# Statistics for Social Data: Class Project

The goal of the class project is to get hands-on experience analyzing network or text data. In analyzing the data, you can choose to use methods we learn about in class, or you can use other methods you find relevant. You will complete the project in three phases: I. Proposal; II. Midterm Progress Report; and III. Final Presentation. You will get feedback after each phase, but you will only be graded on the final presentation. You can either do the project by yourself, or with a partner. Your partner does not need to be enrolled in the class.

## I. Proposal (Due 9/18)

Find a network or text dataset that interests you. This can either be a curated dataset from another source, or it can come from raw data that you process yourself. Describe the dataset:

- What objects do you have measurements for? What is the structure of these measurements? How many measurements are there?
- How was the dataset collected? Is there missing data?
- Is there a natural "population" associated with your dataset? What is this population?
- What aspect of the dataset or population are you interested in learning about? Why are you interested in this?
- How do you propose to learn about the aspect or features of interest?

Your initial proposal can be very informal; it should be about 1–2 pages.

## II. Midterm Progress Report (Due 11/6)

Perform some exploratory analysis on your dataset, possibly related to your proposed goals, but not necessarily. Report on what analysis you have tried, what seems to be working, and what is not working. Summarize what you have learned so far. This should be about 2–4 pages.

## III. Final Presentation (Due 12/4)

Give a 15 minute presentation about your project. In your presentation, you should describe: (a) the data; (b) what you hoped to learn from the data; (c) what analyses you performed; (d) what you discovered.

## Resources

If you're not sure where to start, one way forward is to find a research paper on a topic you're interested in, then reproduce the analysis from the paper using a different dataset. There are many network and text datasets available on the web. In particular, check the following links:

- Jindal and Liu's Amazon product review dataset: `http://liu.cs.uic.edu/download/data/`
- The Enron email corpus: `https://www.cs.cmu.edu/~enron/`
- Crowdflower sentiment data: `https://crowdflower.com/data-repository`
- Jure Leskovec's network datasets: `http://snap.stanford.edu/data/`
- The UCI Network Data Repository: `http://networkdata.ics.uci.edu/resources.php`
- The Yelp Academic datasets: `https://www.yelp.com/academic_dataset` and `http://www.yelp.com/dataset_challenge`.

Alternatively, you can collect your own dataset (beware: this is usually harder than you might think).