The 24th New England Statistics Symposium

Department of Statistics, Harvard University

Friday-Saturday, April 16-17, 2010

Welcome and Acknowledgements

The Harvard University Statistics Department is proud to host NESS 2010. We welcome all attendees to Harvard Yard and hope that this Symposium is valuable and enjoyable for everyone.

In this booklet, you will find:

- concise schedules for both Friday and Saturday on pages 3–5
- keynote speakers information on page 6
- detailed program for parallel paper sessions on pages 7–16
- abstracts for all talks, beginning on page 17.

We are extremely grateful to our sponsors:

Microsoft
Google

for their generous, well-appreciated support of NESS 2010. We also thank

Cambridge University Press
International Press of Boston
John Wiley & Sons
Taylor & Francis
Chapman & Hall
Springer
Facebook
Yahoo!

for their extraordinary contributions.

Dale Rinkel, Local Co-Chair
Steve Finch, Local Co-Chair
Edo Airoldi, Program Chair

Schedule, Friday, April 16, 2010

12:00pm	Registration & Coffee, outside Maxwell-Dworkin G115 (Lessin Lecture Hall)
12:30 - 3:30pm	Short Course I, MD G115: "Quantitative Financial Modelling in the Post-Lehman Landscape" Stephen Blyth, Harvard University and Harvard Management Company
12:30 - 3:30pm	Short Course II, MD 119: "Statistical Software for Genomic and Clinical Data Processing" Curtis Huttenhower, Harvard University; Florian Markowetz, Cancer Research UK Cambridge Research Institute
3:30pm	Coffee Break
3:45 - 6:45pm	Short Course III, MD G115: "Statistical Elements of Complex Networks" Edoardo Airoldi and Joseph Blitzstein, Harvard University
6:45pm	Break
7:00 - 9:00pm	Short Course IV, Science Center Lecture Hall A: "Research Cultivation and Culmination: How to Get Your Paper Published (Eventually)" Joseph Blitzstein and Xiao-Li Meng, Harvard University

Schedule, Saturday, April 17, 2010

9:30am	Registration & Coffee, outside Science Center Lecture Hall D
10:00 - 10:15am	Welcome and Opening Remarks, SC Lecture Hall D: Xiao-Li Meng, Stat Dept Chair and Jeremy Bloxham, Harvard FAS Dean of Science
10:15 - 11:15am	Keynote Presentation 1, SC Lecture Hall D: "Largest Eigenvalues and Eigenvectors in Multivariate Analysis" Iain Johnstone, Stanford University
11:15am	Coffee Break
11:30 - 1:00pm	Parallel Paper Sessions A
A1	Novel Statistical Analysis of Large Datasets (Databases) to Inform Health Policy (Invited Session) SC 309
A2	Systems Biology (Invited Session) SC 309a
A3	Reproducibility (Invited Session) SC B-10

A4 UConn
(Invited Session) SC 221

A5 Machine Learning
(Invited Session) SC 222

A6 MIT
(Invited Session) SC 112

A7 UPenn
(Invited Session) SC 216

A8 Student Paper Competition Session I
SC 113

A9 Contributed Paper Session I
SC 304

A10 Contributed Paper Session II
SC 116

1:00pm Lunch, outside SC Lecture Hall D

2:15 - 3:15pm Keynote Presentation 2, Lecture Hall D:
"Interdisciplinarity in the Age of Networks"
Jennifer Tour Chayes, Microsoft Research New England

3:15pm Coffee Break

3:30 - 5:00pm Parallel Paper Sessions B

B1 Experimental Design
(Invited Session) SC B-10

B2 Google & Yahoo
(Invited Session) SC 309

B3 Social Networks
(Invited Session) SC 309a

B4 Yale
(Invited Session) SC 221

B5 Astrostats
(Invited Session) SC 112

B6 Columbia
(Invited Session) SC 216

B7 Recent Advances and Applications of Bayesian Nonparametric Inference I
SC 222

B8 Student Paper Competition Session II
SC 113

B9 Contributed Paper Session III
SC 304

B10 Contributed Paper Session IV
SC 116

5:00 - 6:30pm Parallel Paper Sessions C

C1 Sports
(Invited Session) SC 221

C2 Bell Labs & Facebook
(Invited Session) SC 309

C3 High Dimensional Data
(Invited Session) SC 309a

C4 Business
(Invited Session) SC B-10

C5 Recent Advances and Applications of Bayesian Nonparametric Inference II
(Invited Session) SC 222

C6 BU
(Invited Session) SC 112

C7 Brown
(Invited Session) SC 216

C8 Student Paper Competition Session III
SC 113

C9 Student Paper Competition Session IV
SC 304

C10 Contributed Paper Session V
SC 116

6:30 - 9:00pm Closing Reception, location to be announced

Plenary Keynote Addresses

Largest Eigenvalues and Eigenvectors in Multivariate Analysis
Iain Johnstone, Stanford University

The eigenvalues of Wishart matrices play a central role in classical multivariate analysis. A new impetus to approximate distribution results has come from methods that imagine the number of variables as large. We focus on the largest eigenvalue in particular, and review null distribution approximations to Gaussian, single and double Wishart problems in terms of the Tracy-Widom laws. If time permits, we will also briefly mention estimation of the eigenvectors associated to the top eigenvalues.

Brief Biography

Iain Johnstone is Marjorie Mhoon Fair Professor of Quantitative Science in the Department of Statistics at Stanford University. He holds a joint appointment in biostatistics in Stanford's School of Medicine. He received his Ph.D. in Statistics from Cornell in 1981.
His work in theoretical statistics aims to provide insight into methods of data analysis in diverse areas of science and medicine. He has used ideas from harmonic analysis, such as wavelets, to understand noise-reduction methods in signal and image processing. More recently, he has applied random matrix theory to the study of high-dimensional multivariate statistical methods, such as principal components and canonical correlation analysis. In biostatistics, he has collaborated extensively with investigators in cardiology and prostate cancer.
A native of Australia, Johnstone is a member of the U.S. National Academy of Sciences and the American Academy of Arts and Sciences and a former president of the Institute of Mathematical Statistics.

Interdisciplinarity in the Age of Networks
Jennifer Tour Chayes, Microsoft Research New England

Everywhere we turn these days, we find that dynamical random networks have become increasing appropriate descriptions of relevant interactions. In the high tech world, we see mobile networks, the Internet, the World Wide Web, and a variety of online social networks. In economics, we are increasingly experiencing both the positive and negative effects of a global networked economy. In epidemiology, we find disease spreading over our ever growing social networks, complicated by mutation of the disease agents. In problems of world health, distribution of limited resources, such as water resources, quickly becomes a problem of finding the optimal network for resource allocation. In biomedical research, we are beginning to understand the structure of gene regulatory networks, with the prospect of using this understanding to manage the many diseases caused by gene mis-regulation. In this talk, I look quite generally at some of the models we are using to describe these networks, and at some of the methods we are developing to indirectly infer network structure from measured data. In particular, I will discuss models and techniques which cut across many disciplinary boundaries.

Brief Biography

Jennifer Chayes is Managing Director of Microsoft Research New England. Her research areas include phase transitions in discrete mathematics and computer science, structural and dynamical properties of self-engineered networks, and algorithmic game theory. She is the coauthor of over 100 scientific papers and the co-inventor of over 20 patents.
Chayes serves on numerous institute boards, advisory committees and editorial boards, including the Turing Award Committee, the US National Committee on Mathematics, and the Board of Trustees of the Mathematical Sciences Research Institute. Chayes received her Ph.D. at Princeton, and held postdoctoral fellowships at Harvard and Cornell. She is the recipient of the NSF Postdoctoral Fellowship, the Sloan Fellowship, and the UCLA Distinguished Teaching Award. Chayes is a Fellow of the AAAS and the Fields Institute, and a National Associate of the National Academies.

Detailed Program, Parallel Paper Sessions

Session A1: **Novel Statistical Analysis of Large Datasets (Databases) to Inform Health Policy**
(Invited Session) SC 309, 11:30 - 1:00pm

Organizer: James O'Malley, Department of Health Care Policy, Harvard Medical School

Talks:

1. Sequential Analytic Methods for Post-marketing Safety Surveillance Using Existing Healthcare Databases
Lingling Li, Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute

2. Assessing Geographical Variations in Hospital Processes of Care Using Multilevel Item Response Models
Yulei He, Department of Health Care Policy, Harvard Medical School

3. Gaussian-Based Routines for Imputing Categorical Variables in Complex Designs
Recai Yucel, Department of Epidemiology and Biostatistics, State University of New York at Albany

Session A2: **Systems Biology**
(Invited Session) SC 309a, 11:30 - 1:00pm

Organizers: Curtis Huttenhower, Harvard University;
Florian Markowetz, Cancer Research UK Cambridge Research Institute

Talks:

1. Evolutionary and Chromatin Signatures for Understanding the Human Genome and its Regulation
Manolis Kellis, MIT

2. The Intersectome: Integration of Knowledge in Systems Biology for Hypothesis Generation
Avi Ma'ayan, Mount Sinai School of Medicine

3. Network Medicine: From Cellular Networks to the Human Diseasome
Albert-László Barabási, Center of Complex Networks Research, Northeastern University & Department of Medicine, Harvard University

Session A3: **Reproducibility**
(Invited Session) SC B-10, 11:30 - 1:00pm

Organizer: Victoria Stodden, Yale Law School

Talks:

1. Barriers to the Practice of Really Reproducible Research
Victoria Stodden, Yale Law School

2. The Importance of Reproducibility in High-Throughput Biology: Case Studies in Forensic Bioinformatics
Keith Baggerly, Univ. of Texas M. D. Anderson Cancer Center

3. Reproducible Research for Genome Scale Biology: Inputs from Statistical Computing
Vincent Carey, Harvard Medical School & Brigham and Women's Hospital

Session A4: **UConn**
(Invited Session) SC 221, 11:30 - 1:00pm

Talks:

1. A Novel Approach to Statistical Modeling of the Output Properties of High Level Auditory Neurons
Zhiyi Chi, Univ. of Connecticut

2. Nonparametric Rank-Based Tests of Bivariate Extreme-Value Dependence
Jun Yan, Univ. of Connecticut

3. Repeated Significance Tests in Presence of Random Costs
Vladimir Pozdnyakov, Univ. of Connecticut

Session A5: **Machine Learning**
(Invited Session) SC 222, 11:30 - 1:00pm

Talks:

1. Recent Progress in MAP Estimation for Computer Vision
Ramin Zabih, Cornell University

2. Nonparametric Bayes Classification and Testing on Manifolds
Abhishek Bhattacharya, Duke University

3. Jointly Primal and Dual Sparse Structured I/O Models
Eric Xing, Cornell University

Session A6: **MIT**
(Invited Session) SC 112, 11:30 - 1:00pm

Talks:

1. One-Shot Learning with a Hierarchical Nonparametric Bayesian Model
Russ Salakhutdinov, MIT

2. An Equivalence between AdaBoost and RankBoost
Cynthia Rudin, MIT

3. Reconstruction of Latent Tree Models
Animashree Anandkumar, MIT

Session A7: **UPenn**
(Invited Session) SC 216, 11:30 - 1:00pm

Talks:

1. Regularization Methods for Sequential Prediction
Alexander Rakhlin, Univ. of Pennsylvania

2. Fuzzy Hypotheses, Hermite Polynomials, and Optimal Estimation of a Nonsmooth Functional
Tony Cai, Univ. of Pennsylvania

3. Causal Inference for Continuous Time Processes When Covariates Are Observed Only at Discrete Times
Dylan Small, Univ. of Pennsylvania

Session A8: **Student Paper Competition Session I**
SC 113, 11:30 - 1:00pm

Talks:

1. Statistical Inference in Factor Analysis for High-Dimensional, Low-Sample Size Data
Miguel Marino, Harvard School of Public Health

2. A Nonparametric Test for the Validation of Surrogate Endpoints
Xiaopeng Miao, Boston University School of Public Health

3. Robust Survival Prediction via Linear Transformation Models
Keith A. Betts, Harvard School of Public Health and Dana Farber Cancer Institute

4. Principled Sure Independence Screening for Cox Models with Ultra-High-Dimensional Covariates
Sihai Dave Zhao, Harvard School of Public Health and Dana Farber Cancer Institute

Session A9: **Contributed Paper Session I**
SC 304, 11:30 - 1:00pm

Talks:

1. Enhancing Interpretation of Patient-Reported Outcomes
Joseph C. Cappelleri, Senior Director - Biostatistics, Pfizer Inc.

