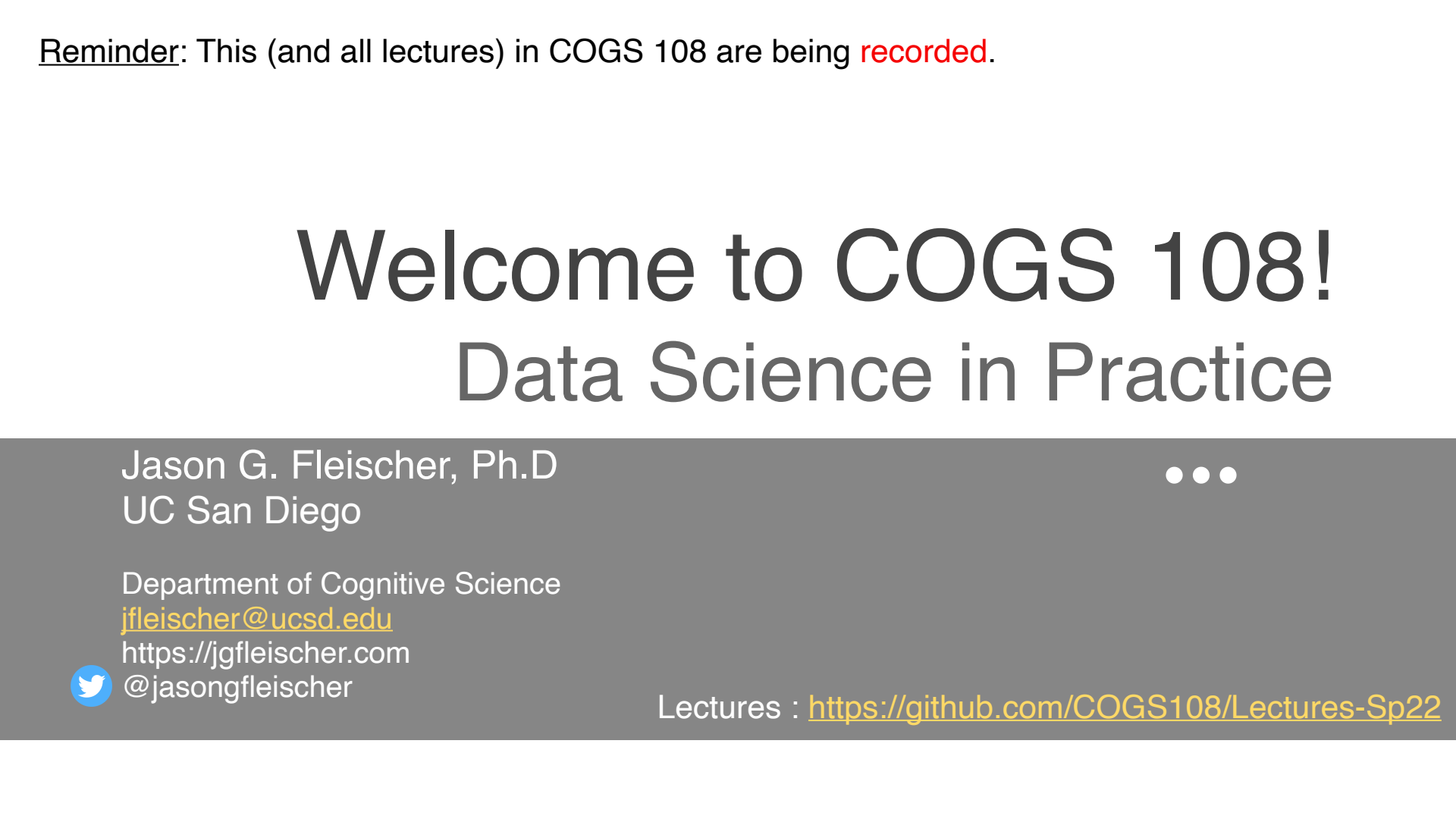
Reminder: This (and all lectures) in COGS 108 are being recorded.

