36-200 Reasoning with Data
Fall and Spring: 9 units
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. Not open to students who have received credit for 36-201, 36-207/70-207, 36-220, 36-247, 36-225, or any upper level course in Statistics. This course is the credit-equivalent to 36-201 and will be honored appropriately as a prerequisite for downstream Statistics courses. As such, this course is not currently open to students who have received credit for 36-201, 36-207/70-207, 36-220, 36-247, or any 300- or 400-level Statistics course.

36-201 Statistical Reasoning and Practice
Intermittent: 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. Examples will be drawn from areas of applications of particular interest to H&SS students. Not open to students who have received credit for 36-207/70-207, 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 Statistics & Data Science Methods
Spring: 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, analysis of variance methods and logistic regression. Other topics may include non-parametric methods and probability models, as time permits. The objectives of this course is to develop the skills of applying the basic principles and methods that underlie statistical practice and empirical research. In addition to three lectures a week, students attend a computer lab once week for "hands-on" practice of the material covered in lecture. Not open to students who have received credit for: 36-208/70-208, 36-300. Students who have completed 36-401 prior to or concurrent with 36-202 will not receive credit for 36-202. Prerequisites: 36-201 or 36-200 or 36-247 or 70-207 or 36-220 or 36-207

Course Website: [http://www.stat.cmu.edu/academics/courselist](http://www.stat.cmu.edu/academics/courselist)

36-207 Probability and Statistics for Business Applications
Spring: 9 units
This is the first 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 21-120 or 21-112

Course Website: [http://www.stat.cmu.edu/academics/courselist](http://www.stat.cmu.edu/academics/courselist)

36-208 Regression Analysis
Spring: 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 time series, regression and forecasting. In addition to two lectures a week, students will attend a computer lab once a week. Not open to students who have received credit for 36-202, 36-626. Cross-listed as 70-208. Students who have completed 36-401 prior to 36-208 will not receive credit for 36-208.

Prerequisites: (21-112 or 21-120) and (36-201 or 70-207 or 36-247 or 36-207 or 36-220) 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 21-123 or 21-122 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
All Semesters: 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 methods 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, 36-236, 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: 21-259 or 21-112 or 21-122 or 21-212 or 21-256

Course Website: [http://www.stat.cmu.edu/academics/courselist](http://www.stat.cmu.edu/academics/courselist)

36-220 Engineering Statistics and Quality Control
All Semesters: 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 36-201, 36-207/70-207, 36-226, 36-626, or 36-247, except when AP credit is awarded for 36-201.

Prerequisites: 21-112 or 21-120 or 21-121

Course Website: [http://www.stat.cmu.edu/academics/courselist](http://www.stat.cmu.edu/academics/courselist)
36-225 Introduction to Probability Theory
Fall: 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 economics, mathematics and statistics. The use of probability theory is illustrated with examples drawn from engineering, the sciences, and management. 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, 21-325, 36-700.
Course Website: http://www.stat.cmu.edu/academics/courselist

36-226 Introduction to Statistical Inference
Spring: 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, and hypothesis testing. 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: 15-359 Min. grade C or 36-225 Min. grade C or 36-217 Min. grade C or 21-325 Min. grade C or 36-218 Min. grade C
Course Website: http://www.stat.cmu.edu/academics/courselist

36-247 Statistics for Lab Sciences
Spring: 9 units
This course is a single-semester comprehensive introduction to statistical analysis of data for students in biology and chemistry. Topics include exploratory data analysis, elements of computer programming for statistics, basic concepts of probability, statistical inference, and curve fitting. In addition to two lectures, students attend a computer lab each week. Not open to students who have received credit for 36-201, 36-207/0-207, 36-220, or 36-226. Prerequisites: 21-112 or 21-120 or 21-121

36-290 Introduction to Statistical Research Methodology
Intermittent: 9 units
This course is designed to introduce statistical research methodology—the procedures by which statisticians go about approaching and analyzing data—to early undergraduates. Students will learn basic concepts of statistical learning—infERENCE 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. Analysis will primarily be carried out using the R programming language, but with attention paid to how one would perform similar analyses using Python. Previous knowledge of R is not required for this course. Space is very limited; there will be an application process. The course is currently open to sophomore statistics students only.

36-300 Statistics & Data Science Internship
Summer: 3 units
The Department of Statistics & 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 dont 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 the Department of Statistics & Data Science for more details.

