

COMMITTEE ON DATA SCIENCE

The Committee on Data Science (CODAS) was established in 2023 to support graduate and undergraduate programs in this emerging discipline at the University of Chicago. Affiliated faculty come from numerous departments across campus with a core group in the departments of Statistics and Computer Science. CDS holds the educational philosophy that a strong program in Data Science should encompass foundational theory, methodological innovations and real-world applications. A Data Science education should draw from the intellectual tradition and key concepts of Computer Science, Applied Mathematics, Statistics, and other fields while providing a new integrative framework for data-driven thinking, discovery, and decision-making.

Click here to view the Committee on Data Science (CODAS) website: <https://codas.uchicago.edu/>

Committee Co-Directors

- Dan L Nicolae (Statistics)
- Michael J Franklin (Computer Science)

Program Faculty

- Luc Anselin (Sociology)
- Luis Bettencourt (Ecology, Sociology)
- Raul Castro Fernandez (Computer Science)
- Aloni Cohen (Computer Science)
- Moon Duchin (Computer Science, Data Science)
- James Evans (Sociology)
- Aaron Elmore (Computer Science)
- Robert Grossman (Medicine, Computer Science)
- Pedram Hassanzadeh (Geophysical Sciences)
- Ari Holtzman (Computer Science, Data Science)
- Nikos Ignatiadis (Statistics, Data Science)
- Hae Kyung Im (Medicine, Human Genetics)
- Alex Kale (Computer Science, Data Science)
- Frederick Koehler (Statistics, Data Science)
- Sanjay Krishnan (Computer Science)
- Mina Lee (Computer Science, Data Science)
- Tian Li (Computer Science, Data Science)
- Edwin Lo (Data Science)
- Cong Ma (Statistics)
- Li Ma (Statistics, Data Science)
- Isaac Mehlhaff (Political Science)
- Sendhil Mullainathan (Computation, Behavioral Science)
- Samantha Riesenfeld (Molecular Engineering, Medicine)
- Veronika Rockova (Econometrics, Statistics)
- Aaron Schein (Statistics, Data Science)
- Nathan Srebro (Computer Science)
- Matthew Stephens (Statistics)
- Chenhao Tan (Computer Science, Data Science)
- David Uminsky (Computer Science)
- Blase Ur (Computer Science)
- Victor Veitch (Statistics, Data Science)
- Jingshu Wang (Statistics)
- Molly Offer-Westort (Political Science)
- Rebecca Willett (Statistics, CAMI, Computer Science)
- Haifeng Xu (Computer Science, Data Science)
- Ce Zhang (Computer Science, Data Science)
- Jian Zhou (Medicine)

PhD in Data Science

Program Overview

The PhD in Data Science was developed to train all students in the mathematical foundations of data science, responsible data use and communication, as well as advanced computational methods. Candidates will be able to explore diverse research opportunities alongside distinguished Data Science faculty at UChicago.

Curriculum

The program requires students to complete nine courses: four required courses (1-4 below); one elective either in mathematical foundations or scalability and computing (5 or 6 below), and four other graduate-level electives that can come from proposed courses in Data Science or existing graduate courses in Computer Science or Statistics. Some students, after consulting with the committee graduate advisor, might decide to take all nine courses over the first two years.

Required Courses:

1. Foundations in Machine Learning & AI - Part I
2. Responsible Use of Data & Algorithms
3. Data Interaction
4. Systems for Data and Computers / Data Design

Required Electives (Choose one of the following):

1. Foundations in Machine Learning & AI - Part II
2. Data Engineering & Scalable Computing

Thesis Advisor and Dissertation Committee

Students typically select a thesis advisor by the beginning of their second year. By the end of the third year, each PhD student shall establish a thesis committee of at least three faculty members, including the advisor, with at least half of the members coming from the Committee on Data Science (CDS).

