COGS 9: Introduction to Data Science
QUARTER YEAR
DAYS TIME
BUILDING ROOM

Instructor: Bradley Voytek (bvoytek@ucsd.edu)
Teaching Assistants:
Office hours:
TA Office hours:
Final exam date:
Grading: Five assignments (12% each) + Class participation (10%) + Final project (30%)

Course Background: We are experiencing an explosion of it: 90% of all digital data didn’t exist two years ago. Researchers are leveraging this data deluge to uncover new insights into human behavior, intelligence and culture (sometimes with surprising findings). Companies are leveraging these data to recommend products to purchase, movies to watch, places to go, and things to do. What are the future implications for science, society, and industry?

Course Overview: In order to understand ***data science***, we first need to talk about ***data***: What counts as data and what doesn't? How do you visualize 1,000,000,000 Facebook friendships? How can you turn numbers on the screen into something meaningful? And how can data lead us astray?

In this class I will *introduce you to* the following topics:

- What are data, information, and knowledge? How are they related?
- How data science can inform or misinform
- Using Python for data science
- A gentle introduction to analysis methods: Statistics, inference, & modeling
- Applications of data science: Culturomics, social networks, text-mining
- Beyond data: Human-based computation, automated science, machine learning
- The art of data science: visualization and storytelling
- The role of data science in science and industry
- The ethics and implications of a data-driven society

Grades: There will be five assignments worth 12% each and a final project worth 30%. Class participation (showing up to lecture, participating in discussions during the lab sections) is worth 10%. Note there is no final exam, the project takes the place of the exam. Late assignments earn fractional credit (75% within one week late; 50% otherwise). A project won’t be accepted late unless it meets the extremely high standards and unusual circumstances that would be required to take a late final exam.

A rough guide to what is in each assignment:

1. Introduction to Python and handling data
2. Exploring data using descriptive statistics, and how not to get fooled
3. Visualizing data, and how not to get fooled
4. How to get fooled: p-hacking your way to the results you want
5. Turn in a draft of the final project, get back comments on it to move you in the right direction

Final project: a research report on how you would handle a complicated analysis from front to back… telling us all about the nitty gritty, whys, and hows of the analysis you choose. You'll write about the problems and issues with data handling and the analysis, and why you choose to overcome the problems in this particular way. If it's appropriate to the problem (e.g., hypothesis testing) you'll write about the expected results, but even if not you'll at least mention the different kinds of outcomes you might see. You WON'T have to actually perform the analysis… just write about it. But if you do make it that far, and can present results that's great and will be taken into that will in

Readings:

- Donoho D, *50 Years of Data Science*
- Tukey JW, *Exploratory Data Analysis*
- Buchanan M, *Depths of Learning*, *Nature Physics* 2015
- Krzywinski M & Cairo A, *Points of view: Storytelling*, *Nature Methods* 2013

Course reserves:

- Huff, D. How to Lie With Statistics, 1954
- Phillips, JL, How To Think About Statistics, 1999

Date	**Title**	**Due In Class**	**Topic**
April 4th	Hello world		Introduction
April 6th	What are data and information?		Data and information
April 11th	How data science informs cognitive science		Data and cognition
April 13th	1,000,000 books and 10,000,000,000 tweets		Culturomics and text mining
April 18th	“import antigravity”		Python
April 20th	Visualizing 1,000,000 Facebook friends		Data visualization and storytelling
April 25th	Making everyone *else* do your work	Assignment #1	Crowdsourcing and wisdom of the crowds
April 27th	Making your play work		Human-based computation
May 2nd	Analyzing stuff: a gentle introduction		Inference and model building
May 4th	Data: The end of theory?	Assignment #2	Hypothesis testing vs. data-driven
May 9th	Lies, Damned Lies, and Statistics		Data reproducability
May 11th	Data science: the boring parts		Data munging/ETL, mining, and dredging
May 16th	How telephones revolutionized neuroscience	Assignment #3	Communication theory & SNR
May 18th	Making computers do your work		Machine learning
May 23rd	Automating science, FASTER		Algorithms and automated science
May 25th	The lack of pirates is causing global warming	Assignment #4	Data errors - Correlation, overfitting, and multiple comparisons
May 30th	How being consistently better can be worse		Data errors - Simpson's paradox and Anscombe's quartet
June 1st	You're illiterate only if you're standing behind this line		Data errors - Ecological fallacy and MAUP
June 6th	Checking your work	Assignment #5	Cross-validation & bootstrapping
June 8th	Privacy in a data-driven world		Privacy and anonymization