Uncertainty, Gary Klein, MITRE
Robustness of Risk Maps and Surveillance Networks to the Knowledge Gaps about a New Invasive Threat, Denys Yemshanov, Canadian Forestry
Assessing risks with Agent-based Models (ABMs). Longitudinal projections and added uncertainty, Georgiy Bobashev, RTI

Inv – 2 ***Statistical Analysis and Data Mining: Frontiers of Problem Data and Solution Methodology*** **Olympic**
Organizer and Session Chair: Arnold Goodman
Panel: Usama Fayyad, Jon Kettenring, Michael Leblanc, Roy Welsch

Inv – 3 ***Recent Computational Topics in Astrostatistics*** **Vashon**
Organizer and Session Chair: Rebecca Nugent
Scaling the sky, Andy Connolly, U. Washington
A statistically rigorous approach to astronomical source detection, David Friedenberg, CMU
Using local likelihoods to estimate gravitational lensing of the CMB, Ethan Anderes, UC Davis

Inv – 4 *Perspectives on Climate Change* **Adams**
Organizer: Yasmin H. Said, Session Chair: Edward J. Wegman
Testing the hypothesis of anthropogenic global warming: A continuing controversy, S. Fred Singer, Science & Environmental Policy Project
Extracting information from large-scale computer model output, Mark Berliner, Ohio State
Discussant: Edward J. Wegman, George Mason

12:00 noon – 1:30 pm **Lunch Break**

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

Inv – 5 *Computational Social Science* **St. Helens**
Organizer and Session Chair: Adrian Raftery
A case study in mixed membership modeling with a focus on model selection, Elena Erosheva, U Washington
Approaches to inference from link-tracing network samples, Krista Gile, Oxford
Matrix models for relational and social network data, Peter Hoff, U Washington

Inv – 6 *Novel Methods: Symbolic Data, Layered Graphics Anatomy, and Census Layered Graphics* **Olympic**
Organizer and Session Chair: Arnold Goodman
Symbolic Data, Lynne Billard, U Georgia
Grammars of Graphics, Hadley Wickham, Rice
OnTheMap Data and Tool, Jeremy Wu, Census Bureau

Inv – 7 *Robust Methods in Regression* **Vashon**
Organizer and Session Chair: David Scott
Robust cross-validation using an L1 criterion, Dennis Cox, Rice
Robust parametric classification and variable selection with minimum distance estimation, Eric Chi, Rice and Baylor College of Medicine
Parametric methods for smooth quantile regression, Jonathan Lane, Rice

Con – 1 *Contributed Session 1: Modeling* **Adams**
Session Chair: Yasmin H. Said
Recursive modeling using xstatR, E. James Harner and Jun Tan, West Virginia U
Bayesian analysis for exponential random graph models using the double Metropolis-Hastings sampler, Ick Hoon Jin and Faming Liang, Texas A&M
Rainbow plots, bagplots and boxplots for functional data, Han Lin Shang, Monash

3:15 pm – 3:45 pm **Afternoon Break**

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

Inv – 8 *Computational Statistics Methods for Social Issues* **St. Helens**
Organizer and Session Chair: Yasmin H. Said
Designing statistical tools with agent-based models in mind, Ben Klemens, Census
Predicting edges and vertices in a network, Walid Sharabati, Purdue
Healthcare utilization among border Hispanic seniors with diabetes: Frequencies and related factors, Xiaohui “Sophie” Wang, U Texas, Pan American

Inv – 9 *Computationally Intensive Dynamic Models* **Olympic**
Organizer: C. Shane Reese, Session Chair: Derek Bingham

Simplified trans-dimensional model jumping with MCMC for complicated models, Dave Campbell, Simon Fraser

Emulating the nonlinear matter power spectrum for the universe, Earl Lawrence, Los Alamos National Laboratory

Estimating time-varying parameters in ODEs, Jiguo Cao, Simon Fraser

Inv – 10 *New Developments in Statistical Data Integration* **Vashon**
Organizer and Session Chair: Michael Schimek

Statistical integration of ranked lists by means of data streams, M. G. Schimek, Danube University Krems, Austria

Space oriented rank aggregation, S. Lin, Ohio State

Application of an integrative analysis using association networks of biological data, K. Kugler, UMIT, Austria

Consensus finding, exponential models, and infinite rankings, Marina Meila, U Washington

Inv – 11 *Bayesian Data Mining and Clustering* **Adams**
Organizer and Session Chair: Roy Welsch

Hiding from decision tree detectors, David Scott, Rice

Multivariate Bayesian logistic regression for analysis of clinical trial safety issues, William DuMouchel, Phase Forward Lincoln Safety Group

A graph-based method for estimating the cluster tree of a density, Werner Steutzle, U Washington

7:00 pm – 10:30 pm **Conference Banquet**

Friday, June 18, 2010

7:30 am – 5:00 pm **Registration**

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

Inv – 12 *Novel Methods: Internet Data and Targeted Marketing*, *Seriation and Pattern Discovery, Plus Collaboration and Social Networks*
Organizer and Session Chair: Arnold Goodman **St. Helens**

Internet data and targeted marketing, Usama Fayyad, Open Insights

Collaborative seriation among disciplines and people, Innar Liiv, Tallinn University of Technology, Estonia

Suggesting enrichment of social network analysis beyond graphs, Arnold Goodman, Collaborative Data Solutions

Inv – 13 *Separating the Wheat from the Chaff - Feature Selection in High-Dimensional Regression* **Olympic**
Organizer and Session Chair: Tim Hesterberg

Feature selection with constrained L1 regularization, Leming Qu, Boise State

Bootstrap inference for network construction, Pei Wang, Fred Hutchinson Cancer Research Center

Boosting for nonparametric high-dimensional models, Lifeng Wang, Michigan State

Inv – 14 *All at See: Snapshots of Modern Visualization Research*
Organizer: Antony Unwin, Session Chair: Jay Emerson Vashon

An object-oriented approach in R for the visualization of functional actigraphy data, Jürgen Symanzik, Utah State
Visualization and statistical modeling, Adi Wilhelm, Jacobs University Bremen
Every plot must tell a story - even in R, Heike Hofmann, Iowa State
iPlots eXtreme - next generation of interactive graphics for analysis of large data, Simon Urbanek, AT&T Labs

Inv – 15 *Policy Issues on Climate Change* Adams
Organizer: Yasmin H. Said, Session Chair: Edward J. Wegman

Global warming: Nexus of politics, economics and science, Jeff Kueter, The Marshall Institute
Global warming--fact, fiction, and fraud, Don Easterbrook, Western Washington
Climate change policy and the climategate scandal, Yasmin H. Said, George Mason

9:45 am – 10:15 am Morning Break

10:15 am – 12:00 noon Technical Sessions

Inv – 16 *Computational and Statistical Issues in ABMs* St. Helens
Organizer and Session Chair: Georgiy Bobashev

Simplifying complex systems into multi-level agent based models, Rainer Hilscher, Altarum Institute
The role of population heterogeneity and human mobility in the spread of pandemic influenza, Marco Ajelli, Fondazione Bruno Kessler, Italy
Statistical issues posed by interaction-based models in social systems, Chris Barrett, VBI

Inv – 17 *Graphical Methods for Classification Based on Dimension Reduction* Organizer and Session Chair: Simon Sheather Olympic

SMVCIR dimensionality test, Charles Lindsey, Texas A&M
Robust dimensional reduction via invariant coordinate selection, David Tyler, Rutgers
Sufficient dimension reduction based on normal and its connection with principal components, Liliana Forzani, Universidad Nacional del Litoral

Inv – 18 *Non-English Text Data Mining Via the Vector Space Model*
Organizer and Session Chair: Jeff Solka Vashon

Computing within the foreign vector space framework, Nick Tucey, NSWCDD
Wikipedia as a test bed for implicit translation, Kristin Ash, NSWCDD
An experiment in implicit translation, David Marchette, NSWCDD

Con – 2 *Contributed Session 2: Inference* Adams
Session Chair: Rida Moustafa

Model-averaged L1-penalized logistic regression, Mark Seligman, Chris Fraley, Insilicos LLC
Regression tree boosting to adjust health care cost predictions for diagnostic mix, John W. Robinson, Statistical and Health Informatics Consulting

12:00 noon – 1:30 pm Lunch Break

1:30 pm – 3:15 pm Technical Sessions

Inv – 19 *Statistical Analysis and Data Mining: Frontiers of Social Science and Network Science* **St. Helens**
Organizer and Session Chair: Arnold Goodman
Panel: Pedro Domingos, Adrian Raftery, Lee Wilkinson

Inv – 20 *Quantitative Horizon Scanning* **Olympic**
Organizer and Session Chair: David Marchette
Detecting anomalies and estimating anomaly characteristics in time series of graphs, Glen Coppersmith, Johns Hopkins
Multi-feature clustering and visualization of large document collections, Jeff Solka, Avory Bryant, NSWCDD
Quantitative horizon scanning for mitigating technological surprise, Avory Bryant, NSWCDD

Inv – 21 *Computational Statistics and Robotics* **Vashon**
Organizer and Session Chair: Barry Bodt
Issues and approaches in on-line modeling of environments from 3D data, Marshal Hebert, Carnegie Mellon
*Partial least squares applications in computer vision***,** Larry S. Davis, U Maryland and Aniruddha Kembhavi, U Maryland
*Test methods and metrology for evaluating human detection and tracking systems***,** Tsai Hong, NIST and Barry Bodt, ARL

Work – 1 *Advances in Machine Learning and Data Mining Technology Workshop* **Adams**
Part 1: Advances in Interaction Detection using New Machine Learning and Data Mining Technology, Presented by Dan Steinberg, the President of Salford Systems

3:15 pm – 3:45 pm **Afternoon Break**

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

Inv – 22 *JCGS Highlights* **St. Helens**
Organizer and Session Chair: David van Dyk
Combining mixture components for clustering, Adrian Raftery, U Washington
Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures, Andrew Cron, Duke
Pairwise display of high dimensional information via Eulerian tours and Hamiltonian decompositions, Wayne Oldford, U Waterloo

Inv – 23 *Computer Models, Virtual Laboratories* **Olympic**
Organizer and Session Chair: Stephan Sain
Predictive modeling of a radiative shock physics, Derek Bingham, Simon Fraser
Statistical analysis of regional climate model ensembles, Steve Sain, NCAR
Posterior exploration for computationally intensive forward models, Shane Reese, BYU

Ref – 1 *Refereed Session* **Vashon**
Session Chair: E. James Harner

ChiD, A χ^2-based discretization algorithm, Ross Bettinger

A Bayesian decision theoretic approach to multiple hypotheses problems, Naveen K. Bansal, Marquette and Klaus Miescke, U Illinois, Chicago

Vulnerability of US hospitals from terrorist attack: Pilot study on hospitals in California, Byeonghwa Park and Yasmin Said, George Mason

Work – 2 *Advances in Machine Learning and Data Mining Technology Workshop* **Adams**

Part 2: Hands-on Introduction to New Machine Learning and Data Mining Technology: Jerome Friedman's TreeNet and Leo Breiman's RandomForests, Presented by Dan Steinberg, the President of Salford Systems

Saturday, June 19, 2010

8:00 am – 9:45 am Technical Sessions

Inv – 24 *Sampling and Inference for Hidden Networked Populations*
Organizer and Session Chair: Mark Handcock **St. Helens**

Respondent-driven sampling: Risks and benefits of a novel sampling strategy, W. Whipple Neely, U Washington