2. Circular Migrations and HIV Transmission Dynamics [CANCELLED]
Aditya Khanna, Quantitative Ecology and Resource Management, Univ. of Washington

3. Identifying Differentially Expressed Genes in Time Series Microarrays
Jonathan J. Smith, MIT

4. A Hierarchical Spherical Radial Quadrature Algorithm for Gene Pathway Analysis
Jacob Gagnon, University of Massachusetts Amherst

Session A10: **Contributed Paper Session II**
SC 116, 11:30 - 1:00pm

Talks:

1. Hedge Fund Replication Using Minimax Filters: Report on a work in Progress
Guillaume Weisang, Bentley University

2. A Maximum Likelihood Estimator for the q-Gaussian Distribution and its Application to Financial Time Series
Claudio D. Antonini, UBS Investment Bank

3. Customer Segmentation Using Nonparametric Clustering Methods of Categorical Time Series
Shan Hu, University of Connecticut

4. On the Pricing of Eurodollar Futures
Balaji Raman, University of Connecticut

Session B1: **Experimental Design**
(Invited Session) SC B-10, 3:30 - 5:00pm

Organizer: Tirthankar Dasgupta, Harvard University

Talks:

1. Designs for Bayesian Model Selection
Dave Woods, University of Southampton, United Kingdom

2. Sequential Numerical Integration and Stochastic Optimization with Statistical Designs
Peter Qian, University of Wisconsin-Madison

3. Multi-Objective fMRI Designs with Unequal Epoch Length via NSGA-II
Ming-Hung Kao, Arizona State University

Session B2: **Google & Yahoo**
(Invited Session) SC 309, 3:30 - 5:00pm

Organizer: Gal Chechik, Google

Talks:

1. Teaching Machines to Understand Signals: A Large Scale Learning Approach
Gal Chechik, Google

2. Domain Adaptation Theory and Algorithms
Mehryar Mohri, Google and New York University

3. Recommender Problems for Web Applications
Deepak Agrawal, Yahoo!

Session B3: **Social Networks**
(Invited Session) SC 309a, 3:30 - 5:00pm

Organizer: Joe Blitzstein, Harvard University

Talks:

1. Using Genes as Instrumental Variables in Analyses of Social Network Data
James O'Malley, Department of Health Care Policy, Harvard Medical School

2. Spatial Process Model for Social Networks
Crystal Linkletter, Brown University

3. Definition, Calculation and Stability of Centrality Measures in Networks
Andrew C. Thomas, Carnegie Mellon University

Session B4: **Yale**
(Invited Session) SC 221, 3:30 - 5:00pm

Talks:

1. The Shannon-McMillan-Breiman Theorem for Log-Concave Distributions
Mokshay Madiman, Yale Univ.

2. Solving Least Squares via Gaussian Belief Propagation
Sekhar Tatikonda, Yale Univ.

3. Optimal Estimation of Large Covariance Matrices
Harry Zhou, Yale Univ.

Session B5: **Astrostats**
(Invited Session) SC 112, 3:30 - 5:00pm

Talks:

1. Analyzing Stellar Populations Using Color-Magnitude Diagrams
Paul Baines, Harvard University

2. Doing Right by Massive Data: Using Probability Modeling to Advance the Analysis of Huge Astronomical Datasets
Alexander Blocker, Harvard University

3. Facing the Supernova Challenge: Complex Theory and Complex Data
Chad Schafer, Carnegie Mellon University

Session B6: **Columbia**
(Invited Session) SC 216, 3:30 - 5:00pm

Talks:

1. Statistical Methods for Studying Social Networks Using Aggregated Relational Data
Tian Zheng, Columbia Univ.

2. On the Stationary Distributions of the Chained Imputations
Jingchen Liu, Columbia Univ.

3. Shape Restricted Function Estimation and Inference in Non-standard Problems
Bodhisattva Sen, Columbia Univ.

Session B7: **Recent Advances and Applications of Bayesian Nonparametric Inference I**
(Invited Session) SC 222, 3:30 - 5:00pm

Organizer: Lorenzo Trippa, Harvard School of Public Health

Talks:

1. On a Class of Normalized Random Measures with Independent Increments
Lorenzo Trippa, Harvard School of Public Health

2. Priors on Topological and Metric Spaces: A Computational Perspective
Daniel Roy, MIT

3. A Bayesian Discovery Procedure
Michele Guindani, University of New Mexico

Session B8: **Student Paper Competition Session II**
SC 113, 3:30 - 5:00pm

Talks:

1. Rasch Model and its Extensions for Analysis of Aphasic Deficits in Syntactic Comprehension
Roee Gutman, Harvard University

2. Spatiotemporal Kriging and Correlation Modeling of Wind Power Generation [CANCELLED] *Ada Lau, Oxford-Man Institute, University of Oxford*

3. A Bayesian Approach to the Analysis of Time Symmetry in Light Curves: Reconsidering Scorpius X-1 Occultations [CANCELLED]
Alexander Blocker, Harvard University

4. How Many Modes Can a Mixture of Two Components Have?
Dan Ren, Boston University

Session B9: **Contributed Paper Session III**
SC 304, 3:30 - 5:00pm

Talks:

1. Multilevel Models and Small Area Estimation in the Context of Vietnam Living Standards Surveys
Dominique Haughton, Bentley University and Toulouse School of Economics

2. Detecting and Understanding Overlapping Community Structure in Network
Guangying Hua, Bentley University

3. Comparing Multiple Treatments to both Positive and Negative Controls
Eleanne Solorzano, University of New Hampshire

4. Rare-Allele Detection Using Compressed Se(que)nsing
Or Zuk, Broad Institute of MIT and Harvard

Session B10: **Contributed Paper Session IV**
SC 116, 3:30 - 5:00pm

Talks:

1. Tracking Multiple Targets Using Binary Decisions from Wireless Sensor Networks
Natallia Katenka, Boston University

2. Factor Rotation of Functional Data for Extracting Periodic Objects
Chong Liu, Boston University

3. Semiparametric Bayesian inference in functional nonlinear regression models for panel time series data
Sylvie Tchumtchoua, University of Connecticut

4. The International Digital Divide: An Application of Kohonen Self Organizing Maps
Maria Skaletsky, Bentley University

Session C1: **Sports**
(Invited Session) SC 221, 5:00 - 6:30pm

Talks:

1. Odds Ratio Models in Sports
Carl Morris, Harvard University

2. Bayesball: Spatial Hierarchical Modeling of Fielding in Major League Baseball
Shane Jensen, University of Pennsylvania

3. Paired Comparison Models with Tie Probabilities and Order Effects as a Function of Strength
Mark Glickman, Boston University

Session C2: **Bell Labs & Facebook**
(Invited Session) SC 309, 5:00 - 6:30pm

Talks:

1. Reading the Social Pages: Understanding and Predicting the Demographics and Behavior of Facebook Users
Jonathan Chang, Facebook

2. Community Detection with Block Models - Some Theory and Applications
Aiyou Chen, Bell Labs

3. Online Analysis of Data Streams
Jin Cao, Bell Labs

Session C3: **High Dimensional Data**
(Invited Session) SC 309a, 5:00 - 6:30pm

Talks:

1. Geometric Representations of Hypergraphs for Prior Specification and Posterior Sampling
Sayan Mukherjee, Duke University

2. Intelligence and Learning Theory: Kernels and Derived Kernels
Tomaso Poggio, MIT

3. On the Geometry of Discrete Exponential Families with Application to Exponential Random Graph Models
Alessandro Rinaldo, Carnegie Mellon University

Session C4: **Business**
(Invited Session) SC B-10, 5:00 - 6:30pm

Talks:

1. To Explain or To Predict?
Galit Shmueli, Smith School of Business, University of Maryland

2. Time-Varying Predictive Systems
Carlos Carvalho, Booth School of Business, University of Chicago

3. Bayesian inference for Sparse Pair-wise Copula Models
Yoonjung Lee, Dept. of Statistics, Harvard University

Session C5: **Recent Advances and Applications of Bayesian Nonparametric Inference II**
(Invited Session) SC 222, 5:00 - 6:30pm

Organizers: Michele Guindani, University of New Mexico;
Fabrizio Leisen, Universidad de Navarra

Talks:

1. Species Sampling Models for Bayesian Nonparametric Inference
Antonio Lijoi, Universita' di Pavia, Italy

2. A Spatial Dirichlet Process Mixture Model for Clustering
Brian Reich, North Carolina State University

3. Vectors of Poisson Dirichlet Processes
Fabrizio Leisen, Universitad de Navarra, Spain

Session C6: **BU**
(Invited Session) SC 112, 5:00 - 6:30pm

Organizer: Surajit Ray, Boston Univ.

Talks:

1. Analysis of Differential Neural Spiking Activity Using Point Process Models
Uri Eden, Boston Univ.

2. Pathway-Based Machine Learning Approaches for Cancer Classification
Mark Kon, Boston Univ.

3. Bayesian Centroid Estimation
Luis Carvalho, Boston Univ.

Session C7: **Brown**
(Invited Session) SC 216, 5:00 - 6:30pm

Talks:

1. Multi-scale Multiple Hypothesis Testing for Spike Trains
Matthew Harrison, Applied Math Division, Brown Univ.

2. Limitations of Point Estimates in Computational Biology
Chip Lawrence, Applied Math Division, Brown Univ.

3. New Statistical and Computational Challenges in Next-Generation DNA and RNA Sequencing
Nicola Neretti, Molecular & Cellular Biology Dept., Brown Univ.

Session C8: **Student Paper Competition Session III**
SC 113, 5:00 - 6:30pm

Talks:

1. Sample Size Calculation for Poisson Endpoint Using the Exact Distribution of Difference between Two Poisson Random Variables
Sandeep Menon, Boston University

2. Estimate Absolute Concentrations of Transcripts from DNA Microarrays
Yunxia Sui, Brown University

3. Semi-Supervised Recusively Partitioned Mixture Models for Identifying Cancer Subtypes
Devin C. Koestler, Department of Community Health, Brown University

4. The Generalized Shrinkage Estimator For Spectral Analysis of Multivariate Time Series
Mark Fiecas, Center for Statistical Sciences, Brown University

Session C9: **Student Paper Competition Session IV**
SC 304, 5:00 - 6:30pm

Talks:

1. Shape Constrained Statistical Estimation via Semidefinite Optimization
David Papp, Rutgers University

2. Multiple Imputation: A Negotiation of Two Parties.
Xianchao Xie, Harvard University

3. A Locally D-Optimal Design for Estimation of Parameters of an Exponential-Linear Growth Curve of Nanostructures
Li Zhu, Harvard University

Session C10: **Contributed Paper Session V**
SC 116, 5:00 - 6:30pm

Talks:

1. Diagnostic of Protein Phosphorylation Site Using Zero-Inflated Poisson Regression
Shu Yang, Boston University

2. Asymptotic Distribution of Poisson Index of Dispersion
Liang Meng, Boston University

3. Dependence of Spike-Field Coherence on Expected Intensity
Kyle Lepage, Boston University

4. A Graph Log-Linear Model for Characterizing Repeated Interactions
Patrick O. Perry, Harvard Statistics and Information Sciences Lab

Abstracts

Recommender Problems for Web Applications
Deepak Agrawal, Yahoo!

Several web applications like content optimization and online advertising involve recommending items from an inventory for each user visit to maximize some yield metric of interest (e.g. click rates). These are instances of large scale recommender system problems that entail several statistical challenges. We provide a mathematical description of the problem followed by modeling solutions for a content optimization problem that arises in the context of Yahoo! Front Page (www.yahoo.com). In fact, we discuss models to a) serve most popular items, b) serve items that are most popular in different user segments and c) provide personalized item recommendations for each user. Our models are based on time series methods, multi-armed bandit schemes and bilinear random effects model.
One class of bilinear random effects model we propose extends reduced rank regression to incomplete matrices, the other class extends matrix factorization to incorporate covariates.

Reconstruction of Latent Tree Models
Animashree Anandkumar, MIT

We consider reconstruction of latent tree models which are tree-structured graphical models with samples available only from a subset of nodes. We propose two consistent algorithms which are guaranteed to recover any minimal latent tree model (without redundant hidden nodes) asymptotically. The algorithms are provably efficient with low computational and sample complexity. Experiments demonstrate the efficacy of our algorithms across a wide spectrum of latent tree models, such as a star, a hidden Markov model and a complete tree. This problem has many applications, and I will briefly describe the on-going work on building a hierarchical contextual tree to predict object co-occurrences in images.
This is joint work with Myung Jin Choi, Vincent Tan and Alan Willsky.

A Maximum Likelihood Estimator for the q-Gaussian Distribution and its Application to Financial Time Series
Claudio D. Antonini, UBS Investment Bank

In the last two decades, the q-Gaussian distribution has become an ubiquitous tool in a multitude of scientific disciplines, with close to 3,000 articles making use of it, trying to explain it, or showing its limitations. Besides being able to reproduce the fat tails observed in real data, statistical mechanical arguments based on nonextensive entropy allow to infer explanations for the mechanisms that give rise to the observed behavior in many applications. The distribution also exhibits properties that makes it more feasible than alternative solutions (like Lévy process), and allows to quantify in an intuitive way the departure from normality. However, in many disciplines, the main tool to find the parameters of the distribution remains least-squares, producing biased parameter estimates.
In this article, we derive a maximum likelihood (ML) procedure to estimate the parameters of the q-Gaussian distribution and its confidence ellipsoid, and compare the results to bootstrapping. We apply the method to study the returns of the S&P500 index from 1950 to 2010 and show how the departure from normality has changed in time.
At the same time, we establish a relationship with recently published results in ARCH(1) and GARCH(1,1), methods which only provide point estimates of their parameters. Consequently, by applying the ML method on the q-Gaussian distribution, we can also find the ARCH(1)/GARCH(1,1) constants and their confidence intervals, thus linking the shape of the distribution with the short-term volatility forecasting, respectively. Moreover, the relationship works both ways: given the ARCH(1)/GARCH(1,1) constants we can find the parameters of the q-Gaussian distribution and thus determine the degree of departure from normality.
By establishing this bridge, it is therefore equivalent to speak of autoregressive behavior (ARCH/GARCH language) and departure of normality (q-Gaussian terminology).

The Importance of Reproducibility in High-Throughput Biology: Case Studies in Forensic Bioinformatics
Keith Baggerly, Univ. of Texas M. D. Anderson Cancer Center

Over the past few years, microarray experiments have supplied much information about the disregulation of biological pathways associated with various types of cancer. Many studies focus on identifying subgroups of patients with particularly aggressive forms of disease, so that we know who to treat. A corresponding question is how to treat them. Given the treatment options available today, this means trying to predict which chemotherapeutic regimens will be most effective.
Several microarray studies have provided such predictions. Unfortunately, ambiguities associated with analyzing the data have made many of these results difficult to reproduce. In this talk, we will describe how we have analyzed the data, and reconstructed aspects of the analysis from the reported results. In some cases, these reconstructions reveal inadvertent flaws that affect the results. Most of these flaws are simple in nature, but their simplicity is obscured by a lack of documentation. We briefly discuss the implications of such ambiguities for clinical findings. We will also describe approaches we now follow for making such analyses more reproducible, so that progress can be made more steadily.

