Data Science Institute

Brown University's Data Science Institute serves as a campus hub for research and education in data science. Through our research and academic programs, we strive to ensure that those most in need are not the last to benefit from fundamental research in data science or data-driven applied research. Brown's DSI is unique because we:

- Equally value both domain-driven and fundamental methodological research in data science;
- Increase data fluency and educate the next generation of data scientists, through our master's program and providing outreach to students and researchers at a variety of career stages;
- Explore the impact of the data revolution on culture, education, health and genomics, society, and social justice by engaging with research partners within the University and beyond.

The DSI offers a master's degree in data science, a doctoral certificate for Brown graduate students and an undergraduate certificate in Data Fluency. The DSI also supports two research centers, the Center for Computational Molecular Biology (https://ccmb.brown.edu/) (CCMB) and the Center for Technological Responsibility, Reimagination, and Redesign (CNTR).

To support data science research and education across Brown's campus, the DSI hosts seminars and public lectures, offers workshops in data science skills, and offers small grants to Brown researchers in all disciplines.

For additional information, please visit the institute's website: http://dsi.brown.edu/.

Data Fluency Concentration Requirements

The Certificate in Data Fluency provides a formal pathway for undergraduates in concentrations other than applied mathematics, computational biology, computer science, math, and statistics who wish to gain fluency and facility with the tools of data science. The driving intellectual question is how we can infer meaning from data whilst avoiding false predictions. The required experiential learning component provides students with the opportunity to apply their data-science skills in applied settings, engage in research that uses data science, teach data science as an undergraduate teaching assistant, or undertake an internship that has a substantive data-science component.

As with all undergraduate certificates (https://www.brown.edu/academics/college/degree/undergraduatecertificates/), the certificate has the following requirements:

- Students may not earn more than **one certificate** and may only have **one declared concentration**.
- Students must be enrolled in or have completed at least **two courses** toward the certificate at the time they declare in ASK.
- No more than **one course** may count toward your concentration and the certificate.
- Students may declare in ASK **no earlier** than the beginning of the fifth semester and must declare **no later** than the last day of classes of the antepenultimate (typically the sixth) semester, in order to facilitate planning for the capstone or other experiential learning opportunity.
- Students must submit a proposal for their experiential learning opportunity by the end of the **sixth semester**.

Excluded Concentrations: Applied Mathematics, Computational Biology, Computer Science, Mathematics, and Statistics (including joint concentrations in these areas).

For more information on the Certificate in Data Fluency, please visit the Data Science Institute (https://dsi.brown.edu/academics/certificate-data-fluency/) website.

Certificate Requirements

Core Courses:		
DATA 0080	Data, Ethics and Society	1
CSCI 0111	Computing Foundations: Data	1
or CSCI 0150	Introduction to Object-Oriented Programming and Computer Science	
or CSCI 0170	Computer Science: An Integrated Introduction	
or CSCI 0190	Accelerated Introduction to Computer Science	
or CLPS 0950	Introduction to programming	
DATA 0200	Data Science Fluency	1
Elective Course: Select one follow-up Applied Math, Biostatistics, Computer Science or domain-specific course with a significant data component from the following list (or another course with approval from the certificate advisor):		1
ANTH 1201	Introduction to Geographic Information Systems and Spatial Analysis	
APMA 1650	Statistical Inference I	
BIOL 0495	Statistical Analysis of Biological Data	
BIOL 1555	Methods in Informatics and Data Science for Health	
BIOL 1565	Survey of Biomedical Informatics	
CLPS 0900	Statistical Methods	
CLPS 1291	Computational Methods for Mind, Brain and Behavior	
CLPS 1580C	Visualizing Information	
CSCI 1450	Advanced Introduction to Probability for Computing and Data Science	
CSCI 1470	Deep Learning	
CSCI 1951A	Data Science	
DATA 1150	Data Science Fellows[1]	
ECON 1620	Introduction to Econometrics	
ECON 1660	Big Data	
EDUC 1230	Applied Statistics for Ed Research and Policy Analysis	
ENVS 1105	Introduction to Environmental GIS	
EEPS 1320	Introduction to Geographic Information Systems for Environmental Applications	
EEPS 1330	Global Environmental Remote Sensing	
MATH 1210	Probability	
MUSC 1210	Seminar in Electronic Music: Real-Time Systems	
PHP 1501	Essentials of Data Analysis	
PHP 1510	Principles of Biostatistics and Data Analysis	
SOC 1020	Methods of Social Research	
SOC 1100	Introductory Statistics for Social Research	
SOC 1340	Principles and Methods of Geographic Information Systems	
Capstone:		0-1
The required experiential learning component provides students with the opportunity to apply their data-science skills in their concentration, engage in research that uses data science, teach data science as UTAs, or undertake an internship that has a data-science component. The capstone may be completed for credit via an independent study course or not for credit.[2]		
Options for fulfilling this requirement include:		
1. Participate in a Brown University credit experience (i.e. independent study).		
2. Participate in a non-credit experience: summer Internship; TA for data-related course; work with a local organization on a data-related project. A 10-12 page reflective paper is required for this option.		
3. Be a Data Science Fellow.[1]		
Total Credits		**4-5**

[1] Students may complete DATA 1150 and the concurrent Data Science Fellows project to fulfill both the elective and experiential components of the certificate.