Network model-assisted inference from respondent-driven sampling data, Krista J. Gile, Oxford

Developments in network sampling, Steven K. Thompson, Simon Fraser

Inv – 25 *Interfacing Text Mining and Image Analysis* **Olympic**
Organizer and Session Chair: Adi Wilhelm

Combining text and image processing in an automatic image annotation system, Iulian Ilies, Jacobs University, Bremen

Extraction of endogenous metadata for text and image databases, Edward Wegman, George Mason

Keeping your own familiar file organization, Dong Cao, CMU

Inv – 26 *Computing on Streams* **Vashon**
Organizer and Session Chair: William Szewczyk

Stream algorithms and workflows, J. David Harris, DoD

Analytics for streaming applications, Daby Sow, IBM/T. J. Watson Research

Index Learning Omid Madani, SRI

9:45 am – 10:05 am Morning Break

10:05 am – 11:50 am Technical Sessions

Inv – 27 *Model Selection Problems in Kernel-Based Methods with Applications* **Organizer and Session Chair: Hamparsum Bozdogan**
St. Helens

Robust Bayesian relevance vector machines using information complexity and the genetic algorithm, Brant Quinton, U Tennessee

A non-inferiority trial design without the need for a conventional margin, Xi Chen, PharmClint Co

Hybridized support vector machine and recursive feature elimination with information complexity, Seung Hyun Baek, U. Tennessee

Inv – 28 *Visualizing Intrusion Detection Data* **Olympic**
Organizer: Rida Moustafa, Session Chair: Yasmin H. Said

On some visualization methods for computer networking data, Rida E.A. Moustafa, dMining Technology, LLC
Visualizing streaming data, John W. Emerson, Yale
Change detection in multivariate streaming data, Kyle A. Caudle, U.S. Naval Academy

Inv – 29 *Learning from Proximity Data* **Vashon**
Organizer and Session Chair: Michael Trosset

Local learning methods for proximity data, Maya Gupta, U Washington
Interactive concept search, Ashish Kapoor, Microsoft Research
Learning from heterogeneous data sources by combining dissimilarities, Brent Castle, Indiana

Abstracts for Interface 2010

Keynote Address

Probabilistic Projection of HIV/AIDS Prevalence Using Bayesian Melding with Incremental Mixture Importance Sampling (IMIS)

Adrian E. Raftery, University of Washington

The Joint United Nations Programme on HIV/AIDS (UNAIDS) has started to use Bayesian melding as the basis for its probabilistic projections of HIV prevalence in countries with generalized epidemics. This combines a mechanistic epidemiological model, prevalence data and expert opinion. Initially, the posterior distribution was approximated by sampling-importance-resampling, which is simple to implement, easy to interpret, transparent to users and gave acceptable results for most countries. For some countries, however, this is not computationally efficient because the posterior distribution tends to be concentrated around nonlinear ridges and can also be multimodal. We propose instead Incremental Mixture Importance Sampling (IMIS), which iteratively builds up a better importance sampling function. This retains the simplicity and transparency of sampling importance resampling, but is much more efficient computationally. It also leads to a simple estimator of the integrated likelihood that is the basis for Bayesian model comparison and model averaging. In simulation experiments and on real data it outperformed both sampling importance resampling and three publicly available generic Markov chain Monte Carlo algorithms for this kind of problem.

Uncertainty and Simplifications in ABMs

Using Simulation Models for Decision Making under Deep Uncertainty

Gary L. Klein, MITRE Corporation
Jennifer J. Mathieu, MITRE Corporation

All models are wrong in that they are always only an abstraction, a simplification of the real world. Because they deal with some of the most complex aspects of our world, Human Social Cultural Behavior (HSCB) models can be especially wrong in a number of ways. First, considering the complexities with which they are dealing, all HSCB models are incomplete: in order to be developmentally and computationally tractable they invariably leave out some factors or some interactions among factors that impact behavior. In addition, the translation of “raw” socio-cultural data from the real world into model parameters and rules is unavoidably imprecise: how precisely can one measure the attitude of one group toward another? The level of uncertainty described has been termed “deep uncertainty.” It is indeed irreducible uncertainty. However, even under deep uncertainty, the models of social science still are our best synthesis of the data at hand into a usable form. Moreover, people need such models to deal with the increasing complexities in forecasting the outcomes of diplomatic, informational, military or economic courses of action. To support robust decision making under deep uncertainty, models are needed to explore a landscape of plausible futures rather than to make a point prediction. This briefing will discuss how systematic data gathering and translation, coupled with appropriate usage can allow us to take best advantage of these models. It will explain how operational usage of HSCB systems will entail a shift from seeking optimal decisions to seeking robust decisions. To illustrate these concepts, the briefing will describe a robust decision making process as applied to a notional information operations mission thread.

Robustness of Risk Maps and Surveillance Networks to the Knowledge Gaps about a New Invasive Threat

Denys Yemshanov, Canadian Forest Service
Frank H. Koch,North Carolina State University
Yakov Ben-Haim, Technion
William D. Smith, USDA Forest Service

In risk assessments of new epidemics and invasive species it is frequently necessary to make management decisions regarding emerging threats under severe uncertainty. Although risk maps provide useful decision support for invasive organisms, they often fail to recognize and quantify uncertainties associated with the underlying risk model and data assumptions or how they may change the risk estimates. Here we apply an information gap concept to evaluate where a pest risk map is “good enough” in the sense that it is robust to uncertainties about a pest’s behavior while also providing adequately stable risk estimates. We generate risk maps with a stochastic spatial model of invasion that simulates potential entries of an invader with international marine trade, their spread through a landscape and subsequent detection through the surveillance network. In particular, we focus on the question of how much uncertainty in risk model assumptions can be tolerated before the risk map loses its value. The approach is illustrated with an example of a quarantine pest recently detected in North America, Sirex noctilio Fabricius. The results provide a spatial representation of the robustness of predictions of S. noctilio invasion risk to uncertainty and show major geographic hotspots where the consideration of uncertainty in model parameters changes surveillance strategies for a new invasive species. We then use the trade-offs between the extent of uncertainties and the degree of robustness of a risk map to select a survey network design that is most robust to knowledge gaps about the invasive organism.

Assessing risks with Agent-based Models (ABMs). Longitudinal projections and added uncertainty

Georgiy Bobashev, RTI International
William A. Zule, RTI International
Robert J. Morris, RTI International

Longitudinal studies of health outcomes could be very costly, cumbersome, and not representative of the risk population. Conversely, cross-sectional approaches could be representative but have to rely on the retrospective information to estimate prevalence and incidence. Agent-based Modeling (ABM) approach can project behavior from a cross-sectional representative study to generate longitudinal-type data and conduct longitudinal analysis. Thus, it is possible to quantify risks of certain behaviors. I will illustrate the application of the ABMs to a cross-sectional HIV study, show how different behaviors are related to the increase or decrease of HIV risks and how to estimate the quantifiable risk measures such as survival HIV free. Because of the simulated nature of the data, the uncertainty of the estimated risks will naturally lead to a higher variance than would be achieved in a real longitudinal study. I will discuss methodology that considers rigorous statistical assessment and the interplay of survey-based standard errors and simulation-based uncertainty.

Recent Computational Topics in Astrostatistics

Scaling the Sky

Andrew Connolly, University of Washington

Astronomy is addressing many fundamental questions about the nature of our universe through a series of ambitious wide-field imaging surveys. These programs will investigate the properties of dark matter, the nature of dark energy, the evolution of large scale structure as well as searching for potentially hazardous asteroids. The volume of data from these experiments, however, presents many fundamental challenges: how do we determine the interdependencies between the observable properties of stars and galaxies in order to better understand the physics of their formation, how do we combine disjoint images, by collating data from several distinct surveys at different wavelengths, scales, and resolutions, and how do we search for moving or variable sources in incomplete and noisy data. In this talk we explore the nature of these surveys, the impact of kernel density estimation, dimensionality reduction through local embedding, anomaly detection and tracking techniques have on these questions and how we might scale these approaches to Petabyte data sets.

A Statistically Rigorous Approach to Astronomical Source Detection

David Friedenberg, Carnegie Mellon University

New high-resolution telescopes will be recording terabytes of image data every night. Fast and accurate detection of astronomical sources (galaxies, quasars, etc) in these images is an essential prerequisite to further scientific analysis of the data. We have developed and implemented a suite of tools for source detection that make probabilistic bounds on the rate of false sources in an image. We examine different testing criterion as well as a new multi-scale test statistic designed for this type of problem and show how these tools can be used on a variety of astronomy datasets. We advocate using algorithms that guarantee that with high probability the rate of false detections is below a preset bound. Several such algorithms will be presented and shown to be competitive with the current algorithms typically used in the field that do not have such explicit error controls. We will exemplify our procedures using data from the Chandra X-ray observatory and simulations from the Atacama Cosmology Telescope Team.

Using Local Likelihoods to Estimate Gravitational Lensing of the CMB

Ethan Anderes, University of California, Davis

This talk will present work on using local stationary approximations to estimate the local dependency structure in nonstationary random fields. We will focus on the application of estimating gravitational lensing of the Cosmic Microwave Background. We develop a weighted local likelihood estimate of the parameters that govern the local sheer of a gravitational distortion of the CMB. The advantage of this local likelihood estimate is that it smoothly downweights the influence of far away observations, works for irregular sampling locations, and when designed appropriately, can trade bias and variance for reducing estimation error.

Perspectives on Climate Change

Testing the Hypothesis of Anthropogenic Global Warming: A Continuing Controversy

S. Fred Singer, Science and Environmental Policy Project

The preferred test compares observed temperature trends with those derived from (greenhouse) climate models. I will discuss the statistical and other uncertainties of both sets of data.

Extracting Information from Large-Scale Computer Model Output

Mark Berliner, Ohio State University

Massive computer models are used in a variety of science and engineering applications. For example global atmospheric models have state spaces on the order of 10,000,000 variables. Earth system models, combining atmospheric, oceanic, cryospheric, and land surface process models, produce massive output. The scales of such models prohibit the production of many runs (ensembles), so establishing the statistical properties of their output is challenging. I review options for incorporating model output into Bayesian statistical analyses. I present two examples in the context of climate change analysis: (1) a simplified approach to detection and attribution of climate change, and (2) using multi-model ensembles in the projection of future climate.

Computational Social Science

Approaches to Inference from Link-Tracing Network Samples

Krista Gile, Oxford University

It is often the case that a population of interest is connected by a network of relations, and that it is beneficial to exploit this network in the sampling process. This typically involves a form of link-tracing sampling, in which subsequent sample units are selected from among the network neighbors of earlier samples. Although the various link-tracing sampling designs have much in common, the foundational assumptions and approaches of existing inferential strategies vary widely. Inference is also affected by the selection procedure for the initial sample, specifics of the link-tracing process, and other information available about the population. In this paper, we present a conceptual review of classical and recent approaches to inference from link-tracing network samples, highlighting the foundational assumptions required by the methods and their implications for inference. We review the current state of research and outstanding issues.

Novel Methods: Symbolic Data, Layered Graphics Anatomy, and Census Layered Graphics

Symbolic Data

Lynne Billard, University of Georgia

Classical data values are single points in p-dimensional space; symbolic data values are hypercubes (broadly defined) in p-dimensional space (and/or a Cartesian product of p distributions). While some datasets, be they small or large in size, naturally consist of symbolic data, many symbolic datasets result from the aggregation of large or extremely large classical datasets into smaller more manageably sized datasets, with the aggregation criteria typically grounded on basic scientific questions of interest. Unlike classical data, symbolic data have internal variation and structure which must be taken into account when analyzing the dataset.