Analyzing Stellar Populations Using Color-Magnitude Diagrams
Paul Baines, Harvard University

Many problems in astrophysics take the form of observations on some process determined by unknown physical parameters, with the relationship between the parameters and the observed quantities defined either by simulations, by a lookup table, or by a many-to-one function. For example, photometric data on a collection of stars can be used to infer the mass, age, and metallicity of those stars. This requires the analyst to specify a mapping between the parameters (mass, age, metallicity) and the observed data. Typically this is done through a lookup table created by one of many competing models of the underlying physics. In the case of Color-Magnitude Diagrams (CMDs) the mapping is many-to-one and, hence, non-invertible. This poses a number of theoretical and computational challenges for the data analysis. We present an example of these challenges in the use of CMDs to infer star formation histories, via Hierarchical Bayesian modelling. Many recent advances in computational methods and computing power enable us to answer previously infeasible questions, and to frame them in greater generality. We present our methodology for the CMD example, and indicate how the theoretical framework and computational techniques involved can be utilized in many applications beyond CMD analysis.

Network Medicine: From Cellular Networks to the Human Diseasome
Albert-László Barabási, Center of Complex Networks Research, Northeastern University and Department of Medicine, Harvard University

The ultimate goal of understanding sub-cellular networks is to gain insights into the normal cellular functions, and understand the microscopic nature of perturbations that could lead to human diseases. A network of disorders and disease genes linked by known disorder- gene associations offers a platform to explore in a single graph-theoretic framework all known phenotype and disease gene associations, indicating the common genetic origin of many diseases. We find that the vast majority of disease genes are nonessential and show no tendency to encode hub proteins, and their expression pattern indicates that they are localized in the functional periphery of the network. We also study the evolution of patient illness using a network summarizing the disease associations extracted from 32 million Medicare claims, demonstrating that the cellular level links between disease causing proteins are amplified in the population as comorbodity patterns.

Robust Survival Prediction via Linear Transformation Models
Keith A. Betts, Harvard School of Public Health and Dana Farber Cancer Institute

For censored time to event data, it is important to develop flexible regression models that can be used to accurately predict future risk. In this article, we develop robust prediction models for event time outcomes by generalizing Cai's estimating equation approach for the linear transformation model (Cai et al., 2000), which includes the proportional odds and proportional hazards model. We demonstrate that under mild

regularity conditions, the solution of the estimating equations possess a stability property which allows for valid predictive inference under possible model misspecification. The proposed procedures are applied to a multiple myeloma dataset to derive a flexible regression model for predicting patient survival based on traditional clinical factors with and without the addition of genetic information. The finite sample properties of the procedures are evaluated through a simulation study.
Co-author: David P. Harrington

Nonparametric Bayes Classification and Testing on Manifolds
Abhishek Bhattacharya, Duke University

We develop general Bayes methods for density estimation, classification and testing on known manifolds. For example, the manifold may correspond to the surface of a hypersphere or a planar shape space. We propose a general kernel mixture model for the joint density of the response and predictors, with the kernel expressed in product form and dependence induced through the unknown mixing measure. We provide simple sufficient conditions on the prior that lead to L1 and Kullback-Leibler (KL) support on the space of continuous joint densities for the predictors and response. Focusing on a Dirichlet process prior for the mixing measure, these conditions hold using von Mises-Fisher kernels when the manifold is the unit hypersphere and complex Watson kernels for planar shape spaces. Bayesian methods are developed for efficient posterior computation using an exact block Gibbs sampler. We also develop nonparametric methods for testing for differences between groups in variables having an unknown density on a manifold, with efficient computational methods proposed for Bayes factor calculation. The methods are evaluated using simulation examples and applied to spherical data and shape applications.
Co-author: David Dunson

Doing Right by Massive Data: Using Probability Modeling to Advance the Analysis of Huge Astronomical Datasets
Alexander Blocker, Harvard University (1)

The analysis of extremely large, complex datasets is becoming an increasingly important task in the analysis of scientific data. This trend is especially prevalent in astronomy, as large-scale surveys such as SDSS, Pan-STARRS, and the LSST deliver (or promise to deliver) terabytes of data per night. While both the statistics and machine-learning communities have offered approaches to these problems, neither has produced a completely satisfactory approach. Working in the context of event detection for the MACHO LMC data, I will present an approach that combines much of the power of Bayesian probability modeling with the the efficiency and scalability typically associated with more ad-hoc machine learning approaches. This provides both rigorous assessments of uncertainty and improved statistical efficiency on a dataset containing approximately 20 million sources and 40 million individual time series. I will also discuss how this framework could be extended to related problems.

A Bayesian Approach to the Analysis of Time Symmetry in Light Curves: Reconsidering Scorpius X-1 Occultations [CANCELLED]
Alexander Blocker, Harvard University (2)

We present a new approach to the analysis of time symmetry in light curves, such as those in the X-ray at the center of the Scorpius X-1 occultation debate. Our method uses a new parameterization for such events (the bilogistic event profile) and provides a clear, physically relevant characterization of each event's key features. We also demonstrate a Markov Chain Monte Carlo algorithm to carry out this analysis, including a novel independence chain configuration for the estimation of each event's location in the light curve. These tools are applied to the Scorpius X-1 light curves presented in Chang et al. (2007), providing additional evidence based on the time series that the events detected thus far are most likely not occultations by TNOs.
Co-authors: Pavlos Protopapas, Charles R. Alcock

Fuzzy Hypotheses, Hermite Polynomials, and Optimal Estimation of a Nonsmooth Functional
Tony Cai, University of Pennsylvania

In this talk I will discuss some recent work on optimal estimation of nonsmooth functionals. These problems exhibit some interesting features that are significantly different from those that occur in estimating conventional smooth functionals. This is a setting where standard techniques fail. I will discuss a newly developed general minimax lower bound technique that is based on testing two fuzzy hypotheses and illustrate the ideas by focusing on the problem of optimal estimation of the l_1 norm of a high dimensional normal mean vector. An estimator is constructed using approximation theory and Hermite polynomials and is shown to be asymptotically sharp minimax. This is joint work with Mark Low.

Online Analysis of Data Streams
Jin Cao, Bell Labs

Massive data streams are becoming increasingly commonplace in many areas of science and technology. Examples of such are network packet traces, or business transaction records. With massive data, and limited storage and computation power, analysis of such data becomes difficult even for very simple task such as simple counting of the number of distinct values. In this talk, I shall give a statistician's perspective for analyzing such data streams using online methods, which means the data cannot be stored and will be seen only once. I will present our recent work on problems such as heavy hitter detection, cardinality (distinct value) counting, and quantile estimation.
This is joint work with Aiyou Chen, Tian Bu, Yu Jin and Li Li.

Enhancing Interpretation of Patient-Reported Outcomes
Joseph C. Cappelleri, Senior Director - Biostatistics, Pfizer Inc.

In December 2009 the Food and Drug Administration released its final guidance on patient-reported outcomes for use in medical product development to support labeling claims. One of the central topics in the Guidance is the interpretation of patient-reported outcomes, a topic that has become an active area of research. This presentation includes several ways to enhance interpretation of patient-reported outcomes: responder analysis with anchor-based methods, cumulative proportions and responder analysis, content-based interpretation, reference-group interpretation, and distribution-based methods. These methodologies are illustrated with several examples from the literature.

Reproducible Research for Genome Scale Biology: Inputs from Statistical Computing
Vincent Carey, Harvard Medical School & Brigham and Women's Hospital

Genome-scale experiments engender new and daunting requirements for reliable statistical analysis of very large numerical data and non-numerical annotation collections. Studies of genome variation previously focused on a few million SNP, but must now confront general idiosyncratic individual level variation measured via exome or full personal genome sequencing. Studies of transcriptome variation have evolved from measurements of thousands of oligonucleotide probes to megabases of mRNA sequence per individual. Current experiments employ complex and sensitive lab protocols, and their interpretation depends on evolving structural and functional annotation, and emerging statistical methodology. I will present case studies of reproducibility failure in microarray and sequencing studies, and will discuss challenges involved in detecting and demonstrating nonreproducibility. Issues involved in designing and deploying tools supporting scalable interaction with billions of statistical tests involved in eQTL discovery and interpretation will be discussed.

Time-Varying Predictive Systems
Carlos Carvalho, University of Chicago

In "Predictive Systems: Living with Imperfect Predictors", Pastor and Stambaugh (2008) develop a framework for estimating expected returns. In "Are Stocks Really Less Volatile in the Long Run" (2009) they use this framework to assess the conventional wisdom that stocks are less volatile over long horizons than short horizons. They show that this conclusion is only reached by ignoring important parameter uncertainty. They also argue that a key component of prior information concerns the correlation between unanticipated expected return and the unpredictable return.

The predictive system framework consists of a vector auto regression in the stock return, the latent expected return for the next period, and a set of variables thought to be able to predict returns. They assume that the innovation covariance is constant over time. This assumption runs counter to much empirical evidence. In this paper we extend the predictive systems framework to account for time varying volatility. We do so in a way that allows us to incorporate complex economically based prior information. In particular, we use prior information about the time series of the correlation between unanticipated expected return and the unpredictable return. In this enriched environment, we examine what kind of prior information is needed to make stock returns less volatile in the long run.

Bayesian Centroid Estimation
Luis Carvalho, Boston University

Maximum likelihood estimators have traditionally dominated discrete inference for a long time. In this work we apply statistical decision theory to derive a new contender that minimizes a posterior generalized Hamming loss: the centroid estimator. The centroid estimator is formally characterized as a solution to a discrete optimization problem having posterior marginal distributions as inputs. We discuss both specific constraints of interest and broad conditions under which this optimization problem becomes tractable and provide further generalizations to centroid estimation. We illustrate centroid estimation with simple applications to stochastic grammar parsing, reconstruction of ancestral states given a phylogeny, and genome-wide association studies. Finally, we offer a few concluding remarks and directions for future work.

Reading the Social Pages: Understanding and Predicting the Demographics and Behavior of Facebook Users
Jonathan Chang, Facebook

Facebook's massive data provide opportunities to answer long-standing questions about hundreds of millions of users – Who are they? What do they do? What do they want to do? Answering such questions at scale requires leveraging advances in data infrastructure, machine learning and data mining. In this talk I will present several approaches pursued by members of the Data Team at Facebook at answering these questions. I will describe how Facebook deals with large data sets, how we can learn and make predictions at scale, and how we can use our unique data to gain insights into the ethnic composition, political inclinations, geographic distribution, and sentiments of our user base.

Teaching Machines to Understand Signals: A Large Scale Learning Approach
Gal Chechik, Google

As information moves from textual to complex signals like images, sounds, videos, it becomes crucial to develop methods to index signals at large scale. Most current model based approaches are still limited to small number of classes.
I will describe our approach to analyze images, sounds and videos by learning from large corpora of weakly labeled data. We focus on using low level features for which we can collect large statistics, rather than explicit shape modelling.
This approach allows us to learn measures of similarity between millions of images, with higher accuracy that has been achieved before. It provides ways to retrieve sounds from text queries, and search for specific events inside videos.

Community Detection with Block Models - Some Theory and Applications
Aiyou Chen, Bell Labs

I'll talk about our recent work about community detection with block models for random graphs, some theory about the consistency of various modularity algorithms with dense graphs, and a variety of applications with network data.
Joint work with Peter Bickel and Jin Cao.

A Novel Approach to Statistical Modeling of the Output Properties of High Level Auditory Neurons
Zhiyi Chi, University of Connecticut

Characterization of response properties of high level auditory neurons is essentially a high dimensional nonlinear regression problem. Due to biological constraints, the number of replicas available from individual neurons is very limited, making regression methods in current statistical literature infeasible. To understand how conspecific vocalizations are represented in the secondary forebrain auditory neurons of starling, the presented approach first isolates important components from acoustic signals behaviorally relevant to the animal, and uses the components as basis functions to reduce acoustic signals into a marked point process. It then treats the modeling of responses as an exercise of using a look-up table, by superimposing predicted responses to individual components to corresponding marked point locations. The talk is based on the work reported in
CD Meliza, Z Chi, D. Margoliash (2010). "Representations of Conspecific Song by Starling Secondary Forebrain Auditory Neurons: Toward a Hierarchical Framework", J. Neurophysiol 103: 1195-1208.

Analysis of Differential Neural Spiking Activity Using Point Process Models
Uri Eden, Boston University

A central problem in neuroscience is to determine whether two sets of spike trains represent information about the outside world in the same way. Such questions arise in assessing whether different neurons maintain similar representations of biological and behavioral signals, or in establishing whether a single neuron responds differently to different stimuli or changing contexts. Previously, ANOVA procedures have been used to determine whether place cells in rat hippocampus fire differentially before left or right turns during a spatial alternation task. However, this approach makes assumptions about the structure of the data, which often go unchecked, and requires an ad-hoc spatial discretization, which can drastically change the ability of the test to detect differential firing. In the meanwhile, point process modeling has been used successfully to characterize the statistical properties of neural firing activity. In this work, we expand on this point process modeling framework, and develop a general testing paradigm for examining if two collections of spike trains are likely to have been generated from the same process. The testing procedure involves fitting conditional intensity models to the observed spiking data and constructing test statistics from the resulting model fits. We identify a few useful test statistics: the Integrated Squared Error (ISE), the maximum difference (MD), and the likelihood ratio (LR) statistic. The sampling distributions associated with each of these test statistics can be estimated using bootstrap methods, or in the case of the LR statistic the asymptotic analytical distribution can be computed exactly. A simulation study and analysis of real data from rat hippocampus suggest that this testing procedure is able to detect differential firing robustly.

