DATA SCIENCE (DS-GA)

DS-GA 1001 Introduction to Data Science (3 Credits)
Typically offered Fall
This course is designed as a survey course to give an overview of how Data Science will allow you to learn and gain insights from data. We introduce the foundational concepts and principles in the field of Data Science - specifically inference and machine learning - to foster a mindset that will unlock the more advanced courses in the program. Moreover, we aim to impart the technical skills to analyze real world datasets. Special focus will be put on building intuitions about the Data Science approach to solving problems and making decisions.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1002 Probability and Statistics for Data Science (3 Credits)
Typically offered Fall
This course introduces fundamental concepts in probability and statistics from a data-science perspective, providing examples with real data. The aim is to become familiarized with probabilistic models and statistical techniques that are widely used in data analysis.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1003 Machine Learning (3 Credits)
Typically offered Spring
This required course for the MS in Data Science should be taken in the first year of study. It covers a wide variety of topics in machine learning, pattern recognition, statistical modeling, and neural computation. It covers the mathematical methods and theoretical aspects, but primarily focuses on algorithmic and practical issues.
Grading: GSAS Graded
Repeatable for additional credit: No
Prerequisites: (Masters or Doctoral student) and (DS-GA 1001 or DS-GA 2003).

DS-GA 1004 Big Data (3 Credits)
Typically offered Spring
Big Data provides unprecedented opportunities to gain insights and to train models. However, this is a double-edged sword, as the scale of truly big datasets also poses new challenges. This class is designed to address problems of data scaling by introducing principles and tools related to distributed computing and distributed storage as well as algorithms and storage schemes that scale well with larger samples. We will do so by using frameworks like Hadoop, Spark and Dask to wrangle big datasets found in the wild.
Grading: GSAS Graded
Repeatable for additional credit: No
Prerequisites: (Masters or Doctoral student) and (DS-GA 1001 or DS-GA 2003).

DS-GA 1005 Inference and Representation (3 Credits)
Typically offered Fall
This course covers graphical models, causal inference, and advanced topics in statistical machine learning.
Grading: GSAS Graded
Repeatable for additional credit: No
Prerequisites: (Masters or Doctoral student) and (DS-GA 1003).

DS-GA 1006 Capstone Project and Presentation (3 Credits)
Typically offered Fall
The purpose of the capstone project is to make the theoretical knowledge acquired by the students operational in realistic settings. During the project, students see through the entire process of solving a real-world problem: from collecting and processing real-world data, to designing the best method to solve the problem, and implementing a solution. The problems and datasets come from real-world settings identical to what the student would encounter in industry, government, or academic research.
Grading: GSAS Graded
Repeatable for additional credit: No
Prerequisites: DS-GA 1001 OR 2003) AND DS-GA 1003 AND DS-GA 1004) AND (Data Science MS OR PhD OR Non-Degree.

DS-GA 1007 Programming for Data Science (3 Credits)
Typically offered Fall
The class will teach students about programming for applications in data science. Students will study the Python language and packages including tools for array operations, table manipulations, time series analysis, visualization, and data extraction. Through a focus on examples, students will learn about query languages, version control systems, and web frameworks. Experience with debugging, testing and documenting programs will enable students to code in integrated development environments and command line interfaces. Generative AI technologies for code completion and question-answering will also be introduced to complement and extend the programming best practices acquired in class.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1008 Deep Learning (3 Credits)
Typically offered Spring
This course concerns the latest techniques in deep learning and representation learning, focusing on supervised and unsupervised deep learning, embedding methods, metric learning, convolutional net and recurrent nets, with applications to computer vision, natural language understanding, and speech recognition. The pre-requisites include DS-GA 1001 Introduction to Data Science and DS-GA 1003 Machine Learning
Grading: GSAS Graded
Repeatable for additional credit: No
Prerequisites: DS-GA 1001 or DS-GA 2003 And 1003.

DS-GA 1009 Practical Training for Data Science (3 Credits)
Typically offered Fall, Spring, and Summer terms
This course provides data science students with an opportunity to apply the knowledge gained in the course work to one or more practical problems in industry, medicine, government, or research. Students can only take this course at most twice.
Grading: GSAS Graded
Repeatable for additional credit: Yes

DS-GA 1010 Independent Study in Data Science (1-3 Credits)
Typically offered Fall, Spring, and Summer terms
This independent study course provides students with the opportunity to work one-on-one with a faculty member on a particular topic or project. The learning objective of the course is to build or to strengthen data science skills through focus on specific issues of interest to the student.
Grading: GSAS Graded
Repeatable for additional credit: Yes

