POIR 613: Computational Social Science

Pablo Barberá

School of International Relations
University of Southern California
pabloobarbera.com

Course website:
pabloobarbera.com/POIR613/
Data is everywhere
Strongly agree □
Agree □
Disagree □
Strongly disagree □
How Obama's data crunchers helped him win

By Michael Scherer

President Obama's campaign manager hired an analytics department five times as large as that of the 2008 operation.
The Data revolution in election campaigns

Data Analyst

BROOKLYN, NY  ANALYTICS  FULL-TIME

We are looking for Data Analysts, at both the junior and senior levels, to join our team at our Brooklyn, NY headquarters. The Analyst will play a pivotal role in developing data-driven strategies for key primary and battleground states. They will be responsible for designing and building tools to guide strategies at all levels of the campaign. By utilizing their statistical expertise, our Analysts will dissect large datasets, synthesize results and present findings to team leaders.

2016

Trump's secret data reversal

Having once dismissed the importance of campaign tech, the mogul is now rushing to catch up with Clinton.

By KENNETH P. VOGEL and DARREN SAMUELSOHN | 06/28/16 05:22 AM EDT

Donald Trump has dismissed political data operations as “overrated,” but his campaign is now bolstering its online fundraising and digital outreach by turning to GOP tech specialists who previously tried to stop him from winning the party’s nomination.
50 Years of Electoral College Maps: How the U.S. Turned Red and Blue

A brief guided tour: Understanding the history of modern American politics means reckoning with the effect of race.

6h ago • By TONI MONKOVIC
Data without borders: why I want to change the world

Data scientist Jake Porway wants to hook up developers with charities and the developing world. Here he explains why.
Data Scientist: The Sexiest Job of the 21st Century

Meet the people who can coax treasure out of messy, unstructured data.
by Thomas H. Davenport and D.J. Patil

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren’t seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, “It was like arriving at a conference reception and realizing you don’t know anyone. So you just stand in the corner sipping your drink—and you probably leave early.”
How can we analyze *Big Data* to answer Political Science questions?
POIR 613

Goals

▶ Read and evaluate research applying computational methods to political science problems
▶ Learn how to collect and manipulate quantitative data
▶ Develop skills necessary to analyze large and heterogeneous datasets

Outline (see detailed scheduled here)

▶ Weeks 3-4: Surveys and experiments
▶ Weeks 5-9: Text as data methods
▶ Week 9-13: Social network analysis
▶ Weeks 11: SQL
Hello!

HELLO

IT'S ME
About me

▶ Assistant Professor in International Relations at Univ. of Southern California
▶ Research scientist at Facebook Core Data Science
▶ PhD in Politics, New York University (2015)
▶ Data Science Fellow at NYU, 2015–2016
▶ My research:
  ▶ Social media and politics, comparative electoral behavior, corruption and accountability
  ▶ Social network analysis, Bayesian statistics, text as data methods
  ▶ Author of R packages to analyze data from social media
▶ Contact:
  ▶ pbarbera@usc.edu
  ▶ www.pablobarbera.com
  ▶ Office hours: Wed 1pm-2pm (VKC 359A)
Your turn!

1. Name?
2. Department, year?
3. Research interests?
4. Previous experience with R?
5. Why are you interested in this course?
The plan for today

- Introductions
- Logistics
- R and RStudio Server
- What is CSS? Opportunities and challenges
- Good practices in scientific computing
- GitHub and version control
Course philosophy

How to learn the techniques in this course?

▶ Lecture approach: not ideal for learning computational social science methods

▶ You can only learn by doing:
  → Reading and criticizing research
  → Applying methods to social science problems

▶ Structure of each session:
  1. Introduction to the topic (30 minutes)
  2. Discussion of research (50 minutes)
  3. Guided coding session (30-40 minutes)
  4. Coding challenges (30 minutes)

▶ You will continue working on the coding challenges after class and submit before beginning of next class
POIR 613

Computational Social Science
University of Southern California, Fall 2019

Citizens across the globe spend an increasing proportion of their daily lives online. Their activities leave behind granular, time-stamped footprints of human behavior and personal interactions that represent a new and exciting source of data to study standing questions about political and social behavior. At the same time, the volume and heterogeneity of digital data present unprecedented methodological challenges. The goal of this course is to introduce students to new computational social science methods and tools required to explore and harness the potential of “Big Data” using the R programming language.