The Generalized Shrinkage Estimator For Spectral Analysis of Multivariate Time Series
Mark Fiecas, Center for Statistical Sciences, Brown University

We develop a new statistical method for estimating functional connectivity between neurophysiological signals represented by a multivariate time series. We use partial coherence as the measure of functional connectivity. Partial coherence identifies the frequency bands that drive the direct linear association between any pair of channels. To estimate partial coherence, one would first need an estimate the spectral density matrix of the multivariate time series. Parametric estimators of the spectral matrix provide good frequency resolution but could be sensitive when the parametric model is misspecified. Smoothing-based nonparametric estimators are robust to model misspecification and consistent but may have poor frequency resolution. In this work, we develop the generalized shrinkage estimator, which is a weighted average of a parametric estimator and a nonparametric estimator. The optimal weights are frequency-specific and derived under the quadratic risk criterion so that the estimator that does better at a particular frequency receives heavier weight. We validate the proposed estimator in a simulation study and apply it on electroencephalogram recordings from a visual-motor experiment.
Co-author: Hernando Ombao

A Hierarchical Spherical Radial Quadrature Algorithm for Gene Pathway Analysis
Jacob Gagnon, University of Massachusetts Amherst

Our work is concerned with the analysis of gene pathways/gene sets. The goal of gene pathway/gene set analysis is to identify gene groups (which are predefined) that are different between differing experimental or environmental conditions. We propose a logistic kernel machine to model the gene pathway effect with a binary response. Kernel machine methods are highly flexible taking into account interactions between genes as well as controlling for clinical covariates. Furthermore, we established a connection between our logistic kernel machine with GLMMs allowing us to use ideas from the GLMM literature. We adopt the spherical radial approach of Clarkson et al to perform the high dimensional integrations required for estimation of the fixed effects in the model and the testing of the genetic pathway effect. Simulation studies show that our estimation performance has comparable to better goodness of fit compared to Bayesian approaches at a much lower computational cost. As for testing of the genetic pathway effect, our REML likelihood ratio test based on spherical radial integration has increased power compared to a score test for simulated non-linear pathways. Additionally, our approach has 3 main advantages over previous methodologies: 1) our testing approach is self-contained rather than competitive, 2) our kernel machine approach can model complex pathway effects and gene-gene interactions, and 3) we test for the pathway effect adjusting for clinical covariates. Motivation for our work is the analysis of an Acute Lymphocytic Leukemia dataset where we test for the genetic pathway effect and provide confidence intervals for the fixed effects.

Paired Comparison Models with Tie Probabilities and Order Effects as a Function of Strength
Mark Glickman, Boston University

Paired comparison models, such as the Bradley-Terry model and its variants, are commonly used to measure competitor strength in games and sports. Extensions have been proposed to account for order effects (e.g., home-field advantage) as well as the possibility of a tie as a separate outcome, but such models are rarely adopted in practice due to poor fit with actual data. We propose a novel paired comparison model that accounts not only for ties and order effects, but recognizes two phenomena that are not addressed with commonly used models. First, the probability of a tie may be greater for stronger pairs of competitors. Second, order effects may be more pronounced for stronger competitors. This model is motivated in the context of tournament chess game outcomes. The models are demonstrated on the 2006 US Chess Open, a large tournament with players of wide-ranging strengths, and to the Vienna 1898 chess tournament, a double-round robin tournament consisting of 20 of the world's top players.

A Bayesian Discovery Procedure
Michele Guindani, University of New Mexico

We discuss a Bayesian discovery procedure for multiple-comparison problems. We show that, under a coherent decision theoretic framework, a loss function combining true positive and false positive counts leads to a decision rule that is based on a threshold of the posterior probability of the alternative. Under a semiparametric model for the data, we show that the Bayes rule can be approximated by the optimal discovery procedure, which was recently introduced by Storey. Improving the approximation leads us to a Bayesian discovery procedure, which exploits the multiple shrinkage in clusters that are implied by the assumed non-parametric model. We compare the Bayesian discovery procedure and the optimal discovery procedure estimates in a simple simulation study and in an assessment of differential gene expression based on microarray data from tumour samples. We extend the setting of the optimal discovery procedure by discussing modifications of the loss function that lead to different single-thresholding statistics. Finally, we provide an application of the previous arguments to dependent (spatial) data.

Rasch Model and its Extensions for Analysis of Aphasic Deficits in Syntactic Comprehension
Roee Gutman, Harvard University

Aphasia is a loss of the ability to produce and/or comprehend language, due to injury to brain areas responsible for these functions. Aphasic patients' performance on comprehension tests has traditionally been related to both personal ability and to difficulty of the test questions. The natural choice to analyze these test results is the Rasch model. It assumes that the probability of a patient responding correctly to a question is the inverse-logit function of the person's ability and the difficulty of the test question. This study

first modeled the way aphasic patients process different sentence types, and their ability to accomplish tasks using Rasch models. However, several scientifically important features of the data such as the correlation of correct responses between two different comprehension tasks, and the association between response patterns in control sentences to response patterns in experimental sentences, were found to be inadequately captured by such models. Using a full Bayesian approach, we explored a mixture of generalized linear mixed models that clustered patients into similar response patterns and abilities. The mixture model was found to better describe the experimental results than any other model examined. The mixture model also expresses the hypothesis that aphasic patients can be classified into different ability and response profile groups, and that patients utilize different cognitive resources in different comprehension tasks. These results are scientifically important and could not have been discovered by using the simple Rasch model.
Co-authors: Gayle DeDe, David Caplan and Jun S. Liu

Multi-scale Multiple Hypothesis Testing for Spike Trains
Matthew Harrison, Brown University

A recurring statistical problem in neuroscience and other fields is the identification of differences across experimental conditions. For neural spike trains, this often means identifying the location(s) in time (relative to some event) and the corresponding time scale(s) for which the firing rates are different across conditions. The multitude of locations, scales, and neurons creates a large multiple-testing problem. We observe that permutation tests using the well-known max-T or min-p methods are well suited for this situation. Unlike traditional permutations tests, however, the multiple testing corrections are not distribution free. We discuss robustness of the procedures to these assumptions.

Multilevel Models and Small Area Estimation in the Context of Vietnam Living Standards Surveys
Dominique Haughton, Bentley University and Toulouse School of Economics

This talk will discuss a methodology to obtain small area estimates in the context of the Vietnam Living Standards Surveys. The presentation will proceed in three parts. First we will introduce the Viet Nam Living Standards Surveys, their historical development, topics covered, sample size issues and challenges. Second, we will briefly review main concepts in small area estimation, including the use of auxiliary data, and will contrast simple small area models with regression small area models. This will then lead to the notion of random effects in small area regression models, and to our proposed multilevel model for small area estimation at the commune level in Vietnam, to our knowledge the first such model built with Vietnam living standards data. The third part of the talk will discuss this model. Our proposed multilevel model for estimating the commune-level mean (log of) household expenditure per capita relies on independent variables available both in the 1999 Census and in the Vietnam Household Living Standards Survey of 2002. Following ideas given in work by Moura (1994, 1999), the small area estimation is performed by plugging the population means of the independent variables into the regression equation, inclusive of suitable random effects both in the intercept and in the coefficient of the dummy variable for the urban location of a household. We will discuss how the random effects in the model can also be used to examine the urban-rural gap across the country. We will also mention how to measure the accuracy of our small area estimators. Finally, we will touch upon the use of sampling weights in models such as presented in the talk.
Co-authors: Phong Nguyen, Irene Hudson and John Boland

Assessing Geographical Variations in Hospital Processes of Care Using Multilevel Item Response Models
Yulei He, Harvard Medical School

National effort is directed toward developing and disseminating comparative information on some standardized processes of care for health care providers. We propose the use of Bayesian multilevel item response models to estimate hospital quality from multiple process measures and to assess their geographical variations. This approach fully incorporates the nesting structure of measures, patients, hospitals, and various levels of geographical units to provide a summary of the hospital quality. We apply the method to a national dataset of patients treated for a heart attack, heart failure, or pneumonia. We demonstrate considerable geographical differences in the quality of hospital care in these conditions. The

variations across census regions and states accounted for slightly more than 10% of the total variation. Some states performed well for all three conditions (e.g., the respective posterior probability of having better than the national average performance was 1 or close to 1 in Iowa, New Jersey, South Dakota, and Wisconsin). In contrast, some states varied across conditions (e.g., the corresponding posterior probability was close to 1 in Massachusetts for the care of heart attack and heart failure, but reduced to less than 0.5 for the care of pneumonia). The study results provide a comprehensive picture of hospital comparison at both regional and national level, and might be informative for policy development.

Customer Segmentation Using Nonparametric Clustering Methods of Categorical Time Series
Shan Hu, University of Connecticut

Customer segmentation enables a company to divide its heterogeneous customer market into several homogeneous groups according to similar needs for products or services provided by this company. Since marketing policies and campaigns can be targeted towards those homogeneous groups more efficiently and effectively, accurate segmentation can enhance the productivity and profitability of the company.
This talk will focus on effective clustering methods of categorical time series using quasi-distances based on first order and second order moment comparisons, and a combined quasi distance matrix as well. A quasi distance matrix based on first order moments will be obtained through likelihood ratio tests of the pairwise means. A quasi distance matrix based on second order moments will be obtained through likelihood ratio tests of the pairwise spectral densities. Hierarchical clustering using these quasi distance matrices enables us to segment customers of a large grocery store who are open to organic/natural grocery and personal products. The results would help the company to understand important characteristics of the target customers in terms of their purchasing in other product categories, keeping in mind health concerns, product uniqueness, product price and quality.
Co-authors: Nalini Ravishanker and Raj Venkatesan

Detecting and Understanding Overlapping Community Structure in Network
Guangying Hua, Bentley University

This study on social networks has been receiving a lot of attention recently. Community structure has been detected in some complex networks and also has been regarded as an important research area. Understanding the community structure of a network is of great importance to further understand the structural and functional properties of the network. While the vast majority of community mining methods cluster the data into mutually exclusive partitions, many real network datasets have inherently overlapping community structures. Overlapping community mining methods break the assumption that each node can only belong to one cluster and thus enables the analysis of a node¡¯s multiple role in the network.
This study provides a review of current overlapping community mining methods with a discussion of issues around community mining. One issue is the omission of a well-recognized community definition; the other is the evaluation of community mining results. As it is the most common method, Clique Percolation Method is discussed. We propose a new method to finding overlapping communities. Our method integrates clique percolation and modularity. The performance of the method will be tested with real network datasets.
The present study proposes a framework to analyze social networking data. First, a new method that extends clique percolation and modularity is proposed to find overlapping communities. Then, we test our results in real network datasets. In the end, content analysis will be used to verify the results. This is an ongoing research study. Our proposed framework will be implemented in R. Text mining is also considered to be part of our research.

Bayesball: Spatial Hierarchical Modeling of Fielding in Major League Baseball
Shane Jensen, University of Pennsylvania

We present sophisticated Bayesian methodology for the analysis of fielding performance in major league baseball. Our approach is based upon high- resolution data consisting of on-field location of batted balls. A key issue is the balance between the personal performance of an individual fielder and the shrinkage to the population performance of similar fielders. We combine spatial probit modeling with a hierarchical structure in order to evaluate individual fielders while sharing information between fielders at each

position. We present results across seven seasons of MLB data and compare our approach to other fielding evaluation procedures.

Multi-Objective fMRI Designs with Unequal Epoch Length via NSGA-II
Ming-Hung Kao, Arizona State University

Functional magnetic resonance imaging (fMRI) is a pioneering technology for studying brain activity in response to mental stimuli. To render precise inference, a design sequence of stimuli that simultaneously achieves high statistical efficiencies and avoids psychological confounds is called for. In this work, we aim at obtaining such a multi-objective design. In contrast to previous studies, we allow different epoch lengths for different stimulus types. In addition, we adapt the non-dominated sorting genetic algorithm II (NSGA-II) to search for good designs, and incorporate knowledge about fMRI designs to improve the effectiveness of the search. We demonstrate that the proposed approach outperforms current methods in use. Moreover, we propose a new criterion for evaluating designs' ability in circumventing psychological effects of anticipation and habituation. Rooted in the overlapping m-tuple test, the proposed criterion is sensitive to patterned design sequences and its value is easy to interpret.

Tracking Multiple Targets Using Binary Decisions from Wireless Sensor Networks
Natallia Katenka, Boston University

Wireless sensor networks (WSN) are a new technology with many applications, including environmental monitoring, surveillance, and health care. This work introduces a novel framework for tracking multiple targets over time using binary decisions collected by a wireless sensor network, and applies the methodology to two case studies: an experiment involving tracking people and a project tracking zebras. Unlike most existing methods, proposed tracking approach is based on a penalized maximum likelihood framework, and allows for sensor failures, targets appearing and disappearing over time, and complex intersecting target trajectories. The results show that binary decisions first corrected locally by a previously developed method known as local vote decision fusion provide the most robust performance in noisy environments and high accuracy in tracking applications.
Co-authors: Elizaveta Levina and George Michailidis

Evolutionary and Chromatin Signatures for Understanding the Human Genome and its Regulation
Manolis Kellis, MIT

Our group at MIT is focused on the computational underpinning of genomics, developing algorithms and machine learning techniques for studying complete genomes, understanding their regulatory constructs, and their evolutionary dynamics. We have defined evolutionary signatures in the nucleotide alignments of multiple related species, enabling the systematic discovery and characterization of diverse classes of functional elements, including protein-coding genes, RNA structures, microRNAs, developmental enhancers, regulatory motifs, and biological networks. We have also defined distinct chromatin signatures, or combinations of chromatin marks, in genome-wide epigenomic datasets, revealing numerous classes of promoter, enhancer, transcribed, and repressed regions, each with distinct functional properties. These techniques have enabled us to discover many new insights into animal gene regulation, including abundant translational read-through in neuronal proteins, functional anti-sense microRNA genes, overlapping functional elements encoded in human protein-coding genes, tissue-specific regulators for chromatin states, and thousands of novel long intergenic non-coding RNAs. Going forward, we are applying such techniques to understand the logic of global gene regulation during development and differentiation in human and fly, in the context of the ENCODE and modENCODE projects.

