

Welcome to COGS 108!

Data Science in Practice

C. Alex Simpkins Jr., Ph.D
RDPRobotics LLC,
UC San Diego, Department of Cognitive Science
rdrobotics@gmail.com
csimpkinsjr@ucsd.edu



Lectures : http://casimpkinsjr.radiantdolphinpress.com/pages/cogs108_ss1_23/index.html

C. Alex Simpkins Ph.D.



But first I want to acknowledge our Instructional Team

TAs

Hari Yadavalli	hyadavalli@ucsd.edu
Rounak Sen	r2sen@ucsd.edu
Abhishek Tanpure	atanpure@ucsd.edu

IAs

Antara Sengupta	asengupt@ucsd.edu

Plan for today

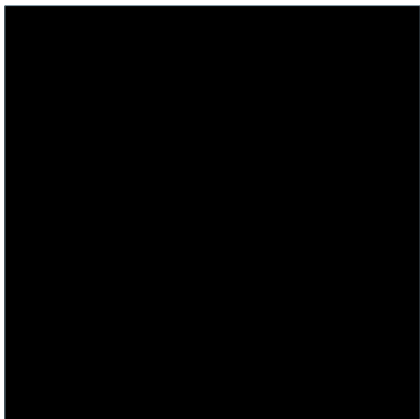
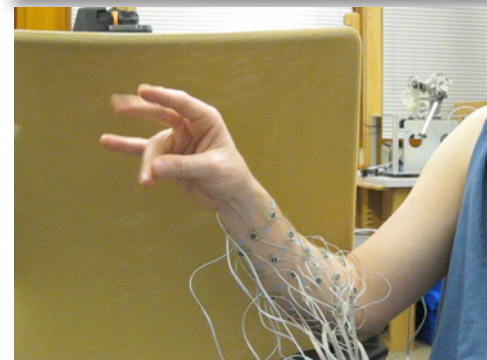
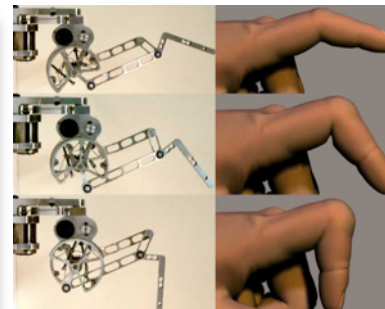
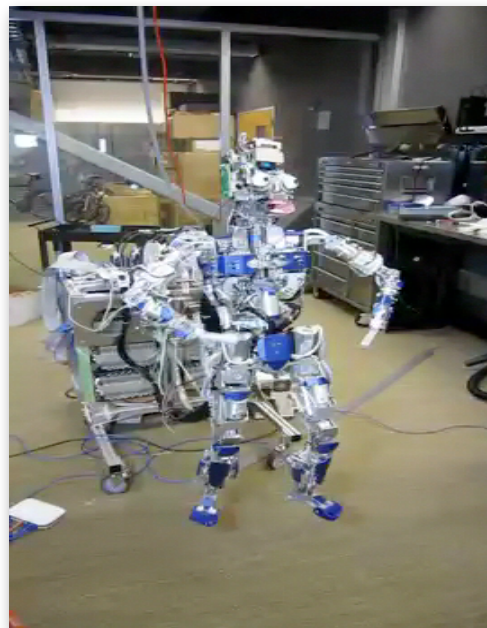
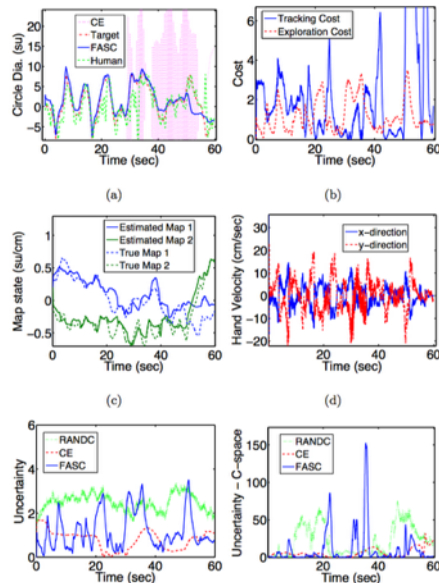
- Introductions
- About your instructor and what I'll try to share
- Motivation for the course
- Defining data science
- The structure and mechanics of the course, assignments, project, etc

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, came back to UCSD Winter quarter - COGS100, 108 and 138
- Been involved in teaching for over 30 years teaching martial arts

