If you walked in here feeling sick,
please turn around and walk back out.
Podcast recording of this class is turned on

In-person illness policy

Please do not attend any in-person activity (lecture/section/office hours) if you are feeling ill, especially if you are sneezing/coughing and have a fever. If you feel mildly ill but without sneezing/coughing, or if you have bad allergies, then you may come to in-person events while wearing a well-fitting mask.

Reminder: This (and all lectures) in COGS 108 are being recorded.

Welcome to COGS 108

Data Science in Practice

Jason G. Fleischer, Ph.D
UC San Diego

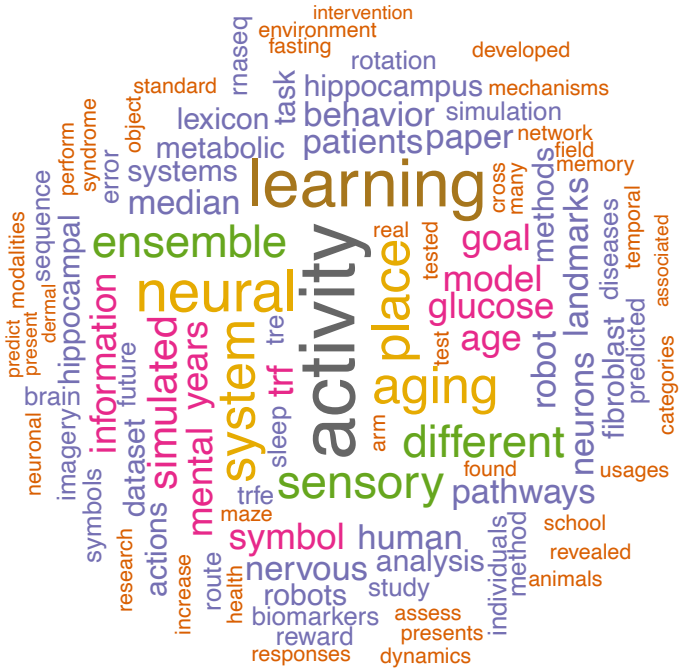
Department of Cognitive Science
jfleischer@ucsd.edu
https://jgfleischer.com

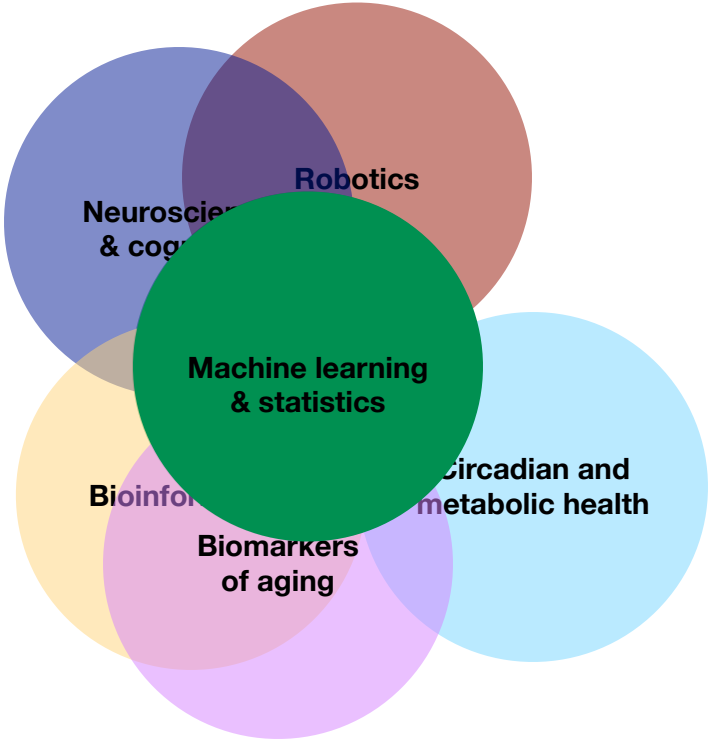
hello

my name is

Jason Fleischer

A bit about me





San Diego Wave FC

5th in NWSL

Overview | Matches | Standings | Players

Matches

Orlando Pride	2	FT	San Diego Wave	3	FT
San Diego Wave	1	Yesterday	Utah Royals FC	2	Sat, Mar 22
Angel City FC	1	FT	San Diego Wave		Sat, Apr 12
San Diego Wave	1	Sun, Mar 16	KC Current		7:00 PM

