COGS 108 - Data Science in Practice
Capstone Project

Project Overview

The 108 Capstone Project will give you the chance to explore a topic of your choice and to expand your analytical skills. By working with real data of your choosing you can examine questions of particular interest to you.

The broad objectives for the project are to:

- Identify the problems and goals of a **real** situation and dataset.
- Choose an appropriate approach for formalizing and testing the problems and goals, and be able to articulate the reasoning for that selection.
- Implement your analysis choices on the dataset.
- Interpret the results of the analyses.
- Contextualize those results within a greater scientific and social context, acknowledging and addressing any potential issues related to privacy and ethics.
- Work effectively to manage a project as part of a team.

To accomplish this you will work in teams of 3 to 6 students to conceive of and carry out an analysis project.

*Note if you wish to participate in the special launch event for the UC San Diego Halicioglu Data Science Institute on Friday, March 02, please **carefully** read the details below in that section! This is **not** a requirement for the course, nor will it affect your grade; it is simply an extra bonus option should you wish to participate.*

Everyone must be part of a group. You will find, in your future careers, you will often need to work on projects in groups (even if you really, really, really, really don't want to).

The basic project steps:

- Find a real world dataset and problem that you believe can be solved with one or more of the techniques we have learned in class.
- After selecting a dataset and identifying the goal, write out a proposed analysis plan and submit it through TritonEd for review (due Sunday, Feb 18).
- Apply the techniques outlined and come up with a result for the dataset that you proposed.
- Assemble a Jupyter notebook that communicates your hypothesis, methods, and results (this is the final product due Thursday, March 22).

Each of the following sections goes into more depth on the components of the project.

Project Teams

It is up to you, the students, to form teams of 3-5 students. We strongly suggest that individuals consider their interests, skills, and schedule availability (including section enrollment) individually and to build teams accordingly. You can use Piazza to try and find potential teammates.

No changes to teams will be made after the Project Proposal is submitted at the end of week 6.

Getting Started

We strongly encourage you to discuss potential project ideas on Piazza, with your TAs and IAs, and with Prof. Voytek! This will give us a chance to make sure you're on the right track even before you submit your draft.

How to Find Datasets

The purpose of this project is to find a real-world problem and dataset that can be analyzed with the techniques learned in class. It is imperative that by doing so you believe extra information will be gained—that you believe you can discover something new!

You must use at least *one* dataset containing at least approximately 1000 observations (if your data are smaller but you feel they are sufficient, email Prof. Voytek). You are welcome (and in fact recommended) to find multiple datasets!

Your question could be just for fun: Using text mining of song lyric websites to identify the most commonly used phrases and sentiments by decade.

You question could be scientific: Scrape data from animal taxonomies and Wikipedia to figure out if larger animals are more likely to be carnivores?.

Or, ideally, your question can be aimed at civic or social good, for example, use mapping, transit, and car accident data to identify which parts of San Diego are most in need of dedicated bike lanes.

To help you find datasets, we have collected a list of websites that have a considerable number of open source data sets and included them at the end of this document. *(Big credit here to Jeremy Karnowski from Insight Data Science)*.

Eventually you will all have to decide on a problem to tackle, with each member of the team having a clear, delineated role in the project.

The Project Proposal

The Project Proposal is a document that does the following things:

> 1) It will present the background and context of your dataset and a description of the specific problem that your team has chosen to address using the data. In particular, you should describe the problem you have chosen and pose some interesting questions relevant to your problem that you would like to explore, as well as acknowledging and addressing any ethics & privacy related issues of your question(s), dataset(s), and/or analyses.
>
> 2) Identify the source(s) of the data you will *actually* be analyzing.
>
> 3) It will give a description of the data analysis techniques that you *intend to use* and how you intend to use them to answer the questions you posed. Your team does not have to have started any analyses at this stage, but you should be specific about the types of problems and goals—and therefore what techniques—you plan to use.

This proposal will be filled out as a Jupyter notebook, that will be uploaded to TritonED. The template and full instructions for doing so are available on the course Github at: https://github.com/COGS108/Projects.

