COGS 108: Data Science in Practice
Spring 2017 MWF 11:00-11:50
Peterson Hall 108

Instructor: Bradley Voytek (bvoytek@ucsd.edu)
Teaching Assistants:
Thomas Donoghue; Harshita Mangal; Larry Muhlstein:
{tdonoghue; hmangal; lmuhlste} @ ucsd.edu
Voytek Office hours: Mondays 10:00-10:50 or by appointment (CSB 169)
TA Office hours:
TBD
Final exam: Saturday, June 10, 2017 11:00a **Note this may be subject to change!**
Grading: Six assignments (10% each) + Final project (30%) + Participation (10%)

Critical note for Spring 2017: *Note this is the very first offering of this course in the history of UCSD, and this is the first time I've tried teaching a class this way. This will be especially tricky given how large the expected enrollment is (up to 500 students!), which is significantly larger than my first trial run of COGS 9 back in Fall 2014 (24 students). Thus, you are all guinea pigs for teaching a class structured this way.*

Additionally, the number of assignments stated above (6) may end up being fewer as we tweak the structure of the course throughout the quarter. In such cases, the benefit will always *go to the students—we won't penalize you if we alter the course on you midway through the quarter.*

Ultimately what you learn (and don't learn) in this course will help me shape future versions of COGS 108 to be better. But for now, caveat emptor.

Course Background and Overview: Who cares about data? We all should! In COGS 9 (Intro to Data Science) you (may have) learned why data and data science are important. The goal of that class was to give you an appreciation for *what can be done with data and where data can even lead you astray*.

In this course we take the educational view that "sometimes the best way to learn something is by doing it," or, more importantly as author Neil Gaiman says, "sometimes the best way to learn something is by doing it wrong and looking at what you did."

In this course, we aim to teach you the joys and frustrations of the *practice* of data science. We won't have you dive deeply into the methods or proofs of machine learning, clustering, etc. *on purpose*. The reasons for that are many-fold:

1. You can take an entire class on pretty much each of the 25 topics we cover below if you want real details.
2. Those classes are taught by true experts in each of those domains.
3. My expertise is *not* machine learning, big data, etc. It is in knowledge discovery and data intuitions.

4. I take an open view to learning: data literacy is critical for modern society, and I don't believe learning these topics should be limited to only those who excel at math, computation, and so on.

That said, we *will* have you try and implement various methods. At times we will even ask you to implement techniques we explicitly *haven't* taught you yet, as there may be times in your data science career you'll be asked to do just that. We want you to build a technical toolkit as well as a skeptical mindset and "data intuition"—that nebulous sense that something in a dataset is "off".

Topics Covered:

1. Introduction
2. Why data analysis? (prediction and classification)
3. Python!
4. Data Science in Python (jupyter, pandas, numpy, scipy, scikit-learn, etc.)
5. Data gathering (How do you find and clean data?)
6. Data wrangling (JSON, CSV, XML, SQL, APIs)
7. Data cleaning
8. Data privacy and HIPAA (anonymization)
9. Basic data visualization
10. Data intuition and the "sniff test" (Fermi estimation)
11. Linear modeling
12. OLS (optimization)
13. Distributions and outliers
14. Distributions and outliers: CDF, PDFs
15. Multiple linear regression and collinearities
16. Model validation (bootstrapping, resampling, k-fold, leave-p-out, train/test)
17. Feature selection
18. Dimensionality reduction (PCA)
19. Clustering (knn and k-means)
20. Classification (SVM)
21. Interpretability (trees!)
22. Non-parametric statistics
23. NLP and text-mining (tf-idf, sentiment analysis)
24. Geospatial analysis
25. Unsupervised learning (dbscan)

Grades: There will be six assignments worth 10% 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).

Final Project: The final project *will be group based, even if you don't want to work with groups.* ***No exceptions.*** The reality of working in technical fields is that you will need to work with others. Thus you need to learn how to communicate with groups effectively, handle disparities in knowledge and skill sets, manage time, and organize.

