CAP 2752: Fundamentals of Data Science

Taught by:
Fahad Saeed
CASE Room xxx
fsaeed@fiu.edu / (305) 348-3131

Catalog Description:
This course will teach data science fundamentals to undergraduate non-CS majors. The focus will be on real-world applications and use of associated analysis, visualization tools, Python programming.

Summary:
This course will teach non-programmers to think in computing terms about modern topics, and to approach real-world applications through data science. The course will enable students to:

- Acquire computational thinking skills that will enable students to represent and reason about complex problems in the digital arena
- Understand different kinds of data in terms of their possibilities and limitations to approach complex problems cast in terms of the emerging field of data science
- Become data science scholars through best practices in data documentation and dissemination

The course is intended for students in disciplines outside of computer science and no prior programming experience is assumed. The course topics will be particularly relevant to students interested in physical sciences and social sciences.

Pre-requisites
None for B.S. or B.A. standing or permission of the instructor.

Textbook:
TBD

Times & Locations:
Lecture will be held twice a week for 1 hour, 15 minutes;
Office hours will be held once a week.

Grading

Homework	20%
Midterm Exam	30%
Presentations	10%
Final Exam	40%

Reading
Reading assignments will be distributed via the website. Readings will be associated with each lecture, and these should be completed in order to completely understand the material.

Homework
There will be at least 4 homework assignments. Homeworks will be available on the website by the beginning of the first class of the week. Homeworks must be submitted electronically as a single pdf file via the website. Late homeworks will be penalized 50%, and not accepted after the start of the first class of the week.

Exams and Evaluation
a) There will be presentation on advanced topics by each individual student which will be evaluated by the instructors
b) There will be two exams: a midterm and a final.

Schedule of Topics

Lecture #	Topic
1	**Computational thinking and data science** •What is computational thinking •Computational thinking for reasoning and analysis •What is data science •Data scientists •The context of data science
2	**Data** •What is data •What is not (yet) data •Time series data •Networked data •Geospatial data •Text data •Labeled and annotated data •Big data
3	**Data Representation (structured)** • Interrelations • Spreadsheets • Databases • Matrices • Graphs • Other perspectives
3	**Data analysis software** •Programs for data analysis •Inputs and Outputs •Program Parameters

	•Programming Languages •Programs as Black Boxes •Algorithms versus software •Data Structures and why they are important for data sciences
4	**Multi-step data analysis as workflows** •Building workflows by composing software •Pre---processing and post---processing data •Workflows for data analysis •Workflow inputs and parameters •Executing workflows •Exploring data through workflows •Workflows in practice
5	**Workflow Jupyter notebook practicum** • The Jupyter notebook workflow system • Jupyter notebook in practice
6	**Data pre-processing** •Data cleaning •Quality control •Data integration •Feature selection •Feature construction
7	**Data lifecycle** •Data collection •Data storage •Data extraction and querying •Data integration •Data presentation
8	**Data visualization** •Quality of visualizations •Major types of visualizations •Time series visualizations •Geospatial visualizations •Multi---dimensional spaces •Network visualizations
9, 10	**Data analysis tasks (I)** • Data analysis tasks in data mining, statistics, and machine learning • Supervised learning o Classification tasks o Classification algorithms o Evaluation of classifiers
11,12	**Data analysis tasks (II)**

	• Unsupervised learning ○ Clustering • Pattern detection • Anomaly detection
13	**Data analysis tasks (III)** • Causality • Probabilistic graphical models • Bayesian networks • Causal models
14,15	**Parallel and Distributed Computing for Big data** • Cost of Computation • Divide and Conquer • Speedup with Parallel Processing • Limits of speedups: Critical Path • Amdahl's law • When problems are not parallelizable • Introduction to Parallel Graph Algorithms
16	**Semantic metadata** • What is metadata • Basic metadata versus semantic metadata • Metadata about data collection • Metadata about data processing • Metadata for search and retrieval • Metadata standards • Domain metadata and ontologies
17	**Ontologies (I)** • What is an ontology • Taxonomies and class inheritance • Properties • Logical constraints
18	**Data formats and standards** • Data formats • Data standards • Data repositories • Data services • The Semantic Web and linked open data
19	**Data stewardship** •Data sharing •Data identifiers •Licenses for data •Data citation and attribution •Software and other work products
20	**Advanced Topics (I)**

	Privacy and Ethics in Data Science
21	**Advanced Topics (I)** Introduction to Databases
22	**Advanced Topics (I)** Multidisciplinary Collaborations between data scientists and domain specialists (best practices)
23,24,25	**Presentation by students on advanced topics**