Standings

NWSL

	Club	MP	W	D	L	GD	Pts
3	Washington Spirit	3	2	0	1	1	6
4	Angel City FC	3	1	2	0	1	5
5	San Diego Wave	3	1	1	1	0	4
6	Seattle Reign FC	3	1	1	1	0	4

San Diego FC

MLS

Overview | Matches | Standings | Players

Matches

MLS

San Diego FC	3	Full-time Yesterday	2	LAFC

Match recap | Match recap

Standings

MLS
Western Conference

	Club	MP	W	D	L	GD	Pts
1	Vancouver	6	4	1	1	5	13
2	Austin	6	4	0	2	2	12
3	San Diego FC	6	3	2	1	4	11
4	Minnesota	6	3	2	1	3	11

Waitlist information

I know this matters to you and is a source of stress and I hate it and wish it could be different!

But please do not email me asking questions about the waitlist... I generally get more than a dozen of these per day during the beginning of quarter. That's annoying and sucks up a lot of time if I reply. Time I can't spend helping students.

I will put everything that I know about the waitlist here

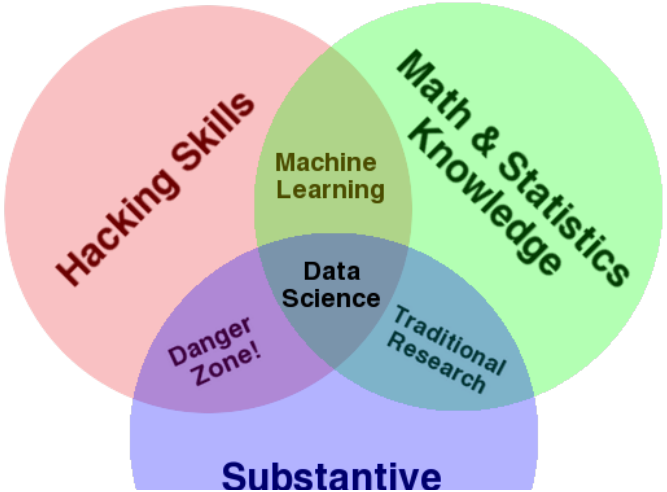
- I have no direct control over the waitlist. If you have questions contact cogsadvising@ucsd.edu
- A few people in each section typically get off the waitlist, but that number varies each quarter.
- We have about 720 students enrolled in two lecture section with 7 lab sections per lecture.
- There are about 180 on the waitlist at last look
- We will likely admit up to about 750 total enrolled. And the exact number admitted will depend on very complicated circumstances.
- We will prioritize students for admission in the following order:
    - COGS students, starting with those with senior status
    - Senior status and graduate students
    - After that its waitlist position
- Your wait list position is in your LAB section. There are 7 labs sections per lecture right now. So if you're 4th on the waitlist of your section, you can expect there are up to 28 people (7 sections x 4 people) in front of you once we admit all the priority students.
- People will be let off the waitlist in the 2nd week. Some people will be let off the waitlist each day that week. The waitlist closes at the end of week 2, if you're not off by then you will not be admitted.
- If YOU REALLY WANT TO BE HERE KEEP UP WITH THE WORK WHILE ON THE WAITLIST!. There will be no chances to hand in things that are past deadline just because you didn't do them while on the waitlist

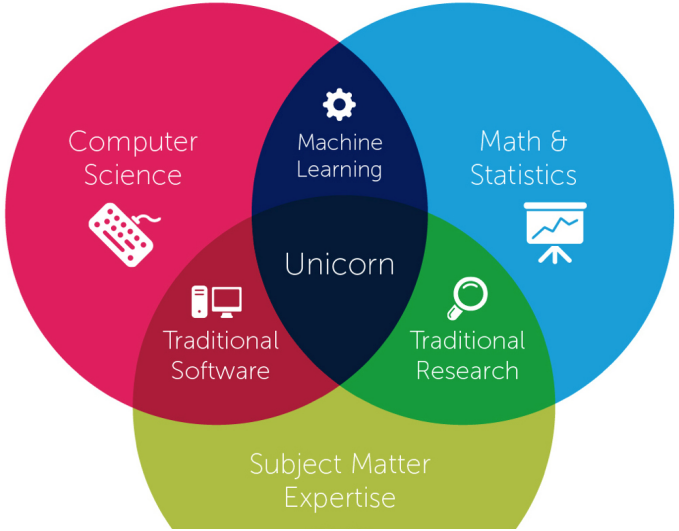
If you have a question not covered above, or if you think there's a good reason why you should be admitted regardless of the waiting list, then you should email cogsadvising@ucsd.edu not me.

