STATISTICS AND DATA SCIENCE (S&DS)

S&DS 1000a or b, Introductory Statistics Ethan Meyers
An introduction to statistical reasoning. Topics include numerical and graphical summaries of data, data acquisition and experimental design, probability, hypothesis testing, confidence intervals, correlation, regression, multiple regression, and ANOVA. Application of statistical concepts to data; analysis of real-world problems. A basic introduction to the R programming language. May not be taken after S&DS 1080 or 1090. QR

S&DS 1080a, Introduction to Statistics: Advanced Fundamentals
Introductory statistical concepts beyond those covered in high school AP statistics. Includes additional concepts in regression, an introduction to multiple regression, and ANOVA. This course is intended as a bridge between AP statistics and courses such as S&DS 2300, 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 1000 or any other full semester Yale introductory statistics courses. Students should take S&DS 1000 rather than S&DS 1080, 1090. ½ Course cr

S&DS 1090a, Introduction to Statistics: Fundamentals
General concepts and methods in statistics. Covers material equivalent to high school AP statistics. Meets for the first half of the term only. May not be taken after or concurrently with S&DS 1000 or any other full semester Yale introductory statistics courses. ½ Course cr

S&DS 1230a or b / CPSC 1230a or b / PLSC 3508a or b / S&DS 5230a or b, YData: An Introduction to Data Science Staff
Computational, programming, and statistical skills are no longer optional in our increasingly data-driven world; these skills are essential for opening doors to manifold research and career opportunities. This course aims to dramatically enhance knowledge and capabilities in fundamental ideas and skills in data science, especially computational and programming skills along with inferential thinking. YData is an introduction to Data Science that emphasizes the development of these skills while providing opportunities for hands-on experience and practice. YData is accessible to students with little or no background in computing, programming, or statistics, but is also engaging for more technically oriented students through extensive use of examples and hands-on data analysis. Python 3, a popular and widely used computing language, is the language used in this course. The computing materials will be hosted on a special purpose web server. QR

* **S&DS 1720a / EP&E 328 / EP&E 4328a / PLSC 2509a, 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 1730b, YData: Analysis of Baseball Data** Ethan Meyers
The fields of data science aim 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 19th 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 include season-level statistics going back to the 1870's, play-by-play statistics going back to the 1930's 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 123. QR

* **S&DS 1790a, Data Science Applications in Insurance** Perry Beaumont
The insurance industry is becoming increasingly data driven. Data Science can be used to inform where new market opportunities are emerging, where risks are growing, how insurance policies can be more optimally structured, and ways claims can be more meaningfully managed. In exploration of these topics, flood insurance claims maintained by the Federal Emergency Management Agency (FEMA) and available in a data set with over 2.6 million rows and 40 data fields, serve as a north star in our analytic journey. We address issues that can arise when working with real-world data collection, along with related strategies for handling data that may be incomplete or simply messy. This course provides opportunities for students to extend the data science skills acquired in computationally-oriented introductory statistics courses, using new applications relevant to the financial services industry. The data science skills presented in a computationally oriented introductory course (S&DS 123, 220, or 230), and previous use of R are suggested. QR

S&DS 2200b, Introductory Statistics, Intensive Robert Wooster
Introduction to statistical reasoning 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 taught and used extensively, as well as application of statistical concepts to analysis of real-world data science problems. MATH 115 is helpful but not required. While no particular prior experience in computing is required, strong motivation to practice and learn computing are desirable. QR

* **S&DS 2240a, Dice, Data, and Decisions - The Statistics of Board Game Strategy** Robert Wooster
This course provides a hands-on application of data analysis, simulation, and probability theory to the world of board games and traditional games of chance. Class lessons will be a combination of lecture, computing labs, and actually learning and playing games! Topics include analyzing board game strategy using probability theory, probabilistic modeling using simulation in R, and exploration and analysis of both simulated and real game board game data. One of S&DS 100, 123, 220, or 230, and experience in the R statistical programming language. QR

S&DS 230a or b, Data Exploration and Analysis Jonathan Reuning-Scherer
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 Robert Wooster
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 240b, 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 Sinho Chewi
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 242b / MATH 242b, Theory of Statistics Zhou Fan
Study of the principles of statistical analysis. Topics include maximum likelihood, sampling distributions, estimation, confidence intervals, tests of significance, regression, analysis of variance, and the method of least squares. Some statistical computing. After S&DS 241 and concurrently with or after MATH 222 or 225, or equivalents. QR

S&DS 265b, 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 3120a, 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 3510b / EENG 434 / MATH 2510b, Stochastic Processes Ilias Zadik
Introduction to the study of random processes including linear prediction and Kalman filtering, Poison counting process and renewal processes, Markov chains, branching processes, birth-death processes, Markov random fields, martingales, and random walks. Applications chosen from communications, networking, image reconstruction, Bayesian statistics, finance, probabilistic analysis of algorithms, and genetics and evolution. Prerequisite: S&DS 241 or equivalent. QR

S&DS 3520b / MB&B 3520b / MCDB 3520b, Biomedical Data Science, Mining and Modeling Mark Gerstein and Matthew Simon
Techniques in data mining and simulation applied to bioinformatics, the computational analysis of gene sequences, macromolecular structures, and functional genomics data on a large scale. Sequence alignment, comparative genomics and phylogenetics, biological databases, geometric analysis of protein structure, molecular-dynamics simulation, biological networks, microarray normalization, and machine-learning approaches to data integration. Prerequisites: MB&B 301 and MATH 115, or permission of instructor. SC 0 Course cr

S&DS 3610b / AMTH 3610b, Data Analysis Brian Macdonald
Selected topics in statistics explored through analysis of data sets using the R statistical computing language. Topics include linear and nonlinear models, maximum likelihood, resampling methods, curve estimation, model selection, classification, and clustering. Extensive use of the R programming language. Experience with R programming (from e.g. S&DS 106, S&DS 220, S&DS 230, S&DS 242), probability and statistics (e.g. S&DS 106, S&DS 220, S&DS 238, S&DS 241, or concurrently with S&DS 242), linear algebra (e.g. MATH 222, MATH 225, MATH 118), and calculus is required. This course is a prerequisite for S&DS 425 and may not be taken after S&DS 425. QR

S&DS 3630b, Multivariate Statistics for Social Sciences Jonathan Reuning-Scherer
Introduction to the analysis of multivariate data as applied to examples from the social sciences. Topics include principal components analysis, factor analysis, cluster analysis (hierarchical clustering, k-means), discriminant analysis, multidimensional scaling, and structural equations modeling. Extensive computer work using either SAS or SPSS programming software. Prerequisites: knowledge of basic inferential procedures and experience with linear models. QR

S&DS 3640b / AMTH 3640b / EENG 454, Information Theory Yihong Wu
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 and finance. After STAT 241. QR

S&DS 3650a or b, Intermediate Machine Learning Staff
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 4000a / MATH 3300a, Advanced Probability Shuangping Li
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 4100a, 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. Prerequisite: level of S&DS 241.

* **S&DS 4250a 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 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). Enrollment limited; requires permission of the instructor. QR

* **S&DS 4800a or b, 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 4910a and S&DS 4920b, 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.