

# DATA SCIENCE

## Overview and Contact Information

The major in data science aims to guide students to be effective, ethical, and judicious consumers, analyzers and communicators of data and data related concepts. The major offers students a foundational understanding of the data generating process, the appropriate and efficient translation of analytic strategies to specific data settings, the potential biases arising from missing data or data collection, the means for drawing accurate conclusions, and the techniques and principles of integrity in data visualization and communication. As part of their data science education, students will develop excellent communication skills and the ability to make clear and persuasive arguments framed by logic and supported by data.

The curriculum is flexible and innovative, broad enough to serve a student population that is diverse in their backgrounds and disciplinary interests, and deep enough to accommodate students who want ultimately to pursue advanced study in statistics and computer science. The Data Science curriculum reflects the increasingly collaborative and interdisciplinary academic landscape.

## See Also

- Computer Science (<http://catalog.mtholyoke.edu/areas-study/computer-science/>)
- Statistics (<http://catalog.mtholyoke.edu/areas-study/statistics/>)
- Data Analytics and Society (<http://catalog.mtholyoke.edu/areas-study/data-analytics-society/>)

## Contact Information

Valerie Barr, Co-chair

Dylan Sheperdson, Co-chair

Connell Heady, Academic Department Coordinator

415A Clapp Laboratory  
413-538-2162

## Learning Goals

- Apply core concepts of statistics, computing, and domain knowledge to extract insight from data sets.
- Understand the ethical challenges and potential privacy issues involved in data analysis.
- Be able to communicate in multiple modalities the results of large scale data analysis.

## Faculty

**This area of study is administered by the Data Science Program Committee:**

Valerie Barr, Jean E. Sammet Professor of Computer Science

Maria Gomez, Elizabeth Page Greenawalt Professor of Chemistry

Benjamin Gebre-Medhin, Assistant Professor of Sociology

Marie Ozanne, Clare Boothe Luce Assistant Professorship in Statistics

## Requirements for the Major

A minimum of 40 credits:

Code	Title	Credits
STAT-140	Introduction to the Ideas and Applications of Statistics	4
STAT-242	Intermediate Statistics	4
STAT-340	Applied Regression Methods	4
As a prerequisite for MATH-211:		
MATH-102	Calculus II (or above)	
MATH-211	Linear Algebra	4
COMSC-151	Introduction to Computational Problem Solving <sup>1</sup>	4
COMSC-205	Data Structures <sup>2</sup>	4
COMSC-335	Machine Learning	4
or 300-level alternative to COMSC-335		
Two courses at the 200 level or above within a single domain area <sup>3</sup>		8
DATA-390	Data Science Capstone <sup>4</sup>	4
<b>Total Credits</b>		<b>40</b>

<sup>1</sup> The combinations of COMSC-150 and COMSC-121 or of COMSC-150 and COMSC-161 are equivalent to COMSC-151.

<sup>2</sup> The combination of COMSC-205PY and COMSC-122 is equivalent to COMSC-205.

<sup>3</sup> A domain area -- chosen in consultation with the student's Data Science advisor -- is defined as any College-defined major excluding mathematics, statistics, and computer science. Course selection must be approved by the student's Data Science advisor.

<sup>4</sup> The study of ethics in relation to data science is integrated throughout the curriculum and emphasized in this integrative capstone course.

## Other Requirements

- At the time of major declaration, a domain area will be selected by the student in consultation with an advisor from Data Science.
- Prior to the DATA-390 Capstone course, each Data Science major will submit to their advisor a brief document of reflection on the domain area, its connection to data science, and topics they might pursue for their capstone research. The Capstone will be offered in the spring term and be run as a research seminar.

## Additional Specifications

- Course substitutions through the Five Colleges require pre-approval in writing by an advisor from Data Science.
- Independent studies cannot be used to satisfy any of the above requirements unless approved by the Data Science Program Committee (with the possible exception of the capstone).
- Students who declare a Data Science major automatically fulfill the College's "outside the major" requirement.

