Analyzing Big Data

Christopher Fariss (cjf20@psu.edu; cjf0006@gmail.com)

The 48th Essex Summer School in Social Science Data Analysis
July, 27, 2015 - August, 7 2015.
Office: TBD
Office Hours: TBD

Introduction to the Course

Course Content

This course focuses on the research design and analysis tools used to explore and understand "big data". The fundamentals of research design are the same throughout the social sciences, however the topical focus of this class is on computationally intensive data generating processes and the research designs used to understand and manipulate such data at scale. By massive or large scale, I mean that there are lots of subjects/connections/units/rows in the data (e.g., social network data like the kind available from Facebook or twitter), or there are lots of variables/items/columns in the data (e.g., text data with many thousands of columns that represent the words in the document corpus), or the selected analytical tool is a computationally complex algorithm (e.g., a Bayesian simulation for modeling a latent variable or a random forest model for exploratory data analysis), or finally some combination of these three issues. The course will provide students with the tools to design observational studies and experimental interventions into large and unstructured data sets at increasingly massive scales and at different degrees of computational complexity.

Course Objectives

Students will lean how to design studies to take advantage of the wealth of information contained in new massive scale online datasets such as data available from Facebook, twitter, and many newly digitized document corpuses now available online. The focus of the course is on designing studies in such away as to maximize the validity of inferences obtained from these complex datasets.

Course Prerequisites

Students should have some familiarity with concepts from research design and statistics. Generally, exposure to these concepts occurs during the first year course at a typical PhD program in political science. Students should have at least some exposure to the R computing environment. The more familiarity with R the better.

Course Details

- I will begin each class day with a short lecture over the class material (approximately 45-60 minutes).
- After each lecture, students will discuss one or two articles as they relate to the lecture (approximately 30-45 minutes).
- On the first day of class, I will introduce students to two large scale datasets. Students will use these data for applied examples over the 10 days of the course.
- The remaining portion of class (approximately 1.5-2 hours) will be devoted to hands on learning with R, simulated data, and the large scale datasets provided by the instructor. Day 7, and Day 9 will consist entirely of in class lab.
- The course schedule section, which is below, provides even more details about the topic of the lecture for each class day, citations for the discussion readings, and chapter entries from the text books for the lab portions of the class.

Reading and Programming Requirements

Required Programs

1. R: http://cran.r-project.org/
2. STAN: http://mc-stan.org/

Required Reading Material

1. Matloff, Norman. 2011. *Art of R Programming: A Tour of Statistical Software Design*. no starch press.
2. Additional required articles and chapters are listed below. Copies of these readings will be provided by the instructor.

Suggested Reading and Reference Material

1. Carpenter, Bob, Andrew Gelman, Matt Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Mrubaker, Jiqiang Guo, Peter Li, and Allen Riddell. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*
2. Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
3. Stan Development Team. 2015. "Stan Modeling Language: Users Guide and Reference Manual. Version 2.6.0." http://mc-stan.org/manual.html

Course Schedule

Day 1: Methods of Observation and Inference

Introduction to Exploratory Data Analysis, Visualization, and Validation

- Adcock, Robert, and David Collier. 2001. ”Measurement Validity: A Shared Standard for Qualitative and Quantitative Research.” *American Political Science Review* 95(3):529–546.
- Lazer, David, Alex (Sandy) Pentland, Lada Adamic, Sinan Aral, Albert-Lszl Barabsi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James H. Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, Marshall Van Alstyne 2009. “Computational Social Science.” *Science* 323(5919): 721-723.
- Shmueli, Galit. 2010. “To Explain or to Predict?” *Statistical Science* 25(3):289-310.

Note: Time permitting, we will get started with the lab material scheduled for day 2.

Day 2: Programming “Big Data”

Introduction to Parallel Programming in R for “Big Data”.

- Lecture and lab material are drawn from the chapters in *Art of R Programming* text book.

Day 3: Experimental Design

Designing and Implementing Randomized Manipulations of Large Scale Data Generating Processes

- Bond, Robert M., Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle, James H. Fowler. 2012. “A 61-Million-Person Experiment in Social Influence and Political Mobilization.” *Nature* 489(7415):295-298.

