The 33rd New England Statistics Symposium
Statistical Data Science in Action

https://symposium.nestat.org/

Department of Statistics, University of Connecticut

May 15–17, 2019, Hilton Hartford, CT

The Department of Statistics of the University of Connecticut is proud to host the first remodeled, 3-day NESS after the establishment of the New England Statistical Society, on May 15–17, 2019.

Four Short Courses Wednesday, May 15, 2019 (Course 1/2 are full-day while 3/4 are half-day)

1. Intermediate Machine Learning: Key Concepts and Techniques, by Dr. David Rosenberg (Bloomberg)
2. Big Data Analytics and Deep Learning, by Dr. Ming Li (Amazon) and Dr. Hui Lin (Netlify)
3. Practical Visualization for Data Scientists, by Dr. Xiaoyue Cheng (University of Nebraska at Omaha)
4. Introduction to Multilevel Modeling, by Dr. Min Zhu (SAS)

Plenary Presentations (https://symposium.nestat.org/plenary-speakers.html)

Opening Keynote Thursday, May 16, 2019

Dr. Sam Kou, Harvard University: Big Data, Google and Disease Detection: A Statistical Adventure

Dr. David Rosenberg / Dr. Amanda Stent, Bloomberg: Machine Learning for Structured and Unstructured Data in Finance

Banquet Talk Thursday, May 16, 2019

Dr. Anthony D'Amico, Dana-Farber Cancer Institute and Brigham & Women's Hospital: Improving Patient Outcomes in Prostate Cancer by Partnering Statistics and Medicine

Chernoff Lecture Friday, May 17, 2019

The inaugural Chernoff Awardee whose identity remains a secret until the closing Award Ceremony.

Invited Sessions A total of 63 sessions cover a wide spectrum of areas of Statistical Data Science in Action. The schedule is online at https://symposium.nestat.org/program.html

Contributed Poster Session Thursday, May 16, 2019. Contribution is welcome from everyone. Presenting students automatically enter the student poster award competition.

Student Competitions Students are encouraged to participate in three events.

Travelers Stat-a-thon (http://statathon.stat.uconn.edu/) A statistical data science invention marathon sponsored by Travelers. Two data challenges have been released: Connecticut Housing; and Customer Retention. Students form teams to compete. The deadline of submission is April 26, 2019.

IBM Student Paper Competition Submission deadline is April 1, 2019.

Liberty Mutual Student Poster Competition Submission deadline is April 26, 2019.

Registration All participants in the conference must register. The registration fee covers conference materials, breakfasts, lunches and tea breaks. Members of the New England Statistical Society (https://nestat.org) receive discounts in pricing. To register, please complete the on-line registration form at https://symposium.nestat.org/registration.php by Monday, May 6, 2019.

Venue Hilton Hartford will offer discount rate to the conference participants until April 18th or until the group block is sold out. To book, please visit https://symposium.nestat.org/venue.html.

Welcoming Reception 5:00-6:30 pm, Wednesday, May 14. Sponsored by Munich Re Group – The Hartford Steam Boiler Inspection and Insurance Co, free and open to all registered participants:

Highlights of the 33rd NESS, 2019

Statistical Data Science in Action

Detailed schedules of 60+ sessions at `https://symposium-dev.nestat.org/program.html`

In addition to the plenary sessions (keynote, banquet, Chernoff, and award), a subset of parallel sessions themed with "Statistical Data Science in Actions" are put in different tracks.

Travelers Stat-a-thon Finalist Presentations Finalists of the 2019 Travelers Stat-a-thon data challenges will present their work in front of a judge panel on Thursday, May 16, in two sessions, one on Connecticut Housing and the other on Customer Retention. Winners will be presented with their prizes in the closing award session on Friday, May 17, 2019.

Education and Career Development

- Career Opportunities in Statistics and Data Sciences
- Effective Communication of Technical Concepts to Non-Technical Audiences
- Teaching Data Science
- Leadership Forum for Statistics/Data Science Professionals

Showcase of Statistical Data Science in Action

- Biomedical Data Science: To Infinity and Beyond!
- Data Science in Actuarial Science
- Data Science and Statistics in Insurance
- Predictive Modeling in Data Science: Methods and Applications
- Some Applications of Statistics in Data Science
- Some Applications of Data Science in Industry
- Application of Statistics and Data Sciences in Pharmaceutical Research and Development
- Manifolds and Anomalies for Data Science
- The Role and Advances of Principled Statistical Inference in the Era of Data Science
- Sports Analytics
- Novel Methodology and Application of Machine Learning Techniques in Data Science
- Data Science in Action at IBM Chief Analytics Office
- Novel Statistical Methods for Data Science: Discrete Data, Time Series and Data Integration
- Robust Statistics for Data Science
- Data Science: Computational Social Science?
- Using Statistics for Health Delivery System Reform in Massachusetts
- Healthcare Data Analysis for Electronic Health Records
- Statistical Methodology for Healthcare Data Analysis
- Novel Statistical Methods for the Analysis of Genomic Data
- Air Pollution and Public Health
- Statistical Tools for Addressing the Opioid Epidemic
- Modern Statistical Methods for Health Outcome Data
- Critical Questions in Drug development
- Statistical and Machine Learning Methods for Large-Scale Biomedical Data Analysis

