COGS 108: Data Science in Practice

Fall 2020
MWF 9-9:50 (remote)

COURSE OVERVIEW

This class is a hands-on practical, technical, and applied data science course intended to get you experience working on data science projects. In COGS 9 (Introduction to Data Science) you (may have) learned why data and data science are important. This class goes beyond appreciation for what can be done to actually *doing it*. Often the best way to learn something is to do it yourself. Often, this process will involve attempting to do something, doing it wrong, learning from your mistakes, and then succeeding. That's part of the data science process. This course is all about the *practice* of data science.

In focusing on the application, there is theory that won't be discussed and mathematical proofs that won't be done. That is by design. In particular:

1. There are entire courses dedicated to each of the topics we'll cover. To have time to do anything, we can't teach all the details in a single course.
2. Experts in each of these domains are out there and excited to teach you the nitty gritty about each topic.
3. My expertise is not machine learning. It's data science, education, human genetics, and the intuition behind data analysis.
4. We're promoting data literacy. We believe that everyone who is data literate is at an advantage as they go out into the modern world. Data literacy is not limited to those who are computational gurus or math prodigies. You do not have to be either of those to excel at this course.

In this course, you will try many methods. Every so often, you'll even be asked to implement a technique that has not been explicitly taught. Again, this is by design. As a data scientist, you'll regularly be asked to step outside of your comfort zone and into something new. Our goal is to get you as comfortable as possible in that space now. We want to provide you with a technical and a data science mindset that will allow you to ask the right questions for the problem at hand and set off alarm bells when something in your dataset or analysis is "off."

A Note About Remote Learning

For some of you, remote learning is new. For others, you've got a bit of practice. For all of us, there is a lot going on in the world. While students have always been under a fair amount of pressure and stress, the struggles students may encounter this quarter (for a whole bunch of different reasons) may go beyond what is typical. I want you all to know that I fully understand this and am here to help you succeed.

While regular deadlines have been established to help keep you all on track, I want you to know up front that I am a very reasonable person. While I ask that you all do your best to meet deadlines that have been set, know that if you're struggling, I absolutely want you to reach out to let me know, to ask for an extension, or to discuss some other accommodation.

Please take care of yourselves and one another, and I'll work as hard as needed to ensure success for all students this quarter.

COURSE STAFF & INFORMATION

Instructor: Shannon Ellis (sellis@ucsd.edu)
Instructor Office Hours: Wednesdays 2-4 PM

Role	Name	Section	Office Hours
TA	Atman Patel	M 4PM; M 6PM	Tue 11:30AM-12:30PM
TA	Ganesh Raghavendran	W 3PM; W 5PM	Thurs 3PM-4PM
TA	Sidharth Suresh	F 3PM; F 4PM	Wed 11AM-12PM
IA	Michael Baluja	M 6PM; W 5PM	–
IA	Fei Dai	–	–
IA	Andrew Nguyen	M 4PM; W 2PM	–
IA	Emily Park	W 2PM; W 3PM	Fri 12-1 PM
IA	Abby Paterson	F 3PM; F 4PM	–

Course GitHub: https://github.com/COGS108
Course Piazza*: http://piazza.com/ucsd/fall2020/cogs108
Course Canvas: https://canvas.ucsd.edu/courses/18744
YouTube Playlist: http://bit.ly/cogs108_youtube_fa20 (videos also on Canvas)
Assignment Submission: https://datahub.ucsd.edu
Course Feedback (anonymous): Google Form

*You will be able to post anonymously on Piazza; however, you will only be anonymous to your classmates. Your Instructor and TAs will be able to see who you are.

