DS4100: Data Collection, Integration and Analysis

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

Pre-Requisites: CS2510

NU Core Designations: AD

After completing this course, students will be able to:

- collect and integrate data from a variety of sources, including Web APIs, XML, CSV, and JSON
- assess the quality of collected data
- devise quality assurance plans
- select data structured appropriate for storing data for analysis
- design and develop programs in R and Python
- transform data objects into representations that are amenable to statistical analysis
- evaluate data structures based on time versus space trade-off considerations
- anonymize data to protect privacy of data
- sort, search, and traverse internal data structures

Achievement of learning outcomes will be assessed through:

- completion of programming assignments in R and Python
- development of a substantial term project requiring collection, integration, traversal, and descriptive analysis of a significant data set
- contributions to discussions
- quizzes, mid-term, and final exams

DS4200: Information Presentation and Visualization

Introduce 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. Teaches 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. Create 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.

Pre-Requisites: DS4100 [May be taken concurrently]

NU Core Designations: AD, WI

After completing this course, students will be able to:

- Assess the quality and effectiveness of a visualization
- Choose appropriate visualization methods for a given data type
- Design an effective visualization by applying design and human perception principles
- Implement a static or interactive visualization
- Implement basic interactivity functions to enable data exploration
- Manipulate data and implement reproducible visualizations in Python and R
- Implement web-based visualizations in JavaScript/HTML/CSS
- Create visualizations in Excel and Tableau

Achievement of learning outcomes will be assessed through:

- Completion of assignments using Python, R, and JavaScript/HTML/CSS using standard libraries, Tableau, and Excel
- Written design critiques and re-design proposals of visualizations
- Completion of a final project that requires the gathering, clean-up, and reduction of data presented in a new web-based interactive visualization
- Mid-term examination

DS4300: Large-Scale Information Storage & Retrieval

Introduces data and information storage approaches for structured and unstructured data. Shows 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. Requires use of non-relational, document, key-column, key-value, and graph databases. Teaches how to model multi-dimensional data. Implements distributed databases. Considers multi-tier 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 programming in R, Python, and C++.

Pre-Requisites: DS3200; DS4100

NU Core Designations: AD

After completing this course, students will be able to:

- classify data storage approaches based on data object type, data retrieval and analysis requirements
- select an appropriate data storage structure depending on object type and analysis goals
- plan an information repository for data analysis, data visualization, and discovery
- implement large-scale non-relational data repositories
- distinguish between storage needs for statistical and non-statistical analysis
- outline tiered information architectures for efficient data retrieval and search
- store social graphs, documents, geographical, non-textual, and time series data
- evaluate distributed data storage and retrieval approaches
- describe hardware and network requirements for large-scale data storage

Achievement of learning outcomes will be assessed through:

- completion of programming assignments in R, Python, and C++
- development of a substantial term project requiring the design and implementation of a storage architecture
- mid-term and final exams

DS4400: Machine Learning and Data Mining 1

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, multi-class data and algorithms, boosting, and decision trees. Teaches 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 problems domains, including recommendation systems, fraud detection, or advertising.

Pre-Requisites: DS4300; Statistics

NU Core Designations: AD

After completing this course, students will be able to:

- gain background in the current state of data mining and machine learning
- set up and execute basic machine learning algorithms
- mine data for patterns and insights
- choose an appropriate data analysis approach
- test learning models on new data sets
- interpret failures and correct training algorithm configurations
- program machine learning algorithms in R and Python

Achievement of learning outcomes will be assessed through:

- completion of programming assignments in R and Python
- development of a substantial term project requiring collection, integration, and analysis through machine learning algorithms of a significant data set
- quizzes, mid-term, and final exams
- peer reviews and code walks

DS4420: Machine Learning and Data Mining 2

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. Teaches 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.

Pre-Requisites: DS4400

NU Core Designations: AD

After completing this course, students will be able to:

- set up and execute advanced machine learning algorithms on medium and large data sets
- assess, evaluate, and present the results of a machine learning algorithm
- see the geometric models from linear algebra that underpin machine learning
- discover patterns and insights in data
- conduct experiments and evaluate learning performance
- program machine learning algorithms in R and Python
- read research papers in machine learning and data mining

Achievement of learning outcomes will be assessed through:

- completion of programming assignments in R and Python
- development of a substantial term project requiring collection, integration, and analysis through machine learning algorithms of a significant data set
- reviews of research papers
- quizzes, mid-term, and final exams
- peer reviews and code walks

DS4900: Data Science Senior Project

Helps students develop a sophisticated understanding of data collection, integration, storage, statistical analysis, visualization, and machine-supported analysis and modeling. Students are required to analyze a substantial data set using statistical and visual methods, and 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++.

Pre-Requisites: DS4200; DS4420

NU Core Designations: AD, WI, CE