May the odds be ever in your favor.

"The scientific process of extracting value from data"

What is data science?





Thinking clearly with help from data

http://

Data scientist is actually MANY jobs

https://hbr.org/2018/11/the-kinds-of-data-scientist

A final piece of advice for those hiring data scientists: Look for people who are in love with solving problems, not with specific solutions or methods, and for people who are incredibly collaborative. No matter what kind of data scientist you are hiring, to be successful they need to be able to work alongside a vast variety of other job functions — from engineers to product managers to marketers to executive teams. Finally, look for people who have high integrity. As a society, we have a social responsibility to use data for good, and with respect. Data scientists hold the responsibility for data stewardship inside and outside the organization in which they work.

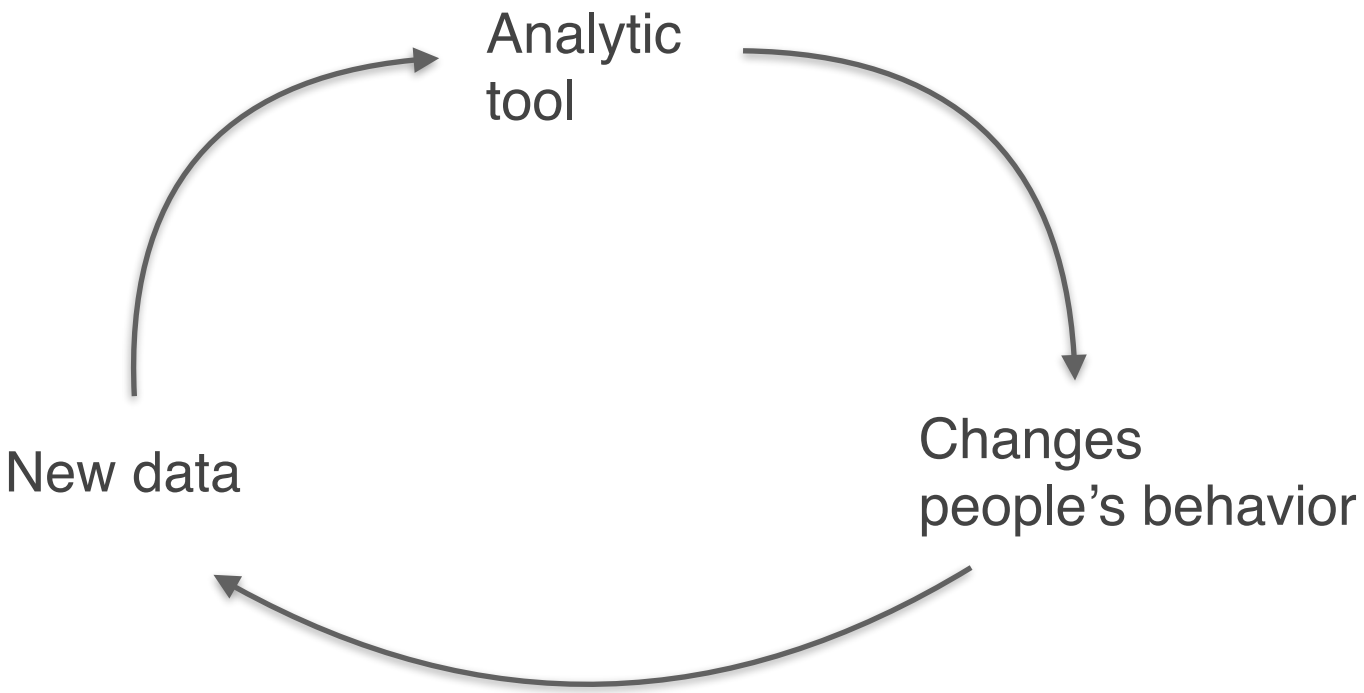
Data science for humans

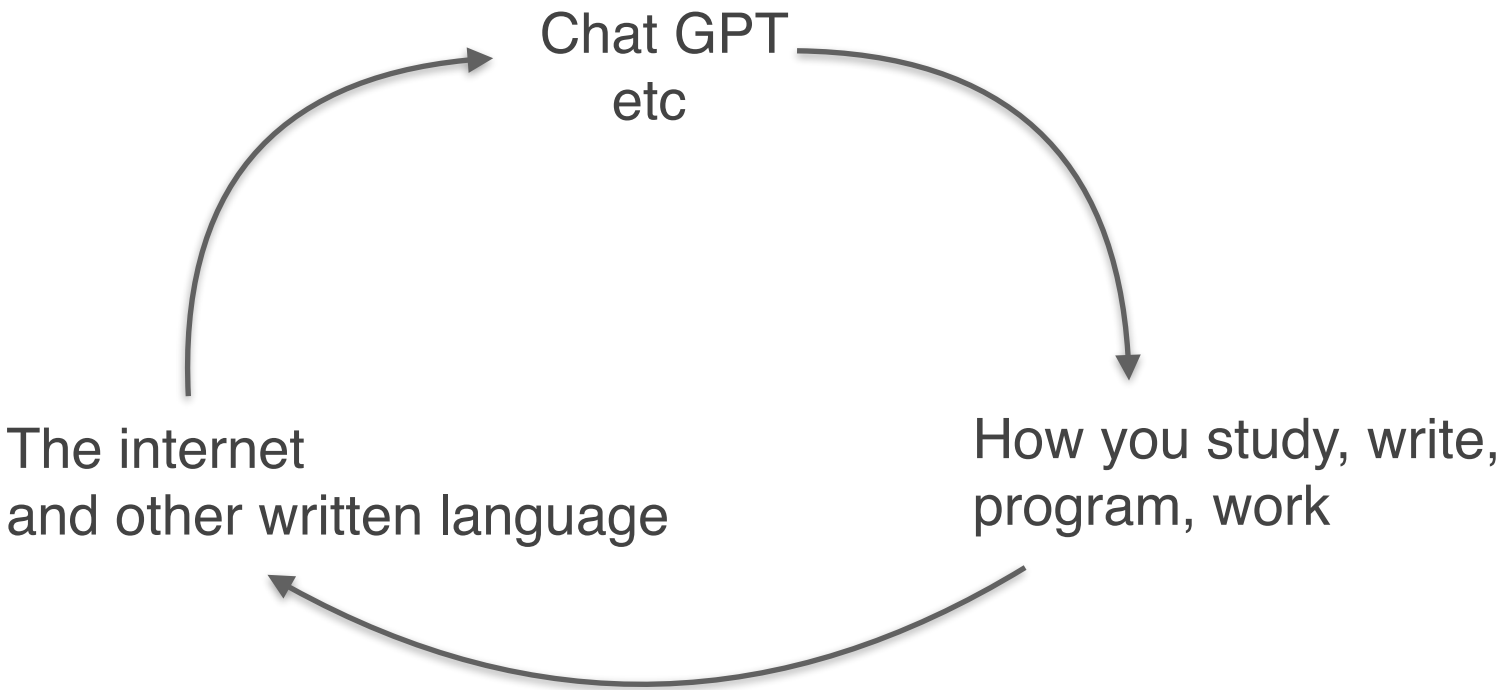
Data science for computers

Data science is inherently social, cultural, and political

When do we care about DS?

Only if the model, analysis, tools, etc are useful to people.


Analytic
tool
Changes
people's behavior
New data


Chat GPT
etc
How you study, write,
program, work
The internet
and other written language

The New York Times

The Hustlers Who Make $6,000 a Month by Gaming Citi Bikes

The bike-sharing program rewards users who help redistribute bikes around New York City. A few riders have figured out how to turn that into profit.

A full dock of Citi Bikes on 34th Street means there's money to be made by moving them elsewhere. Yuvraj Khanna for The New York Times

PDF on Canvas

It was the perfect New York hustle, a scam of subtle perfection. And for three years, it helped Mark Epperson pay his rent.

The hustle, in its simplest form: Borrow a Citi Bike. Ride it one block. Wait 15 minutes, then ride it back.

Earn $6,000 a month (under ideal conditions, and with lots of work).

Occasionally, though, a ride to work ends with the rider's discovery that the docking station nearest the office is full. A dash to brunch is foiled by an empty dock, with no bikes available.