[2] Students must submit a proposal for their experiential component by the end of the sixth semester.

Data Science Graduate Program

Master of Science in Data Science

The Data Science Institute at Brown offers a master's program (ScM) that prepares students from a wide range of disciplinary backgrounds for distinctive careers in Data Science. With connections to departments across campus, in particular Brown's Division of Applied Mathematics (https://appliedmath.brown.edu/) and Department of Computer Science (https://cs.brown.edu/), the master's program offers a unique and rigorous education for people building careers in data science. The program is designed to provide a fundamental understanding of the methods and algorithms of data science, to be achieved through a study of relevant topics in mathematics, statistics, and computer science, including database engineering, visualization, machine learning, and deep learning. The program also provides experience in important, frontline data-science problems in a variety of fields, and introduces students to ethical and societal considerations surrounding data science and its applications.

The program's course structure, including the capstone experience, ensures that students meet the goals of acquiring and integrating foundational knowledge for data science, applying this understanding in relation to specific problems, and appreciating the broader ramifications of data-driven approaches to human activity.

The program can be completed in 12 months (September to August). All students begin the program in September; **there is no option for starting in the spring semester**. Students may elect to complete the program over 16, 21, or 24 months, and most do so. In some cases, exceptionally well-prepared students complete their work in nine months.

The curriculum for the Data Science Master's Program consists of nine credits: eight required courses, one of which is the experiential project course, and one elective. The nine credit-units divide as follows:

- 3 credits in mathematical and statistical foundations
- 3 credits in data and computational science
- 1 credit in societal implications and opportunities
- 1 elective credit to be drawn from a wide range of focused applications or deeper theoretical exploration
- 1 credit capstone experience.

We also offer an option as a 5-th Year Master's Program if you are an undergraduate at Brown. This allows you to substitute maximally 2 credits with courses you have already taken. **5th-Year students must complete the degree in one year.**

Master of Science in Data Science

For more information about the Master's Program curriculum and when courses are offered, please visit the DSI Master's curriculum page (https://dsi.brown.edu/academics/masters-degree/curriculum/) or Courses@Brown (https://cab.brown.edu/).

DATA 1030	Hands-on Data Science	1
DATA 1050	Data Engineering	1
APMA 1690	Computational Probability and Statistics	1
DATA 2020	Statistical Learning	1
CSCI 1951Z	Fairness in Automated Decision Making	1
CSCI 2470	Deep Learning	1
DATA 2060	Machine Learning: from Theory to Algorithms	1
DATA 2050	Data Science Practicum	1
The practicum experience is a hands-on thesis project that entails an in-depth study of a current problem in data science. Students will synthesize their knowledge of probability and statistics, machine learning, and data and computational science. Students will work in teams on projects with Brown faculty members or with external companies. The project will be completed as part of a course that includes additional career-oriented skills development.		
One elective:		1
Domain knowledge relevant to individual interest, 1 credit, must be a graduate level course with 4-digit course number starting with a non-0 digit. Most graduate level CSCI and APMA courses qualify. Please contact the DGS if you plan to take a course from a different department.		
Total Credits		**9**

For more information on admission and program requirements, please visit the following website:

https://www.brown.edu/academics/gradschool/programs/data-science (https://www.brown.edu/academics/gradschool/programs/data-science/)

Courses

DATA 0080. Data, Ethics and Society.
A course on the social, political, and philosophical issues raised by the theory and practice of data science. Explores how data science is transforming not only our sense of science and scientific knowledge, but our sense of ourselves and our communities and our commitments concerning human affairs and institutions generally. Students will examine the field of data science in light of perspectives provided by the philosophy of science and technology, the sociology of knowledge, and science studies, and explore the consequences of data science for life in the first half of the 21st century. Fulfills requirement for Certificate in Data Fluency

Fall	DATA0080	S01	18218	MWF	9:00-9:50(09)	'To Be Arranged'

DATA 0200. Data Science Fluency.
As data science becomes more visible, are you curious about its unique amalgamation of computer programming, statistics, and visualizing or storytelling? Are you wondering how these areas fit together and what a data scientist does? This course offers all students regardless of background the opportunity for hands-on data science experience, following a data science process from an initial research question, through data analysis, to the storytelling of the data. Along the way, you will learn about the ethical considerations of working with data, and become more aware of societal impacts of data science. Course does not count toward CS concentration requirements.