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
@jasongfleischer

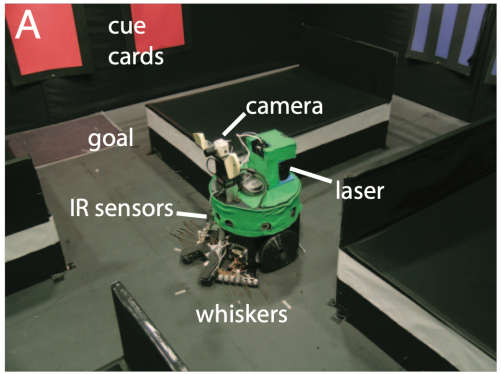


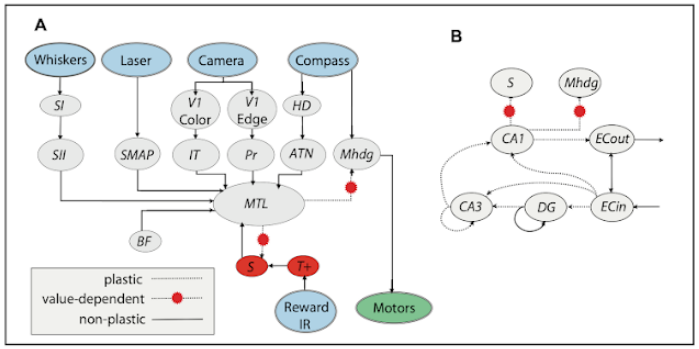
Lectures : https://github.com/COGS108/Lectures-Sp22


hello
my name is
Prof Fleischer

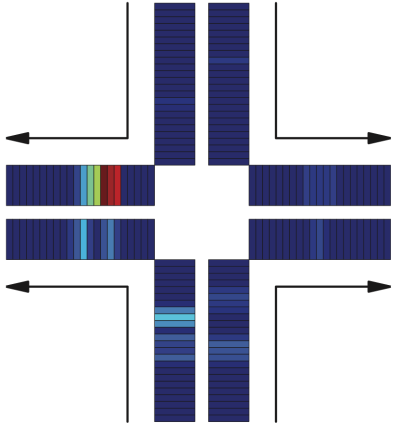
Wikimedia Commons
Sunlight on Colorado National Monument.jpg
By Meelmouse