Grammars of Graphics

Hadley Wickham, Rice University

This talk will summarize recent work in grammars of graphics. A grammar of graphics is an attempt to describe data visualization in terms of independent components that can be combined combinatorially: instead of relying on named graphics (like scatterplots and bar charts) you build custom charts using basic building blocks. I'll provide a brief history of the field, discussing Wilkinson's seminal work, and discuss why developing grammars for graphics is so important. I'll survey current attempts and also discuss efforts to extend the grammar to encompasses dynamic and interactive graphics.

OnTheMap Data and Tool

Jeremy Wu, Census Bureau

The U.S. Census Bureau released the first version of the OnTheMap tool in 2006. It has been updated annually three times to the current Version 4, covering 47 states with seven years of data. The user can select areas by defined layers or freehand to show where U.S. workers live and work with companion reports and resolution at the census block level. OnTheMap is available to the public 24/7 and free via the Internet, with the flexibility of importing and exporting files for the user's own custom analysis. The underlying OnTheMap data are synthesized from an emerging longitudinal national frame of jobs. The interface of OnTheMap data and tool provides easy visualization of complex data with unprecedented details while still protecting confidentiality.

Robust Methods in Regression

Robust parametric classification and variable selection with minimum distance estimation when $n \ll p$

Eric Chi, Rice University and Baylor College of Medicine

We present a robust solution to the classification and variable selection problem when the dimension of the data, or number of predictor variables, may greatly exceed the number of observations. When faced with the problem of classifying objects given many measured attributes of the objects, the goal is to build a model that makes the most accurate predictions using only the most meaningful subset of the available measurements. The introduction of L1 regularized model fitting has inspired many approaches that simultaneously do model fitting and variable selection. If parametric models are employed, the standard approach is some form of regularized maximum likelihood estimation. This is an asymptotically efficient procedure under very general conditions - provided that the model is specified correctly. Correctly specifying a model, however, is not trivial. Even a few outliers among data drawn from an otherwise pure sample of data can result in a very poor model. In contrast, minimizing the integrated square error, while less efficient, proves to be robust to a fair amount of contamination. We propose to fit logistic models using this alternative criterion to address the possibility of model misspecification. The resulting method may be considered a robust variant of regularized maximum likelihood methods for high dimensional data.

Contributed Session 1: Modeling

Recursive Modeling Using xstatR

E. James Harner and Jun Tan, West Virginia University

A large class of statistical models conceptually can be represented in a tree with nodes initially representing prototypes (classes). Inheritance is based on analogy rather than specialization. For example, the nonlinear and generalized linear regression prototypes both inherit from the linear regression prototype even though they are supersets of the latter. This analogy extends to many other prototypes, e.g., prototypes for linear additive models, generalized additive models, linear mixed models, nonlinear mixed models, and semi-parametric models. Model fitting, parameter estimation, observation diagnostics, model validation, etc. are common to all models and form the basis of the inheritance relationships. The tree is represented graphically and model instances are displayed as new nodes spun off from prototypes or from other model instances. Each tree is associated with a dataset and model instances are actually dataset mixings, which can also be displayed graphically.

Model instances are called virtual datasets since they are a mixture of the original dataset and the results of applying the model to the data. Clicking on a model instance node, i.e., selecting a model instance, updates the original dataset to include the derived variables for that node as well as the original variables. This allows model attributes to be dynamically organized so that only those derived variables associated with the selected model are displayed at any given time. Virtual datasets can be treated just like datasets, i.e., diagnostic plots can be generated and further analyses can be done on either the original or derived variables. Virtual dataset also have other attributes, i.e., slots, associated with the model fit. Plots are dynamically linked using the observer/observable design pattern. Thus it is possible to brush or paint the observations, or change their state information, in one plot and see the corresponding change in all other plots.

Once model instances have been created, it is possible to performs operations on selected subtrees. For example, cross validation could be recursively run on all models in a subtree to determine the best performing model. These recursive operations can be done for diagnostics, plots, etc. and provide a powerful mechanism for comparing models or selecting the optimal model. Modeling and dynamic plotting are done in xstatR, a Lisp package for XLISP-STAT. Model computations are done by the statistical computing environment R, which is dynamically linked to XLISP-STAT.

Bayesian Analysis for Exponential Random Graph Models Using the Double Metropolis-Hastings Sampler

Ick Hoon Jin and Faming Liang, Texas A&M University

Social network analysis has received much attention in the recent literature. In this paper, we consider a fully Bayesian analysis for exponential random graph models using the double Metropolis-Hastings sampler, which resolves the intractable normalizing constant problem encountered in Metropolis-Hastings simulations by including an auxiliary variable in its proposal distribution. Since the Gibbs sampler can mix very poorly in simulating social networks, we suggest a sequential parallel tempering algorithm, which partially decomposes the dependence structure of social networks and thus can be much more efficient than the Gibbs sampler in terms of autocorrelation of the resulting samples. Our method is illustrated using the Florentine business network, Kafterer's Taylor shop network, and a high school students friendship network. The results indicate that our method can significantly outperform other social network estimation methods, such as the Markov chain Monte Carlo maximum likelihood estimation (MCMCMLE) method and the exchange algorithm used in Caimo and Friel(2010).

Rainbow Plots, Bagplots, and Boxplots for Functional Data

Han Lin Shang and Rob J. Hyndman, Monash University

We propose new tools for visualizing large amounts of functional data in the form of smooth curves. The proposed tools include functional versions of the bagplot and boxplot, which make use of the first two robust principal component scores, Tukey's data depth and highest density regions. By-products of our graphical displays are outlier detection methods for functional data. We compare these new outlier detection methods with existing methods for detecting outliers in functional data, and show that our methods are better able to identify outliers. The computer code and datasets are available in the rainbow package in R.

Computational Statistics Methods for Social Issues

Designing Statistical Tools with Agent-Based Models in Mind

Ben Klemens, Census Bureau

Tools for statistical modeling and agent-based modeling (ABM) are basically distinct, for reasons practical, social, and sometimes entirely arbitrary. This talk presents some commonalities to all types of model, and one approach to writing statistical software applicable to ABMs. The goal is quantitative descriptions of ABM outputs, such as their distributions and the confidence with which hypotheses can be accepted or rejected.

Predicting Edges and Vertices in a Network

Walid K. Sharabati, Purdue University, Edward J. Wegman and Yasmin H. Said, George Mason University

This paper addresses missing edges and vertices in a network. We discuss interchangeability and duality between vertices and edges in a graph. We use covariate information associated with vertices to estimate the probability of missing edges; likewise, we use covariate information associated with edges to estimate the probability of missing vertices. In order to predict missing vertices, we apply the line graph transformation, which converts edges to vertices and vertices to edges. The probability of an edge is obtained by taking the inner product of the vectors of covariates. Moreover, we have extended the methodology of predicting two edges (dyadic ties) to predict edges in a triad. The method is based on geometry and fuzzy logic.

Healthcare utilization among border Hispanic seniors with diabetes: frequencies and related factors

Xiaohui "Sophie" Wang, University of Texas, Pan American

This study focus on determining personal and social correlates to health care utilization among border Hispanic seniors with diabetes. Access to healthcare is important for managing diabetes; however, little is known about predictors of healthcare utilization among minorities. A community assessment survey ($n = 249$) was conducted. Descriptive and multiple regression analyses were applied. Recruitment settings included a clinic, senior centers, and colonias. We found that older Hispanics residing in a colonia and uninsured had less access to healthcare than their counterparts. Significant correlates to physician utilization were gender, physician visits in Mexico, nativity, insurance, physician fees, obesity, colonia residency, and marital status. Significant correlates to emergency room visits (ER) were age, insurance, heart attack history, and retinopathy. Eye exams were associated with marital status, living situation, insurance, cholesterol level, and diabetes education. Therefore, border older Hispanics may be relying on

community clinics and ERs to treat their diabetes. Public health policies are needed to promote diabetes self-management among this population.

Computationally Intensive Dynamic Models

Simplified Trans-Dimensional Model Jumping with MCMC for Complicated Models

Dave Campbell, Simon Fraser University

Computing the posterior probability of a model from a set of models, MCMC must be able to visit all potential models. Sampling values from different models by Reversible Jump MCMC (RJMCMC) requires altering the dimension of the Markov chain. The main difficulty with RJMCMC is determining a transformation that changes the parameter dimension and proposes parameters into an informative location of a new model. When the models in question are non-nested differential equation systems, this is further complicated by the variety of dynamics that can be produced by a model. I propose a new algorithm to improve the model changing and dimension altering process by stepping through the function spaces of the dynamic systems. The proposed parameter dimension is augmented or reduced through a direction orthogonal to the function space of the previous model but in the space of the proposed model.

Emulating the Nonlinear Matter Power Spectrum for the Universe

Earl Lawrence, Dave Higdon, Katrin Heitmann, Salman Habib, Martin White, and Christian Wagner, Los Alamos National Laboratory

Many of the most exciting questions in astrophysics and cosmology, including most observational probes of dark energy, rely on an understanding of the nonlinear regime of structure formation. In order to fully exploit the information available from this regime and to extract cosmological constraints accurate theoretical predictions are needed. Currently such predictions can only be obtained from costly, precision numerical simulations. This work is aimed at constructing an accurate calibration of the nonlinear mass power spectrum on Mpc scales for a wide range of currently viable cosmological models, including dark energy. We use the Coyote Universe simulation suite which comprises nearly 1,000 N-body simulations at different force and mass resolutions, spanning 38 w-CDM cosmologies. This large simulation suite enables us to construct a prediction scheme for the nonlinear matter power spectrum accurate at the percent level for large wave numbers. We present this scheme and discuss the tests we have done to ensure its accuracy, and discuss how it can be extended to a wider range of cosmological models.

Estimating Time-Varying Parameters in ODEs

Jiguo Cao, Simon Fraser University

Dynamic models, usually written in forms of differential equations (DEs), describe the rate of change of a process. They are widely used in medicine, engineering, ecology and a host of other applications. One central and difficult problem is how to estimate DE parameters from noisy data. Ramsay et al. (2007) proposed the generalized profiling method to solve this problem. DE solutions are approximated by nonparametric functions, which are estimated by penalized smoothing with DE-defined penalty. The computation is much faster than other methods. We have extended the generalized profiling method to estimate time-varying parameters. A roughness penalty term is included to control the smoothness of the time-varying parameters. Simulations show that this method provides better estimates than the two-stage estimation strategy. This method will be demonstrated by estimating two time-varying parameters in an HIV dynamic model. This is joint work with Jianhua Huang, Texas A&M University, and Hulin Wu, Rochester University.

New Developments in Statistical Data Integration

Statistical Integration of Ranked Lists by Means of Data Streams

Michael G. Schimek, Danube University Krems, Austria and Medical University of Graz, Austria

In various application areas, we are confronted with ranked lists representing the same set of distinct objects. Under the assumption of a general decrease of the probability for consensus rankings with increasing distance from the top rank position, we are interested in such objects that highly conform in their rankings across the lists. Here, we model data streams representing the discordance of objects with respect to their rank position in such lists. For inference on the degradation of information and for the statistical integration of the top ranked objects, data stream information is sufficient. As a consequence, the computational demand is reduced to an extent that even long ranked lists can be processed. Based on an inference procedure (Hall and Schimek, 2010) conforming top lists can be identified. Finally, the objects of these partial lists are integrated by graphical means. We introduce a new type of plot and illustrate the whole procedure on various data.

