36-200 Reasoning with Data
All Semesters: 9 units
This course builds on the principles and methods of statistical reasoning that were developed in a first-semester intro statistics course, and will cover regression analysis (simple and multiple), logistic regression, one-way analysis of variance, and some elementary machine learning topics. The course will revisit in more detail the methods for examining the relationship between two variables and will also expand the methods to cases where there is more than one explanatory variable. The course includes an introduction to the R studio statistical programming environment (through labs and homeworks). Skills in course methodology, concepts, and data scientific expression will be synthesized in two written data analysis projects. Students produce and engage with an array of representations in R of descriptions, displays, models, and predictions, of real data. In each weekly lab as well as on all homeworks and projects, data is explored, described, and modeled using R. Coding skills relevant to the course are developed from the ground up. In addition to relevant statistical and machine learning applications, students practice some elementary data cleaning, learn and experience the importance of reproducibility, and are introduced to the importance of commenting code. Issues of model selection, and the danger and issues of overfitting, are covered in a variety of contexts. Issues involving algorithmic bias and data ethics are also discussed.

36-201 Statistical Reasoning and Practice
Fall and Spring: 9 units
This course will introduce students to the basic concepts, logic, and issues involved in statistical reasoning, as well as basic statistical methods used to analyze data and evaluate studies. The major topics to be covered include methods for exploratory data analysis, an introduction to research methods, elementary probability, and methods for statistical inference. The objectives of this course are to help students develop a critical approach to the evaluation of study designs, data and results, and to develop skills in the application of basic statistical methods in empirical research. An important feature of the course will be the use of the computer to facilitate the understanding of important statistical ideas and for the implementation of data analysis. In addition to three lectures a week, students will attend a computer lab once a week. All students will receive credit for 36-201/36-220, 36-225, 36-625, or 36-247.
Course Website: http://www.stat.cmu.edu/academics/courselist/(http://www.stat.cmu.edu/academics/courselist/)

36-202 Methods for Statistics & Data Science
All Semesters: 9 units
This course builds on the principles and methods of statistical reasoning developed in 36-200 (or its equivalents). The course covers simple and multiple regression, basic analysis of variance methods, logistic regression, and introduction to data mining including classification and clustering. Students will also learn the principles of overfitting, training vs testing, ensemble methods, variable selection, and bootstrapping. Course objectives include applying the basic principles and methods that underlie statistical practice and empirical research to real data sets and interdisciplinary problems. Learning the Data Analysis Pipeline is strongly emphasized through structured coding and data analysis projects. In addition to three lectures a week, students attend a computer lab once a week for "hands-on" practice of the material covered in lecture. There is no programming language pre-requisite. Students will learn the basics of R Markdown and related analytics tools. Not open to students who have received credit for 36-201/36-220, 36-208/70-208. Students who have completed or are enrolled in 36-401 prior to completing 36-202, are not able to take/receive credit for 36-202. Prerequisites: 36-207 or 70-207 or 36-247 or 36-220 or 36-200

36-204 Discovering the Data Universe
Intermittent: 3 units
Every day we wake up in the data universe, we use the information around us to make decisions. We are constantly evaluating and interpreting data from our environment, in everything from spreadsheets to Instagram posts. At the same time, our own personal data are being observed and recorded—through websites we visit online, our smart devices, and even our interactions with other students and faculty at CMU. Navigating this data universe requires knowledge of what data is and how to use it responsibly. For example, can a plant be a data set? Discovering the truth behind a piece of data, including who made it, what it looks like, and what we can learn from it, is a critical skill. Understanding data can be the difference between being able to distinguish truth from lies; and the key to identifying your data footprint and succeeding in research and in your career. In this course, we will explore the data universe from multiple angles and across several types of data. We will define, find, and analyze data, and most importantly, identify narratives within data to tell stories about the world around us. We will examine data using the following questions: How can we tell multiple stories from the same dataset? What biases can exist in data? And, who creates or decides what data matters enough to collect, preserve, and share? NOTE: There will be one in person and one virtual pre-recorded lecture each week.

36-207 Probability and Statistics for Business Applications
Fall and Summer: 9 units
This is the second half of a year long sequence in basic statistical methods that are used in business and management. Topics include exploratory and descriptive techniques, probability theory, statistical inference in simple settings, basic categorical analysis, and statistical methods for quality control. Not open to students who have received credit for 36-201, 36-220, 36-625, or 36-247. Cross-listed as 70-207. Prerequisites: 21-121 or 21-112 or 21-120

36-208 Regression Analysis
Fall and Summer: 9 units
This is the second half of a year long sequence in basic statistical methods that are used in business and management. Topics include exploratory and descriptive techniques, probability theory, statistical inference in simple settings, basic categorical analysis, and statistical methods for quality control. Not open to students who have received credit for 36-201, 36-220, 36-625, or 36-247. Cross-listed as 70-208. Students who have completed 36-401 prior to 36-208 will not receive credit for 36-208. Prerequisites: (21-120 or 21-112) and (70-207 or 36-247 or 36-220 or 36-201 or 36-207) and (73-102 or 73-100)
Course Website: http://www.stat.cmu.edu/academics/courselist(http://www.stat.cmu.edu/academics/courselist/)

36-217 Probability Theory and Random Processes
All Semesters: 9 units
This course provides an introduction to probability theory. It is designed for students in electrical and computer engineering. Topics include elementary probability theory, conditional probability and independence, random variables, distribution functions, joint and conditional distributions, limit theorems, and an introduction to random processes. Some elementary ideas in spectral analysis and information theory will be given. A grade of C or better is required in order to use this course as a pre-requisite for 36-226 and 36-410. Not open to students who have received credit for 36-225, or 36-625. Prerequisites: 21-259 or 21-256 or 21-122 or 21-123 or 21-112
Course Website: http://www.stat.cmu.edu/academics/courselist(http://www.stat.cmu.edu/academics/courselist/)
36-218 Probability Theory for Computer Scientists
Fall and Spring: 9 units
Probability theory is the mathematical foundation for the study of both statistics and of random systems. This course is an intensive introduction to probability from the foundations and mechanics to its application in statistical models and modeling of random processes. Special topics and many examples are drawn from areas and problems that are of interest to computer scientists and that should prepare computer science students for the probabilistic and statistical ideas they encounter in downstream courses and research. A grade of C or better is required in order to use this course as a pre-requisite for 36-226, 36-326, and 36-410. Not open to students who have received credit for 36-225, 21-325, or 36-700. If you hold a Statistics primary/additional major or minor you will be required to complete 36-226. For those who do not have a major or minor in Statistics, and receive at least a B in 36-218, you will be eligible to move directly onto 36-401. Prerequisites: (21-112 and 21-111) or 21-120 or 21-256 or 21-259
Course Website: http://www.stat.cmu.edu/academics/courselist (http://www.stat.cmu.edu/academics/courselist/)

36-219 Probability Theory and Random Processes
All Semesters: 9 units
This course provides an introduction to probability theory. It is designed for students in electrical and computer engineering. Topics include elementary probability theory, conditional probability and independence, random variables, distribution functions, joint and conditional distributions, limit theorems, and an introduction to random processes. Some elementary ideas in spectral analysis and information theory will be given. A grade of C or better is required in order to use this course as a pre-requisite for 36-226 and 36-410. Not open to students who have received credit for 36-225, or 36-625. Prerequisites: (21-111 and 21-112) or 21-120 or 21-256 or 21-259

36-220 Engineering Statistics and Quality Control
Fall and Spring: 9 units
This is a course in introductory statistics for engineers with emphasis on modern product improvement techniques. Besides exploratory data analysis, basic probability, distribution theory and statistical inference, special topics include experimental design, regression, control charts and acceptance sampling. Not open to students who have received credit for 36-200, 36-207/70-207, 36-226, 36-626, or 36-247, except when AP credit is awarded for 36-200. Prerequisites: 21-120 or 21-112

36-225 Introduction to Probability Theory
Fall and Summer: 9 units
This course is the first half of a year-long course which provides an introduction to probability and mathematical statistics for students in the data sciences. Topics include elementary probability theory, conditional probability and independence, random variables, distribution functions, joint and conditional distributions, law of large numbers, and the central limit theorem. A grade of C or better is required in order to advance to 36-226, 36-326, and 36-410. Not open to students who have received credit for 36-217, 36-218, 36-219, 21-325, or 36-700. Prerequisites: (21-211 and 21-111) or 21-120 or 21-256 or 21-259
Course Website: http://coursescatalog.web.cmu.edu/schools-colleges/dietrichcollegeofhumanitiesandsocialsciences/department/ste/depn (http://coursescatalog.web.cmu.edu/schools-colleges/dietrichcollegeofhumanitiesandsocialsciences/department/ste/depn/)

