CS46N Project #2: Individual Data Discovery

Feedback meetings: November 5-13
Roundtable discussion: December 3
Final demonstrations: November 30 - December 11

The second project is far less prescriptive than the first one: you are to identify a domain you are interested in for which there is ample data available, and where analyzing that data may yield new insights or answer specific questions. Your goal is to identify the relevant data, formulate a set of questions or goals, then develop data manipulations, analyses, and/or visualizations that aid you in achieving your objective.

This project may be done independently or in pairs; either way, we expect it to be a substantial project on which you devote significant effort.

Components of the project

The project may be from a domain we have discussed in class, or any other data-rich domain that interests you. It is likely that some students will continue the initial work they did for the data visualization assignment, or pursue the topic of their student-led presentation. Your project could even be a more in-depth exploration of movie-rating predictions, if Project #1 was especially interesting for you. This project has four components:

1. Identify the domain you are interested in and the specific data set(s) you will use.
2. Formulate a specific set of questions you want to answer, points you want to make, or issues you wish to explore through the data. Be as concrete as possible.
3. Use techniques and tools such as (but not limited to) those covered in class to manipulate, analyze, and possibly visualize the data in order to achieve your objectives. It is likely you will end up developing a “data processing pipeline”, where in each step you transform or otherwise manipulate some or all of your data to get it into a form that’s suitable for the next step. In the final step your data should be in the best form to answer your questions or otherwise achieve your objectives. Often, but not always, the early steps are more about “cleaning” the data -- correcting mistakes, filling in missing values, creating consistent representations, mapping corresponding values -- while the later steps are more focused on analysis and summarization.
4. Create a writeup describing the above three components, and most importantly discussing the conclusions drawn from your data-driven study.

Timeline

- **November 5-13** (via sign-ups): Individual meetings with Jennifer & Akash to discuss ideas and plans
- **December 3 in class**: Roundtable discussion of all projects
- **November 30 - December 11** (via sign-ups): Individual project demonstrations to Jennifer & Akash