Circular Migrations and HIV Transmission Dynamics [CANCELLED]
Aditya Khanna, Quantitative Ecology and Resource Management, Univ. of Washington

The objective of this work is to investigate the impact of circular migrations on the transmission dynamics of HIV. Circular migrations involve the repetitive movement of people between two or more locations. An example is South Africa, where labourers have been sent from their home villages to mining towns (with periodic returns), and these migrations have been a major component of the South African economy. It is

known that HIV infectivity varies with time since infection. Since AIDS has a long latency period, the high infectivity of HIV before symptoms appear can have a major impact if infected individuals are changing locations and partners frequently, potentially without knowing their infection status. The interaction between timing since infection and variable infectivity is the major focus of the current work. I am investigating this interaction using compartmental ODE models that are an extension of the classic S-I-R structure, and a stochastic network-based framework based on Exponential Random Graph Models (ERGM). ODE models rely on homogeneous mixing and are easy to extend by adding vital dynamics. The ERGM framework allows us to model person to person transmission and consider relational timing, and hence, concurrency. I will present results from the two modeling frameworks that allow us to study the relative properties of the two approaches.
The stochastic model showed that longer average partnership durations slow down progression of disease. There was also evidence of slower migration rates corresponding to a longer time to achieve certain disease prevalence in the population.
Co-author: Steven Goodreau

Semi-Supervised Recusively Partitioned Mixture Models for Identifying Cancer Subtypes
Devin C. Koestler, Department of Community Health, Brown University

Patients with identical cancer diagnoses often have differing responses to therapy, and as a result, progress differently. The disparity we see in disease progression and treatment response can be attributed to the idea that two histologically similar cancers may be completely different diseases on the molecular level. Methods for identifying cancer subtypes associated with patient survival have the capacity to be a powerful instrument for understanding the biochemical processes that underlie disease progression as well as providing an initial step toward more personalized therapy for cancer patients. We propose a method called Semi-Supervised Recursively Partitioned Mixture Models (SS-RPMM) that utilizes array-based genetic and patient-level clinical data for finding cancer subtypes that are associated with patient survival. In the proposed SS-RPMM, cancer subtypes are identified using a selected subset of genes that are associated with survival time. Since survival information is used in the gene selection step, this method is semi-supervised. Unlike other semi-supervised clustering/classification methods, SS-RPMM doesn't require specification of the number of cancer subtypes, which is often unknown. In a small simulation study, our proposed method compared favorably to another competing semi-supervised method. Furthermore, an analysis of mesothelioma data using SS-RPMM, revealed at least 2 distinct methylation profiles that are informative for survival.
Co-author: E. Andres Houseman

Pathway-based Machine Learning Approaches for Cancer Classification
Mark Kon, Boston University

Machine learning methods can be important tools as information integration systems for inference of cancer prognosis, predicted therapy response, and classification. Though such methods have produced good results in classification tasks based on microarrays, the noise and diffuse information in individual gene signals have led to inconsistent biomarker sets for discrimination tasks. A useful approach to this problem has been the development of higher-level biomarkers (e.g. pathway activity markers), which have done a better job of both predicting and consistency across different data sets. As a step toward hierarchical biomarker classification methods, in which gene feature vectors are refined into sparser and more stable biomarker sets, we discuss a pathway-based feature inference algorithm called PathFeature. This approach is hierarchical in that a tree structure is assigned to biomarkers, in which leaves consist of individual data (i.e., individual gene expression features), while higher order nodes combine these features into sparser and more stable coherent markers. Initial importance sorting of pathway features is done using gene set enrichment analysis (GSEA), or a machine-based feature importance algorithm. Both methods give pathway feature selections which can stabilize biomarkers and improve prediction. The motivation is a network-based model in which layers feed through hierarchically to more complex derived feature indicators.

Spatiotemporal Kriging and Correlation Modeling of Wind Power Generation [CANCELLED]
Ada Lau, Oxford-Man Institute, University of Oxford

Wind power spatiotemporal forecasts are important since power grids are integrating an increasing number of wind farms in their portfolio, and spatiotemporal covariance structures should be exploited to generate more reliable probabilistic forecasts. We consider a spatiotemporal kriging approach, which assumes a Gaussian process and optimal predictions are obtained as linear combinations of neighborhood observations in space and time. However, wind power generation at an individual site is well known to have a highly non-Gaussian distribution. There is a discrete probability mass at zero due to their abundance, and some may also have a probability mass at the maximum capacity. Moreover, the continuous distribution is significantly right-skewed. We tackle the problems by considering a modified logistic transformation which could normalize the continuous part into an approximately Gaussian distribution. We then consider a logit model to describe the dynamics of the probability mass at zero and maximum capacity. To apply spatiotemporal kriging, a correlation structure is needed. We construct a non-separable, anisotropic correlation model to describe the correlation structure of wind power generated from 65 farms in Ireland. Our model successfully explains the dynamics due to the movement of weather front. Applying our correlation model to the kriging predictor, together with the discrete probability models at zero and one, we generate spatiotemporal probabilistic and point forecasts at 1-3 hours ahead, which is an important horizon for planning power dispatch strategies. We sum up individual forecasts and obtain aggregated point forecasts of wind power generation. Forecast performances are evaluated by proper scores and results demonstrate that our correlation model is superior to all other classes of models studied in previous literatures. Most importantly, our correlation model outperforms the aggregated point forecasts obtained by considering the univariate aggregated time series directly. Our approach is also computationally efficient and so online updating of forecasts can be easily obtained.

Limitations of Point Estimates in Computational Biology
Chip Lawrence, Brown University

Advances in genomics have rendered increasingly large data sets available for analysis. While the emergence of such large data sets seems to lead to increasingly more precise estimates of parameters, paradoxically just the opposite is becoming increasingly common. This paradoxical circumstance has emerged because these technologies have simultaneously opened opportunities to draw inferences on previously unanswerable high dimensional questions. As a result traditional highest scoring estimation methods such as maximum likelihood estimation or maximum a posteriori (MAP) estimation no longer enjoy the asymptotic favorable properties for which they rightfully became famous, and as a result can be seriously misleading. These unknowns are often of primary interest and often discrete. In computational biology it is common convention to describe these high-D discrete posterior distributions with a single point estimate, to use MAP estimates to obtain these estimates, and not to put confidence limits around these point estimates. Accepting the requirement that such point estimates are often necessary, I'll discuss and illustrate limitations of these conventions, and propose some alternatives.

Bayesian inference for Sparse Pair-wise Copula Models
Yoonjung Lee, Harvard University

Copulas are the functions that link marginal distributions into joint distributions. Pair-wise copula construction provides a flexible and powerful alternative to multivariate copula models that are becoming increasingly popular in modeling financial data. Aas, Czado, Frigessi, and Bakken (2007) demonstrate how one can generate a set of different pair-wise copulas systematically through regular vine structures. In our study, we adopt a D-vine structure and impose a bivariate t-copula for each pair, while introducing some latent variables to describe the conditional independence structure. We develop a Bayesian inference procedure that is suitable for the model selection problem when data are generated from a sparse set of parameters. We also discuss a dynamic extension of the model by incorporating possible structural shifts over time of the conditional independence structure and the model parameters. For an application, we analyze recent credit default swap spread data.
Co-authors: Claudia Czado and Andrew Vesper, Harvard University

Vectors of Poisson Dirichlet Processes
Fabrizio Leisen, Universidad de Navarra

The definition of vectors of dependent random probability measures is a topic of interest in applications to Bayesian statistics. They, indeed, define dependent nonparametric prior distributions that are useful for modelling observables whose values depend on covariates. In this work we propose a vector of two-parameter Poisson-Dirichlet processes. It is well-known that each component can be obtained by resorting to a change of measure of a σ-stable process. Thus dependence is achieved by applying a L\'evy copula to the marginal intensities. In a two-sample problem, we evaluate the corresponding partition probability function which turns out to be partially exchangeable. Moreover, we evaluate predictive and posterior distributions. This is a work in collaboration with Antonio Lijoi (University of Pavia).

Dependence of Spike-Field Coherence on Expected Intensity
Kyle Lepage, Boston University

Within the primate monkey brain, cognitive attention evokes inter-regional synchronization between neuronal firing and the local-field potential (Gregoriou, G. et al, 2009). Important to this discovery is the estimation of coherence between the times at which a neuron spikes and a collection of measurements acquired periodically in time. Hence, this coherence is non-standard, in that it is computed between two time-series that are typically modeled in differing fashions, ie., between a point process and a weak-sense stationary, discrete random process.
Simulations indicate that a standard class of coherence estimators, the non-parametric, direct-type, multiple-taper coherence estimators, respond to the expected number of neuron activations in addition to the per frequency degree of linear dependence between the time-series. This property confounds frequency dependent linear association with overall neuron spiking activity and confuses two quantities with differing physical interpretations. In this talk, the response of these estimators to the overall spiking activity of neurons is demonstrated to be a property of the population coherence as opposed to a manifestation of estimator bias for three statistical models of neuron firing activity. These models are doubly-stochastic point processes and are shown to be sufficiently general to capture salient features of actual neuron firing while remaining stationary.
After introducing and defining relevant quantities, the talk develops the point process models and their stationarity. Population coherence is demonstrated to depend on the expected conditional intensity and this dependence is quantified. Realizations of these point processes are compared to actual spiking data and shown to capture salient features, both in the time-series but also in estimated spectra. Time-permitting, the standard, multiple-taper coherence estimator is presented and shown to be un-biased, up to effects due to the finite length of the time-series.
Co-authors: Uri Eden, Mark Kramer, Georgia Gregoriou, Steve Gotts, Robert Desimone, Nancy Kopell

Sequential Analytic Methods for Post-marketing Safety Surveillance Using Existing Healthcare Databases
Lingling Li, Harvard Medical School & Harvard Pilgrim Health Care Institute

Drug safety surveillance is an important public health need. The Food and Drug Administration (FDA) has launched the Sentinel Initiative aiming to develop and implement a national electronic system to actively query diverse healthcare databases (e.g., administrative and insurance claims, electronic health records, registries) to evaluate possible medical product safety issues in a prompt manner. In this talk, we will introduce a practical group sequential method, a conditional sequential sampling procedure (CSSP), to sequentially examine if the drug of interest has an elevated relative risk for a selected adverse event compared to a comparison drug. The CSSP applies to settings with con-current controls, i.e., information for both the drug of interest and the comparison drug is anticipated to accumulate over time. The CSSP models the selected adverse event(s) and adjusts for population heterogeneity based on selected important confounders (e.g., age groups, gender, race, geographic regions). It also automatically adjusts for temporal trend.

Species Sampling Models for Bayesian Nonparametric Inference
Antonio Lijoi, University of Pavia

Sampling problems from populations which are made of different species arise in a variety of ecological and biological contexts. Basing on a sample of size n, one of the main statistical goals is the evaluation of species richness. In order to deal with this issue, we undertake a Bayesian nonparametric approach based on Gibbs-type priors which include, as special cases, the Dirichlet and the two-parameter Poisson-Dirichlet processes. We show how a full Bayesian analysis can be performed and describe the corresponding computational algorithm.

Spatial Process Model for Social Networks
Crystal Linkletter, Brown University

With concerns of bioterrorism, the advent of new epidemics that spread with person-to-person contact, such as SARS, and the rapid growth of on-line social networking websites, there is currently great interest in building statistical models that emulate social networks. Stochastic network models can provide insight into social interactions and increase understanding of dynamic processes that evolve through society. A major challenge in developing any stochastic social network model is the fact that social connections tend to exhibit unique inherent dependencies. For example, they tend to show a lot of clustering and transitive behavior, heuristically described as "a friend of a friend is a friend." It might be reasonable to expect that covariate similarities, or "closeness" in social space, should somehow be related to the probability of connection for some social network data. The relationship between covariates and relations is likely to be complex, however, and may in fact be different in different regions of the covariate space. Here, we present a new socio-spatial process model that smoothes the relationship between covariates and connections in a sample network using relatively few parameters, so the probabilities of connection for a population can be inferred and likely social network structures generated. Having a predictive social network model is an important step toward the exploration of disease transmission models that depend on an underlying social network.

