RUTGERS
School of Arts and Sciences

Department of Statistics

2019 Rutgers Statistics Symposium:
Tomorrow's Statistics for Today's Data

May 2–3, 2019

Proteomics Building, Room 120, Busch Campus

Rutgers, The State University of New Jersey

Program

Thursday, May 2

8:30am – 9:00am: Breakfast

9:00am – 9:20am: Opening Remarks

Ronald Ransome	**MPS Dean, School of Arts and Sciences, Rutgers University**
Stephen K. Burley	**University Professor and Henry Rutgers Chair; Founding Director, Institute for Quantitative Biomedicine, Rutgers University**
Regina Liu	**Chair, Department of Statistics, Rutgers University**

9:20am – 9:40am:

Speaker: **Stephen K. Burley, M.D., D.Phil. Director, RCSB Protein Data Bank**

"RCSB Protein Data Bank: Sustaining a Living Digital Data Resource"

(9:40am – 10:00am: Break)

Session I: **Chair – William Strawderman, Rutgers University**

10:00am – 11:00am:

Speaker: **Michael Stein, The University of Chicago & Rutgers University**

"*Statistical Extremes: Theory and Practice*"

Discussant: **Ying Hung, Rutgers University**

11:00am – 12:00pm:

Speaker: **Marc Suchard, University of California, Los Angeles**

"Large-scale Evidence Generation across a Network of Databases (LEGEND) for Hypertension: Real-world, Reliable and Reproducible"

Discussant: **Harry Crane, Rutgers University**

(Noon – 1:00pm: Lunch Break)

Session II: **Chair – David Tyler, Rutgers University**

1:00pm – 2:00pm:

Speaker: **Howard Bondell, The University of Melbourne, Australia**

"Tales of Multiple Regression: Informative Missingness, Recommender Systems, and R2-D2"

Discussant: **Robin Gong, Rutgers University**

2:00pm – 3:00pm:

Speaker: **Arnak Dalalyan, ENSAE-CREST (Center for Research in Economics and Statistics, Paris), France**

"Robust Sparse Regression by Convex Programming"

Discussant: **Zijian Guo, Rutgers University**

(3:00pm – 3:20pm: Break)

Session III: **Chair – Rong Chen, Rutgers University**

3:20pm – 4:20pm:

Speaker: **Axel Munk, University of Göttingen, Germany**

"Optimal Transport based Data Analysis: Inference, Algorithms, Applications"

Discussant: **Roy Han, Rutgers University**

4:20pm – 5:20pm:

Speaker: **Richard Davis, Columbia University**

"The Use of Shape Constraints for Modeling Time Series of Counts"

Discussant: **Han Xiao, Rutgers University**

5:20pm – 6:00pm: ***Wine and Cheese Reception***

Friday, May 3

8:30am – 9:00am: Breakfast

Session IV: **Chair – Cun-Hui Zhang, Rutgers University**

9:00am – 10:00am:

Speaker: **Jianqing Fan, Princeton University**

"Communication and Statistical Efficient Distributed Estimation"

Discussant: **Pierre Bellec, Rutgers University**

(10:00am –10:20am: Break)

Session V: **Chair – Minge Xie, Rutgers University**

10:20am – 11:20am:

Speaker: **Dylan Small, The University of Pennsylvania**

"Discovering Effect Modification in Observational Studies"

Discussant: **Tirthankar Dasgupta, Rutgers University**

11:20am – 12:20pm:

Speaker: **Jennifer Hill, New York University**

"Automated versus Do-it-yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition"

Discussant: **Steve Buyske, Rutgers University**

(12:20pm – 1:00pm: Lunch Break)

Session VI: **Chair – John Kolassa, Rutgers University**

1:00pm – 2:00pm:

Speaker: **David Siegmund, Stanford University**

"Detection and Estimation of Local Signals"

Discussant: **Min Xu, Rutgers University**

2:00pm – 3:00pm:

Speaker: **Samuel Kou, Harvard University**

"Big Data, Google and Disease Detection: A Statistical Adventure"

Discussant: **Sijian Wang, Rutgers University**

(3:00pm – 3:20pm: Break)

Session VII: **Chair – Harold Sackrowitz, Rutgers University**

3:20pm – 4:20pm:

Speaker: **John Lafferty, Yale University**

"Statistical Estimation and Machine Learning with Communication and Shape Constraints"

