16th Meeting of New Researchers in Statistics and Probability
Conference Program
Hosted by Harvard University

Thursday, July 31 to Saturday, August 2
Haller Hall, 24 Oxford Street
Cambridge, MA 02138, USA
16th Meeting of New Researchers in Statistics and Probability
Hosted by Harvard University
Haller Hall, 24 Oxford Street
Thursday, July 31 to Saturday, August 2

Schedule
All events are located in Haller Hall unless otherwise noted.

Thursday, July 31

8:00 am to 8:30 am  Continental breakfast, 103B
                    Conference check-in, Lobby

8:30 am to 9:30 am  Essential Concepts for Causal Inference in Randomized Experiments and
                    Observational Studies
                    *Donald B. Rubin (Harvard University)*

9:30 am to 9:45 am  Continental breakfast continued, 103B
                    Conference check-in continued, Lobby

9:45 am to 10:45 am Flash talks
                    Kriging Models with Gaussian Processes: Covariance Function Estimation and
                    Impact of Spatial Sampling by *François Bachoc (University of Vienna)*
                    Estimation of Latent Variable Densities for Exchangeable Network Models by
                    *Sharmodeep Bhattacharyya (University of California, Berkeley)*
                    Survey Fusion for Data that Exhibit Multivariate, Spatio-Temporal Dependencies
                    by *Jonathan R. Bradley (University of Missouri)*
                    Sparse and Orthogonal Factor Regression by *Kun Chen (University of Connecticut)*
                    Log-Concavity and Nonparametric Inference about a Density’s Mode by *Charles
                    Doss (University of Washington)*
                    Genome-Wide Power and Sample Size Calculation for Sequencing Based
                    Experiment by *Yong Seok Park (University of Pittsburgh)*

10:45 am to 11:00 am Coffee break, 103B

11:00 am to 12:00 noon Flash talks
                    Calibrating Sensitivity Analyses to Observed Covariates in Observational Studies
                    by *Jesse Yen-Chih Hsu*
                    Effective Planning in Serial Dilution Design by *Yu-Ting Hsu (Penn State
                    Harrisburg)*
                    Modeling Short- and Long-Term Characteristics of Follicle Stimulating Hormone as
                    Predictors of Severe Hot Flashes in Penn Ovarian Aging Study by *Bei Jiang
                    (University of Michigan)*
                    Restricted Linear Covariance Models for Multivariate Longitudinal Data by *Priya
                    Kohli (Connecticut College)*
                    Variable Screening Based on Combining Quantile Regression by *Linglong Kong
                    (University of Alberta)*
                    Joint Sparse Linear Modeling of Replicate RNA-seq Data for Isoform Discovery and
                    Abundance Estimation (jSLIDE) by *Jingyi Jessica Li (University of California, Los
                    Angeles)*
                    Generalised Particle Filters with Gaussian Mixtures by *Kai Li (Uppsala University)*
                    Integrative Analysis of Prognosis Data on Multiple Cancer Subtypes by *Jin Liu*
16th Meeting of New Researchers in Statistics and Probability
Hosted by Harvard University
Haller Hall, 24 Oxford Street
Thursday, July 31 to Saturday, August 2

*University of Illinois at Chicago*

12:00 noon to 2:00 pm  Buffet lunch, 103B
Seating available in 103B, 105, 1st floor, 4th floor, and outside

2:00 pm to 3:00 pm  Tweedie New Researcher Award Talk
Hidden Markov Models with Applications in Cell Adhesion Experiments
Ying Hung *(Rutgers University)*

