Data Science

Program of Study

The technological revolution has led to an explosion of data in domains of knowledge including medicine, policy, social sciences, commerce, and the natural sciences. Petabytes of data are being collected from a myriad of instruments, like sequencing machines for genomics and mobile devices for quantifying social interactions. In addition to driving research, data are shaping the way people work, live, and communicate. Correspondingly, new methodologies have emerged to power intelligent systems, make more accurate predictions, and gain new insight using the large volumes of data generated by scientists, entrepreneurs, and analysts.

The field of Data Science has emerged to respond to the data revolution, and the necessity to responsibly store, process, analyze, and interpret data. Transdisciplinary by nature, Data Science draws on numerous fields including statistics, computer science and applied mathematics, and also incorporates topics in privacy and ethics, philosophy of science, and economics to better understand the impact of data on society. Therefore, the Data Science curriculum combines computational and analytical skills, extensive knowledge in the domain of application, communication skills, and an emphasis on ethical considerations.

The data science program offers BA and BS degrees, as well as a minor. The minor program in data science is intended to equip students with computational and analytical comprehension and tools that will allow them to work on a variety of data-driven problems in any discipline while emphasizing important issues in data privacy, ethics, and communication. The major has several tracks that can serve different interests. Students have the possibility to pursue a program suited for a double major, and/or suited for graduate studies in data science and related fields. Students majoring in Data Science as well as those who are majoring in other fields of study and want to complete a minor in Data Science are encouraged to discuss their course choices with the Assistant Director of Undergraduate Data Science Studies. Information on the minor follows the description of the major.

Requirements

Students majoring in Data Science must meet the general education requirement in mathematical sciences with courses in calculus. (See summary of requirements table.)

The program for the BA in Data Science consists of 15 courses (1500 units) beyond the general education requirement. The standard curriculum includes an introductory sequence to Data Science and programming, an introductory sequence to mathematical methods, and courses on ethics in data science, data visualization, data engineering, and machine learning. In addition, the curriculum includes exposure to real-world projects through the Data Science Clinic and three electives. The BA track is recommended for students who would like to combine a Data Science major with another minor or major program.

Students interested in mathematically advanced courses and/or graduate research in Data Science can choose the Theory Track, where four MATH and STAT courses replace the mathematical methods sequence. (See Theory Track below.)

Students interested in emphasizing computational aspects of Data Science can choose the Computation Track, which requires the Introduction to Computer Science sequence. (See Computation Track below.)

The program for the BS in Data Science consists of 18 courses (1800 units) beyond the general education requirement. In addition to the BA requirements, students pursuing the BS must meet the following two requirements:

(i) A coherent set of three courses in an independent domain of knowledge to which Data Science can be applied. This set can be in the natural sciences, social sciences, or humanities and sequences in which earlier courses are prerequisites for advanced ones are encouraged. Each of the three should be a course at the 20000 level or higher that counts toward a major in the natural sciences, social sciences, or humanities. Courses in STAT, CMSC, or MATH or courses that focus on methods (mathematical, computational, or statistical) may not be used for this requirement.

(ii) One of the required electives must be in the field of Machine Learning or its applications, e.g., CMSC 25025 Machine Learning and Large-Scale Data Analysis, or CMSC 25440 Machine Learning in Medicine .

Data Science Clinic

The Data Science Clinic is a two-quarter, experiential, project-based sequence where students work in teams as data scientists with real-world clients under the supervision of instructors. The course will serve as a capstone experience that will partner with public interest organizations, industry, NGOs, government agencies, and cutting-edge researchers to source challenging, mission-driven data science projects for the clinic. Teams will vary in size depending on the type and scope of the projects. Students will work with imperfect datasets, apply models and algorithms to real-world data, navigate security and privacy issues, and communicate results to a diverse set of stakeholders, as well as engage in the broader impacts of their work. Teams will be tasked with producing key deliverables, such as data analysis, prototype data science products, and open source software, as well as final client presentations, reports, and manuscripts.

Summary of Requirements for the Major in Data Science

GENERAL EDUCATION

One of the following:		200
MATH 13100-13200	Elementary Functions and Calculus I-II	
MATH 15100-15200	Calculus I-II	
MATH 16100-16200	Honors Calculus I-II	
MATH 16110-16210	Honors Calculus I-II (IBL)	
Total Units		200

BA IN DATA SCIENCE

Data Science		400
DATA 11800 & DATA 11900	Introduction to Data Science I and Introduction to Data Science II	
DATA 25900	Ethics, Fairness, Responsibility, and Privacy in Data Science	
DATA 23700	Visualization for Data Science	
or CMSC 23900	Data Visualization	
Data Science Clinic		200
DATA 27100-27200	Data Science Clinic I-II	
Machine Learning (one of the following)		100
DATA 23100	Machine Learning Fundamentals: Theory and Practice [‡]	
CMSC 25025	Machine Learning and Large-Scale Data Analysis	
CMSC 25300	Mathematical Foundations of Machine Learning	
CMSC 25400	Machine Learning [‡]	
Computer Science		200
DATA 12000	Computer Science for Data Science	
DATA 13600	Introduction to Data Engineering	
Mathematics and Statistics		300
DATA 21100	Mathematical Methods for Data Science I	
DATA 21200	Mathematical Methods for Data Science II	
DATA 21300	Models in Data Science	
or STAT 24320	Applications in Numerical Linear Algebra	
Three Electives		300
Total Units		1500

BS IN DATA SCIENCE

Requirements for BA in Data Science [+]	1500
A coherent set in an independent domain of knowledge	300
Total Units	1800

[+] One of the electives must be in the field of Machine Learning or its applications.