Discussant: **Jason Klusowski, Rutgers University**

Organizing Committee:

Pierre Bellec (co-chair), Steve Buyske, Harry Crane, Tirthankar Dasgupta, Robin Gong, Zijian Guo, Roy Han, Ying Hung, Jason Klusowski, Sijian Wang, Han Xiao, Min Xu, Cun-Hui Zhang (co-chair)

Conference Coordinator:

Marcy Collins: mcollins@stat.rutgers.edu

For more information and registration: ***http://www.stat.rutgers.edu/***

RUTGERS
School of Arts and Sciences

Department of Statistics

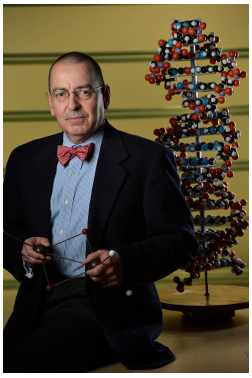
2019 Rutgers Statistics Symposium:

Tomorrow's Statistics for Today's Data

Abstracts

Stephen K. Burley

Director, RCSB Protein Data Bank, Rutgers University



RCSB Protein Data Bank: Sustaining a Living Digital Data Resource

Thursday, May 2, 2019
9:20am – 9:40am

Bio: Stephen Kevin Burley is an expert in structural biology, biophysics, computational biology, data science, structure/fragment-based drug discovery, and clinical medicine/oncology. Burley is the Director of the RCSB Protein Data Bank (rcsb.org). Within Rutgers, The State University of New Jersey he serves as University Professor and Henry Rutgers Chair, Founding Director of the Institute for Quantitative Biomedicine, and a Member of the Rutgers Cancer Institute of New Jersey, wherein he Co-Leads the Cancer Pharmacology Research Program.

Michael Stein

Ralph and Mary Otis Isham Professor, University of Chicago
Distinguished Visiting Professor, Rutgers University

Statistical Extremes: Theory and Practice

Thursday, May 2, 2019
10:00am – 11:00am

Abstract: Extreme outcomes are of great interest in a number of disciplines, including climate science and finance. This talk will discuss some of the basic mathematical ideas underlying the statistical theory of extremes and how it used to study extreme quantiles of distributions. I will then describe recent work of mine on flexible parametric models of distributions that can capture a wide range of behaviors for extreme quantiles. I apply these models to daily temperature data from an ensemble of 50 model runs from a global climate model. The large ensemble provides much more data than one could obtain from observations, thus making it possible to compare different methods for estimating extreme quantiles. At the locations I consider, the approach based on flexible parametric models clearly outperforms traditional extreme value approaches. Professor Stein is also distinguished visiting professor of Department of Statistics and Institute of Earth, Ocean & Atmospheric Sciences of Rutgers University.

Bio: Michael Stein is the Ralph and Mary Otis Isham Professor of Statistics and the College at the University of Chicago. Professor Stein is also distinguished visiting professor of Department of Statistics and Institute of Earth, Ocean & Atmospheric Sciences of Rutgers University, 2018 -- . Most of Professor Stein's research has been in the area of spatial statistics and its applications to environmental sciences and astrophysics. His current research focuses on statistical models and methods for natural processes in space-time and the computational challenges that arise in studying such problems.

Marc Suchard

Professor, University of California, Los Angeles

Large-scale Evidence Generation across a Network of Databases (LEGEND) for Hypertension: Real-world, reliable and reproducible

Thursday, May 2, 2019
11:00am – 12:00pm

Abstract. Concerns over reproducibility in science extend to research using existing healthcare data; many observational studies investigating the same topic produce conflicting results, even when using the same data. To address this problem, we propose a paradigm shift. The current paradigm centers on generating one estimate at a time using a unique study design with unknown reliability and publishing (or not) one estimate at a time. The new paradigm advocates for high-throughput observational studies using consistent and standardized methods, allowing evaluation, calibration, and unbiased dissemination to generate a more reliable and complete evidence base. We demonstrate this new paradigm by comparing all hypertension treatments for a set of effectiveness and safety outcomes, producing 587,020 hazard ratios, each using methodology on par with state-of-the-art studies. We furthermore include control hypotheses to evaluate and calibrate our evidence generation process. Results agree with the limited number of randomized trials. The distribution of effect size estimates reported in literature reveals an absence of small or null effects, with a sharp cutoff at p = 0.05. No such phenomena were observed in our results, suggesting more complete and more reliable evidence.

