COGS 9 | Introduction to Data Science

Meeting: Fall 2023, TuTh 9:30a-10:50p (all times Pacific)
Location: WLH 2005
Instructor: Professor Bradley Voytek (https://voyteklab.com)

Course Piazza*: https://piazza.com/ucsd/fall2023/cogs9
Course GitHub: https://github.com/IntroDataSci
*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.

One-on-one Q&As:

Date & Time	Location	Instructional Staff	email
*** Tuesdays 11:00a-12:00p	HDSI 451	Prof. Bradley Voytek	bvoytek@ucsd.edu
Tuesdays 12:00p-1:00p	SSRB 204	TA: James Michaelov	j1michae@ucsd.edu
Wednesdays 11:00a-12:00p	zoom	TA: Harshada Yadav	hyadav@ucsd.edu
Fridays 1:00p-2:00p	CSB 114	IA: Evelyn Huang	xih037@ucsd.edu
Tuesdays 3:30p-4:30p	zoom	IA: Vicky Li	yil164@ucsd.edu
Wednesdays 1:00p-2:00p	Art of Espresso (Mandeville)	IA: Tommy Shen	twshen@ucsd.edu
Tuesdays 7:00p-8:00p	CSB 114	IA: Vivian Tran	avt001@ucsd.edu

** Also available for virtual appointment
***Also by appointment

Sections

Section	Day	Time	Location	Staff
A01 (267832)	M	5:00-5:50p	PCYNH 120	James & Vivian
A02 (267890)	W	10:00-10:50a	CENTR 222	Harshada & Vicky
A03 (267964)	M	6:00-6:50p	PCYNH 120	Evelyn & Vivian
A04 (267978)	W	12:00-12:50p	CENTR 222	Harshada & Tommy
A05 (267979)	M	3:00-3:50p	YORK 4080A	James & Evelyn

COURSE OBJECTIVES

- Define terminology for core concepts in data science.
- Learn to think critically about data, and how to approach problems with a “data-first” mindset.
- Introduce the basics of data visualization and practice basic storytelling about data.
- Inspect and work through problems demonstrating “p-hacking”, related to ethical data science.
- Discuss data privacy and ethics considerations, using real-world examples.
- Explain examples of real-world data science projects that have been pivotal for understanding aspects of human behavior, language, and society that have helped scientific progress, and business.
- Give students first-hand experience with common pitfalls of data analyses and how to avoid them.

COURSE MATERIALS

- There is no textbook
- All materials will be provided through Canvas

GRADING & ATTENDANCE

Grading:

	Points	**% of Total Grade**
(4) Assignments	40 pts each; **160 total**	40
(2) Exams	50 pts each; **100 total**	25
(5) Readings	20 pts each (one dropped); **80 total**	20
(1) Final Project	**40 total**	10
(*n*) Guest Lectures	20/*n* pts each; **20 total**	5

Letter Grade	**From**	**To**
A+	97.00	100.00
A	93.00	96.99
A-	90.00	92.99
B+	87.00	89.99
B	83.00	86.99
B-	80.00	82.99
C+	77.00	79.99
C	73.00	76.99
C-	70.00	72.99
D+	67.00	69.99
D	63.00	66.99
D-	60.00	62.99
F	0.00	59.99

Notes:

- Final exam date: No final exam, only final project deadline.
- There are 400 possible points to be earned in this course. To determine your final grade, you will add up all of the points for the above categories and divide your grade by 4. Your letter grade will be determined using the standard grading scale. Grades are *not* rounded up.

Semi-synchronous lecture and synchronous discussion sections

Lectures will be given live during the normal class time, however they will also be recorded for later viewing.

Lecture Attendance

Our goal is to make lecture and discussion section worth your while to attend. However, you have the choice to *not* attend lectures *or* discussions. That said, attendance *is required* on guest lecture days, and that attendance will constitute your Guest Lecture attendance grade.

Grades

Grades for assignments and exams will be released on Canvas approximately a week after the submission date. *It is your responsibility to ensure your assignments are submitted on time and to check your grades and get in touch if any are missing or if 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 for an assignment, submit a regrade request on Gradescope within 72 hours of receipt of the grade. This request 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 for us to reconsider.

Discussion Sections

Discussion sections will be used to review content from lectures, discuss readings, and guide your 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 in which you're assigned.

COURSE TOPICS & ASSIGNMENTS

This class is a survey course intended to get you all excited about becoming data scientists! Data are everywhere and they're being used in tried-and-true—as well as in new, awesome, and creative—ways. This course will introduce you to topics in data science, discuss what it means to be a data scientist, and get you on your way to thinking like a data scientist. To see what topics will be introduced in this course, see the calendar at the end.

Assignments

There are four assignments. Assignments will focus on applying the concepts covered in lecture and ensuring you're on the right track for your final project.

Assignments will be posted on the Thursdays of the week before they're due, and you have until *Friday at 11:59 PM* of the following week (so a little over 8 days) to complete each assignment. Assignments will be released on Canvas and submitted on Gradescope. *Assignments will always be due Fridays at 11:59 PM*. Assignments 1, 3, and 4 are submitted individually. The second assignment (and your final project) will be turned in *as a group*. You will receive feedback along with a grade within a week from the due date. *Feedback from A2 should be incorporated into your final project.*

Late assignments earn fractional credit (75% within one week late; no late assignments accepted after one week).

Final Project

The final project is a report on how you would handle a complicated data science project. It's essentially a culmination of the four assignments all tied together in a nice report. This report will include all the nitty gritty, whys, and hows of the analysis you have chosen. You'll specify your data science question, find data that could be used to answer the question, and describe the analysis you *would* carry out to answer your question of interest. You WON'T have to actually perform the analysis to answer the question; you'll just write about it. This will be turned in as a PDF.