## Sample Domain Pathways

At the time of major declaration, the student selects a domain area in consultation with an advisor from Data Science. Some sample pathways are described below:

### Chemistry

Analytical and physical chemists often generate and analyze significant amounts of data. Analysis methods learned in analytical or physical chemistry courses are regularly applied to organic, inorganic, or biochemical systems. Two course sequences highlighting both

methods and systems could include (a) a course in analytical or physical chemistry and (b) a course with a focus on organic, inorganic, or biochemical materials. More data generation and analysis based two course sequences can be two courses in analytical and/or physical chemistry. All first courses in the above sub-areas of chemistry CHEM-150 General Chemistry: Foundations of Structure and Reactivity and some also require CHEM-202 Organic Chemistry I and/or MATH-203 Calculus III.

### Economics

Data touches nearly all parts of economics by informing models and revealing patterns and causal relationships in data. Data science is becoming an essential part of every subfield in economics. For example students interested in: (1) finance might take ECON-270 Accounting and ECON-215 Economics of Corporate Finance; (2) development might take ECON-213 Economic Development and ECON-218 ; (3) theory might take ECON-201 Game Theory and ECON-212 Microeconomic Theory. Almost all 200-level courses in economics require ECON-110 Introductory Economics as a prerequisite.

### English

Digital Humanities and New Media Studies represent two humanities avenues for potential cross-pollination with data analysis. Topic modeling, text mining, and database construction for interactive editions of texts are examples of particular areas of digital humanities that lend themselves to asking interesting questions about large humanities corpora. Students interested in English and Data Science would take courses in literary analysis and at least one upper-level course in digital humanities in the Five Colleges. For example, students interested in: (1) text analysis of literature and the environment might take ENGL-231 British Romanticism: Revolution and Reaction or ENGL-240 American Literature I and ENGL-366 Love, Sex, and Death in the Anthropocene, or Living Through the Age of Climate Change and Other Disasters. Alternately; (2) exploring large corpora might take ENGL-280 and another survey course offering breadth, e.g., ENGL-251 Contemporary African American Literature II or ENGL-241 American Literature II , ENGLISH-302 (UMass) Studies in Textuality and New Media or ENGL-390 (Amherst) Digital Humanities. Ideally, students would also take ENGL-199 Introduction to the Study of Literature.

## Course Advice

The courses listed below form the core of the Data Science curriculum. In addition to core courses, students majoring in Data Science will take courses from their selected domain areas in consultation with their Data Science advisors.

## Course Offerings

### DATA-295 Independent Study

*Fall and Spring. Credits: 1 - 4*

*The department*

*Instructor permission required.*

### DATA-390 Data Science Capstone

*Spring. Credits: 4*

The Capstone is a research seminar that brings together the three pillars of the Data Science curriculum. The course will start with common readings about research projects across a range of disciplines, including readings that address issues of ethics involved with the collection, treatment, and analysis of data. Concurrently, each student will develop an individual research topic and identify relevant data resources. The remainder of the term will be dedicated to exploring these topics through extensive data analysis, visualization, and interpretation, leading to a final report with complete results and a presentation.

*Applies to requirement(s): Math Sciences*

*V. Barr*

*Prereq: COMSC-205 and STAT-340. STAT-340 may be taken concurrently (contact instructor for permission).*

### DATA-395 Independent Study

*Fall and Spring. Credits: 1 - 8*

*The department*

*Instructor permission required.*

### DATA-395P Independent Study w/Practicum

*Fall and Spring. Credits: 1 - 8*

*Instructor permission required.*

## Courses in Other Departments Counting toward the Major in Data Science

Code	Title	Credits
<b>Computer Science</b>		
COMSC-151AA	Introduction to Computational Problem Solving: 'Algorithmic Arts'	4
COMSC-151AR	Introduction to Computational Problem Solving: 'Artificial Intelligence'	4
COMSC-151DS	Introduction to Computational Problem Solving: 'Big Data'	4
COMSC-151EN	Introduction to Computational Problem Solving: 'Environmental Studies'	4
COMSC-151HC	Introduction to Computational Problem Solving: 'Humanities Computing'	4
COMSC-151MD	Introduction to Computational Problem Solving: 'Computers in Medical Technology'	4
COMSC-151SG	Introduction to Computational Problem Solving: 'Computing for Social Good'	4
COMSC-205	Data Structures	4
COMSC-335	Machine Learning	4
<b>Data Science</b>		
DATA-390	Data Science Capstone	4
<b>Mathematics</b>		
MATH-211	Linear Algebra	4
<b>Statistics</b>		
STAT-140	Introduction to the Ideas and Applications of Statistics	4
STAT-242	Intermediate Statistics	4
STAT-340	Applied Regression Methods	4