Bio: Mark Suchard is Professor in the Departments of Biomathematics and of Human Genetics in the David Geffen School of Medicine and in the Department of Biostatistics in the UCLA Fielding School of Public Health. He received Guggenheim Fellowship in 2008 and COPSS Presidents' Award in 2013, among other distinctions. His research interests focus on biomathematical and biostatistical approaches to computational statistics, analyzing stochastic processes in molecular sequence data, longitudinal modeling of biomedical processes, and examining the patient-physician relationship.

Howard Bondell

Professor, University of Melbourne, Australia

Tales of Multiple Regression: Informative Missingness, Recommender Systems, and R2-D2

Thursday, May 2, 2019
1:00pm – 2:00pm

Abstract: In this talk, we briefly discuss two projects tangentially related under the umbrella of high-dimensional regression. The first part of the talk investigates informative missingness in the framework of recommender systems. In this setting, we envision a potential rating for every object-user pair. The goal is to predict the unobserved ratings to then recommend an object that the user is likely to rate highly. A typically overlooked piece is that the data is not missing at random. For example, in movie ratings, a relationship between the user ratings and their viewing history is expected, as human nature

dictates the user would seek out movies that they anticipate enjoying. We model this informative missingness, and place the recommender system in a shared-variable framework which can aid in prediction quality. The second part of the talk deals with a new class of prior distributions for shrinkage regularization in sparse linear regression, particularly the high dimensional case. Instead of placing a prior on the coefficients themselves, we place a prior on the regression R-squared. This is then distributed to the coefficients by decomposing it via a Dirichlet Distribution. We call the new prior R2-D2 in light of its R-Squared Dirichlet Decomposition. Compared to existing shrinkage priors, we show that the R2-D2 prior can simultaneously achieve both high prior concentration at zero, as well as heavier tails. These two properties combine to provide a higher degree of shrinkage on the irrelevant coefficients, along with less bias in estimation of the larger signals.

Bio: Howard Bondell received his Ph.D. in Statistics from Rutgers University in 2005, joined the faculty of the Department of Statistics at North Carolina State University, and moved to the School of Mathematics and Statistics at the University of Melbourne in 2018. He was elected Fellow of the American Statistical Association. His current research interests include: Variable and Model Selection, Robust Estimation, Quantile Regression, Nonparametric Smoothing and Regression, Regularization and Bayesian Methods.

Arnak Dalalyan

Professor, ENSAE-CREST, France

Robust Sparse Regression by Convex Programming

Thursday, May 2, 2019
2:00pm – 3:00pm

Abstract: Motivated by the construction of tractable robust estimators via convex relaxations, we present conditions on the sample size which guarantee an augmented restricted eigenvalue condition for Gaussian designs. Such a condition is suitable for high-dimensional inference in a linear model and in a multivariate Gaussian model when samples are corrupted by outliers (either in the response variable or in the design matrix). Our analysis leads to significantly sharper restricted eigenvalue constant, valid under weaker assumptions than those available in the literature. In particular, we require no condition relating the sparsity of the unknown parameter and the number of outliers. Elaborating on these results, we establish new risk bounds for the l1 penalized Huber M-estimator for robust linear regression. These bounds show that the rate of estimation of s-sparse p-dimensional parameter vectors is of order (s/n) + (o/n)^2, up to logarithmic factors, where o is the number of outliers and n is the sample size. (Joint work with Philip Thompson)

Bio: Arnak Dalalyan is a full professor of Statistics at ENSAE-CREST. He is an associate editor of Electronic Journal of Statistics, Statistical Inference for Stochastic Processes, Journal of the Japan Statistical Society. Arnak is a member on the program commitees of the machine learning conferences COLT and NIPS. He is also a member of the Bernoulli Society council. His research focuses on stochastic optimization, Langevin dynamics, high dimensional statistics and robustness.

