MS IN DATA SCIENCE CURRICULUM

PROGRAM REQUIREMENTS

PREREQUISITE COURSES

Students are expected to satisfy the prerequisite level of knowledge needed to succeed in the Data Science program, and any pre-requisites for specific courses taken during the course of study. This includes calculus and linear algebra, and some prior knowledge of probability and statistics, as well as some programming and database experience.

Students who do not have the required background in programming and databases are expected to take the following courses:

- CS 623 – Database Management Systems
- CS 632P – Python Programming

Students will take seven required courses and three electives. The Analytics Capstone Project is a project-based course that will bring together the different knowledge, ideas and tools learned throughout the program. Students will have flexibility in choosing their electives, thus allowing a limited amount of customization. Only one elective course (maximum three credits) outside of computer science is permitted.

MS IN DATA SCIENCE DEGREE REQUIREMENTS

All courses are required, and it is presumed that students will take the courses in consecutive terms.

Fall 1 (9 credits)	**Spring 1 (9 credits)**
CS 673 Scalable Databases	CS 676 Algorithms for Data Science
CS 675 Introduction to Data Science	CS 619 Data Mining
CS 660 Mathematical Foundations of Analytics	CS 632M Machine Learning
Fall 2 (9 credits)	**Spring 2 (3 credits)**
Elective	CS 668 Analytics Capstone Project
Elective	
Elective	

Total Graduate Credits (<u>excluding</u> prerequisites and bridge courses): 30

Required Courses

Detailed course descriptions of the required courses follow:

CS 673: Scalable Databases

After reviewing relational databases and SQL, students will learn the fundamentals of alternative data storage schemas to deal with large amounts of data (structures and unstructured). The course covers big data and the development of the Hadoop file system, the MapReduce programming paradigm, and database management systems such as Cassandra, HBase and Neo4j. Students will learn about NoSQL, distributed databases, and graph databases. The course emphasizes the differences between traditional database management systems and alternatives with respect to accessibility, cost, transaction speed, and structure. Part of the course is dedicated to access, handle and process data from different sources and of different types using Python.
The course provides hands-on practice.

CS 675: Introduction to Data Science

This course introduces the concepts of data science. The course teaches students the interdisciplinary basis of data science, and the data science process. Additionally the course covers data visualization, data wrangling, ethics of designing and conducting data analysis and research, bias in research, data privacy issues surrounding the use of data, and, research reproducibility. Students will learn about statistical learning methods, and then move on to more advanced topics including: database queries, working with spatial data, text mining, networks and big data. The course also emphasizes writing technical reports and presenting results. The course prepares students for further study in data mining, machine learning, and artificial intelligence, and introduces students to R.

CS 660: Mathematical Foundations of Analytics

This course teaches students the fundamental probability and mathematical statistics necessary for further study in data science. Students learn about probability theory, univariate and multivariate distributions, law of large numbers, the central limit theorem, statistical inference, point and confidence interval estimation, hypothesis testing, linear models including least squares and logistic regression, model selection and assessment, Bayesian statistics, and time series analysis. The emphasis of this course is on theoretical foundations.

CS 676: Algorithms for Data Science

This course focuses on the efficiency and complexity of algorithms needed for data analytics, and has a computational emphasis. Students will develop proficiency in Python and R, as they build algorithms and analyze data. Topics include data reduction: data mapping, data dictionaries, scalable algorithms, Hadoop, and MapReduce; gaining information from data: data visualization, regression modeling, and cluster analysis; and, predictive analytics: k-nearest neighbors, naïve Bayes, time series forecasting, and analyzing streaming data, and optimization with gradient descent.

CS 619: Data Mining

This course will provide an overview of topics such as introduction to data mining and knowledge discovery; data mining with structured and unstructured data; foundations of pattern clustering; clustering paradigms; clustering for data mining; data mining using neural networks and genetic algorithms; fast discovery of association rules; applications of data mining to pattern classification; and feature selection. The goal of this course is to introduce students to current machine learning and related data mining methods. It is intended to provide enough background to allow students to apply machine learning and data mining techniques to learning problems in a variety of application areas.

CS 632M: Machine Learning

This course teaches students machine learning theory and algorithms. Students will learn about probably approximately correct, empirical risk minimization, structural risk minimization, and minimum description length learning rules. Students will then study various machine learning algorithms, such as linear models, gradient descent, support vector machines, kernel methods, and trees, and how they connect to the theoretical framework. Finally the course culminates with additional topics such as clustering, dimensionality reduction, generative models, and feature selection.

CS 668: Analytics Capstone Project:

The purpose of the Capstone Project is for the students to apply the knowledge and skills acquired during the Data Science program to a project involving actual data in a real-world setting. During the project, students will apply the entire data science process from identifying a problem or opportunity, and collecting and processing actual data to applying suitable and appropriate analytic methods to find a solution. Both the problem statement for the project and the datasets will come from real-world domains similar to those that students might typically encounter within industry, government, or academic research. The course will culminate with each student making a presentation of his or her work, and submitting a final paper. This is a largely self-directed course, with guidance and suggestions provided along the way by the instructor.

Elective Courses

Students will take nine credits in elective courses selected from computer science courses, information systems courses, business analytics courses (finance, economics, marketing), and natural sciences courses (environmental science, biochemistry molecular-biology). Only one elective course (maximum three credits) outside of computer science is permitted.