DS-GA 1011 Fundamentals of Natural Language Processing (3 Credits)
How can machines understand and use human languages? This course examines modern computational approaches based on representation learning for processing, understanding, and using human language. These include vector-space models of word meaning, neural network-based deep learning methods, and large language models. Together, they will give students the tools to build state-of-the-art models for language-based applications like machine translation and dialogue systems.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1012 Large Language Models: Evaluation and Applications (3 Credits)
Contemporary language models can conduct conversations with users and perform tasks that go far beyond their historical role of predicting the next word. As these abilities are not explicitly engineered into the architecture, but rather emerge from large-scale training, it is often unclear how the models accomplish those tasks. In this course, we will survey the tasks, methods and metrics that are used to evaluate and understand language models, and to compare these models to humans. Some of the areas of evaluation we will discuss include factuality, pragmatics, reasoning, fairness and safety. We will also discuss interpretability methods that aim to explain the internal mechanisms that underlie the models' behavior. A major aim of the course is to prepare students to do original research in this area, culminating with a substantial final project.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1013 Mathematical Tools for Data Science (3 Credits)
This course provides a rigorous introduction to mathematical tools for data science drawn from linear algebra, harmonic analysis, probability theory, and convex analysis. The main topics are the singular-value decomposition (SVD), the Fourier series, randomized projections, the randomized SVD, convex optimization, duality theory and nonconvex optimization. The material is motivated by multiple data-analysis applications including dimensionality reduction, collaborative filtering, sound and image processing, magnetic-resonance imaging, sparse regression, compressed sensing, and topic modeling.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1014 Optimization and Computational Linear Algebra (3 Credits)
This proof-based course covers the fundamentals of computational linear algebra and optimization used in Data Science. About two thirds of the lectures will be about linear algebra and the remaining third about convex optimization. We will first go over basic linear algebra: vector spaces, linear transformations, rank, norms and inner products, eigenvalues and eigenvectors, building up the singular value decomposition (SVD), which is a cornerstone of many numerical applications: PCA and dimensionality reduction, Markov chains and PageRank, spectral clustering in graphs, linear regression. Lastly, we will go over convex functions, optimality conditions in constrained optimization, and gradient descent.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1015 Text as Data (3 Credits)
Course introduces students to quantitative texts-as-data analysis from an applied perspective, with a focus on political science. Course covers, inter alia, metrics of association between texts, burstiness of words and concepts, measurement of complexity and readability, scaling of political texts, automatic event extraction, dictionary methods for estimating sentiment, clustering, Latent Semantic Analysis, machine learning applications, topic models and LDA. Course also includes special topics such as the estimation of personal characteristics from writings, 'stylometrics' and detection of false statements. Course assumes no prior knowledge of texts-as-data work, though it requires proficiency with programming languages such as R or Python, along with an understanding of elementary statistical theory and regression analysis.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1016 Computational Cognitive Modeling (3 Credits)
This course provides a survey of computational approaches to understanding human intelligence and cognition. Both psychologists and data scientists are working with increasingly large quantities of human behavioral data. Computational cognitive modeling is the project of understanding behavioral data (and the mind and brain, more generally) by building computational models of the cognitive processes that produce the data. The course will cover the goals, philosophy, and technical concepts behind computational cognitive modeling, including model fitting and evaluation. Ideally, students will leave the course with a richer understanding of how computational modeling advances cognitive science, how cognitive science can inform research in machine learning and artificial intelligence, and how to fit and evaluate cognitive models for understanding behavioral data.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1017 Responsible Data Science (3 Credits)
The first wave of data science focused on accuracy and efficiency – on what we can do with data. The second wave focuses on responsibility – on what we should and shouldn't do. Irresponsible use of data science can cause harm on an unprecedented scale. Algorithmic changes in search engines can sway elections and incite violence; irreproducible results can influence global economic policy; models based on biased data can legitimize and amplify racist policies in the criminal justice system; algorithmic hiring practices can silently and scalably violate equal opportunity laws, exposing companies to lawsuits and reinforcing the feedback loops that lead to lack of diversity. Therefore, as we develop and deploy data science methods, we are compelled to think about the effects these methods have on individuals, population groups, and on society at large. Responsible Data Science is a technical course that tackles the issues of ethics, legal compliance, data quality, algorithmic fairness and diversity, transparency of data and algorithms, privacy, and data protection.
Grading: GSAS Graded
Repeatable for additional credit: No
Prerequisites: (Masters or Doctoral student) and (DS-GA 1001 or DS-GA 2003).