36-226 Introduction to Statistical Inference
Spring and Summer: 9 units
This course is the second half of a year-long course in probability and mathematical statistics. Topics include maximum likelihood estimation, confidence intervals, hypothesis testing, and properties of estimators, such as unbiasedness and consistency. If time permits there will also be a discussion of linear regression and the analysis of variance. A grade of C or better is required in order to advance to 36-401, 36-402 or any 36-46x course. Not open to students who have received credit for 36-626. Prerequisites: 36-217 Min. grade C or 15-259 Min. grade C or 36-218 Min. grade C or 36-219 Min. grade C or 21-325 Min. grade C or 36-225 Min. grade C

36-235 Probability and Statistical Inference I
Fall: 9 units
This class is the first half of a two-semester, calculus-based course sequence that introduces theoretical aspects of probability and statistical inference to students. The material in this course and in 36-236 (Probability and Statistical Inference II) is organized so as to provide repeated exposure to essential concepts: the courses cover specific probability distributions and their inferential applications one after another, starting with the normal distribution and continuing with the binomial and Poisson distributions, etc. Topics specifically covered in 36-235 include basic probability, random variables, univariate and multivariate distribution functions, point and interval estimation, hypothesis testing, and regression, with the discussion being supplemented with computer-based examples and exercises (e.g., visualization and simulation). Given its organization, the course is only appropriate for those taking the full two-semester sequence, and thus it is currently open only to statistics majors (primary, additional, dual) and minors. (Check with the statistics advisors for the exact declaration deadline.) Non-majors/minors requiring a probability course are directed to take 36-225 or one of its analogues. A grade of C or better in 36-235 is required in order to advance to 36-236 (or 36-226) and/or 36-410. This course is not open to students who have received credit for 36-217, 36-218, 36-219, or 36-700, or for 21-325 or 15-259. Prerequisites: (21-111 and 21-112) or 21-256 or 21-259 or 21-120

36-236 Probability and Statistical Inference II
Spring: 9 units
This class is the second half of a two-semester, calculus-based course sequence that introduces theoretical aspects of probability and statistical inference to students. The material in this course and in 36-235 (Probability and Statistical Inference I) is organized so as to provide repeated exposure to essential concepts: the courses cover specific probability distributions and their inferential applications one after another, starting with the normal distribution and continuing with the binomial and Poisson distributions, etc. Topics specifically covered in 36-236 include the binomial and related distributions, the Poisson and related distributions, and the uniform distribution, and how they are used in point and interval estimation, hypothesis testing, and regression. Also covered in 36-236 are topics related to multivariate distributions: marginal and conditional distributions, covariance, and conditional distribution moments. All discussion is supplemented with computer-based examples and exercises (e.g., visualization and simulation). Given its organization, the course is only appropriate for those who first take 36-235, and thus it is currently open only to statistics majors (primary, additional, dual) and minors, as well as to CS majors using both 36-235 and 36-236 to complete their probability requirement. All others are directed to take 36-226. A grade of C or better in 36-236 is required in order to advance to 36-401. Prerequisite: 36-235 Min. grade C

36-247 Statistics for Lab Sciences
Fall and Spring: 9 units
This course is equivalent to 36-200 with the exception that 36-247 students take a Lab Sciences-specific laboratory section. This course is an introduction to learning how to make statistical decisions and “reason with data”. The approach will emphasize thinking through an empirical problem from beginning to end and using statistical tools to look for evidence for/against an explicit argument/hypothesis. Types of data will include continuous and categorical variables, images, text, networks, and repeated measures over time. Applications will largely draw from interdisciplinary case studies spanning the humanities, social sciences, and related fields. Methodological topics will include basic exploratory data analysis, elementary probability, hypothesis tests, and empirical research methods. There is no calculus or programming requirement. There will be one weekly computer lab for additional hands-on practice using an interactive software platform that allows student-driven inquiry. This course is the credit-equivalent to 36-201 and will be honored appropriately as a pre-requisite for downstream Statistics courses. As such, this course is not currently open to students who have received credit for 36-200, 36-201, 36-70-207, 36-220, or any 300- or 400-level Statistics course.
36-290 Introduction to Statistical Research Methodology  
Fall: 9 units  
This is a first course in statistical practice, targeted to first-semester sophomores. It is designed as a high-level introduction to the ways in which statisticians go about approaching and analyzing quantitative observational data, thus preparing students for future work in capstone classes. Students in the course are taught the basic concepts of statistical learning—inferring vs prediction, supervised vs. unsupervised learning, regression vs. classification, etc.—and will reinforce this knowledge by applying, e.g., linear regression, random forest, principal components analysis, and/or hierarchical clustering and more to datasets provided by the instructor. Students will also practice disseminating the results of their analyses via oral presentations and posters. Analyses will be carried out using the R programming language.  
Prerequisites: 36-247 or 36-220 or 36-207 or 70-207 or 36-200  
Course Website: http://coursescatalog.web.cmu.edu/schools-colleges/dietrichcollegeofhumanitiesandsocialsciences/depar (http://coursescatalog.web.cmu.edu/schools-colleges/dietrichcollegeofhumanitiesandsocialsciences/depar)  

36-300 Statistics & Data Science Internship  
Summer: 3 units  
The Department of Statistics and Data Science considers experiential learning as an integral part of our program. One such option is through an internship. If a student has an internship, they don't have to register for this class unless they want it listed on their official transcripts. This process should be used by international students interested in Curricular Practical Training (CPT) and should also be authorized by the Office of International Education (OIE). More information regarding CPT is available on OIE’s website. This course will be taken as Pass/Fail, and students will be charged tuition for 3 units. There is an approval process in order to register for this course. Please contact your advisor the Department of Statistics and Data Science for more details.  

36-301 Documenting Human Rights  
Intermittent: 9 units  
This course will teach students about the origins of modern human rights and how they are documented. In the process of documenting the extent to which these rights are being upheld or violated. The need to understand and document human rights issues is at the center of the most pressing current events. From threats to democracy and civil rights at home and abroad to work holding perpetrators of mass harm accountable in legal proceedings to efforts to quantify and advance economic, social, cultural, and environmental rights, making human rights violations visible is fundamental to achieving a more just world. We will begin with an overview of the history of human rights, the main philosophical and political debates in the field, and the most relevant organizations, institutions, and agreements. We will then delve into several specific cases that highlight methodological opportunities and challenges in the context of human rights, including: the identification of mass atrocity victims, the disappeared, and missing migrants; efforts to estimate civilian casualties in war in real-time; the documentation of police brutality and other human rights violations (especially against minority groups) with smartphones and the analysis of human rights media; as well as the use of statistical imaging, drones, and GIS data to detect and record genocide, environmental rights violations, and war crimes. We will critically assess the documentation and measurement challenges that arise in each context and how the human rights and scientific communities have responded. After reviewing these cases, we will conclude by reflection on why the documentation of human rights actually matters and what happens to evidence once it is gathered. Students will then take what they've learned and do a multidisciplinary group project in which they document a rights violation in Western Pennsylvania.  