Both situations are annoying, especially for Citi Bike subscribers, who now pay $220 a year. To fix the imbalance, Citi Bike uses various tactics to move bikes to in-demand stations.

[One] is a program called Bike Angels, in which Citi Bike users move bikes in exchange for points that could be cashed in for swag like water bottles and backpacks, membership discounts and gift cards.

“We imagined people would do it as a recreational fitness kind of thing,” said David B. Shmoys, a data scientist at Cornell University whose research team created Citi Bike’s first rebalancing algorithm in 2014. “We never imagined anyone getting really obsessed.”

Over the years, a few users found ways to maximize the program's financial benefits. [Which works like congestion pricing]

But a few riders realized that by working as a team, and quickly, they could exploit the algorithm. For Mr. Epperson and his fellow hustlers, it "created an opportunity to make a lot of money," he said.

At 10 a.m. seven Bike Angels descended on the docking station at Broadway and 53rd Street, across from the Ed Sullivan Theater. Each rider [..] unlocked a bike [and] rode it one block east, to Seventh Avenue […] docked, ran back to Broadway, unlocked another bike and made the trip again.

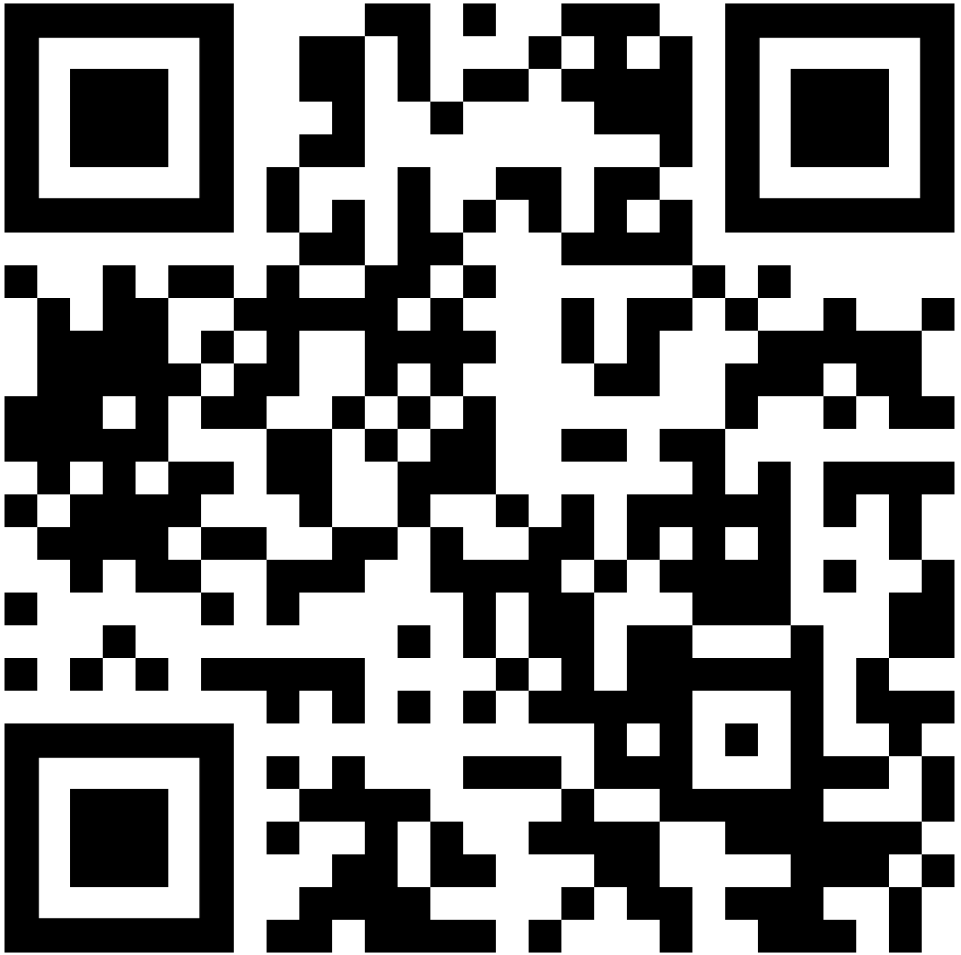
By 10:14, the crew had created an algorithmically perfect situation: One station 100 percent full, a short block from another station 100 percent empty. The timing was crucial, because every 15 minutes, Lyft's algorithm resets, assigning new point values to every bike move.