DS-GA 1018 Probabilistic Time Series Analysis (3 Credits)
This course presents fundamental tools for characterizing data with statistical dependencies over time, and using this knowledge for predicting future outcomes. These methods have broad applications from econometrics to neuroscience. The course emphasizes generative models for time series, and inference and learning in such models. We will cover a range of approaches including Kalman Filter, HMMs, ARMA, Gaussian Processes, RNNs, Transformers, and their application to several kinds of data.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1019 Advanced Python for Data Science (3 Credits)
Typically offered Fall and Spring
This course explores advanced techniques to enhance the performance and scalability of Python programs. Key topics include parallel computing, GPU acceleration, and big data processing using industry-standard frameworks such as Apache Hadoop and Apache Spark. Emphasizing practical application, students will gain hands-on experience developing high-performance Python solutions to real-world problems. The course adopts a student-centered, active learning approach. Each session typically begins with a concise overview of new programming concepts, followed by guided, hands-on coding exercises designed to reinforce and apply those techniques in practice.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1020 Mathematical Statistics (3 Credits)
Typically offered Fall and Spring
This course provides rigorous tools for the mathematical analysis of statistical procedures in data science. Topics include hypothesis testing, confidence sets, regression, classification, and non-parametric statistics. We will focus both on classical asymptotic theory and on modern non-asymptotic techniques and theorems suitable for data science applications. Prerequisites: linear algebra, probability, comfort with mathematical proofs.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 1021 Probability and Statistics 2 (3 Credits)
This course is a continuation of Probability and Statistics for Data Science. It builds upon the contents covered in Probability and Statistics for Data Science to provide a self contained mathematically rigorous description of more advanced concepts from probability and statistics emphasizing their application in practice. These include correlation, the law of large numbers, the central limit theorem, confidence intervals, the bootstrap, hypothesis testing, principal component analysis, low rank models, regression and classification. The course is specifically designed to complement other courses in the Data Science curriculum.
Grading: GSAS Graded
Repeatable for additional credit: No
Prerequisites: (Masters or Doctoral student) and (DS-GA 1002).

DS-GA 1170 Fundamental Algorithms (3 Credits)
Reviews a number of important algorithms, with emphasis on correctness and efficiency. The topics covered include solution of recurrence equations, sorting algorithms, selection, binary search trees and balanced-tree strategies, tree traversal, partitioning, graphs, spanning trees, shortest paths, connectivity, depth-first and breadth-first search, dynamic programming, and divide-and-conquer techniques. Prerequisites: At least one year of experience with a high-level language such as Pascal, C, C++, or Java; and familiarity with recursive programming methods and with data structures (arrays, pointers, stacks, queues, linked lists, binary trees).
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 2001 Research Rotation (1-3 Credits)
Typically offered Fall and Spring
The research rotation course gives PhD students exposure to the research being conducted by CDS faculty. The objective of this course is to broaden students' perspective and make them better rounded data science researchers. During this semester-long course, students will design and carry out original research in a collaborative setting with faculty who will monitor progress on a weekly basis and assign a pass/fail grade at the end of the semester and submit a brief report to the DGS.
Grading: GSAS Pass/Fail
Repeatable for additional credit: Yes

DS-GA 2002 Communication Skills (1 Credit)
This course is a 7-week course for CDS students, particularly PhD students, consisting of two separate and discrete Short Course components -- a 4-week Academic Writing course and a 3-week Great Presentations course. The Academic Writing course component is an intensive introduction to the principles of excellent rhetorical writing with a focus on the development of a clear, interesting, and rigorous science text, the construction of logical arguments, and the reporting of data, as well as the important concepts including reader-oriented writing, genre, precision, tone, the composing process, and strategies useful for redrafting and editing. Some of the sub-genres we analyze and practice include introductions, data commentaries, results/discussion, conclusions, and abstracts. We also practice other professional texts including requests for funding and professional email texts. The fundamental principles discussed and practiced in the Great Presentations course component can be applied in a variety of contexts including the short research talk, a lab talk, a formal conference presentation, poster presentation, job talk, interview, industry pitch. We talk about how to construct a logical and interesting presentation story, the design and best use of visuals, transitioning through the story and visuals, fluent delivery, connecting with the audience, timing, coordination of movement with content, and key linguistic elements such as volume, pitch range, intonation, and ends of utterances. We also look at the art of asking and responding to questions. Students will give talks of varying lengths followed by detailed feedback from the instructor.
Grading: GSAS Pass/Fail
Repeatable for additional credit: No

DS-GA 2003 Introduction to Data Science (3 Credits)
Data Science is a new discipline. This brings challenges and opportunities. On the one hand, its boundaries are poorly defined; on the other hand, this very fact provides unusual latitude for us to establish its intellectual core in real time. This course is about that core. That is, we aim to provide students with an overview of Data Science as an endeavor: its origin, scope, techniques, debate and future. Consequently, this course is both broad and deep. It is broad in that it covers considerable ground in statistics and computer science in a short space of time. The central goal therein is to provide a basic and common vocabulary for students coming from multiple disciplines, enabling them to understand work in the field and to communicate their own work. The course will provide practical hands-on experience with Python and its associated data analysis libraries for this purpose. The course is deep in that we will cover some fundamental Data Science ideas in some detail—both technically and philosophically. This extends to the ethics of Data Science work.
Grading: GSAS Graded
Repeatable for additional credit: No

DS-GA 3001 Special Topics in Data Science (3 Credits)
Typically offered Fall and Spring
This course is always offered in sections, each section around a special topic. The special topics will vary from time to time depending on the availability of suitable instructors. The format will vary by the topic, but will usually include an introduction to the topic and an overview of advanced research in the topic.
Grading: GSAS Graded
Repeatable for additional credit: Yes