Introduction to Computational Thinking and Data Science

USC Viterbi School of Engineering

http://www.datascience4all.org

Term: Fall 2016
Time: Tues-Thur 10am-11:50am
Location: Allan Hancock Foundation Building (AHF) 145D

Instructor: Dr. Yolanda Gil
Office: GER 207
Office Hours: Thursdays 9am-10am
Contact: gil@isi.edu

Instructor: Dr. Atefeh Farzindar
Office: GER 207
Office Hours: Thursdays 11:50 am-12:50 pm
Contact: farzinda@usc.edu

USC course number: INF 549
Units: 4

Catalogue Course Description

Introduction to data analysis techniques and associated computing concepts for non-programmers. Topics include foundations for data analysis, visualization, parallel processing, metadata, provenance, and data stewardship.

Expanded Course Description

This course will teach non-programmers to think in computing terms about modern topics, and to approach real-world phenomena 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, so no prior experience with computer science is assumed. The course topics will be particularly relevant to students interested in physical sciences and social sciences.

This class will include eight homework assignments and a final exam.

Learning Objectives

This course teaches non-programmers to think in computing terms about modern topics, and to approach real-world phenomena through data science. The course introduces different kinds of data and

corresponding approaches to data analysis, including geospatial data, time series, networks, and multimedia data. Students learn to run multi-step analysis through a graphical workflow interface, and will experience first hand complex concepts in data science such as parallel computing, provenance, and visualization. Students also learn to use ontologies and logic representations to capture metadata and other knowledge about complex data. The course includes practical lessons to use workflow and ontology development toolkits, as well as best practices for data stewardship and dissemination.

Prerequisite(s): none
Co-Requisite (s): none
Recommended Preparation: Mathematics and logic undergraduate courses.

Description and Assessment of Homework Assignments

There will be a homework assignment every 3 or 4 lectures. The assignments must be submitted individually and students will receive individual scores. Students may work in groups to complete the tasks. The homework assignments are expected to take 6-8 hours. Each assignment is graded on a scale of 0-100 and the grading criteria will be specified in each assignment. The homework topics are listed in the Course Schedule. The homeworks include a class project that will be developed by the students independently in 3 separate stages, getting feedback from the instructors at each stage.

Syllabus and Class Schedule

	Date	Topic	Material Covered	Homeworks
Section I: Introduction to Computational Thinking and Data Science				
1	**8/23**	**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	**8/25**	**Data**	• What is data • What is not (yet) data • Time series data • Networked data • Geospatial data • Text data • Labeled and annotated data • Big data	**Homework HW1: Project part 1 (due 9/6)**
3	**8/30**	**Data analysis software**	• Programs for data analysis • Inputs and Outputs • Program Parameters • Programming Languages • Programs as Black Boxes • Algorithms versus software	

4	**9/1**	**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	**9/6**	**Workflow practicum**	• The WINGS workflow system • Workflows in practice	**Homework HW2: Exploring data analysis workflows (due 9/15)**
6	**9/8**	**Provenance**	• What is provenance • Provenance concerning objects • Provenance concerning people and institutions • Provenance concerning processes • Provenance models • Provenance standards	
Section II: Data Analysis				
7	**9/13**	**Data pre-processing**	• Data cleaning • Quality control • Data integration • Feature selection • Feature construction	

8	**9/15**	**Data analysis tasks (I)**	• Data analysis tasks in data mining, statistics, and machine learning
			• Supervised learning
			  ○ Classification tasks
			  ○ Classification algorithms
			  ○ Evaluation of classifiers
				Homework HW3: Analyzing data with workflows (due 9/29)

9	**9/20**	**Data analysis tasks (II)**	• Unsupervised learning
			  ○ Clustering
			  ○ Pattern detection
			  ○ Anomaly detection
			• Simulation and prediction	

10	**9/22**	**Data analysis tasks (III)**	• Causality
			  ○ Probabilistic graphical models
			  ○ Bayesian networks
			  ○ Causal models	

11	**9/27**	**Data lifecycle**	• Data collection • Data storage • Data extraction and querying • Data integration • Data presentation	
Section III: Data Analysis in Practice				