Factor Rotation of Functional Data for Extracting Periodic Objects
Chong Liu, Boston University

In this paper, we develop a methodology for rotating factors of functional objects with a goal towards uncovering latent functions of specified structures. Applied to functional principal components our methodology provides the foundation for functional factor analysis. We show that the computationally intensive step of rotation of functions is mathematically equivalent to rotation of vectors in the basis space and thus can be solved using standard multivariate canonical correlation analysis. Here, we apply our technique to decompose the annual and extra annual variations of vegetation indices obtained from remote sensing data.
Co-authors: Surajit Ray, Giles Hooker and Mark Friedl

On the Stationary Distributions of the Chained Imputations
Jingchen Liu, Columbia University

Chained imputation is a widely used imputation procedure especially for computer packages. It provides substantial convenience in designing the imputation procedures by means of regression models. However, the theoretical properties are largely unexplored. In this paper, we provide analysis for the existence and characterization of the stationary distributions of chained imputations. In particular, we give a set of sufficient conditions under which the Markov chain is positive recurrent and the invariant distribution converges in total variation to the posterior distribution of a Bayesian model. This paper also provides explanations to the empirical findings that it is usually hard to construct an example in which chained imputation results in a transient Markov chain when there is enough observed data. The analysis technique consists of constructing coupling of Markov processes and bound the total variation distance between their stationary distributions by their convergence rates.
Co-authors: Andrew Gelman, Jennifer Hill and Yu-Sung Su

The Shannon-McMillan-Breiman Theorem for Log-Concave Distributions
Mokshay Madiman, Yale University

Suppose a random vector taking values in n-dimensional Euclidean space has a log-concave density f. We show that the nonparametric log-likelihood function, namely $log f(X)$, is highly concentrated around its mean, namely the negative entropy, with the strength of concentration increasing exponentially with dimension n. This concentration property implies in particular an extension of the Shannon-McMillan-Breiman theorem to the class of discrete-time stochastic processes with log-concave marginals.

Statistical Inference in Factor Analysis for High-Dimensional, Low-Sample Size Data
Miguel Marino, Harvard School of Public Health

Cancer researchers are keen on tracking trends in cancer mortality rates and studying the cross relationship of these trends not only for scientific reasons of understanding the cancers as a complex dynamical system, but also for practical reasons such as prevention, planning and resource allocation. Factor analysis which studies such cross-correlation matrices is an effective means of data reduction, whose inference typically requires the number of random variables, p, to be relatively small and fixed, and the sample size, n, to be approaching infinity. However, contemporary surveillance techniques have yielded large matrices in both dimensions, limiting the usage of existing factor analysis techniques due to the poor estimate of the covariance/correlation matrix. We develop methods, in the framework of random matrix theory, to study the cross-correlation of cancer mortality annual rate changes in the setting where p > n. We propose methodology to test complete independence across cancer sites. We develop an approach based on group sequential theory to determine the number of significant factors in a factor model. Sparse principal components analysis is studied on the principal components deemed to be significantly different than random matrix theory prediction to aid in the interpretation of the underlying factors. Methods are implemented on SEER cancer mortality rates from 1969 through 2005.
Co-author: Yi Li

The Intersectome: Integration of Knowledge in Systems Biology for Hypothesis Generation
Avi Ma'ayan, Mount Sinai School of Medicine

I will discuss how we utilize data collected from the public domain, describing regulatory interactions in mammalian cells, to analyze results from experiments that profile cells using a variety of cutting-edge genome-wide technologies. The results from our analyses produce rational hypotheses for further experimental validation as well as provide a global view of cell regulation across multiple layers. Specifically, I will show GATE, a program we developed for the analysis of time-series expression data used to analyze stem cell differentiation. I will also demonstrate Genes2Networks a program we developed and used to predict components and pathways essential for CB1R induced neurite outgrowth in Neuro2A cells, as well as to predict a novel disease gene that causes Noonan-like syndrome. Finally, I will discuss our tools KEA for the analysis of SILAC phosphoproteomics, and ChEA for the analysis of gene expression data using a ChIP-X database we are developing. I will conclude with a proposition of ideas about developing robust theoretical models for candidate drug/gene/protein rankings for functional experimental validation.

Asymptotic Distribution of Poisson Index of Dispersion
Liang Meng, Boston University

In neuroscience, the index of dispersion, or Fano factor, is widely used for determining whether neural spiking signals are Poisson in nature. In order to construct probability bounds on this statistic for the Poisson case or to perform an explicit hypothesis test, it is first necessary to determine its sampling distribution. Hoel (1943) and N. Kathirgamatamby (1953) previously argued that under the Poisson assumption, the index of dispersion asymptotically has a asymptotic chi square distribution with degrees of freedom equal to the number of observations minus one, as both the number of observations and the expected number of occurrences per observation go to infinity. However, this work only considered the convergence of the first four moments of the index of dispersion rather than the whole distribution. In this paper, an alternative method is proposed to show the convergence of Poisson index of dispersion to the chi square in distribution. Moreover, the convergence rate and accuracy of a chi square approximation for the index are discussed. Examples of applications of this result to the analysis of neural spike train data are discussed as well.

Co-author: Uri Eden

Sample Size Calculation for Poisson Endpoint Using the Exact Distribution of Difference between Two Poisson Random Variables
Sandeep Menon, Boston University

In many clinical trials, the clinical endpoint, especially MRI endpoints in neurology trials, follows a Poisson distribution. We propose a method that uses the exact distribution of the difference between two Poisson variables to calculate sample size at the protocol design stage. When the difference between the two Poisson rates is more than 1.2 units, the number of subjects and events needed at the desired power and type I error rate is slightly less than that computed by simulation based on the normal approximation method. The exact sample size calculations are more comparable to the normal approximation when the difference between the rates is less than 1.2 units. The proposed method is more intuitive, efficient and less subjective compared to the normal approximation method. A simple code is developed in R-software to estimate the sample size and critical values.
Co-authors: Joseph Massaro, Jerry Lewis, Michael Pencina, Yong-Cheng Wang and Philip Lavin

A Nonparametric Test for the Validation of Surrogate Endpoints
Xiaopeng Miao, Boston University School of Public Health

We present a novel nonparametric test to validate surrogate endpoints based on multivariate density estimation and permutation test. This test is the first in literature to directly verify the Prentice statistical definition for surrogacy. The test does not impose distributional assumptions on the endpoints and it is robust to model misspecification. Our simulation study show that the proposed nonparametric test outperforms the practical test of the Prentice criterion in terms of both robustness of size and test power. The method is applied to the validation of MRI lesions as the surrogate endpoint for clinical relapses in multiple sclerosis trials.
Co-authors: Yong-Cheng Wang, Ashis Gangopadhyay

Domain Adaptation Theory and Algorithms
Mehryar Mohri, Google & New York University

Earlier learning theory and algorithms were developed for an ideal world. Modern large-scale data sets and many large-scale applications bring forth problems that must be addressed for learning to be effective, e.g., training points are often poorly labeled, the sample can be biased, the distributions may drift with time, and the sample points may not be i.i.d.
This talk will address the specific problem of domain adaptation which arises when the distribution of the source labeled data somewhat differs from that of the target domain. It will present novel theoretical results for adaptation and provide algorithmic solutions derived from that theory. It will also report experimental results in a sentiment analysis task.
Joint work with Yishay Mansour and Afshin Rostamizadeh.

Odds Ratio Models in Sports
Carl Morris, Harvard University

In paired comparison settings, odds ratio and logistic regression models arise theoretically and as approximations. These models are used to predict winning and other binary events in sports, via estimates of model parameters that measure team and player strengths. Applications include baseball (e.g. Bill James' "Pythagorean" formula), basketball, football, and tennis. Part of this talk is from a paper with Jason Rosenfeld, Dan Adler, and Jake Fisher of the Harvard Sports Analysis Collective (http://harvardsportsanalysis.wordpress.com) entitled "Predicting Overtime with the Pythagorean Formula", soon to appear in the Journal of Quantitative Analysis in Sports.

Geometric Representations of Hypergraphs for Prior Specification and Posterior Sampling
Sayan Mukherjee, Duke University

A parametrization of hypergraphs based on the geometry of points in a compact space is developed. Informative prior distributions on hypergraphs are induced through this parametrization by priors on point configurations via spatial processes. This prior specification is used to infer conditional independence models or Markov structure of multivariate distributions. Specifically, we can recover both the junction tree factorization as well as the hyper Markov law. This approach offers greater control on the distribution of graph features than Erdös-Rényi random graphs, supports inference of factorizations that cannot be retrieved by a graph alone, and leads to new Metropolis/Hastings Markov chain Monte Carlo algorithms with both local and global moves in graph space. We illustrate the utility of this parametrization and prior specification using simulations.
Joint work with Simon Lunagomez and Robert Wolpert.

Using Genes as Instrumental Variables in Analyses of Social Network Data
James O'Malley, Harvard Medical School

We describe methodology for evaluating peer effects in a social network with respect to health behavior using an "instrumental variable" approach. The key idea is that a person (an "ego") may have peers (or "alters") who are randomly "assigned" genes predisposing the alters to certain health behaviors, and that this random assignment can be seen as a kind of natural experiment, exposing the ego to peers who either exhibit or do not exhibit the pertinent behaviors. Using non-parametric two-stage IV estimation and alternative structural equations models, we will assess whether obesity and BMI in an ego's alters (e.g., friends, siblings), are causally related to similar behaviors in the ego and if so test the more complex hypothesis that similarity effects will vary according to the closeness of the relationship between the ego and the alter.

New Statistical and Computational Challenges in Next-Generation DNA and RNA Sequencing
Nicola Neretti, Brown University

Next-generation DNA and RNA sequencing is currently one of the driving forces in genomics and transcriptomics. While sequencing costs are constantly decreasing, they are still limiting the use of biological replicates in most experimental studies. In this talk I will focus on the computational challenges of studying protein binding in the repetitive regions of the genome through sequencing data, and on the statistical problem of testing differences between experimental conditions. I will show that, because of the large number of sequenced fragments available, it is crucial to include biological replicates in the analysis to reliably detect these differences.

Shape Constrained Statistical Estimation via Semidefinite Optimization
David Papp, Rutgers University

Statistical estimation problems often involve functions that must satisfy certain shape constraints, which usually reduce to the nonnegativity of linear functionals of the approximating function. If the estimator is a spline, then shape constraints take the form of conic inequalities with respect to cones of nonnegative polynomials. Some of these shape constraints are tractable (especially in the univariate case), while in other cases we need to consider tractable restrictions, such as those involving weighted-sum-of-squares cones. We present theoretical justifications of the proposed approach, as well as computational results.
Co-author: Farid Alizadeh

A Graph Log-Linear Model for Characterizing Repeated Interactions
Patrick O. Perry, Harvard Statistics and Information Sciences Lab

We are surrounded by interaction data. More and more, this data is being harnessed for inferential purposes. Phone and email records are being used to study communications networks. Records of Legislation cosponsorship and journal coauthorship are being used to study collaboration networks. Animal co-occurrence data is being used to study herding and association behaviors. Network scientists have risen to the challenge of analyzing these new types of data, delivering a host of promising and powerful new methods and models. Most network models are designed for modeling binary relations (e.g. whether or not two people are friends). These methods are not directly applicable to data with varying frequencies of

interactions between actors. We propose a simple modeling framework to handle data with repeated interactions.
We introduce a "graph log-linear linear model", or graph-LLM, to predict the frequency of interaction between pairs of actors. This is a graph with a log-linear model on each edge. The number of interactions on edge (i,j) is modeled in terms of the edge's endpoints and other edge-specific covariates. We show how to estimate the parameters of the model and also give asymptotic results showing that the estimates are consistent when the observation interval or the number of nodes goes to infinity. We validate our modeling framework using a subset of the emails sent within the Enron Corporation. Over small windows, we demonstrate the interactions between people to be approximately pairwise independent and time homogeneous. Using past patterns of interaction to estimate the connectivity graph and model parameters, the graph-LLM model is able to outperform a log-linear model that ignores the network structure.
Co-author: Patrick J. Wolfe

Intelligence and Learning Theory: Kernels and Derived Kernels
Tomaso Poggio, MIT

Learning is the gateway to understanding intelligence and to reproducing it in machines. In this context, I will introduce modern learning theory and sketch some of its mathematical foundations centered on conditions for prediction. A classical example of learning algorithms is provided by regularization in Reproducing Kernel Hilbert Spaces. Neuroscience, however, suggests a hierarchical architecture for learning which classical algorithms do not have. I will describe a new attempt (with S. Smale) to develop a mathematics for hierarchical kernel machines - centered around the notion of a recursively defined "derived kernel" - and directly suggested by the neuroscience of the visual cortex. I will conclude with a proposal to extend this feedforward architecture to reflect the recursive organization of cortex.

Repeated Significance Tests in Presence of Random Costs
Vladimir Pozdnyakov, University of Connecticut

For many tasks of reconnaissance and surveillance, networks of spatially distributed sensors provide the low-cost, low-risk solution, and the rapidly growing field of distributed sensing now provides many interesting challenges for probabilistic modeling . The main purpose of this talk is to examine a simple model that joins the rudiments of sequential decision making with the engineering constraint of a fixed budget for the cost of the transmissions sent to and from the distributed sensors. Here the costs associated with transmissions from the sensors are intended to either real physical expenditures, such as battery life, or to capture more subtle costs, such as the cumulative risk of a remote sensor (or bug) being found by an adversary.
Co-authors: Joseph Glaz, Marco Guerriero, J. Michael Steele and Peter Willett

Sequential Numerical Integration and Stochastic Optimization with Statistical Designs
Peter Qian, University of Wisconsin-Madison

Numerical integration and stochastic optimization are at the heart of many estimation problems in statistics, data mining and machine learning. Examples include marginalization of nonlinear likelihood functions, imputation of missing data in regularized variable selection, maximum likelihood estimation of hierarchical models, and posterior mode calculation of Bayesian models.
The first part of my talk is devoted to sequential numerical integration with statistical designs. I will discuss several new sampling designs for accurately estimating multi-dimensional integrals in a sequential fashion. These designs are combinational in nature and are constructed by exploiting nesting in stratified permutations, difference schemes, linear codes and other discrete structures.
The second part of my talk deals with stochastic programming with space-filling designs. A stochastic program is an optimization problem in which the objective function is a multi-dimensional integral. The popular Sample Average Approximation method solves a stochastic program by constructing a sampling based approximation to the objective function and then finding the solution of the approximated problem. I will report some recent advances in using statistical designs to enhance the accuracy of this method.

