Data Science (DS)

DS 1990. Elective. 1-4 Hours.
Offers elective credit for courses taken at other academic institutions. May be repeated without limit.

DS 2000. Programming with Data. 2 Hours.
Introduces programming for data and information science through case studies in business, sports, education, social science, economics, and the natural world. Presents key concepts in programming, data structures, and data analysis through Python and Excel. Integrates the use of data analytics libraries and tools. Surveys techniques for acquiring and programmatically integrating data from different sources. Explains the data analytics pipeline and how to apply programming at each stage. Discusses the programmatic retrieval of data from application programming interfaces (APIs) and from databases. Introduces predictive analytics for forecasting and classification. Demonstrates the limitations of statistical techniques.

DS 2001. Practicum for DS 2000. 2 Hours.
Accompanies DS 2000. Applies topics from the course through various experiments and in a variety of contexts.

DS 2990. Elective. 1-4 Hours.
Offers elective credit for courses taken at other academic institutions. May be repeated without limit.

DS 3990. Elective. 1-4 Hours.
Offers elective credit for courses taken at other academic institutions. May be repeated without limit.

DS 4100. Data Collection, Integration, and Analysis. 4 Hours.
Studies how to collect data from multiple sources and integrate them into consistent data sets. Covers how to use semi-automated and automated classification to integrate disparate data sets; how to parse data from files, XML, JSON, APIs, and structured data stores to construct analyzable data sets that are stored in databases; and how to assess and ensure quality of data. Introduces key concepts of algorithms and data structures, including divide-and-conquer, sorting and selection, and graph traversal and descriptive analysis of data through descriptive statistics and plotting. Analyzes complexity and run-time behavior of programs. Presents approaches for data anonymization and protecting data privacy. Studies data shaping and manipulation techniques for data analysis and the R and Python programming languages.

DS 4200. Information Presentation and Visualization. 4 Hours.
Introduces foundational principles, methods, and techniques of visualization to enable creation of effective information representations suitable for exploration and discovery. Covers the design and evaluation process of visualization creation, visual representations of data, relevant principles of human vision and perception, and basic interactivity principles. Studies data types and a wide range of visual data encodings and representations. Draws examples from physics, biology, health science, social science, geography, business, and economics. Emphasizes good programming practices for both static and interactive visualizations. Creates visualizations in Excel and Tableau as well as R, Python, and open web-based authoring libraries. Requires programming in Python, JavaScript, HTML, and CSS. Requires extensive writing including documentation, explanations, and discussions of the findings from the data analyses and the visualizations.

DS 4300. Large-Scale Information Storage and Retrieval. 4 Hours.
Introduces data and information storage approaches for structured and unstructured data. Covers how to build large-scale information storage structures using distributed storage facilities. Explores data quality assurance, storage reliability, and challenges of working with very large data volumes. Studies how to model multidimensional data. Implements distributed databases. Considers multitier storage design, storage area networks, and distributed data stores. Applies algorithms, including graph traversal, hashing, and sorting, to complex data storage systems. Considers complexity theory and hardness of large-scale data storage and retrieval. Requires use of nonrelational, document, key-column, key-value, and graph databases and programming in R, Python, and C++.

DS 4400. Machine Learning and Data Mining 1. 4 Hours.
Introduces supervised and unsupervised predictive modeling, data mining, and machine-learning concepts. Uses tools and libraries to analyze data sets, build predictive models, and evaluate the fit of the models. Covers common learning algorithms, including dimensionality reduction, classification, principal-component analysis, k-NN, k-means clustering, gradient descent, regression, logistic regression, regularization, multiclass data and algorithms, boosting, and decision trees. Studies computational aspects of probability, statistics, and linear algebra that support algorithms, including sampling theory and computational learning. Requires programming in R and Python. Applies concepts to common problem domains, including recommendation systems, fraud detection, or advertising.

DS 4420. Machine Learning and Data Mining 2. 4 Hours.
Continues with supervised and unsupervised predictive modeling, data mining, and machine-learning concepts. Covers mathematical and computational aspects of learning algorithms, including kernels, time-series data, collaborative filtering, support vector machines, neural networks, Bayesian learning and Monte Carlo methods, multiple regression, and optimization. Uses mathematical proofs and empirical analysis to assess validity and performance of algorithms. Studies additional computational aspects of probability, statistics, and linear algebra that support algorithms. Requires programming in R and Python. Applies concepts to common problem domains, including spam filtering.

DS 4900. Data Science Senior Project. 4 Hours.
Designed to help students develop a sophisticated understanding of data collection, integration, storage, statistical analysis, visualization, and machine-supported analysis and modeling. Requires students to analyze a substantial data set using statistical and visual methods and to build machine-learning models to discover patterns in the data. Results must be communicated in writing. Requires substantial programming in R, Python, Java, or C++.

