CS 418 INTRODUCTION TO DATA SCIENCE

University of Illinois at Chicago, Spring 2019

Lecture time: MW 4:30-5:45pm

Location: LC D5

Instructor: Prof. Elena Zheleva (../../)

Office hours: Tue 3-5pm, SEO 1140

Contact: ezheleva@uic.edu (mailto:ezheleva@uic.edu)

TA1: Usman Shahid

Office hours: TBD

Contact: hshahi6@uic.edu

TA2: Mao Li

Office hours: TBD

Contact: mli206@uic.edu

*"Our ability to collect, manipulate, analyze, and act on vast amounts of data is having a profound impact on all aspects of society. This transformation has led to the emergence of **data science** as a new discipline. The explosive growth of interest in this area has been driven by research in social, natural, and physical sciences with access to data at an unprecedented scale and variety, by industry assembling huge amounts of operational and behavioral information to create new services and sources of revenue, and by government, social services and non-profits leveraging data for social good. This emerging discipline relies on a novel mix of mathematical and statistical modeling, computational thinking and methods, data representation and management, and domain expertise."*

--Committee on Data Science, Computing Research Association (https://cra.org/data-science/)

COURSE DESCRIPTION

This course provides an in-depth overview of data science from a computer science perspective. Topics include modeling, storage, manipulation, integration, classification, analysis, visualization, information extraction, and big data. The course is programming-intensive and an emphasis will be placed on tying data science concepts to specific real-world applications through hands-on experience.

PREREQUISITES

Working knowledge of probability, data structures and algorithms, and ability to (learn to) program in Python.

COURSE MATERIALS

We will use Piazza (https://piazza.com/uic/spring2019/cs418) for the course schedule, discussions, and materials, and Gradescope (https://gradescope.com/courses/35901) for grading.
Python is the programming language used for homework assignments. We will use Google Cloud for big data computing thanks to a generous grant that Google Cloud provided for this course.

STUDENT DELIVERABLES

Programming-based homework assignments - 30%
Midterm exam - 20%
Bi-weekly quizzes - 15%
Class project - 35%

TEXTBOOKS

No textbook is required. Readings will be assigned, using multiple online sources, including:
[PTDS] *Principles and techniques of data science (https://www.textbook.ds100.org/)*. Lau, Gonzalez, Nolan.
[MMD] *Mining of massive datatasets (http://infolab.stanford.edu/~ullman/mmds/bookL.pdf)*. Leskovec, Rajaraman, Ullman.
[FDV] *Fundamentals of data visualization (https://serialmentor.com/dataviz/)*. Wilke.
[CIT] *Computational and Inferential thinking (https://www.inferentialthinking.com)*. Adhikari, DeNero.
[CIML] *A course in machine learning (http://ciml.info/) [Errata (https://github.com/hal3/ciml/issues)]*. Hal Daume III.

SCHEDULE

Date	Topic	Assigned Reading	Announcements
1/14	Welcome to Data Science		Quiz 0 out
1/16	Data science lifecycle, data design and sampling	Syllabus, PTDS: Chapters 1-2	Quiz 0 due 11:59pm
1/21	MLK Day - no class		
1/23	Hypothesis testing	CIT: Chapter 11	Pre-lab assignment out
1/28	Lab 1: Data processing with pandas	PTDS: Chapter 3	Project Requirements out Quiz 1 out
1/30	Class canceled due to inclement weather		Quiz 1 due 11:59pm, Jan. 30
2/4	Lab 2: Web scraping and data collection	PTDS: Chapter 7	
2/6	Data cleaning and exploratory data analysis	PTDS: Chapters 4-5	HW 1 (the two labs) due 11:59pm, Feb. 10
2/11	Data visualization	PTDS: Chapter 6	Quiz 2 out
2/13	Data visualization	FDV: Chapters 1-2	Quiz 2 due 2/13 11:59pm Project proposal due 2/14 11:59pm HW 2 out
2/18	Models and estimation	PTDS: Chapter 10	Sign up for project check-in slot on 3/5
2/20	Probability and generalization	PTDS: Chapter 12	
2/25	Supervised learning: decision trees	CIT: Chapter 17	Quiz 3 out
2/27	Supervised learning: geometric view	CIML: Chapter 1	Quiz 3 due 11:59pm 2/27 HW 2 due 11:59pm 2/28
3/4	Supervised learning: linear and non-linear models	PTDS: Chapter 13	Project check-in 3/5
	Supervised learning: bias-		Post-check-in

3/6	variance tradeoff, ensembles	PTDS: Chapter 8 CIML: 5.6	proposal slides due 11:59pm 3/8
3/11	Supervised learning: practical issues, regression	PTDS: Chapter 13 CIML: Chapter 5.5, 6.2	Quiz 4 out
3/13	Unsupervised learning	CIML: 3.4	Quiz 4 due 11:59pm HW 3 out
3/18	Midterm review		
3/20	Midterm exam		
3/25	Spring break		
3/27	Spring break		
4/1	Databases and SQL	PTDS: Chapter 9	
4/3	Data science in the real world Invited talk by Dr. Plamen Petrov, VP of AI for Anthem (https://www.anthem.com/)		HW 3 due 11:59pm 4/3
4/8	Databases and SQL	PTDS: Chapter 9	
4/10	Large-scale data processing	MMD: Chapter 2	HW 4 out Project progress report due 11:59pm, April 11
4/15	Recommender systems	MMD: Chapter 9	Quiz 5 out
4/17	A/B testing	PTDS: Chapter 18	Quiz 5 due 11:59pm
4/22	Ethics in data science Guest lecture by Dr. Emanuelle Burton	Main reading: Who's Using Your Face - The Ugly Truth About Facial Recognition (https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e) Optional: Why You Can No Longer Get Lost in the Crowd (https://www.nytimes.com/2019/04/17/opinion/data-privacy.html) Microsoft Denied Police Facial Recognition Tech Over Human Rights Concerns (https://www.theverge.com/2019/4/17/18411757/microsoft-facial-recognition-sales-refused-police-access) NYPD Claws Back Documents On Facial Recognition (https://www.nydailynews.com/new-york/ny-nypd-facial-recognition-disclosures-20190414-ifwgro76cje5tgiyq6wqsty7ou-story.html)	
4/24	Social network analysis	MMD: Chapter 10	HW 4 due 11:59pm 4/25
4/29	Final project presentations		Presentations due at noon Quiz 6 out
5/1	Final project presentations		Quiz 6 due 11:59pm
5/7	Final project due		Final project due 11:59pm