Regularization Methods for Sequential Prediction
Alexander Rakhlin, University of Pennsylvania

Prediction of individual sequences is studied in several fields: machine learning, information theory, prequential statistics, and game theory. First, we show that regularization methods play a central role in the abstract setting of sequential prediction as they provide Hannan-consistent strategies. Second, by considering the minimax value of the sequential prediction game, we can address questions such as 'under which assumptions is an i.i.d. data source close to the worst-case sequence crafted by Nature?'

On the Pricing of Eurodollar Futures
Balaji Raman, University of Connecticut

Heath, Jarrow and Morton's term structure model is a standard tool for the analysis of fixed-income securities and their associated derivatives, example -- modeling Eurodollar futures prices. A Eurodollar is a Dollar deposited in banks outside the United States. A specific HJM model is fully determined by a choice of volatility structure. This is attributed to the forward rate drift restriction of HJM models. A plot of monthly quadratic variation of the price process is useful to measure the adequacy of different HJM structures to model these prices. However, this graphical tool does not identify "one suitable" model.
In this talk, we will discuss how the price of options on Eurodollar futures can be used for validation of a specific choice of an HJM model. Moreover, a different usage of options is to enhance the estimation of the parameters of an HJM model. For example, within Gaussian framework, options data is often useful to fit a multi-parameter HJM model.
Co-author: Vladimir Pozdnyakov

A Spatial Dirichlet Process Mixture Model for Clustering
Brian Reich, North Carolina State University

Recent advances in tools for molecular genetics, along with greater computational power, has led to a developing field known as landscape genetics. This emerging area examines the interactions between environmental features and microevolutionary processes, such as gene flow, genetic drift, and selection. In this paper we develop a Bayesian clustering algorithm based on the Dirichlet process prior that uses both genetic and spatial information to classify individuals into homogeneous clusters for further study. We study the performance of our method using a simulation study and use our model to cluster wolverines in Western Montana using microsatellite data.

How Many Modes Can a Mixture of Two Components Have?
Dan Ren, Boston University

Multivariate normal mixtures are widely used for modeling homogeneous population and clustering. The density shapes arising from multivariate mixtures are often very complex. Ray and Lindsay (2005) presents a unified theory for understanding the topography of high dimensional normal mixtures. Their main result shows that their topography, in the sense of their key features as a density, can be analyzed rigorously in lower dimensions by use of a ridgeline manifold that contains all critical points as well as the ridges of the density. Modes of densities is often of primary interests. Analysis show that the number of modes depends on the means and eigenvalues of the mixture components. In this paper we provide a tight upper bound (equal to Dimension+1) on number of modes of a two component mixture of normals with arbitrary variance covariance matrix. The recursive process of constructing such a normal mixture with maximum number of modes in any dimension is also provided.
Co-author: Surajit Ray

On the Geometry of Discrete Exponential Families with Application to Exponential Random Graph Models
Alessandro Rinaldo, Carnegie Mellon University

There has been an explosion of interest in statistical models for analyzing network data, and considerable interest in the class of exponential random graph (ERG) models. In this talk I will relate the properties of

ERG models to the properties of the broader class of discrete exponential families. I will describe a general geometric result about discrete exponential families with polyhedral support. Specifically, I will show how the statistical properties of these families can be well captured by the normal fan of the convex support. I will discuss the relevance of such results to maximum likelihood estimation and apply them to the analysis of ERG models. By means of a detailed example, I will provide some characterization and a partial explanation of certain pathological features of ERG models known as degeneracy.
Joint work with S.E. Fienberg and Y. Zhou.

Priors on Topological and Metric Spaces: A Computational Perspective
Daniel Roy, MIT

Most popular Bayesian nonparametric priors admit closed form expressions for posterior inference. Is our focus too narrow? Using results in computable analysis and recent results in computable probability theory (Freer and Roy 09, Horyrup and Rojas 09, Freer and Roy 10, Ackerman, Freer and Roy, preprint), I will describe a general (and in some sense complete) framework for constructing priors on (nice) topological and metric spaces in such a way that important operations are computable, though not necessarily of closed form.

An Equivalence between AdaBoost and RankBoost
Cynthia Rudin, MIT

In machine learning, "classification" and "bipartite ranking" are distinct problems with distinct goals. Classification algorithms try to minimize the number of examples on the wrong side of a decision boundary, whereas bipartite ranking algorithms try to rank as many positives above as many negatives as possible, without respect to a decision boundary. We prove an equivalence between two well-known algorithms for these separate goals. Specifically, we show that the classification algorithm "AdaBoost" (Freund & Schapire, 97) achieves a ranking error that is equally as good as the one produced by the ranking algorithm "RankBoost" (Freund, Iyer, Schapire, Singer 03). Furthermore, the solution of RankBoost can be trivially altered to produce a classification error as good as that of AdaBoost.
This is joint work with Rob Schapire.

One-Shot Learning with a Hierarchical Nonparametric Bayesian Model
Russ Salakhutdinov, MIT

In typical applications of machine classification algorithms, learning curves are measured in tens, hundreds or thousands of training examples. For humans learners, however, the most interesting regime occurs when the training data are very sparse. Just a single example of a novel category is often sufficient for people to grasp a concept and make meaningful generalizations to novel instances. In this talk we will present a nonparametric hierarchical Bayesian model that aims to capture this human-like pattern of one-shot learning. The proposed model leverages higher-order knowledge abstracted from previously learned categories to estimate the new category's prototype as well as an appropriate similarity metric from just one example. We provide an efficient MCMC algorithm and show on several real-world datasets that the model is able to learn useful representations of novel categories based on a single training example.
Joint work with Josh Tenenbaum and Antonio Torralba.

Facing the Supernova Challenge: Complex Theory and Complex Data
Chad Schafer, Carnegie Mellon University

Optimal utilization of the increasing flood of astronomical data will require the development of statistical methods for combining physical understanding and observational evidence into optimal constraints on cosmological models. Analytical and simulation models, built upon well-refined theory, are able to resolve the evolution of the Universe to unprecedented detail. Immense effort is placed on data collection in astronomy. I will describe initial steps taken by our group towards bridging the gap between these two by constructing and implementing methods for statistical inference which fully utilize the available data, while adhering to the constraints imposed by current theoretical understanding. These steps run the range from dimension reduction techniques to classically-motivated confidence set construction. I will use the recently-

posed Supernova Photometric Classification Challenge as a framework for exploring the issues and our recent work.
This work is in collaboration with the InCA Group, including Peter Freeman, Ann Lee, and Joseph Richards.

Shape Restricted Function Estimation and Inference in Non-standard Problems
Bodhisattva Sen, Columbia University

The talk will focus on estimation and bootstrap based inference for nonparametric shape restricted functions (e.g., monotonicity/convexity restrictions). The first part of the talk will consider some issues with the consistency of different bootstrap methods for constructing confidence intervals in non-standard problems, i.e., statistical problems where estimators converge at non-root-n rate and/or have non-normal distribution. As the estimation of shape restricted functions exhibit such non-standard behavior, they will be the focus of our study. Further examples, that include change point models, will also be shown that illustrate the poor finite sample and asymptotic properties of the naive bootstrap method.
In the second part of the talk we consider the nonparametric least squares estimation of a multivariate convex regression function (under the known convexity constraint) and discuss its computation, characterization and consistency.

To Explain or To Predict?
Galit Shmueli, University of Maryland

Statistical modeling is at the core of many scientific disciplines. Although a vast literature exists on statistical modeling, on good practices, and on abuses of statistical models, the literature lacks the discussion of a key component: the distinction between modeling for explanatory purposes and modeling for predictive purposes. This omission exacts considerable cost in terms of advancing scientic research in many fields. In this talk I highlight these two uses of empirical modeling in scientific research and the differences that arise in each of the two statistical modeling paths.

The International Digital Divide: An Application of Kohonen Self Organizing Maps
Maria Skaletsky, Bentley University

We employ the method of Kohonen Self Organizing Maps (SOM) to examine the digital divide on a panel of 180 countries. The SOM is a data analysis technique that presents multi-dimensional data on a two-dimensional hexagonal grid. We examine the digital divide indicators alongside with economic, infrastructure and demographic variables to identify clusters between countries over a range of time. The SOM provides visualization of the digital divide and the changes in country groupings over time. We extend the work of Deichmann et al (2007), who used similar indictors over the period of three years, by considering a wider range of seven years and by using additional ICT indicators, such as mobile phone subscriptions per 100 people.
Co-author: Mayokun Soremekun

Causal Inference for Continuous Time Processes When Covariates Are Observed Only at Discrete Times
Dylan Small, University of Pennsylvania

Most of the work on the structural nested model and g-estimation for causal inference in longitudinal data assumes a discrete time underlying data generating process. However, in some observational studies, it is more reasonable to assume that the data are generated from a continuous time process, and are only observable at discrete time points. When these circumstances happen, the sequential randomization assumption in the observed discrete time data, which is essential in justifying discrete time g-estimation, may not be reasonable. Under a deterministic model, we discuss other useful assumptions that guarantee the consistency of discrete time g-estimation. In more general cases, when those assumptions are violated, we propose a controlling-the-future method that provides at least as good performance as g-estimation in most scenarios, and provides consistent estimation in some cases in which g-estimation is severely inconsistent. We apply the methods discussed in this paper to simulated data, as well as to a data set

collected following a massive flood in Bangladesh, estimating the effect of diarrhea on children's height. Results from different methods are compared in both simulation and the real application. This is joint work with Mingyuan Zhang and Marshall Joffe.

Identifying Differentially Expressed Genes in Time Series Microarrays
Jonathan J. Smith, MIT

Modern microarray technologies make it feasible to assess in parallel more than 50,000 gene expressions. When biological samples are drawn from different phenotypes, e.g., cancerous vs. benign, we can identify differentially-expressed genes that are crucial for distinguishing such phenotypes. In non-time series studies, many methods have been proposed for gene selection, such as fold change, t-statistics, and Bayes factors. In time series studies, gene selection methods are ad hoc -- merely selecting genes that are differentially expressed at isolated times. These methods neglect an important fact that genes interact with each other over time. To capture such longitudinal dependencies, we consider a first-order Markov chain to model the series, using continuous random variables to represent gene expression and discrete binomial random variables to represent phenotypes. The Markov chain model and the gene-phenotype dependence model are complementary, and can be unified in the framework of Bayesian networks to achieve identification of crucial genes for biological studies. The network is structured such that the gene expression at time T forms a node, which links to the gene expression at time T+1. There is an additional node for the phenotype variable, which is isolated if the time series is independent of it; otherwise, the phenotype node is linked to any node in the time series. When a gene is a crucial gene, its expression time series must be dependent on its phenotype, thus creating a Bayesian conditional probability relationship between the gene expression time series and its phenotype. Using this Bayesian network, we can calculate the Bayes factor by calculating the likelihood of a phenotype-dependent time series and comparing it to the likelihood of a phenotype-independent time series. We evaluate our method using breast cancer data, which is publicly available in Gene Expression Omnibus with accession number GSE11352. Our method identifies 40 genes crucial to breast cancer progression. The biological analysis of these genes confirms that they involve in cell death, developmental disorder, and endocrine system disorder, which are prerequisites of breast cancer.
Co-authors: Hsun-Hsien Chang and Marco F. Ramoni

Comparing Multiple Treatments to Both Positive and Negative Controls
Eleanne Solorzano, University of New Hampshire

In the past, most comparison to control problems have dealt with comparing k test treatments to either positive or negative controls. Dasgupta et al. [2006. Using numerical methods to find the least favorable configuration when comparing k test treatments to both positive and negative controls. Journal of Statistical Computation and Simulation 76, 251–265] enumerate situations where it is imperative to compare several test treatments to both a negative as well as a positive control simultaneously. Specifically, the aim is to see if the test treatments are worse than the negative control, or if they are better than the positive control when the two controls are sufficiently apart. To find critical regions for this problem, one needs to find the least favorable configuration (LFC) under the composite null. In their paper, Dasgupta et al. [2006. Using numerical methods to find the least favorable configuration when comparing k test treatments to both positive and negative controls. Journal of Statistical Computation and Simulation 76, 251–265] came up with a numerical technique to find the LFC. In this paper we verify their result analytically. Via Monte Carlo simulation we compare the proposed method to the logical single step alternatives: Dunnett's [1955. A multiple comparison procedure for comparing several treatments with a control. Journal of the American Statistical Association 50, 1096–1121] or the Bonferroni correction. The proposed method is superior in terms of both the Type I error and the marginal power.
Co-authors: N. Dasgupta and T. Tong

Barriers to the Practice of Really Reproducible Research
Victoria Stodden, Yale University

As computation becomes more pervasive in scientific research, if code and data are not made available we miss a crucial opportunity to control for error, the central motivation of the scientific method, through

reproducibility. In this talk I present results from a survey of the Machine Learning community uncovering the factors underlying the decision whether or not to share code and data. Intellectual property issues are surprising salient barriers and the "Reproducible Research Standard" is presented to realign intellectual property law with longstanding scientific norms through open licensing.

Estimate Absolute Concentrations of Transcripts from DNA Microarrays
Yunxia Sui, Brown University

Microarrays quantify the abundance of nucleic acids via hybridization. As the binding efficiency of probes can vary greatly, the apparent expression measurement is affected by the gene expression level as well as by the probe effects. Thus, most microarray studies provide only a relative measurement of gene expression for the same gene in different samples. Neither the expression levels of different genes within a sample nor the measurements of gene expression taken on different microarray platforms can be directly compared. Consequently, methods using probe sequences to predict probe behavior in hybridization have been proposed to extract absolute concentration on gene expression. However, the small amount of calibration data limits the power of sequence-only models. We demonstrate that, by taking advantage of a large database of samples in a combination of sequence models, the probe efficiency can be estimated with smaller variance. Gene expression measures adjusted for probe efficiency allow the comparison of expression between genes, as well as the comparison of the same gene measured on different platforms. Co-author: Zhijin Wu

Solving Least Squares via Gaussian Belief Propagation
Sekhar Tatikonda, Yale University

Belief propagation is a simple distributed algorithm for computing marginal distributions. We show how Gaussian belief propagation (GaBP) can be used to solve many least square problems. We discuss the convergence and correctness aspects of GaBP.