The course will follow a “learning-by-doing” approach and will place emphasis on gaining experience in analyzing data with the R programming language. Students are expected to do the required readings and coding exercises for each week. The lectures will build upon the content of the readings with a series of data challenges that will introduce new statistical and programming concepts, which will then be applied to the analysis of data from published research papers or common tasks in computational social science. Most of the applications will be related to Political Science and International Relations questions, but the course should be of interest to social science students more generally.

pablobarbera.com/POIR613
Evaluation

- **Class participation:** 10%
  - Do all “readings for discussion” (required)
  - If unfamiliar with topic, also background reading

- **Referee reports and presentations:** 20%
  - TWO peer reviews (800-1000 words) of readings for discussion, due 8pm day before the class via email
  - 10-minute presentation in class (slides optional)

- **Coding challenges:** 20%
  - Not graded but submission (.Rmd + html/pdf files) of at least FIVE is required before next class

- **Research project:** 50%
  - Original research paper (8,000 words) that employs computational methods in political science. Individual or group project (up to 3 people)
Research project

**Goal:** demonstrate ability to conduct research that applies computational methods to political science questions.

**Constant progress** throughout semester:

09/20  Project idea (one paragraph)
10/07  Project summary (2 pages)
10/15  Feedback from peers
11/04  Summary with descriptive statistics (5 pages)
11/25  First full draft (10-15 pages)
12/03  Student presentations
12/18  Final paper

See [course website](#) for more information.
Why we’re using R

- Becoming *lingua franca* of statistical analysis in academia
- What employers in private sector demand
- It’s free and open-source
- Flexible and extensible through *packages* (over 10,000 and counting!)
- Powerful tool to conduct automated text analysis, social network analysis, and data visualization, with packages such as quanteda, igraph or ggplot2.
- Command-line interface and scripts favors reproducibility.
- Excellent documentation and online help resources.

R is also a full programming language; once you understand how to use it, you can learn other languages too.
library(ggplot2)
source("plots/FormatPlot.R")
View(diamonds)
summary(diamonds)
summary(diamonds$price)
aveSize <- round(mean(diamonds$carat), 4)
clarity <- levels(diamonds$clarity)
p <- ggplot(carat, price, 
  data=diamonds, color=clarity,
  xlab="Carat", ylab="Price",
  main="Diamond Pricing")
format.plot(p, size=24)
Big Data: Opportunities and Challenges
BUT WHAT IS
BIG DATA??
The Three V’s of Big Data

Dumbill (2012), Monroe (2013):

1. **Volume**: 6 billion mobile phones, 1+ billion Facebook users, 500+ million tweets per day...
2. **Velocity**: personal, spatial and temporal granularity.
3. **Variability**: images, networks, long and short text, geographic coordinates, streaming...

**Big data**: data that are so large, complex, and/or variable that the tools required to understand them must first be invented.
“We have life in the network. We check our emails regularly, make mobile phone calls from almost any location ... make purchases with credit cards ... [and] maintain friendships through online social networks ... These transactions leave digital traces that can be compiled into comprehensive pictures of both individual and group behavior, with the potential to transform our understanding of our lives, organizations and societies”.

**Lazer et al (2009) Science**

“*Digital footprints* collected from online communities and networks enable us to understand human behavior and social interactions in ways we could not do before”.

**Golder and Macy (2014) ARS**
Computational Social Science

Two different approaches in the growing field of computational social science:

1. Big data as a new source of information
   - Behavior, opinions, and latent traits
   - Interpersonal networks
   - Elite behavior
   - Affordable online experiments

2. How big data and social media affect social behavior
   - Collective action and social movements
   - Political campaigns
   - Social capital and interpersonal communication
   - Political attitudes and behavior
Big data and social science: challenges

1. Big data, big bias?
2. The end of theory?
3. Spam and bots
4. The privacy paradox
5. Generalizing from online to offline behavior
6. Ethical concerns
Computational *social* science

**Challenge for social scientists:** need for advanced technical training to collect, store, manipulate, and analyze massive quantities of semistructured data.

Discipline *dominated by computer scientists* who lack theoretical grounding necessary to know where to look.

Even if analysis of big data requires thoughtful measurement, careful research design, and creative deployment of statistical techniques (Grimmer, 2015).

**New required skills for social scientists?**

▶ Manipulating and storing large, unstructured datasets
▶ Webscraping and interacting with APIs
▶ Machine learning and topic modeling
▶ Social network analysis
For next week

1. Sign up for TWO peer reviews. Email with link will be sent tomorrow at 2pm.
2. Do reading for discussion: Kramer et al 2014 (and “Editorial Expression of Concern”) and Hargittai 2018
3. New to CSS? Do background readings