The clock struck 10:15. The algorithm, mistaking this manufactured setup for a true emergency, offered the maximum incentive: $4.80 for every bike returned to the Ed Sullivan Theater. The men switched direction, running east and pedaling west.

Think-Pair-Share

1. Fill out google form
2. Pair up, introduce yourself, and discuss this topic
3. 3+ hands up and we discuss

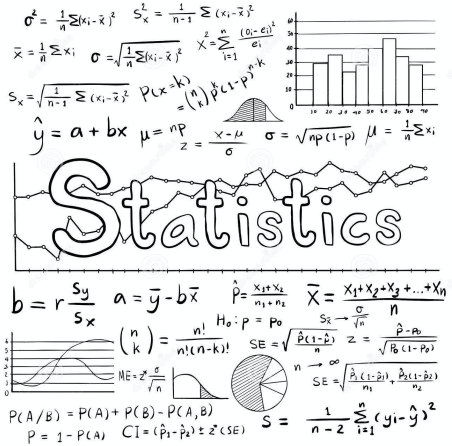
https://forms.gle/4UG7nugVvQrS4zQr7

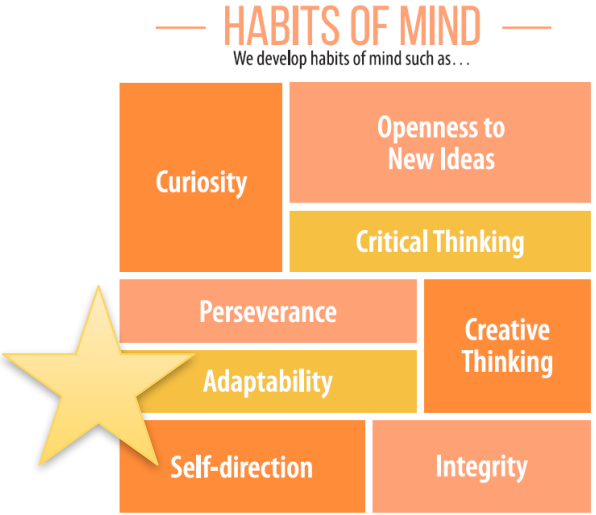


Data scientists are people who try to think clearly with data.

Good data scientists try to understand how their tools work (technically) but also how their tools affect society, culture, and politics

How this class will train you for DS





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

Programming Prerequisite

- MAE 8 - MATLAB
- CSE 8A or 11 - Python/Java
- COGS 18 - Python
- DSC 10 - Python

Bottom line: we will assume programming knowledge.

Python will be used for all labs/projects/assignments.

No programming experience (or you forgot it all)?

- *Preferred option*
  - Take a programming course first
  - COGS 18 : Introduction to Python
- *Can't wait?*
  - Use online sites like *codecademy.com* or *LearnPython.org*
  - *Python Data Science Handbook*

COGS 108: General Plan

Week	Topic(s)
1	Data Science tools
2	Data Intuition & Wrangling
3	Data Ethics & Questions
4	Data Visualization & Data Analysis
5	Inference
6	Text Analysis
7	Machine Learning
8	Nonparametric Analysis
9	Geospatial Analysis
10	Data Science Communication & Jobs

Course links

GitHub	https://github.com/COGS108	lecture/section materials & final projects
datahub	https://datahub.ucsd.edu	assignment submission
EdStem	https://edstem.org/us/join/an6dhT	questions, discussion, and regrade requests. There was an email invite you must accept!
Canvas	https://canvas.ucsd.edu/courses/64550	grades, lecture videos
Anonymous Feedback	https://forms.gle/MnBrvofZY7YxnwYMA	general feedback on what's going well or badly

Discussion Section

- Goals:
  - MORE chance for individual contact
  - help with technical aspects of the course
  - assignment & project help
- Can I switch sections? Yes, but stick with one for the duration
- You'll never be required to go to section.
  - Do lab exercises on your own if you feel comfortable with material
  - Questions via EdStem if you can't attend
- At least one section is always recorded

Discussion Sections start in Week 1 with a Python review!!