Proposal Presentation & Admission to Candidacy

By the end of the third year, students should have scheduled and completed a proposal presentation to their committee in order to be advanced to candidacy. The proposal presentation is typically an hour-long meeting that begins with a 30-minute presentation by the student followed by a question and discussion period with the committee.

Admissions

The PhD in Data Science admits students each year for the Fall quarter only; a full list of admission requirements and a link to start your application can be found here (<https://codas.uchicago.edu/how-to-apply/>). If you have any questions regarding your application or the admissions process, please send all inquiries to kkritikos34@uchicago.edu (data-science@uchicago.edu) for a timely response.

Master's in Data Science (MSDS)

Program Overview

The Master's in Data Science (MSDS) was developed for students interested in pursuing a research career in Data Science with courses taught by faculty in Statistics, Computer Science, and other departments across the university. Students must complete a total of 9 core and elective courses plus 3 foundational courses as follows:

- Foundational Courses (3 courses): These cover essential concepts and practices in (1) programming, (2) mathematics, and (3) statistical knowledge that is central to data science and provide the groundwork for advanced coursework. Students can qualify or be placed out of these courses. If needed to take these (online, free) courses, they should be completed in the summer before they begin with the MSDS program in the Autumn, or during the beginning of the Autumn quarter if the course is not offered in the summer.
- Core Courses (5 courses): These delve into key data science techniques, methodologies, and tools.
 - Introduction to Data Science
 - Data Interaction
 - Introduction to Data Systems and Data Design OR Data Engineering & Scalable Computing
 - Responsible Use of Data and Algorithms
 - Introduction to ML and Neural Networks OR Foundations of Machine Learning and AI - Part I

- Elective Courses (4 courses): Students choose from a wide array of electives to tailor their learning to specific interests or career goals. Most graduate level courses (300xx) in the DATA, STAT, and CMSC programs taken with a letter grade (not pass/fail nor audit) are eligible as electives. Reading/Research courses do not count. Graduate courses from other departments/programs that are related to data science could be counted as well. Contact the director for approval with such course elections.

You can review our Student Handbook (<https://physicsland.github.io/dsi-msds-student-handbook/structure.html>) for additional questions about our curriculum overview, research requirements, and much more.

Admissions

The Master's in Data Science (MSDS) program admits students each year for the Fall quarter only; a full list of admission requirements and a link to start your application can be found here (<https://codas.uchicago.edu/how-to-apply/>). If you have any questions regarding your application or the admissions process, please send all inquiries to kkritikos34@uchicago.edu (data-science@uchicago.edu) for a timely response.

DATA SCIENCE COURSES

DATA 30100. Introduction to Data Science. 100 Units.

The course will focus on the analysis of real life data and on statistical and machine learning methods to perform inference and to predict future outcomes. It will cover topics from the whole data life cycle, ranging from data collection (including wrangling, cleaning, and sampling) to summarizing results through visualization and interpretable summaries, with a focus on extracting meaning, value and information from data. Important aspects in data science, such as bias, fairness, privacy while building algorithms and predictive models, will also be explored.

Instructor(s): D. Nicolae Terms Offered: Autumn

Prerequisite(s): Consent of Instructor unless graduate student in Data Science

DATA 30120. Technical Presentation. 100 Units.

This course is intended for PhD students in CS and Data Science. This seminar will focus on giving technical presentations, emphasizing presenting results at a conference or workshop. We will cover topics such as structuring and designing talks, audience identification, setting context, introductions, body language, pacing, slideshow visualizations, explaining experiments and results, conclusions, and other general tips. Students will be expected to give short snippets of talks and provide active feedback on others.

Equivalent Course(s): CMSC 30120

DATA 30195. Data and the State. 100 Units.

In order to regulate and govern, states must marshal 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.

Instructor(s): M. Duchin Terms Offered: Spring

Equivalent Course(s): DATA 20195

DATA 30332. Thinking with Deep Learning for Complex Social & Cultural Data Analysis. 100 Units.

