INF 549: Introduction to Computational Thinking and Data Science

USC Viterbi School of Engineering

Units: 4
Term—Day—Time:
Fall 2015 – TBD

Location: TBD

Instructor: Dr. Yolanda Gil
Office: Outside classroom
Office Hours: Immediately after class
Contact Info: gil@isi.edu, 310-448-8794

Teaching Assistant: TBD
Office: plaza between RTH cafe and EEB
Office Hours: TBD
Contact Info: TBD

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 XXXX homework assignments, a midterm exams, and a final.

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
Concurrent Enrollment: ??
Recommended Preparation: Mathematics and logic undergraduate courses.

Methods of Teaching:

Students will be assigned articles and videos ahead of each class. During class, those assigned materials will be discussed through practical data analysis examples. In the first homework, each student will choose a dataset based on their interests (social media dataset, geospatial dataset, health dataset, etc). Subsequent homeworks will be devoted to exercise on that dataset what has been learned in class. All of the course materials, including the readings, videos, slides, and homework materials will be posted online.

Technological Proficiency and Hardware/Software Required

Students are required for the homework assignments to access Web sites to download, process, and deposit data. Students who have basic programming skills (whether Python, MatLab, or other languages) will be able to use more advanced techniques in their homeworks if they desire.

Required Readings and Supplementary Materials

There is no textbook. Handouts of all required readings will be made freely available to students electronically. All required software is freely available for students to install on their personal computers or to access through a web interface.

The main reference for the class is this book (freely available online), where selected chapters will be covered:

- The Fourth Paradigm: Data-Intensive Scientific Discovery" T. Hey, S. Tansley, and K. Tolle(Eds), Microsoft Research, 2009. Available free online at http://research.microsoft.com/en-us/collaboration/fourthparadigm/.

Representative course readings that will be used in addition (all will be made freely available to students) include:

- "Computational Thinking." J. M. Wing. Communications of the ACM, viewpoint, vol. 49, no.3, March 2006.
- "Ten Simple Rules for the Care and Feeding of Scientific Data." Goodman, A.; Pepe, A.; Blocker, A. W.; Borgman, C. L.; Cranmer, K.; Crosas, M.; Stefano, R. D.; Gil, Y.; Groth, P.; Hedstrom, M.; Hogg, D. W.; Kashyap, V.; Mahabal, A.; Siemiginowska, A.; and Slavkovic, A. PLOS Computational Biology, 10, 2014.
- "Reusing Scientific Data: How Earthquake Engineering Researchers Assess the Reusability of Colleagues' Data." Faniel, I. M., & Jacobsen, T. E. Journal of Computer-Supported Cooperative Work, 19, 355–375, 2010.
- "Intelligent Workflow Systems and Provenance-Aware Software." Y. Gil. In Proceedings of the Seventh International Congress on Environmental Modeling and Software, San Diego, CA, 2014.
- "A Metadata Best Practice for a Scientific Data Repository." Greenberg, J., White, H. C., Carrier, S., & Scherle, R. Journal of Library Metadata, 9, 194–212, 2009.
- "Introduction: An overview of the knowledge commons." Hess, C., & Ostrom, E. In C. Hess & E. Ostrom (Eds.), Understanding knowledge as a commons: from theory to practice (pp. 3–26). Cambridge, Mass.: MIT Press, 2007.

- "OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors & Common Patterns." "Alan L. Rector, Nick Drummond, Matthew Horridge, Jeremy Rogers, Holger Knublauch, Robert Stevens, Hai Wang, Chris Wroe. Proceedings of EKAW 2004, pp 63-81. doi:10.1007/978-3-540-30202-5_5
- "The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web." Hastings, J., Chepelev, L., Willighagen, E., Adams, N., Steinbeck, C., & Dumontier, M. PLoS ONE, 6(10), 2011. e25513. doi:10.1371/journal.pone.0025513
- "Requirements for Provenance on the Web." Groth, P.; Gil, Y.; Cheney, J.; and Miles, S. International Journal of Digital Curation, 7(1), 2012.
- "A Primer for the PROV Provenance Model." Gil, Y.; Miles, S.; Belhajjame, K.; Deus, H.; Garijo, D.; Klyne, G.; Missier, P.; Soiland-Reyes, S.; and Zednik, S. World Wide Web Consortium (W3C) Technical Report, 2013.
- "The Ethics of Data Sharing and Reuse in Biology." Duke, C. S., & Porter, J. H. BioScience, 63(6), 483–489, 2013. doi:10.1525/bio.2013.63.6.10
- "Information-sharing in academia and the industry: A comparative study." Haeussler, C. Research Policy, 40, 105–122, 2011. doi:10.1016/j.respol.2010.08.007
- "Ensuring the Data-Rich Future of the Social Sciences." King, G. Science, 331, 719–721, 2011. doi:10.1126/science.1197872
- "Cool URIs and Dynamic Data." Sanderson, R., & Van de Sompel, H. IEEE Internet Computing, 16(4), 76–79, 2012. doi:10.1109/MIC.2012.78

Description and Assessment of Assignments

There will be a homework assignment every 2 to 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.

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.
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

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 cannot be submitted.

