

POIR 613: Computational Social Science

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Course website:
pablobarbera.com/POIR613/

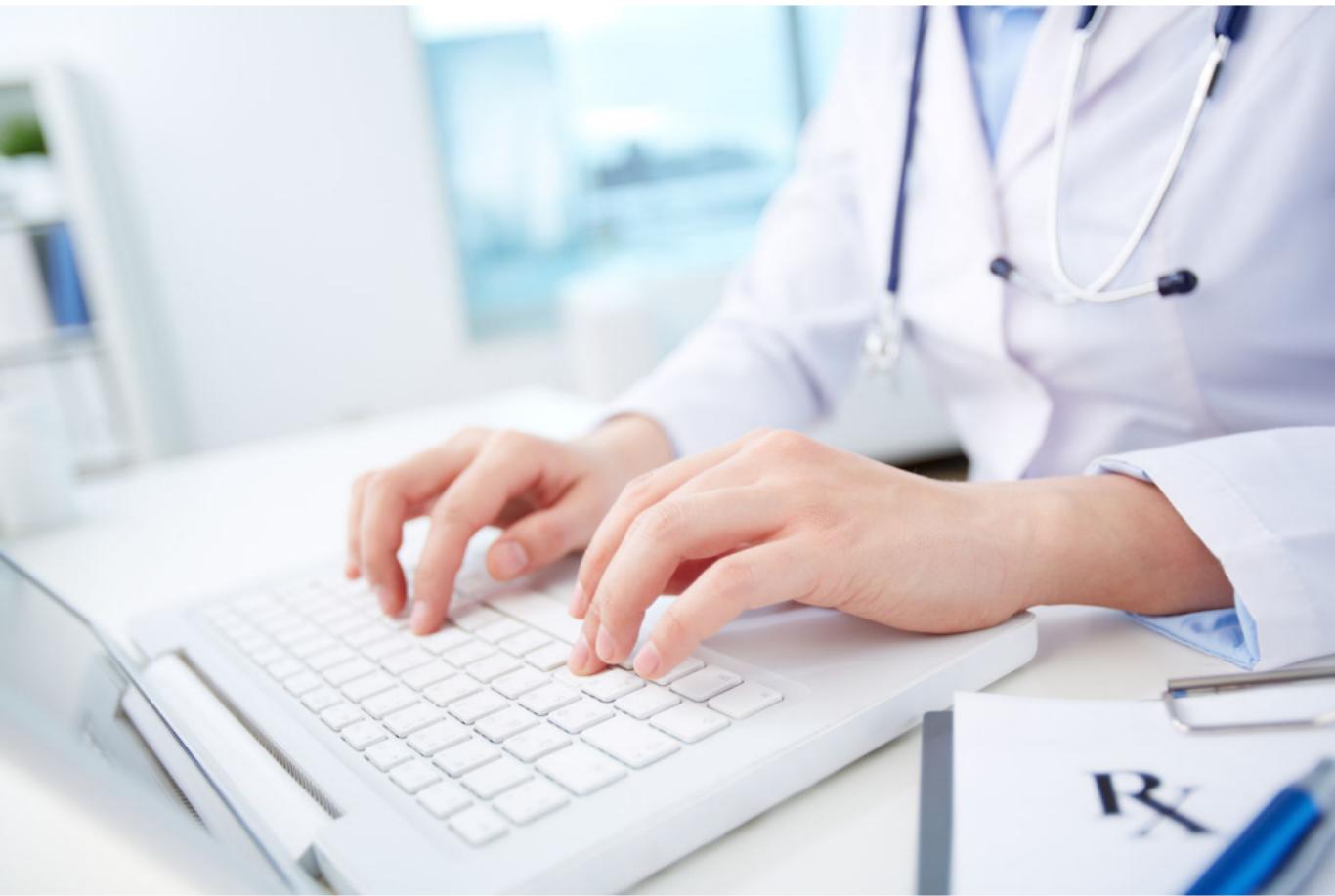
The background of the image is a dense, colorful blur of digital data. It appears as a grid of small, multi-colored pixels in shades of blue, green, red, and yellow, resembling a computer screen or a data visualization. Interspersed among these colors are several words and patterns that are partially visible and out of focus, such as "click", "rate", "eval", "image", and "data".

Data is everywhere



243.26

TOMACCO
JUICE



2005



2013





Google Books Ngram Viewer

Graph these comma-separated phrases: case-insensitive

between and from the corpus with smoothing of



Strongly agree

Agree

Disagree

Strongly disagree

The Data revolution in election campaigns



Tech » Gadgets | Cyber Security | Innovation Nation

Live TV

U.S. Edition



menu

How Obama's data crunchers helped him win

By Michael Scherer

Updated 11:45 AM ET, Thu November 8, 2012



President Obama's campaign manager hired an analytics department five times as large as that of the 2008 operation.

Top stories



Top US commander warns Russia, Syria



Is NBC's Olympics coverage really that bad?

The Data revolution in election campaigns



Data Analyst

[APPLY FOR THIS JOB](#)

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 SAMUELSON | 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.