Application of an Integrative Analysis using Association Networks of Biological Data

Karl G. Kugler, UMIT and Danube University, Laurin A. Mueller, Armin Graber, and Matthias Dehmer, Danube University

Meta-analysis of biomedical data has become a standard approach in bioinformatics. The broad abundance of gene expression data within public repositories enables researches to make use of a vast amount of information. The combination and integration of this data in a systematic manner allows us to discover new insights into various diseases. Another current trend in computational biology is the application of network based analysis. The network approach makes it feasible to overcome the static nature of investigating single detached features. Instead, it is then possible to represent the dynamic and complex nature of underlying processes and connections. For our analysis, we utilize correlation measures to infer association networks. In the presented work, we combine classical meta-analysis with quantitative network analysis methods in an integrative approach. To demonstrate these methods we make use of a set of seven prostate cancer studies. The results hint at systematic differences between these data sets. However, these differences are not only measurable between the networks, but are also present in the newly inferred common network representations. We present current findings from our integrative network analysis approach and illustrate its application on a biomedical data set.

Consensus Finding, Exponential Models, and Infinite Rankings

Marina Meila, University of Washington

This talk is concerned with summarizing -- by means of statistical models -- of data that expresses preferences. This data is typically a set of rankings of n items by a panel of experts; the simplest summary is the "consensus ranking", or the "centroid" of the set of rankings. Such problems appear in many tasks, ranging from combining voter preferences to boosting of search engines. We study the problem in its more general form of estimating a parametric model known as the Generalized Mallows (GM) model. I will present an exact estimation algorithm, non-polynomial in theory, but extremely effective in comparison with existing algorithms. Then we introduce the infinite GM model, corresponding to "rankings" over an infinite set of items, and show that this model is both elegant and of practical significance. Finally, the talk will touch upon the subject of multimodal distributions and clustering. Joint work with: Alnur Ali, Harr Chen, Bhushan Mandhani, Le Bao, Kapil Phadnis, Arthur Patterson and Jeff Bilmes

Bayesian Data Mining and Clustering

Hiding from Decision Tree Detectors

David W. Scott, Rice University

While richer feature spaces allow for improved detection, overly simplistic rules can be exploited. In this talk, some basic ideas about feature spaces are explored and vulnerabilities described.

Multivariate Bayesian Logistic Regression for Analysis of Clinical Trial Safety Issues

William DuMouchel, Phase Forward Lincoln Safety Group

This paper describes a method for a model-based analysis of clinical safety data called multivariate Bayesian logistic regression (MBLR). Parallel logistic regression models are fit to a set of medically related issues, or response variables, and MBLR allows information from the different issues to "borrow strength" from each other. The method is especially suited to sparse response data, as often occurs when fine-grained adverse events are collected from subjects in trials sized more for efficacy than for safety investigations. A combined analysis of data from multiple trials can be performed and the method enables a search for vulnerable subgroups based on the covariates in the regression model. An example involving 10 medically related issues from a pool of 8 trials is presented, as well as simulations showing distributional properties of the method.

A Graph Based Method for Estimating the Cluster Tree of a Density

Werner Steutzle, University of Washington

Clustering problems occur in may domains, from genomics and astronomy to document analysis and marketing. The general goal is to identify distinct groups in a collection of objects. To cast clustering as a statistical problem we regard the feature vectors characterizing the objects as a sample from some unknown probability density. The premise of nonparametric clustering is that groups correspond to modes of this density. I will introduce the cluster tree of a density as a summary statistic reflecting the group structure, and I will present a graph-based method for estimating the cluster tree. This is joint work with Rebecca Nugent, CMU.

Novel Methods: Internet Data and Targeted Marketing, Seriation and Pattern Discovery, Plus Collaboration and Social Networks

Collaborative Seriation among Disciplines and People

Innar Liiv, Tallinn University of Technology, Estonia

Seriation is an unsupervised data mining technique to reorder objects into a sequence along a one-dimensional continuum to make sense of the whole series. With clustering, objects are assigned to groups, but with seriation, objects are assigned to a position within a sequence Seriation has been applied to a variety of disciplines including archaeology and anthropology; cartography, graphics and information visualization; sociology and sociometry; psychology and psychometrics; ecology; biology and bioinformatics; cellular manufacturing, and operations research. Interestingly, across those different disciplines, there are several commonly emerging similar structural patterns. In this talk, we will present a prototype to support cross-discipline collaboration among people to discover patterns from the data and accumulate knowledge.

Suggesting Enrichment of Social Network Analysis beyond Graphs

Arnold Goodman, Collaborative Data Solutions

Social networks are social structures of individuals or organizations which are connected by their interdependencies. Collaboration and value creation may be viewed as not only productive means of increasing the success and value of social networks, but also themselves generating social networks. How might we support social networks in increasing the likelihood of their success and subsequent value of their success? We may support the planning and managing of any activities they undertake, as well as then evaluating the performance of those activities and coaching network members in improving their performance in the future. To advance progress in this direction, we introduce model structures for collaboration, value creation and their likely interactions. The structures may also be used as checklists of what to do and how to do it. In addition, these checklists may be transformed into scorecards showing relative importance. Steps which likely lead to collaboration are conversation within the network, combination of efforts, coordination of objectives as well, cooperation on behaviors as well, and ideal collaboration on attitudes as well. This ideal collaboration may be characterized by members' community of relationships, drive to participate, motivation to commit, trust of each other and wisdom to transcend. The likely success factors for collaboration are communicating the meaning, connecting the members, contributing the results, comparing the alternatives, and challenging the status quo within needed situations. Basic stages for value creation are defining needs, specifying resources, designing tools, producing results, creating value, and advocating needed change. Change may be needed for enterprise-wide, mission-critical or paradigm-shifting situations within business, government and science. Social networks typically solve problems, make products or participate in projects. Finally, we introduce a novel structure to support increasing the success and value of social networks. It is based on our process model for the cell's protein cycle, from DNA through RNAs to proteins and beyond. This model and its exploration for uncertainty were envisioned by Paul Silverman, who directed first United States genome laboratory at Lawrence Berkeley Laboratory. We co-developed the model with Silverman until he passed away, and have led both model development and its exploration for uncertainty ever since. His vision is yet to be accepted by cell biologists.

Separating the Wheat from the Chaff - Feature Selection in High-Dimensional Regression

Feature Selection with Constrained L1 Regularization

Leming Qu, Boise State University

Regularized regression with the L1 penalty, known as LASSO, is a popular approach for feature selection and coefficient estimation. In some cases, constraints on regression coefficients are available. For example, in hyperspectral unmixing problem, the regression coefficients are all nonnegative and summing to one, yet only a few coefficients are positive. Methods incorporating these constraints into L1 regularization would yield better selected model than those ignoring them. A fast algorithm is proposed for L1 regularization subject to linear constraints (including both box and linear equality constraints) on the regression coefficients. The algorithm iteratively solves a subproblem: a constrained L1 regularized denoising problem. With only box constraints, the subproblem is easily solved by a closed form formula. With both box and linear equality constraints, an iterative algorithm will be used to solve the subproblem. The basic operation of the algorithm is matrix vector multiplication, hence it can handle high-dimensional problems. A simulation study illustrates the advantages of constrained L1 regularization relative to those ignoring constraints.

Bootstrap Inference for Network Construction

Pei Wang, Fred Hutchinson Cancer Research Center

We are interested in constructing genetic networks based on high dimensional microarray data through regularized regressions. This is essentially a variable selection problem. In the literature, variable selection is usually achieved as a result of model selection. Many techniques, such as AIC, BIC, and cross validation, have been developed for model/variable selection based on prediction errors or likelihood scores. However, since the goal of network construction is to identify edges between variables (nodes) rather than predicting an outcome, models that have optimal prediction errors or likelihood scores may not give the best variable (here, edge) selection results. In this talk, we will introduce a new inference procedure which aims at selecting edges directly by controlling the false positive rate. The idea is to repeatedly estimate the network based on perturbed data through bootstrapping, and then examine the distribution of the frequency of each edge being selected. As expected, the frequency distribution of true edges being selected would be different from that of false edges. Thus by fitting the overall frequency distribution with a mixture model, we are able to estimate the proportion of false edges at each frequency cutoff. We illustrate the performance of proposed method on both simulated and real data sets.

Boosting for Nonparametric High-dimensional Models

Lifeng Wang, Michigan State University

In regression analysis, variables can often be combined into groups based on prior knowledge. Such a group structure of the predictor variables can be effectively utilized in regression analysis in order to improve identification of relevant groups of variables and to improve the prediction performance. In this paper, we propose a boosting method to perform nonparametric regression and feature selection for high-dimensional group additive models. We investigate the learning theory for the proposed boosting algorithm, and illustrate its finite sample performance via both simulated and real data.

All at See: Snapshots of Modern Visualization Research

An Object-Oriented Approach in R for the Visualization of Functional Actigraphy Data

Juergen Symanzik, Abbass Sharif, Utah State University and William D. Shannon, Washington University School of Medicine

Actigraphy, a technology for measuring a patient's overall activity level almost continuously over time, has gained a lot of momentum over the last few years. An actigraph, a watch-like device that can be attached to the wrist or ankle of a patient, uses an accelerometer to measure human movement every minute or even every 15 seconds. Actigraphy data is often treated as functional data. In this talk, we will present a prototype of our object-oriented approach in R for the visualization of functional actigraphy data. We will demonstrate possible user interfaces (from the Web via an R package called Rpad and from Excel via another R package called Rcmdr), we will outline our object-oriented approach using the R.oo package, and we will demonstrate some possible applications of our work.

Visualization and Statistical Modeling

Adalbert Wilhelm, Jacobs University, Bremen, Germany

The development and availability of efficient statistical software has also led to a wider distribution, a wider use and a broader presence of both exploratory statistical graphics and sophisticated, complex modeling approaches. However, there are only a few publications that link

the two fields together as, for example; Gelman (2004) does with a particular focus on Bayesian analysis. The graphical representation of data quite often requires the corresponding statistical modeling phase to yield practically significant results. On the other hand there are plenty of examples such as the famous Anscombe quartet based on four different data sets but resulting in one identical regression model, that show that statistical modeling without a corresponding visualization misses out on some fundamental features. It seems obvious that the one can't be reasonable performed without the other. Doing statistical modeling without a proper graphical representation of data and model is risky and problematic. Exploring the data graphically without the attempt to model them properly usually falls short and leaves the analyst with isolated insights and anecdotes. The systematic approach of modeling combined with the flexible use of exploratory graphics combines the strengths of both fields and constitutes a powerful research tool. This paper will illustrate this by providing an eclectic tour through the modeling process and illustrating the potential applications of exploratory graphics in the various steps. We will focus on three main stages of modeling: visualization prior to the modeling to check data quality and model adequacy, during the modeling process to check for model assumptions and model quality and after the modeling process to enhance interpretation of the modeling parameters as well as comparing between competing models.

References:

Andrew Gelman. Exploratory data analysis for complex models (with Discussion by Andreas Buja and Rejoinder). Journal of Computational and Graphical Statistics, 13(4):755-779, 2004.