COURSE OBJECTIVES

- Formulate a plan for and complete a data science project from start (question) to finish (communication)
- Explain and carry out descriptive, exploratory, inferential, and predictive analyses in Python
- Communicate results concisely and effectively in reports and presentations
- Identify and explain how to approach an unfamiliar data science task

---

COURSE MATERIALS

- There is no textbook
- All materials will be provided on GitHub/datahub
- iclickers will NOT be used this quarter

CLASS TECHNOLOGY

- Python (>= 3.6; Anaconda distribution)
- Jupyter Notebooks
- git and GitHub (option to use SourceTree, GitHub Desktop, or other GUI)

GRADING

	% of Grade	Requirement
Assignments	45%	Complete 6 Assignments
Content Engagement	20%	Complete Video Quizzes
Project Proposal*	8%	Complete Project Proposal
Final Project*	25%	Complete Final Project
Final Project Survey*	2%	Complete Final Project Survey

* This quarter there will be two options for the COGS 108 Final Project. Students will choose between the following: (1) Data Science Project on a topic of your choosing completed throughout the quarter in a group of 4-5 people OR (2) Data Science Project completed individually duringweek 10 + finals week on a dataset and question provided to you (to mimic the data science interview process). If option 2 is chosen, the Project Proposal will also be completed individually.

Final exam date: No final exam, only final project deadline (Wednesday, 12/16 at 11:59 PM). You do not have to show up anywhere on the date/time of the final exam.

Grades

All grades will be released on Canvas. It is *your responsibility to check that your assignment was submitted, that your grade is accurate, and to get in touch if any are missing and/or you think there is a problem.*

Your final letter grade will be determined using the standard grading scale. Grades are not rounded up.

Attendance

Given the unique situation of this quarter, lecture and discussion section attendance will be neither required nor incentivized. We do not want to negatively impact the learning (or grades) of students in different timezones.

LECTURE (CONTENT ENGAGEMENT)

Starting Monday of Week 1, all lectures will be pre-recorded and shared (asynchronous learning). Students are encouraged to watch these during regular lecture time if that works for their schedules, but are free to watch them at a more convenient time, should that be necessary. They are designed so that each days' videos and quizzes can be completed in the normal class meeting time (50 min).

For every pre-recorded video, there will be a corresponding quiz. These are to be completed individually. There will be 1-3 videos per day, resulting in ~65 quizzes in total throughout the quarter. These are designed to help keep students on track. Links to videos will be posted on each week's README on GitHub and on Canvas. Corresponding quizzes will be taken on Canvas.

You will have a single attempt for each quiz. Each quiz will have between 3 and 15 questions. Quizzes will be timed, and you will have a little more than 1 minute per question.

Every student has 10 "free" video quiz points for the quarter. This means that if there are 200 possible quiz points throughout the quarter (there may be a few more or a few less than 200 in reality), and at the end of the quarter you have >=190 total video quiz points, you will receive full credit for the "Content Engagement" portion of your grade. If, in this scenario, you end the quarter with 180 total quiz points, you will receive

a 95.0% (190/200) for "Content Engagement". This allows for flexibility if people miss a quiz one day or perform worse than they wanted to on a particular quiz.

Starting with Mon of Week 1, there will be video quizzes due every Mon, Wed, and Fri throughout the quarter. While quizzes are designed to be taken each MWF, all assigned quizzes for a given week must be completed on Canvas by the Friday of the week on which they're assigned to receive credit (11:59 PM PST).

All videos and quizzes will be released on the Friday of the preceeding week. (For example, any video quizzes due week 2 will be released by Friday of week 1.)

ASSIGNMENTS

Assignments are hands-on in this course. They will be completed individually in Jupyter Notebooks and both released and submitted on datahub.

The practice of data science involves writing code to answer questions and accomplish tasks. Thus, to get practice, your assignments will require you to use Python to do just that. Not everything will be explicitly mapped out step-by-step for you. This is intentional. Figuring things out when it's not entirely clear what to do next is part of the practice here. You'll attempt things that won't work and become comfortable with this. You'll get stuck and work to get unstuck. Not quite knowing exactly what's going on at all times is part of the process. And, to be honest, part of the job of being a data scientist.

That said, the first two assignments will be the simplest assignments and aim to get you up to speed in Python and familiar with `pandas`. If the first two assignments are particularly difficult for you, that's ok. But, it's then up to you to determine if you want to put in the work to make it through the rest of the quarter. **Assignments will take more time and be more difficult starting with the third assignment**.

As assignments become more difficult, we don't want you to get or feel totally lost. If you've thought long and hard, gone down a long rabbithole on Stack Overflow, and can't even get a sense of what the next step may be, take a step away. Take a break. Then, come back and see if you can't solve it with a refreshed mind. If you're still totally stuck, ask on Piazza, talk to a classmate, and/or attend office hours for help.

