DATA SCIENCE

Overview

Data Science, the art and science of extracting information from large datasets, is an interdisciplinary field that offers many exciting and challenging opportunities. Formed from the amalgamation of Computer Science, Statistics, and Mathematics, Data Science aims to solve the world's problems by revealing information hidden in what has become known as "Big Data". As today's businesses and IT systems continue to produce massive and ever-increasing amounts of digital data, the need for data scientists is greater than ever. Whether you are interested in analyzing consumer transactions, tweets, call data records, text corpuses, or sounds in nature, or creating stunning data visualizations, you will find that the concepts, techniques, and tools covered in our Data Science program will be extremely employable in a wide range of industrial domains and disciplines where data scientists are in high demand. These skills will also form a strong foundation for advanced graduate studies.

The area of concentration starts with courses that form the foundational knowledge and skills in mathematics, statistics, computer programming, databases, and data munging, followed by more advanced courses on statistical models, algorithms, distributed computing, software engineering, and machine learning. Students will then have the option to carry out an applied data science thesis in any domain of their interest under the supervision of one or more interdisciplinary faculty, including from the Divisions of Natural Sciences, Humanities, or Social Sciences depending on the area of focus.

Faculty in Data Science

Melissa Crow (https://www.ncf.edu/directory/melissa-a-crow/), Instructor of Statistics
David Gillman (https://www.ncf.edu/directory/david-w-gillman/), Associate Professor of Computer Science
Bernhard Klingenberg (https://www.ncf.edu/directory/bernhard-klingenberg/), Professor of Statistics/Interim Director of Data Science
Patrick McDonald (https://www.ncf.edu/directory/patrick-t-mcdonald/), Professor of Mathematics/Vice Chair of the Faculty
Tiago Perez (https://www.ncf.edu/directory/tiago-perez/), Assistant Professor of Data Science
Eirini Poimenidou (https://www.ncf.edu/directory/eirini-poimenidou/), Professor of Mathematics (On Leave)
Tania Roy (https://www.ncf.edu/directory/tania-roy/), Associate Professor of Human Centered Computing
Andrey Skripnikov (https://www.ncf.edu/directory/andrey-skripnikov/), Assistant Professor of Statistics
Necmettin Yildirim (https://www.ncf.edu/directory/necmettin-yildirim/), Professor of Mathematics/Soo Bong Chae Chair of Applied Mathematics

Requirements for the AOC in Data Science

A minimum of sixteen (16) academic units.

Code	Title
Core Requirements	
MATH 2311	Calculus I*
CSCI 2200	Introduction to Programming in Python*
CSCI 3250	Intermediate Python
or CSCI 2400	Object-Oriented Programming
STAN 2700	Dealing with Data I*
STAN 2800	Dealing with Data II
MATH 2500 & MATH 3510	Probability I and Probability II* (Mods I & II)
MATH 3105	Linear Algebra
Data Science Area Courses	
DATA 3120	Algorithms for Data Science
DATA 4300	Databases for Data Science
DATA 3300	Software Engineering in Data Science
STAN 3275	Applied Linear Models
CSCI 4210	Artificial Intelligence and Data Mining
DATA 2400	Ethics in Data Science
Thesis Preparation Courses	
Select **three** elective courses either from either **Pool A** or **Pool B**: [1]	
Pool A:	
3xxx and 4xxx courses	In CSCI, STAN, MATH or the Graduate Program
DATA 3131	Distributed Computing
Pool B:	
3xxx and 4xxx courses	In Humanities, Social Sciences, or Natural Sciences not in **Pool A**
Additional Requirements	
Data Science Internship or Community Project [2]	
Senior Thesis or Senior Capstone Project in Data Science, and Baccalaureate Exam	

[1] Students will conduct their thesis either as a theoretical/methodological Data Science thesis, or as an applied Data Science thesis that combines skills acquired earlier in the program with skills and knowledge that will be gained by taking cross-disciplinary courses (e.g. courses from Humanities or Social Sciences). Hence, the student is expected to select all three elective courses either from **Pool A** or from **Pool B**.

[2] Data Science is a practical field. As such, each AOC student is expected to do an internship or a community project in applied Data Science, preferably following the completion of their third year in the program. The internship or project topic must be approved by the student's advisor or internship coordinator.

Requirements for a Secondary Field in Data Science

A minimum of eight (8) academic units.

Code	Title
Core Requirements	
CSCI 2200	Introduction to Programming in Python*
CSCI 3250	Intermediate Python
or CSCI 2400	Object-Oriented Programming
STAN 2700	Dealing with Data I*
STAN 2800	Dealing with Data II
DATA 4300	Databases for Data Science

CSCI 4210	Artificial Intelligence and Data Mining
Electives	
Select **two** from the following examples:	
DATA 3120	Algorithms for Data Science
DATA 3300	Software Engineering in Data Science
STAN 3275	Applied Linear Models
DATA 3131	Distributed Computing
STAN 3000	Statistical Learning
STAN 3230	Data Visualization and Communication

Sample Pathways

The sample pathway starts with the first-year introductory courses for Data Science including three courses that involve programming in Python and R and a two-module course sequence on probability. These courses are intended to provide prospective Data Science students an initial view into the discipline and allow them to decide whether they would like to pursue the AOC. In the second year, students are expected to take the remaining foundation courses (Calculus and Linear Algebra), a Python continuation course, and also three of the core courses of Data Science (Databases, Algorithms, and Software Engineering). With this background, students can go on to take the remaining core courses of Data Science and elective courses oriented towards their thesis.

