STATISTICS AND DATA SCIENCE

219 Prospect Street, 203.432.0666
http://statistics.yale.edu
M.A., M.S., Ph.D.

Chair
Yihong Wu

Acting Chair
Daniel Spielman [Sp]

Directors of Graduate Studies
John Emerson (219 Prospect, john.emerson@yale.edu)
Zhou Fan (219 Prospect, zhou.fan@yale.edu)

Professors Donald Andrews (*Economics*), Andrew Barron, Jeffrey Brock (*Mathematics*), Joseph Chang, Katarzyna Chawarska (*Child Study Center*), Xiaohong Chen (*Economics*), Nicholas Christakis (*Sociology*), Ronald Coifman (*Mathematics*), James Duncan (*Radiology and Biomedical Imaging*), John Emerson (*Adjunct*), Alan Gerber (*Political Science*), Mark Gerstein (*Molecular Biophysics and Biochemistry*), Anna Gilbert, John Hartigan (*Emeritus*), Edward Kaplan (*School of Management*), Harlan Krumholz (*Internal Medicine*), John Lafferty, Zongming Ma, David Pollard (*Emeritus*), Nils Rudi (*School of Management*), Jasjeet Sekhon, Donna Spiegelman (*Biostatistics*), Daniel Spielman, Hemant Tagare (*Radiology and Biomedical Engineering*), Van Vu (*Mathematics*), Yihong Wu, Heping Zhang (*Biostatistics*), Hongyu Zhao (*Biostatistics*), Harrison Zhou, Steven Zucker (*Computer Science*)

Associate Professors P.M. Aronow, Forrest Crawford (*Biostatistics*), Joshua Kalla (*Political Science*), Amin Karbasi (*Electrical Engineering*), Vahideh Manshadi (*School of Management/Operations*), Ethan Meyers (*Visiting*), Sekhar Tatikonda

Assistant Professors Elisa Celis, Sinho Chewi, Zhou Fan, Melody Huang (*Political Science*), Roy Lederman, Lu Lu, Theodor Misiakiewicz, Omar Montasser, Fredrik Savje (*Political Science*), Dustin Scheinost (*Radiology and Biomedical Imaging*), Ramina Sotoudeh (*Sociology*), Andre Wibisono (*Computer Science*), Zhuoran Yang, Ilker Yildirim (*Psychology*), Ilias Zadik

FIELDS OF STUDY

Fields of study include the main areas of statistical theory (with emphasis on foundations, Bayes theory, decision theory, nonparametric statistics), probability theory (stochastic processes, asymptotics, weak convergence), information theory, bioinformatics and genetics, classification, data mining and machine learning, neural nets, network science, optimization, statistical computing, and graphical models and methods.

SPECIAL REQUIREMENTS FOR THE PH.D. DEGREE IN STATISTICS AND DATA SCIENCE

There is no foreign language requirement. Students take at least twelve courses, usually during the first two years. The department requires that students take S&DS 625,

Statistical Case Studies, and S&DS 626, Practical Work. The department strongly recommends that students take:

S&DS 551	Stochastic Processes	1
S&DS 600	Advanced Probability	1
S&DS 610	Statistical Inference	1
S&DS 612	Linear Models	1
S&DS 631	Optimization and Computation	1
S&DS 632	Advanced Optimization Techniques	1
S&DS 661	Data Analysis	1

Substitutions are possible with the permission of the director of graduate studies (DGS); courses from other complementary departments such as Mathematics and Computer Science are encouraged. With the permission of the DGS and under special circumstances, appropriate courses may be taken at the undergraduate level in departments outside of Statistics and Data Science to fulfill these elective requirements.

The qualifying examination consists of three parts: a written report on an analysis of a data set, one or more written examination(s), and an oral examination. The examinations are taken as scheduled by the department. All parts of the qualifying examination must be completed before the beginning of the third year. A prospectus for the dissertation should be submitted no later than the first week of March in the third year. The prospectus must be accepted by the department before the end of the third year if the student is to register for a fourth year. Upon successful completion of the qualifying examination and the prospectus (and meeting of graduate school requirements), the student is admitted to candidacy. Students are expected to attend weekly departmental seminars.