Spr	DATA0200	S01	26366	TTh	1:00-2:20(08)	(L. Clark)

DATA 0250. Applied Statistics in Python.
As more students engage in data science there is a need to provide guidance on conducting basic statistical analysis in Python. This course will provide a non-specialist approach to applied statistics, specifically linear models Python. Students will learn how to conduct linear modules using the Statsmodels package in Python. Students should have good working knowledge of descriptive statistics (equivalent to a high school AP level). Python coding experience is helpful but not required. Student learning would be assessed through hands-on Python coding activities and written interpretation of statistical reports.

DATA 1010. Probability, Statistics, and Machine Learning.
An introduction to the mathematical methods of data science through a combination of computational exploration, visualization, and theory. Students will learn scientific computing basics, topics in numerical linear algebra, mathematical probability (probability spaces, expectation, conditioning, common distributions, law of large numbers and the central limit theorem), statistics (point estimation, confidence intervals, hypothesis testing, maximum likelihood estimation, density estimation, bootstrapping, and cross-validation), and machine learning (regression, classification, and dimensionality reduction, including neural networks, principal component analysis, and unsupervised learning).

DATA 1030. Hands-on Data Science.
Develops all aspects of the machine learning pipeline: data acquisition and cleaning, handling missing data, exploratory data analysis, visualization, feature engineering, modeling, interpretation, presentation in the context of real-world datasets. Fundamental considerations for data analysis are emphasized (the bias-variance tradeoff, training, validation, testing). Classical models and techniques for classification and regression are included (linear and logistic regression with regularization, support vector machines, decision trees, random forests, XGBoost). Uses the Python data science ecosystem (e.g., sklearn, pandas, matplotlib). Prerequisites: A course equivalent to CSCI 0050, CSCI 0150 or CSCI 0170 are strongly recommended.

Fall	DATA1030	S01	18121	TTh	10:30-11:50(13)	(A. Zsom)

DATA 1050. Data Engineering.
The course will cover the storage, retrieval, and management of various types of data and the computing infrastructure (such as various types of databases and data structures) and algorithmic techniques (such as searching and sorting algorithms) and query languages (such as SQL) for interacting with data, both in the context of transaction processing (OLTP) and analytical processing (OLAP). Students will be introduced to measures for evaluating the efficacy of different techniques for interacting with data (such as 'Big-Oh' measure of complexity and the number of I/O operations) and various types of indexes for the efficient retrieval of data. The course will also cover several components of the Hadoop ecosystem for the processing of 'big data.' Additional topics include cloud computing and NoSQL databases. Introduction to concepts and techniques of computer science essential for data science will also be covered.

Fall	DATA1050	S01	18222	MWF	11:00-11:50(16)	(S. Pradhan)

DATA 1150. Data Science Fellows.
DATA 1150 for juniors and seniors possessing data science skills, seeking to apply these skills and collaborate with faculty to integrate data science content into Brown courses. The course teaches communication, teaching and learning strategies, and determining project requirements. Qualified students have a combination of programming experience (intermediate level or above in R or Python), statistical knowledge (intermediate level or above) and knowledge of how data and computing can be used in applied fields. Students in the data fluency certificate must have DATA 0200 prior to DATA 1150. Students need to complete the application (url below) no later than August 1st for consideration. Qualified students must participant in an interview with the instructor and override requests will be granted only to students by instructor approval. https://forms.gle/Je3Prrzs3NDEo4eG9

Fall	DATA1150	S01	18117	TTh	1:00-2:20(06)	(L. Clark)

DATA 1200. Reality Remix - Experimental VR.
This course pursues collaborative experimentation with virtual and augmented reality (AR and VR). The class will work as a team to pursue research (survey of VR/AR experiences, scientific and critical literature review), reconnaissance (identifying VR/AR resources on campus, in Providence and the region), design (VR/AR prototyping). Research findings are documented in a class wiki. The course makes use of Brown Arts Initiative facilities in the Granoff Center where an existing VR laboratory will be expanded through the course of the semester based on student needs. Class culminates in the release the class wiki as a resource for the Brown community.

DATA 1340. Machine Learning for the Earth and Environment (EEPS 1340)..
Interested students must register for EEPS 1340.