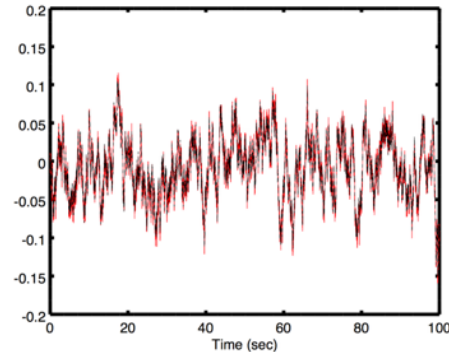
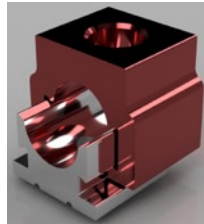
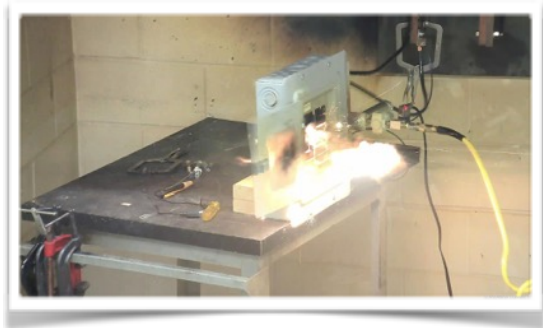


- PhD in Control Theory (and Design) is based on cybernetics
- Robotics



Who am I?

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



What to expect in this class

- What is my (our) role?
 - I am (we are) 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
 - Watch/attend lectures, do the readings, complete assignments and tests, and think about it all
 - 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 a great deal 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 on canvas and web page)


COGS108 student survey (SS1 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**

Why Data Science?

- Jobs
 - We need work, you can apply this in many ways and many fields
 - Big companies, small companies, startup/entrepreneurial, research, art, more!
- Powerful applications
- It's interesting and can benefit many people
- New challenges require new solutions!
 - We are in the era of Big Data, we have an opportunity to improve society and solve world problems that are not obvious otherwise (even the questions)

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.

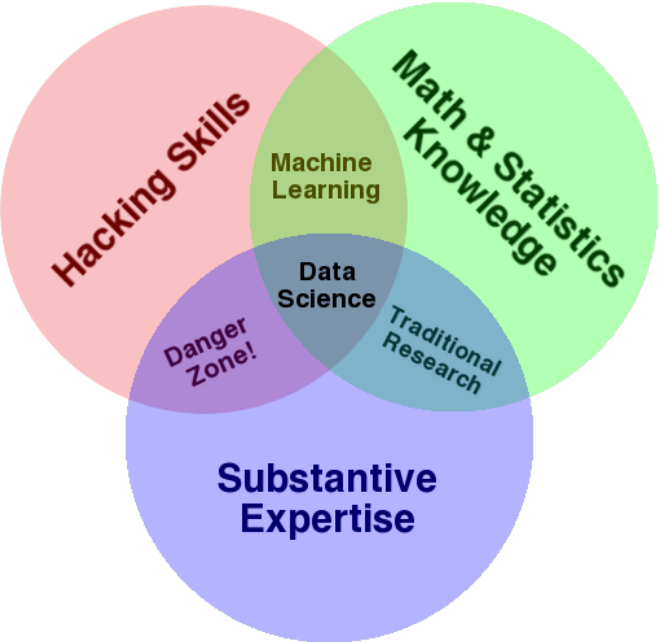


Data science for humans



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: M, W 11am-1:50pm in SOLIS 107

Discussion Sections: M, W 2-2:50pm in SOLIS 107

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

TAs

Hari Yadavalli	hyadavalli@ucsd.edu
Rounak Sen	r2sen@ucsd.edu
Abhishek Tanpure	atanpure@ucsd.edu

IAs

Antara Sengupta	asengupt@ucsd.edu

COGS 108: General Plan

Week	Topic(s)
1	Data Science & Version Control, Datahub, Jupyter, python I
1	Data Intuition & Wrangling
2	Data Ethics & Questions
2	Data Visualization & Data Analysis
3	Inference
3	Text Analysis
4	Machine Learning
4	Nonparametric Analysis
5	Geospatial Analysis
5	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

