COGS 108: Data Science in Practice
Fall 2019
T/TH 9:30a-11:50a (Price Center Theater)

Instructor: Prof. Bradley Voytek

Office hours*:

Day	Time	Location	Instructor name	Instructor email
*Thu	11:00a-12:00p	CSB 169	Prof. Bradley Voytek	bvoytek@ucsd.edu
Mon	10:00a-11:00a	CSB 114	Akshansh Chahal (TA)	a3chahal@eng.ucsd.edu
Wed	10:00a-11:00a	SSRB 100	Samuel Lau (TA)	lau@ucsd.edu
Fri	12:00p-2:00p	Price Center study room 4	Enlin Wei (TA)	enwei@ucsd.edu
Wed	3:00p-4:00p	CSB 114	Miguel Garcia (IA)	mig053@ucsd.edu
Mon	10:00a-11:00a	CSB 114	Anona Gupta (IA)	arg060@ucsd.edu
Tue	3:00-4:00p	CSB 114	Xirui Li (IA)	xil475@ucsd.edu
Tue	2:00p-3:00p	CSB 114	Xuanchen Lu (IA)	xul076@ucsd.edu
Fri	9:00a-11:00a	CSB 114	Sutianyi Wen (IA)	s5wen@ucsd.edu
Wed	3:00p-4:00p	CSB inner courtyard	Ziwen Zeng (IA)	ziz236@ucsd.edu

*also by appointment

Instructional staff and section times

Section	ID	Day	Time	Location	TA/IA name(s)
A01	983873	Mon	2:00p	PCYNH 121	Enlin Wei, Xirui Li
A02	983874	Wed	9:00a	PETER 104	Sam Lau, Anona Gupta
A03	983875	Wed	4:00p	WLH 2111	Sam Lau, Ziwen Zeng
A04	983876	Fri	2:00p	PCYNH 121	Enlin Wei
A05	983877	Fri	3:00p	PCYNH 121	Xuanchen Lu, Miguel Garcia
A06	983878	Fri	4:00p	PCYNH 121	Suitanyi Wen, Xirui Li
A07	996902	Fri	11:00a	PCYNH 121	Akshansh Chahal, Ziwen Zeng, Anona Gupta
A08	996903	Fri	12:00p	PCYNH 121	Akshansh Chahal, Miguel Garcia

Course email: cogs108@gmail.com
Course GitHub: https://github.com/COGS108/
DataHub: https://datahub.ucsd.edu/
Piazza*: https://piazza.com/ucsd/fall2019/cogs108
Course podcast/screencast: https://podcast.ucsd.edu/?q=COGS108
Course Gradescope Entry Code: TBD

*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
- iClickers are not used
- All materials will be provided on GitHub

CLASS TECHNOLOGY

- Python 3 (Anaconda distribution)
- Jupyter Notebooks
- git and GitHub (option to use SourceTree GUI)

COURSE OVERVIEW

This class is a hands-on practical programming course intended to get you experience working on data science projects. Specifically, our goal is to give you the time and critical thinking skills necessary to identify, engage with, synthesize, and analyze publicly available datasets. We want to train you on what we call the “automation-resistant” data science skills: data-first thinking; data literacy, intuition, and creativity; integration of heterogeneous data types; data communication, visualization, and storytelling; and team-oriented projects. We want you to walk away with a stronger data science portfolio that shows off your skills, cleverness, and creativity.

In COGS9 (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. This process frequently involves attempting to do something, doing it wrong, learning from your mistakes, and then iterating until success. 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 neuroscience, data science, and data parameterization.
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."

GRADING and ATTENDANCE

Grading:

	% of Total Grade
Five (5) Assignments	50
One (1) Project Proposal	10
One (1) Final Project	35
Guest Lectures	5

Final exam date: No final exam (or any exams, for that matter), only a final project deadline, with the final project due by 11:59p on the class final exam date: **Thursday, December 12**.

Grades

All grades will be released on Canvas. We will try to send out automated alerts if we do not receive a submission and/or if it fails to be processed, but *ultimately it is your responsibility to check your grades and get in touch if any are missing or you think there is a problem*.

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 on Gradescope (for project proposal and final project submitted to Gradescope) or on Piazza (for programming assignments) *within 72 hours of your receipt of the grade*. This post 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. When submitting on Piazza, address it to "Instructors," and select the tag "regrades."

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," you will not be rewarded credit upon regrade. This is because (1) following instructions and being detail-oriented is important and, (2) there are almost 500 students taking the course this quarter.