Day due		% of Grade
M	(8/9) Weekly Quizzes (lecture content)	8
F	(8) Discussion Labs (technical)	16
W	(4+1) Assignments	33
W	Final Group Project	44
	(1) Project Review*	5
	(1) Project Proposal*	9
	(2) Project Checkpoints*	10
	(1) Final Report*	15
	(1) Final Video*	3
	(1) Team evaluation survey	1

Weekly Lecture Quizzes:

- (9) weekly quizzes (first one due Monday of Week 2, covering Week 1)
- Goal: to help you keep on top of the material covered in lecture
- Why?: experience + student feedback
- How:
  - Taken on Canvas
  - Single Attempt
  - ~10 Questions
  - Posted by Friday sometime after class and before midnight; due the following Mon
  - Meant to test concepts from previous week's lecture

Lecture quizzes will be due on Mon by 11:59 PM.

Lowest quiz score will be dropped.

NO LATES

8 Discussion Lab exercises

Completed individually and graded partly manually (for effort and good thinking) and partly programmatically (for correctness).

- These are meant to get you practice programming around the topics covered in class.
- You will have to look some stuff up on your own. This is by design.
- Instructions must be followed perfectly to receive credit.
- You'll have the opportunity to practice in discussion section.

Discussion labs will be due on Fridays by 11:59PM

75% credit if submitted less than 5 days after deadline.

7 LATE DAYS allowed per person without penalty
to be used for Discussion Labs + Assignments

(4 + 1 practice) Assignments

Completed individually and graded almost completely programmatically.

- These are meant to get you practice programming around the topics covered in class.
- The first two are much simpler/shorter, the last two are harder/longer.
- You will have to look some stuff up on your own. This is by design.
- Instructions must be followed PERFECTLY to receive credit.
- You'll have the opportunity to practice in discussion section.

Assignments will be due on Wednesdays by 11:59 PM

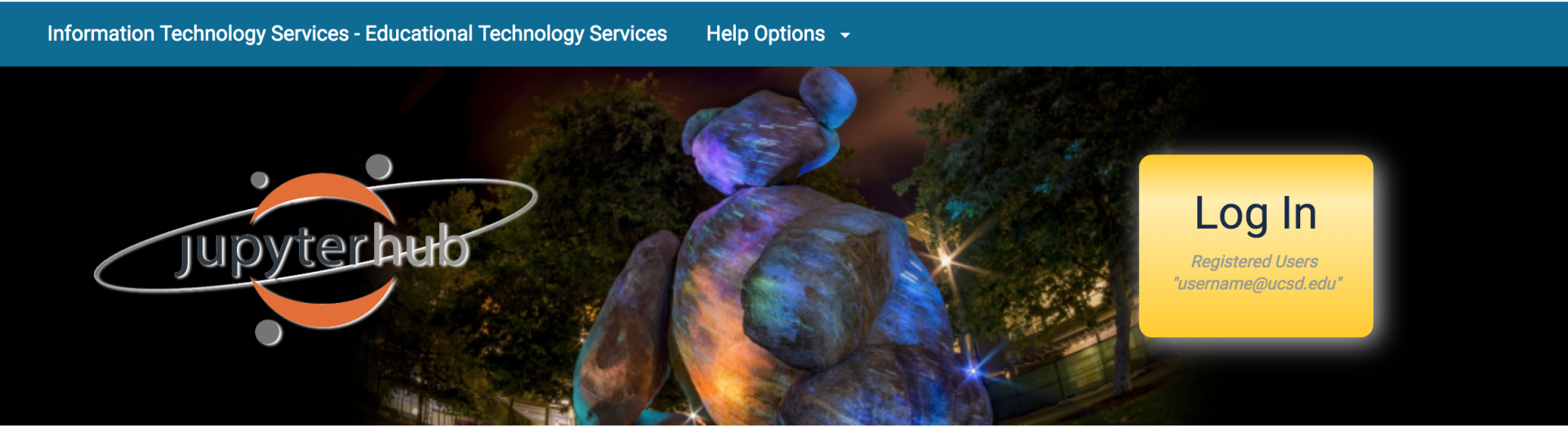
75% credit if submitted less than 5 days after deadline.

7 LATE DAYS allowed per person without penalty
to be used for Discussion Labs + Assignments

Assignment Submission @ Datahub: https://datahub.ucsd.edu

DATA SCIENCE / MACHINE LEARNING PLATFORM

UC San Diego



UC San Diego Jupyterhub (Data Science) Platform

Before next Mon: log onto datahub & have a working installation of Jupyter on your computer

Group Projects: the main focus of COGS 108

1) review a previous project, 2) make a proposal, 3) get feedback and fix things, 4) hit some checkpoints, 5) turn in a final project w/ video

Groups of 4-5 people, figure it out!

