Title: Syllabus-DRAFT Author: Pavlos Protopapas Slug: syllabus Date: 2020-08-16

Introduction to Data Science (Fall 2020)

CS 109a, AC 209a, Stat 121a, or CSCI E-109a

Course Heads

Pavlos Protopapas (SEAS), Kevin Rader (Statistics), & Chris Tanner (SEAS)

Lab Instructor: Eleni Kaxiras (SEAS)

Lectures: Mon, Wed, Fri at 9am-10:15am and 3pm-4:15pm (tentative)

Sections: Fri 1:30-2:45 pm and Mon 8:30-9:45 pm.

Advanced Sections: Wed at 12pm (tentative)

Office Hours: (TBD)

Format: Exclusively online, but we aim to foster enriching interactions and collaboration as much as possible.

Prerequisites: You are expected to have programming experience at the level of CS 50 or above, and statistics knowledge at the level of Stat 100 or above (Stat 110 recommended). HW #0 is designed to test your knowledge on the prerequisites. Successful completion of this assignment will show that this course is suitable for you. HW #0 will not be graded but you are required to submit.

Welcome to CS109a/STAT121a/AC209a, also offered by the DCE as CSCI E-109a, Introduction to Data Science. This course is the first half of a one-year course in data science. The course focuses on the analysis of messy, real-life data to perform predictions using statistical and machine learning methods.

Throughout the semester, our content continuously centers around five key facets:

1. data collection - data wrangling, cleaning, and sampling to get a suitable data set;

2. data management - accessing data quickly and reliably;

3. exploratory data analysis – generating hypotheses and building intuition;

4. prediction or statistical learning; and

5. communication – summarizing results through visualization, stories, and interpretable summaries.

Only one of CS109a, AC209a, or STAT121a can be taken for credit. Students who have previously taken CS109, AC209, or STAT121 cannot take CS109A, AC 209A, or STAT121A for credit.

Course Components

The lectures will be live-streamed and can be accessed through the Zoom section on Canvas. Video recordings of the live stream will be made available within 24 hours after the event, and will be accessible from the Lecture Video section on Canvas.

Lectures

The class meets, virtually, three days a week for lectures (M, W, F). The same lecture will be given twice each day: once in the morning and again in the afternoon, to accommodate students in different time zones. Mondays and Wednesdays will be mostly lecture content with some hands-on lab work, whereas Fridays will be the inverse (mostly hands-on lab work). Attending and participating in lectures is a crucial component of learning the material presented in this course.

Quizzes

At the end of each lecture, there will be a short, graded quiz that will cover the pre-class and in-class material; there will be no AC209a content in the quizzes. The quizzes will be available for 36 hours after posting.

50% of the quizzes will be dropped from your grade.

Sections

Lectures are supplemented by sections led by teaching fellows. There are two types of sections:

Standard Sections :

This will be a mix of review of material and practice problems similar to the HW. The material covered on Friday and Monday is identical.

Advanced Sections

The course will include advanced sections for 209a students and will cover a different topic per week. These are 75-min lectures and will cover advanced topics like the mathematical underpinnings of the methods seen in lecture and hands-on exercises, along with extensions of those methods. The material covered in the advanced sections is **required** for all AC209a students.

Note: Sections are not held every week. Consult the course calendar for exact dates.

Exams

There are no exams in this course.

Projects

Students will work in groups of 2-4 to complete a final group project, due during the Exams period. See Calendar for specific dates.

Homework Assignments

There will be 9 graded homework assignments. Some of them will be due one week after being assigned, and some will be due two weeks after being assigned. You have the option to work and submit in pairs for all the assignments **except** HW4 and HW7, which you will do individually.

You will be working in Jupyter Notebooks, which you can run in your own environment or in the SEAS JupyterHub cloud.

Instructions for Setting up Your Environment

Instructions for Using JupyterHub

On weeks with new assignments, the assignments will be released by Wednesday 3pm.

Standard assignments are graded out of 5 points.

AC209a students will have additional homework content for most assignments worth 1 point.