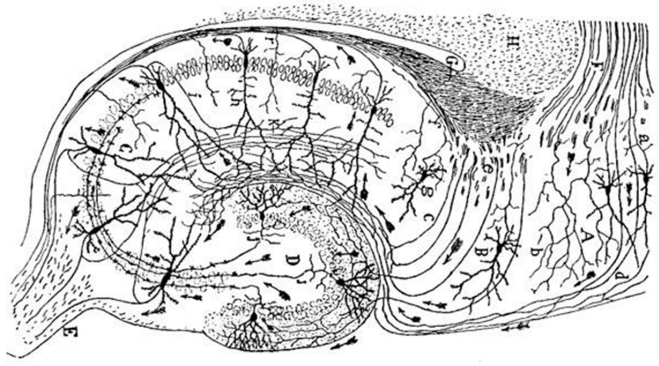

TO LET
43

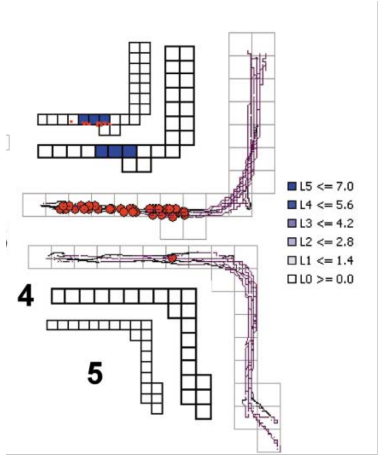


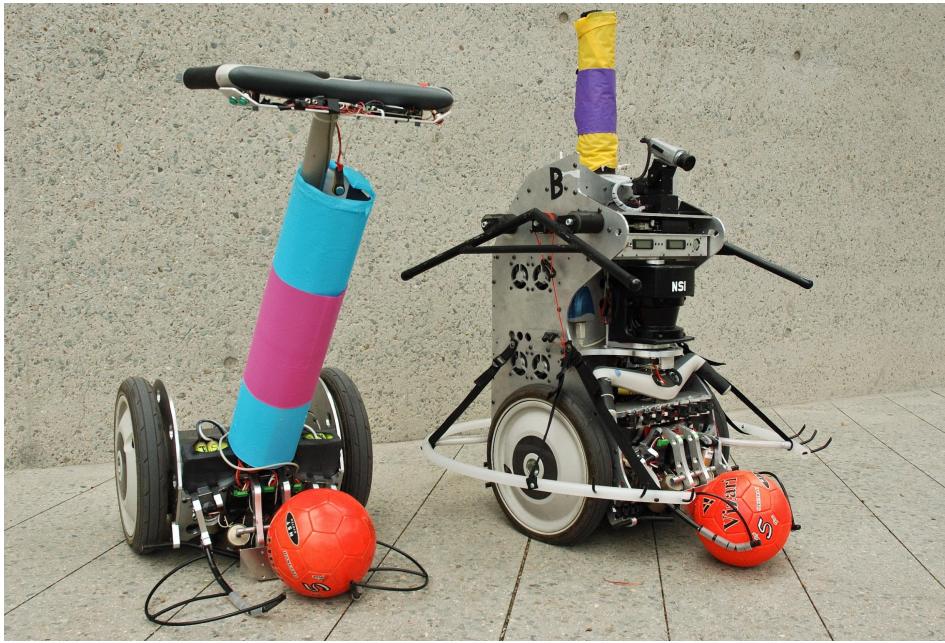


PNAS 104:3556-3561









RESEARCH ARTICLE | DEVELOPMENTAL NEUROSCIENCE

Sleepmore in Seattle: Later school start times are associated with more sleep and better performance in high school students

Gideon P. Dunster[1], Luciano de la Iglesia[1], Miriam Ben-Hamo[1], Claire Nave[1], Jason G. Fleischer[2], Satchidan…
+ See all authors and affiliations

Science Advances 12 Dec 2018:
Vol. 4, no. 12, eaau6200
DOI: 10.1126/sciadv.aau6200