12	**9/29**	**Analyzing different kinds of data (I)**	• Analyzing multimedia data
			  ○ Pre-processing images
			  ○ Segmentation
			  ○ Edge detection
			  ○ Object detection
			  ○ Video analysis
			• Analyzing geospatial data
			  ○ Coordinate systems
			• GIS systems
				Homework HW4: Project part 2 (due 10/11)

13	**10/4**	**Analyzing different kinds of data (II)**	• Analyzing time series data
			  ○ Collecting time series data
			  ○ Pre-processing time series data
			  ○ Event detection
			  ○ Granger causality	

14	**10/6**	**Parallel and distributed computing for big data (I)**	• Cost of computation • Divide and conquer • Parallel computing •	
15	**10/11**	**Parallel and distributed computing for big data (II)**	• Multi-core computing • Distributed computing • Cluster computers • Cloud computing • Grid computing • Virtual machines • Web services • Speedup with parallel computing • Dependencies and message passing • Limits of speedup: Critical path • Amdahl's law • Embarrassingly parallel computations • When problems are not parallelizable • Execution failures • Reduction through	

			MapReduce/Hadoop	
16	**10/13**	**Data visualization**	• Quality of visualizations • Major types of visualizations • Time series visualizations • Geospatial visualizations • Multi-dimensional spaces • Network visualizations	**Homework HW5: Data visualization (due 10/25)**

17	**10/18**	**Analyzing different kinds of data (III)**	• Analyzing text data
			  ○ Pre-processing text
			  ○ Document classification
			  ○ Document clustering
			  ○ Topic detection
			• Sentiment analysis	

18	**10/20**	**Analyzing different kinds of data (IV)**	• Analyzing network data
			  ○ Network structure
			  ○ Dynamic networks
			  ○ Scale-free networks
			  ○ Network analysis	

19	**10/25**	**Analyzing different kinds of data (V)**	• Social media Analysis • Geolocation • Reading:NLP for social media, Ch1 and Ch3	
20	**12/27**	**Project management**	• Multidisciplinary collaborations • Project management • Lean Sixsigma • Reading:NLP for social media, Ch4	
Section IV: Metadata				
21	**11/1**	**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	**Homework6: Parallel Processing (due 11/8)**
22	**11/3**	**Ontologies (I)**	• What is an ontology • Taxonomies and class inheritance • Properties • Logical constraints • **Ontologies (II)** ○ Logical reasoning and	

			inference ○ Expressivity and computation ○ The Semantic Web	
23	**11/8**	**Ontologies (III)**	• Practicum: the PROTÉGÉ ontology editor	**Homework HW7: Developing ontologies (due 11/22)**
Section V: Data Dissemination				
24	**11/10**	**Data stewardship**	• Data sharing • Data identifiers • Licenses for data • Data citation and attribution • Software and other work products	
25	**11/15**	**Data formats and standards**	• Data formats • Data standards • Data repositories • Data services • The Semantic Web and linked open data	
26	**11/17**	**Tracking metadata and provenance**	• Combining computation with metadata and provenance • Validating a data analysis method • Tracking provenance during data analysis • Automatically generating metadata for data analysis	
Section VI: Advanced Topics				
27	**11/22**	**Privacy and ethics in data science**	• Privacy ○ Fair Information Practices ○ Managing sensitive data ○ Anonymizing sensitive data, k-anonymity, differential privacy ○ Re-identifying datasets • Reproducibility • Societal value of data	**Homework HW8: Project part 3-final report (due 12/1)**
28	**11/29**	**Databases**	• File systems vs databases • Relational databases ○ Data models ○ SQL ○ Transactions	

			• NoSQL databases	
29	**12/1**	**Review**	• Selected topics	

Assignment Submission Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Blackboard. Homework will be accepted up to one week late as long as the student requested a late submission ahead of the deadline, and in that case the assignment will be graded at 20% less than the possible points for the assignment. After one week, the assignment will not be graded.

Grading Breakdown

Quizzes: There will be weekly quizzes based on the material from the week before. There is no mid-term for this class.
Homework: There will be eight homework assignments throughout the course. The homework topics are listed in the Syllabus and Class Schedule.
Final Exam: There is a final exam at the end of the semester covering all of the material covered in the class.

Grading Schema:

Quizzes	20%
Homework assignments	50%
Class participation	10%
Final:	20%
Total	100%

Grades will range from A through F.