Patrick J.F. Groenen and A.J. Koning. A new model for visualizing interactions in analysis of variance. Econometric Institute Report No EI 2004-06, Erasmus University Rotterdam, Econometric Institute" http://econpapers.repec.org/RePEc:dgr:eureir:1765001189" 2004.

Jonathan Kastellec and Eduardo Leoni. Using graphs instead of tables in political science. Perspectives on Politics, 5(4):755-771, 2007.

Adalbert F. X. Wilhelm. Interactive statistical graphics: The paradigm of linked views. In C.R. Rao, E.J. Wegman, and J. Solka, editors, Handbook of Statistics, volume 24, pages 437-537. Elsevier, 2005.

Every Plot Must Tell a Story - Even in R

Heike Hofmann, Dianne Cook, Iowa State University and Antony Unwin, Augsburg University, Germany

R makes it easy to distribute datasets and up-to-date statistical methodology. Vignettes and help files make use of real data sets. This comes with an obligation - examples have to adhere to the same standards of good practice that we would expect from other publication channels. We will give an overview of some selected example data sets in currently available R packages. Given the power of R's graphical tools and how easy it is to draw graphics, there is no need to restrict ourselves to a single all-encompassing display for a dataset. Instead, we recommend a set of graphics to ensure that most aspects of the dataset are presented.

iPlots eXtreme - Next Generation of Interactive Graphics for Analysis of Large Data

Simon Urbanek, AT&T Laboratories

Interactive graphics provide a very important tool in real-world applied statistics. Although various interactive graphics software exists they are often not scalable to large data that we encounter today. The two major constraints are the fact that interactive graphics needs to respond very quickly to be truly interactive and that new ways of displaying data need to be considered. In this talk we present a new interactive graphics system iPlots eXtreme that was designed specifically

to address the issues of scalability and flexibility in prototyping new visual methods. It leverages the power of modern graphics processing units via OpenGL for interactive visualization and is aggressively optimized to perform well on large data. Is features a consistent user interface to flatten the learning curve. It also embeds seamlessly into the R system for statistical computing to provide extensibility, ease of use in analytical workflow, prototyping for new plots but also to give R a very fast graphics device. We will explain the design of the system, highlight the key advances and comparison to existing interactive graphics and illustrate its scalability and use on several practical examples. We will also showcase its extensibility in conjunction with R and its general role as a very fast interactive graphics system.

Policy Issues on Climate Change

Global Warming: Nexus of Politics, Economics and Science

Jeff Kueter, President, George C. Marshall Institute

The United States Congress is actively considering legislation to cap greenhouse gas emissions. Independently, the Environmental Protection Agency is moving to impose regulations on emissions as well. Pursuit of an international agreement to limit emissions continues. The belief that anthropogenic activities are negatively transforming the Earth's climate motivates each of these efforts. Debate over the certainty of that conclusion as well as the economic cost and consequences of proposed mitigation efforts is generating opposition to these legislative, regulatory and international efforts. The presentation will review the economic and scientific aspects of the ongoing public policy debate.

Global Warming, Fact, Fiction and Fraud

Don Easterbrook, Western Washington University

The global warming debate is filled with facts, fiction, and fraud. The facts are that (1) the Earth has experienced natural global warming and cooling 4 times in the past century, 40 times in the past 500 years, and 60 times in the past 5000 years, long before CO_2 could possibly have been a factor, (2) at least 10 warm/cool climate fluctuations between 10,000 and 15,000 years ago were far more intense than recent warming, including warming of 15°F in 40 years, (3) from 1945 to 1977, while CO_2 was soaring, we had 30 years of global cooling, (4) although we've had global warming (1977 to 1999), Antarctic ice is not melting, (5) nothing that humans are doing can significantly affect global climate. The fiction is that (1) CO_2 is capable of producing warming of the atmosphere 10°F by the end of the century, (2) sea level will rise 20 feet this century, (3) global warming is causing extinction of polar bears, (4) carbon cap and trade will reduce atmospheric CO_2, (5) carbon cap and trade will affect global warming. The fraud is (1) faking data, (2) changing climate data to make it appear warmer, (3) lying about Himalayan glacier retreat, (4) deliberate suppression of data that doesn't support CO_2 as the cause of global warming.

Climate Change Policy and the Climategate Scandal

Yasmin H. Said, George Mason University

The release of emails from the East Anglia University Climate Research Unit just before the Copenhagen Climate summit has had a damaging effect on public support for action on global warming. The lack of transparency by some climate researchers, the willingness to bend the peer review process, and the willingness to destroy data rather than share it with researchers of a different perspective all raise fundamental issues of climate change policy. Perhaps the best thing to come from the climategate scandal is the formal recommendation of engaging

statisticians. In this talk I will discuss some of the implications of climategate on climate change policy.

Computational and Statistical Issues in ABMs

Simplifying Complex Systems into Multi-level Agent Based Models

Rainer Hilscher, Altarum Institute

Every mathematical model is necessarily an abstraction of reality but different modeling approaches have different simplification assumptions. Agent based modeling simplifies in a way that sets it apart from most traditional approaches to simplification. With ABM the focus is not on reducing reality to a tractable number of system variables but on reducing reality to a set of individual level decision making processes and variables that these processes may update. This shift in focus results in a bottom up capturing of the core processes that individual employ to make decisions rather than in a top down state variable description of the system of investigation. Multi-level agent based models push the process modeling method even further by integrating interactions within humans (e.g. metabolic, neuro-biological), interactions between humans (social networks) and interactions between humans and their physical environment (e.g. built environment). This presentation will first provide a brief overview of agent based models that highlight the process focused modeling method. It will then introduce the notion of a multi-level agent based model and the underlying ecological approach to complex systems. Our own work is in the domain of modeling the obesity epidemic from an ecological community health perspective using a multi-level agent based model. The main part of this presentation will provide a detailed overview of the different interaction levels we are modeling and the different decision making theories that we are building into our agents. One of the most important research areas in multilevel ABMs is the nature of multi-dimensional feedback loops between these various levels. These feedback loops are addressed in the last part of the presentation.

The Role of Population Heterogeneity and Human Mobility in the Spread of Pandemic Influenza

Stefano Merler, Piero Manfredi, and Marco Ajelli, Fondazione Bruno Kessler

Little is known on how different levels of population heterogeneity and different patterns of human mobility affect the course of pandemic influenza and, more in general, of epidemics. Thanks to a highly detailed model of the European populations and of their movements, based on specific sociodemographic, air and railway transportation data, a large-scale spatially-explicit individual-based model has been developed and parametrized. This allows providing quantitative measures of the influence of such factors at the European scale. Our results show that Europe has to be prepared to face a rapid diffusion of a pandemic influenza, because of the high mobility of the population, resulting in the early importation of the first cases from abroad and highly synchronized local epidemics. The impact of the epidemic in the European countries is highly variable because of marked differences in the sociodemographic structure of the European populations. Final epidemic size, basic reproductive number and peak day incidence depend heavily on sociodemographic parameters, such as the size of household groups and the fraction of workers and students in the population.

Graphical Methods for Classification Based on Dimension Reduction

SMVCIR Dimensionality Test

Charles Lindsey, Texas A&M University

The sliced mean variance-covariance inverse regression (SMVCIR) algorithm takes grouped multivariate data as input and transforms it to a new space where the group mean, variance, and covariance differences are more apparent. A dimensionality test is developed for SMVCIR, telling us how many coordinates the new discrimination space needs. Simulations are shown to verify accuracy of the test. An example is provided, contrasting SMVCIR with sliced average variance estimation (SAVE) and sliced inverse regression (SIR) on a handwriting dataset used in machine learning.

Robust Dimensional Reduction via Invariant Coordinate Selection

David E. Tyler, Rutgers, The State University of New Jersey

In this talk, a general dimension reduction method recently introduced in Tyler et al. (2009) is discussed and is shown to have applications to cluster analysis as well as to independent components analysis (ICA). The general method is based upon the comparison of any two different robust estimates of a multivariate scatter matrix . In particular, a new coordinate system is obtain from the eigenvectors associated with the eigenvalue-eigenvector decomposition of one estimate of scatter relative to another. An important property of this decomposition is that the corresponding eigenvectors generate an affine invariant coordinate system (ICS) for the multivariate data. By plotting the data with respect to this new invariant coordinate system, various data structures can be revealed. For example, when the data arises from a mixture of elliptical distributions, a subset of the invariant coordinates correspond to Fisher's linear discriminant subspace, even though the class identification of the data points are unknown. As another example, under certain ICA models, the invariant coordinates correspond to the independent components. Finally, in the cluster analysis setting it is noted that the classical ACECLUS methods due to Art, Gnanadesikan, and Kettenring (1982) can be viewed as a special case of the ICS method, and in the ICA setting it is noted that the classical FOBI algorithm due to Cardoso (1989) can also be viewed as a special case of the ICS method.

References:

Art, D., Gnanadesikan, R. and Kettenring, J. R. (1982). Data-based metrics for cluster analysis. Util. Math. A, 21, pp. 75–99.

Cardoso, J.-F. (1989) Source separation using higher order moments. In: Proc. Int. Conf. Acoustics, Speech and Signal Processing, pp. 2109–2112. Glasgow: Institute of Electrical and Electronics Engineers.

Tyler, D.E., Critchley, F., Duembgen, L. and Oja, H. (2009). Invariant co-ordinate selection (with discussion). J. R. Statist. Soc. B, 71, pp. 549–592.

Sufficient Dimension Reduction Based on Normal and its Connection with Principal Components

Liliana Forzani, Universidad Nacional del Litoral, Argentina

Using as a starting point the hypothesis of normality of $X|(Y=y)$, we developed a methodology for dimension reduction for the forward regression of $Y|X$ under the sufficient dimension reduction paradigm. This methodology includes finding the dimension of the central subspace,

maximum likelihood estimation for a basis of the central subspace, testing for predictors and prediction.

Non-English Text Data Mining Via the Vector Space Model

Computing Within the Foreign Vector Space Framework

Nick Tucey, NSWCDD

This talk will discuss how to expand a text data mining application to allow for the processing of documents in foreign languages, such as Arabic, Chinese, Russian, etc. Many text analysis algorithms, such as hierarchical clustering, latent semantic indexing (LSI), and dimensionality reduction, utilize a vector space encoding of the document collection, and thus are language independent. These techniques, however, are all dependent on the tokenization of multilingual documents. This talk will explore topics such as language identification, word segmentation, indexing, and document display. In order to develop this methodology necessary and sufficient conditions for each method are studied. After that, maximum likelihood estimators for the parameters are found. This allows us to develop inference procedures that are shown to work in simulation studies. The models are applied to get conclusions in real data set examples. Finally, its connection with principal components is given.

Implicit Translation: An experiment in English and Farsi

David Marchette, NSWCDD

Implicit translation can be defined as matching a document in one language with the correct paired document in another. The idea is that the paired document may be a translation, or it may simply be the document closest in topic and meaning from a in a given corpus of documents. I will discuss a method based on embedding the documents from both languages into a single space. This approach utilizes the inter document dissimilarity matrices and is thus independent of the actual text processing methodology utilized to define the dissimilarities. I will illustrate the approach using a subset of the English and Farsi (Persian) Wikipediae, and compare various variants of the algorithm. Simulations from a model of word count histograms will also be provided to illustrate the properties of the approach.

Contributed Session 2: Inference

Model-Averaged L1-Penalized Logistic Regression

Mark Seligman and Chris Fraley, Insilicos LLC

We discuss recent work in which the leaps-and-bounds subset selection algorithm of Bayesian model averaging (BMA) for logistic regression is replaced by subsets chosen via L1-penalization. This permits BMA to be applied efficiently to much larger predictor sets, with predictive quality that compares favorably with results obtained using iterative BMA.