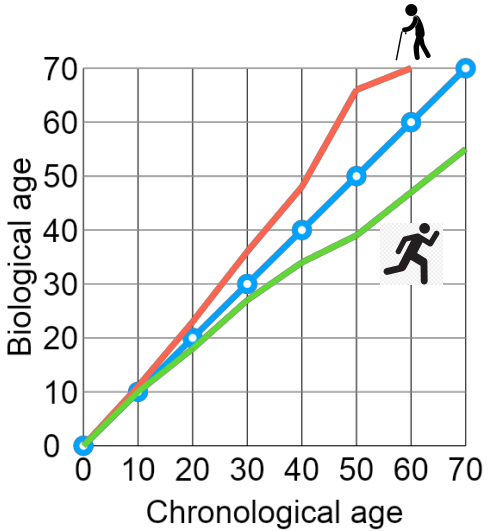
Abstract

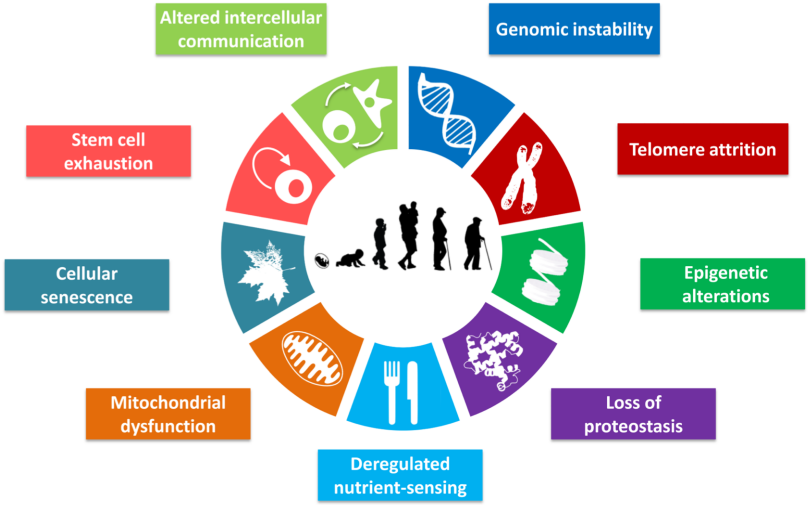
Most teenagers are chronically sleep deprived. One strategy proposed to lengthen adolescent sleep is to delay secondary school start times. This would allow students to wake up later without shifting their bedtime, which is biologically determined by the circadian clock, resulting in a net increase in sleep. So far, there is no objective quantitative data showing that a single intervention such as delaying the school start time significantly increases daily sleep. The Seattle School District delayed the secondary school start time by nearly an hour. We carried out a pre-/post-research study and show that there was an increase in the daily median sleep duration of 34 min, associated with a 4.5% increase in the median grades of the students and an improvement in attendance.

Cell Metabolism
Clinical and Translational Report

Ten-Hour Time-Restricted Eating Reduces Weight, Blood Pressure, and Atherogenic Lipids in Patients with Metabolic Syndrome

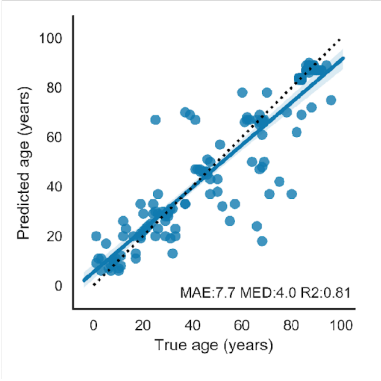
Michael J. Wilkinson,[1,3] Emily N.C. Manoogian,[2,3] Adena Zadourian,[1] Hannah Lo,[1] Savannah Fakhouri,[2] Azarin Shoghi,[2] Xinran Wang,[2] Jason G. Fleischer,[2] Saket Navlakha,[2] Satchidananda Panda,[2,4,*] and Pam R. Taub[1,*]



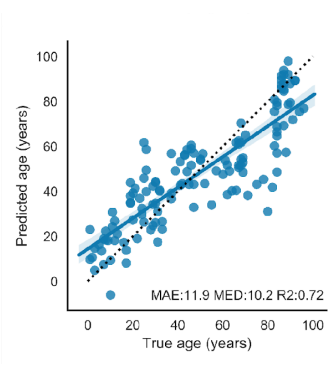


Hallmarks of Aging, López-Otín et al. Cell. 2013 Jun 6; 153(6): 1194–1217

Ensemble LDA



Support vector regression



3

FINAL

2


STONE
BREWING
STONE
BREWING

SAN DIEGO WAVE FC

Why this course?

You are going to be analyzing lots of data because you're studying to be a:

cognitive scientist
data scientist
computer scientist
neuroscientist, biologist, or chemist
social scientist (linguist?)
statistician or biostatistician
CEO/small business owner
something else really awesome

50 Best Jobs in America

Awards

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors:

	Job Title	Median Base Salary	Job Satisfaction	Job Openings
#1	Front End Engineer	$105,240	3.9/5	13,122
#2	Java Developer	$83,589	3.9/5	16,136
#3	Data Scientist	$107,801	4.0/5	6,542