With regards to asking questions of instructional staff, we're here to help you, but there are way more students than there are instructors. So, help each other. Ask one another first. It's awesome that we all have different backgrounds and experiences - let's use that to our advantage. In fact, this is how the best data science gets done. Diverse minds solving a problem invariably improves the solution. Also, teaching something to someone else is the best way to determine if you really know something. So, it's win-win. The person who's stuck gets unstuck and the person who helped is more sure in their knowledge. Help one another! Section and office hours are meant to be collaborative.

Also, your instructional staff may not know the answer to everything off the top of their head right away. There is an entire field of data science topics and programming out there - we'll do our best to help and show you where to look and how to figure out the answer (or steer you in a different direction if your approach is going to lead you in a messy and intractable direction), but know that we may not have all the answers.

Deadlines

Assignments will be submitted individually on datahub. We'll talk about the details for submission in class. Assignments will always be released at least a week before the assignment due date. On weeks with assignment deadlines, they will always be due Friday at 11:59 PM of the week specified (see Course Schedule below).

Check to ensure that your file shows up under "Submitted assignments" on datahub after you click submit. If the file is the incorrect file, corrupted, or otherwise unreadable, we cannot grade it and we will mark your assignment as late.

Late assignments will be accepted at 75% credit for 72 hours (3 days) after the assignment's due date. Once the late deadline passes, assignments will be graded, feedback will be made available on datahub, and assignments will no longer be able to be submitted for credit.

Feedback & Grades

It is your responsibility to ensure that we receive a submission from you on datahub and that you submit the correct file (a Jupyter notebook with the .ipynb extension with the same file name as the assignment) for each assignment. If you identify that a mistake has been made, it is your responsibility to get in touch on Piazza should a problem arise. You will receive individualized feedback via email with your grade and feedback about a week after each assignment is due.

Assignment Questions on Piazza

Piazza will be used for all general questions. For example, if you are confused by what a question is asking or are unsure where to start to look for the answer and need direction, Piazza is the place to go. However, when asking or answering questions on Piazza, code that answers assignment questions should **not** be provided. Instead, answer with suggestions as to what topics/ideas/lectures to look into or vague pseudocode that helps move the person asking the question in the right direction. For general programming questions (unrelated to the assignment answers), feel free to share minimal code segments.

Assignment Regrades

We will work hard to grade everyone fairly and return assignments quickly. But, we know you also work hard and want you to receive the grade you've earned. Occasionally, grading mistakes do happen, and it's important to us to correct them.

If you think there is a mistake in your grade, request a regrade within 72 hours of your receipt of the grade on Piazza via a *Private* message to "Instructors" and use the "regrades" tag. This message should include evidence of why you think your answer was correct (i.e. a specific reference to something said in lecture) and should point to the specific part of the assignment in question.

Note that points will *not* be rewarded if you fail to follow instructions. For example, if the instructions say to name the variable `orange` and you name it `ornage` (misspelled), you will not be rewarded credit upon regrade. This is because (1) following instructions and being detail-oriented is important and (2) there are hundreds of students taking the course this quarter. It would be an unfair burden to place on TAs if we didn't have this policy.

COURSE PROJECT

This quarter there will be two options for the final project, each of which is described below:

Option 1: Final Group Project

Your course project will be completed in a group of 4-5 people. The reality of data science is that you will have to work with others. You'll need to work together to communicate effectively, manage time, organize your projects, and accomplish a goal. People will have different knowledge and skills sets. It is your job as a group to work together to figure out how to maximize each group member's skills to make sure that your differences are helpful to accomplishing your goal, rather than a hindrance. For example, some of you will

find the programming aspects of the class assignments very easy, while others will struggle. Alternatively, some of you may find experimental research and hypothesis testing intuitive, while others find it confusing and frustrating. It is best for your project if you choose a team with a mix of background and experience.

Finding A Group

Finding a group may be a tad trickier this quarter. As such, we'll offer additional support. Groups can be found in a few different ways:

1. If you have people in the class you know you want to work with, chat with one another and if you're all on board, form a group.
2. If you don't know people in the class or don't have people you want to work with, no problem. Piazza has a feature where you can look for group mates - check for that post and look through there to find group mates
3. There will be time to find groups in discussion section during week 1.
4. If you are struggling to find a group by week 2, there will be a form to fill out. If you sign up for help finding a group, you will be assigned to a group. We will do our best to form groups among individuals with similar schedules and interests. Groups will be assigned by the middle of week 3.