36-303 Sampling, Survey and Society  
Spring: 9 units  
This course will revolve around the role of sampling and sample surveys in the context of U.S. society and its institutions. We will examine the evolution of survey taking in the United States in the context of its economic, social, and political uses. This will eventually lead to discussions about the accuracy and relevance of survey responses, especially in light of various kinds of nonsampling error. Students will be required to design, implement and analyze a survey sample.  
Prerequisites: 36-309 or 36-202 or 36-226 or 36-326 or 70-208 or 36-208 or 36-220  

36-304 Biostatistics  
Fall: 9 units  
TBD  

36-309 Experimental Design for Behavioral & Social Sciences  
Fall and Summer: 9 units  
Statistical aspects of the design and analysis of planned experiments are studied in this course. A clear statement of the experimental factors will be emphasized. The design aspect will concentrate on choice of models, sample size and order of experimentation. The analysis phase will cover data collection and computation, especially analysis of variance and will stress the interpretation of results. In addition to a weekly lecture, students will attend a computer lab once a week.  
Prerequisites: 36-226 or 36-220 or 36-218 or 15-260 or 36-200 or 36-247 or 36-220  
Course Website: http://www.stat.cmu.edu/academics/counsellist (http://www.stat.cmu.edu/academics/counsellist)  

36-311 Statistical Analysis of Networks  
Intermittent: 9 units  
Networks are omnipresent. In this course, students will get an introduction to network science, mainly focusing on social network analysis. The course will start with some empirical background, and an overview of concepts used when measuring and describing networks. We will also discuss network visualization. Most traditional models cannot be applied straightforwardly to social network data, because of their complex dependence structure. We will discuss random graph models and statistical network models, that have been developed for the study of network structure and growth. We will also cover models of how networks impact individual behavior.  
Prerequisite: 36-226  

36-312 Statistics of Inequality and Discrimination  
Intermittent: 9 units  
Many social questions about inequality, injustice and unfairness are, in part, questions about evidence, data, and statistics. This class lays out the statistical methods which let us answer questions like “Does this employer discriminate against members of that group?” “Is this standardized test biased against that group?” “Is this decision-making algorithm biased and does that even mean?” and “Did this policy which was supposed to reduce this inequality actually help?” We will also look at inequality within groups, and at different ideas about how to explain inequalities between groups. The class will interweave discussion of concrete social issues with the relevant statistical concepts.  
Prerequisite: 36-202  

36-313 Biostatistics  
Fall: 9 units  
This course is an introduction to methods used frequently in biostatistics and public health applications.  
Prerequisites: 36-226 or 36-208 or 88-250 or 36-225 or 36-625 or 70-208 or 36-303 or 36-202 or 36-309  

36-315 Statistical Graphics and Visualization  
All Semesters: 9 units  
Graphical displays of quantitative information take on many forms, and they help us understand data and statistical methods by (hopefully) clearly communicating arguments, results, and ideas. This course introduces students to the most common forms of graphical displays and their uses and misuses. Ideally, graphs are designed according to three key elements: The data structure, the graph’s audience, and the designer’s intended message. Students will learn how to create well-designed graphs and understand them from a statistical perspective. Furthermore, the course will consider complex data structures that are becoming increasingly common in data visualizations (temporal, spatial, and text data); we will discuss common ways to process these data that make them easy to visualize. As time permits, we may also consider more advanced graphical methods (e.g., interactive graphics and computer-generated animations). In addition to two weekly lectures, there will be weekly computer labs and homework assignments where students use R to visualize and analyze real datasets. Along the way, students also make monthly Piazza posts discussing the strengths and weaknesses of a graph they found online, thereby critiquing real graphical designs found in the wild. The course culminates in a group final project, where students make public-facing data visualizations and analyses for a real dataset. All assignments will be in R; although this is not a programming class, using programming-based statistical software like R is essential to create modern-day graphics, and this class will give you practice using this kind of software. Throughout, communication skills (usually written or visual, but sometimes spoken) will play an important role. Indeed, if it’s true that “a picture speaks a thousand words,” then ideally the one thousand words you are communicating with your graphics are statistically correct, clear, and compelling.  
Prerequisites: 15-259 or 36-218 or 36-225 or 70-208 or 21-325 or 36-219 or 36-208 or 36-202 or 36-309
36-318 Introduction to Causal Inference
Intermittent: 9 units
Many social science and scientific inquiries can be framed as causal questions. Does a new cancer treatment cause a reduction in mortality? Do financial grants cause students to do better in college? Does a new public policy cause an increase in voter turnout? When tackling these questions, we frequently come across the phrase “correlation does not imply causation.” If that’s the case, then what does imply causation? In this course, we will discuss causal inference methods for measuring causal effects of different interventions (e.g., drug treatments, financial grants, and public policies). First, we will discuss how experiments—where interventions are randomized among subjects—can imply causation when an appropriate experimental design and statistical analysis is used. Then, we will discuss how observational studies—where interventions are not randomized—can also imply causation when approaches like propensity score methods, matching, and doubly robust estimation are employed. Finally, we will discuss instrumental variables and regression discontinuity designs—which are frequently used in medicine and public policy for establishing causal inferences—as well as the nuances of conducting causal inferences when there are multi-valued and time-varying interventions. Throughout we will use R to conduct causal analyses. A working knowledge of regression is encouraged, but regression will also be discussed and taught during much of the course.
Prerequisites: 36-225 Min. grade C or 21-325 Min. grade C or 36-218 Min. grade C or 15-259 Min. grade C or 36-219 Min. grade C

36-326 Mathematical Statistics (Honors)
Spring: 9 units
This course is a rigorous introduction to the mathematical theory of statistics. A good working knowledge of calculus and probability theory is required. Topics include maximum likelihood estimation, confidence intervals, hypothesis testing, Bayesian methods, and regression. A grade of C or better is required in order to advance to 36-401, 36-402 or any 36-46x course. Not open to students who have received credit for 36-625.
Prerequisites: 15-359 or 21-325 or 36-217 or 36-225 with a grade of A AND advisor approval. Students interested in the course should add themselves to the waitlist pending review.
Prerequisites: 36-218 Min. grade A or 36-217 Min. grade A or 36-225 Min. grade A or 15-359 Min. grade A or 21-325 Min. grade A

36-350 Statistical Computing
All Semesters: 9 units
Statistical Computing is a one-semester course that will introduce you to the fundamentals of computational data analysis, as carried out in the R programming language, and to the fundamentals of working with relational databases, as illustrated with PostgreSQL. No previous knowledge of either is required.
Prerequisites: 15-259 Min. grade C or 36-225 Min. grade C or 36-217 Min. grade C or 36-218 Min. grade C or 36-219 Min. grade C or 21-325 Min. grade C

36-375 Data Ethics & Responsible Conduct of Research
Intermittent: 3 units
TBD

36-401 Modern Regression
Fall: 9 units
This course is an introduction to the real world of statistics and data analysis. We will explore real data sets, examine various models for the data, assess the validity of their assumptions, and determine which conclusions we can make (if any). Data analysis is a bit of an art; there may be several valid approaches. We will strongly emphasize the importance of critical thinking about the data and the question of interest. Our overall goal is to use a basic set of modeling tools to explore and analyze data and to present the results in a scientific report. A grade of C is required to move on to 36-402 or any 36-46x course.
Prerequisites: (36-218 Min. grade B or 36-326 Min. grade C or 36-226 Min. grade C) and (21-242 or 21-240 or 21-241)

36-402 Advanced Methods for Data Analysis
Spring: 9 units
This course introduces modern methods of data analysis, building on the theory and application of linear models from 36-401. Topics include nonlinear regression, nonparametric smoothing, density estimation, generalized linear and generalized additive models, simulation and predictive model-checking, cross-validation, bootstrap uncertainty estimation, multivariate methods including factor analysis and mixture models, and graphical models and causal inference. Students will analyze real-world data from a range of fields, coding small programs and writing reports.
Prerequisite: 36-401 Min. grade C

36-410 Introduction to Probability Modeling
Spring: 9 units
An introductory-level course in stochastic processes. Topics typically include Poisson processes, Markov chains, birth and death processes, random walks, recurrent events, and renewal theory. Examples are drawn from reliability theory, queuing theory, inventory theory, and various applications in the social and physical sciences.
Prerequisites: 36-625 or 36-225 or 36-217 or 21-325

36-424 Natural Science Applications of Stochastic Processes
Intermittent: 6 units
Much of the natural world is governed by the evolution of temporally and spatially dependent systems; the study of stochastic processes provides statisticians tools to predict, succinctly explain, and quantify the uncertainty in these systems. This is a course in stochastic processes at the upper undergraduate and masters level. Basic theory will be covered, but the course focuses primarily on simulation and real-world, mostly temporally-dependent data applications. Topics include finite-state Markov chains, branching processes, temporal and spatial Poisson processes. Case studies will be drawn from several fields, including neuroscience, epidemiology, seismology, molecular biology, and computer science. Familiarity with R is assumed.

