Exploration and Analysis of Social Media Data

Christopher Fariss (cjf0006@gmail.com)

The 50th Essex Summer School in Social Science Data Analysis
July, 24, 2017 - August, 4 2017.

Introduction to the Course

Course Content

This course focuses on the research design and analysis tools used to explore and understand social media and "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 Readings

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 Readings and Reference Material

1. Bolker, Ben. 2007. Ecological Models and Data in R. Princeton NJ: Princeton University Press.
2. 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*
3. Dunning, Thad. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press.
4. Google's R Style Guide: https://google.github.io/styleguide/Rguide.xml
5. Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
6. Lin, Winston, Donald P. Green, and Alexander Coppock. "Standard operating procedures for Don Greens lab at Columbia." Version 1.05: June 7, 2016. https://github.com/acoppock/Green-Lab-SOP
7. Stan Development Team. 2015. "Stan Modeling Language: Users Guide and Reference Manual. Version 2.6.0." http://mc-stan.org/manual.html
8. Teetor, Paul. 2011. *R Cookbook* O'Reily. https://ase.tufts.edu/bugs/guide/assets/R%20Cookbook.pdf
9. Trochim, William and James P. Donnelly. 2007.*The Research Methods Knowledge Base*, 3rd Edition. Cincinnati, OH, Atomic Dog Publishing. http://www.socialresearchmethods.net/kb/

Course Schedule

Suggested Background Readings

Symposium: Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?

- Clark, William Roberts and Matt Golder. 2015. "Big Data, Causal Inference, and Formal Theory: Contradictory Trends in Political Science?: Introduction" *PS: Political Science & Politics* 48(1):65-70.
- Monroe, Burt L., Jennifer Pan, Margaret E. Roberts, Maya Sen, Betsy Sinclair. 2015. "No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):71-74.
- Titiunik, Rocío. 2015. "Can Big Data Solve the Fundamental Problem of Causal Inference?" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):75-79.
- Grimmer, Justin 2015. "We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):80-83.
- Nagler, Jonathan and Joshua A. Tucker. 2015. "Drawing Inferences and Testing Theories with Big Data" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):84-88.
- Ashworth, Scott, Christopher R. Berry, Ethan Bueno de Mesquita. 2015. "All Else Equal in Theory and Data (Big or Small)" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):89-94.
- Patty, John W., Elizabeth Maggie Penn. 2015. "Analyzing Big Data: Social Choice and Measurement" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):95-101.
- Keele, Luke. 2015. "The Discipline of Identification" *PS: Political Science & Politics PS: Political Science & Politics* 48(1):102-106.

Day 1: Methods of Observation and Inference

Introduction to Exploratory Data Analysis, Visualization, and Validation

- Lazer, David and Jason Radford. 2017. "Data ex Machina: Introduction to Big Data" *Annual Review of Sociology*
- 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.

Day 2: Designing Validity

Introduction to Key Research Design Concepts

- Rubin, Donald B. 2008. "For Objective Causal Inference, Design Trumps Analysis." *Annals of Applied Statistics* 2(3):808-840.
- Shadish, William R. 2010. "Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings." *Psychological Methods* 15(1):3-17.

Day 3: Programming

Introduction to Parallel Programming in R for Exploration and Analysis of Social Media Data.

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

Day 4: Experimental and Quasi-Experimental Designs

Designing and Implementing Experiments with Social Media Data

- Bond, Robert M., Jaime E. Settle, Christopher J. Fariss, Jason J. Jones, and James H. Fowler. "Social Endorsement Cues and Political Participation in an Experiment Involving 61 Million Facebook Users" *Political Communication* (Forthcoming).
- Carlson, Taylor N. and Jaime E. Settle. "Political Chameleons: An Exploration of Conformity in Political Discussions" *Political Behavior* (Forthcoming).
- Crabtree, Charles D., Christopher J. Fariss, and Holger L. Kern. "Truth Replaced by Silence: A Field Experiment on Private Censorship in Russia." working paper.
- Hobbs, William and Moira Burke. 2017. "Connective Recovery in Social Networks After the Death of a Friend" *Nature Human Behaviour* 1 (0092).

Day 5: Designing Measurement for Social Media Data part 1

Introduction to Construct Validity and Measurement Models

- Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95(3):529–546.
- Jackman, Simon. 2008. "Measurement." In *The Oxford Handbook of Political Methodology*, edited by Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier. Oxford University Press.

Day 6: Designing Measurement for Social Media Data part 2

Measurement 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.
- Imai, Kosuke, James Lo, and Jonathan Olmsted. 2016. "Fast Estimation of Ideal Points with Massive Data" *American Political Science Review* 110(4):631-656.

Day 7: Designing Measurement for Social Media Data part 3

Measurement and Assessment of Incomplete and Missing Data

- Croicu, Mihai and Joakim Kreut. 2017. "Communication Technology and Reports on Political Violence: Cross-National Evidence Using African Events Data" *Political Research Quarterly* (Forthcoming).
- Driscoll, Jesse and Elaine Denny. "Fear of Anarchy or Fear of a Predatory State?: Using Survey Non-Response To Assess Somali State Legitimacy." working paper.
- Jesse, Stephen A. "'Don't Know' Responses, Personality and the Measurement of Political Knowledge" *Political Science Research and Methods* (Forthcoming).
- Steinert-Threlkeld, Zachary C., 2016, "Longitudinal Network Centrality Using Incomplete Data" *Political Analysis* (Forthcoming).

Day 8: Exploratory Data Analysis part 1

Random Forests

- Jones, Zachary M. and Fridolin Linder. 2016. "edarf: Exploratory Data Analysis using Random Forests" *Journal of Open Source Software*.
- Jones, Zachary M. and Yonatan Lupu. "Is There More Violence in the Middle?" working paper.

Day 9: Exploratory Data Analysis part 2

Validating Discovered Relationships

- Fariss, Christopher J. and Zachary M. Jones. "Enhancing Validity in Observational Settings When Replication is Not Possible" *Political Science Research and Methods* (Forthcoming).
- Grimmer, Justin and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21(3):267-297.

Day 10: Ethical Responsibilities for the Social Data Scientist

Issues Relating to Transparency and Research Ethics

- Driscoll, Jesse. 2015. "Prison States & Games of Chicken" in S. Desposato, *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*, Taylor and Francis.
- Margetts, Helen. 2017. "Political Behaviour and the Acoustics of Social Media" *Nature Human Behaviour* 1 (0086).

Next Steps

In this course, we have focused exclusively on the research design tools and analysis techniques necessary for working with and understanding social media 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 optimal 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 an Assistant Professor in the Department of Political Science at the University of Michigan. Prior to beginning this appointment I was 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.