For students who choose Option 1, groups will be submitted via Google Form by the Friday of week 3 (see Course Schedule). One form will be submitted per group. This will be required of all groups who choose Option 1 (even those we help form), as you will be required to include your GitHub username in this submission.

Project Proposal

You will have to submit a group project proposal by the end of week four (see Course Schedule below) on GitHub. This means that you will have to have to have met as a group by then, have determined what question you want to ask in your project, started to identify the data you'll need/the data you'll use to answer your group's question of interest, have laid out a plan for your project's completion going forward. You will receive feedback on your project proposal to help guide your final project. However, we strongly encourage you to chat with the instructional staff throughout the quarter as you work on your project to elicit more feedback and ensure you're going in the right direction.

Group Project Survey

Every individual in the class will assess how working with their group as a whole and each individual in the group throughout the quarter has been via a short survey. Link to survey will be provided to students. Surveys will be completed individually and are due at the same time as your Final Project (date of the final at 11:59 PM).

Final Project

The final project will be a full, detailed data science report in the form of a Jupyter notebook that carries out an analysis from start to finish. This report will answer the data science question your group has chosen to answer. The topic will be up to you and more details will be provided in class, but generally this report will include (1) background research and ethical considerations, (2) your data science question(s) and hypothesis/hypotheses, (3) data & data wrangling, (4) a descriptive & an exploratory data analysis, (5) your full analysis, (6) your results, and your (7) conclusion(s).

Final Project Extra Credit

Being an effective data scientist requires effective communication. The report you submit will demonstrate your ability to communicate in the written form; however, oral communication is equally important. Extra credit on the final project is optional and can be earned by creating a 3-5 minute video that communicates your group project's question, analysis, and results. This can be a filmed presentation or a video that is more creative. Most importantly, it must effectively communicate your team's project. Videos must be submitted on Canvas by the final project deadline.

Option 2: Individual Final Project

Option 2 will be completed individually and has been designed to mimic the data science interview process. During data science interviews, applicants are often given a dataset, a question, and tasks and sent home to complete the task. This is what students who choose Option 2 will be asked to do. Monday night of Week 10, students will be given a dataset, a topic, and tasks to complete individually. Students will have until the Final Deadline to carry out the data science project on their own. For all aspects of this project, students will have full access to course materials, their own brains and information on the Internet but are not allowed to discuss their approach or analysis with any other humans (this includes, but is not limited to: family members, members of the class, friends, or people online).

Choosing this option

Students who choose Option 2 will have to specify this choice via Google Form by the Friday of week 3 (see Course Schedule). One form will be submitted per individual.

Project Proposal

Students who choose Option 2 will still submit a project proposal by the end of week 3 (see Course Schedule below) on GitHub. This will be completed individually on a topic of *our* choosing.

Final Project Survey

Every individual in the class will provide feedback about their experience completing this option. Surveys will be completed individually and are due at the same time as your Final Project (date of the final at 11:59 PM).

Final Project

The final project will be a full, detailed data science report in the form of a Jupyter notebook that carries out an analysis from start to finish. This report will answer the data science question provided to you during finals week. You will have 5 days to complete your individual final project. We do not anticipate it taking you 5 days straight; however, we know you'll have other finals to study for and take during this time.

More details will be provided in class, but generally this report will include (1) background research and ethical considerations, (2) your data science question(s) and hypothesis/hypotheses, (3) data & data wrangling, (4) a descriptive & an exploratory data analysis, (5) your full analysis, (6) your results, and your (7) conclusion(s).

Final Project Extra Credit

Being an effective data scientist requires effective communication. The report you submit will demonstrate your ability to communicate in the written form; however, oral communication is equally important. Extra

credit on the final project is optional and can be earned by creating a 2-3 minute video that communicates your final project's question, analysis, and results. This can be a filmed presentation or a video that is more creative. Most importantly, it must effectively communicate your project. Videos must be submitted on Canvas by the final project deadline.