36-428 Time Series
Spring: 6 units
The course is designed for graduate students and advanced undergraduate students. It will introduce the analysis and some of the theory of sequences of serially-dependent random variables (known as time series). Students should already have learned mathematical probability and statistics, including multivariate and conditional distributions, linear regression, calculus, matrix algebra, and the fundamentals of complex variables and functions. The focus will be on popular models for time series and the analysis of data that arise in applications.
Prerequisite: 36-401 Min. grade C

36-431 Foundations of Causal Inference
Intermittent: 6 units
This course will provide an introduction to the fundamentals of causal inference. Causal inference is concerned with whether and how one can go beyond statistical associations to draw causal conclusions from observational data. Topics will include: counterfactuals (potential outcomes and graphs), identification and estimation of average treatment effects in experiments and observational studies, nonparametric bounds, sensitivity analysis, instrumental variables, effect modification, and longitudinal studies. Special permission is required for undergraduate students.

36-432 Modern Causal Inference
Intermittent: 6 units
This course will provide an in-depth look at modern causal inference. Topics will include: optimal treatment regimes, mediation, principal stratification, stochastic interventions, accounting for complex confounding and exposures, and methods for efficient nonparametric estimation. Some background in mathematical statistics is advised. Special permission is required for undergraduate students.
36-459 Statistical Models of the Brain
Spring: 12 units
This new course is intended for CNBC students, as an additional option for fulfilling the computational core course requirement, but it will also be open to Statistics and Machine Learning students. It should be of interest to anyone wishing to see the way statistical ideas play out within the brain sciences, and it will provide a series of case studies on the role of stochastic models in scientific investigation. Statistical ideas have been part of neurophysiology and the brain sciences since the first stochastic description of spike trains, and the quantal hypothesis of neurotransmitter release, more than 50 years ago. Many contemporary theories of neural system behavior are built with statistical models. For example, integrate-and-fire neurons are usually assumed to be driven in part by stochastic noise; the role of spike timing involves the distinction between Poisson and non-Poisson neurons; and oscillations are characterized by decomposing variation into frequency-based components. In the visual system, V1 simple cells are often described using linear-nonlinear Poisson models; in the motor system, neural response may involve direction tuning; and CA1 hippocampal receptive field plasticity has been characterized using dynamic place models. It has also been proposed that perceptions, decisions, and actions result from optimal (Bayesian) combination of sensory input with previously-learned regularities; and some investigators report new insights from viewing whole-brain pattern responses as analogous to statistical classifiers. Throughout the field of statistics, models incorporating random "noise" components are used as an effective vehicle for data analysis. In neuroscience, however, the models also help form a conceptual framework for understanding neural function. This course will examine some of the most important methods and claims that have come from applying statistical thinking.
Prerequisite: 36-401 Min. grade C

36-461 Special Topics: Statistical Methods in Epidemiology
Intermittent: 9 units
Epidemiology is concerned with understanding factors that cause, prevent, and reduce diseases by studying associations between disease outcomes and their suspected determinants in human populations. Epidemiologic research requires an understanding of statistical methods and design. Epidemiologic data is typically discrete, i.e., data that arise whenever counts are made instead of measurements. In this course, methods for the analysis of categorical data are discussed with the purpose of learning how to apply them to data. The central statistical themes are building models, assessing fit and interpreting results. There is a special emphasis on generating and evaluating evidence from observational studies. Case studies and examples will be primarily from the public health sciences.
Prerequisite: 36-401 Min. grade C
Course Website: http://coursescatalog.web.cmu.edu/schools-colleges/dietrichcollegeofhumanitiesandsocialsciences/depar (http://coursescatalog.web.cmu.edu/schools-colleges/dietrichcollegeofhumanitiesandsocialsciences/depar/)

36-462 Special Topics: Methods of Statistical Learning
Intermittent: 9 units
Data mining is the science of discovering patterns and learning structure in large data sets. Covered topics include information retrieval, clustering, dimension reduction, regression, classification, and decision trees.
Prerequisite: 36-401 Min. grade C
Course Website: http://www.stat.cmu.edu/academics/cour selist (http://www.stat.cmu.edu/academics/cour selist/)

36-463 Special Topics: Multilevel and Hierarchical Models
Intermittent: 9 units
Multilevel and hierarchical models are among the most broadly applied "sophisticated" statistical models, especially in the social and biological sciences. They apply to situations in which the data "cluster" naturally into groups of units that are more related to each other than they are the rest of the data. In the first part of the course we will review linear and generalized linear models. In the second part we will see how to generalize these to multilevel and hierarchical models and relate them to other areas of statistics, and in the third part of the course we will learn how Bayesian statistical methods can help us to build, estimate and diagnose problems with these models using a variety of data sets and examples.
Prerequisite: 36-401 Min. grade C
Course Website: http://www.stat.cmu.edu/academics/cour selist (http://www.stat.cmu.edu/academics/cour selist/)

36-464 Special Topics: Applied Multivariate Methods
Intermittent: 9 units
This course is an introduction to applied multivariate methods. Topics include a discussion of the multivariate normal distribution, the multivariate linear model, repeated measures designs and analysis, principle component and factor analysis. Emphasis is on the application and interpretation of these methods in practice. Students will use at least one statistical package.
Prerequisite: 36-401 Min. grade C
Course Website: http://www.stat.cmu.edu/academics/cour selist (http://www.stat.cmu.edu/academics/cour selist/)

36-465 Special Topics: Conceptual Foundations of Statistical Learning
Intermittent: 9 units
This class is an introduction to the foundations of statistical learning theory, and its uses in designing and analyzing machine-learning systems. Statistical learning theory studies how to fit predictive models to training data, usually by solving an optimization problem, in such a way that the model will predict well, on average, on new data. The course will focus on the key concepts and theoretical tools, at a level accessible to students who have taken 36-401 and its pre-requisites. The course will also illustrate those concepts and tools by applying them to carefully selected kinds of machine learning systems (such as kernel machines). Students wanting exposure to a broad range of algorithms and applications would be better served by 36-462/662 ("Data Mining"). This class is for those who want a deeper understanding of the principles underlying all machine learning methods.
Prerequisite: 36-401 Min. grade C

36-466 Special Topics: Statistical Methods in Finance
Intermittent: 9 units
Financial econometrics is the interdisciplinary area where we use statistical methods and economic theory to address a wide variety of quantitative problems in finance. These include building financial models, testing financial economics theory, simulating financial systems, volatility estimation, risk management, capital asset pricing, derivative pricing, portfolio allocation, proprietary trading, portfolio and derivative hedging, and so on and so forth. Financial econometrics is an active field of integration of finance, economics, probability, statistics, and applied mathematics. Financial activities generate many new problems and products, economics provides useful theoretical foundation and guidance, and quantitative methods such as statistics, probability and applied mathematics are essential tools to solve quantitative problems in finance. Professionals in finance now routinely use sophisticated statistical techniques and modern computation power in portfolio management, proprietary trading, derivative pricing, financial consulting, securities regulation, and risk management.
Prerequisite: 36-401

36-467 Special Topics: Data over Space & Time
Intermittent: 9 units
This course is an introduction to the opportunities and challenges of analyzing data from processes unfolding over space and time. It will cover basic descriptive statistics for spatial and temporal patterns; linear methods for interpolating, extrapolating, and smoothing spatio-temporal data; basic nonlinear modeling; and statistical inference with dependent observations. Class work will combine practical exercises in R, a little mathematics on the underlying theory, and case studies analyzing real problems from various fields (economics, history, meteorology, ecology, etc.). Depending on available time and class interest, additional topics may include: statistics of Markov and hidden-Markov (state-space) models; statistics of point processes; simulation and simulation-based inference; agent-based modeling; dynamical systems theory.
Prerequisite: 36-401 Min. grade C
Course Website: http://coursescatalog.web.cmu.edu/schools-colleges/dietrichcollegeofhumanitiesandsocialsciences/depar (http://coursescatalog.web.cmu.edu/schools-colleges/dietrichcollegeofhumanitiesandsocialsciences/depar/)
36-468 Special Topics: Text Analysis
Intermittent: 9 units
The analysis of language is concerned with how variables relate to people (their gender, age, and location, for example), how variables relate to use (such as writing in different academic disciplines), and how variables change over time. While we are surrounded by data that might potentially shed light on many of these questions, working with real-world linguistic data can present some unique challenges in sampling, in the distribution of features, and, in their high dimensionality. In this course, we work through some of these issues, paying particular attention to the aligning of the statistical questions we want to investigate with the choice of statistical models, as well as focusing on the interpretation of results. Analysis will be carried out in R and students will develop a suite of tools as they work through their course projects.

