STATISTICS AND DATA SCIENCE (S&DS)

S&DS 100a 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. May not be taken after S&DS 101–106 or 109. QR

S&DS 123a or b / CPSC 123a or b / PLSC 351a or b / 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; 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 150a, Data Science Ethics** Elisa Celis
In this course, we introduce, discuss, and analyze ethical issues, algorithmic challenges, and policy decisions that arise when addressing real-world problems via the lens of data science. We grapple with the normative questions of what constitutes bias, fairness, discrimination, or ethics when it comes to data science and machine learning in applications such as policing, health, journalism, and employment. We incorporate technical precision by introducing quantitative measures that allow us to study how algorithms codify, exacerbate and/or introduce biases of their own, and study analytic methods of correcting for or eliminating these biases. Lastly, we study the social implications of these decisions, and understand the legal, political and policy decisions that could be used to govern data-driven decision making by making them transparent and auditable. We read critical commentary by practitioners, state-of-the-art technical papers by data scientist and computer scientists, and samples of legal scholarship, moral and ethical philosophy, readings in sociology, and policy documents. We often ground our discussions around recent case studies, controversies, and current events. Prerequisites: One from S&DS 238, S&DS 241, S&DS 242, or the equivalent; and one from S&DS 230, ECON 131, or the equivalent. Suggested courses: one from: CPSC 470, S&DS 365, ECON 429, CPSC 365, CPSC 366, or equivalent; and one from: EP&E 215, PHIL 175, PHIL 177, SOCY 144, PLSC 262, PLSC 320, or equivalent. SO

*** 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 173b, 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 220b, 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 230a 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. 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 240b, 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 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 Harrison Zhou
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 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 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 317b, Applied Machine Learning and Causal Inference** P Aronow
We cover approaches to causal inference using machine learning. Machine learning methods include bagging, boosting, random forests, and neural networks. Causal topics include randomized experiments with and without noncompliance, observational studies with and without ignorable treatment assignment, instrumental variables, and regression discontinuity. 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 230, 238, 241 and 242; previous programming experience (e.g., R, Matlab, Python, C++), R preferred. Strong knowledge of OLS is assumed. SO

S&DS 351b / EENG 434b / MATH 251b, 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 352b / MB&B 452b / MCDB 452b, 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

S&DS 361b / AMTH 361b, 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. 106, 220, 238, 241, or concurrently with 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 363b, 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 364b / AMTH 364b / EENG 454b, 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 and finance. After STAT 241. 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 Zhou Fan
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 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 431a / AMTH 431a / ECON 431a, Optimization and Computation Zhuoran Yang
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 432b, 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 accepted from statistics & data science, economics, engineering, and the sciences. Prerequisites: Knowledge of linear algebra, such as MATH 222, 225; multivariate calculus, such as MATH 120; probability, such as S&DS 241/541; optimization, such as S&DS 431/631; and, comfort with proof-based exposition and problem sets.

* **S&DS 480a 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 491a and S&DS 492b, 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.