[‡] CMSC 25400 and DATA 23100 cannot both count toward the Data Science major or minor. Likewise, CMSC 25400 and DATA 22100 cannot both count toward the Data Science major or minor.

Please note: Prerequisites for (almost) all CMSC courses include CMSC 14200 Introduction to Computer Science II. Typically, these prerequisites are not waived without a placement exam that ensures that students have the necessary background. Please consult the lists of prerequisites well in advance so as to have time to adequately plan your registration.

Electives

Electives in Data Science could come from the list of representative approved courses below. Additional courses with a strong Data Science focus can be approved by the Assistant Director of Undergraduate Data Science Studies or the Director of Undergraduate Studies for Data Science. These may include appropriate graduate-level Data Science courses. A successful petition requires students to obtain approval from the Assistant Director of Undergraduate Data Science Studies or the Director of Undergraduate Studies, who will contact College Advising on the student's behalf.

Data Science	
DATA 24100	Software Engineering for Data Science

DATA 26100	Statistical Pitfalls and Misinterpretation of Data
Statistics	
STAT 22200	Linear Models and Experimental Design
STAT 24400	Statistical Theory and Methods I
STAT 24500	Statistical Theory and Methods II ‡
STAT 24620	Multivariate Statistical Analysis: Applications and Techniques
STAT 25100	Introduction to Mathematical Probability
STAT 26300	Introduction to Statistical Genetics
STAT 27400	Nonparametric Inference
STAT 27725	Machine Learning
STAT 28000	Optimization
Computer Science	
CMSC 22240	Computer Architecture for Scientists
CMSC 23300	Networks and Distributed Systems
CMSC 23310	Advanced Distributed Systems
CMSC 23500	Introduction to Database Systems
CMSC 25025	Machine Learning and Large-Scale Data Analysis
CMSC 25040	Introduction to Computer Vision
CMSC 25300	Mathematical Foundations of Machine Learning
CMSC 25400	Machine Learning ‡
CMSC 25440	Machine Learning in Medicine
CMSC 25610	Undergraduate Computational Linguistics
CMSC 27620	Introduction to Bioinformatics
Biological Sciences	
BIOS 21216	Introduction to Statistical Genetics
BIOS 28407	Genomics and Systems Biology
BIOS 29331	Clinical Research Design and Interpretation of Health Data
Business	
BUSN 20810	Machine Learning
Public Policy	
PBPL 28829	Artificial Intelligence for Public Policy (Public Policy)
Psychology	
PSYC 26010	Big Data in the Psychological Sciences
Sociology	
SOCI 20519	Spatial Cluster Analysis

‡ Only one of CMSC 25400, DATA 23100, or STAT 24500 can be counted toward the Data Science major or minor.

Theory Track

Data Science in and of itself is a field on the frontier of scientific inquiry. Students interested in mathematically advanced courses and/or graduate research in Data Science can choose to emphasize the mathematical foundations of Data Science. The theoretical curriculum will replace four courses:

- DATA 21100 Mathematical Methods for Data Science I: STAT 25100 Introduction to Mathematical Probabilityis approved as a substitute. Please make note of the appropriate prerequisites needed to enroll in this course on the Statistics Catalog page.
- DATA 21200 Mathematical Methods for Data Science II: MATH 19620 Linear Algebra or STAT 24300 Numerical Linear Algebra are acceptable substitutes.
- DATA 21300 Models in Data Science*: STAT 24400 Statistical Theory and Methods I, when STAT 24500 Statistical Theory and Methods II is also taken, is approved as a substitute.
- DATA 23100 Machine Learning Fundamentals: Theory and Practice*: STAT 24500 Statistical Theory and Methods II, when STAT 24400 Statistical Theory and Methods Iis also taken, is approved as a substitute.

*This substitution is approved as a set. Both STAT 24400 Statistical Theory and Methods I and STAT 24500 Statistical Theory and Methods II must be taken for these courses to be approved as a substitution.

Please make note of the appropriate prerequisites needed to enroll in these courses on the Statistics Catalog page.

In addition, students opting for this curriculum will be required to take at least one elective in the field of Machine Learning. Combined, these courses satisfy the prerequisites for a wide selection of mathematically advanced courses in Data Science, Statistics, and Mathematics. Substitutions made within tracks are transferable across different tracks and are applicable, even if a student chooses not to pursue a particular track.