3:00 pm to 3:15 pm  Coffee and snack break, 103B

3:15 pm to 4:15 pm  Flash talks
Semiparametric Deconvolution Density Estimation *by Cornelis Potgieter (Southern Methodist University)*
Valid Inference from Non-Ignorable Network Sampling Designs *by Simon Lunagomez (Harvard University)*
Minimax Solution for the Two-Stage Group Testing Problem *by Yaakov Malinovsky (University of Maryland, Baltimore County)*
On Dimensionality Reduction Techniques for Multivariate Time Series: Application to Energy Markets *by Carolina Garcia-Martos (Technical University of Madrid)*
Logistic Generalized Survey Regression Estimators with the Lasso *by Kelly McConville (Swarthmore College)*
Weighted Estimating Equations for Semiparametric Transformation Models with Missing Covariates *by Yang Ning (University of Waterloo)*
Causal Inference with Social Network Data: Inflated Effective Sample Sizes, Deflated Standard Errors, and Other Perils *by Elizabeth L. Ogburn (Johns Hopkins University)*
Efficient Inference of Population Size Trajectories from the Sequential Markovian Coalescent *by Julia A. Palacios (Harvard University and Brown University)*
Estimating State Changes for Epileptic Seizure Data *by Ivor Cribben (University of Alberta)*

4:15 pm to 5:00 pm  Coffee and snack break continued, 103B

5:00 pm to 7:00 pm  Poster session and reception, 103A, 103B, 105, and 1st Floor
*Thursday flash talk presenters*
Friday, August 1

8:00 am to 8:30 am  Continental breakfast, 103B

8:30 am to 9:30 am  Cross-study Reproducibility of Predictions, with Application to Genomics
Giovanni Parmigiani (Dana Farber Cancer Institute)

9:30 am to 9:45 am  Continental breakfast continued, 103B

9:45 am to 10:45 am  Flash talks
Robust Heritability Estimation in Plant Candidate Gene Association Studies by
Vanda Lourenço (New University of Lisbon)
Finite Group Invariance by Alexander Volfovsky (Harvard University)
Threshold Regression with Censored Covariates by Jing Qian (University of
Massachusetts, Amherst)
Bayesian Marked Point Process Modeling for Generating Fully Synthetic Public Use
Data with Point-Referenced Geography by Harrison Quick (University of Missouri)
The Structural Topic Model and Applied Social Science by Molly Roberts
(University of California, San Diego)
A Multivariate Two-Sample Test Using Regular Minimum-Weight Spanning
Subgraphs by David Ruth (United States Naval Academy)
Interference in Deformation Compensation for 3D Printing by Arman Sabbaghi
(Purdue University)
Spectral Embedding of Stochastic Blockmodel Graphs by Daniel Sussman (Harvard
University)
Multi-Agent Inference in Social Networks: A Finite Population Learning Approach
by Xin Tong (University of Southern California)

10:45 am to 11:00 am  Coffee and snack break, 103B

11:00 am to 12:00 noon  Flash talks
Learning with Low Rank Matrices: Flexible Modeling and Scalable Computation by
Rahul Mazumder (Columbia University)
Multiply Robust Estimation in Regression Analysis with Missing Data by Peisong
Han (University of Waterloo)
A Covariance Matrix-Valued State-Space Model by Jesse Windle (Duke University)
A Bayesian Spatio-Temporal Geostatistical Model with an Auxiliary Lattice for
Large Datasets by Ganggang Xu (Texas A&M University)
Robust Nonnegative Matrix Factorization: Modern Dimension Reduction Procedure
for Big Noisy Data Set by Yifan Ethan Xu (Case Western Reserve University)
Regularized Learning of High-Dimensional Graphical Models by Lingzhou Xue
(Pennsylvania State University)
Principal Flow on Manifolds by Zhigang Yao (Swiss Federal Institute of
Technology, Lausanne)
FMEM: Functional Mixed Effects Modeling with Applications to the Analysis of
Longitudinal White Matter Tract Data by Ying Yuan (St. Jude Children’s Research
Hospital)
16th Meeting of New Researchers in Statistics and Probability
Hosted by Harvard University
Haller Hall, 24 Oxford Street
Thursday, July 31 to Saturday, August 2