Highest Paying Jobs

Oddball Interview Questions

Median Base Salary

Job Openings

View Jobs

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

Data scientist is actually MANY jobs

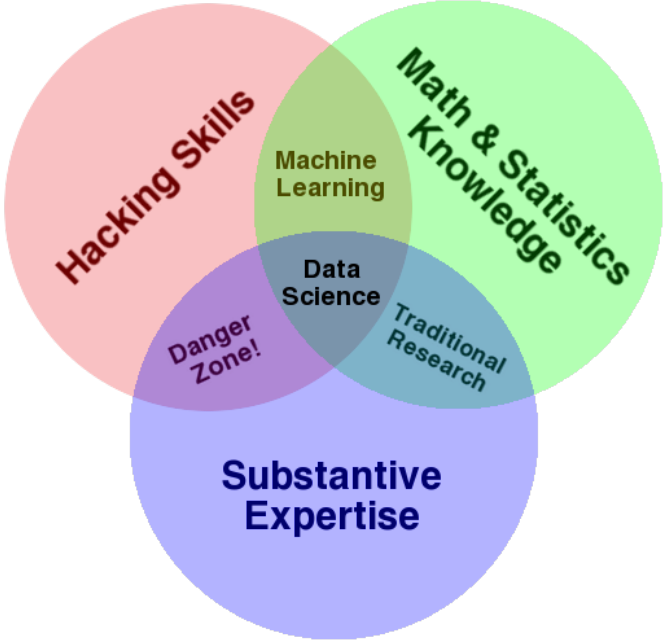
https://hbr.org/2018/11/the-kinds-of-data-scientist

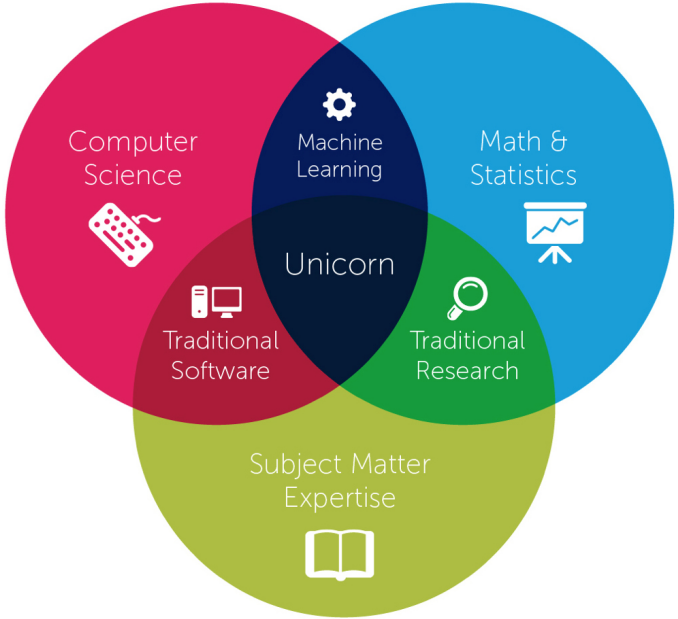
A final piece of advice for those hiring data scientists: Look for people who are in love with solving problems, not with specific solutions or methods, and for people who are incredibly collaborative. No matter what kind of data scientist you are hiring, to be successful they need to be able to work 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

What is data science?





http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data.[3] *It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.* -Wikipedia

"This coupling of scientific discovery and practice involves the collection, management, processing, analysis, visualization, and interpretation of vast amounts of heterogeneous data associated with a diverse array of scientific, translational, and interdisciplinary actions." -David Donoho ("*50 years of Data Science*

"an emerging discipline that draws upon knowledge in statistical methodology and computer science to create impactful predictions and insights for a wide range of traditional scholarly fields" - from a panel Rafael Irizarry moderated, shared on SimplyStatistics ("*The role of academia in data science education"*)

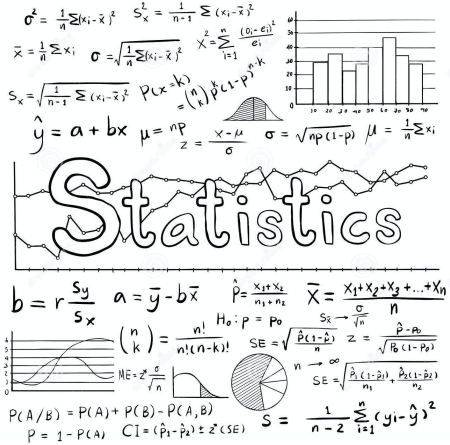
"an umbrella term used by organizations to describe the processes used to extract value from data" -Rafael Irizarry's personal definition in "*The role of academia in data science education"*