Computation Track

DATA 12000 Computer Science for Data Science could be replaced with CMSC 14200 Introduction to Computer Science II. Conditional on a minimum grade of C+ in CMSC 14200 Introduction to Computer Science II, CMSC 14300 Systems Programming I and CMSC 14400 Systems Programming II could count as two Data Science electives. This will help accommodate students interested in emphasizing computational aspects of Data Science through advanced Computer Science electives such as CMSC 23310 Advanced Distributed Systems or CMSC 23500 Introduction to Database Systems. Substitutions made within tracks are transferable across different tracks and are applicable, even if a student chooses not pursue a particular track.

Double Majors (and Other)

The program makes it possible to pursue a double major in four years. Examples of possible combinations include majors in the Social Sciences, Humanities, Biological Sciences, Computer Science, or Statistics. Students interested in a double major are strongly encouraged to discuss their course plans and obtain advice from their academic advisers and the Assistant Director of Undergraduate Data Science Studies.

Students who have taken STAT 23400 Statistical Models and Methods may use it to replace DATA 21100 Mathematical Methods for Data Science I. Additionally, students who have taken MATH 19620 Linear Algebra or STAT 24300 Numerical Linear Algebra may use either of these courses to replace DATA 21200 Mathematical Methods for Data Science II.

Grading and Advising for the Major Program

Prospective majors are strongly encouraged to meet with the Assistant Director of Undergraduate Data Science Studies for a preliminary discussion of their plans. This preliminary meeting will address frequently asked questions and offer advice when appropriate. In contrast to the minor program, a consent form is not required for declaring the major.

Courses in the major must be taken for quality grades, and more than half of the requirements for the minor must be met by registering for courses bearing University of Chicago course numbers. Students who are majoring or minoring in Data Science must receive a quality grade of at least C in all of the courses counted toward their major or minor program in Data Science. Subject to College and divisional regulations, and with the consent of the instructor, students may register for either quality grades or for P/F grading in any 20000-level Data Science course, other than DATA 27100-27200 Data Science Clinic, that is not counted toward a major or minor in Data Science. A grade of P is given only for work of C- quality or higher. Students who earn at least a 3.6 GPA within the major qualify for graduating with honors.

The following policy applies to students who wish to receive a grade of Incomplete for a Data Science course: In addition to submitting the official Incomplete Form required by the College, students must have completed at least half of the total required course work with a grade of C- or better, and they must be unable to complete the remaining course work by the end of the quarter due to an emergency.

Minor in Data Science

The minor in data science targets students from all disciplines and consists of four required courses and two electives drawn from an approved list. Students may petition to take electives other than those listed below, if they can demonstrate substantial data science content in those courses. A successful petition requires students to obtain approval from the Assistant Director of Undergraduate Data Science Studies or the Director of Undergraduate Studies, who will contact College Advising on the student's behalf.

1. Introductory Sequence (four courses required):

DATA 11800	Introduction to Data Science I	100
DATA 11900	Introduction to Data Science II	100
One of the following:		100
CMSC 25300	Mathematical Foundations of Machine Learning	
or DATA 22100	Introduction to Machine Learning: Concepts and Applications	
DATA 25900	Ethics, Fairness, Responsibility, and Privacy in Data Science	100

2. Elective Sequence (two of the following courses required):

Two of the following:		200
DATA 13600	Introduction to Data Engineering	

DATA 22700	Data Visualization and Communication
or CMSC 23900	Data Visualization
CMSC 25025	Machine Learning and Large-Scale Data Analysis
or DATA 22100	Introduction to Machine Learning: Concepts and Applications
STAT 22200	Linear Models and Experimental Design

Grading and Advising for Minor Program

Courses in the minor may not be double-counted with the student's major(s) or with other minors. Courses in the minor must be taken for quality grades, and more than half of the requirements for the minor must be met by registering for courses bearing University of Chicago course numbers.

Prospective minors may meet with the Assistant Director of Undergraduate Data Science Studies to discuss their course plans and to obtain advice, and subsequently fill out a Consent to Complete a Minor Form (https://humanities-web.s3.us-east-2.amazonaws.com/college-prod/s3fs-public/documents/Consent_Minor_Program.pdf). The complete form should be sent to the Assistant Director of Undergraduate Data Science Studies for approval. Students should submit completed, signed forms to their College adviser by the end of Spring Quarter of their third year.

No courses in the minor can be double-counted with the student's major(s) or with other minors, nor can they be counted toward general education requirements.