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: 88-250 or 36-208 or 36-218 or 36-226 or 36-220 or 36-625 or 36-225 or 73-261 or 36-309 or 70-208
Course Website: http://www.stat.cmu.edu/academics/courselist

36-304 Biostatistics
Fall: 9 units
TBD

36-309 Experimental Design for Behavioral & Social Sciences
Fall: 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: 36-207 or 36-220 or 36-217 or 36-200 or 36-201 or 36-247
Course Website: http://www.stat.cmu.edu/academics/courselist

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. Prerequisites: 36-226

36-314 Biostatistics
Fall: 9 units
This course is an introduction to methods used frequently in biostatistics and public health applications. Prerequisites: 36-226 or 88-250 or 36-225 or 70-208 or 36-303 or 36-309 or 36-202 or 36-208 or 36-625

36-315 Statistical Graphics and Visualization
Spring: 9 units
Graphical displays of quantitative information take on many forms as they help us understand both data and models. This course will serve to introduce the student to the most common forms of graphical displays and their uses and misuses. Students will learn both how to create these displays and how to understand them. As time permits the course will consider some more advanced graphical methods such as computer-generated animations. Each student will be required to engage in a project using graphical methods to understand data collected from a real scientific or engineering experiment. In addition to two weekly lectures there will be lab sessions where the students learn to use software to aid in the production of appropriate graphical displays. Prerequisites: 21-325 or 36-625 or 36-225 or 36-309 or 36-303 or 36-208 or 36-218 or 36-217 or 88-250 or 36-226 or 36-208 or 36-202
Course Website: http://www.stat.cmu.edu/academics/courselist

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 36-225 or 36-217 or 36-200 or 36-201 or 36-247

36-328 Biostatistics
Fall: 9 units
This course will cover models of how networks impact individual behavior.

36-350 Statistical Computing
Fall and Spring: 9 units
Statistical Computing: An introduction to computing targeted at statistics majors with minimal programming knowledge. The main topics are core ideas of programming (functions, objects, data structures, flow control, input and output, debugging, logical design and abstraction), illustrated through key statistical topics (exploratory data analysis, basic optimization, linear models, graphics, and simulation). The class will be taught in the R language. No previous programming experience required. 36-225 is a prerequisite.
Prerequisites: 21-325 Min. grade C or 36-217 Min. grade C or 36-225 Min. grade C or 15-259 Min. grade C or 36-218 Min. grade C
Course Website: http://www.stat.cmu.edu/academics/courselist
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: 36-226 Min. grade C or 36-218 Min. grade B or 36-625 Min. grade C or 36-362 Min. grade C and (21-240 or 21-241)
Course Website: http://www.stat.cmu.edu/academics/courselist

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.
Prerequisites: 36-401
Prerequisite: 36-401 Min. grade C
Course Website: http://www.stat.cmu.edu/academics/courselist

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: 36-225 or 36-217 or 21-325 or 36-625
Course Website: http://www.stat.cmu.edu/academics/courselist

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 the 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, model choice operates on the 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://www.stat.cmu.edu/academics/courselist

36-462 Special Topics: Data Mining
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.
Prerequisites: 36-401 (C or better).
Prerequisite: 36-401 Min. grade C
Course Website: http://www.stat.cmu.edu/academics/courselist

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 to 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/courselist

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.
Prerequisites: 36-401 (C or better).
Prerequisite: 36-401 Min. grade C
Course Website: http://www.stat.cmu.edu/academics/courselist
36-465 Special Topics: An Introduction to Bayesian Inference
Intermittent: 9 units
The aim of this course is to introduce students to theory and application of Bayesian statistical modeling and inference. The course starts with epistemological differences between the Bayesian and Frequentist paradigms and the treatment of simple models, such as those based on well-known distributions. Concepts of conjugate and non informative priors are illustrated, for single- and multi-parameters models. Basic treatment of hierarchical models and linear regression models are also covered. Bayesian computational methods such as the Gibbs sampler and Metropolis-Hastings algorithms, are briefly presented with an emphasis on their implementation and use on simple cases. 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.

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. Co-requisite: For undergraduates taking the course as 36-467, 36-401. For graduate students taking the course as 36-467, 36-401. For graduate students taking the course as 36-667, consent of the professor.

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 will work through some of these issues, paying particular attention to the alignment 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.