Regression Tree Boosting to Adjust Health Care Cost Predictions for Diagnostic Mix

John W. Robinson, Statistical and Health Informatics Consulting

Background: Systems for risk-adjusting health care cost, described in the literature, have consistently employed deterministic models to account for interactions among diagnostic groups, simplifying their statistical representation, but sacrificing potentially useful information. An alternative is to use a statistical learning algorithm such as regression tree boosting that

systematically searches the data for consequential interactions, which it automatically incorporates into a risk-adjustment model that is customized to the population under study.

Methods: Using administrative data for over 2 million members of indemnity, preferred provider organization (PPO), and point-of-service (POS) plans, AHRQ's Clinical Classification Software (CCS) was applied to sort diagnoses from year 2001 into 260 diagnosis categories (DCs). For each plan type (indemnity, PPO, and POS), boosted regression trees and main effects linear models were fitted to predict concurrent (year 2001) and prospective (year 2002) total health care cost per member, given DCs and demographic variables.

Results: Regression tree boosting explained 49.7-52.1 percent of concurrent cost variance and 15.2-17.7 percent of prospective cost variance in independent test samples. Corresponding results for main effects linear models were 42.5-47.6 percent and 14.2-16.6 percent.

Conclusions: The combination of regression tree boosting and a diagnostic grouping scheme, such as CCS, represents a competitive alternative to risk adjustment systems that use complex deterministic models to account for interactions among diagnostic groups.

Quantitative Horizon Scanning

Detecting Anomalies and Estimating Anomaly Characteristics in Time-Series of Graphs

Glen A. Coppersmith, The Johns Hopkins University

Detecting anomalies and change-points in graphs (communications graphs, social networks, coauthorship graphs, etc.) is a crucial component to quantitative horizon scanning. When actors from diverse fields begin collaborating, communicate more than they have previously, or shift the focus of their work, this may indicate research at the juncture between fields. Such research has been implicated in sizable advances like the fields of biology and optimization yielding genetic algorithms or physics and medicine yielding MRI techniques. We demonstrate results in theory, simulations, and real data (the Enron email corpus) for anomaly detection and anomaly estimation (e.g. the number of actors that changed their communication patterns or the amount their communication pattern has changed). Specifically, we investigate "chatter" type anomalies and change-points, where an unknown subset of actors communicate more frequently (by an unknown amount) at an unknown point in time.

Multi-feature Clustering and Visualization of Large Document Collections

Jeff Solka and Avory Bryant, NSWCDD

Sometimes one is interested in ascertaining important developments in a particular topic area over a period of time. Depending on the breadth of the topic area and the length of the period of time the size of the document set can exceed millions of records. One of the first steps in conducting such a study is the development of a taxonomy of the topic. This talk will discuss the use of multi-feature clustering along with multi-resolution visualization strategies as a means to successfully create such a taxonomy. The inner workings of this methodology will be illustrated using a large collection of PubMed documents.

Quantitative Horizon Scanning for Mitigating Technological Surprise

Avory Bryant, NSWCDD

Using online data repositories of journal publications and conference proceedings to represent global scientific research, we are interested in finding instances of emerging collaboration between disparate topics. Author graphs are used to represent co-authorship between a set of

authors present in some document collection. Given that each edge represents a document which belongs to some topic we can define an author a as being b-privy to some topic t if the number of edge hops to an author adjacent to an edge of type t equals b. We hope to empirically show that the probability of authors being b-privy to topics $t_1$and t_2 at year y_j given $(b+1)$-priviness at year $y_j - 1$ is greater than the associated unconditional probability. In other words it is our assumption that future scientific collaborations between disparate topics can be predicted by analyzing changes in the author graph over time.

Computational Statistics and Robotics

Issues and Approaches in on-line Modeling of Environments from 3D Data

Marshal Hebert, Carnegie Mellon University

Building models of environments involves distinguishing between different classes of objects and terrain in the static case, and differentiating between different types of moving objects and the static surroundings in the case of dynamic environments. In both cases, the fundamental problem is to build models of different classes from training data. I will focus on the problems associated with building models from 3D data. This type of data presents unique challenges in terms of its distribution, i.e., unstructured point cloud as opposed to regularly sampled signal in the case of images, in addition to the usual robustness issues with respect to in-class variation and observation conditions. I will describe different approaches that we have used in the context of on-line data from sensors on unmanned ground vehicles.

Partial Least Squares Applications in Computer Vision

Larry Davis and Aniruddha Kembhavi, University of Maryland

It is often the case in computer vision applications that visual entities are described using very rich and high dimensional representations based on multi-scale shape, texture and color. At the same time, the number of samples available to train a classifier is ordinarily orders of magnitude smaller than the size of the representations. This is an ideal setting for using partial least squares to build compact models for classification and matching. I will describe the use of partial least squares for a variety of computer vision problems including human detection, tracking, vehicle detection and face recognition.

Test Methods and Metrology for Evaluating Human Detection and Tracking Systems

Barry Bodt and Richard Camden, Army Research Laboratory, Harry Scott, Adam Jacoff, Tsai Hong, Tommy Chang, Rick Norcross, Tony Downs, and Ann Virts, National Institute of Standards and Technology

The Army Research Laboratory (ARL) Robotics Collaborative Technology Alliance (CTA) conducted an assessment and evaluation of multiple algorithms for real-time detection of pedestrians in Laser Detection and Ranging (LADAR) and video sensor data taken from a moving platform. The algorithms were developed by Robotics CTA members and then assessed in field experiments jointly conducted by the National Institute of Standards and Technology (NIST) and ARL. A robust, accurate and independent pedestrian tracking system using ultra wide-band tracking and real-time display were developed to provide ground truth and data analysis. The ground truth was used to assess the CTA members' algorithms for the performance of their detection and tracking results. In addition, measurements were taken to support comparative analysis of the tracking consistency and the recognition accuracy. The assessment scenarios included moving and articulated humans acting as targets for detection from a moving vehicle which was equipped with two pairs of stereo cameras, two image LADARs, one SICK and a navigation sensor in unstructured environments. The test procedures for the assessment

mainly focused on choreograph repeatable human movement scenarios relative to the speed of a vehicle and scene complexity. The test results are intended to support comparative analysis across treatment conditions and to help developers to advance their algorithms. The presentation will describe the measurement system in this challenging environment and provide some analysis results.

Advances in Machine Learning and Data Mining Technology Workshop

Dan Steinberg, Salford Systems

Part 1: Recent advances in machine learning technology make it possible to determine definitively whether interactions of any degree need to be included in a predictive model. We can thus establish conclusively, for example, for a given set of predictors, whether or not a model with interactions will outperform a model without them. Further, we can now identify precisely which interactions are supported by the data, and also the degree of interaction, even in very high-dimensional data. The tools we use to achieve these results are extensions of Stanford Professor Jerome Friedman's TreeNet. We illustrate the concepts in the context of a real world regression model where we are able quickly to identify all the important interactions with a modest number of boosted tree ensemble models.

Part 2: We will walk you through Leo Breiman's Random Forests and Jerome Friedman's TreeNet/MART (also known as TreeNet Stochastic Gradient Boosting). Random Forests and MART/TreeNet are new advances to classification and regression tree software, which enable the modeler to construct predictive models of extraordinary accuracy. Random Forest is a tree-based procedure that makes use of bootstrapping and random feature generation. In TreeNet, classification and regression models are built gradually through a potentially large collection of small trees, each of which improves on its predecessors through an error-correcting strategy. We will discuss theory, and what is novel in both RandomForests and MART/TreeNet, and show how to each of them to solve real-world data mining problems. Finally, we will compare the two methodologies and show where they fit in terms of other machine learning and data mining software.

JCGS Highlights

Understanding GPU Programming for Statistical Computation: Studies in Massively Parallel Massive Mixtures

Marc Suchard, Quanli Wang, Cliburn Chan, Jacob Frelinger, Andrew Cron and Mike West, Duke University

We describe advances in statistical computation for large-scale data analysis in structured Bayesian mixture models via GPU (graphics processing unit) programming. The developments are partly motivated by computational challenges arising in increasingly prevalent biological studies using high-throughput flow cytometry methods, generating many, very large data sets and requiring increasingly high-dimensional mixture models with large numbers of mixture components. This presentation describes the strategies and process for GPU computation in Bayesian simulation and optimization approaches, examples of the benefits of GPU implementations in terms of processing speed and scale-up in ability to analyze large data sets, while providing a detailed, tutorial-style exposition that will benefit readers interested in developing GPU-based approaches in other statistical models.

Pairwise Display of High-Dimensional Information via Eulerian Tours and Hamiltonian Decompositions

Wayne Oldford, University of Waterloo

Every statistical graphic is a construct of display elements laid out in some arrangement. Sometimes these arrangements require an ordering of their elements. For example, a parallel coordinate layout must order its parallel axes, a multivariate star glyph its radial axes. Whenever this is the case, the display items to be ordered can be represented as nodes in a formal graph. Edges between nodes appear if and only if the corresponding display items may follow one another in the display's layout. In many statistical graphics any item can follow any other (e.g. parallel coordinates) and so a complete graph results. Much is known about paths on complete graphs that is of immediate relevance to statistical display layout. Every ordering of the display items is a Hamiltonian path on the complete graph. A display that ensures that every item immediately follows every other item is equivalent to an Eulerian tour on the graph. If statistically meaningful weights can be attached to each edge, then the graph is a weighted graph and some paths may be preferred to others. We illustrate how these and other graph theoretic results can be put to good use by showing how several statistical displays can be improved. Examples include star glyphs, parallel coordinate plots, and multiple comparison plots. These new displays will be illustrated on data. (N.B. All methods and graphical displays are now available from the PairViz R package.) This is based on joint work with Catherine Hurley of the National University of Ireland, Maynooth.

Computer Models, Virtual Laboratories

Predictive Modeling of a Radiative Shock Physics

Derek Bingham, Simon Fraser University

Predictive science, in some circles, has become synonymous with the use of physics modeling, often realized in complex computer codes, and field data to forecast what would be observed in a physical process. In the statistics community, this endeavor falls in the area of computer experiments. This talk deals with the design and analysis of a series of computer and field experiments to form a predictive model for a radiative shock hydrodynamics system. Specific issues addressed are (i) the role of experiment design in efficient predictive model building and (ii) uncertainty quantification. The interplay between (i) and (ii) is discussed.

Posterior Exploration for Computationally Intensive Forward Models

Shane Reese, BYU, David Higdon, David Moulton and Jasper Vrugt, Los Alamos National Laboratory, and Colin Fox, University of Otago, New Zealand

While standard single-site Metropolis updating proves effective in a variety of applications, it has the drawback of requiring many calls to the simulation model. Here we compare two MCMC schemes simulation. We use highly multivariate updates to sample from the posterior: the multivariate random walk Metropolis algorithm and the distributed evolution-MCMC sampler. Such schemes are alluring for computationally demanding inverse problems since they have the potential to update many components at once, while requiring only a single evaluation of the simulator. We consider new formulations based on faster, approximate simulators created by altering the multi-grid solver used in the simulator.