This year's final project will be *extra special* as it will be judged by a guest star, so I strongly encourage you to put some serious thought and effort into it. You will be briefed well ahead of time.

Piazza Discussion Board: This term we will be using the Piazza website for class discussion. The system is highly catered to getting you help fast and efficiently from classmates, the TA, and myself. Rather than emailing questions to the teaching staff, I encourage you to post your questions on Piazza. Please assist your fellow classmates, but be sure to do so in a manner that is consistent with the academic integrity policy (discuss the issue, but do not write the code needed or the answer). In addition to encouraging peer teaching and learning, engagement on Piazza can be a deciding factor for borderline grading cases (e.g., B+ vs. A-) since active Piazza citizens can earn 1% extra credit.

Find our class piazza page at: https://piazza.com/ucsd/spring2017/cogs108/home

Readings: There are no reading materials to be purchased. We will provide all necessary materials online on TritonEd. You will be asked to read and occasionally watch video tutorials online before attending class or for further learning afterwards.

Additional Readings "for fun":

- 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

Sections & Assignments: This is the hands on section of the course. The *practice* of data science largely boils down to writing code, and so, although this is not a programming class *per se*, the majority of the section and assignments will center around using Python to do data science.

Not everything you are expected to do will be explicitly mapped out, step by step. Not only would that level of direction be prohibitively long to prepare, it would ultimately be dishonest and unhelpful—data science is ultimately about figuring things out. If clear step-by-step instructions could be written down that always worked, data science would be automated and this class would be moot. The human level of figuring things out when it is not entirely clear how to proceed is the actual practice here. You are going to have to be okay with trying things out that don't work, with getting stuck, with not quite knowing what is going on all the time. That's the job.

What we don't want you to do is get or feel totally lost. If you do find yourself wandering through a problem that you really have no idea about, 17 layers into stackoverflow, Jupyter running at a sluggish pace due to an uncountable number of browser tabs—pause, and ask someone for help. Most of the time a quick intervention from someone who can identify the problem, give you an overview as to what is happening, and suggest where to look can save you tons of time (and frustration).

As instructors, we are here to help you, but keep in mind that the # students >>> # instructors. Thus we encourage you to help each other. If you are unsure about something, ask your fellow students. Everyone has different levels of experience in different domains. This is how {data science, programming, science, research, industry} works—teams of people with different backgrounds and expertise sharing knowledge and ideas and working toward a common goal. The best way to learn something, or to check if you really know it, is to teach it. So if you do think you know something, offer help! Sections and office hours are meant to be collaborative experiences.

Ask questions on piazza, and come to office hours. Please be aware that we may not know the answer to every question, at least not right away. Data science / programming is broad and dynamic and no single person can be on top of everything. If nothing else, we will try and give you pointers on what might be happening, and whether it is likely to be tractable, or potentially very messy and worth circumventing. Knowing which problems you can solve, and which you can't (at least in the context of a given project) is an important skill.

The goal of this class is not to teach you (all of) data science - we cannot possibly do that in 10 weeks. The goal is to give you a meaningful introduction and hands-on experience of what data science is, to teach you how to continue to learn data science, and to give you the desire and skills needed to do so. There are an amazing amount of resources out there for this material. The difficulty, as a newcomer, is to figure out what it is you're looking for, and how to find it - that is, figuring out what you don't know and what you need to learn. Technical experts do not know all the answers, they just know where to find the answers. The goal of these sections and assignments is to show you what is available, so you know where to look if you want to keep going in data science.

Section Administrative Details:
If your question/request is general, please post on Piazza. If it is about grading, please contact us through the course e-mail (COGS108 {at} gmail.com).