36-490 Undergraduate Research
Intermittent: 9 units
This course is designed to give undergraduate students experience using statistics in real research problems. Small groups of students will be matched with clients and do supervised research for a semester. Students will gain skills in approaching a research problem, critical thinking, statistical analysis, scientific writing, and conveying and defending their results to an audience. Prerequisite: 36-401
Course Website: http://www.stat.cmu.edu/academics/courselist

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 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 of 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

36-494 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 astrostatiscticians 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 la 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. Students will gain skills in approaching a real world problem, critical thinking, advanced statistical analysis, scientific writing, collaborating in an industry setting, communicating results, and meeting expectations with respect to deliverables and timelines. The industry clients will change and rotate each semester; available projects will be advertised prior to registration. 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 corporate capstone experience with the goal of providing experiences to a broad group of qualified students.

36-601 Perspectives in Data Science I
Fall: 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’s in Statistical Practice students only.

36-602 Perspectives in Data Science II
Spring: 9 units
This course is a continuation of 36-601 and covers interactive data visualization with Shiny, advanced R programming techniques, intermediate SAS (macros), web scraping, Hadoop, and Spark. For Master’s in Statistical Practice students only. Prerequisite: 36-601 Min. grade C
36-611 Professional Skills for Statisticians I
Fall: 6 units
This course covers a variety of professional skills including resumes and cover letters, writing reports, oral presentations, teamwork, and project planning. Consulting skills are developed in the form of a whole-class consulting project. For Master's in Statistical Practice students only.

36-612 Professional Skills for Statisticians II
Spring: 6 units
This course is a continuation of 36-611 and covers additional writing and presentation skills, as well as interview skills. For Master's in Statistical Practice students only.
Prerequisite: 36-611 Min. grade C

36-617 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.

36-618 Experimental Design & Time Series
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.
Prerequisites: 36-601 Min. grade C and 36-617 Min. grade C

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. Prerequisite: 21-122 and 21-241 and (21-256 or 21-259).
Prerequisites: 21-123 or 21-256 or 21-118 or 21-122

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

36-635 Applied Survival Analysis
Intermittent: 6 units
TBD

36-636 Methods for Clinical Trials
Intermittent: 6 units
TBD

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 advisor and instructor 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

36-661 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: Bayesian Methods
Intermittent: 9 units
TBD

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
TBD

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. 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.

36-699 Statistical Immigration
Fall: 3 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 program 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.

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.

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 & Data Science; it requires an appropriate background for entering that program.

36-708 Statistical Methods in Machine Learning
All Semesters: 12 units
TBD
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 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
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 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.

36-721 Statistical Graphics and Visualization
Intermittent: 6 units
An effective statistical graphic is a powerful tool for analyzing data and communicating insights. This course will introduce students to creating, understanding, and critiquing such graphical displays, choosing the right visual tool for the task at hand. Students will learn how to produce legible, self-contained, informative graphics using statistical software, as well as how to plan effective statistical graphics by following the principles of human visual perception. Beyond the most commonly used graphs for univariate and bivariate data, we will cover useful visualizations for statistical model diagnostics; cartographic maps; network- and tree-structured data; and interactive exploration of high-dimensional datasets. Through project assignments, students will practice applying the principles of graphic design and interaction design. Course materials will primarily use R (including ggplot2 and Shiny), but we will also introduce illustrator/inkscape and Tableau, and students may complete assignments using other software if they wish (Python, MATLAB, etc.).

36-725 Convex Optimization
Intermittent: 12 units
Nearly every problem in machine learning can be formulated as the optimization of some function, possibly under some set of constraints. This universal reduction may seem to suggest that such optimization tasks are intractable. Fortunately, many real world problems have special structure, such as convexity, smoothness, separability, etc., which allow us to formulate optimization problems that can often be solved efficiently. This course is designed to give a graduate-level student a thorough grounding in the formulation of optimization problems that exploit such structure, and in efficient solution methods for these problems. The main focus is on the formulation and solution of convex optimization problems. These general concepts will also be illustrated through applications in machine learning and statistics. Students entering the class should have a pre-existing working knowledge of algorithms, though the class has been designed to allow students with a strong numerate background to catch up and fully participate. Though not required, having taken 10-701 or an equivalent machine learning or statistics class is strongly encouraged, since we will use applications in machine learning and statistics to demonstrate the concepts we cover in class. Students will work on an extensive optimization-based project throughout the semester; those wanting to take the class without the project can register under the 9 unit option.
Course Website: http://www.stat.cmu.edu/~ryanibs/convexopt/

36-726 Statistical Practice
Spring: 12 units
Students are taught how to structure a consulting session, elicit and diagnose a problem, manage a project, and report an analysis. The class will participate in meetings with industrial and academic clients. For Master's in Statistical Practice students only.
36-727 Modern Experimental Design
Intermittent: 6 units
Designed experiments are crucial to draw causal conclusions with minimum expense and maximum precision. This course introduces the basic principles and theory of experimental design, including randomized designs, blocking, analysis of covariance, factorial designs, and power analysis, along with a selection of more advanced topics, which may include sequential and adaptive designs, A/B testing, the design of observational studies, or other topics depending on time and class interest. Students will learn to design appropriate experiments for a variety of research scenarios, and practice these skills through a course project. Coursework will primarily use R for analysis of experimental data. Students will be expected to have taken a graduate course in regression or being taking a graduate course in regression concurrently.