How to find a group:

1. go to discussion section, talk to people there
2. post on Looking for Teammates on EdStem
3. talk to people you are sitting near after class

Demographic information for research purposes

You are being invited to participate in a research study while you are attending COGS 108 Data Science in Practice. All students in the course are invited to participate in this study, which is being conducted by the course's instructor Dr. Jason Fleischer. **By filling out this optional section of the survey you are agreeing to participate in this study.**

The purpose of this research study is to investigate how different ways of assigning students to project groups affects student learning and satisfaction. The results of this study may help improve learning outcomes for future students in this course and others like it.

By filling out this section of the survey you agree to have the above data included in analyses that will appear in publication. Analysis of your data will be done by myself and members of my research lab. Any publications from this study will report only aggregated, deidentified data. Your personally identifiable information will never appear in public. We will avoid reporting data on demographic groups of < 3 students to avoid potentially revealing personal information through aggregate reporting. This data will be secured with access limited to a small team of researchers responsible for this study. The data will be deleted after 10 years. You may also contact us at any time to revoke your permission and have your data deleted.

Your participation in this study is completely voluntary and you can withdraw at any time. **Choosing not to participate or withdrawing will not affect your grade or relationships with the course instruction team.** If you do participate you are free to skip any question you do not wish to answer.

If you have questions about this project, have a research-related problem, or wish to revoke your permission to use your data, you may contact Jason Fleischer at jfleischer@ucsd.edu. If you have any questions concerning your rights as a research subject, you may contact the UC San Diego Office of IRB Administration at irb@ucsd.edu or 858-246-4777.

By participating in this research study you are indicating that you are at least 18 years old, have read this consent form, and agree to participate in this research study.

Confusion and struggling is expected and ITS OK!!!

This class is a mile wide and an inch deep…. you will need to teach yourself!

If something is unclear:

First try for a while to understand it yourself, or to educate yourself from the internet

Still struggling?

- *ask in class*
- *ask during section*
- *post on EdStem*
- *ask a classmate*
- *come to office hours*

Academic integrity

Don't cheat. Please review academic integrity policies here.

YYou will work together on projects. You should help one another learn in general. Assignments and discussion labs should be completed individually, although you may seek help from your fellow students. However you may not give answers to each other at any time. THERE IS NO ASKING QUESTIONS OF EACH OTHER ON QUIZZES AND EXAMS!!

Examples of good collaboration on assignments:

- Student posts non-working code to get help, others send a link to a good reference page in sklearn documentation, or point out the generic kind of mistake being made (e.g., you've messed up the order of operations). Nobody just writes the correct code for the student.
- Student posts a question about a theory or concept. If its not directly related to an assignment question you can choose to answer in full. However, it's generally more helpful for learning if you use the Socratic method: ask the student questions that lead them to find the answer themselves. Also doing this helps you cement your own knowledge of the subject!
- Student posts a question about an assignment problem. Others point out important principles that we have learned in class that can be used to solve it. They describe the important points, or mention important pitfalls to avoid. References to book pages, lecture slides, or lecture video times are helpful. Nobody posts the correct answer for the student.

For group projects, you will work together but every person in the group is expected to understand every aspect of the project. People may be asked to individually explain any aspect of the project and your grade may be reduced compared to the rest of the group if you are unable to do so. Projects may include ideas and code from other sources—but these other sources must be documented with clear attribution.

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-/B+). 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 Campuswire.

Cheating and plagiarism have been and will be strongly penalized.

Policy on getting help from other people or AIs

1. You can help someone learn, but be Socratic wherever possible
2. Straight up giving someone answers or code is cheating.
3. MORE HELP OK ———————————— LESS HELP OK

   | | |

   Projects Labs Assignments
4. If you get help please credit the software package, webpage, publication, person or AI that produced the code or writing you are reusing.
5. Don’t use stuff you don’t understand! You are responsible for being able to explain your solutions or you may lose points

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 considered 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, your TAs, or IAs. If you are uncomfortable doing so, the OPHD and/or CARE are wonderful resources on campus.

What COGS 108 logistics questions do you have?

Extra time? Then we can look at datahub

I’m excited to have
you all in COGS 108!