

Welcome to COGS 108!

Data Science in Practice

C. Alex Simpkins Jr., Ph.D
UC San Diego, RDPRobotics LLC



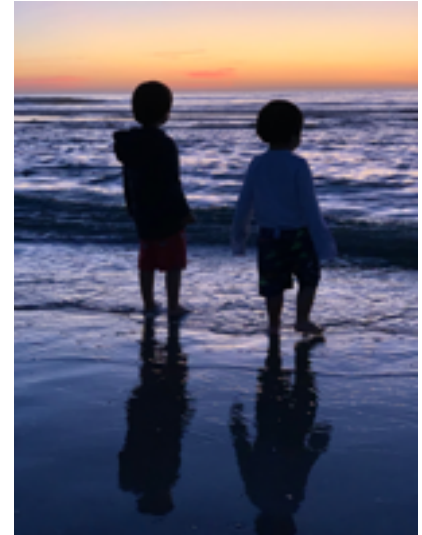
Department of Cognitive Science
rdrobotics@gmail.com
csimpkinsjr@ucsd.edu

Lectures : <https://github.com/COGS108/Lectures-Wi23>

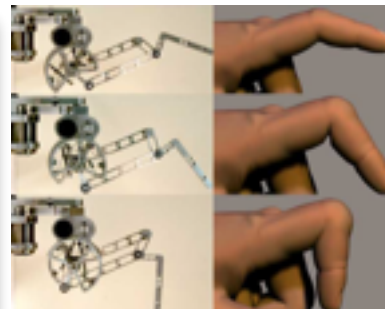
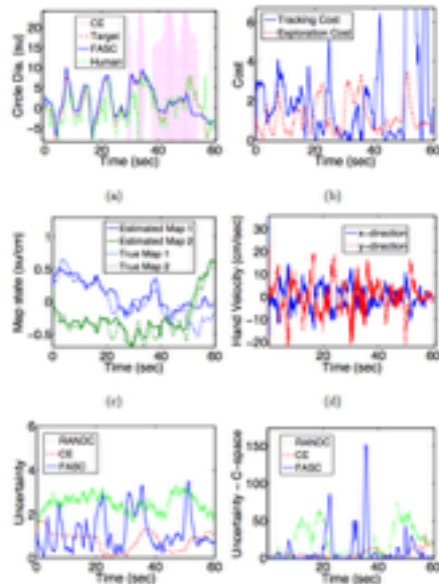


Who am I?

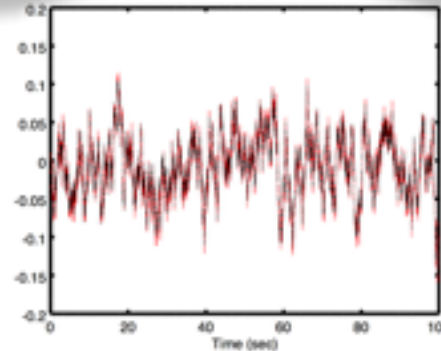
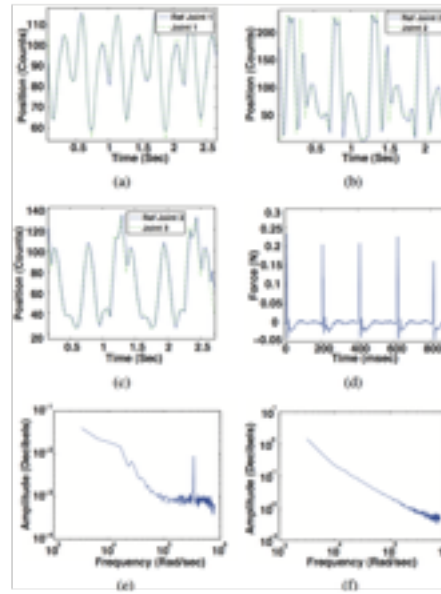
- C. Alex Simpkins Jr. Ph.D.
- BS/BS/MS/PhD UCSD Psyc, AMES, MAE, MAE, 2 postdocs UCSD Cogs and UW CSE
- Taught as TA ~20 times as a student, taught COGS109 as a grad student, taught at SDSU for a year in ME in Design
- Been involved in teaching for over 30 years teaching martial arts



- PhD in Control Theory (and Design)
- Robotics
- ML/System Identification
- Interdisciplinary



- Many industrial projects - robotics, AI, drones
- Consulting
- Entrepreneurial startup work
- Work with larger companies
- Research



What to expect in this class

- What is my role?
 - I am here to help you to learn and succeed, to open the door
 - NOT here to weed anybody out
 - NOT here to compete with you
 - Mutual respect

What to expect in this class II

- What is your role?
 - Learn! Open your mind
 - Put in the effort - you must walk through the door
 - Treat each other well, help each other to succeed (but do your own work of course)

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**
- **engineer**
- **something else really awesome**

Survey (link also on canvas)

COGS108 student survey (Winter 2023)

This survey is used to help me get to know you better! Thanks for your participation

Complete before Fri. week1 @ 11:59pm - opportunity for a little extra credit

If any data is used in class, the data will be anonymized. How you respond will not affect how you do in the class. Many are not required questions, please do not answer questions that make you uncomfortable

your email address will be recorded when you submit your form



**“Due” 11:59
PM Friday**

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

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.