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, Markov chains, Markov chain Monte Carlo (MCMC), Poisson processes, birth and death processes, renewal processes, and queueing processes. 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. Students will be expected to have taken a graduate course in regression or being taking a graduate course in regression concurrently. Knowledge of R or similar statistical packages is advised.

36-734 Methods for Clinical Trials
Intermittent: 6 units
This course will introduce the fundamentals 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.

36-750 Statistical Computing
Fall: 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., 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
Department of Statistics and Data Science Courses

36-759 Statistical Models of the Brain
Intermittent: 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 quantum 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.

36-762 Data Privacy
Fall: 6 units
Protection of individual data is a growing problem due to the large amount of sensitive and personal data being collected, stored, analyzed, and shared across multiple domains and stakeholders. Researchers are facing new policies and technical requirements imposed by funding agencies on accessing and sharing of the research data. This course will introduce students to (1) key principles associated with the concepts of confidentiality and privacy protection, and (2) techniques for data sharing that support useful statistical inference while minimizing the disclosure of sensitive personal information. Methodologies to be considered will include tools for disclosure limitation used by government statistical agencies and those associated with the approach known as differential privacy which provides a formal privacy guaranteed. Students will explore specific techniques using special tools in R.

36-763 Multilevel and Hierarchical Models
Fall: 6 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 to 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 these models using a variety of data sets and examples.

36-765 Writing in Statistics
Intermittent: 6 units
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 such that it makes 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.).

36-771 Martingales 1: Concentration Inequalities, The Basics
Intermittent: 6 units
Martingales are a central topic in statistics, but are even more relevant today due to modern applications to sequential learning and decision-making problems. This course will present a unified derivation of a wide-variety of new and old concentration inequalities for martingales. We will prove inequalities for scalars and matrices, that hold under a wide variety of nonparametric assumptions. For example, we will encounter exponential concentration inequalities for martingales whose increments have heavy-tails, for continuous-time martingales, and for martingales in general. This course will be a pre-requisite for the second mini, which focuses more on applications.

36-772 Martingales 2: Concentration Inequalities, Applications to Sequential Analysis
Intermittent: 6 units
This second mini will focus on deriving guarantees for a variety of important problems in sequential analysis using the tools developed in the first mini, as well as new tools such as uniform nonasymptotic versions of the law of the iterated logarithm for scalars and matrices. Applications include sequential analogs of the t-test, that are valid without a Gaussian assumption, best-arm identification in multi-armed bandits, average treatment effect estimation in sequential clinical trials, sequential covariance matrix estimation, and other such problems.

36-775 Data Ethics & Responsible Conduct of Research
Intermittent: 3 units
TBD

36-777 Multivariate Analysis I
Intermittent: 6 units
This is the second part of a semester long course on multivariate analysis. The aim of the class is to provide fundamental tools in understanding multivariate (including high dimensional) data. In this MINI we will study in detail the multivariate Gaussian distribution, the Wishart and Hotelling distributions. Time permitting we will cover principal component analysis (PCA) as well as discriminant analysis.

36-778 Multivariate Analysis II
All Semesters: 6 units
This is the second part of the multivariate analysis class. This MINI will discuss asymptotic inequalities for eigenvalues of Gaussian matrices, quadratic form concentration inequalities, and matrix estimation (including multivariate regression, covariance matrix estimation, PCA). Time permitting the class might also cover dimension reduction and graphical models.

36-779 Topics in Modern Multivariate Analysis II
Intermittent: 6 units
This is the second part of a semester-long course on modern multivariate analysis. In this MINI we will introduce recent research results focusing on high dimensional multivariate analysis. Topics include high dimensional mean and covariance testing, kernel based methods, structured high dimensional subspace estimation (sparse PCA, functional data), and network data.

36-791 Central Limit Theorem in High-Dimensions
Intermittent: 6 units
TBD
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.