Lecture Attendance
Our goal is to make lecture and discussion section worth your while to attend. However, most days it will be up to you whether or not you show up. Attendance is only *required* on guest lecture days—these will be announced at least a week in advance in lecture and on Piazza.

Section Attendance
Attendance is not required. But, to get the most out of this course and get the most amount of practice possible, we suggest attending section regularly. You'll want to try to attend the section to which you're assigned; however, if something comes up one week or you can't make your assigned section, you are free to attend another section *as long as you first clear it with your section TA/IA*.

ASSIGNMENTS
Assignments are hands-on in this course. They will be completed individually in Jupyter Notebooks and 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 you'll become more comfortable with this. You'll get stuck, and then 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 assignment will be the simplest assignment. If the first assignment is 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 second 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 rabbit hole 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, or set up 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 one other as well! 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 offers new solutions. 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 surer 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 and section.) Assignments will always be released on Mondays, and you will always have just shy of two weeks from assignment release date to assignment due date. On weeks with assignment deadlines, they *will always be due Sunday at 11:59p of the week specified (see Course Outline below)*.

Check the files you submit immediately after you submit them. 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 earn fractional credit (75% within one week late; no late assignments accepted after one week).

Feedback and 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) 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.

COURSE PROJECT
Your course project will be completed in a group of 4-6 (assigned to you during the second week of class), *even if you don't want to work with a group.* No exceptions. 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 skillsets. 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. Teams will be chosen, to the best of our ability, so that there is a mix of background and experience.

If your assigned team has members who drop the class after group assignment, it *is* okay if your group ends up with fewer than 4 members. However, if this occurs and you would *like* additional members, please reach out as soon as possible to Professor Voytek to have a new member added to your team.

Project Proposal
You will have to submit a group project proposal by the end of week three (see course outline below) on Gradescope. 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, and started to identify the data you'll need/the data you'll use to answer your group's question of interest. 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.

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 and data wrangling, (4) a descriptive and an exploratory data analysis, (5) your full analysis, (6) your results, and your (7) conclusion(s). Along with your report, each individual will submit feedback about the team and each individual members' contributions to the project. Details about this will be sent out during week 5.

DISCUSSION SECTIONS
Section will be used to review material from lecture by getting hands-on programming experience. You will be guided through Jupyter notebooks during section that will allow you to get practice. There are two general types of workbooks that will be used in discussion section: 1) closely-guided step-by-step workbooks that walk you through data science concepts, getting hands-on practice and, 2) case study workbooks that provide you with data but less guidance so that you have to make data decisions. Workbooks with "answers" will be provided each week following that notebook's completion in section.

Note that 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. See course outline below for what general topic will be covered in section each week.

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*
- Joel Grus, *Data Science from Scratch*

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. Voytek, 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 us via Piazza and emailing your assignment directly to Prof. Voytek (bvoytek@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. Voytek privately to arrange accommodations.

Contacting the OSD can help you further:
858.534.4382 (phone)
osd@ucsd.edu (email)
http://disabilities.ucsd.edu

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, we *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 (>60min) 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, check or post on Piazza or ask a classmate.

- **Questions about a grade:** Post as a question on Piazza, address it to “Instructors,” and select the folder “regrades”.

- **A specific section-related question:** Send a direct message on Piazza to your TA.

- **Something super cool to share related to class:** Feel free to email Dr. Voytek (bvoytek@ucsd.edu) or come to office hours. Be sure to include COGS108 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 email to schedule a time to meet. Be sure to include COGS108 in the email subject line. (bvoytek@ucsd.edu).

COURSE OUTLINE

Week	Topic(s)	Due	Covered in Section
0	Intro	—	—
1	Guest lecture: Ben Cipollini (Facebook) Parameterization	Course Survey	Intro
2	Python, Data Science, and Ethics Data Wrangling and Visualization	A1	Python Basics
3	Introduction to Analysis	Project Proposal*	Data Wrangling
4	EDA	A2	EDA and Data Visualization
5	Probability and Statistics	—	EDA
6	Inference	A3	Inference
7	Nonparametric Approaches Text Analysis (NLP)	—	Inference
8	Machine Learning	A4	Text Analysis
9	Geospatial Analysis Dimensionality Reduction	—	Machine Learning
10	Future of Data Science	A5	Work on Projects

Final Project*: due **Thursday, December 12** of finals week by 11:59pm.

*indicates group submission. All other assignments/surveys are completed and submitted individually.