Summary of Requirements for the Minor in Data Science

Introductory Sequence: Four courses	400
Electives: Two courses	200
Total Units	600

Data Science Courses

DATA 11800. Introduction to Data Science I. 100 Units.
Data science provides tools for gaining insight into specific problems using data, through computation, statistics and visualization. This course introduces students to all aspects of a data analysis process, from posing questions, designing data collection strategies, management+storing and processing of data, exploratory tools and visualization, statistical inference, prediction, interpretation and communication of results. Simple techniques for data analysis are used to illustrate both effective and fallacious uses of data science tools. Although this course is designed to be at the level of mathematical sciences courses in the Core, with little background required, we expect the students to develop computational skills that will allow them to analyze data. Computation will be done using Python and Jupyter Notebook.
Instructor(s): M.Hutch; K.Smalenberger; W. Trimble; J. Blanc; V. Shevade Terms Offered: Autumn Spring Winter
Equivalent Course(s): STAT 11800

DATA 11900. Introduction to Data Science II. 100 Units.
This course is the second quarter of a two-quarter systematic introduction to the foundations of data science, as well as to practical considerations in data analysis. A broad background on probability and statistical methodology will be provided. More advanced topics on data privacy and ethics, reproducibility in science, data encryption, and basic machine learning will be introduced. We will explore these concepts with real-world problems from different domains.
Instructor(s): A. Nussbaum; K. Sehgal; A. Kube Terms Offered: Autumn Spring Winter
Prerequisite(s): DATA 11800 or consent of instructor.
Equivalent Course(s): STAT 11900

DATA 12000. Computer Science for Data Science. 100 Units.
This course teaches computational thinking and programming skills to students in the Data Science program. Topics include control structures and basic data types, abstraction and functional decomposition, classes and objects in Python, basic algorithms, and an introduction to computer structure and low level representation of data types. Examples will include the application of tools such as scraping web pages and rudimentary machine learning to a variety of fields.
Instructor(s): A. Kube, A. Nussbaum Terms Offered: Spring Winter
Prerequisite(s): DATA 11800

DATA 13600. Introduction to Data Engineering. 100 Units.
Data-driven models are revolutionizing science and industry. Scalable systems are needed to collect, stream, process, and validate data at scale. This course is an introduction to "big" data engineering where students will receive hands-on experience building and deploying realistic data-intensive systems. It will cover streaming, data cleaning, relational data modeling and SQL, and Machine Learning model training. A core theme of the course is "scale," and we will discuss the theory and the practice of programming with large external datasets that cannot fit in main memory on a single machine. The course will consist of bi-weekly programming assignments, a midterm examination, and a final.
Instructor(s): W. Trimble Terms Offered: Autumn Spring Winter

Prerequisite(s): DATA 11900, DATA 12000 or CMSC 14200

DATA 13820. Data Science in Quantitative Finance and Risk Management. 100 Units.
Have you started or are about to start your investment journey? Do you want to know more about terms like "recession" and "volatility," and how they might affect your own bank account? Are you interested in mathematics and its application to human emotions? This course introduces the leading statistical models and methods which financial data researchers use to understand ever-evolving markets and build insightful financial strategies, such as machine learning, risk calculation, and portfolio management . At first, students will learn about the theoretical and applied foundations of regression and classification designs for predicting market patterns. Next, students will gain exposure to proprietary metrics such as Value-at-Risk(VaR) used to evaluate returns/losses of both single and multi-asset portfolios. Lastly, they will experiment with portfolio allocation tactics by visualizing risk-to-reward graphs under various buying and selling conditions. These techniques can be applied to the U.S. and foreign asset classes, including equities, commodities, and cryptocurrencies. Students will experience how professionals in quantitative trading, hedge funds, and risk analytics collaborate to pitch asset strategies to their clients, and form research teams to play a stock market game using the skills they learned throughout the course with the objective of maximizing the teams' portfolio returns. All implementations will be done using Python.
Terms Offered: Summer
Equivalent Course(s): STAT 13820

DATA 20002. Cognitive Models. 100 Units.
A foundational principle of cognitive science is that the workings of cognitive systems--whether biological, mechanical, or digital--can be productively represented by the operation of formal computational models. This course provides a survey of popular modeling frameworks (such as Bayesian rational agents, connectionist networks, dynamical systems, etc.), as well as the cognitive phenomena that these models have been used to simulate. We will discuss the theoretical commitments of these models, assess strengths and weaknesses of each framework for addressing different types of cognitive questions, and analyze the implications of these models' successes and failures for our understanding of the mind.
Instructor(s): Yu Ji, Eugene Terms Offered: Spring
Equivalent Course(s): PSYC 22002, COGS 20002, LING 20002, LING 30002

DATA 20027. Quantitative Research Methods in Linguistics. 100 Units.
This course provides an introduction to how quantitative methods are used in the analysis of linguistic data. This will include a foundation in statistical methods that can be applied to experimental and psycholinguistic data, including probability theory, hypothesis testing, regression models and the use of Bayesian statistics. Further topics will include a brief introduction to the use of basic machine learning algorithms in linguistic research and techniques that can be used in the analysis of large linguistic datasets. The class will be grounded in case studies from a variety of subfields of linguistics and provide hands-on examples through a guided introduction to programming. This class is intended for students who are interested in jump-starting a path into linguistic data science and is designed to be accessible to those with no experience in data science or programming.
Instructor(s): Craig Thorburn Terms Offered: Autumn
Equivalent Course(s): LING 22500, COGS 20027, LING 32500