Web page	http:// casimpkinsjr.radiantdolphinpress.com/ pages/cogs108_ss1_23/index.html	Central hub: lecture/section materials, readings, links and jumping off point to the other resources
datahub	https://datahub.ucsd.edu	assignment submission
Piazza	https://piazza.com/ucsd/summer2022/ cogs108_s123_a00 (course code on canvas home page)	questions , discussion, and regrade requests
Canvas	https://canvas.ucsd.edu/courses/47460	grades, lecture videos, quizzes, mirrors of lectures
Course 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
(4/5) Weekly Quizzes (lecture content)	8
(8) Discussion Labs (technical)	16
(4) Assignments	32
Final Group Project	44
(1) Previous Project Review*	5
(1) Project Proposal*	9
(2) Project Checkpoints*	10
(1) Final Report*	15
(1) Final Video*	3
(1) Project Survey	2
EXTRA CREDIT (SONA, Surveys, 1? quiz)	10

* indicates group

Attendance is not required, but live interaction is strongly encouraged

- All lectures will be recorded (available by 2PM every MW; Canvas Media Gallery)
- The technical discussion section each MW will be recorded

Assignment deadline timing

- When possible (which is most of the time), we will give you a week for an assignment
- Summer session is a little compressed, so for some assignments we recommend you complete a week after an assignment
 - Soft deadline (week after assignment)
 - Hard deadline (Fridays) - for simplicity

Weekly Lecture Quizzes:

- (4-5) 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
 - Multiple Attempts (3)
 - ~10 Questions
 - Timed : 30 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.

If we get through all 5, one will be extra credit

All deadlines Fri at 11:59 PM

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

Final Project(Report*, Video*, Survey): due Fr week 5 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, subject to instructor/TA judgment

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

DATA SCIENCE / MACHINE LEARNING PLATFORM

[UC San Diego](https://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 We: 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	D1		--
1	Data Intuition & Wrangling	D2	Q1	A1-python, Group proj. survey*
2	Data Ethics & Questions	D3		Project Review*
2	Data Visualization & Data Analysis	D4	Q2	A2-pandas/viz, Project Proposal*
3	Inference	D5		
3	Text Analysis	D6	Q3	A3 - Inference, Checkpoint #1: Data*
4	Machine Learning	D7		
4	Nonparametric Analysis	D8	Q4	A4 - NLP/ML, Checkpoint #2: EDA*
5	Geospatial Analysis			
5	Data Science Communication & Jobs		Q5?	--

Final Project(Report*, Video*, Survey): due Fr week 5 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 today and Wed
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	D1		--
1	Data Intuition & Wrangling	D2	Q1	A1-python, Group proj. survey*
2	Data Ethics & Questions	D3		Project Review*
2	Data Visualization & Data Analysis	D4	Q2	A2-pandas/viz, Project Proposal*
3	Inference	D5		
3	Text Analysis	D6	Q3	A3 - Inference, Checkpoint #1: Data*
4	Machine Learning	D7		
4	Nonparametric Analysis	D8	Q4	A4 - NLP/ML, Checkpoint #2: EDA*
5	Geospatial Analysis			
5	Data Science Communication & Jobs		Q5?	--

Final Project(Report*, Video*, Survey): due Fr week 5 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? Nope there's only one!
 - Do lab exercises on your own if you are comfortable with material
 - Questions via piazza if can't attend
 - The section is always recorded

Discussion Sections will start today!

All deadlines Fri at 11:59 PM

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

Final Project(Report*, Video*, Survey): due Fr week 5 by 11:59 PM

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

Course Confusion and getting questions answered

- 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*

Canvas messages are less ideal - email or piazza first.

We are going to focus on piazza first then email.

Clarification on communications

- I do my very best to be approachable as we all will
- I have many 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 considerate and respectful. Always refrain from demeaning, discriminatory, or harassing behavior and speech, including to ANY of the teaching staff in class or online. Last of all, **take care of each other**.
- If you have a concern, please speak with Dr. Simpkins, your TAs, or IA. If you are uncomfortable doing so, the OPHD and/or CARE are excellent resources on campus.

The (dreaded) waitlist

1. I know this matters to you and is a source of stress (sorry about 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. We already increased the class size from 60 to 165.
 - a. I understand why when you can do it remotely you'd expect us to let everyone in.
 - b. But, this is project-based. I already have 165 students in this class.
 - c. Winter 2023 I had 800+ students and it was just at the end of the TA strike. I worked through spring break on grades and barely got any sleep.
 - d. I'll add as many as possible but we have to cap the class size relative to the number of TAs we have
4. The waitlist settles after week 1.
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!

(Let's take a 10min break, then back for part II)