

# DATA SCIENCE (DSC)

## **DSC 323 | DATA ANALYSIS AND REGRESSION | 4 quarter hours (Undergraduate)**

Topics include multiple regression and correlation methods, model building and validation processes, analysis of variance, logistic regression and regularized regression techniques.

**IT 223 or MAT 351 or MAT 137 is a prerequisite for this course.**

## **DSC 324 | ADVANCED DATA ANALYSIS | 4 quarter hours (Undergraduate)**

The course will teach advanced statistical techniques to discover information from large sets of data. The course topics include visualization techniques to summarize and display high dimensional data, dimensional reduction techniques such as principal component analysis and factor analysis, clustering techniques for discovering patterns from large datasets, and classification techniques for decision making. The methods will be implemented using standard computer packages.

**CSC 324 or DSC 323 or consent of instructor is a prerequisite for this class.**

## **DSC 333 | INTRODUCTION TO BIG DATA PROCESSING | 4 quarter hours (Undergraduate)**

This course will explore different approaches and a framework for performing data analytics on a dynamic, heterogeneous cluster of computing nodes. The course will begin with studying principles behind MapReduce and implementation of custom distributed queries using Hadoop. It will then expand to cover higher-level languages and tools within Hadoop ecosystem (e.g., Pig, Hive) and cluster configuration techniques. Finally, the course will delve into a comparative evaluation of several NoSQL and NewSQL databases that make fundamentally different assumptions for data processing (e.g., OLAP vs OLTP, disk-bound vs in-memory or real-time streaming data). The primary focus of the course will be hands-on implementation and tuning performance for large-scale clusters and data sets.

**CSC 355 is a prerequisite for this class.**

## **DSC 341 | FOUNDATIONS OF DATA SCIENCE | 4 quarter hours (Undergraduate)**

The course is an introduction to the Data Mining (DM) stages and its methodologies. The course provides students with an overview of the relationship between data warehousing and DM, and also covers the differences between database query tools and DM. Possible DM methodologies to be covered in the course include: multiple linear regression, clustering, k-nearest neighbor, decision trees, and multidimensional scaling. These methodologies will be augmented with real world examples from different domains such as marketing, e-commerce, and information systems. If time permits, additional topics may include privacy and security issues in data mining. The emphasis of this course is on methodologies and applications, not on their mathematical foundations.

**IT 223 (or MAT 137 or MAT 242 or MAT 341 or MAT 353) is a prerequisite for this class.**

## **DSC 345 | MACHINE LEARNING | 4 quarter hours (Undergraduate)**

This course introduces students to machine learning techniques and builds upon the background and skills learned in the previous data science and statistics courses. The course topics include advanced methods and algorithms for supervised and unsupervised learning, and ensemble methods. Through research paper discussion and hands-on assignments, the course will also cover recent applications of machine learning, such as autonomous navigation, biomedical informatics, biometrics, and text and web mining.

**(CSC 367 or DSC 341) and (CSC 334 or DSC 324) are prerequisites for this class.**

## **DSC 365 | DATA VISUALIZATION | 4 quarter hours (Undergraduate)**

This course will be an introduction to data visualization techniques for exploration and analysis of data sets from a wide range of fields including commercial, financial, medical, scientific and engineering applications. Topics will include visual encoding of numeric data, effective visualization design, graphical integrity, visualizing distributions and correlation, false-color techniques for feature extraction and enhancement, basic network graph visualization, geospatial visualization and some additional topics.

**(IT 223 or MAT 137) and CSC 241 and CSC 241 are prerequisites for this class.**

## **DSC 390 | TOPICS IN DATA SCIENCE | 4 quarter hours (Undergraduate)**

Specific topics will be selected by the instructor and may vary each quarter. This course is repeatable.

