COGS 9: Introduction to Data Science
Fall 2017, MWF 10:00-10:50a
Pepper Canyon Hall 106

Instructor: Bradley Voytek (bvoytek@ucsd.edu)
Teaching Assistants: Richard Gao; Isaac Shamie ({rigao; isshamie}@ucsd.edu)
Voytek's Office hours: Mondays, 11:00a-12:00p or by appointment (CSB 169)
TA Office hours: TBA
Final exam date: NO FINAL EXAM, ONLY FINAL PROJECT DEADLINE
Grading: Four assignments (15% each) + Final project (30%) + Participation (10%)

Course Background: Who cares about data? We all should! 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 data science? Soon, we will move beyond targeted ads and product recommendations to profound transformations in business, science, and society.

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?

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

- Data and Information
- Python
- Data-mining
- Text-mining and analytics
- Communication theory
- Human-based computation
- Automated science
- Data visualization and storytelling

Grades: NOTE THIS IS STILL A WORK IN PROGRESS. DETAILS MAY CHANGE BUT, IF THEY DO, YOU WILL BE GIVEN PLENTY OF ADVNACE NOTICE.

There will be four assignments worth 15% each and a final project worth 30%. 10% of your grade is for class participation (attendance taken during guest lectures). Late assignments earn fractional credit (75% within one week late; 50% otherwise up until assignment answers have been posted after which no late credit can be earned).

A rough guide to what is in each assignment:

- Introduction to Python and handling data
- Exploring data using descriptive statistics, and how not to get fooled
- Visualizing data, and how not to get fooled

- How to get fooled: p-hacking your way to the results you want
- Turn in a draft of the final project, get back comments on it to move you in the right direction

Final project: The final project is 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 account.

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 Calendar:

Date	Title	Assignment Due
09-29	Hello world!	
10-02	What is Data Science?	
10-04	Data and information	
10-06	Python	
10-09	Culturomics and text-mining	
10-11	Geospatial Analysis	
10-13	*Tentative guest lecture*	
10-16	Data visualization	
10-18	**No class!**	
10-20	Data journalism	
10-23	*Tentative guest lecture*	
10-25	Probability and statistics	
10-27	Statistical inference	
10-30	Algorithms and computability	
11-01	Hypothesis-testing vs. exploratory data analysis	
11-03	*Tentative guest lecture*	
11-06	Inference errors	
11-08	Data extraction	
11-10	**Veterans Day - No class!**	
11-13	**No class!**	
11-15	Signals and noise	
11-17	Version control and reproducability	
11-20	*Tentative guest lecture*	
11-22	Machine learning	
11-24	Cross validation and bootstrapping	
11-27	**Thanksgiving - No class!**	
11-29	Databases	
12-01	*Tentative guest lecture*	
12-04	Wisdom of the crowds and crowdsourcing	
12-06	Privacy and ethics	
12-08	The future of Data Science	