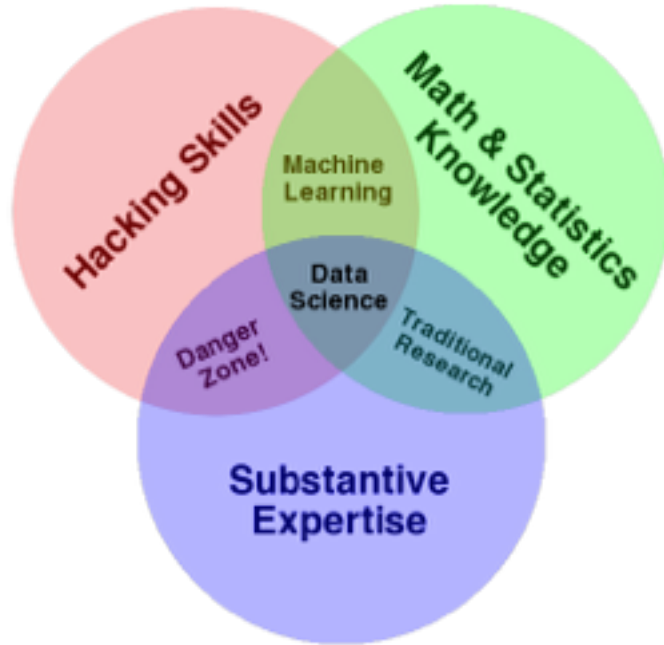


Data science for humans



Data science for computers

What is data science?



Copyright © 2014 by Steven Geringer Raleigh, NC.
Permission is granted to use, distribute, or modify this image,
provided that this copyright notice remains intact.

Defining Data Science

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: Tu-Th 9:30-10:50

Discussion Sections: M, W, F

Office Hours: TBA(Dr. Simpkins, by appt.);

IAs

TAs

Matthew Fain	mfain@ucsd.edu
Ruby Ying (Zhuojun)	z5ying@ucsd.edu
Aditya Jeswani	ajeswani@ucsd.edu
Tapan Sahni	tsahni@ucsd.edu

Xiaotong (Tony)	x6zeng@ucsd.edu
Robert Kovar	rkovar@ucsd.edu
Tian Gao	tgao@ucsd.edu
Valli Nachiappan	vnachiap@ucsd.edu
Zikang Chen	zic017@ucsd.edu
Zairan Xiang	zaxiang@ucsd.edu

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](https://www.codecademy.com) or [LearnPython.org](https://www.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/winter2023/ cogs108 (course code on canvas home page)	questions, discussion, and regrade requests
Canvas	https://canvas.ucsd.edu/courses/44466	grades, lecture videos
Anonymous Feedback	Submit via Google Form	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
(8/9) Discussion Labs (technical)	16
(4) Assignments	32
Final Group Project	44
(1) Project Planning Survey*	1
(1) Project Review*	5
(1) Project Proposal*	8
(2) Project Checkpoints*	10
(1) Final Report*	15
(1) Final Video*	3
(1) Project Survey	2

* indicates group

Attendance is neither required nor incentivized

- All lectures will be recorded (available by 2PM every TuTh; Canvas Media Gallery)
- One TBA 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
 - Timed : 15 minutes
 - Posted by Friday @ 11:59 PM (after each week of lecture); due the following Friday
 - Meant to test concepts from previous week's lecture

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

Lowest quiz score will be dropped.