Students normally serve as teaching fellows for several terms to acquire professional training. All students are required to be teaching fellows for a minimum of two terms, regardless of the nature of their funding. The timing of this teaching is at the discretion of the DGS.

COMBINED PH.D. PROGRAM

The Department of Statistics and Data Science also offers, in conjunction with the Department of Political Science, a combined Ph.D. in Statistics and Data Science and Political Science. For further details, see Political Science.

MASTER'S DEGREES

Three different M.A. in Statistics are offered. All require completion of eight term courses approved by the DGS; of which one must be in probability, one must be in statistical theory, and one must be in data analysis. The remaining five elective courses may include courses from other departments and, with the permission of the DGS and under special circumstances, appropriate courses may be taken at the undergraduate level in departments outside of Statistics and Data Science.

M.A. in Statistics (en route to the Ph.D. in Statistics and Data Science) This degree requires an average grade of HP or higher, and two terms of residence.

M.A. in Statistics (en route to the Ph.D. in other areas of study) Pursuit of this degree requires an application process managed by the DGS of Statistics and Data Science followed by approval from the DGSs from both programs and the cognizant Graduate School dean. All eight courses for this degree must earn grades of HP or higher. This degree also has an academic teaching fellow requirement, to be determined by the DGSs from both programs and the cognizant graduate school dean.

Terminal M.A. in Statistics Students are also admitted directly to a terminal master of arts program in Statistics. Students must earn an average grade of HP or higher and receive at least one grade of Honors. Full-time students must take a minimum of four courses per term. Part-time students are also accepted into the program. All students are expected to complete two terms of full-time tuition and residence, or the equivalent, at Yale. See Degree Requirements: Terminal M.A./M.S. Degrees, under Policies and Regulations.

Terminal M.S. in Statistics and Data Science Students are also admitted directly to a terminal master of science program in Statistics and Data Science. To qualify for the M.S., the student must successfully complete an approved program of twelve term courses with an average grade of HP or higher and receive at least two grades of Honors, chosen in consultation with the DGS. With the permission of the DGS and under special circumstances, appropriate courses may be taken at the undergraduate level in departments outside of Statistics and Data Science to fulfill elective requirements. Full-time students must take a minimum of four courses per term. Part-time students are also accepted into the program. All students are expected to complete three terms of full-time tuition and residence, or the equivalent, at Yale. See Degree Requirements: Terminal M.A./M.S. Degrees, under Policies and Regulations.

Program information is available online at http://statistics.yale.edu.

COURSES

S&DS 500a or b, Introductory Statistics Robert Wooster
An introduction to statistical reasoning. Topics include numerical and graphical summaries of data, data acquisition and experimental design, probability, hypothesis testing, confidence intervals, correlation and regression. Application of statistical concepts to data; analysis of real-world problems.

S&DS 517b, Applied Machine Learning and Causal Inference P Aronow
Approaches to causal inference using machine learning. Covers randomized experiments with and without noncompliance, observational studies with and without ignorable treatment assignment, instrumental variables, and regression discontinuity. Machine-learning methods include bagging, boosting, tree-based methods such as random forests, and neural networks. Assignments provide students with hands-on experience with the methods. Applications are drawn from a variety of fields including political science, economics, public health, and medicine. Programming is central to the course and is based on the R programming language. Prerequisites: the equivalent of at least two of the following courses: S&DS 530, S&DS 538, S&DS 541, and S&DS 542; and previous programming experience (e.g., R, MATLAB, Python, C++), R preferred. Strong knowledge of OLS is assumed.

S&DS 520b, Intensive Introductory Statistics Robert Wooster
An introduction to statistical reasoning designed for students with particular interest in data science and computing. Using the R language, topics include exploratory data analysis, probability, hypothesis testing, confidence intervals, regression, statistical modeling, and simulation. Computing is taught and used extensively throughout the course. Application of statistical concepts to the analysis of real-world data science problems.

