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 your formalizing and testing you problem 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.
- Work effectively to manage a project as part of a team.

To accomplish this you will work in teams of 3 to 5 students to conceive of and carry out an analysis project.

*Note if you wish to participate in the special judging panel on Saturday, June 10 with Dr. DJ Patil, please read carefully 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 (see the Methods list below).
- After selecting a dataset and identifying the goal, write out a proposed analysis plan and submit it through TritonEd for review (this is Assignment 4, due Friday, May 19).
- 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 Tuesday, June 13).

Each of the following sections will go 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 draft, Assignment 4, is turned in.

Getting Started

We strongly encourage you to discuss potential project ideas on Piazza, with your TAs, 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!

Your question could be just for fun (*e.g.*, "What are commonly misheard song lyrics?"), scientific (*e.g.*, "How do different cultures perceive different colors?"), or, ideally, aimed at civic or social good (*e.g.*, "What 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 Analysis Plan (Assignment 4)

The project analysis plan 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.

2) 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 should not actually complete the analyses at this stage, but you should be specific about the types of problems and goals—and therefore what techniques—you will utilize.

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 a multilayer perceptron, 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 Assignment 4 (the proposal) *you are not you are not expected to have already done any analyses for the proposal submission, but you submit should be a plan for what you do plan to actually do for the project*. (Of course, for the final project you *will* need to actually do the analyses.)

Specifically, for Assignment 4 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 *six* sections:

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 the dataset in terms of number of observations, number of features, etc. 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! If you do so, describe each one, and briefly explain how you will combine them together.
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) Discussion: What are the pitfalls and potential confounds of your methods? For example, how might biases in your data sources or analyses influence your interpretations? What will you do if you methods don't work and/or your hypothesis are wrong?

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 assignment, meant to give us time to assess and criticize your Final Project (further described below), to give you time to improve upon it before your Final Project.

You are fully expected to make the changes suggested by the Professor, TAs, and your classmates on this assignment before submitting your Final Project.

Remember to proofread your Essays 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 get to work analyzing the data! We may set aside time in class for working on projects and getting more feedback from Prof. Voytek, your peers, and the TAs.

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 competition on Sat, Jun 10).

Jupyter Notebook

The main product of the application project is a Jupyter Notebook. You can work on your project how you wish, but ultimately you will be graded on the one group notebook. Each team will upload *one* Notebook to GitHub in your group-specific private repo. Your notebook should contain a complete walkthrough of your project.

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 may be opened to the general public, so others may read what you've done!

DJ Patil and Data Science Expert Competition

The *special objectives* for the *optional judging* by Dr. DJ Patil and the panel of Data Science experts 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*.

Reminder, DJ Patil was the first US Chief Data Scientist appointed by President Obama. He was the Head of Data Products and Chief Scientist at LinkedIn, as well as a UCSD alumnus. He and a panel of local Data Science experts will evaluate 10 projects, selected by Prof. Voytek for their potential for addressing critical questions of civic utility and/or social good.

More details to follow as we sort them out, but this is a great opportunity to get feedback from Data Science experts, and it will be a useful experience overall.

Timeline

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

Friday, May 12: This document is released.

Week 7 (Monday, May 15 to Friday, May 20, in sections): Mixing time in section is provided for project ideation and implementation.

Friday, May 19 11:59p (23:59): Assignment 4 outline is due. You will pull the assignment from the COGS108/Assignments GitHub repo and upload the .ipynb file to TritonEd.

Week 9 (Tue, May 30 to Friday, Jun 2, in class and sections): Time is provided to work on projects in class and sections.

Monday, Jun 5: Submission deadline *only for those who wish to be considered for judging at the special DJ Patil panel.* 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.

Thur, Jun 8: The 10 finalists chosen to participate in the DJ Patil judging will be notified by this date.

Saturday, Jun 10: DJ Patil judgment day!

Tuesday, Jun 13 11:59p (23:59): Real Due Date for all projects, for everyone (including DJ Patil participants). Your group GitHub will be locked immediately after this time.

Grading

The capstone project is worth 30% of your grade according to the course syllabus. The project grading is broken down as follows:

The grading rubric for Assignment 4 is as follows:

Category	Percentage of Assignment Grade
File Naming and ID Inclusion	**5%**
Research Question	**10%**
Hypothesis	**10%**
Datasets	**15%**
Background	**20%**
Proposed Methods	**20%**
Discussion	**20%**

The rubric for the Final Project Notebook will be given later.

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 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!

Example 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
- 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
- Google Public Data
- 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