DATA 20195. Data and the State. 100 Units.
In order to regulate and govern, states must marshall data: the whos, whats, and wheres of a nation get rendered in records and statistics, at least approximately. One of the most fundamental instruments for the "datafication" of the population is a census - and the very first Article of the U.S. Constitution calls for one to be conducted every ten years. That same sentence is also the source of the notorious "Three-Fifths Compromise" establishing an enumeration rule where enslaved people count fractionally as much as free people: political arithmetic. In this course we will use Census data as a jumping-off point for an investigation of data practices of governance, looking at a host of frameworks and mechanisms to classify, measure, and locate. Students will learn to use a suite of modern packages for geospatial data science in Python, including Pandas, GeoPandas, Shapely, and Matplotlib, and readings will include texts from geography, anthropology, sociology, and science and technology studies. No prior knowledge assumed.
Equivalent Course(s): DATA 30195

DATA 20519. Spatial Cluster Analysis. 100 Units.
This course provides an overview of methods to identify interesting patterns in geographic data, so-called spatial clusters. Cluster concepts come in many different forms and can generally be differentiated between the search for interesting locations and the grouping of similar locations. The first category consists of the identification of extreme concentrations of locations (events), such as hot spots of crime events, and the location of geographical concentrations of observations with similar values for one or more variables, such as areas with elevated disease incidence. The second group consists of the combination of spatial observations into larger (aggregate) areas such that internal similarity is maximized (regionalization). The methods covered come from the fields of spatial statistics as well as machine learning (unsupervised learning) and operations research. Topics include point pattern analysis, spatial scan statistics, local spatial autocorrelation, dimension reduction, as well as spatially explicit hierarchical, agglomerative and density-based clustering. Applications range from criminology and public health to politics and marketing. An important aspect of the course is the analysis of actual data sets by means of open source software, such as GeoDa, R or Python.

Instructor(s): L. Anselin and P. Amaral Terms Offered: Winter
Prerequisite(s): STAT 22000 or equivalent; SOCI 20253/30253 (or equivalent) Introduction to Spatial Data Science required.
Equivalent Course(s): GISC 30519, MACS 30519, SOCI 20519, GISC 20519, SOCI 30519, MACS 20519

DATA 20559. Spatial Regression Analysis. 100 Units.
This course covers statistical and econometric methods specifically geared to the problems of spatial dependence and spatial heterogeneity in cross-sectional data. The main objective for the course is to gain insight into the scope of spatial regression methods, to be able to apply them in an empirical setting, and to properly interpret the results of spatial regression analysis. While the focus is on spatial aspects, the types of methods covered have general validity in statistical practice. The course covers the specification of spatial regression models in order to incorporate spatial dependence and spatial heterogeneity, as well as different estimation methods and specification tests to detect the presence of spatial autocorrelation and spatial heterogeneity. Special attention is paid to the application to spatial models of generic statistical paradigms, such as Maximum Likelihood and Generalized Methods of Moments. An import aspect of the course is the application of open source software tools specifically those contained in the Python package PySal.
Instructor(s): L. Anselin Terms Offered: Winter. Not Being Offered in 2025/26
Prerequisite(s): An intermediate course in multivariate regression or econometrics. Familiarity with matrix algebra
Equivalent Course(s): SOCI 20559, GISC 20559, GISC 30559, SOCI 30559

DATA 20602. Thinking like a Computational Social Scientist. 100 Units.
The movement of much of our social lives online has created exciting new opportunities for social science research. This course provides a broad survey of computational methods used to make sense of this data. Students will learn how to collect online data and analyze this data using contemporary techniques from natural language processing, supervised/unsupervised machine learning, and generative AI. Students will also cultivate analytical skills through formal paper presentations, oral exams, and an original research project. The course will be taught in Python. This is an intuitive introduction without prerequisites, although previous experience with probability, statistics, and/or programming will be helpful. This course has a shared lecture on Thursdays and a separate graduate and undergraduate sections on Tuesdays(required).
Instructor(s): B. Koch Terms Offered: Spring
Equivalent Course(s): MACS 30267, HIST 49307, SOCI 40267, MACS 20267, PSYC 38520, PSYC 28520, SOCI 20602