Syllabus and Class Schedule

Lecture	Topic	Material Covered	Homework Assigned After Lecture(s), Due 2 Weeks After
Section I: Introduction to Computational Thinking and Data Science			
1	**Why computational thinking and data science**	• What is computational thinking • Importance of computational thinking for reasoning and analysis • What is data science • Importance of data science for reasoning and analysis	
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**: Find on the Web a dataset of personal/career interest, characterize it, formulate questions that this dataset can and cannot help answer, hypothesize some computational approach to answer those questions with this and/or other data, discuss computational thinking with respect to the dataset.
3	**Data analysis software**	• Programs for data analysis • Inputs and Outputs • Program Parameters • Programming Languages • Programs as Black Boxes • Algorithms versus software	
4,5	**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	**Homework HW2:** Run an existing workflow for data analysis, experiment with different parameter settings, run it with other datasets, discuss the exploration process.

		• Workflows in practice	
Section II: Data Analysis			
6,7,8	**Data analysis tasks**	• Data analysis tasks in data mining, statistics, and machine learning • Classification • Clustering • Pattern detection • Anomaly detection • Simulation and prediction	**Homework HW3:** Run classification and clustering workflows for different kinds of data (e.g., tabular, text), compare the results using different algorithms and parameters, analyze different subsets of the data, discuss the exploration process.
9	**Data pre-processing**	• Data cleaning • Quality control • Data integration • Feature selection • Feature construction	
10	**Data post-processing**	• Filtering • Data exploration • Data presentation	
11	**Data visualization**	• Types of visualizations • Time series visualizations • Geospatial visualizations • Multi-dimensional spaces	**Homework HW4:** Explore different visualization software, create different visualizations of the same data, qualify and compare the utility of alternative visualizations, discuss the role of visualizations in computational thinking and data analysis.
Section III: Data Analysis in Practice			
12,13, 14,15	**Analyzing different kinds of data**	• Analyzing time series data • Analyzing networked data • Analyzing geospatial data • Analyzing text • Analyzing images • Analyzing video	**Homework HW5:** For a set of selected scientific articles, describe and compare how the data was analyzed, discuss alternative approaches and tradeoffs in data analysis.
16, 17	**Parallel and distributed computing for big data**	• Cost of computation • Divide and conquer • Parallel computing • 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	**Homework HW6:** Run workflows with different kinds of parallel algorithms, run workflows in distributed computers, compare the speed of parallel versus serial execution when data size grows, discuss implications of parallel and distributed computing.

		• Amdahl's law • Embarrassingly parallel computations • When problems are not parallelizable • Execution failures • Reduction through MapReduce and 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	
19,20, 21	**Ontologies**	• What is an ontology • Taxonomies and class inheritance • Properties • Logical constraints • Logical reasoning and inference • Expressivity and computation • The Semantic Web • Practicum: the PROTÉGÉ ontology editor	**Homework HW7:** Download an existing ontology, extend it with new classes and properties, design and implement a small ontology for a problem of personal/career interest, discuss the challenges in representing real-world data and knowledge.
22	**Tracking metadata with semantic workflows**	• Semantic workflows: Combining computation with metadata and provenance • Validating a data analysis method as a workflow • Automatically generating metadata for data analysis • Tracking provenance during data analysis • Publishing workflows • Finding workflows	
Section V: Data Dissemination			
23	**Data formats and standards**	• Data formats • Data standards • Data services • The Semantic Web and linked open data	

24	**Provenance**	• What is provenance • Provenance concerning objects • Provenance concerning people and institutions • Provenance concerning processes • Provenance models • Provenance standards	
25	**Data stewardship**	• Data sharing • Data identifiers • Licenses for data • Data citation and attribution	**Homework HW8:** Specify a problem where data provenance would be useful, design a representation for that provenance, describe how this representation would be used, describe how the provenance would be useful for others to reuse the data for their own purposes, discuss the utility of provenance on science, journalism, and society.
Section VI: Advanced Topics			
26,27, 28,29, 30	**Advanced topics**	• Introduction to Web technologies • Introduction to databases • Introduction to information integration • Introduction to natural language processing and grammars	
31	**Final Exam – Finals week (see the University Schedule of Classes)**		

Statement on Academic Conduct and Support Systems

Academic Conduct

Plagiarism – presenting someone else's ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Section 11, *Behavior Violating University Standards* https://scampus.usc.edu/1100-behavior-violating-university-standards-and-appropriate-sanctions. Other forms of academic dishonesty are equally unacceptable. See additional information in *SCampus* and university policies on scientific misconduct, http://policy.usc.edu/scientific-misconduct.

Discrimination, sexual assault, and harassment are not tolerated by the university. You are encouraged to report any incidents to the *Office of Equity and Diversity* http://equity.usc.edu or to the *Department of Public Safety* http://capsnet.usc.edu/department/department-public-safety/online-forms/contact-us. This is important for the safety of the whole USC community. Another member of the university community – such as a friend, classmate, advisor, or faculty member – can help initiate the report, or can initiate the report on behalf of another person. *The Center for Women and Men* http://www.usc.edu/student-affairs/cwm/ provides 24/7

confidential support, and the sexual assault resource center webpage http://sarc.usc.edu describes reporting options and other resources.

Support Systems

A number of USC's schools provide support for students who need help with scholarly writing. Check with your advisor or program staff to find out more. Students whose primary language is not English should check with the *American Language Institute* http://dornsife.usc.edu/ali, which sponsors courses and workshops specifically for international graduate students. *The Office of Disability Services and Programs* http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially declared emergency makes travel to campus infeasible, *USC Emergency Information* *http://emergency.usc.edu* will provide safety and other updates, including ways in which instruction will be continued by means of blackboard, teleconferencing, and other technology.