Axel Munk

Professor, University of Göttingen, Germany

Optimal Transport based Data Analysis: Inference, algorithms, applications

Thursday, May 2, 2019
3:20pm – 4:20pm

Abstract: Recent developments in statistical data analysis based on (regularized) empirical optimal transport (EOT) will be discussed. Fundamental to our approach for optimal transport based data analysis (OTDA) is that we restrict to finite and discrete spaces. We provide risk bounds and limit laws for EOT plans and distances. These are notably different compared to the continuous world. We argue that for most real world applications restricting to discrete measurements might be even more appropriate as it takes into account the degree of discretization. The distributional limits of EOTs are characterized by dual optimal transport problems over a Gaussian process. Proofs for this are based on a combination of sensitivity analysis from convex optimization and discrete empirical process theory. We examine an upper bound for such limiting distributions based on a spanning tree approximation which can be computed explicitly. This can be used for various tasks of statistical inference and also fast simulation. Bases on this we discuss error bounds for the EOT. This allows to balance computational and statistical error for simple resampling schemes to compute optimal transport in large scale data applications. A major finding is that up to spatial dimension 3 resampling is scalable with the size of the image, for larger dimensions this is not the case anymore. Our methodology is illustrated in computer experiments and on biological images from super-resolution cell microscopy. This is joint work with Marcel Klatt, Max Sommerfeld, Carla Tameling an Yoav Zemel.

Bio: Axel Munk is the Felix-Bernstein Chair for Mathematical Statistics. His is also a Max-Planck fellow at the Max-Planck institute for biophysical chemistry since 2010. Professor Munk received the German Section award of the Biometric Society in 1992. He is an elected fellow of the Institute of Mathematical Statistics, the Bernoulli Society, the Göttingen Academy of Sciences and Humanities and The International Statistical Institute (ISI). He serves or has served as Associate Editor of various statistical journals, including the Annals of Statistics, Bernoulli and the Journal of the Royal Statistical Society, Series B. His research focus on nonparametrics, optimal transport and their applications in biometrics and engineering.

Richard Davis

Professor, Columbia University

The Use of Shape Constraints for Modeling Time Series of Counts

Thursday, May 2, 2019
4:20pm – 5:20pm

Abstract: For many formulations of models for time series of counts, the specification of a family of probability mass functions relating the observation Yt at time t to a state variable Xt must be explicitly specified. Typical choices are the Poisson and negative binomial distributions. One of the principal goals of this research is to relax this parametric framework and assume that the requisite pmf is a one-parameter exponential family in which the reference distribution is unknown but log-concave. This class of distributions includes many of the commonly used pmfs. The serial dependence in the model is governed by specifying the evolution of the conditional mean process. The particular link function used in the exponential family model depends on the specification of the reference distribution. Using this semi-parametric model formulation, we are able to extend the class of observation-driven models studied in Davis and Liu (2016). In particular, we show there exists a stationary and ergodic solution to the state-space model. In this new semi-parametric framework, we compute and maximize the likelihood function over both the parameters associated with the mean function and the reference measure subject to a concavity constraint. On top of this we can "smooth" the pmf using the Skellam distribution in order to obtain an estimated distribution defined on all the non-negative integers. In general, the smooth version has better performance than existing methods. The estimator of the mean function and the conditional distribution are shown to be consistent and perform well compared to a full parametric model specification. Further limit theory in other situations will be

described. The finite sample behavior of the estimators are studied via simulation and empirical examples are provided to illustrate the methodology. This is joint work with Jing Zhang and Thibault Vatter.

Bio: Richard Davis is Chair and Howard Levine Professor of Statistics at Columbia University. He has held academic positions at MIT, Colorado State University, and visiting appointments at numerous other universities. Recently he was Hans Fischer Senior Fellow at the Technical University of Munich and Villum Kan Rasmussen Visiting Professor at the University of Copenhagen. Davis is a fellow of the Institute of Mathematical Statistics and the American Statistical Association, and is an elected member of the International Statistical Institute. His research interests lie primarily in the areas of applied probability, time series, and stochastic processes. While Professor Davis's research interests have gravitated towards problems in time series analysis, extreme value theory still has a strong influence in his approach to solving problems.

Jianqing Fan

Professor, Princeton University

Communication and Statistical Efficient Distributed Estimation

Friday, May 3, 2019
9:00am – 10:00am

Abstract: When the data are stored in a distributed manner, direct application of traditional statistical inference procedures is often prohibitive due to communication cost and privacy concerns. This paper develops and investigates iterative algorithms for distributed statistical estimation. In each iteration, node machines carry out computation in parallel and communicates with the central processor, which then broadcasts aggregated gradient vector to node machines for new updates. The algorithms can adapt to the similarity among loss functions on node machines, and converge rapidly when each node machine has large enough sample size. Moreover, they do not require good ini-tialization and enjoy linear converge guarantees under general conditions. In addition, the improved statistical accuracy per iteration is derived. Extensive numerical experiments on both synthetic and real data validate the theoretical results and demonstrate the superior performance of our algorithms.