DATA 20620. Artificial Intelligence, Innovation, and Growth. 100 Units.
Social and cultural innovation, alongside economic growth, are among the most compelling, critical and challenging phenomena in modern social science. Innovation has always been associated with unleashing transformative growth in art, science, and the economy, and in this class we explore these issues in the context of the contemporary emergence of Artificial Intelligence (AI). AI represents a novel source of innovation in economy and society, but also a powerful tool for understanding, modeling, and steering innovation in new ways. The primary purpose of this course is to enable students to understand innovation and growth in the age of AI, and with tools from AI alongside theoretical frameworks and methods from economics, sociology, evolution, and complex systems necessary for study them. The course strives to provide students with a background in dynamic analysis, data analysis, and modern AI requisite for studying innovation in the modern age. We will also consider a number of compelling theoretical and empirical challenges, ranging from the paradox of institutionalizing innovation to inequalities that emerging AI capacities could create or remove to advances it could unleash in science and technology to the spread of misinformation to consequences of AI tools and "agents" in all domains of modern life to existential risks associated with AI. We will cover theories and models at an abstract and advanced level.
Equivalent Course(s): MACS 23050, ECMA 33050, SOCI 20620, MACS 33050, SOCI 30620, ECON 23050

DATA 21100. Mathematical Methods for Data Science I. 100 Units.
This course introduces topics in probability and calculus (III) for data science students. Topics covered include random variables and probability distributions, independence, conditional probability, expected values, the central limit theorem, the definite integral, integration of functions of several variables, sums and transformations of random variables, multivariable calculus - partial derivatives, gradients, and gradient descent.
Instructor(s): B. Trok; K. Smalenberger Terms Offered: Autumn Winter
Prerequisite(s): MATH 13200 or 15200 or 16200 (could be taken concurrently), DATA 11800, DATA 11900 (could be taken concurrently) or CMSC 14100

DATA 21200. Mathematical Methods for Data Science II. 100 Units.
This course introduces topics in linear algebra for data science students. Topics covered include Vectors spaces, linear transformations and associated subspaces, orthogonality and projections, orthogonal (orthonormal) matrices, eigenvectors and eigenvalues, diagonalization, Det and Tr, matrix decompositions (LU decomposition, singular value decomposition), and transfer matrices and discrete stochastic processes. Assignments and examples focus on results and algorithms that are relevant to machine learning and data science.
Instructor(s): A. Nussbaum; J. Antunes Terms Offered: Spring Winter
Prerequisite(s): MATH 13200 or 15200 or 16200 (could be taken concurrently), DATA 11800, DATA 11900 (could be taken concurrently) or CMSC 14100

DATA 21300. Models in Data Science. 100 Units.
This course introduces fundamental aspects of modeling data such as regression, linearization and linear models, discrete dynamical systems, Markov chains, continuous dynamical systems, and stability analysis.
Instructor(s): B. Trok; S. Bhowmick; D. Biron Terms Offered: Autumn Spring Winter
Prerequisite(s): DATA 21100 (or equivalent; could be taken concurrently), DATA 21200 or STAT 24300 or MATH 19620

DATA 22100. Introduction to Machine Learning: Concepts and Applications. 100 Units.
This course introduces topics in current applications of machine learning for Data Science minor students. Topics include machine learning models, supervised and unsupervised learning, loss functions, risk, empirical risk and overfitting, regression and classification, clustering, gradient boosting, decision trees and random forests, and a brief introduction to Neural Networks and deep learning.
Instructor(s): A. Kube; L. House Terms Offered: Autumn Spring Winter
Prerequisite(s): DATA 11900 (could be taken concurrently) or CMSC 14200
Note(s): DATA 22100 and DATA 23100 are antirequisites. Students may count at most one of them, but not both, toward the forty-two credits required for graduation.

DATA 22700. Data Visualization and Communication. 100 Units.
This course introduces Data Science minors to best practices for presenting and communicating quantitative data. Principles of data visualization include the use of colors and negative spaces, drawing attention to important details, repetition of design motifs, appropriately using figures and tables, and combining different scales in a single figure. The course also discusses how to avoid common distortions resulting in misleading plots and figures and how to effectively communicate findings. Examples are chosen from a variety of fields, such as the biological sciences, the social sciences, and the media.
Instructor(s): D. Biron; S. Bhowmick; M. Hutch Terms Offered: Autumn Spring Winter
Prerequisite(s): DATA 11800 or CMSC 14100

DATA 23100. Machine Learning Fundamentals: Theory and Practice. 100 Units.
This course introduces topics in current applications of machine learning for Data Science major students. Topics include machine learning models, supervised and unsupervised learning, loss functions, risk, empirical risk and overfitting, regression and classification, clustering, gradient boosting, decision trees and random forests, and a brief introduction to Neural Networks and deep learning.
Instructor(s): D. Biron; A. Schein; K. Sehgal Terms Offered: Autumn Spring Winter
Prerequisite(s): DATA 12000 (could be taken concurrently) or CMSC 14200; DATA 21100 (or equivalent); DATA 21300 (could be taken concurrently) and DATA 21200 (or equivalent) recommended
Note(s): DATA 22100 and DATA 23100 are antirequisites. Students may count at most one of them, but not both, toward the forty-two credits required for graduation."