A deluge of digital content is generated daily by web-based platforms and sensors that capture digital traces of human communication and connection, and complex states of society, culture, economy, and the world. Emerging deep learning methods enable the integration of these complex data into unified social and cultural "spaces" that enable new answers to classic social and cultural questions, and also pose novel questions. From the perspective of deep learning, everything can be viewed as data-novels, field notes, photographs, lists of transactions, networks of interaction, theories, epistemic styles-and our treatment examines how to configure deep learning architectures and multi-modal data pipelines to improve the capacity of representations, the accuracy of complex predictions, and the relevance of insights to substantial social and cultural questions. This class is for anyone wishing to analyse textual, network, image or arbitrary structured and unstructured data, especially in concert with one another to solve complex social and cultural analysis problems (e.g., characterize a culture; predict next year's ideology).

Instructor(s): James Evans Terms Offered: Winter

Prerequisite(s): The course uses Python and the widely popular PyData ecosystem to demonstrate all motivating examples and includes working code, accompanying exercises, relevant datasets and additional analytics and visualization that facilitate social and cultural interpretation and communication. Familiarity with Python is required.

Equivalent Course(s): SOCI 30332, MACS 37000, MACS 27000

DATA 31500. Data Interaction. 100 Units.

This course provides core knowledge and technical skills around data interfaces, with an emphasis on visualization and front-end software development. Graduate students in Data Science and Computer Science will engage in project-based learning to become fluent with visualization APIs, computational notebooks, web development, technical writing, and presentation. Topics of interest include data visualization design, cartography, interactive articles, data storytelling, uncertainty communication, model interpretability, and appropriate use of AI assistance.

Instructor(s): A. Kale Terms Offered: Autumn Spring

Prerequisite(s): Consent of Instructor unless graduate student in Data Science

Equivalent Course(s): CMSC 31500

DATA 32100. Generative Models: Theory and Applications. 100 Units.

This graduate topics course covers classical and modern generative models and their applications. Students will read and present research papers, implement representative algorithms and models, and study recent developments in generative modeling. Topics include latent variable models (e.g., mixture models, mixture-of-experts, latent factor models, variational autoencoders), tree-based models, generative adversarial networks, flow-based models, diffusion models, as well as assessment methods for these models. The applications will emphasize the generation of non-text data and their use in scientific inquiries, with a particular focus on biomedical problems. Prerequisites: probabilistic modeling, statistical machine learning, and experience in Python programming (in particular, with PyTorch). Experience with Bayesian inference and computation will also be helpful.

Instructor(s): L. Ma Terms Offered: Spring

Prerequisite(s): Must be graduate student in Data Science, Statistics, or Computer Science. All others should email faculty directly for instructor consent to enroll.

Equivalent Course(s): STAT 32100

DATA 33255. Modeling Democracy. 100 Units.

This is a graduate course that develops a mathematical/computational toolkit for analyzing democratic systems. I'll provide a self-contained introduction to social choice theory, computational social choice theory (also known as COMSOC), and applied modeling for democratic mechanisms.

Instructor(s): M. Duchin Terms Offered: Winter

Prerequisite(s): Consent of instructor unless graduate student in data science

Equivalent Course(s): CMSC 33255

DATA 34100. Introduction to Data Systems and Data Design. 100 Units.

The goal of this course is to teach students: (1) how to think about data , its logical semantics, and what is a query; (2) how to practically handle data, both in relational databases and other more flexible data processing frameworks (e.g. Spark); (3) practical design principles about schema, integrity constraints, etc. (4) an introduction to systems that allows students to understand performance, and helps them become better users.

Instructor(s): C. Zhang Terms Offered: Autumn

Prerequisite(s): Consent of Instructor unless graduate student in Data Science

DATA 34200. Data Engineering and Scalable Computing. 100 Units.

