Data Science (DS)

Search DS Courses using FocusSearch (http://catalog.northeastern.edu/class-search/?subject=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.

Corequisite(s): DS 2001
Attribute(s): NUpath Analyzing/Using Data

DS 2001. Data Science Programming Practicum. (2 Hours)
Applies data science principles in interdisciplinary contexts, with each section focusing on applications to a different discipline. Involves new experiments and readings in multiple disciplines (both computer science and the discipline focus of the particular section). Requires multiple projects combining interdisciplinary subjects.

Corequisite(s): DS 2000

DS 2500. Intermediate Programming with Data. (4 Hours)
Offers intermediate to advanced Python programming for data science. Covers object-oriented design patterns using Python, including encapsulation, composition, and inheritance. Advanced programming skills cover software architecture, recursion, profiling, unit testing and debugging, lineage and data provenance, using advanced integrated development environments, and software control systems. Uses case studies to survey key concepts in data science with an emphasis on machine-learning (classification, clustering, deep learning); data visualization; and natural language processing. Additional assigned readings survey topics in ethics, model bias, and data privacy pertinent to today's big data world. Offers students an opportunity to prepare for more advanced courses in data science and to enable practical contributions to software development and data science projects in a commercial setting.

Prerequisite(s): DS 2000 with a minimum grade of D-
Corequisite(s): DS 2501
Attribute(s): NUpath Analyzing/Using Data

DS 2501. Lab for DS 2500. (1 Hour)
Practices the programming techniques discussed in DS 2500 through hands-on experimentation.

Corequisite(s): DS 2500

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

DS 2991. Research in Data Science. (1-4 Hours)
Offers an opportunity to conduct introductory-level research or creative endeavors under faculty supervision.

DS 3000. Foundations of Data Science. (4 Hours)
Introduces core modern data science technologies and methods that provide a foundation for subsequent Data Science classes. Covers: working with tensors and applied linear algebra in standard numerical computing libraries (e.g., NumPy); processing and integrating data from a variety of structured and unstructured sources; introductory concepts in probability, statistics, and machine learning; basic data visualization techniques; and now standard data science tools such as Jupyter notebooks.

Prerequisite(s): CS 2510 with a minimum grade of D- or DS 2500 with a minimum grade of D-
Attribute(s): NUpath Analyzing/Using Data, NUpath Natural/Designed World

DS 3500. Advanced Programming with Data. (4 Hours)
Offers a deep dive into the design and implementation of enterprise-grade software systems with an emphasis on software architectures for more complex data-driven applications. Covers extensible architectures that support testing, data provenance, reuse, maintainability, scalability, and robustness and building software APIs and libraries for wide-scale adoption and ease of use. Students design, implement, and test complex loosely coupled service-oriented architectures using distributed processing, stream-based data processing, and interprocess communication via message passing. Explores the features, capabilities, and underlying design of popular data analysis and visualization frameworks.

Prerequisite(s): DS 2500 with a minimum grade of D-

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

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.

Prerequisite(s): CS 2510 with a minimum grade of D- or DS 2500 with a minimum grade of D-
Attribute(s): NUpath Analyzing/Using Data, NUpath Writing Intensive

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++.

Prerequisite(s): CS 3200 with a minimum grade of D- ; (DS 4100 with a minimum grade of D- or DS 3000 with a minimum grade of D-)
Attribute(s): NUpath Analyzing/Using Data

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.

Prerequisite(s): ((DS 4100 with a minimum grade of D- or DS 3000 with a minimum grade of D-); (CS 2810 with a minimum grade of D- or ECON 2350 with a minimum grade of D- or ENVR 2500 with a minimum grade of D- or MATH 3081 with a minimum grade of D- or MGSC 2301 with a minimum grade of D- or PHTH 2210 with a minimum grade of D- or PSYC 2320 with a minimum grade of D-)) or (CS 2810 with a minimum grade of D- ; CS 3500 with a minimum grade of D-)
Attribute(s): NUpath Analyzing/Using Data, NUpath Capstone Experience, NUpath Writing Intensive

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.

Prerequisite(s): DS 4400 with a minimum grade of D-
Attribute(s): NUpath Analyzing/Using Data, NUpath Capstone Experience, NUpath Writing Intensive

DS 4440. Practical Neural Networks. (4 Hours)
Offers a hands-on introduction to modern neural network ("deep learning") tools and methods. Covers the fundamentals of neural networks and introduces standard and new architectures from simple feed forward networks to recurrent neural networks. Also covers stochastic gradient descent and backpropagation, along with related fitting techniques. Emphasizes using these technologies in practice, via modern toolkits. Specifically introduces Keras (together with TensorFlow) and PyTorch, which are illustrative of static and dynamic network implementations, respectively. Reviews applications of these models to various types of data, including images and text.

Prerequisite(s): DS 4400 (may be taken concurrently) with a minimum grade of D-
Attribute(s): NUpath Analyzing/Using Data

DS 4970. Junior/Senior Honors Project 1. (4 Hours)
Focuses on in-depth project in which a student conducts research or produces a product related to the student's major field. Combined with Junior/Senior Project 2 or college-defined equivalent for 8 credit honors in the discipline project.

DS 4971. Junior/Senior Honors Project 2. (4 Hours)
Focuses on second semester of in-depth project in which a student conducts research or produces a product related to the student's major field.

Prerequisite(s): DS 4970 with a minimum grade of D-

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.

Attribute(s): NUpath Integration Experience

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.

Attribute(s): NUpath Integration Experience

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.

Attribute(s): NUpath Integration Experience

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)
Introduces students to the core tasks in data science, including data collection, storage, tidying, transformation, processing, management, and modeling for the purpose of extracting knowledge from raw observations. Programming is a cross-cutting aspect of the course. Offers students an opportunity to gain experience with data science tasks and tools through short assignments. Includes a term project based on real-world data.

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-world data sets. Requires proficiency 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 proficiency in a programming language such as Python, R, or MATLAB.

DS 5500. Capstone: Applications in Data Science. (4 Hours)
Offers students a capstone opportunity to practice data science skills learned in previous courses and to build a portfolio. Students practice visualization, data wrangling, and machine learning skills by applying them to semester-long term projects on real-world data. Students may either propose their own projects or choose from a selection of industry options. Emphasizes the overall data science process, including identification of the scientific problem, selection of appropriate machine learning methods, and visualization and communication of results. Lectures may include additional topics, including visualization, communication, and data science ethics.

Prerequisite(s): (CS 5800 with a minimum grade of C- or EECE 7205 with a minimum grade of C-); DS 5110 with a minimum grade of C- ; DS 5220 with a minimum grade of C- ; DS 5230 (may be taken concurrently) with a minimum grade of C-

DS 6050. Seminar in Data Science. (4 Hours)
Offers students an opportunity to learn how to approach data analysis problems in a systematic manner and to learn how to design data analysis pipelines, as well as how to implement them at scale in the context of real-world problems. Data science is at the intersection of statistics, machine learning, and software development. Data analysis problems are solved in a series of datacentric steps: data acquisition, data cleaning, data transformation, data modelling, and data visualization.

DS 6962. Elective. (1-4 Hours)
Offers elective credit for courses taken at other academic institutions. May be repeated without limit.

DS 7990. Thesis. (4 Hours)
Offers selected work with the agreement of a project supervisor.

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.

DS 8982. Readings. (1-8 Hours)
Offers selected readings under the supervision of a faculty member. May be repeated without limit.