You can choose your final project groups of 4-5 people. If you do not have a group, Professor Voytek will assign one. There will be time to work on and discuss your second assignment and projects in section, so we recommend (but do not require) you form groups within the section you plan to attend.

Exams

There will be two exams covering material in lecture (including guest lectures!) and the readings discussed in section. They will be closed notes; you may not use any outside resources. Exams will be primarily multiple choice with a few short answer questions. See schedule below for in-class exam dates.

Readings

There will be five weeks where readings are assigned. Just like the Assignments, Readings will be posted on the Thursdays of the week before they're due, and you have until *Friday at 11:59 PM* of the following week (so a little over 8 days) to complete the reading quiz assignment *on Canvas*. You are not timed. **You *must* click submit** to submit your reading quizzes. Your **most recent submission** will be graded—**you only get three attempts** (with unlimited attempts it would be possible to figure out all the answers simply through trial-and-error). If you fail to finish and submit your quiz answers before the deadline, it will not be graded **and you will receive an automatic zero**. Your lowest reading quiz score will be dropped.

***No late credit will be given if Reading quiz assignments are submitted after the due date*.**

Planned Readings:

All of the below readings are available on the class GitHub page at:
https://github.com/IntroDataSci/Readings

- R1: Donoho D, *50 Years of Data Science*
- R2: Keyes O, Hutson J, & Durbin M, *A Mulching Proposal*
- R3: Wickham H*, Tidy Data*
- R3: Woo K & Broman K, *Data in Spreadsheets*
- R4: Peck, E, Ayuso S, & El-Etr O, *Data Is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania*
- R5: Angwin J, Larson J, Mattu S & Kirchner L*, Machine Bias*

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: https://ucsd.edu/about/principles.html

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 Dr. 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.

OPHD: https://blink.ucsd.edu/HR/policies/sexual/OPHD.html
CARE: https://care.ucsd.edu/

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 at: https://academicintegrity.ucsd.edu/

We anticipate you all doing well in this course; however, if you are feeling lost or overwhelmed, that's ok! Should that occur, we recommend: (1) attending discussions and leveraging the time there, (2) attending weekly one-on-one hours with Dr. Voytek and the course TAs and IAs and/or, (3) browsing Piazza.

Cheating and plagiarism have been and will be strongly penalized. If, for whatever reason, Canvas 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 (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 *Pepper Canyon Hall Suite 300*. Please make arrangements to contact Dr. 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, 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 weekday. As such, do your best not to wait until the last minute to ask a question.

Finally…

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!
- **questions about course logistics**: first, check the syllabus. If you can't find the answer, ask a classmate. If still unsure, post on Piazza.
- **questions about a grade**: If for an assignment, submit a regrade request on Gradescope. For anything else, post as a question on Piazza, address it to "Instructors," and select the folder "regrades"
- **something super cool to share related to class**: feel free to email Dr. Voytek (bvoytek@ucsd.edu) or come to one-on-one hours. Be sure to include COGS9 in the email subject line and your full name in your message.
- **something you want to talk about in-depth**: meet during weekly one-on-one hours or schedule a time to meet by email. Be sure to include COGS9 in the email subject line. (bvoytek@ucsd.edu).

Schedule (exact guest lecture dates subject to change, but assignment and exam dates are stable)

Week	Date	Title	Schedule
0	Thu Sep 28	01 - Introduction	class lecture
1	Tue Oct 03	02 - What is Data Science?	class lecture
1	Thu Oct 05	03 - Privacy and ethics	class lecture
1	Fri Oct 06	--	--
2	Tue Oct 10	04 - Ethics (cont.)	class lecture
2	Thu Oct 12	05 - Data visualization	class lecture
2	Fri Oct 13	--	***R1: Data Science quiz***
3	Tue Oct 17	06 - Data and information	class lecture
3	Thu Oct 19	07 - Data and storytelling	class lecture
3	Fri Oct 20	--	***A1: Data Viz due***
4	Tue Oct 24	*Guest lecture:* Alexandra Keamy NLP Data Scientist: *Leidos*	***guest lecture 1***
4	Thu Oct 26	08 - Algorithms and crowds	class lecture
4	Fri Oct 27	--	***R2: Data Ethics quiz***
5	Tue Oct 31	EXAM 1	***exam 1***
5	Thu Nov 02	*Guest lecture:* Parker Addison Consultant, AI & Data Science: *Deloitte*	***guest lecture 2***
5	Fri Nov 03	--	***A2: Final Outline due***
6	Tue Nov 07	09 - Data wrangling	class lecture
6	Thu Nov 09	10 - Hypothesis testing and EDA	class lecture
6	Fri Nov 10	--	***R3: Tidy Data quiz***
7	Tue Nov 14	11 - Prog and version control	class lecture
7	Thu Nov 16	12 - Stats and probability	class lecture
7	Fri Nov 17	--	***A3: p-values due***
8	Tue Nov 21	13 - Inference and machine learning	class lecture
8	Thu Nov 23	*Thanksgiving holiday: no class!*	***holiday: no class!***
8	Fri Nov 24	--	***R4: Data Viz quiz***
9	Tue Nov 28	*Guest lecture:* Christopher Hannemann NLP Data Scientist: *Dexcom*	***guest lecture 3***
9	Thu Nov 30	14 - Geospatial analysis and text-mining	class lecture
9	Fri Dec 01	--	***A4: ML due***
10	Tue Dec 05	EXAM 2	***exam 2***
10	Thu Dec 07	15 - The Future of Data Science	class lecture
10	Fri Dec 08	--	***R5: Algorithms quiz***
finals	Fri Dec 15	Final Project Deadline (DO NOT SHOW UP)	***Final Project deadline***