## **DSC 394 | DATA SCIENCE PROJECT | 4 quarter hours (Undergraduate)**

This course provides students with the opportunity to apply and integrate the knowledge they have acquired during the degree program. Students may work in teams and will work on real world data analytics projects using their skills and knowledge. At the end of the course, they submit a complete report summarizing analyses and study outcomes, and present results to the class.

**(CSC 367 or DSC 341) and CSC 301 are prerequisites for this class.**

## **DSC 423 | DATA ANALYSIS AND REGRESSION | 4 quarter hours (Graduate)**

Multiple regression and correlation, residual analysis, analysis of variance, and robustness. These topics will be studied from a data analytic perspective, supported by an investigation of available statistical software.

**IT 403 is a prerequisite for this class.**

## **DSC 424 | ADVANCED DATA ANALYSIS | 4 quarter hours (Graduate)**

The course will teach advanced statistical techniques to discover information from large sets of data. The course topics include visualization techniques to summarize and display high dimensional data, dimensional reduction techniques such as principal component analysis and factor analysis, clustering techniques for discovering patterns from large datasets, and classification techniques for decision making. The methods will be implemented using standard computer packages.

**CSC 423 or DSC 423 or consent of instructor is a prerequisite for this class.**

**DSC 425 | TIME SERIES ANALYSIS AND FORECASTING | 4 quarter hours (Graduate)**

The course introduces students to statistical models for time series analysis and forecasting. The course topics include: autocorrelated data analysis, Box-Jenkins models (autoregressive, moving average, and autoregressive moving average models), analysis of seasonality, volatility models (GARCH-type, GARCH-M type, etc.), forecasting evaluation and diagnostics checking. The course will emphasize applications to financial data, volatility modeling and risk management. Real examples will be used throughout the course.

**CSC 423 or DSC 423 or MAT 456 or consent is a prerequisite for this class.**

**DSC 430 | PYTHON PROGRAMMING | 4 quarter hours (Graduate)**

This course builds the skills necessary to use Python to develop larger programs and libraries. Students will learn to design, implement and debug Python functions and programs, including stochastic and object-oriented techniques. The course will cover Python data structures, and Python facilities for working with files, strings, regular expressions, databases and URLs. The course will also include an introduction to the Pandas package for data management, the NumPy package for scientific computing, and the Matplotlib package for visualization.

**CSC 401 or IS 411 is a prerequisite for this class.**

**DSC 433 | SCRIPTING FOR DATA ANALYSIS | 4 quarter hours (Graduate)**

Data access and transformation with modern statistical software such as SAS and R. Report writing, data graphing and visualization, writing macros and functions to automate tasks and statistical analyses.

**IT 403 and (CSC 401 or IT 411) are prerequisites for this class.**

**DSC 441 | FUNDAMENTALS OF DATA SCIENCE | 4 quarter hours (Graduate)**

An introduction to the Knowledge Discovery Technologies covering all stages of a data mining process: domain understanding, data collection and selection, data cleaning and transformation, dimensionality reduction, pattern discovery, evaluation, and knowledge extraction. The course provides a comprehensive overview of data mining techniques used to realize these stages, including traditional statistical analysis and machine learning techniques. Students will analyze large datasets and develop modeling solutions to support decision making in various domains such as healthcare, finance, security, marketing, customer relationship management (CRM), and multimedia.

**IT 403 or DSC 423 or ECO 520 is a prerequisite for this class.**

**DSC 450 | DATABASE PROCESSING FOR LARGE-SCALE ANALYTICS | 4 quarter hours (Graduate)**

The course covers core concepts of database systems with focus on applications in large-scale analytics. Topics include relational databases, scheme normalization, SQL queries for data integration and data cleaning, database programming for ETL, and nontraditional database systems for unstructured data.