DISCUSSION SECTION

Discussion Section will begin week 1.

Section will be used to review material from lecture by getting hands-on programming experience. You will be given tips for working in Python, guided through Jupyter notebooks to clarify topics presented in class, and will be given time to get additional practice. There will be information covered in section that are not covered in lecture and that will be needed (or at least very helpful) for the assignments.

You should be signed up for a section for which you can attend. However, if you are unable to attend the section for which you are signed up, you are free to attend a different section any given week than the one to which you're assigned.

COURSE SCHEDULE

Date	Week	Lecture	Day	Topic	Assignments (due: 11:59 PM PST)
10/2	0	1	F	Welcome!	–
10/5	1	2	M	Data Science	–
10/7	1	3	W	Ethics	–
10/9	1	4	F	Version Control	–
10/12	2	5	M	Data & Intuition	–
10/14	2	6	W	Python Review	–
10/16	2	7	F	Data Wrangling	**A1**: git + python; **Project Planning Survey***
10/19	3	8	M	DataViz I	–
10/21	3	9	W	`pandas`	–
10/13	3	10	F	Intro to Analysis	**Project Proposal***
10/26	4	11	M	EDA	–
10/28	4	12	W	DataViz II	–
10/30	4	13	F	Inference I	**A2**: `pandas`
11/2	5	14	M	Inference II	–
11/4	5	15	W	Inference III	–
11/6	5	16	F	Text Analysis I	**A3**: Data Exploration
11/9	6	17	M	Guest Lecture I	–
11/11	6	–	W	**No Class**	–
11/13	6	18	F	Text Analysis II	**A4**: Data Privacy
11/16	7	19	M	Machine Learning I	–
11/18	7	20	W	Machine Learning II	–
11/20	7	21	F	Text + ML	**A5**: Data Analysis
11/23	8	22	M	Nonparametric	–
11/25	8	23	W	Geospatial I	–
11/27	8	–	F	**No Class**	**A6**: NLP
11/30	9	24	M	Geospatial II	–
12/2	9	25	W	Dimensionality Reduction	–

Date	Week	Lecture	Day	Topic	Assignments (due: 11:59 PM PST)
12/4	9	26	F	Data Science Jobs	–
12/7	10	27	M	Guest Lecture II	–
12/9	10	28	W	Communication	–
12/11	10	29	F	Future of Data Science	–

* indicates (possible) group submission. All other assignments/quizzes/surveys are completed & submitted individually.

OTHER GOOD STUFF

Class Conduct

In all interactions in this class, you are expected to be respectful. This includes following the UC San Diego principles of community .

This class will be a welcoming, inclusive, and harassment-free experience for everyone, regardless of gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, ethnicity, religion (or lack thereof), political beliefs/leanings, or technology choices.

At all times, you should be considerate and respectful. Always refrain from demeaning, discriminatory, or harassing behavior and speech. Last of all, take care of each other.

If you have a concern, please speak with Prof Ellis, your TAs, or IAs. If you are uncomfortable doing so, that's ok! The OPHD (Office for the Prevention of Sexual Harassment and Discrimination) and CARE (confidential advocacy and education office for sexual violence and gender-based violence) are wonderful resources on campus.

Academic Integrity

Don't cheat.

You are encouraged to (and at times will have to) work together and help one another. However, you are personally responsible for the work you submit. For assignments, it is also your responsibility to ensure you understand everything your group has submitted and to make sure the correct file has been uploaded, that the upload is uncorrupted, and that it renders correctly. Projects may include ideas and code from other sources—but these other sources must be documented with clear attribution. Please review academic integrity policies here.

Know that a third of the class typically feels overwhelmed at the start of the quarter. That said, the average is quite high in this course typically (A-). So, while we anticipate you all doing well in this course, if you are feeling lost or overwhelmed, that's ok! Should that occur, we recommend: (1) asking questions in class, (2) attending office hours and/or (3) asking for help on Piazza.

Cheating and plagiarism have been and will be strongly penalized. If, for whatever reason, datahub is down or something else prohibits you from being able to turn in an assignment on time, immediately contact me by emailing your assignment by email (sellis@ucsd.edu), or else it will be graded as late.

