

CAP 5768: Introduction to Data Science

Catalog Description

Foundations of databases, analytics, visualization and management of data. Practical data analysis with applications. Introduction to Python, SQL, R, and other specialized data analysis toolkits. **3 Credits**

Prerequisites

Undergraduate course in statistics such as STA 3164 or equivalent

Type

This is a core course for MS – Data Science

Objectives

The goal of this course is to provide students with an overview of the entire data analysis process, while providing them with basic tools (programming languages and toolkits) to navigate through the program. Students will also be exposed to many different applications of the data science approach.

Topics

1. Introduction
 - What is Data Science? Examples, History, Application areas
2. Data Science Environments and Tools
 - Anaconda, Python, R Studio, R, SQL, Hadoop, MapReduce
3. Data Analytics
 - Statistical foundations, exploratory data analysis
 - Data summarizations
 - Visualizations and visual data analytics
 - PCA, Dimensionality Reduction, Clustering and Data Mining
 - Outliers, Normality, Hypothesis Testing
 - Time Series Analysis
 - Network Analytics
 - Text Analytics
 - Machine Learning
 - Causal Inferencing with Data
 - Big Data Analytics
 - PageRank
 - Similarity, MinHash & Bloom Filters
 - Stream Analytics
4. Data management
 - Provenance, security, privacy, data cleaning, and data curation
5. Miscellaneous Topics
 - Ethics, Privacy, Security, Fairness
 - Case studies from different application areas

Textbook(s)

- *Mining of Massive Datasets*, by Leskovec, Rajaraman, and Ullman