S&DS 523a or b, YData: An Introduction to Data Science Ethan Meyers
Computational, programming, and statistical skills are no longer optional in our increasingly data-driven world; they are essential for opening doors to manifold research and career opportunities. This course aims to dramatically enhance students' knowledge and capabilities in fundamental ideas and skills in data science, especially computational and programming skills and inferential thinking. It emphasizes the development of these skills while providing opportunities for hands-on experience and practice. The course is designed to be accessible to students with little or no background in computing, programming, or statistics, but also engaging for more technically oriented students through extensive use of examples and hands-on data analysis. Python 3 is the computing language used. Enrollment is limited.

S&DS 530a or b / PLSC 530a or b, Data Exploration and Analysis Staff
Survey of statistical methods: plots, transformations, regression, analysis of variance, clustering, principal components, contingency tables, and time series analysis. The R computing language and web data sources are used.

S&DS 538a, Probability and Statistics Joseph Chang
Fundamental principles and techniques of probabilistic thinking, statistical modeling, and data analysis. Essentials of probability: conditional probability, random variables, distributions, law of large numbers, central limit theorem, Markov chains. Statistical inference with emphasis on the Bayesian approach: parameter estimation, likelihood, prior and posterior distributions, Bayesian inference using Markov chain Monte Carlo. Introduction to regression and linear models. Computers are used throughout for calculations, simulations, and analysis of data. Prerequisite: after or concurrently with MATH 118 or MATH 120.

S&DS 540b, An Introduction to Probability Theory Elisa Celis
Introduction to probability theory. Topics include probability spaces, random variables, expectations and probabilities, conditional probability, independence, discrete and continuous distributions, central limit theorem, Markov chains, and probabilistic modeling. *This course may be appropriate for non-S&DS graduate students.* Prerequisite: MATH 115 or equivalent.

S&DS 541a, Probability Theory Harrison Zhou
A first course in probability theory: probability spaces, random variables, expectations and probabilities, conditional probability, independence, some discrete and continuous distributions, central limit theorem, Markov chains, probabilistic modeling. Prerequisite: calculus of functions of several variables.

S&DS 542a or b, Theory of Statistics Staff
Principles of statistical analysis: maximum likelihood, sampling distributions, estimation, confidence intervals, tests of significance, regression, analysis of variance, and the method of least squares. Prerequisite: S&DS 541.

S&DS 551b / ENAS 502b, Stochastic Processes Ilias Zadik
Introduction to the study of random processes, including Markov chains, Markov random fields, martingales, random walks, Brownian motion, and diffusions. Techniques in probability such as coupling and large deviations. Applications chosen from image reconstruction, Bayesian statistics, finance, probabilistic analysis of algorithms, genetics, and evolution.

S&DS 563b, Multivariate Statistical Methods for the Social Sciences Jonathan Reuning-Scherer
An introduction to the analysis of multivariate data. Topics include principal components analysis, factor analysis, cluster analysis (hierarchical clustering, k-means), discriminant analysis, multidimensional scaling, and structural equations modeling. Emphasis on practical application of multivariate techniques to a variety of examples in the social sciences. Students complete extensive computer work using either SAS or SPSS. Prerequisites: knowledge of basic inferential procedures, experience with linear models (regression and ANOVA). Experience with some statistical package and/or familiarity with matrix notation is helpful but not required.

S&DS 565a, Introductory Machine Learning John Lafferty
This course covers the key ideas and techniques in machine learning without the use of advanced mathematics. Basic methodology and relevant concepts are presented in lectures, including the intuition behind the methods. Assignments give students hands-on experience with the methods on different types of data. Topics include linear regression and classification, tree-based methods, clustering, topic models, word embeddings, recurrent neural networks, dictionary learning, and deep learning. Examples come from a variety of sources including political speeches, archives of scientific articles, real estate listings, natural images, and others. Programming is central to the course and is based on the Python programming language.

S&DS 572a, YData: Data Science for Political Campaigns Joshua Kalla
Political campaigns have become increasingly data driven. Data science is used to inform where campaigns compete, which messages they use, how they deliver them, and among which voters. In this course, we explore how data science is being used to design winning campaigns. Students gain an understanding of what data is available to campaigns, how campaigns use this data to identify supporters, and the use of experiments in campaigns. The course provides students with an introduction to political campaigns, an introduction to data science tools necessary for studying politics, and opportunities to practice the data science skills presented in S&DS 523.