Disability Access

Students requesting accommodations due to a disability must provide a current Authorization for Accommodation (AFA) letter. These letters are issued by the Office for Students with Disabilities (OSD), which is located in University Center 202 behind Center Hall. Please make arrangements to contact Prof Ellis privately to arrange accommodations. If you are struggling to get a meeting with OSD, you can let Prof Ellis know and she's likely able to help accommodate while you work to get official documentation. Contacting the OSD can help you further: 858.534.4382 (phone) osd@ucsd.edu (email) http://disabilities.ucsd.edu

Optional Readings:

There are no required readings for this course; however, if you're interested in learning more and reading about data science topics, we recommend the following:

- Donoho D, 50 Years of Data Science
- Wickham H, Tidy Data
- Woo K & Broman K, Data in Spreadsheets
- Tukey JW, Exploratory Data Analysis
- Grus J, Data Science from Scratch

How to Get Your Question(s) Answered and/or Provide Feedback

It's great that we have so many ways to communicate, but it can get tricky to figure out who to contact or where your question belongs or when to expect a response. These guidelines are to help you get your question answered as quickly as possible and to ensure that we're able to get to everyone's questions.

That said, to ensure that we're respecting their time, TAs and IAs have been instructed they're only obligated to answer questions between normal working hours (M-F 9am-5pm). However, I know that's not when you may be doing your work. So, please feel free to post whenever is best for you while knowing that if you post late at night or on a weekend, you may not get a response until the next day. As such, do your best not to wait until the last minute to ask a question.

If you have...

- **Questions about course content**: these are awesome! We want everyone to see them and have their questions answered too...so post these to Piazza!
- **A technical assignment question**: Come to office hours (or post to Piazza). Answering technical questions is often best accomplished in person where we can discuss the question and talk through ideas. However, if that is not possible, post your question to Piazza. Be as specific as you can in the question you ask. And, for those answering, help your classmates as much as you can without just giving the answer. Help guide them, point them in a direction, provide pseudo code, but do not provide code that answers assignment questions.
- **Been stuck on something for a while (>30min) and aren't even really sure where to start**: Programming can be frustrating and it may not always be obvious what is going wrong or why something isn't working. That's ok - we've all been there! IF you are stuck, you can and should reach out for help, even if you aren't exactly sure what your specific question is. To determine when to reach out, consider the 2-hour rule. This rule states that if you are stuck, work on that problem for an hour. Then, take a 30 minute break and do something else. When you come back after your break, try for another 30 minutes or so to solve your problem. If you are still completely stuck, stop and contact us (office hours, post on Piazza). If you don't have a specific question, include the information you have (what you're stuck on, the code you've been trying that hasn't been happening, and/or the error messages you've been getting).
- **Questions about course logistics**: First, check the syllabus. If the answer is not there, ask a classmate. If you still are unsure, post on Piazza

- **Questions about a grade**: For programming assignments, reply to the COGS 108 email directly; For project-related regrades, post a note to instructors on Piazza and select the 'regrades' tag. Include specifics as to why you feel you mistakenly/unfairly lost points in that post.
- **A specific section-related question**: send a direct message on Piazza to your TA/IA
- **Something super cool to share related to class**: feel free to email Prof Ellis or come to office hours. Be sure to include COG S108 in the email subject line and your full name in your message.
- **Something you want to talk about in-depth**: meet in person during office hours or schedule a time to meet by email. Be sure to include COGS 108 in the email subject line.
- **Some feedback about the course you want to share anonymously**: If you've been offended by an example in class, really liked or disliked a lesson, or wish there were something covered in class that wasn't but would rather not share this publicly, etc., please fill out the anonymous Google Form*

*This form can be taken down at any time if it's not being used for its intended purpose; however, you all will be notified should that happen.

What should you call me?

Most students call me Professor/Prof Ellis, and that's great! This is how I typically sign emails to students. I'm also totally OK with you addressing me as Shannon or Dr. Ellis. I would prefer you *not* address me as Ms./Miss/Mrs. Ellis (but I likely won't correct you. . . .it's really not that big of a deal).

What I should call you

I should call you by your preferred name, with the correct pronunciation. Please correct me (either in the chat, out loud on zoom, or via email/Piazza after the fact. . . however you're most comfortable) if I ever make a mistake.