Department of Statistics and Data Science Courses

Note on Course Numbers

Each Carnegie Mellon course number begins with a two-digit prefix which designates the department offering the course (76-xxx courses are offered by the Department of English, etc.). Although each department maintains its own course numbering practices, typically the first digit after the prefix indicates the class level: xx-1xx courses are freshmen-level, xx-2xx courses are sophomore level, etc. xx-6xx courses may be either undergraduate senior-level or graduate-level, depending on the department. xx-7xx courses and higher are graduate-level. Please consult the Schedule of Classes (https://enr-apps.as.cmu.edu/open/SOC/SOCServlet) each semester for course offerings and for any necessary pre-requisites or co-requisites.

36-200 Reasoning with Data
Spring: 9 units
This course will serve as an introduction to learning how to "reason with data". While still an introductory-level course in the Statistics Department, the focus will be more on thinking about the relationship between the application and the data set and extracting useful statistical information rather than taking primarily a formula-driven approach. There will be an emphasis on thinking through a empirical research problem from beginning to end. Types of data will include continuous and categorical variables, images, text, and networks. Applications will largely drawn 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 weekly computer labs for additional hands-on practice. This course is the credit-equivalent and substitute for 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 36-201, 36/70-207, 36-220, 36-247, or any 300- or 400-level Statistics course.

36-201 Statistical Reasoning and Practice
All Semesters: 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

36-202 Methods for Statistics and Data Science
Spring: 9 units
This course builds on the principles and methods of statistical reasoning developed in 36-201 (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, 36-309. 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: 36-201 or 70-207 or 36-247 or 36-220 or 36-207

Course Website: http://www.stat.cmu.edu/academics/courselist

36-207 Probability and Statistics for Business Applications
Fall: 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

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: (21-120 or 21-112) and (36-220 or 36-201 or 70-207 or 36-207 or 36-247) and (73-100 or 73-102)

Course Website: 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 21-122 or 21-123 or 21-112 or 21-256

Course Website: 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: 21-121 or 21-120 or 21-112

Course Website: 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 and 36-410. Not open to students who have received credit for 36-217 or 36-625.

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, 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: 21-325 Min. grade C or 36-217 Min. grade C or 36-225 Min. grade C or 15-359 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, 36-207/70-207, 36-220, or 36-226.
Prerequisites: 21-121 or 21-120 or 21-112

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 73-261 or 70-208 or 36-625 or 36-208 or 36-226 or 36-202 or 36-225 or 88-250

Course Website: http://www.stat.cmu.edu/academics/courselist

36-304 Biostatistics
Fall: 9 units
TBD

36-309 Experimental Design for Behavioral and 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 36-217 or 36-220 or 36-247 or 36-201

Course Website: http://www.stat.cmu.edu/academics/courselist

36-314 Biostatistics
Fall: 9 units
Tbd

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: 36-202 or 36-208 or 36-226 or 88-250 or 36-309 or 36-625 or 70-208 or 36-225 or 36-303

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 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: 15-359 Min. grade A or 36-217 Min. grade A or 21-325 Min. grade A or 36-225 Min. grade A

36-350 Statistical Computing
Fall: 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. Pre-requisites: (36-202, 36-208, or 36-309), plus ("computing at Carnegie Mellon" or consent of instructor) and 36-225 co-requisite.
Prerequisites: 36-315 or 36-226 or 36-303 or 36-309 or 70-208 or 36-208 or 36-202

Course Website: http://www.stat.cmu.edu/academics/courselist

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 minimum grade of C in any one of the pre-requisites is required. A grade of C is required to move on to 36-402 or any 36-46x course.
Prerequisites: (36-226 Min. grade C or 36-326 Min. grade C or 36-625 Min. grade C) and (21-241 or 21-240)

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: 36-625 or 36-217 or 36-225 or 21-325

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

36-459 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 brainsciences 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://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 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

36-468 Special Topics
Intermittent: 9 units
TDB

36-490 Undergraduate Research
Spring: 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. Eligible students will receive information about the application processes for this course early in the fall.
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 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

36-494 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

36-625 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

36-665 Special Topics
Intermittent: 9 units
TDB

36-668 Tbd
Intermittent: 9 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.

36-700 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.

36-721 Statistical Graphics and Visualization
Intermittent: 6 units
Graphical displays of quantitative information take on many forms to 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. The class will also cover some principles of visual perception and estimation. We will start with univariate and bivariate data, looking at some commonly used graphs and, after discussing their advantages/disadvantages, then turning to more sophisticated tools. We will then explore some three-dimensional tools, group structure/clustering, and projections of higher dimensional data. As time permits, the course will consider some more advanced graphical models such as statistical maps, networks, and the usage of icons.

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).

36-762 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.

36-765 Writing in Statistics
Intermittent: 3 units
TBD

36-777 Topics in Modern Multivariate Analysis I
Intermittent: 6 units
This is the first part of a semester-long course on modern multivariate analysis. In this MINI we will cover basic concepts about random vectors, multivariate Gaussian, and inference tools such as mean and covariance estimation and testing, multivariate analysis of variance, discriminant analysis, principal components analysis, and, if time permits, canonical correlation analysis, clustering analysis. Relevant matrix algebra results will be emphasized as a useful tool.

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.