S&DS 573b, YData: Analysis of Baseball Data Ethan Meyers
The field of data science aims to extract insights from large data sets that often contain random variation. Baseball is a game that contains a high degree of randomness, and because professional baseball has been played since the nineteenth century, a large amount of data has been collected about players' performance. In this class we use baseball data to understand key concepts in data science including data visualization, data wrangling, and statistical inference. To understand these concepts, we analyze data that include season-level statistics going back to the 1870s, play-by-play statistics going back to the 1930s, and pitch trajectory statistics going back to 2006. The course uses the Python programming language and is paced to be accessible to students who have previously taken or are currently enrolled in S&DS 523. Co-requisite: S&DS 523.

S&DS 600a, Advanced Probability Sekhar Tatikonda
Measure theoretic probability, conditioning, laws of large numbers, convergence in distribution, characteristic functions, central limit theorems, martingales. Some knowledge of real analysis is assumed.

S&DS 602a, High-Dimensional Probability and Applications Zhou Fan
This course covers techniques for studying high-dimensional probabilistic problems, with a focus on non-asymptotic methods that find common use in applications across statistics, machine learning, computer science, and engineering. Topics covered include tail bounds for i.i.d. sums and martingale differences, concentration inequalities for non-linear functions, matrix concentration inequalities, suprema of Gaussian processes, and interpolation techniques for understanding universality of high-dimensional phenomena. Prerequisite: S&DS 351b/551b, S&DS 400/600 (may be taken concurrently), or permission of instructor.

S&DS 605a, Sampling and Optimal Transport Sinho Chewi
MCMC sampling and variational inference have long been utilized in Bayesian statistics and machine learning; what can we say about the convergence of these methods? Recently, a modern theory has emerged which blends principles from convex optimization with a geometric perspective on the space of probability distributions based on optimal transport. This course provides an introduction to this theory, as well as to related tools used for modern algorithmic analysis: Markov semigroup theory and stochastic calculus, coupling, and functional inequalities. Much of the course focuses on the complexity of log-concave sampling, but we also discuss applications to diffusion models and variational inference. Prerequisite: Advanced Probability (S&DS 400 / S&DS 600 MATH 330). The following are helpful but not required: Optimization (S&DS 431 / S&DS 631, S&DS 432 / S&DS 632) and Stochastic Processes (S&DS 351 / S&DS 551). Enrollment is limited; requires permission of the instructor.

S&DS 610a, Statistical Inference Theodor Misiakiewicz
A systematic development of the mathematical theory of statistical inference covering methods of estimation, hypothesis testing, and confidence intervals. An introduction to statistical decision theory. Knowledge of probability theory at the level of S&DS 541 is assumed.

S&DS 612a, Linear Models Zongming Ma
The geometry of least squares; distribution theory for normal errors; regression, analysis of variance, and designed experiments; numerical algorithms (with particular reference to the R statistical language); alternatives to least squares. Prerequisites: linear algebra and some acquaintance with statistics.

S&DS 625a or b, Statistical Case Studies Staff
Statistical analysis of a variety of statistical problems using real data. Emphasis on methods of choosing data, acquiring data, assessing data quality, and the issues posed by extremely large data sets. Extensive computations using R. Enrollment limited; requires permission of the instructor.

S&DS 626b, Practical Work Jay Emerson
Individual one-term projects, with students working on studies outside the department, under the guidance of a statistician.

S&DS 627a and S&DS 628a or b, Statistical Consulting Jay Emerson
Statistical consulting and collaborative research projects often require statisticians to explore new topics outside their area of expertise. This course exposes students to real problems, requiring them to draw on their expertise in probability, statistics, and data analysis. Students complete the course with individual projects supervised jointly by faculty outside the department and by one of the instructors. Students enroll for both terms (S&DS 627 and 628) and receive one credit at the end of the year. Enrollment limited; requires permission of the instructor. ½ Course cr per term

S&DS 631a / AMTH 631a, Optimization and Computation Zhuoran Yang
An introduction to optimization and computation motivated by the needs of computational statistics, data analysis, and machine learning. This course provides foundations essential for research at the intersections of these areas, including the asymptotic analysis of algorithms, an understanding of condition numbers, conditions for optimality, convex optimization, gradient descent, linear and conic programming, and NP hardness. Model problems come from numerical linear algebra and constrained least squares problems. Other useful topics include data structures used to represent graphs and matrices, hashing, automatic differentiation, and randomized algorithms. Prerequisites: multivariate calculus, linear algebra, probability, and permission of the instructor. Enrollment is limited, with preference given to graduate students in Statistics and Data Science.

