CS 391 E1 - Fall '19 - Foundations of Data Science – Syllabus

Official Course Description

This course is intended as the first to take for students interested in the aspects of computer science related to data analysis and data management. It specifically serves as a preparation including, but not limited, to the courses CS460, CS506, CS542 and CS565. Course topics will cover data collection, cleaning and visualization. Data modeling and basics of databases. Mathematical foundations of data science including linear algebra, (multivariate) calculus and convex optimization. Topics in data mining, such as similarity and distance functions, clustering, ranking, networks. Introduction to machine learning. Prediction methods, e.g. regression and common measures.

Piazza (Q&A, discussion, as well as distribution of lecture notes and homework):
piazza.com/bu/fall2019/cs391e1/home
Gradescope (homework and programming exercise submission):
https://www.gradescope.com/courses/61210(Entry code: M6XN33)

Google calendar with labs and lectures:

Prerequisites

CS 112 and CS 131 (or MA 293).

Instructors and Teaching Fellows

Name	**Office Hours**	**Email@bu.edu** *
Prof. Dora Erdos	Thurs 3-4 pm and Wed 9:30-11am in MCS288	edori
Yida Xin (TF)	TBA in EMA 302 (or 303 if 302 is very busy)	yxin

* Messaging via Piazza is preferable to email (and will get a faster response).

Textbook

Joel Grus: Data Science From Scratch: First Principles with Python, 2nd edition, ISBN: 9781492041122 (github)

Useful additional resources:

- Wes McKinney: Python for Data Analysis
- Leskovec, Rajaraman, Ullman: Mining massive Datasets (link)
- James, Witten, Hastie, Tibshirani: An Introduction to Statistical Learning (available from BU network on this link)

Course Structure and Communication

We will be using Piazza for all discussions outside of class. The system is highly catered to getting you answers to your questions fast and efficiently from classmates, the TFs, and instructors. Please do not email questions to the teaching staff -- post your questions on Piazza instead. We also encourage you to post answers to other students' questions there (but obviously, not answers to problems on the problem sets!). *Your fastest route to get an answer to most questions is via Piazza.* Office hours are not to be used to fill you in on a class you skipped or to re-explain entire topics. Office hours are scheduled at times to provide the most help to students who start the homework before the last minute.

Lectures

Tues/Thurs 12:30 - 13:45 pm, KCB 106

We expect students to come to class, and to come on time. While the class is large, class participation and questions will be encouraged. Also, while our textbook will be very helpful, it is an imperfect substitute for in-class learning, which is the fastest (and easiest) way to learn the material. If you miss a class, please get the notes and work through the material with a fellow student.

Discussion Labs

Lab E2 (Mon 9:05 - 9:55 am) CAS 233
Lab E3 (Mon 10:10 - 11:00 am) CAS 233

Labs will be an invaluable part of the course involving interactive problem-solving sessions, tips on homework questions, and supplemental material not covered in lecture.

Coursework and Grading

The course grade will break down as follows:

40% weekly homework assignments, due Thursdays starting September 12.
25% in-class midterm exam (in-class, planned for **Tuesday, October 22**).
35% comprehensive final, in the normal exam slot for classes in our respective time blocks.
Up to 5% bonus for participation in lecture, lab, and on Piazza.

Last day to drop without a “W”: October 7, 2019. With a “W”: November 8, 2019. Incompletes for this class will not be granted.

Exams: There will be one eighty minute in-class midterm held during the middle of the semester on **Tuesday, October 22, 2019**. The cumulative final will be held during the normal two-hour final exam slot. Please make your travel plans accordingly.

Attendance: We will not take formal attendance in this course. However, while our textbook will be very helpful, it is an imperfect substitute for in-class learning, which is the fastest (and easiest) way to learn the material. Some material covered in lecture and lab may not be in our textbooks. You are in all cases responsible to be up to date on the material. We ask that you please arrive in class on time, since it is disruptive to have students flowing in throughout the class period. While the class is large, class participation and questions are very much encouraged.

Homework problems: Homework problem sets, assigned weekly, allow you to practice (a) solving problems using the ideas from class, often in a new way, (b) communicating your ideas using technical language (precise descriptions, pseudocode, formal claims, proofs), (c) applying techniques to real world data. Most problems require written solutions, but some of the problems will contain small-scale implementations and simulations.

The homework is probably the most useful learning tool in the course—take it seriously, allow yourself time to do it, and have fun! Limited collaboration is permitted; see below.

Homework Submission: Assignments will typically be due Thursdays by 11:59PM, electronically via Gradescope. Solutions should be typeset (preferred) or neatly hand-written and scanned.

Late Policy for homework assignments**:** During the course, you will have **two** chances to electronically submit assignments on Gradescope up to 48 hours late with no penalty, but Saturday 11:59PM is a hard deadline. Any assignment arriving between Thursday 11:59PM and Saturday 11:59PM is considered late. Please do not send emails to the staff about late submissions (not necessary) or requesting additional time.

Regrade Policy: If, after reviewing your solution, you still believe a portion of your homework was graded in error, you may request a regrade, **via Gradescope**, *NOT* through email. One of the staff will consider your request and adjust your grade if appropriate. Note that when we regrade a problem, your score may go up or down.

Workload: The workload in this course will be medium heavy. There is a problem set (almost) every week. As you likely already know, assignments requiring substantial creativity can take more time than you expect, so plan to finish a day early.

Collaboration Policy

Collaboration on homework problems, is permitted and even encouraged! If you choose to collaborate on some problems, you are allowed to discuss each problem with at most 5 other students currently enrolled in the class. Before working with others on a problem, you should think about it yourself for at least 45 minutes.

You must write up each problem solution by yourself (using your own words) without assistance, even if you collaborate with others to solve the problem. You must also identify your collaborators. If you did not work with anyone, you should write ``Collaborators: none.'' It is a violation of this policy to submit a problem solution that you cannot orally explain to an instructor or TF. You may get help on Piazza, from the TFs and instructors for the class for specific problems. (Don't expect them to do it for you, however!)

Finding answers to problems on the Web or from other outside sources (these include anyone not enrolled in the class) is strictly forbidden.

No collaboration whatsoever is permitted on exams. The collaboration policy for programming problems will be specified in the assignments.

Collaboration strategies: If you do collaborate, use it as an opportunity to practice group work skills: give everyone a chance to speak, listen carefully, acknowledge good suggestions. If you have a tendency to be shy, speak up! If you have the tendency to dominate conversations, make sure to give others the floor.

Academic Conduct:

Academic standards and the code of academic conduct are taken very seriously by our University, by the College of Arts and Sciences, and by the Department of Computer Science. Course participants must adhere to the CAS Academic Conduct Code -- please take the time to review this document if you are unfamiliar with its contents.

If in doubt, our department has an extensive compilation of **examples** with regard to Academic Conduct and permissible collaboration.

Violations of this policy will be dealt with according to University regulations.

Course topics:

This list is subject to change. Many topics will not show up by themselves but in combination with others. We will keep an up-to-date schedule on Piazza.

1. Data handling in Python: collecting, cleaning visualizing using Python tools.

2. SQL and data modeling: relational algebra, schemes, indexing basics
3. Similarity and distance functions, clustering
4. Linear algebra, dimensionality reduction: SVD, least squares
5. Probability and statistics: interpreting results
6. Convex optimization: gradient (one and multi dimensional), gradient descent, regressions
7. Machine Learning basics: linear and ridge regression, SVM
8. Graphs: node importance, connectivity, centrality - Pagerank, social and web graphs, community detection