Officially, we will be using, assignments will require, and we will support the use of:

- python3 with the anaconda platform
- Jupyter Notebooks
- git & Github (optionally using the SourceTree GUI)

The assignments must be completed using these tools. You are welcome to explore other tools as you explore these topics, and to use different tools for the project. Note however, that we offer no guarantees that we can help with other languages / modules / tools, etc.

Section Topics & Hands-On Materials:
Section topics will follow the topics outlined for the lectures. For each lecture topic outlined in the syllabus, there will be a Jupyter notebook covering that topic, with a basic demonstration. We may not cover all the section materials explicitly in section, we will prioritize covering material that is needed for the assignments - that is, sections will serve as hands-on sessions targeted at fulfilling the assignments.

The section materials are not intended, or written, as tutorials. Instead they are more like an index, or a map. For each topic, we aim to give a cursory overview and simple demonstration of what the topic under investigation is, and guide you to bigger and better resources to really dive into it. You are not expected to, and will not be able to, follow every link for every topic—pick the ones you are most interested in and/or the most helpful ones to get you unstuck for the assignments and/or project.

Section Attendance & Switching:
Section attendance is not strictly mandatory but is recommended. Sections will be used for hands-on tutorials and overviews of the assignments - they are the best way to get detailed instructions on how to do the assignments. You should make sure you are enrolled in a section that you can attend, and plan to attend that section each week. If something comes up, and you have a conflict, you may instead attend another section. Each section, in a given week, will cover the same material.

Assignments:
Assignments will be done in Jupyter Notebooks. They will be released on Github (https://github.com/COGS108) and submitted to TritonED.

Assignment Schedule:
The (tentative) schedule for assignments is as follows:

> **Name:**
> A1 - Set Up & Github A2 - Data Exploration A3 - Data Identification A4 - Machine Learning A5 - Project Outline A6 - Data Types
>
> **Due Date:**
> 11:59 pm, Sunday, April 16th (end of Week 2) 11:59 pm, Sunday, April 23rd (end of Week 3) 11:59 pm, Sunday, April 30th (end of Week 4) 11:59 pm, Sunday, May 14th (end of Week 6) 11:59 pm, Sunday, May 21st (end of Week 7) 11:59 pm, Sunday, June 2nd (end of Week 9)

Assignment Questions & Using Piazza:
Please use Piazza for all general questions - for example, if you find a question unclear, or are totally lost and need some direction. When answering questions on Piazza, you may answer with suggestions on what topics / ideas / lectures to look into, and/or vague pseudocode but do **not** provide code that answers assignment questions. For questions that are tangential or unrelated to answering assignments, you may post minimal code segments. If you have specific questions about your assignment (for example, you are missing a grade), email the class email (COGS108 {at} gmail.com)

Naming Conventions:
We will be tracking assignments and files using the first letter of your last name and the last four digits of your student ID. This is necessary due to the size of the class - we have many students with the same first and/or last name. You must follow the file name

conventions for assignment submission. Assignments are auto-graded and these scripts may fail with improperly named files. Triple check your submitted files meet the specified naming conventions. If you fail to do so, it is likely your submission won't make it through the grading procedure and you will not receive a grade.

Grades:
Grades will be released on TritonED a week after the submission date. We will try to send out automated alerts if we do not receive a submission and/or if it fails to be processed, but ultimately it is your responsibility to check your grades and get in touch if any are missing or you think there is a problem.

Assignment Regrades & Solutions:
After grading is complete, we will release the assignment solutions at the same time that we release the grades (1 week after the submission deadline). These solutions will be our solutions and the full test-suite that was used for grading. Note that there may be multiple possible solutions, and variations may receive full credit, provided they pass the test criteria that are specified in the question set up. If you think there is a mistake or ambiguity (for example, your different solution meets the question specifications, but fails on an unexpected test) email the course email (COGS108 {at} gmail.com), and we will look into it.

Late Submissions:
You may submit late assignments, up until assignment solutions are posted, but at a 50% penalty.