36-469 Special Topics: Statistical Genomics and High Dimensional Inference
Intermittent: 9 units
The field of computational and statistical genomics focuses on developing and applying computationally efficient and statistically robust methods to sort through increasingly rich and massive genome wide data sets to identify complex genetic patterns, gene interactions, and disease associations. Because the genome is vast, analytical approaches require high dimensional statistical approaches such as multiple testing, dimension reduction techniques, regularization and high dimensional regression analysis, best linear unbiased prediction models, networks and graphical models. In this course, we will motivate these topics using data obtained from the human genetic and genomic literature. No prior knowledge in biology is required. Prerequisite: 36-401 Min. grade C

36-490 Undergraduate Research
Fall and Spring: 9 units
This course is designed to give undergraduate students experience using statistics in real research problems. Small groups of students are matched with clients and do supervised research for a semester. From an academic perspective, the course presents an opportunity for students to gain skills in approaching a research problem, critical thinking, and statistical analyses. Additionally, the course will help students develop the professional skills necessary for successfully navigating team-based project delivery roles. Client-facing and collaborative skills will be emphasized within a team setting, and students will learn leading practices for engaging stakeholders as well as gain a conceptual understanding of leading practices for project delivery.

36-492 Topic Detection and Document Clustering
Intermittent: 6 units
Imagine if someone read all your email. Everything you sent, everything you received. What would they find? Do you have repeating topics? How do the topics change over time? The Enron Corporation was an energy, commodities, and services company in Houston, Texas that went spectacularly bankrupt in 2001, after it was revealed that it was engaging in systematic, planned accounting fraud. At its peak, it owed over 20,000 people with revenues over $100 billion. Its downfall was related to deregulation of California's energy commodity trading and a series of rolling power blackouts over months. For example, Enron traders encouraged the removal of power during the energy crisis by suggesting plant shutdowns. The resulting increase in the price for power made them a fortune. After Enron's collapse, journalists used the Freedom of Information Act to release the emails sent/received by the employees of Enron. Subsequently, the emails were analyzed to see who knew what and when. Every news article, email, letter, blog, tweet, etc can be thought of as an observation. We characterize these documents by their length, what words they use and how often, and possibly extra information like the time, the recipient, etc. Topic detection and document clustering methods are statistical and machine learning tools that extract and identify related documents, possibly over time. These methods need to be flexible enough to handle both very small and very large clusters of documents, topics that change in importance, and topics that appear and disappear. This class will emphasize application of methods and real-world data analysis. Class time will be split into lecture and "lab". (Bring your laptop.) Occasional homeworks and final project, but mostly we'll focus on the downfall of Enron as our overarching case study. Prerequisite: 36-401

36-493 Sports Analytics Capstone
Intermittent: 9 units
This course is designed to give undergraduate students experience applying statistics and data science methodology to real industry projects. Small groups of students will be matched with clients in the Carnegie Mellon Athletics Department and do supervised projects for a semester. Students will gain skills in approaching a real world problem, critical thinking, advanced statistical analysis, scientific writing, collaboration with clients, communicating results, and meeting expectations with respect to deliverables and timelines. The projects will change and rotate each semester. The course size is limited, and students will submit an application including their project preferences. Students with skill sets matching project needs will be given priority. We will also take into consideration whether or not a student has had a recent prior data science experience with the goal of providing experiences to a broad group of qualified students. Students do not need to be experts in sports analytics or have extensive knowledge in sports.

36-494 Astrostatistics
Intermittent: 6 units
Since a young age, many of us have pondered the vastness and beauty of the Universe as we gazed up at the night sky. Planets, moons, stars, galaxies, and beyond have fascinated humanity for centuries. It turns out it also provides a plethora of interesting and complex statistical problems. In this course, problems in astronomy, cosmology, and astrophysics are going provide motivation for learning about some advanced statistical methodology. Possible topics include computational statistics, topological data analysis, nonparametric regression, spatial statistics, and statistical learning. While exploring newer statistical methodology, we will get to sample a variety of problems that appeal to astrostatisticians Statistical problems related to exoplanets (planets orbiting stars outside our Solar System), the large-scale structure of the Universe (the "Cosmic Web"), dark matter (over 80% of the matter in the Universe is thought to be invisible), Type Ia supernova (a dying star eats its companion star until explodes), cosmic microwave background (a.k.a. "baby pictures of the Universe") are some possibilities. This course will be suitable for advanced undergraduate statistics majors through Ph.D. level statistics students, and astronomy Ph.D. students with some background in statistics. Prerequisite: 36-401 Min. grade C

36-497 Corporate Capstone Project
Fall and Spring: 9 units
This course is designed to give undergraduate students experience applying statistics data science methodology to real industry projects. Small groups of students will be matched with industry clients and do supervised projects for a semester. From an academic perspective, the course presents an opportunity for students to gain skills in approaching a research problem, critical thinking, and statistical analyses. Additionally, the course will help students develop the professional skills necessary for successfully navigating team-based project delivery roles. Client-facing and collaborative skills will be emphasized within a team setting, and students will learn leading practices for engaging stakeholders as well as gain a conceptual understanding of leading practices for project delivery. The industry clients will change and rotate each semester; available projects will be advertised prior to the first week of class. The course size is limited; students apply the previous semester and placed on the course waitlist until project matching is performed. Students with skill sets matching project needs will be given priority. We will also take into consideration whether or not a student has had a recent prior corporate capstone experience with the goal of providing experiences to a broad group of qualified students. Note that there is no guarantee a waitlisted student will be matched to a project in any given semester.

36-498 Corporate Capstone II
Fall and Spring
This course allows students to continue work on projects begun as part of 36-497, Corporate Capstone Project. Enrollment is at the discretion of the external advisor for the 36-497 project and the Department of Statistics and Data Science.
36-600 Overview of Statistical Learning and Modeling
Fall: 12 units
This is a first course in statistical practice, targeted specifically to CMU graduate students outside of statistics and machine learning. It is designed as a high-level introduction both to fundamental concepts of probability and statistics and to the ways by which statisticians go about approaching and analyzing data. The course will cover exploratory data analysis, parameter estimation and hypothesis testing, clustering, and common regression and classification models. If time permits, additional topics such as text mining, experimental design, and time series may be covered. Students will carry out all work using the R programming language. (Previous exposure to R is not required, nor is previous programming experience.) When covering probability and statistics concepts, basic facility with univariate calculus (at the level of AP Calc AB) will be assumed, but not explicitly required.

36-601 Professional Skills for Statisticians I
Fall: 6 units
The Professional Skills for Statisticians I course develops essential skills and required competencies for successfully navigating the workplace. Topics in this course include professional written and oral communication, communication approaches for varied audiences, leading practices for enabling high-functioning teams, managing project work, job search leading practices, and managing one's career. There will be a focus on the context of project work, and servicing and interacting with internal and external clients in the workplace. Consultative soft-skills are developed via group work and class presentations. This course is for Master's in Statistical Practice students only.

36-602 Perspectives in Data Science I
Spring: 6 units
This course covers the principles and practice of Data Science including data input and cleaning, exploratory data analysis, intermediate R programming, beginning SAS programming, beginning to intermediate python programming, and SQL. For Master in Statistical Practice (MSP) students only.

36-603 Perspectives in Data Science II
Spring: 6 units
This course is a continuation of 36-601 and includes exploring, analyzing, and interpreting data throughout the lifecycle of a data science project as well as introductory knowledge of specific applications of data science across various industries. For Master in Statistical Practice students only.

36-611 Professional Skills for Statisticians II
Fall: 6 units
The Professional Skills for Statisticians II course develops essential skills and required competencies for successfully navigating the workplace. Topics in this course include professional written and oral communication, communication approaches for varied audiences, leading practices for enabling high-functioning teams, managing project work, job search leading practices, and managing one's career. There will be a focus on the context of project work, and servicing and interacting with internal and external clients in the workplace. Consultative soft-skills are developed via group work and class presentations. This course is for Master's in Statistical Practice students only.

36-612 Data Visualization
Fall: 6 units
For Master of Statistical Practice (MSP) students only. This course focuses on data visualization.

36-614 Data Engineering and Distributed Environments
Fall: 6 units
For Master of Statistical Practice (MSP) students only. This course focuses on data engineering.

36-615 Applied Linear Models
Fall: 12 units
This course covers the theory and practice of linear models in matrix form with emphasis on practical skills for working with real data and communicating results to technical and non-technical audiences. For Master's in Statistical Practice students only.

