

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 Director of Undergraduate 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 table of summary requirements.)

The program for the BA in Data Science consists of 15 courses 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 beyond the general education requirement. In addition to the BA requirements, students pursuing the BS must meet the following two requirements:

(i) A coherent three-quarter sequence in an independent domain of knowledge to which Data Science can be applied. This sequence 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 2000 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 25700 Natural Language Processing, 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

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 or CMSC 25900	Ethics, Fairness, Responsibility, and Privacy in Data Science Ethics, Fairness, Responsibility, and Privacy in Data Science	
DATA 22700 or CMSC 23900	Data Visualization and Communication Data Visualization	
Data Science Clinic		200
DATA 27100-27200	Data Science Clinic I-II	
Machine Learning (one of the following)		100
DATA 22100	Introduction to Machine Learning [‡]	
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	
CMSC 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	
Three Electives		300
Total Units		1500

BS IN DATA SCIENCE

Requirements for BA in Data Science ⁺		1500
A coherent sequence 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 22100 cannot both count toward the Data Science major or minor.

ELECTIVES

Electives in Data Science may come from the list of representative approved courses below or may be selected in consultation with the Program Administrator. This list is updated regularly.

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 25700	Natural Language Processing
CMSC 27620	Introduction to Bioinformatics
Biological Sciences	
BIOS 21216	Introduction to Statistical Genetics
BIOS 28407	Genomics and Systems Biology

† Only one of CMSC 25400, DATA 22100, 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-DATA 21200 Mathematical Methods for Data Science I-II
- DATA 21300 Models in Data Science
- DATA 22100 Introduction to Machine Learning

with

- MATH 16300 Honors Calculus III (or MATH 15300 Calculus III + MATH 18400 Mathematical Methods in the Physical Sciences II)
- STAT 24300 Numerical Linear Algebra or MATH 19620 Linear Algebra
- STAT 25100 Introduction to Mathematical Probability
- STAT 24400 Statistical Theory and Methods I and STAT 24500 Statistical Theory and Methods II

Conditional on quality grades in these courses, students will be allowed to count STAT 24500 Statistical Theory and Methods II as one of their Data Science electives. 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.

COMPUTATION TRACK

Students interested in emphasizing computational aspects of Data Science may be interested in advanced Computer Science electives such as CMSC 23310 Advanced Distributed Systems or CMSC 23500 Introduction to Database Systems. To accommodate their needs, such students will be allowed to replace DATA 12000 Computer Science for Data Science with CMSC 15200 Introduction to Computer Science II. Conditional on a minimum grade of C+ in CMSC 15200, these students will be able to count CMSC 15400 Introduction to Computer Systems as one of their Data Science electives. Similarly, 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, CMSC 14300 Systems Programming I and CMSC 14400 Systems Programming II could count as two Data Science electives.

DOUBLE MAJORS

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 encouraged to discuss their course plans and obtain advice from their academic

advisers and the Data Science program director. In addition, they must meet with the Data Science program director to obtain approval for their course plan.

GRADING AND ADVISING FOR MAJOR PROGRAM

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. Prospective majors must meet with the data science program director to discuss their course plans and to obtain advice and approval. Together the student and the program director will fill out a Consent to Complete a Major Form. Students should submit completed, signed forms to their College adviser by the end of the Spring Quarter of their third year.

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.

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 program director, 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 or DATA 22100	Mathematical Foundations of Machine Learning Introduction to Machine Learning	
One of the following:		100
DATA 25900 or CMSC 25900	Ethics, Fairness, Responsibility, and Privacy in Data Science Ethics, Fairness, Responsibility, and Privacy in Data Science	

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

Two of the following:		200
CMSC 13600	Introduction to Data Engineering	
DATA 22700 or CMSC 23900	Data Visualization and Communication Data Visualization	
CMSC 25025 or DATA 22100	Machine Learning and Large-Scale Data Analysis Introduction to Machine Learning	
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 must meet with the data science program director to discuss their course plans and to obtain advice and approval. Together the student and the program director will 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). 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): William Trimble / TBD Terms Offered: Autumn Winter

Prerequisite(s): None

Equivalent Course(s): CMSC 11800, 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): William L Trimble / TBD Terms Offered: Spring Winter

Prerequisite(s): DATA 11800 , or STAT 11800 or CMSC 11800 or consent of instructor.

Equivalent Course(s): CMSC 11900, 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.

Terms Offered: Spring

Prerequisite(s): DATA 11800

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): D. Biron Terms Offered: Winter

Prerequisite(s): MATH 13200 or 15200 or 16200 (could be taken concurrently), DATA 11800 or CMSC 12100 or CMSC 15100 or CMSC 16100

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): D. Biron Terms Offered: Spring

Prerequisite(s): MATH 13200 or 15200 or 16200 (could be taken concurrently), DATA 11800 or CMSC 12100 or CMSC 15100 or CMSC 16100

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.

Terms Offered: Spring

Prerequisite(s): DATA 21100 (may be replaced by MATH 13300 and STAT 23400), DATA 21200 or STAT 24300 or MATH 19620

DATA 22100. Introduction to Machine Learning. 100 Units.

This course introduces topics in current applications of machine learning for Data Science 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 (time permitting) a brief introduction to Neural Networks and deep learning.

Instructor(s): W. Trimble Terms Offered: Spring

Prerequisite(s): DATA 11900 or CMSC 12200 or CMSC 15200 or CMSC 16200, DATA 21100 or MATH 13300 or 15300 or 16300 or STAT 23400 or STAT 24400 (can be taken concurrently)

DATA 22700. Data Visualization and Communication. 100 Units.

This course introduces 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): W. Trimble Terms Offered: Autumn

Prerequisite(s): DATA 11800 or CMSC 12100 or CMSC 15100 or CMSC 16100

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.

Prerequisite(s): CMSC 15400

Equivalent Course(s): DATA 35422, CMSC 35422, CMSC 25422

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): Raul Castro Fernandez Terms Offered: Spring

Prerequisite(s): CMSC 11900

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.

Instructor(s): D. Biron Terms Offered: Autumn Spring

Prerequisite(s): MATH 13200 and DATA 11800 or CMSC 12100 or CMSC 15100 or CMSC 16100

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

Data Science I-II

DATA 27100. Data Science Clinic I. 100 Units.

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): D. Uminsky Terms Offered: Autumn-Winter, Winter-Spring

Prerequisite(s): DATA 11900, CMSC 13600, DATA 21200 or MATH 13300 or 15300 or 16300, DATA 21200 or STAT 24300 or MATH 20250 or CMSC 25300 or CMSC 25400, DATA 12000 or CMSC 12200 or CMSC 15200. Alternatively, by permission of the instructor.

Note(s): DATA 22100 or CMSC 25025 or CMSC 25300, DATA 22700 or CMSC 23900, DATA 25900 or CMSC 25900 are recommended.

DATA 27200. Data Science Clinic II. 100 Units.

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): D. Uminsky Terms Offered: Spring Winter

Prerequisite(s): DATA 27100