Statistical Analysis of Regional Climate Model Ensembles

Steve Sain, NCAR

Atmosphere-ocean general circulation models (AOGCMs or, more simply, GCMs) have become an integral part of climate science and are key components in studies of global climate change. However, the spatial resolution of GCMs limit their use for impacts studies on regional and local scales. To overcome this, downscaling methods have been developed that generate local-scale climate information from the coarse-scale output from the GCMs. One class of such methods is referred to as dynamic downscaling and involves the use of high-resolution climate models, such as nested regional climate models (RCMs). The North American Regional Climate Change Assessment Program (NARCCAP) is generating a multi-model ensemble of regional climate models whose output is focused on North America. In this talk, I will present a statistical approach for combining the early runs from NARCCAP that leads to probabilistic projections based on the ensemble as well as a framework for intercomparisons between the individual models in the ensemble.

Refereed Session

ChiD, A χ^2-Based Discretization Algorithm

Ross Bettinger, Seattle, WA

We have developed a discretization algorithm, based on Kerber's ChiMerge and Liu and Setiono's Chi2, that automatically chooses the best set of cutpoints for dividing a continuous variable into a set of contiguous discrete intervals. The algorithm, ChiD, uses class information to perform supervised discretization based on maximizing the logworth of the significance of a statistic computed from adjacent intervals of the continuous variable being discretized. The ChiD algorithm generates cutpoints that match the quality of those computed by the Enterprise Miner Decision Tree algorithm as measured by the accuracy of classification models built using ChiD cutpoints versus original, undiscretized data.

A Bayesian Decision Theoretic Approach to Multiple Hypotheses Problems

Naveen K. Bansal, Marquette University and Klaus Miescke, University of Illinois, Chicago

A multiple hypothesis problem with directional alternatives is considered in a decision theoretic framework. Skewness in the alternatives is considered, and it is shown that this skewness permits the Bayes rules to possess certain advantages when one direction of the alternatives is more important or more probable than the other direction. Bayes rules subject to certain constrains on the directional false discovery rates are obtained, and their performances are compared with a traditional FDR rule through simulation.

Vulnerability of US Hospitals from Terrorist Attack: Pilot Study on Hospitals in California

Byeonghwa Park and Yasmin Said, George Mason University

A terrorist strategy could very well be a two-phase attack. First disable the medical infrastructure so that a second attack would leave individuals with minimal medical assistance. Attacking the largest, most capable hospitals is a natural way to achieve the first phase. Identifying hospitals most vulnerable to a terrorist attack is crucial for preventing terrorist attack and for preparing strategies or policy for response or evacuation systems for the hospital. In this paper, we identify vulnerable hospitals in California, leading to discover certain rules and apply discovered rules to detect vulnerable hospitals from the terrorist attack nationwide. Based on hospitals in California

as a test bed, exploratory data analysis and machine learning technique are used mutually to identify vulnerable hospitals and discover rules in order to apply the finding to other states.

Sampling and Inference for Hidden Networked Populations

Respondent-Driven Sampling: Realistic Models and Appropriately Conservative Variance Estimates or 'RDS For Skeptics'

W. Whipple Neely, University of Washington

We examine the problem of making conservative variance estimates and dealing with situations in which the basic assumptions of Respondent-Driven Sampling (RDS) fail. The current RDS statistical theory can be derived from a trio of assumptions: (1) respondents' self-reported personal network size can serve as a proxy for an unknown sample inclusion probability, (2) any dependence between observations is completely explained by a homogeneous first order Markov model, (3) inclusion probabilities are conditionally independent given any outcome of interest. In this paper we examine how treating (1) and (2) as exact representations of human behavior leads to a model that cannot account for variance due to individual-level variation amongst respondents. We then show how recasting assumptions (1) and (2) as descriptions of mean modeled behaviors allows us to account for such variability and can lead to more conservative variance estimates. In the process we introduce a class of dynamic models that can be used to model the RDS referral process, and a sensitivity analysis that can be used to assess the impact of deviations from of assumptions (1) and (2). This work can serve as an introduction to the mathematical statistics of the classical RDS estimators and as an introduction to the problem of making conservative inferences based on data collected using RDS.

Inference for Hidden Populations Based on Network Sampling

Krista J. Gile, Nuffield College, Oxford University

Survey sampling in hidden populations is complicated by the lack of a practical sampling frame. If the target population is connected by an underlying network of social relations, a link-tracing network sampling design can often be employed to collect a sizeable sample. Respondent-Driven Sampling is a widely-used type of link-tracing sampling, which is often effective in recruiting large and diverse samples from hidden populations.

Current estimation relies on sampling weights estimated by treating the sampling process as a random walk on the underlying network of social relations. These estimates are based on strong assumptions allowing the data to be treated as a probability sample. In particular, existing estimators assume a with-replacement sample or small sample fraction, and ignore bias induced by the initial convenience sample. We illustrate the impact of violations of these assumptions, and introduce two new estimators. The first uses a without-replacement approximation to the sampling process model, and corrects for large sample fractions. The second uses a model-assisted approach and leverages the fitting of a parametric model for the social network to correct for both large sample fractions and biases introduced by the initial convenience sample.

Recent Developments in Network Sampling

Steven K. Thompson, Simon Fraser University

Hidden human populations often can be reached only through network sampling methods which obtain the sample by following links from one member of the hidden population to another. Recent developments in network sampling designs and inference methods for such situations provide increased flexibility for selecting the sample of people from the hidden population. For inference from the sample to the wider hidden population and its network structure a choice of

design based and model based methods are available. Most of these methods are computational intensive.

Interfacing Text Mining and Image Analysis

Combining Text and Image Processing in an Automatic Image Annotation System

Iulian Ilies, Arne Jacobs, Otthein Herzog, Adalbert Wilhelm, Jacobs University Bremen, Germany

The continuously increasing quantity of image data available on the Internet necessitates efficient classification and indexing methods for easy access and usage. However, progress within the field of image understanding remains limited, with algorithms for automatic concept recognition being successful only in restricted settings (e.g. Schober et al., 2005). Consequently, the prevalent approach in current web search engines is to associate images with text, e.g. surrounding articles, captions, file names, or other metadata such as manual annotations, thus allowing access to the image database via text queries. This procedure restricts the set of searchable images to those associated with text, and additionally can lead to errors if the existent associations are of poor quality. We propose an automatic annotation system for images that combines textual and visual processing techniques into a semi-supervised classification framework (first outlined in Jacobs et al., 2007). As learning basis, our procedure uses a large set of images and related text – e.g. captions and article titles, harvested from popular news web sites. Such sites feature strongly structured articles that have well defined relations between images and co-occurring text, and can therefore be parsed automatically. The acquired textual data is analyzed with standard natural language processing tools, focusing on the extraction of relevant keywords or concepts – either by using specialized detectors such as named entity recognizers (Drozdzynski et al., 2004), or by using term relevance measures such as TF-IDF (Salton & Buckley, 1988). Similarly, the images are examined using image understanding techniques that search for specific features, and extract descriptors of such interest points (SIFT, Lowe, 1999). A visual vocabulary of prototype visual features is constructed by clustering the resulting set of descriptors (Sivic and Zisserman, 2003). Consequently, each image is associated with the concepts extracted from the related text, and with the visual words representing its detected features. These two sets of associations can then be combined into a direct linkage scheme between textual concepts and visual words. We developed methods for the propagation of associations with concepts between co-occurrence-based groups of features, i.e. images, and similarity-based groups of features, i.e. visual words, in a natural way. This allows the construction of an automatic image classifier that can annotate new images with textbased concepts using only their visual features. Therefore, images with no textual information can be included in the search space of standard queries, while images that show a concept, but whose textual description is inaccurate, could have their associations corrected.

As an initial application, we investigated the performance of different concept propagation and vocabulary construction algorithms in a person classification task. We used a set of approximately 1000 images from German news sites, with named entities detected in the captions serving as categories (see Jacobs et al., 2008). The results demonstrated that the proposed approach is feasible, being able to classify new images with an overall accuracy of up to 70% (for preliminary results, see Ilies et al., 2009). A generalization to nonspecific concepts obtained via latent semantic analysis (LSA, Deerwester et al., 1990) of the captions, conducted on a significantly extended set of images, is currently in progress.

References:

Deerwester S.C., Dumais S.T., Landauer T.K., Furnas G.W., and Harshman R.A. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, vol. 41, pp. 391-407.

Drozdzynski W., Krieger H.-U., Piskorski J., Schäfer U., and Xu F. (2004). Shallow Processing with Unification and Typed Feature Structures - Foundations and Applications. Künstliche Intelligenz, vol. 1, pp. 17-23.

Ilies I., Jacobs A., Wilhelm, A.F.X., and Herzog, O. (2009). Classification of News Images Using Captions and a Visual Vocabulary. Technical Report No. 50, TZI, Universität Bremen.

Jacobs A., Hermes T., Wilhelm A.F.X. (2007). Automatic Image Annotation by Association Rules. Electronic Imaging and the Visual Arts EVA 2007, Berlin, Germany, pp. 108-112.

Jacobs A., Herzog O., Wilhelm A. F.X., and Ilies I. (2008). Relaxation-Based Data Mining on Images and Text from News Web Sites. 4th World Conference of the IASC, Yokohama, Japan, pp. 736–743.

Lowe D. G. (1999). Object Recognition from Local Scale-Invariant Features. Proceedings of the International Conference on Computer Vision, Kerkyra, Greece, pp. 1150-1157.

Salton G., and Buckley, C. (1988). Term-Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, vol. 24, pp. 513-523.

Schober J.-P., Hermes T., and Herzog O. (2005). PictureFinder: Description Logics for Semantic Image Retrieval. IEEE International Conference on Multimedia and Expo, pp. 1571-1574.

Sivic J., and Zisserman A. (2003). Video Google: A Text Retrieval Approach to Object Matching in Videos. Proceedings of the 9th IEEE International Conference on Computer Vision, Nice, France, pp. 1470-1477.

Extraction of Endogenous Metadata for Text and Image Databases

Edward J. Wegman, George Mason University

Recent years have witnessed an explosion in the amount of digitally stored data, the rate at which data is being generated, and the diversity of disciplines relying on the availability of stored data. Massive datasets are increasingly important in a wide range of applications, including observational sciences, product marketing, and the monitoring and operations of large systems. Massive datasets are collected routinely in a variety of settings in astrophysics, particle physics, genetic sequencing, geographical information systems, weather prediction, medical applications, telecommunications, sensors, government databases, and credit card transactions. Mining of massive data sets presents a major problem to the serious data miner. Datasets on the scale of terabytes or more preclude any possibility of serious effort by individual humans at manually examining and characterizing the data objects. This research addresses the challenges of autonomous discovery and triage of the contextually relevant information in massive and complex datasets. The aim is to extract feature vectors from the datasets, which will function as digital objects and then effectively reduce the volume of the datasets. I have developed an automated metadata system for scanning the database for certain statistically appropriate feature vectors, recording them as digital objects, and subsequently augmenting the metadata with appropriate digital objects. The result is that the data miner can do a Boolean search on the augmented metadata and quickly reduces the number of objects to be scanned to a much smaller dataset. Two datasets are considered. The first dataset is text data, and the second dataset is remote sensing data. The text data used in my research are documents from Topic Detection and Tracking (TDT) Pilot Corpus collected by Linguistic Data Consortium, Philadelphia, PA, which is taken directly from CNN and Reuters. The TDT corpus comprises a set of 15863 documents spanning the period from July 1, 1994 to June 30, 1995. Four features are extracted from text dataset, topics feature, discriminating words feature, verbs feature, and bigrams feature. The four features were attached to each document in the dataset as digital objects, which help in retrieving the information related to each document on the dataset. The remote sensing images used

consisted of 50 gigabytes of the Multi-angle Imaging SpectroRadiometer (MISR) instrument delivered by the Jet Propulsion Laboratory (JPL). The MISR instrument of NASA JPL?s satellite TERRA is an excellent prototype database for demonstrating feasibility. The instrument captures radiance measurements that can be converted to georectified images. In my research I developed a set of features part of it is based on Gray Level Co-occurrence Matrix (GLCM). Adjacent pairs of pixels (assuming 256 gray levels) are used to create 256 by 256 matrix with all possible pairs of gray levels reflected. Images with similar GLCM are expected to be similar images. Some of these features are constructed based on the GLCM such as Homogeneity, Contrast, Dissimilarity, Entropy, Angular Second Moment (ASM), and Energy. Other computed features include Histogram-based Contrast, Alternate Vegetation Index (AVI), and Normalized Difference Vegetation Index (NDVI), are also taking into consideration as part of the features extracted on this research. This is joint work with Faleh Alshameri.