36-616 Time Series and Experimental Design
Spring: 12 units
This course covers fundamentals of experimental design including various ANOVA models, Latin squares and factorial and fractional factorial designs. The time series components covers exponential smoothing models and ARIMA, including seasonal models and transfer function models. Special topics are intermittent. For Master's in Statistical Practice students only.

36-625 Probability and Mathematical Statistics I
Fall: 12 units
This course is a rigorous introduction to the mathematical theory of probability, and it provides the necessary background for the study of mathematical statistics and probability modeling. A good working knowledge of calculus is required. Topics include combinatorial analysis, conditional probability, generating functions, sampling distributions, law of large numbers, and the central limit theorem. Undergraduate students studying Computer Science, or considering graduate work in Statistics or Operations Research, must receive permission from their advisor and from the instructor. Prerequisites: 21-122 and 21-241 and (21-256 or 21-259). Prerequisites: 21-122 or 21-123 or 21-256 or 21-118

36-626 Probability and Mathematical Statistics II
Intermittent: 12 units
An introduction to the mathematical theory of statistical inference. Topics include likelihood functions, estimation, confidence intervals, hypothesis testing, Bayesian inference, regression, and the analysis of variance. Not open to students who have received credit for 36-226. Students studying Computer Science should carefully consider taking this course instead of 36-220 or 36-226 after consultation with their advisor. Prerequisite: 36-625. Prerequisite: 36-625

36-631 Data Engineering and Distributed Environments
Fall: 6 units
For Master of Statistical Practice (MSP) students only. This course focuses on data engineering.

36-632 Probability and Mathematical Statistics II
Intermittent: 6 units
The topic of the course will be applied methods for the analysis of survival data. Survival time data arises when the random quantity of interest is time until an event occurs. These types of data arise in longitudinal studies, such as randomized controlled trials, as well as in longitudinal, observational studies, such as cohort studies. The emphasis of this course will be on understanding the class of problems and data in which these methods are applicable and the use of these methods in practice. Although this is not a theory class some foundational knowledge of probability theory and mathematical statistics are necessary. Basics concepts from probability and mathematical statistics will be reviewed as needed.

36-636 Methods for Clinical Trials
Intermittent: 6 units
Clinical trials have become an essential research tool for the evaluation of the benefit and risk of new interventions. This course will focus primarily on interventions for the treatment or prevention of disease, however, the principles and methods have also found applications in the social and behavioral sciences. The fundamental principles of clinical trials are heavily based on statistical principles related to experimental design, quality control, and sound analysis. No analytical methods can rescue a trial with poor experimental design and the conclusions from a trial with proper design can be invalid if sound analytical principles are not adhered to. This course, collection appropriate and high quality data is essential. The principles presented in this course are an introduction to important statistical concepts in design, conduct, and analysis. Although this is not a theory class some foundational knowledge of probability theory and mathematical statistics are necessary. Basics concepts from probability and mathematical statistics will be reviewed as needed.

36-650 Statistical Computing
Spring: 9 units
A detailed introduction to elements of computing relating to statistical modeling, targeted to advanced undergraduates, masters students, and doctoral students in Statistics. Topics include important data structures and algorithms; numerical methods; databases; parallelism and concurrency; and coding practices, program design, and testing. Multiple programming languages will be supported (e.g., C, R, Python, etc.). Those with no previous programming experience are welcome but will be required to learn the basics of at least one language via self-study. There are very limited spots for undergraduates; special permission from both instructor and advisor required.

36-651 Advanced Statistical Computing
Intermittent: 6 units
A project-based course in statistical computing. Students will choose individual projects on computing topics related to statistical modeling and practice, including databases, parallel and cluster programming, big data frameworks (e.g. Spark or Hadoop), algorithms and data structures, numerical methods, and other topics based on student interest. The course will include introductions to each topic as well as student presentations on the results of their projects. Multiple programming languages will be supported. Recommended prerequisite: 36-650 or 36-750
Prerequisite: 36-650 Min. grade B
36-652 Statistical Computing II
Spring: 9 units
A detailed discussion of computing elements relating to statistical modeling, targeted to advanced undergraduates and master's students in Statistics. Topics include databases, cloud storage models, data integration concepts; numerical methods; big data analytics frameworks; machine learning models; graphical user interface; low-code application development; and DevOps. The course will focus on using Python programming language. Comparable programming experience in Python -as judged by the instructor- is required. There are very limited spots for undergraduates; special permission from both instructor and advisor required. Prerequisite: 36-650 Min. grade C

36-661 Special Topics: Statistical Methods in Epidemiology
Intermittent: 9 units
Epidemiology is concerned with understanding factors that cause, prevent, and reduce diseases by studying associations between disease outcomes and their suspected determinants in human populations. Epidemiologic research requires an understanding of statistical methods and design. Epidemiologic data is typically discrete, i.e., data that arise whenever counts are made instead of measurements. In this course, methods for the analysis of categorical data are discussed with the purpose of learning how to apply them to data. The central statistical themes are building models, assessing fit and interpreting results. There is a special emphasis on generating and evaluating evidence from observational studies. Case studies and examples will be primarily from the public health sciences.

36-663 Multilevel and Hierarchical Models
Intermittent: 9 units
Multilevel and hierarchical models are among the most broadly applied “sophisticated” statistical models, especially in the social and biological sciences. They apply to situations in which the data “cluster” naturally into groups of units that are more related to each other than they are the rest of the data. In the first part of the course we will see how to generalize linear models to multilevel and hierarchical models and relate them to other areas of statistics, and in the last part of the course we will learn how Bayesian statistical methods can help us to build, estimate and diagnose problems with these models using a variety of data sets and examples.

36-665 Special Topics: Conceptual Foundations of Statistical Learning
Intermittent: 9 units
This class is an introduction to the foundations of statistical learning theory, and its uses in designing and analyzing machine-learning systems. Statistical learning theory studies how to fit predictive models to training data, usually by solving an optimization problem, in such a way that the model will predict well, on average, on new data. The course will focus on the key concepts and practical tools, at a level accessible to students who have taken 36-401 and its prerequisites. The course will also illustrate those concepts and tools by applying them to carefully selected kinds of machine learning systems (such as kernel machines). Students wanting exposure to a broad range of algorithms and applications would be better served by 36-462/662 (“Data Mining”). This class is for those who want a deeper understanding of the principles underlying all machine learning methods.

36-666 Special Topics: Statistical Methods in Finance
Intermittent: 9 units
Financial econometrics is the interdisciplinary area where we use statistical methods and economic theory to address a wide variety of quantitative problems in finance. These include building financial models, testing financial economics theory, simulating financial systems, volatility estimation, risk management, capital asset pricing, derivative pricing, portfolio allocation, proprietary trading, portfolio and derivative hedging, and so on and so forth. Financial econometrics is an active field of integration of finance, economics, probability, statistics, and applied mathematics. Financial activities generate many new problems and products, economics provides useful theoretical foundation and guidance, and quantitative methods such as statistics, probability and applied mathematics are essential tools to solve quantitative problems in finance. Professionals in finance now routinely use sophisticated statistical techniques and modern computation power in portfolio management, proprietary trading, derivative pricing, financial consulting, securities regulation, and risk management.

36-667 Special Topics: Data over Space & Time
Intermittent: 9 units
This course is an introduction to the opportunities and challenges of analyzing data from processes unfolding over space and time. It will cover basic descriptive statistics for spatial and temporal patterns; linear methods for interpolating, extrapolating, and smoothing spatio-temporal data; basic nonlinear modeling; and statistical inference with dependent observations. Class work will combine practical exercises in R, a little mathematics on the underlying theory, and case studies analyzing real problems from various fields (economics, history, meteorology, ecology, etc.). Depending on available time and class interest, additional topics may include: statistics of Markov and hidden-Markov (state-space) models; statistics of point processes; simulation and simulation-based inference; agent-based modeling; dynamical systems theory.

36-668 Special Topics: Text Analysis
Intermittent: 9 units
The analysis of language is concerned with how variables relate to people (their gender, age, and location, for example), how variables relate to use (such as writing in different academic disciplines), and how variables change over time. While we are surrounded by data that might potentially shed light on many of these questions, working with real-world linguistic data can present some unique challenges in sampling, in the distribution of features, and in their high dimensionality. In this course, we will work through some of these issues, paying particular attention to the aligning of the statistical questions we want to investigate with the choice of statistical models, as well as focusing on the interpretation of results. Analysis will be carried out in R and students will develop a suite of tools as they work through their course projects.