If you think you will need any special resources or training outside what we have covered in COGS 108 to solve your problem, then your proposal should state these clearly. For example, if you have selected a problem that involves implementing multiple neural networks, please state this so we can make sure you know what you're doing and so we can point you to resources you will need to implement your project. *Note that you are not required to use outside methods*.

To reemphasize: for the Project Proposal *you are not expected to have already done any analyses for the proposed project, but what you submit should be a plan for what you will actually do for the project*. (Of course, for the final project you *will* need to actually do the analyses.)

Of the 35% of the course grade that is made up by your final project, 5% of this is directly from your project proposal.

Project Proposal - Detailed Description

For the Project Proposal you need to write a report, in the style outlined below, about how you might approach your question of interest. Specifically, every Report must contain seven sections, briefly outlined here, with more specific direction provided in the proposal template notebook:

1) Research Question: What's your question?
2) Hypothesis: What's your prediction?
3) Dataset(s): What data will you use to answer your question? Describe your dataset(s).
4) Background: Why is this question of interest, what background information led you to your hypothesis, and why is this important?
5) Proposed Methods: What methods will you use to analyze your data?
6) Ethics: Acknowledge and address any potential ethics and privacy issues related to your project.
7) Discussion: Discuss the potential impact of your project, as well as trying to anticipate any problems you may encounter.

The proposal should be written as if to a fellow student. You may assume that your audience is familiar with the material we have covered as a class this semester. The proposal is to be submitted electronically on TritonEd *as a group*. That is, one person from your group will submit a file *including the names and IDs of each group member*.

This is a short proposal meant to give us time to assess and criticize your Final Project (further described below), in order to give you time to *improve upon* it before your Final Project.

You will receive feedback on your project proposal, and you are fully expected to make the changes suggested by the Professor, TAs, IAs, and your classmates on this assignment before submitting your Final Project.

Remember to proofread your Project Proposal and do not using overly flowery and/or vague language.

Working on the Problem

Once you've settled on a problem and approach, it's time to actually analyze the data!

Note: It is very important that you get right to work on the problem and don't procrastinate. This is not a homework set—this is a large, complex problem that will take concerted effort to complete (and present effectively if you wish to compete in the judged event on Friday, March 02).

Final Project Details: Jupyter Notebook & Submission

The main product of the project is a single Jupyter Notebook. You can work on your project how you wish, but ultimately you will be graded on the one group notebook. After the Project Proposal submission, each group will be assigned a private repository on the COGS108 Github Organization. You can use this repository as you wish, but for the purposes of grading, each team will upload *one* Notebook to GitHub in the group-specific private repo. Your notebook should contain a complete walkthrough of your project.

This notebook should include all the code you used in your project for all components of the project (cleaning, visualization, analysis) that you wrote and used in your project. We will not be running the code in your notebook—make sure your notebook as uploaded to Github has the code evaluated and outputs present so that we can read the notebook as is. Each Notebook must contain a cell outlining who each member of the group is (including student ID), and what their contributions to the Final Project were. This notebook must be in your COGS108 Project Repository by the due date, and should be self-contained, so that we can evaluate your entire project from the notebook alone.

Your final project notebook must be present in your project groups Github repository, as of the due date: Thu, Mar 22 11:59p (23:59). Your group repo will be frozen at this time so no further changes will be allowed.

This file must have the filename:

FinalProject.ipynb

These notebooks may be opened to the general public, so others may read what you've done! You will have the option to opt out of making your project public.

Grading

The final project is worth 35% of your grade (as noted on the course syllabus). 5% of this is from your project proposal. The other 30% is based on your project notebook that will be uploaded on Github.

Your project will be graded based on the rubric below. Make sure you address each rubric section in the notebook, in an organized manner, using cell Markdowns for textual descriptions.

