Economics 1660: Big Data Prof. Daniel Björkegren

Final Project

The final project will give you a chance to apply the skills you've learned in this class to a topic of interest. This is a course on big data, so what qualifies as a project? We don't measure 'big data' by the number of observations, but more by the richness of behavior that it captures.

Your question should be of economic interest (not simply predicting outcome Y with data X, unless your application is novel and important), and should tell us something about human behavior. We encourage projects that integrate theory and data.

We want you to push the boundaries, to think about a problem in a new way. Hopefully this course has sparked some ideas. You may use methods introduced in the course or learned elsewhere.

You are welcome to work in a group of up to 3. If you would like help finding group mates please contact Prof. Björkegren.

Schedule:

- April 4: Project proposal, turn in with PS7 (short, ~2 pages)
  Get feedback from teaching staff on project direction.
- April 26: In class presentations (short)
  Discuss project with classmates and get feedback.
- May 10: Written final project due on Canvas
  Share your idea in all its brilliance.

Your project can take one of several forms:[1]

1. **New research paper**
   Identify a topic of interest, get a relevant dataset, and perform analysis to answer one or more research questions. There are many places to find data; you can start with the list of ideas at the end of this document.

[1] Thanks to Jon Bittner for ideas on these project forms.

2. **Replicate a research paper**
   Identify a research paper of interest where the underlying data (or similar data) could be used to replicate the analysis. Follow the steps in the original paper, and compute the main graphs and tables selected by the authors. If you find the same results, great! If you find different results, determine why (for example, you may have made different assumptions, results may be different in different settings, or you or the author may have made a mistake).

   Then, either:

   - **Advance on the paper**
     You can test something new that was not tested in the original paper, explore a smaller conclusion, or try the method in a different setting.

   - **Develop a business plan for turning the method into a service**
     Cost out the expenses that would be associated turning the data collection method, or algorithm into a service. Would the value generated by providing this as a service justify these expenses?
     Come up with a strategy to finance these expenses. Analyze whether this service is, or could potentially become financially viable. If not, estimate how much value is being neglected.

We expect most students will do projects of the first two forms. If you are interested in this third type of project, please check in with Professor Björkegren.

3. **Preparation of a novel dataset**
   While some data is public, a lot of data that has been collected is not. Data that is owned by private entities is likely to stay private, but many public entities have obligations to provide data to researchers, with some exceptions. For example, such provisions are outlined by the U.S. Federal Freedom of Information Act and similar laws elsewhere (NY State FOIL, India Right to Information).

   Using these laws, independent researchers have identified and made public valuable data that helps us understand our world. (For example, Chris Whong has made available rich data on NYC Taxi trips obtained through a FOIL request: http://chriswhong.com/open-data/foil_nyc_taxi/ which has been used in several research projects)

   Your task here is to identify data that would be useful for the world that could feasibly be obtained with an information act request, write up a proposal for using that data, and then request that data.

Data Set Sources
Datahub: https://datahub.io
UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/

City Open Data Census: http://us-city.census.okfn.org/
Police Open Data Census: https://codeforamerica.github.io/PoliceOpenDataCensus/

Kiva: http://build.kiva.org/
Yelp: http://www.yelp.com/dataset_challenge
Wikipedia Clickstream: https://datahub.io/dataset/wikipedia-clickstream/resource/be85cc68-d1e6-4134-804a-fd36b94dbb82

Million Song Dataset: http://labrosa.ee.columbia.edu/millionsong/

NYC Taxi pickups: http://chriswhong.com/open-data/foil_nyc_taxi/