12:00 noon to 2:00 pm
Buffet lunch, 103B
Seating available in 103B, 105, 1st floor, 4th floor, and outside

2:00 pm to 3:30 pm
Panel on Funding
Michelle Dunn (National Cancer Institute), Predrag Neskovic (Office of Naval Research), Sujit Ghosh (National Science Foundation)

3:30 pm to 3:45 pm
Coffee and snack break, 103B

3:45 pm to 4:45 pm
Flash talks
Partially Missing at Random and Ignorable Inferences for Parameter Subsets with Missing Data by Sahar Zangeneh
EM Algorithm for Generalized Odds-Rate Hazard Models with Interval Censored Data by Bin Zhang (Cincinnati Children’s Hospital Medical Center)
Sampling for Conditional Inference on Network Data by Emma Jingfei Zhang (University of Miami)
Improving the Convergence Rate in Linear Mixed Models with Multiple Random-Effects by Guangxiang Zhang (University of Hawaii at Manoa)
Gene-Disease Associations via Sparse Simultaneous Signal Detection by Sihai Dave Zhou (University of Illinois at Urbana-Champaign)
Competing Risks Model Selection with Application to Kidney Transplant Data by Bingqing Zhou (Yale University)
Improvement of Cancer Prevention in Lynch Syndrome by Wei Vivian Zhuang (Creighton University)
Efficient Support Recovery via Weighted Maximum-Contrast Subagging by Jelena Bradic (University of California, San Diego)

4:45 pm to 5:30 pm
Coffee and snack break continued, 103B

5:30 pm to 7:30 pm
Poster session and reception, 103A, 103B, 105, and 1st Floor
Friday flash talk presenters
Saturday, August 2

8:00 am to 8:30 am  Continental breakfast, 103B

8:30 am to 9:00 am  Let Us Own Data Science  
*Bin Yu (University of California, Berkeley)*

9:00 am to 9:15 am  Continental breakfast continued, 103B

9:15 am to 10:45 am  Panel on Leadership  
*Alan Karr (National Institute of Statistical Sciences), Xiao-Li Meng (Harvard University), Susan Murphy (University of Michigan, Ann Arbor), Giovanni Parmigiani (Dana Farber Cancer Center)*

10:45 am to 11:00 am  Coffee and snack break, 103B

11:00 am to 12:30 pm  Panel on Publishing  
*Stephen E. Fienberg (Carnegie Mellon University), Jun S. Liu (Harvard University), Runze Li (Pennsylvania State University)*

12:30 pm to 2:00 pm  Buffet lunch, 103B  
Seating available in 103A, 103B, 105, 1st floor, 4th floor, and outside

2:00 pm to 3:00 pm  Micro-Randomized Trials & mHealth  
*Susan A. Murphy (University of Michigan, Ann Arbor)*

3:00 pm to 3:15 pm  Coffee and snack break, 103B

3:15 pm to 4:45 pm  Panel on Data Science  
*Michael I. Jordan (University of California, Berkeley), Sham Kakade (Microsoft), Steven Scott (Google), Bin Yu (University of California, Berkeley)*

4:45 pm to 5:00 pm  Coffee and snack break continued, 103B

5:00 pm to 6:00 pm  On the Statistical-Computational Interface  
*Michael I. Jordan (University of California, Berkeley)*

6:00 pm  Departure
Speaker Abstracts

Essential Concepts for Causal Inference in Randomized Experiments and Observational Studies
Donald B. Rubin, PhD
John L. Loeb Professor of Statistics
Harvard University

There are several essential concepts for causal inference in randomized experiments and observational studies. These concepts were formulated only recently, in the twentieth century, and are important to keep in mind when trying to understand the causal effects of past interventions or new proposed interventions. Some historical connections will be emphasized, and the reasons for the inapposite focus on regression based methods for causal inference will be discussed.