Bio: Jianqing Fan is Frederick L. Moore '18 Professor of Finance, Professor of Statistics, and Professor of Operations Research and Financial Engineering at the Princeton University. He received The 2000 COPSS Presidents' Award, Morningside Gold Medal for Applied Mathematics (2007), Guggenheim Fellow (2009), Pao-Lu Hsu Prize (2013) and Guy Medal in Silver (2014). He is elected member of Academia Sinica in 2012. His research interests include: statistical theory and methods in data science, statistical machine learning, finance, economics, computational biology, biostatistics with particular skills on high-dimensional statistics, nonparametric modeling, longitudinal and functional data analysis, nonlinear time series, wavelets, among others.

Dylan Small

Professor, University of Pennsylvania

Discovering effect modification in observational studies

Friday, May 3, 2019
10:20am – 11:20am

Abstract: There is effect modification if the magnitude of a treatment effect varies with the level of an observed covariate. A larger treatment effect is typically less sensitive to bias from unmeasured covariates, so it is important to recognize effect modification when it is present. Additionally, effect modification is of interest for personalizing treatments based on an individual's covariates. We present a method for conducting a sensitivity analysis in an observational study that empirically discovers effect modification by exploratory methods, but controls the family-wise error rate or false discovery rate in discovered groups. We will discuss an application of the method to an observational study of the effect of superior nursing at a hospital on surgical mortality.

Bio: Dylan Small is the Class of 1965 Wharton Professor in the Department of Statistics at The Wharton School, University of Pennsylvania. He is an elected fellow of the American Statistical Association. His research interests include causal inference, design and analysis of experiments and observational studies for comparing treatments, longitudinal data, measurement error and applications of statistics to medicine and health. Professor Small is the founding editor of the journal Observational Studies.

Jennifer Hill

Professor, New York University

Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition

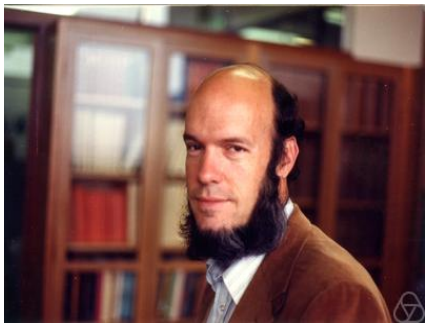
Friday, May 3, 2019
11:20am – 12:20pm

Abstract: Statisticians have made great progress in creating methods that reduce our reliance on parametric assumptions. However this explosion in research has resulted in a breadth of inferential strategies that both create opportunities for more reliable inference as well as complicate the choices that an applied researcher has to make and defend. Relatedly, researchers advocating for new methods typically compare their method to at best 2 or 3 other causal inference strategies and test using simulations that may or may not be designed to equally tease out flaws in all the competing methods. The causal inference data analysis challenge launched as part of the 2016 Atlantic Causal Inference Conference, sought to make progress with respect to both of these issues. The researchers creating the data testing grounds were distinct from the researchers submitting methods whose efficacy would be evaluated. Results from 30 competitors across the two versions of the competition (black box algorithms and do-it-yourself analyses) are presented along with post-hoc analyses that reveal information about the characteristics of causal inference strategies and settings that affect performance. The most consistent conclusion was that methods that flexibly model the response surface perform better overall than methods that fail to do so. Finally new methods are proposed that combine features of several of the top-performing submitted methods.

Bio: Jennifer Hill is Professor of Applied Statistics and Data Science at NYU. She is also the Co-Director of the Center for Practice and Research at the Intersection of Information, Society, and Methodology (PRIISM) and Co-Director of and the Master's of Science Program in Applied Statistics for Social Science Research (A3SR). Professor Hill develops and evaluates methods that help us answer the causal questions that are vital to policy research and scientific development. In particular she focuses on situations in which it is difficult or impossible to perform traditional randomized experiments, or when even seemingly pristine study designs are complicated by missing data or hierarchically structured data. Most recently Professor Hill has been pursuing two intersecting strands of research. The first focuses on Bayesian nonparametric methods that allow for flexible estimation of causal models and are less time-consuming and more precise than competing methods (e.g. propensity score approaches). The second line of work pursues strategies for exploring the impact of violations of typical causal inference assumptions such as ignorability (all confounders measured) and common support (overlap).