Computing on Streams

Stream Algorithms and Workflows

J. David Harris, Department of Defense

Stream algorithms are typically defined to be algorithms that are designed to accept as input a sequence of items that can be examined in at most a few passes. Stream algorithms are generally assumed to have limited memory, exercise minimal control over the order and rate of arrival of the input, and have limited processing time per item. Given these constraints, why would one want to design algorithms to compute on streams? A streaming algorithm is almost certainly disadvantaged with respect to its cousin that isn't subject to such constraints. If we restrict our assessment of value to the correctness or accuracy of an individual algorithm, we'll oftentimes be disappointed with our result. But in most real-world applications, we're not interested in the case where data is presented to a single algorithm. Rather, we wish to consider the value of a workflow consisting of a sequence of stream algorithms. In this scenario, we can describe a measure of effectiveness for the workflow that is the combination of how well we label or characterize or understand each item as a result of presenting it to a particular stream algorithm and how well we determine what to do next with the data items, given the output of the stream algorithm. In this talk, we will consider various actions that we might wish to take while computing on streams, and we will describe a monitoring and control framework that can be used to dynamically alter the topology of a stream algorithm workflow.

Analytics for Streaming Applications

Daby Sow, IBM/T.J. Watson

In this paper, we present our experiences in building an infrastructure supporting analytics for streaming medical applications to monitor for the early onset detection of medical complications. In building this infrastructure, we identified the need to provide an environment to help domain experts create domain-specific applications. This environment has resulted in three key components: signal processing, time-series analysis, and data mining. In this paper, we will present these three components and illustrate their use in medical research as well as other domains.

Index Learning

Omid Madani, SRI International

We will describe online indexing algorithms for efficient supervised learning when the number of classes, in addition to features, can be huge, that is in the tens of thousands and beyond. The essence of indexing is limiting the number of prediction connections that each feature can make,

for efficient learning and classification. The algorithms are presented in the context of several applications, including text classification tasks and desktop action prediction. If time permits, we will also touch on an exciting application to unsupervised learning, wherein the system constructs its own classes over time (to predict and help predict), from processing much data.

Model Selection Problems in Kernel-Based Methods with Applications

Robust Bayesian Relevance Vector Machines using Information Complexity and the Genetic Algorithm

Brant Quinton and Hamparsum Bozdogan, University of Tennessee

Relevance vector machine (RVM) is a Bayesian technique that offers an improvement over the usual support vector machine (SVM) by the introduction of Bayesian paradigm to provide a probabilistic classification and regression modeling in machine learning. However, current RVM methods do not address how to choose the subset selection of best predictor variables without the use of computationally intensive cross-validation (CV) technique and the choice of the kernel function is left to be arbitrary to the discretion of the analyst. The introduction of model selection techniques such as information complexity (ICOMP) criterion of Bozdogan, layered with the genetic search algorithm (GA), allows us optimum subset selection of best predictor variables; choosing the optimal form of the kernel function among a portfolio of kernel choices; and tuning the parameters of the kernel function in a simultaneous fashion. In this paper, we introduce and develop a new hybridized approach for subset selection of the best predictor variables, kernel choice and its parameter data-adaptively based on the Smoothed Complexity Mahalanobis Distance (SCMD). We also propose the use of robust or smoothed covariance estimators to improve the model selection procedure and prevent an ill conditioned Hessian matrix used in the expectation maximization (EM) algorithm in the computation of the hyperparameters. We show numerical examples on real data sets including that of a medical data in diagnosing the early detection of heart attack in 418 patients to illustrate the performance and the flexibility of these methods.

A Non-inferiority Trial Design without the Need for a Conventional Margin

Xi Chen, PharmClint Co. and Hamparsum Bozdogan, University of Tennessee

The model selection methodology in information theory has been introduced into the design of the non-inferiority (NI) trial. The new trial set up eliminates the dependency on the conventional NI margin, and it explicitly uses the minimum clinical important difference (MCID) that links the statistical analysis to the clinical sense. Different from the conventional trial design, the new methodology is self-adaptive to the change in the sample size and overall cure rate, and it has an asymptotic property. Through the conventional likelihood ratio test, it is shown that the trial design along with AIC4 criterion the mode selection methodology consistently reached 5% type I and type II error control, which meet the requirement of ICH and other guidance. Besides, these error controls are independent to the choice of the model selection for statistical inference, so that it is more objective. The model selection methodology has revived the concept of equivalence in confirmative trial set up.

Hybridized Support Vector Machine and Recursive Feature Elimination with Information Complexity

Seung Hyun Baek and Hamparsum Bozdogan, University of Tennessee

In statistical data mining research, datasets often have nonlinearity and at the same time high-dimensionality. It has become difficult to analyze such datasets in a comprehensive manner using traditional statistical methodologies. In this paper, a novel wrapper method called SVM-

ICOMPPERF-RFE based on a hybridized support vector machine (SVM) and recursive feature elimination (RFE) with information-theoretic measure of complexity (ICOMP) is introduced and developed to classify high-dimensional data sets and to carry out subset selection of the features in the original data space for finding the best subset of features which are discriminating between the groups. Recursive feature elimination (RFE) ranks features based on information complexity (ICOMP) criterion. ICOMP plays an important role not only in choosing an optimal kernel function from a portfolio of many other kernel functions, but also in selecting important subset(s) of features. The potential and the flexibility of our approach are illustrated on two real benchmark data sets, one is ionosphere data which includes radar returns from the ionosphere, and another is aorta data which is used for the early detection of atheroma most commonly resulting heart attack. Also, the proposed method is compared with other RFE based methods using different measures (i.e., weight and gradient) for feature rankings.

Visualizing Intrusion Detection Data

Change Detection in Multivariate Streaming Data

Kyle A. Caudle, United States Naval Academy

The problem of detecting distribution changes in streaming data arises in many applications. Change detection is particularly challenging in multivariate settings in which little is known about underlying data distributions. This talk outlines a nonparametric sequential testing method for change detection in multivariate streaming data that controls the expected number of false positive detections through alpha-investing. In this talk, I present a sequential testing method for change detection in multivariate streaming data. This method addresses the two main issues in multivariate sequential testing: (1) the multivariate aspect of the data and (2) the multiple testing implications of sequential tests. This method uses multivariate density estimation through wavelets to reduce the multivariate data to a single dimension and then implements an alpha-investing algorithm to handle sequential tests.

Learning from Proximity Data

Local Learning Methods for Proximity Data

Maya Gupta, University of Washington

We discuss the importance of local learning given non-positive-definite proximity data, focusing on supervised learning. We discuss new and old state-of-the-art local learning algorithms, including discriminative, generative, and smoothing approaches, such as local support vector machines, local similarity discriminant analysis, kernel ridge interpolation, and kernel ridge regression. Results are discussed for experiments on eight real data sets, including applications to computer vision, bioinformatics, and marketing.

Learning from Heterogeneous Data Sources by Combining Dissimilarities

Brent Castle, Indiana University

Suppose that a feature space is equipped with m pairwise measuresof dissimilarity. We combine m dissimilarity matrices by forming a heterogeneous polynomial of degree 2 in the original dissimilarities. If the coefficients are entries in a copositive matrix A, then this operation maintains closure in the set of dissimilarity matrices. Now suppose that the combined dissimilarities will be used by a KNN classifier, in which case we can attempt to optimize classifier performance by learning an optimal A. Motivated by previous work on distance metric learning, we propose a plausible optimality criterion. Optimization over the set of copositive matrices is NP-hard; however, instead of restricting A to be positive semidefinite (the case of distance metric learning),

we relax the restriction on A to obtain a tractable problem. Numerical experiments confirm the viability of our approach.

Interface 2010 Program
Late Changes

Non-English Text Data Mining Via the Vector Space Model

Wikipedia as a Testbed for Implicit Translation

Krtistin Ash, NSWDD

We present an overview of the Wikipediae as a testbed for methods such as implicit translation. Discussion includes the kinds of data available within the Wikipediae (i.e. article text, inter-article links, etc.), the linguistic makeup of the Wikipediae, and how the Wikipedia data may be obtained and processed. Because a random graph embedding is used to perform implicit translation, we give particular attention to the Wikipediae as graphs.

New Developments in Statistical Data Integration

Space Oriented Rank Aggregation

Shili Lin, Ohio State University

One of the major challenges facing researchers studying complex biological systems is integration of data from omics platforms. Omic-scale data include DNA variations, transcriptom profiles, and RAomics. Selection of an appropriate approach for a data integration task is problem dependent, primarily dictated by the information contained in the data. In situations where modeling of multiple raw data sets jointly might be extremely challenging due to their vast differences, rankings from each data set would provide a commonality based on which results could be integrated. Because the underlying spaces of genes (elements) from which each ranked list come from are likely to be different, taking the underlying spaces into consideration is paramount, as failure to do so would lead to inefficient use of data and might render biases and/or sub-optimal results. However, this important aspect is usually overlooked in the literature on rank-based integration methods for omic-scale data. Nevertheless, although no assumptions about the underlying spaces are explicitly stated, carefully dissections of the algorithms reveal implicit assumptions about the spaces regardless of whether such assumptions are valid for a particular integration problem. In this talk, I will discuss a number of space oriented methods, including Markov chain based heuristic algorithms and optimization based cross entropy Monte Carlo methods for integrating ranking data. Examples will be shown to dissect the methods and to demonstrate the effects of assumptions about the underlying spaces.

Novel Methods: Internet Data and Targeted Marketing, Seriation and Pattern Discovery, Plus Collaboration and Social Networks

New Challenges in Big Data: Social Networking, Direct-Response Marketing, and Understanding Customer Behavior

Usama Fayyad, Open Insights

The rise of the interactive media represented by web, social media, search and behavioral targeting have created new challenges and opportunities for predictive analytics. While these new media offer a deeper mechanism for approaching the holy grail in marketing and advertising - understanding the customer's intent - the rich structure of available data, from social graph data to time-series from interactions to reputation and other behavioral traits, expand the complexity of prediction in dimensions where we have little experience and a poor understanding of the terrain.

We present examples of such applications as well as challenges, and will relate some case-studies to illustrate the power of understanding and harnessing this data. However, the context will also be used to illustrate the stronger need to build up new sciences to help us better understand these new powerful dimensions. Building up the science underlying these new capabilities is necessary for the future of these new media and the success of predictive and descriptive analytics in these new fields.

Index