"The study of how the quantification of observable phenomena can lead to human understanding of the processes giving rise to those phenomena—or even the ability to predict future outcomes absent human understanding—and why certain phenomena require more or less data to lead to human understanding and/or prediction accuracy". -Brad Voytek's definition

"The scientific process of extracting value from data"

Data scientists ask interesting questions & answer them with data

The goal in COGS 108 is to *do* data science.



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

How we'll approach learning about *and doing* data science in COGS 108

Scheduling & Staff

Lecture: MWF 10-10:50pm
Discussion Sections: M, W, F
Office Hours: 11-12 MWF in person, 4-5pm MTWHF Zoom, ALWAYS BOOK on gcal!

TAs	IAs
Shivani	Xiqiang
Yueyan	Sizhe
Shanay	Suzy
Heeket	

COGS 108: General Plan

Week	Topic(s)
1	Data Science, Python, & Version Control
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

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 forget 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*

Course links

GitHub	https://github.com/COGS108	lecture/section materials & final projects
datahub	https://datahub.ucsd.edu	assignment submission
Piazza	**https://piazza.com/ucsd/spring2022/cogs108**	questions, discussion, and regrade requests
Canvas	https://canvas.ucsd.edu/courses/36591	grades, lecture videos
Anonymous Feedback	https://forms.gle/dXs8TSfDM86CLN3H8	if I ever offend you, use an example you hate, or to provide general feedback

General grading:

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

Attendance is neither required nor incentivized

- All lectures will be recorded (available end of day Canvas Media Gallery)
- One technical discussion section each week will be recorded

Weekly Lecture Quizzes:

- (9) weekly quizzes (first one due Friday of Week 2)
- 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 w/o a dr's note or similar

Why polling questions in COGS 108?

- There are a whole lot of you!
- Checks understanding
- Provides me with feedback
- Aids in critical thinking & allows for application of concepts
- Give you all a break from listening to me (we humans need this!)

(4 + 1 practice) Assignments

Assignments are completed individually and graded programmatically.

- These are meant to get you practice programming around the topics covered in class.
- The first two are much simpler than the following two and should take less time.
- You will have to look some stuff up on your own. This is by design.
- Instructions must be followed to receive credit.
- You'll have the opportunity to practice in discussion section.

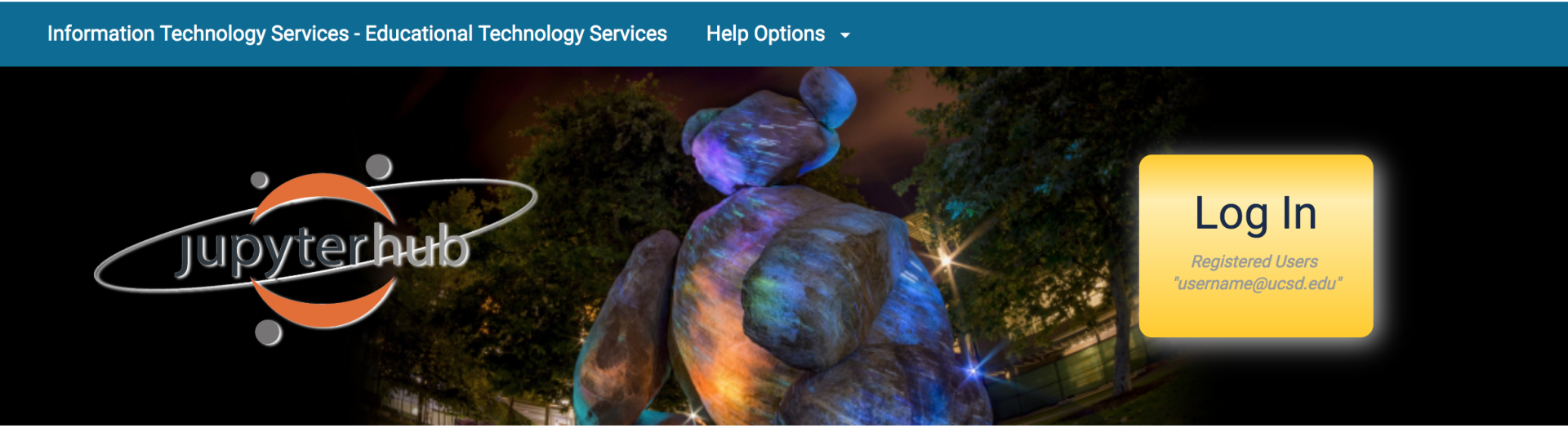
Assignments will be due on Fridays by 11:59 PM.