David Siegmund

Professor, Stanford University



Detection and Estimation of Local Signals

Friday, May 3, 2019
1:00pm – 2:00pm

Abstract: We study the maximum score statistic to detect and estimate local signals in the form of change-points in the level, slope, or other local property of a sequence of observations, and to segment the sequence when there appear to be multiple signals. We find that change-points in level or slope can lead to upwardly biased estimates of autocorrelations that result in a loss of power. Applications to temperature variations, atmospheric CO2 levels, disease incidence, and fluctuations in the size of animal populations illustrate the general theory. This is joint research with Xiao Fang.

Bio: David Siegmund is the John D. and Sigrid Banks Chair Professor at Stanford University. Professor Siegmund is a statistician who is comfortable in both the airy heights of theory and the practicalities of real-world applications. He works at the interface between probability and statistics, applying the tools he develops to topics as diverse as the design of medical clinical trials and mapping the locations of genes that are involved in specific physiological traits. He has received several awards and distinctions, including a Guggenheim Fellowship in 1974, the Humboldt Prize in 1980, and membership in the American Academy of Arts and Sciences in 1994. In 2002 he was elected to the National Academy of Sciences.

Sam Kou

Professor, Harvard University

Big data, Google and disease detection: a statistical adventure

Friday, May 3, 2019
2:00pm – 3:00pm

Abstract: Big data collected from the internet have generated significant interest in not only the academic community but also industry and government agencies. They bring great potential in tracking and predicting massive social activities. We focus on tracking disease epidemics in this talk. We will discuss the applications, in particular, Google Flu Trends, some of the fallacy and the statistical implications. We will propose a new model that utilizes publicly available online data to estimate disease epidemics. Our model outperforms all previous real-time tracking models for influenza epidemics at the national level of the US. An extended version of the model gives accurate tracking of Dengue fever in Asian and South American countries. We will also draw some lessons for big data applications.

Bio: Samuel Kou is Professor of Statistics and Professor of Biostatistics at Harvard University. He received COPSS Presidents' Award in 2012, the Guggenheim Fellowship in 2013, among other distinctions. He is an elected Fellow of the American Statistical Association, an elected member of the International Statistical Institute, and an elected Fellow and a Medallion Lecturer of the Institute of Mathematical Statistics. His research interests include stochastic inference in biophysics, chemistry and biology; protein folding; big data analytics; digital disease tracking; Bayesian inference for stochastic models; nonparametric statistical methods; model selection and empirical Bayes methods; and Monte Carlo methods.

John Lafferty
Professor, Yale University

Statistical Estimation and Machine Learning Under Communication and Shape Constraints

Friday, May 3, 2019
3:20pm – 4:20pm

Abstract: Imagine that I estimate a statistical model from data, and then want to share my model with you. But we are communicating over a resource constrained channel. By sending lots of bits, I can communicate my model accurately, with little loss in statistical risk. Sending a small number of bits will incur some excess risk. What can we say about the tradeoff between statistical risk and the communication constraints? We provide a sharp analysis in certain nonparametric settings under centralized and distributed communication protocols.

Now suppose that I wish to use a complex machine learning algorithm for prediction, but want my prediction rule to obey natural shape constraints suggested by domain knowledge. We present two methods for high-dimensional shape-constrained regression and classification that "reshape" the original prediction rule. The first method can be applied to any pre-trained prediction rule, while the second method deals specifically with random forests. In both cases, efficient algorithms are developed for computing the estimators to enforce the shape constraints without compromising predictive accuracy.

Joint work with Rina Barber, Matt Bonakdarpour, Sabyasachi Chatterjee, and Yuancheng Zhu.

Bio: John D. Lafferty is the John C. Malone Professor of Statistics and Data Science at Yale University. He was previously the Louis Block Professor of Statistics and Computer Science at the University of Chicago before joining Yale in July 2017. His honors include four “Test of Time” awards from the International Conference on Machine Learning. In 2015, he was elected to deliver an Institute of Mathematical Statistics Medallion Lecture. He conducts research on statistical machine learning, with a focus on computational and statistical aspects of nonparametric methods, high-dimensional data, graphical models, and statistical language modeling.