This course covers the principles and practices of managing and processing data at scale. Students will learn about distributed systems, cloud computing, and big data technologies. Topics include data storage architectures, data catalogs and governance, distributed computing frameworks like Apache Spark, streaming data processing, and data transformation pipelines. The course will provide hands-on experience with state-of-the-art tools and techniques for building end-to-end data engineering solutions to support large-scale data science, analytics and AI applications.

Instructor(s): M. Franklin Terms Offered: Spring Winter

Prerequisite(s): DATA 34100; Consent of Instructor unless graduate student in Data Science

DATA 35422. 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, DATA 25422, CMSC 35422

DATA 35900. Responsible Use of Data and Algorithms. 100 Units.

The goal of this course is to cultivate a societally oriented mindset and to train students critically about the contexts into which data science is deployed. It will be organized around a series of modules consisting of three components: (i) a broad challenge, (ii) mathematical / technical approaches that have been used to address that challenge, and (iii) a real-world case study. The modules will cover a diverse set of topics, including for example: data privacy; algorithmic fairness; AI safety; explainable and interpretable AI; and robustness to strategic or adversarial behavior.

Instructor(s): A. Cohen Terms Offered: Spring

Prerequisite(s): Consent of Instructor unless graduate student in Data Science

DATA 37000. Introduction to Machine Learning and Neural Networks. 100 Units.

This course is an introduction to machine learning (ML) for students to build a solid foundation in modeling and data science. It will cover both unsupervised and supervised ML algorithms, with the latter focusing on both regression and classification models. Python is the programming language of choice for implementing various models to solve complex problems across multiple domains. The course will also introduce basic neural network architectures, including Single-Layer Perceptron (SLP), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). Students will apply these techniques in contexts where they are most effective. A strong understanding of linear algebra, multivariable calculus, and statistics/probability theory is expected. Python coding assignments and projects will be integral to the course.

Instructor(s): E. Lo Terms Offered: Autumn

Prerequisite(s): Consent of Instructor unless graduate student in Data Science

DATA 37005. 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 37005, DATA 27005, MACS 27005

DATA 37100. Introduction to AI: Deep Learning and GAI. 100 Units.

Artificial Intelligence is transforming industries and daily life, permeating almost every aspect of modern society. This course builds on technical knowledge from previous foundations in Machine Learning and Neural Networks to provide a deep understanding of current AI platforms. Emphasizing hands-on experience in Generative Artificial Intelligence, students will learn to implement and train advanced AI models, including but not limited to transformers, diffusion models, and Large Language Models (LLMs). Additionally, the course will critically examine the ethical implications of AI, exploring the benefits, challenges, and potential risks associated with its deployment. Students enrolling in this course should have proficiency in Python programming, and a solid foundation in mathematics (including linear algebra and multivariable calculus) as well as statistics.

Instructor(s): E. Lo Terms Offered: Winter

Prerequisite(s): DATA 37000; Consent of Instructor unless graduate student in Data Science

DATA 37110. Foundations of Machine Learning and AI for Scientists. 100 Units.

This course introduces PhD students in PSD to core concepts and methods in machine learning and artificial intelligence, with a focus on applications to scientific data. Topics include supervised and unsupervised learning, regression, classification, clustering, dimensionality reduction, and introductory perspectives on neural networks and generative models. Emphasis is placed on both conceptual understanding and hands-on practice with real-world scientific datasets. Students will gain experience implementing standard algorithms, evaluating model performance, and critically assessing the role of machine learning in research. Designed for participants without prior data science training, this one-quarter course provides a rigorous but accessible foundation for applying modern computational methods to scientific problems and prepares students for further study in data science.

Instructor(s): E. Lo

Prerequisite(s): Consent of Instructor unless graduate student in the Data Science Ph.D. Certificate program

DATA 37200. Learning, Decisions, and Limits. 100 Units.