The grading rubric for the Final Project is as follows:

Category	Percentage of Project Grade
Introduction and Background	**10%**
Data Description	**10%**
Data Cleaning/Pre-processing	**10%**
Data Visualization	**15%**
Data Analysis and Results	**25%**
Privacy/Ethics Considerations	**15%**
Conclusions and Discussion	**15%**

UC San Diego Halicioglu Data Science Institute Launch Event

The *special objectives* for the *optional* UC San Diego Halicioglu Data Science Institute launch event are to:

- Communicate your results effectively to both experts and laypersons.
- Use data scientific approaches to address questions *specifically concerning civic utility and social good*.

A panel of local Data Science experts from the university, government, and industry will evaluate 4-8 projects, selected by Prof. Voytek for their potential for addressing critical questions of civic utility and/or social good.

These Projects *need not be the complete and final project you will submit for grading*, however they do need to be relatively thorough and complete to be considered for presentation on the afternoon of the launch event.

Deadline: To be considered eligible for presenting at this event, you will need to submit your Project Notebook by Sunday, Feb 25 at 23:59.

This *optional* submission, to be considered for the event, should follow the same outline and rubric as above for the final project notebook. You must have preliminary results, but it can be be a work-in-progress (for example, discussion section and conclusions need not necessarily be fleshed out).

One member from your team must submit this notebook on TritonED, with filename format (filled in with your group number):

'Pr_0XX_HDSIevent.ipynb'

Timeline

To make sure we are all progressing well toward the end of the project, use the following timing guidelines and deadlines:

Sunday, Feb 04: This document is released.

Week 5: Time in section is provided for project ideation and implementation.

Week 6: Project Proposal due Sunday, Feb 18 11:59p (23:59). You will pull the assignment from the COGS108/Projects GitHub repository and upload the .ipynb file to TritonEd.

Week 7: To be considered eligible for presenting at this event, you will need to submit your Project Notebook by Sunday, Feb 25 at 23:59. Groups who do not want to be considered for participation need not submit anything by this date. Remember if you do wish to participate here, your project should focus on questions of a civic/social good nature.

Week 8: 4-8 groups will present their Projects at the UC San Diego Halicioglu Data Science Institute launch event, on March 2nd.

Wed, Feb 28: The finalists chosen to participate in the launch event judging will be notified by this date.

Fri, Mar 02: HDSI launch!

Week 9: Time is provided to work on projects in sections.

Thu, Mar 22 11:59p (23:59): Due Date for all projects, for everyone (including HDSI launch event participants). Your group GitHub will be locked immediately after this time.

Resources and Advice

The main pieces of advice are:

- Start early
- Work consistently
- Be a good teammate
- Work as a team
- Seek advice when you are unsure, and see it early and often!
- Use Piazza
- Email your TAs, IAs, and Prof. Voytek; we're here to help!
- Choose a general interest domain, but then choose a dataset and decide on a problem, not vice versa. I promise it will go much better.
- Start early!!!!!

As far as resources go, it is okay to ask other teams what they are doing in terms of sources, presentation plans, and so on. As long as you are not using another team's work and claiming it as your own, collaboration with classmates is encouraged. If you find a good source of datasets, please share with everyone on Piazza!

Previous Class Projects (received perfect scores)

See Prof. Voytek's write up of excellent class projects from the Spring 2017 instantiation of COGS 108 here. A selection of these projects from previous iterations of the class are available here.

Example External Projects

Note these aren't civic/social good focused, but they are fun examples of what can be done with publicly available data.

- Visualizing The Hobbit
- Most Trendy Names in US History
- The Largest Vocabulary in Hip Hop
- A Map of Where NFL Quarterbacks Throw the Ball
- Every Shot Kobe Bryant Ever Took

Dataset Resource List

Below is a list of potential locations to find datasets and problems to investigate.
If you have another dataset or search location, that is great!

- Local San Diego Data Sets
- Data.gov
- Competitions | Kaggle
- Datasets « Deep Learning
- City of Chicago | Data Portal
- DataKind | Blog
- Code for America | Brigade
- Free Datasets - RDataMining.com: R and Data Mining
- 30 Places to Find Open Data on the Web
- 20 Free Big Data Sources Everyone Should Know
- Data Sources for Cool Data Science Projects
- UCSD behavioral mobile data