DATA 23700. Visualization for Data Science. 100 Units.
This course imparts design knowledge and technical skills around data visualization that data scientists need for analysis and communication. Design principles for data visualization are based on frameworks for thinking about chart construction, empirical research on human perception and cognition, a repertoire of design strategies, and theories about what makes an effective chart. Data Science majors will build practical skills around visualization APIs, computational notebooks, graphics editing software, and technical writing. Topics of interest include visualization software, spatial and visual reasoning, cartography, making data interactive, persuasion and deception, uncertainty communication, and model interpretability.
Instructor(s): A. Kale Terms Offered: Autumn
Prerequisite(s): DATA 11900 or CMSC 14100, DATA 12000 (could be taken concurrently), DATA 21300 (could be taken concurrently)

DATA 24100. Software Engineering for Data Science. 100 Units.
This course is designed to equip students with the practical skills and theoretical knowledge necessary to excel at the intersection of data science and software engineering. Through a hands-on approach, students will delve into the core tools and concepts that form the backbone of this interdisciplinary field, including data modeling, building data pipelines and software development best-practices. Emphasis will be placed on real-world applications, enabling students to work on projects that simulate professional scenarios and challenges. This course is ideal for those looking to deepen their understanding of how data-focused technologies are developed and deployed.
Instructor(s): N. Ross Terms Offered: Autumn
Prerequisite(s): DATA 12000

DATA 25422. Machine Learning for Computer Systems. 100 Units.
This course will cover topics at the intersection of machine learning and systems, with a focus on applications of machine learning to computer systems. Topics covered will include applications of machine learning models to security, performance analysis, and prediction problems in systems; data preparation, feature selection, and feature extraction; design, development, and evaluation of machine learning models and pipelines; fairness, interpretability, and explainability of machine learning models; and testing and debugging of machine learning models. The topic of machine learning for computer systems is broad. Given the expertise of the instructor, many of the examples this term will focus on applications to computer networking. Yet, many of these principles apply broadly, across computer systems. You can and should think of this course as a practical hands-on introduction

to machine learning models and concepts that will allow you to apply these models in practice. We'll focus on examples from networking, but you will walk away from the course with a good understanding of how to apply machine learning models to real-world datasets, how to use machine learning to help computer systems operate better, and the practical challenges with deploying machine learning models in practice."
Instructor(s): Nick Feamster
Prerequisite(s): CMSC 14300 or CMSC 15400
Equivalent Course(s): CMSC 25422, CMSC 35422, DATA 35422

DATA 25900. Ethics, Fairness, Responsibility, and Privacy in Data Science. 100 Units.
This course takes a technical approach to exploring societal issues of ethics, fairness, responsibility, and privacy related to the collection, use, and generalization of data. The course introduces fundamental techniques related to data acquisition, data cleaning, sampling, statistical modeling, experimental design, feature engineering, and modeling with machine learning. It then explores the problems that arise in different ways of performing those tasks, the fairness and bias of machine learning models, data visualizations, and user interfaces. In addition, the course covers anonymization and deanonymization, conceptions of privacy from a number of perspectives (statistical, legal, and philosophical), and compliance with contractual or legal requirements around data. The course concludes by discussing current controversies around the use and misuse of data. Through both programming assignments and discussions, students who complete the course will learn how to design systems that are inclusive and respectful of all data subjects.
Instructor(s): A. Nussbaum, A. Kube Terms Offered: Autumn Spring
Prerequisite(s): DATA 11900
Note(s): Students may not take DATA 25900 if they have taken CMSC 25900 or CMSC 25910.

DATA 26100. Statistical Pitfalls and Misinterpretation of Data. 100 Units.
This course will provide tools for thinking critically about data and models that constitute evidence, e.g., for the purpose of making predictions, reaching decisions, judging the quality of information, or in the context of scientific research. The course will examine examples of misleading language and graphics and discuss good practices in representing and communicating data. Examples will include such pitfalls as data size effects, false linearity, biased or correlated samples, mistaking correlation for causation, regression to the mean, and Simpson's paradox.
Terms Offered: Spring
Prerequisite(s): MATH 13200 and DATA 11800 or CMSC 14100

DATA 27005. AI Agents for Social Science & Society. 100 Units.
This course takes the position that AI agents represent a fundamental transformation in both society and social science methodology: from cartoons of social life to another dimension of sociality, and from tools that merely process data to autonomous systems that can formulate hypotheses, conduct literature reviews, design studies, analyze complex multimedia data, engage in theoretical reasoning, simulate human behavior and social dynamics, and reveal their own behaviors that are playing an increasingly important role in the human social world. In this course, students will learn to understand and construct AI agents that serve as research assistants (automating data collection and analysis), research subjects (simulating human responses and social processes or revealing their own authentic behavior), research advisors (synthesizing literature and proposing theoretical frameworks), research scientists (generating and testing hypotheses), and workers within organizations, institutions, and societies for study, but also productive work and life.
Instructor(s): James Evans Terms Offered: Winter
Equivalent Course(s): MACS 27005, MACS 37005, DATA 37005

