Introduction to Computational Thinking and Data Science

USC Viterbi School of Engineering

DSCI 549 **Term: Spring 2021**

Syllabus

Term: Spring 2021
Units: 4
Time: Thu, Thu 10:00-11:50 AM
Location: Online

Instructor: Dr. Anna Farzindar
Office Hours: Thu before class, arranged by appointment only via email
Office hours location: Zoom meeting
Contact Info: farzinda@usc.edu

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, a midterm 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.

Software and Supplementary Readings

All required software is freely available for students to install on their personal computers or to access through a web interface.

There is no textbook. Students can find all the supplementary readings online. Supplementary readings include:

- "Computational Thinking." J. M. Wing. Communications of the ACM, viewpoint, vol. 49, no.3, March 2006.
- "Data Science in the News: Advances and Challenges for the Era of Big Data." Kate Musen, Alyssa Deng, Taylor Alarcon, Yolanda Gil. Technical Report ISI-TR-702, Information Sciences Institute, University of Southern California. August 24, 2015.
- "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.
- "Intelligent Workflow Systems and Provenance-Aware Software." Y. Gil. Proceedings of the Seventh International Congress on Environmental Modeling and Software, San Diego, CA, 2014.
- "Data Science for Business", Foster Provost and Tom Fawcett. O'Reilly Media publishers, 2013.
- "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

Description and Assessment of Homework Assignments

There will be 8 homework assignments. The homeworks include a class project that will be developed by the students independently in 3 separate stages, getting feedback from the instructor at each stage. The assignments must be submitted individually and students will receive individual scores. Students may NOT 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. The homework topics are listed in the Course Schedule.

Assignment Submission Policy

Homework assignments are due at 11:59pm on the due date and should be submitted in Desire2Learn (D2L). Homework will be accepted up to 1 week late *as long as the student requested a late submission ahead of time*, 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. Exceptions will only be made with a note from a professional: for illness or family caregiving due to illness, religious observances, USC athletic event.

Syllabus and Class Schedule

Week	Topic	Material Covered	Homework assigned
Section I: Introduction to Computational Thinking and Data Science			
Week 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	**HW1: Project part 1 – Finding data**
	Data	• What is data • What is not (yet) data • Time series data • Networked data • Geospatial data • Text data • Labeled and annotated data • Big data	
Week 2	**Data analysis software**	• Programs for data analysis • Inputs and Outputs • Program Parameters • Programming Languages • Programs as Black Boxes • Algorithms versus software	
	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	**Homework HW2: Exploring data analysis workflows**
Week 3	**Project presentation**	• Project presentation by students	
Section II: Data Analysis			
	Logic and probability for statistics	• Basic probability for statistics • Logic for statistics • Null hypothesis significance testing • Sampling distributions	
Week 4	**Basic statistics**	• Descriptive statistics • Inferential statistics	

		○ T-tests ○ ANOVAs ○ Chi-squared tests ○ Correlation	
Week	**Project presentation**	• Project presentation by students	
Week 5	**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**
	Data analysis tasks (II)	• Unsupervised learning ○ Clustering ○ Pattern detection ○ Anomaly detection • Simulation and prediction	
Week 6	**Data analysis tasks (III)**	• Causality ○ Probabilistic graphical models ○ Bayesian networks ○ Causal models	
	Data analysis tasks (IV)	• Networks ○ Network structure ○ Dynamic networks ○ Scale-free networks	
	Analyzing different kinds of data (I)	• Time series ○ Collecting time series data ○ Pre-processing time series data ○ Event detection ○ Granger causality	
Week 7	**MIDTERM EXAM**	○ Midterm exam	**Homework HW4: Processing Different Types of Data**
Week 8	**Analyzing different kinds of data (II)**	• Analyzing text data ○ Pre-processing text ○ Document classification ○ Document clustering ○ Topic detection ○ Sentiment analysis	

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

	Data visualization	• Quality of visualizations • Major types of visualizations • Time series visualizations • Geospatial visualizations • Multi-dimensional spaces • Network visualizations	
Section IV: User interfaces and user studies			
Week 9	**User experience, user interfaces, user studies**	• UX/UI Design Principles • AB testing • User study design	**Homework HW5: Project part 2 – Design of data analysis approach**
Week 10	**Analysis for experiments**	• Advanced analysis for experiments • Appropriate statistical tests	**Homework HW6: Data analysis and research methods**
	Causal claims from user studies	• Correlational research • Comparing correlational research to experiments • Ensuring internal validity	
Section V: Data analysis at scale			
Week 11	**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	
	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	
Section VI: Metadata			
Week 12	**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	**Homework HW7: Project part 3 – Final report**
	Ontologies (I)	• What is an ontology • Taxonomies and class inheritance • Properties • Logical constraints	
Week 13	**Ontologies (II)**	• Logical reasoning and inference • Expressivity and computation • The Semantic Web • The PROTÉGÉ ontology editor	**Homework HW8: Data Science Scenarios**
	Provenance and standards	• What is provenance • Provenance models • Provenance standards • Data formats and standards • Data repositories and services • Data sharing • Data identifiers • Licenses for data • Data citation and attribution • Software and other work products	
Section VII: Data lifecycle			
Week 14	**Data lifecycle**	• Data collection and storage • Data cleaning • Data extraction and querying • Data preparation • Quality control • Data integration	
	Privacy and ethics	• Privacy • Sensitive data • Anonymization • Research ethics	