Sample Four-Year Pathway

First Year

Fall Term	Spring Term
Dealing with Data 1	Dealing with Data 2
Intro. to Programming in Python	CYC 1
Probability 1 & 2	CYC 2

Second Year

Fall Term	Spring Term
Calculus 1	Linear Algebra
Database for Data Science	Algorithms for Data Science
Intermediate Python **or** Object-Oriented Programming	Software Eng. for Data Science

Third Year

Fall Term	Spring Term	Summer
Applied Linear Models	Artificial Intelligence and Data Mining	Internship or Community Project
Ethics in Data Science	Elective 2	
Elective 1	Elective 3	

Fourth Year

Fall Term	ISP	Spring Term
Thesis	Thesis	Thesis

Sample Two-Year Pathway

This pathway assumes a student has completed two statistics courses, two programming courses (at least one in Python), Calculus 1, and Linear Algebra.

First Year

Fall Term	Spring Term	Summer
Probability 1 & 2	Algorithms for Data Science	Internship or Community Project
Database for Data Science	Software Eng. for Data Science	
Ethics in Data Science	Elective 1	

Second Year

Fall Term	ISP	Spring Term
Applied Linear Models	Thesis	Artificial Intelligence and Data Mining
Elective 2		Elective 3
Thesis		Thesis

Requirements for 3+2 Pathway for Combined Undergraduate + Graduate Degrees (BA and MS in Data Science)

This pathway is intended for high-performing students who aspire to complete a combined sequence of undergraduate + graduate studies faster than the normal duration of 6 years. This is for current and future New College majors (other than Data Science) who would like to pursue a graduate degree in Data Science. Undergraduate students in this track can take additional courses in their third and fourth years from the Data

Science graduate program, followed by the second and final year of the Graduate Program itself, earning the two degrees by the end of fifth year.

A student is eligible for this pathway after entering the undergraduate program and showing sufficiently high performance. In other words, acceptance into this 3+2 pathway is not automatically granted at the time of undergraduate admission; students will have to apply and seek admission only after they satisfy certain minimum conditions:

- Complete 2 years of study with no Unsatisfactory grade
- Complete prerequisite courses (see below)
- Be recommended for the 3+2 pathway by a faculty member

The Data Science Graduate Program admissions committee will also review applications for this pathway and make admission decisions. Other application requirements of the Graduate Program will still apply.

Code	Title
Prerequisites	
In addition to the chosen AOC requirements, the following courses must be completed during the first two years of undergraduate study: [1]	
MATH 2311	Calculus I*
MATH 2312	Calculus II*
CSCI 2200	Introduction to Programming in Python*
CSCI 3250	Intermediate Python
or CSCI 2400	Object-Oriented Programming
MATH 2500	Probability I
MATH 3510	Probability II*
MATH 3105	Linear Algebra
Third Year: Fall Term	
IDC 5204	Applied Statistics I
IDC 5110	Data Munging and Exploratory Data Analysis
Fourth Year: Fall Term [2]	
IDC 5120	Algorithms for Data Science
IDC 5130	Databases for Data Science
IDC 5251	Industrial Seminar Series I
Fourth Year: January Interterm	
IDC 5295	Industrial Workshops
Fourth Year: Spring Term	
IDC 5205	Applied Statistics II
IDC 5112	Data Visualization
IDC 5210	Applied Machine Learning
IDC 5131	Distributed Computing
IDC 5297	Industrial Seminar Series II
Fourth Year: Summer or Fifth Year: January Interterm	
IDC 6293	Industrial Practicum I
Fifth Year: Fall Term	
IDC 6200	Advanced Applied Statistics
IDC 6215	Advanced Applied Computing
IDC 6250	Practical Data Science
IDC 6253	Industrial Seminar Series III
Fifth Year: Spring Term	
IDC 6294	Industrial Practicum II

1 If the student's AOC already includes some or all of the prerequisite courses, these courses can be counted towards fulfilling the prerequisite course requirements for the 3+2 track. They also count towards satisfying the IDC 5100 Introduction to Data Science Bootcamp course in the graduate program. However, IDC courses must be taken in addition to the AOC requirements and can only be counted towards the graduate program requirements.

2 The undergraduate program is completed after the Fourth Year Fall Term.

Data Science Facilities

New College has a number of servers that support students and faculty in the computer science and data science programs. These include 5 HP physical servers with NVIDIA graphics processing units (Tesla, Titan X, and 1080 Ti); 1 SuperMicro physical server with 4 NVIDIA graphics processing units (Quadro RTX 6000); 1 SuperMicro physical server with 4 NVIDIA graphics processing units (RTX A5000 and 1080 Ti); and 12 virtual servers used in a variety of computer science, data science, and statistics courses.