All deadlines Fri at 11:59 PM

Week	Quiz	Discussion Lab	Lecture Quiz	Assignment
1	Data Science, Python, & Version Control			--
2	Data Intuition & Wrangling	D1	Q1	A1 - python, group proj. survey*
3	Data Ethics & Questions	D2	Q2	Project Review*
4	Data Visualization & Data Analysis	D3	Q3	Project Proposal*
5	Inference	D4	Q4	A2 - pandas/viz
6	Text Analysis	D5	Q5	Checkpoint #1: Data*
7	Machine Learning	D6	Q6	A3 - Inference
8	Nonparametric Analysis	D7	Q7	Checkpoint #2: EDA*
9	Geospatial Analysis	D8	Q8	A4 - NLP/ML
10	Data Science Communication & Jobs	D9	Q9	--

Final Project(Report*, Video*, Survey): due Tu March 23rd of finals week by 11:59 PM

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

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) 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 things up on your own. This is by design, and a skill to develop!
- 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](https://www.ucsd.edu)

Information Technology Services - Educational Technology Services

Help Options ▾



Log In

Registered Users
"username@ucsd.edu"

UC San Diego Jupyterhub (Data Science) Platform

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

All deadlines Fri at 11:59 PM

Week	Quiz	Discussion Lab	Lecture Quiz	Assignment
1	Data Science, Python, & Version Control	--	--	--
2	Data Intuition & Wrangling	D1	Q1	A1 - python, group proj. survey*
3	Data Ethics & Questions	D2	Q2	Project Review*
4	Data Visualization & Data Analysis	D3	Q3	Project Proposal*
5	Inference	D4	Q4	A2 - pandas/viz
6	Text Analysis	D5	Q5	Checkpoint #1: Data*
7	Machine Learning	D6	Q6	A3 - Inference
8	Nonparametric Analysis	D7	Q7	Checkpoint #2: EDA*
9	Geospatial Analysis	D8	Q8	A4 - NLP/ML
10	Data Science Communication & Jobs	D9	Q9	--

Final Project (Report*, Video*, Survey): due Wed June 9th of finals week by 11:59 PM

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

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 - Wed or Friday
2. post on group formation piazza thread
3. Use Zoom chat *at the end of class*

All deadlines Fri at 11:59 PM

Week	Quiz	Discussion Lab	Lecture Quiz	Assignment
1	Data Science, Python, & Version Control			--
2	Data Intuition & Wrangling	D1	Q1	A1 - python, group proj. survey*
3	Data Ethics & Questions	D2	Q2	Project Review*
4	Data Visualization & Data Analysis	D3	Q3	Project Proposal*
5	Inference	D4	Q4	A2 - pandas/viz
6	Text Analysis	D5	Q5	Checkpoint #1: Data*
7	Machine Learning	D6	Q6	A3 - Inference
8	Nonparametric Analysis	D7	Q7	Checkpoint #2: EDA*
9	Geospatial Analysis	D8	Q8	A4 - NLP/ML
10	Data Science Communication & Jobs	D9	Q9	--

Final Project(Report*, Video*, Survey): due Wed June 9th of finals week by 11:59 PM

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

Discussion Section

- Goals:
 - More opportunity for individual contact
 - Help with technical aspects of the course
 - Assignment & project help
 - Labs submitted by Fri @ 11:59 PM (2pt/lab; lowest lab dropped)
- Can I switch sections? Yes but stick to one for the quarter
 - Do lab exercises on your own if you are comfortable with material
 - Questions via piazza if can't attend
 - At least one section is always recorded

Discussion Sections will start week 2, week1 is just for group formation.

All deadlines Fri at 11:59 PM

Week	Quiz	Discussion Lab	Lecture Quiz	Assignment
1	Data Science, Python, & Version Control			--
2	Data Intuition & Wrangling	D1	Q1	A1 - python, group proj. survey*
3	Data Ethics & Questions	D2	Q2	Project Review*
4	Data Visualization & Data Analysis	D3	Q3	Project Proposal*
5	Inference	D4	Q4	A2 - pandas/viz
6	Text Analysis	D5	Q5	Checkpoint #1: Data*
7	Machine Learning	D6	Q6	A3 - Inference
8	Nonparametric Analysis	D7	Q7	Checkpoint #2: EDA*
9	Geospatial Analysis	D8	Q8	A4 - NLP/ML
10	Data Science Communication & Jobs	D9	Q9	--

Final Project(Report*, Video*, Survey): due Wed June 9th of finals week by 11:59 PM

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

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.

We are going to focus on piazza first then email. I have 800+ students. Please use piazza when possible.)

Clarification on communications

- I do my very best to be approachable as we all will
- I have ~800 students this quarter (not to mention two small children and a small company, research, and more going on)
- If I am every slow or brief in a response, it's probably that I have many communications to respond to, it is never to be taken as anything else

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 Dr. Simpkins, 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
 - 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.
 - a. I understand why when we're remote you'd expect me to let everyone in.
 - b. But, this is project-based. I already have 400 students in this class.
 - c. I tried letting everyone Fall quarter. I barely got any sleep.
4. The waitlist settles after week 2.
5. Our staff (cogsadvising@ucsd.edu) take care of this.

What COGS 108 logistics
questions do you have?

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