Final Exam
TBD

Attendance
Attendance will not be tracked. All aspects of the course – including the midterm and final exam – can be done asynchronously at your own time via Desire2Learn (D2L). While you can watch the recordings of the lecture on D2L, I encourage you to attend the lectures online in real time so you can ask questions; **for exams, I also suggest that you complete the exam during the assigned time so you can ask clarification questions**. If you choose not to complete the exam during class time, you won't be able to ask clarification questions. You do, however, have a 24 hour window after the start of the assigned in-person time in which to complete the exam. The WebEx links, which allow you to attend lectures online in real time, will all be available on the course page itself within the "Access to Online Lecture" module. Once the network control staff populates a specific lecture link inside the module, students will receive the notification. *If WebEx is not supported by DEN staff, we will use zoom instead. Please see D2L course page for any updates.*

Grading Breakdown
Homework: There will be eight homework assignments throughout the course (see description above).
Midterm Exam: A short answer *closed book* midterm exam will cover all of the material up to that point.
Final Exam: A short answer *closed book* final exam will cover <u>all</u> of the material covered in the class.

Grading Schema:	
Homework assignments	50%
Midterm:	25%
Final:	25%
Total	100%

Grades will range from A through F. The following is the breakdown for grading:

94 - 100 = A	74 – 76.99 = C
90 – 93.99 = A-	70 – 73.99 = C-
87 – 89.99 = B+	67 – 69.99 = D+
84 – 86.99 = B	64 – 66.99 = D
80 – 83.99 = B-	60 – 63.99 = D-
77 – 79.99 =C+	59.99 and below = F

Academic Conduct and Support Systems

Honor Code
In response to recommendations made by the Academic Integrity Task Force to the Dean, the USC Viterbi School of Engineering now has an Honor Code. The Code was developed by Viterbi students, and its text is as follows:

> *Engineering enables and empowers our ambitions and is integral to our identities. In the Viterbi community, accountability is reflected in all our endeavors.*
>
> *Engineering+ Integrity.*
> *Engineering+ Responsibility.*
> *Engineering+ Community.*
> *Think good. Do better. Be great.*

These are the pillars we stand upon as we address the challenges of society and enrich lives.

During your time here at Viterbi, please know that academic and personal resources are available to help:

- The student-driven and student-written Honor Code is here: http://viterbi.usc.edu/academics/integrity/.
- An introductory video is posted at https://myviterbi.usc.edu/ under the link "Academic Integrity Introduction" and serves as a reminder of the school's emphasis in maintaining a high level of academic integrity.
- Master's and PhD students can contact the GAPP office in OHE 106 (https://gapp.usc.edu/) for other helpful resources.
- The Viterbi Academic and Resource Center (VARC) (http://viterbi.usc.edu/students/undergrad/varc) has a variety of services available.

Academic Integrity
The Viterbi School takes academic integrity violations seriously. Most of the violations that have been reported in the past fall into four categories: unauthorized collaboration, plagiarism, code sharing, and cheating on an exam. Specifically:

- Unauthorized collaboration - Unauthorized collaboration on a project, homework or other assignment. (section 11.14.B) All homework assignments must be individually developed. Students that collaborate on assignments will be referred to the Academic Integrity Coordinator.
- 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.
- Code sharing - Obtaining for oneself or providing for another person a solution to homework, a project or other assignment, without the knowledge and expressed consent of the instructor. (section 11.14.A)
- Cheating in an exam - this may involve a number of violations, such as looking at class notes during the exam, looking at other student's exam, "texting" with other students during the exam. See the section titled Two Exams for a list of specific violations.

Please note that that these are only the basic violations that we have encountered in the past, and there are many more. Please familiarize yourself with the discussion of plagiarism in SCampus in Section B.11.00, Behavior Violating University Standards and Appropriate Sanctions available at https://scampus.usc.edu/b/11-00-behavior-violating-university-standards-and-appropriate-sanctions/.

All academic integrity violations will be referred to the Academic Integrity Coordinator of the Viterbi School of Engineering. The process for adjudicating these cases is available in SCampus, Part B, Section 13.

Other Misconduct
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 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.

Diversity
The diversity of the participants in this course is a valuable source of ideas, problem solving strategies, and engineering creativity. The instructors encourage and support the efforts of all of our students to contribute freely and enthusiastically. As members of an academic community, it is our shared responsibility to cultivate a climate where all students and individuals are valued and where both they and their ideas are treated with respect, regardless of their differences, visible or invisible.

Students with Disabilities
Any student requesting academic accommodations based on a disability is required to register with Disability Services and Programs (DSP) each semester. A letter of verification for approved accommodations can be obtained from DSP. Please be sure the letter is delivered to me as early in the semester as possible. DSP is located in STU 301 and is open 8:30 a.m. - 5:00 p.m., Monday through Friday. Website and contact information for DSP: http://sait.usc.edu/academicsupport/centerprograms/dsp/home_index.html, (213) 740-0776 (Phone), (213) 740-6948 (TDD only), (213) 740-8216 (FAX), ability@usc.edu.

Emergency Preparedness/Course Continuity in a Crisis
In case of a declared emergency if travel to campus is not feasible, USC executive leadership will announce an electronic way for instructors to teach students in their residence halls or homes using a combination of Blackboard, teleconferencing, and other technologies.