75% credit if submitted w/n 72h after deadline.

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

DATA SCIENCE / MACHINE LEARNING PLATFORM

UC San Diego



UC San Diego Jupyterhub (Data Science) Platform

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

Group Projects: the main focus of COGS 108

Groups of 4-5 Individuals

How to find a group:

1. go to discussion section week 1
2. post on group formation campuswire category
3. Use Zoom chat *at the end of class*

Discussion Section

- Goals:
  - help with technical aspects of the course
  - assignment & project help
- Technical Discussion Section
  - Rotates among the different lab sections
  - Labs exercise submitted by Fri @ 11:59 PM (2pt/lab)
- Project-Focused Discussion Sections
  - Each Project group will be assigned a staff member as their point of contact/ grader

Why is it like this? What about the section I'm assigned to?

- You'll never be required to go to section
- Have labs to help those struggling technically
- Technical section is always recorded
- Questions via Piazza if you can't attend

Discussion Sections start today! First week is just for group formation.

COURSE SCHEDULE

Date	Week	Day	Topic	Section	Assignment	Lecture Quiz
2/28	1	M	Welcome!	--	--	--
2/30	1	W	Python Review	--	--	--
4/01	1	F	Version Control I	--	Practice assignment	--
4/04	2	M	Version Control II	--	--	Q1
4/06	2	W	Data & Intuition	--	--	--
4/08	2	F	Data Wrangling (`pandas`)	D1	A1; Group signup*	--
4/11	3	M	Ethics	--	--	Q2
4/13	3	W	Data Science ?s	--	--	--
4/15	3	F	Dataviz I	D2	Project Review*	--
4/18	4	M	Intro to Analysis	--	--	Q3
4/20	4	W	Descriptive Analysis	--	--	--
4/22	4	F	EDA	D3	Project Proposal*	--
4/25	5	M	Inference I	--	--	Q4

Course Confusion

- If something in lecture, a section workbook, or an assignment is unclear:
  - *ask in class*
  - *ask during section*
  - *post on Piazza*
  - *ask a classmate*
  - *come to office hours*

Please do not use Canvas messages.

(The UI is the worst. I miss messages all the time. I will not look at them first. I look at Piazza first every day. Then email. Please use Piazza when possible.)

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

The (dreaded) waitlist

1. I know this matters to you and is a source of stress (and I hate that).
2. I have no control over the waitlist. If you have questions contact cogsadvising@ucsd.edu
    a. I know in other departments profs have control of this
    b. I quite literally do not have access to the system
3. A few people in each section typically get off the waitlist, but that number varies each quarter.
4. We have 417 enrolled with 175 on the waitlist at last look
5. We will likely admit up to about 430 or 440 total enrolled. So not everyone
6. Your wait list position is in your section. There are 7 sections. So if you're 6th on the waitlist of your section, you can expect there are up to 41 people in front of you
7. The waitlist settles after week 2.

What COGS 108 logistics questions do you have?

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