Instructor Office Hours

Pavlos: (TBD)

Kevin: (TBD)

Chris: (TBD)

Eleni: (TBD)

Participation

Students are expected to be actively engaged with the course. This includes:

1. Attending and participating in lectures (or the follow-up session later in the day)
2. Making use of resources such as office hours, labs, and sections
3. Participating in the Ed discussion forum — both through asking thoughtful questions and by answering the questions of others

Despite being remote, we aim to make this course as interactive, stimulating, and fun as always, and we rely on each of you to contribute your awesome uniqueness.

Recommended Textbook

An Introduction to Statistical Learning by James, Witten, Hastie, Tibshirani.

The book is available here:

Free electronic version: http://www-bcf.usc.edu/~gareth/ISL/ (Links to an external site).

HOLLIS: http://link.springer.com.ezp-prod1.hul.harvard.edu/book/10.1007%2F978-1-4614-7138-7

Amazon: https://www.amazon.com/Introduction-Statistical-Learning-Applications-Statistics/dp/1461471370 (Links to an external site).

Deep Learning, Vol. 1: From Basics to Practice by Andrew Glassner

Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville (MIT Press, 2016)

Course Policies

Getting Help

For questions about homework, course content, package installation, JupyterHub, and after you have tried to troubleshoot yourselves, the process to get help is:

1. Post the question in **Ed** and get a response from your peers. Note that in Ed questions are visible to everyone. The teaching staff monitors the posts.

2. Go to **Office Hours**; this is the best way to get direct help.

3. For private matters send an email to the **Helpline:** cs109a2020@gmail.com. The Helpline is monitored by the teaching staff.

4. For personal and confidential matters send an email to the **instructors**.

Collaboration Policy

We expect you to adhere to the Harvard Honor Code at all times. Failure to adhere to the honor code and our policies may result in serious penalties, up to and including automatic failure in the course and reference to the ad board. If you work with a partner on an assignment make sure both parties solve all the problems. Do not divide and conquer. You are expected to be intellectually honest and give credit where credit is due. In particular:

- if you work with a fellow student but decide to submit individual assignments, include the name of each other in the designated area of the submission.
- if you work with a fellow student and want to submit the same assignment, you need to form a group prior to the submission. Details in the assignment. Remember, not all assignments will permit group submissions.
- you need to write your solutions entirely on your own or with your collaborator (e.g., not entirely from Google search results)
- you are welcome to take ideas from code presented in labs, lecture, or sections, but you need to change it, adapt it to your style, and ultimately write your own. We do not want to see code copied verbatim from the above sources.
- if you use code found on the internet, books, or other sources you need to cite those sources.
- you should not view any written materials or code created by other students for the same assignment;
- you may not provide or make available solutions to individuals who take or may take this course in the future.
- if the assignment allows it you may use third-party libraries and example code, so long as the material is available to all students in the class and you give proper attribution. Do not remove any original copyright notices and headers.

Late or Wrongly Submitted Assignments

There are **no late** days in homework submission. We will accept late submissions only for medical (if accompanied by a doctor's note) or other official University-excused reasons.

To submit after Canvas has closed or to ask for an extension, send an email to the Helpline with subject line "Submit HW1: Reason=the flu" replacing 'HW1' with the name of the current assignment and "the flu" with your reason. You need to attach the note from your medical provider otherwise we will not accept the request. Email the instructors if you have other University-excused reasons.

If you forgot to join a Group with your peer and are asking for the same grade we will accept this with no penalty up to HW3. For homeworks beyond that we feel that you should be familiar with the process of joining groups. After that there will be a penalty of -1 point for both members of the group provided the submission was on time.

Grading Guidelines

Homework will be graded based on:

1. How correct your code is (the Notebook cells should run, we are not troubleshooting code)

2. How you have interpreted the results — we want text not just code. It should be a report.

3. How well you present the results.

The scale is 0 to 5 for each assignment.

Re-grade Requests

Our graders and instructors make every effort in grading accurately and in giving you a lot of feedback.

If you discover that your answer to a homework problem was correct but it was marked as incorrect, send an email to the Helpline with a description of the error. **Please do not submit regrade requests based on what you perceive is overly harsh grading**, The points we take off are based on a grading rubric that is being applied uniformly to all submissions.