Semiparametric Bayesian Inference in Functional Nonlinear Regression Models for Panel Time Series Data
Sylvie Tchumtchoua, University of Connecticut

Panel time series consist of measurements recorded on a large number of individuals over a large number of time points. Such data are becoming increasingly popular in various fields (e.g. online auction prices in electronic commerce, functional magnetic resonance images in psychology, scanner data in marketing, repeated observations in clinical trials, etc.). When the number of time points is large, it is natural to expect that the nature of the relationships among variables will change over time as well as vary across individuals. We propose a semiparametric Bayesian model for multivariate panel time series data that incorporates both time and individual heterogeneity. The framework allows joint modeling of continuous and categorical response variables. A nonlinear regression model is specified in which the covariates, the regression coefficients, and the error covariance matrices are smooth functions of time. In addition, the trajectories of the regression coefficients and error covariance matrices vary across individuals. Our aim is to (a) estimate the covariates and regression parameter trajectories; (b) cluster trajectories with similar shape; and (c) predict the time evolution of the categorical response variables from observations of the continuous covariates. To this end, the covariates and parameter trajectories are represented as linear combinations of some basis functions where the weights are drawn from an unknown distribution which has a dependent Dirichlet process prior. This nonparametric specification induces clustering across the trajectories and a nonparametric link function for the categorical response variables. We develop a Markov chain Monte Carlo algorithm for fitting the model and illustrate the methodology using simulated data and a panel data on online auction prices.

Definition, Calculation and Stability of Centrality Measures in Networks
Andrew C. Thomas, Carnegie Mellon University

"Centrality" is one of the most widely-used notions network analysis, attempting to measure the importance of a node. Yet despite being such a central idea to the study of networks, there is no standard definition;

rather, there are many incompatible definitions, e.g., based on degrees, shortest paths, and eigenvectors of the adjacency matrix. We examine this menagerie of definitions of centrality from the perspective of stability and estimability. That is, if the network is perturbed slightly or is measured with noise, how much does centrality change?

On a Class of Normalized Random Measures with Independent Increments
Lorenzo Trippa, Harvard School of Public Health

We define and investigate a new class nonparametric prior distributions. This class of priors is dense in the class of the homogeneous normalized random measures with independent increments and it is characterized by a predictive structure which is more elaborate than those arising from other priors widely used in the literature. A natural area of application is represented by species sampling problems and, in particular, prediction problems in genomics. We study distributional results related to the prior and posterior probability of discovering a certain number of new species when the observation process is driven by a model in the introduced class of random distributions. Finally, by using the coupling from the past method, we provide an exact sampling algorithm for the predictive distributions.
Joint with Stefano Favaro.

Hedge Fund Replication Using Minimax Filters: Report on a Work in Progress
Guillaume Weisang, Bentley University

Of practical importance both for the alternative investment industry and for regulation purposes, hedge fund replication offers to investors and regulators an insight into the sources of performance of hedge funds' portfolios. While Roncalli and Teiletche (2008) demonstrated that the Kalman filter provides a better approach than the classical rolling OLS-regression windows to capture the dynamic allocation in standard asset classes of hedge funds' investment profiles, hedge fund returns nonlinearities have yet to be successfully replicated. In this work in progress, I report on my work on a hedge fund replication strategy that is robust to the existence of nonlinearities in the hedge funds returns distributions, and hence robust to the use of non standard asset classes in hedge funds investment profiles. To do so, I use minimax filters. In this exposé, I first review the results obtained when using the Kalman filter. Second, I expose what I have done and what remains to be solved in this endeavour.

Designs for Bayesian Model Selection
Dave Woods, University of Southampton

In the early stages of experimentation, the aim is often to choose an appropriate linear model, potentially including interactions, for the dependence of a response on a set of factors. Motivated by an example from materials science, we describe a Bayesian approach to this problem using an expected loss for model selection which is a weighted sum of posterior model probabilities. The weights in this loss can be chosen according to the aim of the experiment. We introduce the Penalized Model Discrepancy (PMD) criterion for design selection, that minimizes an objective function based on the expected loss incurred by the model selected, averaged over the distribution of the data. The use of this criterion is explored through a variety of issues pertinent to screening experiments, including the choice of initial and follow-up designs and the robustness of design performance to prior information.
Designs from the PMD criteria are compared with those from existing approaches in the literature through illustrative examples. We also investigate reducing the computational burden of the method for experiments with a large number of contending models, through both the use of informative prior distributions and the approximation of the PMD objective function.
Joint work with Andy Rose (Lubrizol), Sue Lewis and Jon Forster (University of Southampton).

Multiple Imputation: A Negotiation of Two Parties
Xianchao Xie, Harvard University

Multiple imputation (Rubin, 1987), a principled method to conduct statistical inference in the presence of missing data, has been extensively studied and applied in various areas. Recent research in multiple imputation has seen significant attempts to better understand the theory underpinning the method and the

suitability of its use in circumstances other than those it was originally intended for. One recurrent criticism of multiple imputation has been that the confidence interval constructed by the procedure may not achieve the nominal coverage rate when there is model uncongeniality (Meng, 1994). In this paper we further examine this phenomena and study properties of multiple imputation as a general statistical methodology. Our major contributions are two-fold: first we show precisely how multiple imputation can be treated as an integration of the knowledge of two parties (the imputer and the user), and why the procedure may fail when model uncongeniality takes presence, i.e., the knowledge and assumptions of the two parties are "incompatible'". Second, we identify circumstances under which the procedure produces confidence intervals that will at least have the declared nominal rate, along with extensions that can be used when the original procedure is suspected to fail. We believe these theoretical results helps in deepening our understanding of previous findings as well as in developing new ones.
Co-author: Xiao-Li Meng

Jointly Primal and Dual Sparse Structured I/O Models
Eric Xing, Carnegie Mellon University

Sparsity has been a desirable property for a variety of predictive models. Methods that attain primal (sample) sparsity through optimizing hinge losses (i.e., SVM), or dual (feature) sparsity via regularization have been popular in the machine learning and statistical literature. When these methods are applied to train the so-called structured input/output models, which concern correlated multivariate output from dependent input features typically encountered in NLP and vision, they have led to the well-known CRF and structured SVM (a.k.a. M^3N), respectively, each enjoys some advantages, as well as weaknesses.
In this talk, I present a new general framework called Maximum Entropy Discrimination Markov Networks (MEDN), which integrates the above two approaches under a Bayesian framework, and thereby combines and extends their merits. I will discuss a number of theoretical properties of this model, and show applications of MEDN to learning fully supervised structured i/o model, max-margin structured i/o models with hidden variables, and a max-margin topic model for jointly discovering discriminative latent topic representations and predicting document label/score of text documents, with compelling performance in each case.

Nonparametric Rank-Based Tests of Bivariate Extreme-Value Dependence
Jun Yan, University of Connecticut

A new class of tests of extreme-value dependence for bivariate copulas is proposed. It is based on the process comparing the empirical copula with another natural nonparametric rank-based estimator of the unknown copula derived under extreme-value dependence. A multiplier technique is used to compute approximate p-values for several candidate test statistics. Extensive Monte Carlo experiments are carried out to compare the resulting procedures with the tests of extreme-value dependence recently studied in Ben Ghorbal et al. (2009) and Kojadinovic and Yan (2010). The finite-sample performance study of the tests is complemented by local power calculations.
Co-author: Ivan Kojadinovic

Diagnostic of Protein Phosphorylation Site Using Zero-inflated Poisson Regression
Shu Yang, Boston University

Differential expression analysis has been a basic and popular approach in the area of bioinformatics. Most works of detecting differential expression are done on microarray expression data. Recently, detection methods of phosphorylated residues are being actively developed. Comparing with microarray data, the data from protein phosphorylation sites provides more direct information on underlying differences in cellular processes, which helps to identify cancer apart from normal tissue and enables future treatment. However, this new data pose several statistical challenges: it contains of low-level counts and is frequently inflated with artificial zeros. Standard permutation tests under these conditions can be rather conservative. To better identify differential protein sites, we propose a framework based on zero-inflated Poisson regression. We show in both simulation and application to phosphorylation data from lung cancers that significantly more accurate identification of differentially expressed protein sites is possible.
Co-authors: Eric Kolaczyk, Simon Kasif and Martin Steffen

Gaussian-Based Routines for Imputing Categorical Variables in Complex Designs
Recai Yucel, State University of New York at Albany

Among many potential complexities, two common problems encountered in health surveys or administrative databases used to inform health policy are complexity of the design or data structure and missing data. This work modifies widely-used inferential tools to derive inferences via multiple imputation (MI) to be applied in such settings. The underlying methodology is widely-accepted and is based on computational techniques to sample imputations from a proposed imputation model with Gaussian errors. These models are flexible enough to take into account of complexities due to clustering. This work proposes rounding rules to be used with these existing MVN-based imputation methods, allowing practitioners to obtain usable imputation with small biases. These rules are calibrated in the sense that values re-imputed for observed data have distributions similar to those of the observed data. The methodology is demonstrated using a sample data from the NewYork Cancer Registry database.

Recent Progress in MAP Estimation for Computer Vision
Ramin Zabih, Cornell University

Over 25 years ago Geman and Geman popularized the formulation of computer vision problems as computing the MAP estimate of a Markov Random Field. The computational challenges involved proved quite daunting, but in the last decade rapid progress has been made. I will provide an overview of these methods, with a focus on combinatorial optimization techniques based on max flow.

Principled Sure Independence Screening for Cox Models with Ultra-High-Dimensional Covariates
Sihai Dave Zhao, Harvard School of Public Health and Dana Farber Cancer Institute

It is rather challenging for current variable selectors to handlesituations where the number of covariates under consideration is ultra-high. Consider a motivating study of clinical trials of bortezomib for the treatment of multiple myeloma, where overall survival and expression levels of 44760 probesets, encompassing more than 22000 genes, were measured for each of 188 patients with the goal of identifying genes that predict survival after treatment. This dataset defies analysis even with regularized regression. Some remedies have been proposed for the linear model and for generalized linear models, but there are few solutions in the survival setting and, to our knowledge, no theoretical support. Furthermore, existing strategies often involve tuning parameters that are difficult to interpret. In this paper we propose and theoretically justify a principled method for reducing dimensionality in the analysis of censored data by selecting only the important covariates. Our procedure involves a tuning parameter has a simple interpretation as the desired false positive rate. Simulation studies show that our method performs well even under model misspecification. We apply the proposed procedure to analyze the aforementioned myeloma study and identify biologically important and predictive genes.
Co-author: Yi Li

Statistical Methods for Studying Social Networks Using Aggregated Relational Data
Tian Zheng, Columbia University

Questions of the form "How many X's do you know?" collect aggregated relational data from one's social network and are a common means of learning about populations that are hard to reach or survey directly. McCarty et al. (2001), for example, take X to be individuals who are HIV positive, are homeless, or were recently raped to estimate the size of these traditionally hard-to-count populations. In this talk, we will discuss several recent statistical methodological developments for analyzing ARD to study features of social networks.

Optimal Estimation of Large Covariance Matrices
Harry Zhou, Yale University

With the emergence of high dimensional data from modern technologies, estimating large scale covariance matrices as well as their inverse is becoming a crucial problem in many fields. In this talk we give some

theories to unveil the precision to which (inverse) covariance matrices can be estimated and to develop general methodologies for optimal estimation of the (inverse) covariance matrices under various settings.

A Locally D-Optimal Design for Estimation of Parameters of an Exponential-Linear Growth Curve of Nanostructures
Li Zhu, Harvard University

We consider the problem of determining an optimal experimental design for estimation of parameters of a complex curve characterizing nanowire growth that is partially
exponential and partially linear. A locally D-optimal design for the non-linear change-point growth model is obtained by using a geometric approach. Further, a sequential algorithm is proposed for obtaining the D-optimal design. The convergence of the sequential algorithm to the D-optimal design is demonstrated using Monte-Carlo simulations. Some guidelines for the choice of initial design are also proposed.
Co-authors: Tirthankar Dasgupta, Qiang Huang

Rare-Allele Detection Using Compressed Se(que)nsing
Or Zuk, Broad Institute of MIT and Harvard

Detection of rare variants by resequencing is important for the identification of individuals carrying disease variants. Rapid sequencing by new technologies enables low-cost resequencing of target regions, although it is still prohibitive to test more than a few individuals. In order to improve cost trade-offs, it has recently been suggested to apply pooling designs which enable the detection of carriers of rare alleles in groups of individuals. However, this was shown to hold only for a relatively low number of individuals in a pool, and requires the design of pooling schemes for particular cases.
We propose a novel pooling design, based on a compressed sensing approach, which is both general, simple and efficient. We model the experimental procedure and show via computer simulations that it enables the recovery of rare allele carriers out of larger groups than were possible before, especially in situations where high coverage is obtained for each individual.
Our approach can also be combined with barcoding techniques to enhance performance and provide a feasible solution based on current resequencing costs. For example, when targeting a small enough genomic region (100 base-pairs) and using only 10 sequencing lanes and 10 distinct barcodes, one can recover the identity of 4 rare allele carriers out of a population of over 4000 individuals.
Co-authors: Noam Shental and Amnon Amir