Probability, Statistics, and Geometry: A Special Session Honoring Professor Rick Vitale This session is to honor Professor Rick Vitale for his academic achievements and to celebrate his 75th birthday. Chaired by David Pollard (Yale University), it starts at 1:00 pm on Friday, May 17:

- Andrew Barron (Yale University): Gaussian Complexity, Metric Entropy, and Risk of Deep Nets
- David Donoho (Stanford University): The Statistical Significance of Perfect Linear Separation
- Subhashis Ghoshal (North Carolina State University): Coverage of Credible Intervals for Monotone Regression

NESS | New England Statistical Society
New England Statistics Symposium since 1987 & New England Statistical Society since 2017

Statistical Data Science LAB UCONN

UCONN
COLLEGE OF LIBERAL ARTS AND SCIENCES
DEPARTMENT OF STATISTICS

Travelers Stat-a-thon 2019

http://statathon.stat.uconn.edu/

Join us for the first ever Stat-a-thon Competition!

Co-sponsored by the New England Statistical Society, Connecticut Open Data, Travelers, UConn Department of Statistics, and UConn Statistical Data Science Lab, Stat-a-thon is a statistical data science invention marathon with realworld data science problems. It emphasizes the statistical aspects (insight, interpretation, significance, etc.) of data science problems that are often overlooked in many hackathons. The 2019 stat-a-thon is a companion event of the 33rd New England Statistics Symposium (NESS), named after our major sponsor Travelers.

Theme 1: Connecticut Housing Using the Connecticut Open Data (`https://data.ct.gov/`), find interesting insights, trends, correlations, relationships, or patterns in housing in Connecticut.

Theme 2: Customer Retention Predict insurance policies that are most likely to cancel as well as understand what variables are most influential in causing a policy cancellation.

Timeline

- March 4: Team registration opens
- March 25: Registration deadline for individuals looking for a team
- April 15, Team registration closes
- April 26: Submission deadline
- May 3: Notification of finalist teams
- May 16: Finalist teams present to the review panel at NESS
- May 17: Awards to the winning teams at the closing Award Ceremony of NESS

CASH PRIZES for winning teams!

Ct DATA

TRAVELERS

NESS | New England Statistical Society
New England Statistics Symposium *since 1987* & New England Statistical Society *since 2017*

Four Short-Courses at
The 33rd New England Statistics Symposium
https://symposium.nestat.org/short-courses.html

Wednesday, May 15, 2019 **8:40am — 4:20pm** **Hilton Hartford**

Course 1 (Full Day): Intermediate Machine Learning: Key Concepts and Techniques

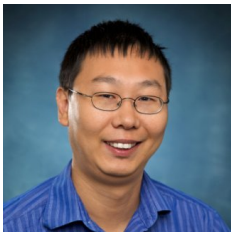
Instructor Dr. David Rosenberg is a Data Scientist in the data science group in the Office of the CTO at Bloomberg, and an Adjunct Associate Professor at the Center for Data Science at New York University, where he has repeatedly received NYU's Center for Data Science "Professor of the Year" award. He received his Ph.D. in statistics from UC Berkeley, where he worked on statistical learning theory and natural language processing. David received a Master of Science in applied mathematics, with a focus on computer science, from Harvard University, and a Bachelor of Science in mathematics from Yale University.

Outline There is no shortage of tutorials describing a huge variety of machine learning models and techniques. In this short course, we take a step back and present a coherent framework for thinking about supervised machine learning models in general. We examine multiple examples of the four fundamental components of a machine learning method: loss function, regularization, hypothesis space, and optimization method. Within this framework, we'll study linear regression (regression, lasso, and elastic net) and classification methods. We'll also introduce the most important nonlinear models, including tree-based ensemble methods and neural network models. Time-permitting, we may discuss conditional probability models, noting that the vast majority of contemporary deep learning models are of this type, as well as an approach to multiclass classification that generalizes to structured prediction and ranking problems, among many other applications. Throughout our discussion, we'll introduce the terminology and notation used by experts in machine learning to help bridge the gap between introductory-level tutorials and the more advanced materials you can find at conferences and in graduate-level courses.

Prerequisites Familiarity with basic mathematical notation (e.g., $\sum$ for summation, arg min), basic linear algebra (e.g., matrix multiplication, projections, inner products, norms, and spans), and introductory probability (probability distributions, conditional probability and conditional expectation).

Course 2 (Full Day): Big Data Analytics and Deep Learning

Instructors Dr. Hui Lin is leading and building data science department at Netlify. Before Netlify, she was a Data Scientist at DuPont. She was a leader in the company of applying advanced data science to enhance Marketing and Sales Effectiveness. She provided data science leadership for a broad range of predictive analytics and market research analysis from 2013 to 2018. She is the co-founder of Central Iowa R User Group, blogger of scientistcafe.com and 2018 Program Chair of ASA Statistics in Marketing Section. She enjoys making analytics accessible to a broad audience and teaches tutorials and workshops for practitioners on data science. She holds MS and Ph.D in statistics from Iowa State University.