DATA 1450. Text Analytics.
This course will first cover techniques for compiling textual corpora from web pages, pdfs, scanned pdfs, images, audio clips, etc. Secondly, it will look at processes for extracting some common types of information from these corpora. In particular, we will cover extracting named entities (persons, locations, organizations, etc.), relations between entities, events, transactions, topics, document summaries, abstracts, legal clauses, etc. This course is different from standard courses in Natural Language Processing and Computational Linguistics in that we will spend significant amount of course time on compiling textual corpora from documents in a variety of formats and our emphasis will be on extracting information that can be fed to analytics pipelines.

Spr	DATA1450	S01	26368	TTh	2:30-3:50(11)	(S. Pradhan)

DATA 1500. Data Visualization & Narrative.
Data visualization is an essential tool in both discovering and communicating key analytical findings. However, data practitioners and developers can sometimes undervalue the visual polish that goes into creating the most effective graphics. This course will act as a technical primer for building data visualization using code, but a core focus will be the graphic design decisions – color, hierarchy, font selection, labeling – that elevate visualizations. Additional topics will include cartography, web design, and interaction.

DATA 1720. Tackling Climate Change with Machine Learning (EEPS 1720).
Interested students must register for EEPS 1720.

DATA 2020. Statistical Learning.
A modern introduction to inferential methods for regression analysis and statistical learning, with an emphasis on application in practical settings in the context of learning relationships from observed data. Topics will include basics of linear regression, variable selection and dimension reduction, and approaches to nonlinear regression. Extensions to other data structures such as longitudinal data and the fundamentals of causal inference will also be introduced.

Spr	DATA2020	S01	26362	TTh	10:30-11:50(09)	(R. DeVito)

DATA 2040. Deep Learning and Special Topics in Data Science.
A hands-on introduction to neural networks, reinforcement learning, and related topics. Students will learn the theory of neural networks, including common optimization methods, activation and loss functions, regularization methods, and architectures. Topics include model interpretability, connections to other machine learning models, and computational considerations. Students will analyze a variety of real-world problems and data types, including image and natural language data.

DATA 2050. Data Science Practicum.
This course is a requirement for master's students in Data Science and is only open to them. The course includes a semester-long capstone project as well as instruction on topics that prepare students for working as data scientists, such as requirements gathering, version control, bug tracking, software deployment, and other professional development. Capstone projects will be sourced from entities in Brown (departments, labs, researchers, etc.) as well as external entities (companies, nonprofits, etc.) and will engage most of the core components of data science: Data sculpting (data cleaning, formatting, feature selection, etc.), exploratory data analysis, data modeling, and data visualization. Students will also address any social and ethical issues raised by their project. Students will usually work in teams of 3 - 4.

Fall	DATA2050	S01	18362	TTh	2:30-3:50(12)	(S. Pradhan)
Spr	DATA2050	S01	26364	W	3:00-5:30(10)	(S. Pradhan)

DATA 2060. Machine Learning: from Theory to Algorithms.
Data science techniques and tools are all around us. Machine learning is a term used across many different disciplines, and often people use machine learning tools without a thorough understanding of how and why the tools work. This course will provide a foundation of machine learning grounded in the mathematical models behind the techniques. We will cover the theory, computational methods, and visualization inherent in the application of machine learning models. Students will learn the statistical learning framework, common assumptions in the data generation process, the mathematics behind machine learning models, including supervised and unsupervised techniques, as well as how to implement machine learning models in Python from scratch. For DSI and CS master's students, no prerequisites required. Prerequisites for others are: MATH0520 (linear algebra) and one of APMA 1650, 1655, CSCI 1450, or DATA1030. Equivalencies will be considered by the instructor.

Fall	DATA2060	S01	18227	MW	3:00-4:20(10)	(A. Zsom)

DATA 2080. Data and Society.
A course on the social, political, and philosophical issues raised by the theory and practice of data science. Explores how data science is transforming not only our sense of science and scientific knowledge, but our sense of ourselves and our communities and our commitments concerning human affairs and institutions generally. Students will examine the field of data science in light of perspectives provided by the philosophy of science and technology, the sociology of knowledge, and science studies, and explore the consequences of data science for life in the first half of the 21st century.

DATA 2110. Topics in Econometrics.
This course will begin with a survey of the literature on identification using instrumental variables, including identification bounds, conditional moment restrictions, and control function approaches. The next part of class will cover some of the theoretical foundations of machine learning, including regularization and data-driven choice of tuning parameters. We will discuss in some detail the canonical normal means model, Gaussian process priors, (empirical) Bayes estimation, and reproducing kernel Hilbert space norms. We will finally cover some selected additional topics in machine learning, including (deep) neural nets, text as data (topics models), multi-armed bandits, and data visualization.

DATA 2980. Research in Data Science.
Section numbers vary by instructor. Please check Banner for the correct section number and CRN to use when registering for this course.