If you decide to send a regrade request, send an email to the Helpline with subject line "Regrade HW1: Grader=johnsmith" replacing 'HW1' with the current assignment and 'johnsmith' with the name of the grader within **48 hours of the grade release**.

Zoom Expectations:

Despite being remote and distributed, we want everyone to participate and engage with fellow classmates and staff. To this end, we **request for everyone to leave their video on during class.** We fully understand that some may be uncomfortable or unable to do so, for various personal reasons. We respect your decision and empathize for any situation that may arise. Nonetheless, if possible, we strongly ask for you to please leave your video on.

Many are getting accustomed to Zoom for the first time. While it is understandable to have some glitches, mistakes, and faux pas, some mistakes can be largely disruptive to such a large audience like that of our class. Please familiarize yourself to Zoom before the semester begins. In particular, please:

- keep your video on (see above)
- keep your mic muted by default, until you speak

- be mindful of overtly distracting actions that your video may cause (this isn't a large issue, but it's akin to students' laptops distracting those who sit nearby)
- be mindful of your mic settings after returning to the main room from a break-out room, as it tends to un-mute each person's mic.

Communication from Staff to Students

Class announcements will be through **Ed**. All homework and will be posted and submitted through **Canvas**. Quizzes are completed on **Ed** as well as all feedback forms.

NOTE: make sure you adjust your account settings so you can receive emails from Canvas.

Submitting an assignment

Please consult Homework Policies & Submission Instructions [Update link]

Course Grade

Your final score for the course will be computed using the following weights:

Assignment	Final Grade Weight
Homework 0	1%
Paired Homework (7)	42%
Individual Homework (2)	20%
Quizzes	12%
Project	25%
Total	**100%**

Software

We will be using Jupyter Notebooks, Python 3, and various python modules. You can access the notebook viewer either on your own machine by installing the Anaconda platform (Links to an external site) which includes Jupyter/IPython as well all packages that will be required for the course, or by using the SEAS JupyterHub from Canvas. Details in class.

Auditing the Class

If you would like to audit the class, please send an email to the Helpline indicating who you are and why you want to audit the class. You need a HUID to be included to Canvas. Please note that auditors may not submit assignments for grading or make use of other limited student resources such as office hours.

Academic Integrity

Ethical behavior is an important trait of a Data Scientist, from ethically handling data to attribution of code and work of others. Thus, in CS109A we give a strong emphasis to Academic Honesty. As a student your best guidelines are to be reasonable and fair. We encourage teamwork for problem sets, but you should not split the homework and you should work on all the problems together. For more detailed expectations, please refer to the Collaborations section above.

Accommodations for Students with Disabilities

Students needing academic adjustments or accommodations because of a documented disability must present their Faculty Letter from the Accessible Education Office (AEO) and speak with the professor by the end of the second week of the term, (fill in specific date). Failure to do so may result in the Course Head's inability to respond in a timely manner. All discussions will remain confidential, although Faculty are invited to contact AEO to discuss appropriate implementation.

Diversity and Inclusion Statement

Data Science, like many fields of science, has historically only been represented by a small sliver of the population. This is despite some of the early computer scientist pioneers being women (see Ada Lovelace and Grace Hopper for two examples). Recent initiatives have attempted to overcome some barriers to entry: Made w/ Code.

We want to attempt to discuss diversity in data science from time to time, where appropriate and possible. Please contact us (in person or electronically) or submit anonymous feedback if you have any suggestions to improve the diversity of the course materials. Furthermore, we would like to create a learning environment for our students that supports a diversity of thoughts, perspectives and experiences, and honors your identities (including but not limited to race, gender, class, sexuality, religion, ability, etc.) To help accomplish this:

- If you have a name and/or set of pronouns that differ from those that appear in your official Harvard records, please let us know!
- If you feel like your performance in the class is being impacted by your experiences outside of class, please do not hesitate to come and talk with us. We want to be a resource for you. Remember that you can also submit anonymous feedback (which will lead to me making a general announcement to the class, if necessary, to address your concerns). If you prefer to speak with someone outside of the course, you may find helpful resources at the Harvard Office of Diversity and Inclusion.

- We (like many people) are still learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please talk to us about it. (Again, anonymous feedback is always an option.)
- As a participant in course discussions, you should also strive to honor your classmates' diversity.