STATISTICS AND DATA SCIENCE (S&DS)

S&DS 101a / E&EB 210a, Introduction to Statistics: Life Sciences Jonathan Reuning-Scherer
Statistical and probabilistic analysis of biological problems, presented with a unified foundation in basic statistical theory. Problems are drawn from genetics, ecology, epidemiology, and bioinformatics. QR

S&DS 102a / PLSC 452a, Introduction to Statistics: Political Science Jonathan Reuning-Scherer
Statistical analysis of politics, elections, and political psychology. Problems presented with reference to a wide array of examples: public opinion, campaign finance, racially motivated crime, and public policy. QR

S&DS 103a / PLSC 453a, Introduction to Statistics: Social Sciences Jonathan Reuning-Scherer
Descriptive and inferential statistics applied to analysis of data from the social sciences. Introduction of concepts and skills for understanding and conducting quantitative research. QR

S&DS 105a, Introduction to Statistics: Medicine Jonathan Reuning-Scherer
Statistical methods used in medicine and medical research. Practice in reading medical literature competently and critically, as well as practical experience performing statistical analysis of medical data. QR

S&DS 106a, Introduction to Statistics: Data Analysis Robert Wooster and Jonathan Reuning-Scherer
An introduction to probability and statistics with emphasis on data analysis. QR

S&DS 108a, Introduction to Statistics: Advanced Fundamentals Jonathan Reuning-Scherer
Introductory statistical concepts beyond those covered in high school AP statistics. Includes additional concepts in regression, an introduction to multiple regression, ANOVA, and logistic regression. This course is intended as a bridge between AP statistics and courses such as S&DS 230, Data Exploration and Analysis. Meets for the second half of the term only. Prerequisites: A previous statistics course in high school. May not be taken after S&DS 100, S&DS 101–106, PSYC 100, or any other full semester Yale introductory statistics courses. Students should consider S&DS 103 or both S&DS 108, 109. ½ Course cr

S&DS 109a, Introduction to Statistics: Fundamentals Jonathan Reuning-Scherer
General concepts and methods in statistics. Meets for the first half of the term only. May not be taken after or concurrently with S&DS 100 or 101–106. ½ Course cr

*** S&DS 110a, R for Statistical Computing and Data Science**
Intensive introduction to the R language, widely-accepted for statistical computing and graphics, and used by the data science industry as well as in a wide range of academic disciplines. It is a useful complement (concurrently or in advance) to many courses in

S&DS. Prerequisite: Some prior programming experience (in any language), even at the level of S&DS 100, or 101-109, or 123, 220, or 230. QR

* **S&DS 172a / EP&E 328a / PLSC 347a, 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. This 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 123, YData.
QR

* **S&DS 178a / SOCY 362a, Sociogenomics** Ramina Sotoudeh
Since the first human genome was sequenced in 2003, social and behavioral data have become increasingly integrated with genetic data. This has proven important not only for medicine and public health but also for social science. In this course, we cover the foundations of sociogenomics research. We begin by surveying core concepts in the field, from heritability to gene-by-environment interactions, and learning the computational tools necessary for producing sociogenomics research. In later weeks, we read some of the latest applied work in the field and discuss the value and limitations of such research. The course culminates in a final project, in which students are tasked with using empirical data to answer a social genetics question of their own. SO

S&DS 230a, Data Exploration and Analysis Ethan Meyers
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. Prerequisite: a 100-level Statistics course or equivalent, or with permission of instructor. QR

S&DS 238a, Probability and Bayesian Statistics Joseph Chang
Fundamental principles and techniques of probabilistic thinking, statistical modeling, and data analysis. Essentials of probability, including conditional probability, random variables, distributions, law of large numbers, central limit theorem, and 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 for calculations, simulations, and analysis of data. After or concurrently with MATH 118 or 120. QR

S&DS 240a, An Introduction to Probability Theory Robert Wooster
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 counts towards the Data Science certificate but not the Statistics and Data Science major. Prerequisite: MATH 115. QR

S&DS 241a / MATH 241a, Probability Theory Yihong Wu
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. After or concurrently with MATH 120 or equivalent. QR

S&DS 265a, 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 several others. Programming is central to the course, and is based on the Python programming language. Prerequisites: Two of the following courses: S&DS 230, 238, 240, 241 and 242; previous programming experience (e.g., R, Matlab, Python, C++), Python preferred. QR

* **S&DS 280a / NSCI 280a, Neural Data Analysis** Ethan Meyers
We discuss data analysis methods that are used in the neuroscience community. Methods include classical descriptive and inferential statistics, point process models, mutual information measures, machine learning (neural decoding) analyses, dimensionality reduction methods, and representational similarity analyses. Each week we read a research paper that uses one of these methods, and we replicate these analyses using the R or Python programming language. Emphasis is on analyzing neural spiking data, although we also discuss other imaging modalities such as magneto/electro-encephalography (EEG/MEG), two-photon imaging, and possibility functional magnetic resonance imaging data (fMRI). Data we analyze includes smaller datasets, such as single neuron recordings from songbird vocal motor system, as well as larger data sets, such as the Allen Brain observatory's simultaneous recordings from the mouse visual system. Prerequisite: S&DS 230. Background in neuroscience is recommended but not required (e.g., it would be useful to have taken at the level of NSCI 160).

S&DS 312a, 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. After S&DS 242 and MATH 222 or 225. QR

S&DS 365a, 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. Prerequisites: a background in probability and statistics at the level of S&DS 242; familiarity with the core ideas from linear algebra, for example through Math 222; and computational skills at the level of S&DS 265 or CPSC 200. QR

S&DS 400a / MATH 330a, 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 assumed. QR

S&DS 410a, Statistical Inference Harrison Zhou
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. Prerequisite: level of S&DS 241.

* **S&DS 425a, Statistical Case Studies** Brian Macdonald
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 statistical software. Prerequisites: S&DS 361, and prior course work in probability, statistics, and data analysis (e.g. 363, 365, 220, 230, etc., equivalent courses, or equivalent research/ internship experience). QR

S&DS 431a / AMTH 431a / ECON 431a, Optimization and Computation Yang Zhuoran
This course is designed for students in Statistics & Data Science who need to know about optimization and the essentials of numerical algorithm design and analysis. It is an introduction to more advanced courses in optimization. The overarching goal of the course is teach students how to design algorithms for Machine Learning and Data Analysis (in their own research). This course is not open to students who have taken S&DS 430. Prerequisites: Knowledge of linear algebra, multivariate calculus, and probability. Linear Algebra, by MATH 222, 223 or 230 or 231; Graph Theory, by MATH 244 or CPSC 365 or 366; and comfort with proof-based exposition and problem sets, such as is gained from MATH 230 and 231, or CPSC 366.

* **S&DS 480a, Individual Studies** Sekhar Tatikonda
Directed individual study for qualified students who wish to investigate an area of statistics not covered in regular courses. A student must be sponsored by a faculty member who sets the requirements and meets regularly with the student. Enrollment requires a written plan of study approved by the faculty adviser and the director of undergraduate studies.

S&DS 491a, Senior Project Brian Macdonald
Individual research that fulfills the senior requirement. Requires a faculty adviser and DUS permission. The student must submit a written report about results of the project.