Introduction to Computational Thinking and Data Science

USC Viterbi School of Engineering

Instructor: Dr. Yolanda Gil

Syllabus

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.

Syllabus and Class Schedule

Lecture	Topic	Material Covered	Homework assigned
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	**Homework HW1: Formulating questions about data**
3	**Data analysis software**	• Programs for data analysis • Inputs and Outputs • Program Parameters • Programming Languages • Programs as Black Boxes • Algorithms versus software	
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 practicum**	• The WINGS workflow system • Workflows in practice	**Homework HW2: Exploring data analysis workflows**

6	**Data analysis tasks (I)**	• Data analysis tasks in data mining, statistics, and machine learning
		• Supervised learning
		  ○ Classification tasks
		  ○ Classification algorithms
		  ○ Evaluation of classifiers	

7	**Data analysis tasks (II)**	• Unsupervised learning
		  ○ Clustering
		  ○ Pattern detection
		  ○ Anomaly detection	

		• Simulation and prediction	

8	**Data analysis tasks (III)**	• Causality
		  o Probabilistic graphical models
		  o Bayesian networks
		  o Causal models
			Homework HW3: Analyzing data with workflows

9	**Data pre-processing**	• Data cleaning • Quality control • Data integration • Feature selection • Feature construction	
10	**Data lifecycle**	• Data collection • Data storage • Data extraction and querying • Data integration • Data presentation	
11	**Data visualization**	• Quality of visualizations • Major types of visualizations • Time series visualizations • Geospatial visualizations • Multi-dimensional spaces • Network visualizations	**Homework HW4: Data visualization**

12	**Analyzing different kinds of data (I)**	• Analyzing text data
		  o Pre-processing text
		  o Document classification
		  o Document clustering
		  o Topic detection
		  o Sentiment analysis	

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

14	**Analyzing different kinds of data (III)**	• Analyzing network data
		  o Network structure
		  o Dynamic networks
		  o Scale-free networks
		  o Network analysis
			Homework HW5: Data analysis in scientific articles

15	**Analyzing different kinds of data (IV)**	• Analyzing multimedia data
		  o Pre-processing images
		  o Segmentation
		  o Edge detection
		  o Object detection
		  o Video analysis
		• Analyzing geospatial data
		  o Coordinate systems
		  o GIS systems	

16	**Parallel and**	• Cost of computation	

	distributed computing for big data (I)	• Divide and conquer • Speedup with parallel processing • Limits of speedup: Critical path • Amdahl's law • When problems are not parallelizable	
17	**Parallel and distributed computing for big data (II)**	• Multi-core computing • Distributed computing • Cluster computing • Cloud computing • Grid computing • Virtual machines • Web services • Practical concerns in distributed computing • Parallel programming languages o MapReduce/Hadoop	**Homework HW6: Data analysis with parallel processing**
18	**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	
19	**Ontologies (I)**	• What is an ontology • Taxonomies and class inheritance • Properties • Logical constraints	
20	**Ontologies (II)**	• Logical reasoning and inference • Expressivity and computation • The Semantic Web	
21	**Ontologies (III)**	• Practicum: the PROTÉGÉ ontology editor	**Homework HW7: Developing ontologies**
22	**Provenance**	• What is provenance • Provenance concerning objects • Provenance concerning people and institutions • Provenance concerning processes • Provenance models • Provenance standards	
23	**Data formats and standards**	• Data formats • Data standards • Data repositories • Data services • The Semantic Web and linked open	

		data	
24	**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	**Homework HW8: Describing provenance for data**
25	**Data stewardship**	• Data sharing • Data identifiers • Licenses for data • Data citation and attribution • Software and other work products	
26	**Advanced topics (I)**	• Privacy and ethics in data science	
27	**Advanced topics (II)**	• Introduction to databases	
28	**Advanced topics (III)**	• Crowdsourcing data collection	
29	**Advanced topics (IV)**	• Multidisciplinary collaborations	
30	**Review**	• Review of real-world data science projects	

Description and Assessment of Assignments

There will be a homework assignment every 3-4 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.

Grading Breakdown

Quizzes: There will be weekly quizzes based on the material from the week before. There is no mid-term.
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. The following is the breakdown for grading:

94 - 100 = A	74 - 76 = C
90 – 93 = A-	70 - 73 = C-
87 – 89 = B+	67 - 69 = D+
84 – 86 = B	64 - 66 = D
83 – 83 = B-	60 - 63 = D-
77 – 79 =C+	Below 60 is an F