DS 4990. Elective. 1-4 Hours.
Offers elective credit for courses taken at other academic institutions. May be repeated without limit.

DS 4991. Research. 4 Hours.
Offers an opportunity to conduct research under faculty supervision. .

DS 4992. Directed Study. 1-4 Hours.
Offers independent work under the direction of members of the department on a chosen topic. May be repeated without limit.

DS 4993. Independent Study. 1-4 Hours.
Offers independent work under the direction of members of the department on a chosen topic. May be repeated without limit.

DS 4994. Internship. 4 Hours.
Offers students an opportunity for internship work. May be repeated without limit.

DS 4996. Experiential Education Directed Study. 1-4 Hours.
Draws upon the student's approved experiential activity and integrates it with study in the academic major. Restricted to those students who are using it to fulfill their experiential education requirement. May be repeated without limit.

DS 4997. Data Science Thesis. 4 Hours.
Offers students an opportunity to prepare an undergraduate thesis under faculty supervision. .

DS 4998. Data Science Thesis Continuation. 4 Hours.
Focuses on student continuing to prepare an undergraduate thesis under faculty supervision. .

DS 5010. Introduction to Programming for Data Science. 4 Hours.
Offers an introductory course on fundamentals of programming and data structures. Covers lists, arrays, trees, hash tables, etc.; program design, programming practices, testing, debugging, maintainability, data collection techniques, and data cleaning and preprocessing. Includes a class project, where students use the concepts covered to collect data from the web, clean and preprocess the data, and make it ready for analysis.

DS 5020. Introduction to Linear Algebra and Probability for Data Science. 4 Hours.
Offers an introductory course on the basics of statistics, probability, and linear algebra. Covers random variables, frequency distributions, measures of central tendency, measures of dispersion, moments of a distribution, discrete and continuous probability distributions, chain rule, Bayes' rule, correlation theory, basic sampling, matrix operations, trace of a matrix, norms, linear independence and ranks, inverse of a matrix, orthogonal matrices, range and null-space of a matrix, the determinant of a matrix, positive semidefinite matrices, eigenvalues, and eigenvectors.

DS 5110. Introduction to Data Management and Processing. 4 Hours.
Discusses the practical issues and techniques for data importing, tidying, transforming, and modeling. Offers a gentle introduction to techniques for processing big data. Programming is a cross-cutting aspect of the course. Offers students an opportunity to gain experience with data science tools through short assignments. Course work includes a term project based on real-world data. Covers data management and processing—definition and background; data transformation; data import; data cleaning; data modeling; relational and analytic databases; basics of SQL; programming in R and/or Python; MapReduce fundamentals and distributed data management; data processing pipelines, connecting multiple data management and analysis components; interaction between the capabilities and requirements of data analysis methods (data structures, algorithms, memory requirements) and the choice of data storage and management tools; and repeatable and reproducible data analysis.

DS 5220. Supervised Machine Learning and Learning Theory. 4 Hours.
Introduces supervised machine learning, which is the study and design of algorithms that enable computers/machines to learn from experience or data, given examples of data with a known outcome of interest. Offers a broad view of models and algorithms for supervised decision making. Discusses the methodological foundations behind the models and the algorithms, as well as issues of practical implementation and use, and techniques for assessing the performance. Includes a term project involving programming and/or work with real-life data sets. Requires profiency in a programming language such as Python, R, or MATLAB.

DS 5230. Unsupervised Machine Learning and Data Mining. 4 Hours.
Introduces unsupervised machine learning and data mining, which is the process of discovering and summarizing patterns from large amounts of data, without examples of data with a known outcome of interest. Offers a broad view of models and algorithms for unsupervised data exploration. Discusses the methodological foundations behind the models and the algorithms, as well as issues of practical implementation and use, and techniques for assessing the performance. Includes a term project involving programming and/or work with real-life data sets. Requires profiency in a programming language such as Python, R, or MATLAB.

DS 5500. Information Visualization: Applications in Data Science. 4 Hours.
Offers students an opportunity to develop effective communication skills with data by drawing from different disciplines including physics, biology, health science, social science, geography, business, and economics. Introduces principles of effective oral and written communication and a wide range of visual data encodings and representations. Covers the foundational principles for visual representations, including human vision and perception and basic interactivity. A semester-long project requires students to translate the domain science or technology problem into the language of data science; design, evaluate, implement, and deploy both static and interactive visualizations of data and data analysis results; translate the results into the language of the original science or technology problem; communicate the findings in oral and written form; and provide constructive criticism of other examples of data communication and visualization.

DS 7995. Project. 1-4 Hours.
Offers students an opportunity to participate in a direct data science project under the supervision of a faculty member. May be repeated once for a total of 8 credits.