Hidden Markov Models with Applications in Cell Adhesion Experiments
Ying Hung, PhD
Assistant Professor of Statistics
Rutgers University

Cell adhesion experiments refer to biomechanical experiments that study protein, DNA, and RNA at the level of single molecules. The study of cell adhesion plays a key role in many physiological and pathological processes, especially in tumor metastasis in cancer research. Motivated by the analysis of a specific type of cell adhesion experiments, a new framework based on hidden Markov model is proposed. A double penalized order selection procedure is introduced and shown to be consistent in estimating the number of hidden states in hidden Markov models. Simulations show that the proposed framework outperforms existing methods. Applications of the proposed methodology to real data demonstrate the accuracy of estimating receptor-ligand bond lifetimes and waiting times, which are essential in kinetic parameter estimation.

Cross-study Reproducibility of Predictions, with Application to Genomics
Giovanni Parmigiani, PhD
Chair of the Department of Biostatistics and Computational Biology
Dana Farber Cancer Institute

Numerous gene signatures of patient prognosis for late-stage, high-grade ovarian cancer have been published, but diverse data and methods have made these difficult to compare objectively. However, the corresponding large volume of publicly available expression data creates an opportunity to validate previous findings and to develop more robust signatures. We thus built a database of uniformly processed and curated public ovarian cancer microarray data and clinical annotations, and re-implemented and validated 14 prognostic signatures published between 2007 and 2012. In this lecture I will describe the methodology and tools we developed for evaluating published signatures in this context. I will also use this application as the springboard for a more general discussion on how to evaluate statistical learning methods based on a collection of related studies.

Let Us Own Data Science
Bin Yu, PhD
Chancellor's Professor
Departments of Statistics and Electrical Engineering and Computer Science
University of California, Berkeley
IMS Past President
This talk is a repeat of my IMS presidential address at the IMS joint meeting with Statistical Society of Australia (SSA) in July, 2014. In the talk, I will start with a historical review of people in our profession who have been practicing “data science.” We will start with the founding of SSA (1947) and IMS (1930, Ann. Math. Stats) by Helen Turner and Harry Carver and move back to Hollerith. More recent “data scientists” like Cochran, Tukey, Wu, and Breiman will also be discussed. Furthermore, an argument is made that name does matter and we will end with a list of suggestions for action for a substantial engagement of statistics with data science.

Micro-Randomized Trials & mHealth
Susan A. Murphy, PhD
H.E. Robbins Professor of Statistics
University of Michigan, Ann Arbor
Micro-randomized trials are trials in which individuals are randomized 100’s or 1000’s of times over the course of the study. The goal of these trials is to assess the impact of momentary interventions, e.g. interventions that are intended to impact behavior over small time intervals. A fast growing area of mHealth concerns the use of mobile devices for both collecting real-time data, for processing this data and for providing momentary interventions. We discuss the design and analysis of these types of trials.

On the Statistical-Computational Interface
Michael I. Jordan, PhD
Pehong Chen Distinguished Professor
University of California, Berkeley
The rapid growth in the size and scope of datasets in science and technology has created a need for novel foundational perspectives on data analysis that blend the statistical and computational sciences. A particularly critical need is that of obtaining guarantees on statistical risk that incorporate notions of computational complexity, and in particular of runtime. Taking statistical decision theory as a point of departure, we treat “externalities,” such as computation, communication, and privacy, as constraints on estimation procedures, and explore the tradeoffs that result. [Joint work with Venkat Chandrasekaran, John Duchi, Martin Wainwright, and Yuchen Zhang.]
16th Meeting of New Researchers in Statistics and Probability
Hosted by Harvard University
Haller Hall, 24 Oxford Street
Thursday, July 31 to Saturday, August 2

Map

Conference Center: Haller Hall, 24 Oxford Street
Dormitory Housing: Perkins Hall
Dormitory Check-In/Check-Out and Statistics Department: Science Center, 7th Floor Suite
Notes