DATA 27100-27200. Data Science Clinic I-II.
Data Science Clinic I-II

> **DATA 27100. Data Science Clinic I. 100 Units.**
> In order to enroll in this class, students must first submit an application and be matched with a project. Visit the Data Science Clinic site for application deadlines, how to apply, and information session details: https://github.com/dsi-clinic/the-clinic. The Data Science Clinic partners with public interest organizations to leverage data science research and technology to address pressing social and environmental challenges. The Clinic also provides students with exposure to real-world projects and problems that transcend the conventional classroom experience including: working with imperfect datasets, applying models and algorithms to real-world data, navigating security and privacy issues, communicating results to a diverse set of stakeholders (e.g., industry, public interest, government agencies), and translating information into actionable insights, policy briefs and software prototypes. The Clinic is an experiential project-based course where students work in teams as data scientists with real-world clients under the supervision of instructors. Students will be tasked with producing key deliverables, such as data analysis, open source software, as well as final client presentations, and reports.
> Instructor(s): N. Ross Terms Offered: Autumn Spring Winter
> Prerequisite(s): DATA 12000, DATA 13600, DATA 21100, DATA 21200, DATA 23100 (or equivalent) and by permission of instructor
> Equivalent Course(s): MACS 30300, PPHA 30581, MPCS 57300, CAPP 30300

DATA 27200. Data Science Clinic II. 100 Units.
In order to enroll in this class, students must first submit an application and be matched with a project. Visit the Data Science Clinic site for application deadlines, how to apply, and information session details: https://github.com/dsi-clinic/the-clinic. This is the second course in a sequence of Data Science Clinics. The Data Science Clinic partners with public interest organizations to leverage data science research and technology to address pressing social and environmental challenges. The Clinic also provides students with exposure to real-world projects and problems that transcend the conventional classroom experience including: working with imperfect datasets, applying models and algorithms to real-world data, navigating security and privacy issues, communicating results to a diverse set of stakeholders (e.g., industry, public interest, government agencies), and translating information into actionable insights, policy briefs and software prototypes. The Clinic is an experiential project-based course where students work in teams as data scientists with real-world clients under the supervision of instructors. Students will be tasked with producing key deliverables, such as data analysis, open source software, as well as final client presentations, and reports.
Instructor(s): N. Ross Terms Offered: Autumn Spring Winter
Prerequisite(s): DATA 27100

DATA 27400. Nonparametric Inference. 100 Units.
Nonparametric inference is about developing statistical methods and models that make weak assumptions. A typical nonparametric approach estimates a nonlinear function from an infinite dimensional space rather than a linear model from a finite dimensional space. This course gives an introduction to nonparametric inference, with a focus on density estimation, regression, confidence sets, orthogonal functions, random processes, and kernels. The course treats nonparametric methodology and its use, together with theory that explains the statistical properties of the methods.
Instructor(s): Staff Terms Offered: Winter
Prerequisite(s): STAT 24400 or STAT 24410 w/B- or better is required; alternatively STAT 22400 w/B+ or better and exposure to multivariate calculus (MATH 16300 or MATH 16310 or MATH 18400 or MATH 19520 or MATH 20000 or MATH 20500 or MATH 20510 or MATH 20800) and linear algebra (MATH 18600 or 19620 or 20250 or 20700 or STAT 24300 or equivalent). Master's students in Statistics can enroll without prerequisites.
Equivalent Course(s): STAT 37400, DATA 37400, STAT 27400

DATA 27420. Introduction to Causality with Machine Learning. 100 Units.
This course is an introduction to causal inference. We'll cover the core ideas of causal inference and what distinguishes it from traditional observational modeling. This includes an introduction to some foundational ideas---structural equation models, causal directed acyclic graphs, and then do calculus. The course has a particular emphasis on the estimation of causal effects using machine learning methods.
Instructor(s): V. Veitch Terms Offered: TBD
Prerequisite(s): [STAT 24500 or STAT 24510 or STAT 27725] with a grade of B or higher or consent of instructor.
Equivalent Course(s): STAT 27420

DATA 27751. Trustworthy Machine Learning. 100 Units.
Machine learning systems are routinely used in safety critical situations in the real world. However, they often dramatically fail! This course covers foundational and practical concerns in building machine learning systems that can be trusted. Topics include foundational issues---when do systems generalize, and why, essential results in fairness and domain shifts, and evaluations beyond standard test/train splits. This is an intermediate level course in machine learning; students should have at least one previous course in machine learning.
Terms Offered: TBD
Prerequisite(s): STAT 27700 or STAT 37710 or consent of instructor.
Equivalent Course(s): STAT 27751, STAT 37787

DATA 29700. Reading and Research. 100 Units.
Instructor consent required. Students are required to submit the signed College Reading and Research Course Form to the Director of Undergraduate Studies. Forms will be reviewed only after a student has secured a faculty sponsor on the Committee on Data Science. Must be taken for a quality grade.