This is a graduate course on theory of machine learning. While ML theory has multiple branches in general, this course is designed to cover basics of online learning, along with basics of reinforcement learning. It aims to establish the foundation for students who are interested in conducting research related to online decision making, learning, and optimization. The course will introduce formal formulations for fundamental problems/ models in this space, describe basic algorithmic ideas for solving these models, rigorously discuss performances of these algorithms as well as these problems' fundamental limits (e.g., minimax/lower bounds). En route, we will develop necessary toolkits for algorithm development and lower bound proofs.

Instructor(s): F. Koehler Terms Offered: Winter

Prerequisite(s): Requires linear algebra (at the level of CMSC 25300 or its equivalent), algorithms (CMSC 27200 or its equivalent) and probability (STAT 25100 or its equivalent). If not sure, consult with the instructor. Note that no background on learning theory is required.

Equivalent Course(s): STAT 37201

DATA 37400. 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): DATA 27400, STAT 27400, STAT 37400

DATA 37711. Foundations of Machine Learning and AI - Part I. 100 Units.

This course is an introduction to machine learning targeted at students who want a deep understanding of the subject. Topics include modern approaches to supervised learning, unsupervised learning, and the use of machine learning in estimating real-world effects. In principle, no previous exposure to machine learning is required. However, students are expected to have mathematical maturity at the level of an advanced undergraduate, including being comfortable with linear algebra, multivariate calculus, and (non-measure theoretic) statistics and probability. Assignments include programming in python (and pytorch).

Instructor(s): V. Veitch Terms Offered: Autumn

Prerequisite(s): Consent of Instructor unless graduate student in Data Science

Equivalent Course(s): STAT 37711, CAAM 37711

DATA 37712. Foundations of Machine Learning and AI - Part II. 100 Units.

Deep generative models have become a staple of modern machine learning research. This course is meant as an introduction to the way generative models are structured and trained: students will learn the mechanics of generative models as well as getting their hands dirty building them. We will discuss open questions for which we lack complete theoretical or empirical answers, with importance placed on analyzing, interpreting, and making arguments from necessarily incomplete empirical evidence. We will have a specific focus on Autoregressive Transformers and their use as Large Language Models (LLMs). The goal of this course is to get students to be proficient enough with the inner workings of deep generative models to be able to understand and reason about cutting-edge research. This is an advanced machine learning course, and assumes a familiarity with basic machine learning concepts (generalization, overfitting, etc.) and techniques (regularization, stochastic gradient descent, etc).

Instructor(s): A. Holtzman Terms Offered: Winter

Prerequisite(s): DATA 37711; Consent of Instructor unless graduate student in Data Science

Equivalent Course(s): CMSC 37712

DATA 37784. Representation Learning in Machine Learning. 100 Units.

This course is a seminar on representation learning in machine learning. The core questions in this are: how do machine learning systems recover the structure present in real-world data, how can we expose this recovered structure to human analysts, and how does this help us in real-world applications? In this seminar, we will read and discuss papers from the modern research literature on these subjects. Students should have previous exposure to machine learning and deep learning.

Terms Offered: Spring

Equivalent Course(s): STAT 37784

DATA 39900. Reading/Research: Data Science. 300.00 Units.

Directed reading and research in data science, under the guidance of a faculty member.

Instructor(s): Staff Terms Offered: Winter

DATA 41551. Empirical Bayes. 100 Units.

In an empirical Bayes analysis, we imitate inferences made by an oracle Bayesian with extensive knowledge of the data-generating distribution. Empirical Bayes provides a principled approach for "learning from the experience of others" and is widely used in application domains such as genomics, small-area estimation, economics, and large-scale experimentation. In this graduate topics course, we provide an overview of empirical Bayes. We revisit the original papers that introduced the core ideas and explain how empirical Bayes is applied in practice. We also develop mathematical techniques to study empirical Bayes procedures from a theoretical perspective.

Terms Offered: Winter

Prerequisite(s): STAT 30100 or consent of instructor

Equivalent Course(s): STAT 41551

