Data Science

College of Engineering

The College of Engineering offers a Master of Science degree in Data Science with a concentration in Biomedical Informatics.

Application Requirements

- CV
- Personal Statement
- GRE or GMAT
- Three letters of recommendation
- Additional requirements:
    - A GPA of at least 3.0 on a 4.0 scale
    - Successful completion of the following courses: multivariate calculus (MA 213), linear algebra (MA 322), and computer programming (CS 215)
    - Exceptions may be made if persuasive evidence of applicant's potential is presented. Students without sufficient mathematical or programming background will be considered for admission but may be required to take preparatory coursework in addition to required courses.

Application Deadlines

- Spring: December 9 (domestic applicants), October 15 (international applicants)
- Fall: July 16 (domestic applicants), March 15 (international applicants)

Degree Requirements

Core Requirement:

Earn a B or better in each core course in the program (see table below).

Course Number	Course Title		Credit
CS626	Large Scale Data Science	Core	3
CPH580/STA580	Biostatistics I	Core	3
CPH630/STA681	Biostatistics II	Core	3

DS710	Data Science Seminar (three credits of this course are required; will be repeated three times)	Core	1
DS711	MS project in Data Science	Core	3

Guided Electives

Each student completes at least one course from the following list.

Course Number	Course Title		Credit
CS405G	Introduction to Database Systems	Elective	3
CS460G	Machine Learning	Elective	3
CS515	Algorithm Design	Elective	3

Concentration Requirement

Each student completes BMI633 and one of the elective courses listed below.

Course Number	Course Title		Credit
BMI633	Introduction to Bioinformatics	Core	3
BMI730	Principles of Clinical Informatics	Elective	3
BMI733	Biomedical Natural Language Processing	Elective	3
BMI734	Introduction to Biomedical Image Analysis	Elective	3

Elective Requirement

Each student completes three elective courses (9 credit hours), with at least two courses having a significant data science component in the subject of the student's MS project.

Project Requirement

Complete an MS project, which will be supervised by a faculty member, that includes a written report and oral presentation to a committee.

Course Descriptions

BMI 633 INTRODUCTION TO BIOINFORMATICS. (3) This is an introductory course aimed at a multi-disciplinary audience with an interest in applying the principles of information sciences for obtaining insight into biological processes and systems that can eventually be used to make informed decisions.

BMI 730 PRINCIPLES OF CLINICAL INFORMATICS. (3) This course offers an overview of Clinical informatics, which is the application of informatics principles, methods, and tools to support healthcare practice and research activities as well as business processes.

BMI 733 BIOMEDICAL NATURAL LANGUAGE PROCESSING. (3) This course is a technical introduction to the area of biomedical natural language processing (NLP). In the field of biomedical informatics, this focuses on the common steps in extracting information from textual data that arises from biomedical literature and clinical documents. Topics involve n-gram models, tokenization, POS tagging, and parsing. Prereq: MA 123 (or equivalent) or consent of instructor.

BMI 734 INTRODUCTION TO BIOMEDICAL IMAGE ANALYSIS. (3) This class aims to give students a broad overview of biomedical image analysis and imaging informatics. We will introduce the state-of-the-art knowledge to understand, develop, and apply existing methods and software to handle biomedical image data to extract quantitative matrices.

CPH 580 BIOSTATISTICS I. (3) CPH 580 covers univariate statistical methods commonly encountered in public health studies. This includes descriptive statistics, hypothesis testing, paired and unpaired t tests, ANOVA, contingency tables, log rank test, regression and correlation. Prereq: MA 109 or higher. (Same as STA 580.)

CPH 630 BIOSTATISTICS II. (3) Students will learn statistical methods used in public health studies. This includes receiver operator curves, multiple regression, logistic regression, confounding and stratification, the Mantel-Haenzel procedure, and the Cox proportional hazards model. Prereq: STA 570, CPH 603, STA 580/CPH 580, or equivalent. (Same as STA 681.)

CS 626 LARGE SCALE DATA SCIENCE. (3) This course will offer an opportunity for students to learn big data techniques and apply them to tackle real-world data science challenges (e.g., processing, storing, querying, exploring, and mining big data). Topics include big data systems and programming models, parallel computing framework, scalable data management and processing solutions, scalable data mining techniques for large datasets, and advanced applications. Prereq: CS 505 or consent of the instructor.

CS 405G INTRODUCTION TO DATABASE SYSTEMS. (3) Study of fundamental concepts behind the design, implementation and application of database systems. Brief review of entity relationship, hierarchical and network database models and an in-depth coverage of the relational model including relational algebra and calculi, relational database theory, concepts in schema design and commercial database languages. Prereq: CS 315 and graduate or engineering standing.

CS 460G MACHINE LEARNING. (3) Study of computational principles and techniques that enable software systems to improve their performance by learning from data. Focus on fundamental algorithms, mathematical models and programming techniques used in Machine Learning. Topics include: different learning settings (such as supervised, unsupervised and reinforcement learning), various learning algorithms (such as decision trees, neural networks, k-NN, boosting, SVM, k-means) and crosscutting issues of generalization, data representation, feature selection, model fitting and optimization. The course covers both theory and practice, including programming and written assignments that utilize concepts covered in lectures. Prereq: Strong programming ability (CS 315), basic probability and statistics (STA 281), and basic concepts of linear algebra (MA/CS 321 or MA/CS 322), or instructor's consent.

CS 515 ALGORITHM DESIGN. (3) The design and analysis of efficient algorithms on data structures for problems in sorting, searching, graph theory, combinatorial optimization, computational geometry, and algebraic computation. Algorithm design techniques: divide-and-conquer, dynamic programming, greedy method, and randomization, approximation algorithms. Prereq: CS 315 and engineering standing.

DS 710 RESEARCH SEMINAR IN DATA SCIENCE. (1) An independent study course that aims to expose students to a broad range of research and professional topics in data science, and enhance students' ability to do research. Students read research papers and professional articles, attend colloquium/seminar talks on topics in data science as directed by the instructor, prepare written reports on topics and make in-class presentations. Prereq: Enrollment in the MS or PhD program in Data Science.

DS 711 MASTER'S PROJECT IN DATA SCIENCE. (3) Design and implementation of a large data science project under the supervision of a faculty member. Prereq: The course is open only to students in the MS program in Data Science who completed the program's core requirement.