**DSC 430 is a prerequisite for this class.**

**DSC 465 | DATA VISUALIZATION | 4 quarter hours (Graduate)**

An introduction to data visualization techniques to enhance the exploration and analysis of large data sets from a wide range of fields including commercial, financial, medical, scientific and engineering applications. Topics include visual encoding of numeric data, graphical integrity and effective visualization design, visualizing distributions and correlation, false-color techniques for feature extraction and enhancement, basic network visualization and graph layout, isosurface generation, geospatial visualization and volumetric rendering techniques. The course explores both existing visualization software packages and code interfaces for data visualization.

**(IT 403 or MAT 453) and (CSC 401 or IT 411 or IS 411 or MAT 449) are prerequisites for this class.**

**DSC 478 | PROGRAMMING MACHINE LEARNING APPLICATIONS | 4 quarter hours (Graduate)**

The course will focus on the implementations of various data mining and machine learning techniques using a high-level programming language. Students will have hands on experience developing both supervised and unsupervised machine learning algorithms and will learn how to employ these techniques in the context of popular applications including automatic personalization, recommender systems, searching and ranking, text mining, group and community discovery, and social media analytics.

**(DSC 441 and DSC 430) or CSC 480 are prerequisites for this class.**

**DSC 480 | SOCIAL NETWORK ANALYSIS | 4 quarter hours (Graduate)**

This course is an introduction to the concepts and methods of social network analysis. Students will learn to extract and manage data about network structure and dynamics, and to analyze, model and visualize such data. Students will use software tools to model and visualize network structure and dynamics. Specific network applications to be discussed include online social networks, collaboration networks, and communication networks.

**(DSC 423 or SOC 412 or PSY 411) is a prerequisite for this class.**

**DSC 484 | WEB DATA MINING | 4 quarter hours (Graduate)**

An in-depth study of the knowledge discovery process and its applications in Web mining, Web analytics and business intelligence. The course provides coverage of various aspects of data collection and preprocessing, as well as basic data mining techniques for segmentation, classification, predictive modeling, association analysis, and sequential pattern discovery. The primary focus of the course is the application of these techniques to Web analytics, user behavior modeling, e-metrics for business intelligence, Web personalization and recommender systems. Also addressed are privacy and ethical issues related to Web data mining. Students can choose from three types of final course projects: implementation projects, research papers, or data analysis projects. Throughout the course, the students will learn and use a variety of data mining tools to analyze sample data sets as part of class assignments.

**IT 403 and (CSC 451 or CSC 453 or CSC 455 or DSC 450) are prerequisites for this class.**

**DSC 510 | HEALTH DATA SCIENCE | 4 quarter hours****(Graduate)**

The course will focus on data science methods used in clinical studies and public health applications. Students will be introduced to a variety of health care data from electronic health records to payer data, geospatial and unstructured data, and will learn how to solve data science problems in the health sector. Topics include overview of healthcare analytics and typical research questions, epidemiology, data ethics, governance and security, applications of modeling techniques and machine learning methods to a variety of case studies in health care.

**DSC 441 is a prerequisite for this class.**

**DSC 540 | ADVANCED MACHINE LEARNING | 4 quarter hours****(Graduate)**

The course is for students with prior background in data mining or machine learning techniques, and covers more advanced modeling techniques, including ensemble learning, extended linear models such as support vector machines, probabilistic graphical models, mixture and latent variable models, matrix factorization and link analysis. Application of the models will be presented in popular domains such as Web and social media analytics, text mining, crime analysis, community discovery, and health informatics.

**CSC 412 and (DSC 441 or CSC 480) are prerequisites for this class.**

**DSC 590 | TOPICS IN DATA SCIENCE | 4 quarter hours****(Graduate)**

Specific topics will be selected by the instructor and may vary each quarter. This course is repeatable.

**DSC 672 | DATA SCIENCE CAPSTONE | 4 quarter hours****(Graduate)**

The capstone course provides an opportunity for students to integrate and apply the analytics skills and knowledge learned in the classroom to real world data. Students work in teams on a large scale analytics project. At the end of the course, students submit a report summarizing their analyses and study outcomes, and present results to the class.

**PREREQUISITE(S):** Instructor consent required.