Introduction to Computational Thinking and Data Science

Yolanda Gil, gil@isi.edu
Information Sciences Institute and Department of Computer Science, University of Southern California

Course Syllabus

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. Students will learn 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.

	Lesson	Topics
Section I: Introduction to Computational Thinking and Data Science		
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 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
Section II: Data Analysis		

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
		  ○ Probabilistic graphical models
		  ○ Bayesian networks
		  ○ Causal models

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
Section III: Data Analysis in Practice		

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

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

14	**Analyzing different kinds of data (III)**	• Analyzing network data
		  ○ Network structure
		  ○ Dynamic networks
		  ○ Scale-free networks
		  ○ Network analysis

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

16	**Parallel and distributed computing for big data (I)**	• Cost of computation • 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 ○ MapReduce/Hadoop
Section IV: Metadata		
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
22	**Provenance**	• What is provenance • Provenance concerning objects • Provenance concerning people and institutions • Provenance concerning processes • Provenance models • Provenance standards
Section V: Data Dissemination		
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
25	**Data stewardship**	• Data sharing • Data identifiers • Licenses for data • Data citation and attribution • Software and other work products

Section VI: Advanced Topics		

26	**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 and data science

27	**Databases**	• File systems vs databases
		• Relational databases
		  ○ Data models
		  ○ SQL
		  ○ Transactions
		• NoSQL databases

28	**Multidisciplinary collaborations**	• Discipline-specific data, cultures, and methodologies • Indentifying common goals • Developing shared terminology • Structuring and focusing discussions • Use cases and example scenarios • Project planning