S&DS 632b, Advanced Optimization Techniques Staff
This course covers fundamental theory and algorithms in optimization, emphasizing convex optimization. Topics covered include convex analysis; duality and KKT conditions; subgradient methods; interior point methods; semidefinite programming; distributed methods; stochastic gradient methods; robust optimization; and an introduction to nonconvex optimization. Applications from statistics and data science, economics, engineering, and the sciences. Prerequisites: knowledge of linear algebra, such as MATH 222 or MATH 225; multivariate calculus, such as MATH 120; probability, such as S&DS 541; optimization, such as S&DS 631; and comfort with proof-based exposition and problem sets.

S&DS 661b, Data Analysis Brian Macdonald
By analyzing data sets using the R statistical computing language, a selection of statistical topics are studied: linear and nonlinear models, maximum likelihood, resampling methods, curve estimation, model selection, classification, and clustering. Prerequisite: after or concurrent with S&DS 542.

S&DS 663a, Computational Mathematics Situational Awareness and Survival Skills Roy Lederman
Are you using a computer to analyze data? Will the computer ever finish processing the data? Will the result be junk? Will you recognize that it is junk? We discuss the difference between math on paper and math on a computer and the difference between general programming and implementing mathematics on a computer. We experience benign mathematical operations failing catastrophically without any error message. We experience mathematically equivalent operations taking anywhere between a fraction of a second and a lifetime. We develop situational awareness and survival skills for this harsh environment. We discuss algorithms, complexity, numerical computation, linear algebra, data analysis, programming, and prototyping. Assignments include theory, programming, data analysis, individual work, and collaborative work. We use C

(optionally, FORTRAN) and Python. Making mistakes on assignments and respectful class discussions of insights from such mistakes are integral parts of the course. Prerequisites: linear algebra, multivariate calculus, and programming experience (any language). Prior experience with C, FORTRAN, or Python is recommended but not required; students unfamiliar with these languages must be comfortable independently learning them during the course. Limited size. Instructor permission is required.

S&DS 664b, Information Theory Staff
Foundations of information theory in communications, statistical inference, statistical mechanics, probability, and algorithmic complexity. Quantities of information and their properties: entropy, conditional entropy, divergence, redundancy, mutual information, channel capacity. Basic theorems of data compression, data summarization, and channel coding. Applications in statistics.

S&DS 665a, Intermediate Machine Learning John Lafferty
S&DS 365 is a second course in machine learning at the advanced undergraduate or beginning graduate level. The course assumes familiarity with the basic ideas and techniques in machine learning, for example as covered in S&DS 265. The course treats methods together with mathematical frameworks that provide intuition and justifications for how and when the methods work. Assignments give students hands-on experience with machine learning techniques, to build the skills needed to adapt approaches to new problems. Topics include nonparametric regression and classification, kernel methods, risk bounds, nonparametric Bayesian approaches, graphical models, attention and language models, generative models, sparsity and manifolds, and reinforcement learning. Programming is central to the course, and is based on the Python programming language and Jupyter notebooks.

S&DS 669a, Statistical Learning Theory Omar Montasser
This course covers classical topics and recent advances in statistical learning theory. This includes topics such as PAC learning, VC theory, boosting, and online learning. We explore statistical and computational aspects, with an emphasis on developing a rigorous quantitative understanding of key machine learning concepts. A second aim is to introduce technical tools that help with designing learning algorithms and proving learning guarantees. Prerequisites: Mathematical maturity and comfort with proof-oriented courses. Background in probability (e.g., S&DS 241), machine learning (e.g., S&DS 265), and algorithms (e.g., CPSC 365). Familiarity with basic concepts in computational complexity (e.g., NP-hardness) is helpful but not required.

