INTRODUCTION TO DATA SCIENCE

STAT 4410/8416

Course Description:

Topics covered in this course include Data Technology, Methods of gathering and cleaning structured or unstructured data, Exploratory data analysis & Dynamic and interactive data visualization, Modeling data for prediction, forecasting or classification. **3 credits**

Prerequisites:

Undergraduate and Graduate: MATH 4750 or STAT 3800 or permission of instructor. Students planning to enroll in this course should be comfortable with computer programming and have knowledge of data structures and preliminary statistical methods.

Overview of content and purpose of the course:

This course emphasizes the developments of the solutions related to the big data problem. It introduces the data technology and how that helps to handle data problems especially the big data problems.

This course covers exploratory methods associated with the structured or unstructured data and provides basic training on how to handle them. It prepares students for working with massive amounts of data. It provides hands on experience with real data provided by the industries that are dealing with data problems

Anticipated audience/demand:

Undergraduate or Graduate students in Statistics, Mathematics, Engineering, Computer Science or Business needing appropriate training on data technology, exploratory data analysis, dynamic and interactive data visualization and statistical modeling of data..

Major topics:

1) Introduction
- a. What is data science?
- b. History and development of the terms

2) Data technology
- a. Data Management
  - i. Relational Database
  - ii. Unstructured Data
  - iii. Big Data
- b. Hadoop, MapReduce, Hive, Google BigQuery, NoSQL, etc.
- c. Accessing Data from Different Sources
  - i. Text Data, Sound Data, Video Data
  - ii. Web Data (Twitter)
  - iii. Other Unstructured Data

3) Exploratory Data Analysis & Visualization

a. Statistical Summary of Data
b. Reshaping Data
c. Handling Missing Information and Data Cleaning
d. Visualizing Data
e. Interactive Visualization
f. Simple Statistical Modeling

4) Advanced Statistical Data Analysis

a. Linear, Non-Linear, and Mixed Models
b. Clustering Analysis
c. Supervised and Unsupervised Learning
d. Crowd Computing / Using Human Intelligence for Large Scale Parallel Computing

5) Communication

a. Collaborative Project, Work with Real Data
b. Project Presentation

Methods:

The course will be presented in a lecture-discussion format.

Student role:

Students are expected to contribute to class discussions and complete course materials as assigned.

Textbook: Paul Murrell. *Introduction to Data Technologies*. Boca Raton: Chapman & Hall, 2009.

September 2014