Data Manipulation and Analysis -- Elective

Course Description

This course introduces students to the emerging field of Data Science. Instructional units cover the standard practices for effective data manipulation, analysis and interpretation as well as necessary concepts in the three disciplines involved (mathematics, statistics and computing.) Numerous examples of typical problems encountered in Big Data are provided. The emphasis on this course is in the application of the concepts rather than the theory. In the second semester, students will work in teams on large projects in which they will use programming to analyze large datasets and create predictive models. The students will summarize their findings for each project in a written report and will also present them orally.

Course Objectives

- Create models driven by large amounts of data
- Interpret graphs and tables, and make sense of data by visualizing it
- Create interactive presentations for visualizing data
- Automate routine data manipulation tasks, such as formatting, collection and storage
- Describe basic algorithms for clustering, regression and classification of data.
- Explain typical trade-offs that data scientists must make when analyzing data.
- Demonstrate effective communication skills, through team working, oral presentations and good written communication.

Assessing Performance

In the first semester, formative assessment includes worksheets, several practice activities for each lesson, and unit quizzes, and summative assessment includes a small programming project at the end of each four-week unit. In the second semester, projects are assessed on accomplishment, originality, code sophistication, presentation and team work. Milestones are set for each project, so that teams receive feedback on progress before the final evaluation of each project.

Course Essentials

Equipment	Cost/Unit
Classroom set of computers	$0 if you already have some, $500-600 per computer if you need to purchase

First Semester

Unit 1: Overview of Data Science	Small data and big data. Sources of data. The curse of dimensionality. Mean and variance. Overfitting and underfitting. The Variance-Bias trade-off.
Unit 2: Review of Python and Scipy	Syntax. Semantics. Conditions. Loops. Lists. The Scipy suite of packages.
Unit 3: Data visualization	Static graphs and charts. Interactive graphs and live feeds. Visualization for exploration vs. visualization for presentation.
Unit 4: Data collection and organization	Tables, matrices and databases. Text files and the CSV format. Web scraping. JSON and XML Web APIs. Cleaning and summarizing data.

Second Semester

Unit 5: Exploratory data analysis	Projects will use either Descriptive statistics or Inferential statistics (Hypothesis testing) to explore large datasets of interest to the students.
Unit 6: Regression	Students will analyze datasets to create linear models that describe relationships between features of interest.
Unit 7: Clustering	Students will use clustering techniques, such as k-means, to divide the data into meaningful subsets. Students will explore the effect of changing the number of clusters and will try to find an optimal subdivision.
Unit 8: Classification	Students will run training sets of data using decision trees, naive Bayes or, in some cases, neural networks that learn a predictive model fitting the data.