Day 4: Quasi-Experimental Designs

Exploiting Exogenous Shocks to Large Scale Data Generating Processes to Improve the Validity of Inferences

- Settle, Jaime E., Robert M. Bond, Lorenzo Coviello, Christopher J. Fariss, James H. Fowler, Jason J. Jones. 2015. “From Posting to Voting: The Effects of Political Competition on Online Political Engagement” *Political Science Research and Methods* (Forthcoming).

Day 5: Measurement Theory: Data, Validity, and Reliability

Construct Validity and Models for Reducing High Dimensional Data for Visualization and Analysis

- Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76-91.

Day 6: Ethical Responsibilities for the Social Data Scientist

Issues Relating to Transparency, Anonymity, Replication, and Reproduction in the Analysis of "Big Data"

- Driscoll, Jesse. "Prison States & Games of Chicken" working paper.
- Fariss, Christopher J. and Zachary M. Jones. "Enhancing Validity in Observational Settings When Replication is Not Possible" working paper.
- Jones, Jason J., Robert M. Bond, Christopher J. Fariss, Jaime E. Settle, Adam D. I. Kramer, Cameron Marlow, and James H. Fowler. 2013. "Yahtzee: An Anonymized Group Level Matching Procedure" *PLoS ONE* 8(2):e55760.

Day 7: Text as Data Part 1

Introduction to Text as Data

- Roberts, Margaret E, Brandon Stewart, and Dustin Tingley. "Navigating the Local Modes of Big Data: The Case of Topic Models." In *Data Analytics in Social Science, Government, and Industry*, New York: Cambridge University Press.

Day 8: Social Network Analysis Part 1

Introduction to Social Network Data and Analysis

- Christakis, Nicholas A. and James H. Fowler. 2012. "Social contagion theory: examining dynamic social networks and human behavior." *Statistics in Medicine* 32(4):556-577.

Day 9: Text as Data Part 2

Text as Data Lab

Day 10: Social Network Analysis Part 2

Social Network Data and Analysis Lab

Next Steps

In this course, we have focused exclusively on the research design tools and analysis techniques necessary for working with and understanding "big data". We have not considered tools designed to gather and maintain large scale datasets. R is a useful tool for some of these tasks but other more powerful tools exist. Students should consider devoting additional energy to learning tools such as the python, json, and java programming languages and other tools for database management such as hive, hadoop, and SQL (Structured Query Language).

Recommended Programs

1. Python: https://www.python.org/
2. Java: https://www.oracle.com/java/index.html
3. Json: http://json.org/
4. Hive: https://hive.apache.org/
5. Hadoop: https://hadoop.apache.org/

Biographical Details

I am currently the Jeffrey L. Hyde and Sharon D. Hyde and Political Science Board of Visitors Early Career Professor in Political Science and Assistant Professor in the Department of Political Science at Penn State University. To date, I have taught courses on research design, measurement, and human rights. For my research, I use computational methods to understand why governments around the world choose to torture, maim, and kill individuals within their jurisdiction. Other projects cover a broad array of themes, ranging from foreign aid to American voting behavior, but share a focus on computationally intensive methods and research design. These methodological tools, essential for analyzing ”big data”, open up new insights into the micro-foundations of state repression.

Syllabus Acknowledgments

This syllabus is based on several courses that I have taken and designed over the last several years. Some of the material is based on the Research Design (PL SC 501) course that I developed at Pennsylvania State University when I began teaching there in the fall of 2013, which itself is based on similar course developed by David Lake and Mathew McCubbins at the University of California, San Diego. It is also based on material that I developed for a graduate measurement theory class (PL SC 597) and undergraduate Social Data Analysis and Design class (SO DA 308) that I also developed at Pennsylvania State University. Elements of the syllabus and other class materials created for this class are also based in part on the Bayesian Statistics class offered by Seth Hill at University of California, San Diego and the Measurement class offered by Keith Poole at UCSD and now the University of Georgia.