S&DS 674b, Applied Spatial Statistics Timothy Gregoire
An introduction to spatial statistical techniques with computer applications. Topics include modeling spatially correlated data, quantifying spatial association and autocorrelation, interpolation methods, variograms, kriging, and spatial point patterns. Examples are drawn from ecology, sociology, public health, and subjects proposed by students. Four to five lab/homework assignments and a final project. The class makes extensive use of the R programming language as well as ArcGIS.

S&DS 685b, Theory of Reinforcement Learning Zhuoran Yang
There has been a surge of research interest in reinforcement learning recently, fueled by exciting applications of reinforcement learning techniques to various challenging decision-making problems in artificial intelligence, robotics, and natural sciences. Many of these advances were made possible by a combination of innovative use of

flexible neural network architectures, modern optimization techniques, and new and classical RL algorithms. However, a systematic understanding of when, why, and to what extent these algorithms work remains active ongoing research. This course aims to introduce the theoretical foundations of reinforcement learning, with the goal of equipping students with necessary tools for conducting research. This graduate level course focuses on theoretical and algorithmic foundations of reinforcement learning. Specifically, there are four main themes of the course:(a) fundamentals of RL (Markov decision process, planning algorithms, Q-learning and temporal difference learning, policy gradient), (b) online RL (bandit algorithms, online learning, exploration), (c) offline RL (off-policy evaluation, offline policy learning), and (d) further topics (multi-agent RL, partial observability). Prerequisites: knowledge of linear algebra (MATH 225/226/240), multivariate calculus (MATH 255/256), probability (S&DS 241), and statistics (S&DS 242). Comfort with proof-based exposition and problem sets is also required.

S&DS 688a, Computational and Statistical Trade-offs in High Dimensional Statistics
Ilias Zadik

Modern statistical tasks require the use of both computationally efficient and statistically accurate methods. But, can we always find a computationally efficient method that achieves the information-theoretic optimal statistical guarantees? If not, is this an artifact of our techniques, or a potentially fundamental source of computational hardness? This course surveys a new and growing research area studying such questions on the intersection of high dimensional statistics and theoretical computer science. We discuss various tools to explain the presence of such "computational-to-statistical gaps" for several high dimensional inference models. These tools include the "low-degree polynomials" method, statistical query lower bounds, and more. We also discuss connections with other fields such as statistical physics and cryptography. Prerequisites: maturity with probability theory (equivalent of 241/541) and linear algebra and a familiarity with basic algorithms and mathematical statistics.

S&DS 689a, Scientific Machine Learning Lu Lu

There are two main branches of technical computing: scientific computing and machine learning. Recently, there has been a convergence of the two disciplines in the emerging scientific machine learning (SciML) field. The main objective of this course is to teach theory, algorithms, and implementation of SciML techniques to graduate students. This course entails various methods to solve a broad range of computational problems frequently encountered in solid mechanics, fluid mechanics, nondestructive evaluation of materials, systems biology, chemistry, and non-linear dynamics. The topics in this course cover multi-fidelity learning, physics-informed neural networks, deep neural operators, uncertainty quantification, and parallel computing. Certain materials are discussed through student presentations of selected publications in the area. Students should have prior coursework in advanced calculus, linear algebra, and probability. Having a background in scientific computing, Python, and/or machine learning is helpful but not mandatory.

S&DS 690a or b, Independent Study Jay Emerson

By arrangement with faculty. Approval of DGS required.

S&DS 695b, Summer Internship in Statistics and Data Science Jay Emerson

The purpose of this course is to provide students with the opportunity to gain practical experience in statistics and data science. Students who identify a suitable summer

internship consult with the DGS and prepare a one-page description of the plan. The internship must be full-time: 35–40 hours per week for 10–12 weeks during the summer. Upon completion of the internship, the student must submit a written report of the work to the instructor no later than October 1. Prerequisites: completion of at least one term of the M.S. program (or the M.A. program if transferring into the M.S. program) and permission of the DGS.

S&DS 700a or b, Departmental Seminar Staff
Presentations of recent breakthroughs in statistics and data science. 0 Course cr