36-669 Stat/High Dim Genetics
Intermittent: 9 units
The field of computational and statistical genomics focuses on developing and applying computationally efficient and statistically robust methods to sort through increasingly rich and massive genome wide data sets to identify complex genetic patterns, gene interactions, and disease associations. Because the genome is vast, analytical approaches require high-dimensional statistical approaches such as multiple testing, dimension reduction techniques, regularization and high dimensional regression analysis, best linear unbiased prediction models, networks and graphical models. In this course, we will motivate these topics using data obtained from the human genetic and genomic literature. No prior knowledge in biology is required.

36-675 Data Ethics & Responsible Conduct of Research
Intermittent: 3 units
TBD

36-692 Topic Detection and Document Clustering
Intermittent: 6 units
Imagine if someone read all your email. Everything you sent, everything you received. What would they find? Do you have repeating topics? How do the topics change over time? The Enron Corporation was an energy, commodities, and services company in Houston, Texas that went spectacularly bankrupt in 2001 after it was revealed that it was engaging in systematic, planned accounting fraud. At its peak, it employed over 20,000 people with revenues over $100 billion. Its downfall was related to deregulation of California’s energy commodity trading and a series of rolling power blackouts over months. For example, Enron traders encouraged the removal of power during the energy crisis by suggesting plant shutdowns. The resulting increase in the price for power made them a fortune. After Enron’s collapse, journalists used the Freedom of Information Act to release the emails sent/received by the employees of Enron. Subsequently, the emails were analyzed to see who knew what and when. Every news article, email, letter, blog, tweet, etc can be thought of as an observation. We characterize these documents by their length, what words they use and how often, and possibly extra information like the time, the recipient, etc. Topic detection and document clustering methods are statistical and machine learning tools that extract and identify related documents, possibly over time. These methods need to be flexible enough to handle both very small and very large clusters of documents, topics that change in importance, and topics that appear and disappear over time. This class will focus on methods and real-world data analysis. Class time will be split into lecture and “lab”. (Bring your laptop.) Occasional homeworks and final project, but mostly we’ll focus on the downfall of Enron as our overarching case study.
36-699 Statistical Immigration
Fall: 6 units
Students are introduced to the faculty and their interests, the field of
statistics, and the facilities at Carnegie Mellon. Each faculty member gives
at least one elementary lecture on some topic of his or her choice. In the
past, topics have included: the field of statistics and its history, large-
scale sample surveys, survival analysis, subjective probability, time series,
robustness, multivariate analysis, psychiatric statistics, experimental
design, consulting, decision-making, probability models, statistics and the
law, and comparative inference. Students are also given information about
the libraries at Carnegie Mellon and current bibliographic tools. In addition,
students are instructed in the use of the Departmental and University
computational facilities and available statistical software packages. THIS
COURSE IS FOR PHD STUDENTS IN THE DEPT OF STATISTICS ONLY.

36-700 Probability and Mathematical Statistics
Fall: 12 units
This is a one-semester course covering the basics of statistics. We will
first provide a quick introduction to probability theory, and then cover
fundamental topics in mathematical statistics such as point estimation,
hypothesis testing, asymptotic theory, and Bayesian inference. If time
permits, we will also cover more advanced and useful topics including
nonparametric inference, regression and classification. Prerequisites: one-
and two-variable calculus and matrix algebra. Graduate students in degree-
seeking programs are given priority.

36-705 Intermediate Statistics
Fall: 12 units
This course covers the fundamentals of theoretical statistics. Topics include:
probability inequalities, point and interval estimation, minimax theory,
hypothesis testing, data reduction, convergence concepts, Bayesian
inference, nonparametric statistics, bootstrap resampling, VC dimension,
prediction and model selection. This course is primarily for PhD students in
Statistics and Data Science, Machine Learning, and Computer Science; it
requires an appropriate background for entering those programs.

36-707 Regression Analysis
All Semesters: 12 units
This is a course in data analysis. Topics covered include: Simple and
multiple linear regression, causation, weighted least-squares, global and
case diagnostics, robust regression, exponential families, logistic regression
and generalized linear models; Model selection: prediction risk, bias-
variance tradeoff, risk estimation, model search, ridge regression and
lasso, stepwise regression, maybe boosting; Smoothing and nonparametric
regression: linear smoothers, kernels, local regression, penalized regression,
regularization and splines, wavelets, variance estimation, confidence
bands, local likelihood, additive models; Classification: parametric and
nonparametric regression, LDA, QDA, trees. Practice in data analysis is
obtained through course projects. This course is primarily for first year
PhD students in Statistics and Data Science; it requires an appropriate
background for entering that program.

36-708 The ABCDE of Statistical Methods in Machine Learning
All Semesters: 12 units
This course focuses on statistical methods for machine learning, a decades-
old topic in statistics that now has a life of its own, intersecting with many
other fields. While the core focus of this course is methodology (algorithms),
the course will have some amount of formalization and rigor (theory/
derivation/proof), and some amount of interacting with data (simulated
and real). However, the primary way in which this course complements
related courses in other departments is the joint ABCDE focus on (A)
Algorithm design principles, (B) Bias-variance thinking, (C) Computational
considerations (D) Data analysis (E) Explainability and interpretability.
Prerequisite: 36-705 Min. grade A

36-709 Advanced Statistical Theory I
All Semesters: 12 units
This is a core Ph.D. course in theoretical statistics. The class will cover a
selection of modern topics in mathematical statistics, focussing on high-
dimensional parametric models and non-parametric models. The main
goal of the course is to provide the students with adequate theoretical
background and mathematical tools to read and understand the current
statistical literature on high-dimensional models. Topics will include:
concentration inequalities, covariance estimation, principal component
analysis, penalized linear regression, maximal inequalities for empirical
processes, Rademacher and Gaussian complexities, non-parametric
regression and minimax theory. This will be the first part of a two semester
sequence.
Prerequisite: 36-705 Min. grade A

36-710 Advanced Statistical Theory II
All Semesters: 12 units
This is a core Ph.D. course in theoretical statistics. The class will cover a
selection of modern topics in mathematical statistics, focussing on high-
dimensional parametric models and non-parametric models. The main
goal of the course is to provide the students with adequate theoretical
background and mathematical tools to read and understand the current
statistical literature on high-dimensional models. Topics will include:
concentration inequalities, covariance estimation, principal component
analysis, penalized linear regression, maximal inequalities for empirical
processes, Rademacher and Gaussian complexities, non-parametric
regression and minimax theory.
36-730 Graphical Models and Its Applications
Intermittent: 6 units
Probabilistic graphical models (PGMs) lie at the intersection of probability and graph theory. Its application to real world problems has served useful in the process of understanding, formulating and solving problems, and in particular as tools for making decisions and calculating the probability of a particular based on (often incomplete) collections of prior knowledge. This course will introduce the fundamentals of graphical models and probability propagation algorithms; demonstrate how to build and model (PGMs) using R, focusing on DAGs. The aim will be to learn and demonstrate the versatility of PGMs, through applications and methodology, including its use in decision support, causal and temporal problems. Applications will focus on areas of public policy including criminal justice/forensic science, health/medical, environment, etc.

36-731 Foundations of Causal Inference
Intermittent: 6 units
This course will provide an introduction to the fundamentals of causal inference. Causal inference is concerned with whether and how one can go beyond statistical associations to draw causal conclusions from observational data. Topics will include: counterfactuals (potential outcomes and graphs), identification and estimation of average treatment effects in experiments and observational studies, nonparametric bounds, sensitivity analysis, instrumental variables, effect modification, and longitudinal studies.

36-732 Modern Causal Inference
Intermittent: 6 units
This course will provide an in-depth look at modern causal inference. Topics will include: optimal treatment regimes, mediation, principal stratification, stochastic interventions, accounting for complex confounding and exposures, and methods for efficient nonparametric estimation. Some background in mathematical statistics is advised.

36-733 Probability Models and Stochastic Processes
Intermittent: 6 units
By the end of this course you will be able to handle basic discrete and continuous time stochastic processes, including random walks, branching processes, discrete and continuous Markov chains, Markov chain Monte Carlo (MCMC), Poisson and renewal processes, birth and death processes, and queuing systems. This class is not overly mathematical, but techniques such as generating functions, difference and differential equations, linear systems of equations, are needed at a basic level.