Dr. Ming Li is currently a Research Scientist at Amazon. He organized and presented 2018 JSM Introductory Overview Lecture: Leading Data Science: Talent, Strategy, and Impact. He was the Chair of Quality & Productivity Section of ASA for 2017. He was a Data Scientist at Walmart and a Statistical Leader at General Electric Global Research Center. He obtained his Ph.D. in Statistics from Iowa State University in 2010. With deep statistics background and a few years' experience in data science, he has trained and mentored numerous junior data scientists with different background such as statistician, programmer, software developer, database administrator and business analyst. He is also an Instructor of Amazon's internal Machine Learning University and was one of the key founding member of Walmart's Analytics Rotational Program which bridges the skill gaps between new hires and productive data scientists.

Outline In the past couple of years, deep learning has gained traction in many areas. It becomes an essential tool in data scientist's toolbox. In this course, students will develop a clear understanding of the big data cloud platform, technical skills in data sciences and machine learning, the motivation and use cases of deep learning through hands-on exercises. We will also cover the "art" part of data science: data science project flow, general pitfalls in data

science and machine learning, and soft skills to effectively communicate with business stakeholders. The course is for audience with statistics background. We use real-world data science and machine learning problems to illustrate data science workflow, pitfalls, and soft skills. The hands-on sessions use Databricks community edition cloud platform. Specific modules are: (1) big data platform using Spark through R **sparklyr** package; (2) introduction to deep neural network, convolutional neural network recurrent neural networks, and their applications; (3) deep learning examples using TensorFlow through R **keras** package.

Prerequisites Introductory statistics or practical experience in data science ; entry level of R knowledge; a free Databrick Community Edition account through `https://databricks.com/try-databricks`; a laptop.

Course 3 (Half Day): Practical Visualization for Data Scientists

Instructor Dr. Xiaoyue Cheng is an Assistant Professor in the Department of Mathematics, University of Nebraska at Omaha. She received her Ph.D. in Statistics from Iowa State University in 2015. Her research interests include data visualization, interactive graphics, image recognition, machine learning, statistical computing and simulation, exploratory data analysis, and missing data analysis. She has extensive interdisciplinary research experience in a variety of fields including education, ethnicity population, psychology, medical clinics, public health, engineering, aviation, agronomy, and business marketing. Cheng is the main author of five R packages: **MissingDataGUI**, **cranvas**, **cartogram**, **ePort**, and **MergeGUI**.

Outline Visualization has an important role in data science as it is widely used for data exploration, information delivery, and communication among people at different positions or from different backgrounds. Specifically, statistical graphics focuses on revealing the patterns, trends, and relationships from dataset with complex challenges including the massive amount, high dimensions, and various formats of data. This short course will introduce the advanced programming skills to accurately and attractively communicate data information with visualization. Topics of the course will include (1) the elements and grammar of graphics via the **ggplot2** package; (2) interactive web apps by the **shiny** package; (3) time series data visualization using the **dygraphs** package; (4) geographic/map/GIS data visualization using the **leaflet** package; and (5) interactive graphics with the **plotly** package. Depending on the audience, example data from different research topics such as the US Census Bureau, business marketing, clinical trials, images, etc., will be applied to demonstrate the visual methods. R language will be employed for this course, and the attendees will have the chance to generate graphs on their own for all of the topics.

Prerequisites Introductory statistics; basic programming experience with R; a laptop.

Course 4 (Half-Day): Introduction to Multilevel Modeling

Instructor Dr. Min Zhu is a senior research statistician developer at SAS Institute Inc. She joined SAS in 2009 after receiving her Ph.D in Statistics from University of New Mexico. Her research and development focus is the generalized linear mixed model procedure (PROC GLIMMIX). She works closely with statisticians and researchers in both academia and industry on performance and functionality improvement in PROC GLIMMIX. Her recent work include enhancements for small sample inference, efficient numerical integration, and multilevel weighting for complex survey analysis. Dr. Zhu enjoys advocating the tool of mixed models to the community of statisticians and data analysts. She has organized sessions and taught courses on generalized linear mixed models at Joint Statistical Meeting, SAS Global Forum, and ASA Biopharmaceutical Section FDA-Industry Workshop.

Outline Hierarchical data are common in many fields, from pharmaceuticals to agriculture to sociology. As you collect more and more data, information is likely to be observed on nested units at multiple levels, calling for a multilevel modeling approach. This course will show you how to construct a multilevel model to account for variability at each level through both explanatory and random variables, in a way that shows the close relationship between multilevel models and mixed models. You will learn how to use generalized linear mixed models to estimate multilevel models for both continuous and discrete responses. You will see examples that illustrate the flexibility multilevel offers for modeling within-cluster correlation, for disentangling multilevel explanatory variables, and for differentiating between-cluster and within-cluster effects. You will also learn about weighted multilevel models that handle weights at different levels. Finally, you will see how to apply weighted multilevel models to the analysis of complex survey data that are collected by multistage sampling with unequal sampling probabilities.

Prerequisites Introductory courses in statistical modeling and statistical inference; experience with multilevel data.