36-736 Methods for Clinical Trials
Intermittent: 6 units
Clinical trials have become an essential research tool for the evaluation of the benefit and risk of new interventions. This course will focus primarily on interventions for the treatment or prevention of disease, however, the principles and methods have also found applications in the social and behavioral sciences. The fundamental principles of clinical trials are heavily based on statistical principles related to experimental design, quality control, and sound analysis. No analytical methods can rescue a trial with poor experimental design and the conclusions from a trial with proper design can be invalid if sound analytical principles are not adhered to. Of course, collection of appropriate and high quality data is essential. The principles presented in this course are an introduction to important statistical concepts in design, conduct, and analysis. Although this is not a theory class some foundational knowledge of probability theory and mathematical statistics are necessary. Basics concepts from probability and mathematical statistics will be reviewed as needed.

36-741 Statistics meets Optimization: Randomized Sketching Methods
All Semesters: 6 units
In this mini, we will discuss some aspects of the interface between statistics and optimization. The goal of these lectures is to touch on various evolving areas at this interface. The objectives of optimization can be influenced by underlying statistical objectives in many ways, for example, the statistics precision caused by not having enough sample size is often of higher order than the machine precision; worst-case instance can be too conservative compared to the random ensemble; polynomial-time complexity may still be too large to be tractable. To further discuss these issues, we will start with a dimension reduction technique based on random projections and analyze how this technique helps us achieve faster optimization convergence without hurting statistical precision.

36-742 Statistics meets Optimization: Approximate Message Passing Algorithm
All Semesters: 6 units
In this mini, we focus our attention on the recent development of the approximate message passing algorithm. We follow a rigorous approach that builds upon ideas from statistical physics, information theory and graphical models, and is based on the analysis of an highly efficient reconstruction algorithm. We start with some basics for the probability graphical model, introduce the message passing algorithm and motivate the AMP algorithm along the way. Then we will discuss the exact asymptotic characterization in terms of the so-called state evolution and talk about the applications in LASSO and more generally, high-dimensional robust M-estimation.

36-743 Statistical Methods for Reproducibility and Replicability: Static Settings
Intermittent: 6 units
See URL
Course Website: http://www.stat.cmu.edu/~aramdas/reproducibility19/

36-744 Statistical Methods for Reproducibility and Replicability: Dynamic Settings
All Semesters: 6 units
See URL
Course Website: http://www.stat.cmu.edu/~aramdas/reproducibility19/

36-746 Statistical Methods for Neuroscience and Psychology
Intermittent: 12 units
This course provides a survey of basic statistical methods, emphasizing motivation from underlying principles and interpretation in the context of neuroscience and psychology. Though 36-746 assumes only passing familiarity with school-level statistics, it moves faster than typical university-level first courses. Vectors and matrices will be used frequently, as will basic calculus. Topics include Probability, Random Variables, and Important Distributions (binomial, Poisson, and normal distributions; the Law of Large Numbers and the Central Limit Theorem); Estimation and Uncertainty (standard errors and confidence intervals; the bootstrap); Principles of Estimation (mean squared error; maximum likelihood); Models, Hypotheses, and Statistical Significance (goodness-of-fit, p-values; power); General methods for testing hypotheses (permutation, bootstrap, and likelihood ratio tests); Linear Regression (simple linear regression and multiple linear regression); Analysis of Variance (one-way and two-way designs; multiple comparisons); Generalized Linear and Nonlinear Regression (logistic and Poisson regression; generalized linear models); and Nonparametric regression (smoothing scatterplots; smoothing histograms).

36-750 Statistical Computing
Fall: 9 units
A detailed introduction to elements of computing relating to statistical modeling, targeted to PhD students and masters students in Statistics and Data Science. Topics include important data structures and algorithms; numerical methods; databases; parallelism and concurrency; and coding practices, program design, and testing. Multiple programming languages will be supported (e.g., C, R, Python, etc.). Those with no previous programming experience are welcome but will be required to learn the basics of at least one language via self-study.

36-751 Advanced Statistical Computing
Intermittent: 6 units
A project-based course in statistical computing. Students will choose individual projects on computing topics related to statistical modeling and practice, including databases, parallel and cluster programming, big data frameworks (e.g. Spark or Hadoop), algorithms and data structures, numerical methods, and other topics based on student interest. The course will include introductions to each topic as well as student presentations on the results of their projects. Multiple programming languages will be supported. Recommended prerequisite: 36-650 or 36-750 Prerequisite: 36-750 Min. grade B

36-752 Statistical Computing II
All Semesters: 9 units
A detailed discussion of computing elements relating to statistical modeling. Topics include databases; cloud storage models; data integration concepts; numerical methods; big data analytics frameworks; machine learning models; graphical user interface; low-code application development; and DevOps. The course will focus on using Python programming language. Comparable programming experience in Python -as judged by the instructor- is required.
This new course is intended for CNBC students, as an additional option for fulfilling the computational core course requirement, but it will also be open to Statistics and Machine Learning students. It should be of interest to anyone wishing to see the way statistical ideas play out within the brain sciences, and it will provide a series of case studies on the role of stochastic models in scientific investigation. Statistical ideas have been prevalent in neurophysiology and the brain sciences since the first stochastic description of spike trains, and the quantal hypothesis of neurotransmitter release, more than 50 years ago. Many contemporary theories of neural system behavior are built with statistical models. For example, integrate-and-fire neurons are usually assumed to be driven in part by stochastic noise; the role of spike timing involves the distinction between Poisson and non-Poisson neurons; and oscillations are characterized by decomposing variation into frequency-based components. In the visual system, V1 simple cells are often described using linear-nonlinear Poisson models; in the motor system, neural response may involve direction tuning; and CA1 hippocampal receptive field plasticity has been characterized using dynamic place models. It has also been proposed that perceptions, decisions, and actions result from optimal (Bayesian) combination of sensory input with previously-learned regularities; and some investigators report new insights from viewing whole-brain pattern responses as analogous to statistical classifiers. Throughout the field of statistics, models incorporating random "noise" components are used as an effective vehicle for data analysis. In neuroscience, however, the models also help form a conceptual framework for understanding neural function. This course will examine some of the most important methods and claims that have come from applying statistical thinking.

There is no one correct way to write. But there are things you can do that tend to make it difficult for a reader to absorb the ideas you are writing about, or make it easier for the reader. Thus, it is important to focus on the reader, and the constraints and habits of mind that most readers (even in the rarefied population of academics who can understand the technical details of your work) bring to the task of reading what you have written. The goals for students in this course are: to understand that writing requires an intellectual investment similar to the investment that you put into other areas of your research, from developing research questions, data collection, and data analysis, to writing and testing algorithms, and formulating and proving theorems; to understand ways of organizing your writing that make it more likely that the reader will interpret and understand your ideas in the way that you intend; and to gain experience writing with these ideas in mind. The course is most suitable for graduate students in statistics who are engaged in a writing project (ADA paper, journal article, thesis work, etc.).
36-792 Topic Detection and Document Clustering
Intermittent: 6 units
Imagine if someone read all your email. Everything you sent, everything you received. What would they find? Do you have repeating topics? How do the topics change over time? The Enron Corporation was an energy, commodities, and services company in Houston, Texas that went spectacularly bankrupt in 2001 after it was revealed that it was engaging in systematic, planned accounting fraud. At its peak, it employed over 20,000 people with revenues over $100 billion. Its downfall was related to deregulation of California’s energy commodity trading and a series of rolling power blackouts over months. For example, Enron traders encouraged the removal of power during the energy crisis by suggesting plant shutdowns. The resulting increase in the price for power made them a fortune. After Enron’s collapse, journalists used the Freedom of Information Act to release the emails sent/received by the employees of Enron. Subsequently, the emails were analyzed to see who knew what and when. Every news article, email, letter, blog, tweet, etc can be thought of as an observation. We characterize these documents by their length, what words they use and how often, and possibly extra information like the time, the recipient, etc. Topic detection and document clustering methods are statistical and machine learning tools that extract and identify related documents, possibly over time. These methods need to be flexible enough to handle both very small and very large clusters of documents, topics that change in importance, and topics that appear and disappear. This class will emphasize application of methods and real-world data analysis. Class time will be split into lecture and “lab”. (Bring your laptop.) Occasional homeworks and final project, but mostly we’ll focus on the downfall of Enron as our overarching case study.

36-795 Interdisciplinary Applied Research
Fall: 12 units
This course is the second course of the advanced applied data analysis sequence and is for PhD students in Statistics and Data Science only.