Data Journalism

The Upshot

THE 2016 RACE

50 Years of Electoral College Maps: How the U.S. Turned Red and Blue

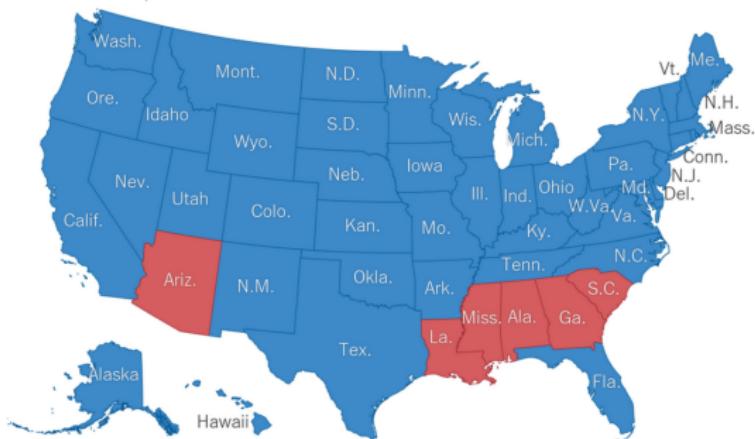
A brief guided tour: Understanding the history of modern American politics means reckoning with the effect of race.

6h ago • By TONI MONKOVIC

The 1964 Election: Johnson Defeats Goldwater
In the popular vote, Lyndon Johnson defeated Barry Goldwater, 61.1 to 38.5.

The blue states reflect a total of 486 electoral votes for Lyndon Johnson

■ Dem. ■ Rep.



Non-profit sector

Development
data
Datablog

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

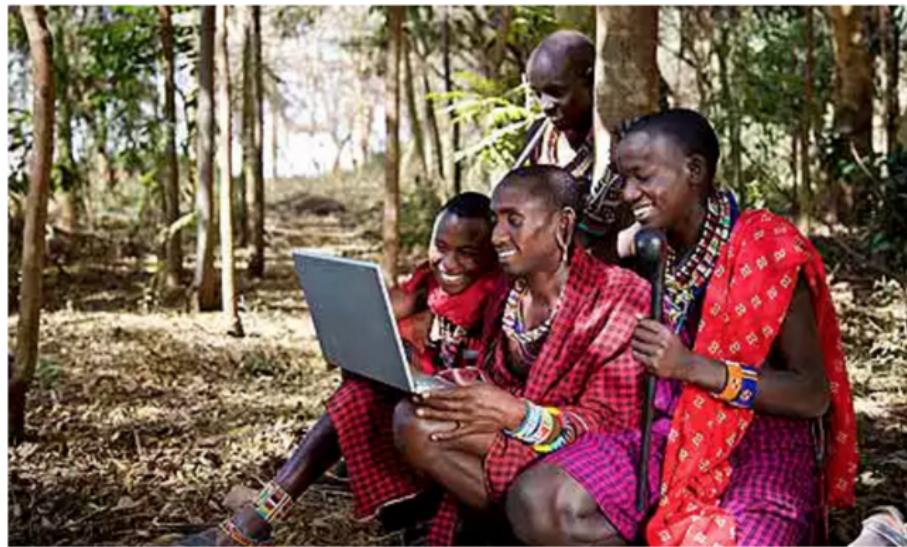
Jake Porway

Thursday 23 June 2011
11.10 EDT



8 Shares 0 Comments

Save for later



Data without borders: Men on the Samburu National Reserve, Kenya, using a laptop. Photograph: Scott Stulberg/Corbis

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

W

hen 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 schedule [here](#))

- ▶ Weeks 1-2: Introduction. Ethics.
- ▶ Weeks 3-4: Surveys and experiments
- ▶ Weeks 5-9: Text as data methods
- ▶ Week 9-13: Social network analysis
- ▶ Weeks 11: SQL

Hello!



HELLO

A black and white photograph of a woman with blonde hair, wearing a fur-trimmed coat, singing with her eyes closed and mouth open. The word "HELLO" is overlaid in large, bold, white letters across the center of the image.

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

Course website

The screenshot shows a course website layout. At the top, there is a navigation bar with links for POIR 613, Overview, Syllabus, Project, Code, Resources, and a user icon. The main content area has a title "POIR 613" and subtitle "Computational Social Science" followed by the text "University of Southern California, Fall 2019". On the left, a sidebar titled "POIR 613" contains links for Instructor, Schedule, Prerequisites, Course structure, Course requirements and grading, Software, License and credit, and Feedback.

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.

RStudio Server

RStudio

File Edit Code View Project Workspace Plots Tools Help

Go to file/function

Project: (None)

diamondPricing.R* | formatPlot.R* | diamonds*

Source On Save

1 library(ggplot2)
2 source("plots/formatPlot.R")
3
4 view(diamonds)
5 summary(diamonds)
6
7 summary(diamonds\$price)
8 aveSize <- round(mean(diamonds\$carat), 4)
9 clarity <- levels(diamonds\$clarity)
10
11 p <- qplot(carat, price,
12 data=diamonds, color=clarity,
13 xlab="Carat", ylab="Price",
14 main="Diamond Pricing")
15

15:1 (Top Level) R Script

Console

	x	y	z
Min. :	0.000	0.000	0.000
1st Qu.:	4.710	4.720	2.910
Median :	5.700	5.710	3.530
Mean :	5.731	5.735	3.539
3rd Qu.:	6.540	6.540	4.040
Max. :	10.740	58.900	31.800

> summary(diamonds\$price)

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
326	950	2401	3933	5324	18820	

> aveSize <- round(mean(diamonds\$carat), 4)

> clarity <- levels(diamonds\$clarity)

> p <- qplot(carat, price,
+ data=diamonds, color=clarity,
+ xlab="Carat", ylab="Price",
+ main="Diamond Pricing")

> format.plot(p, size=24)

> |

Workspace History

Load Save Import Dataset Clear All

Data diamonds 53940 obs. of 10 variables

Values aveSize 0.7979

clarity character [8]

p ggplot

Functions format.plot(plot, size)

Files Plots Packages Help

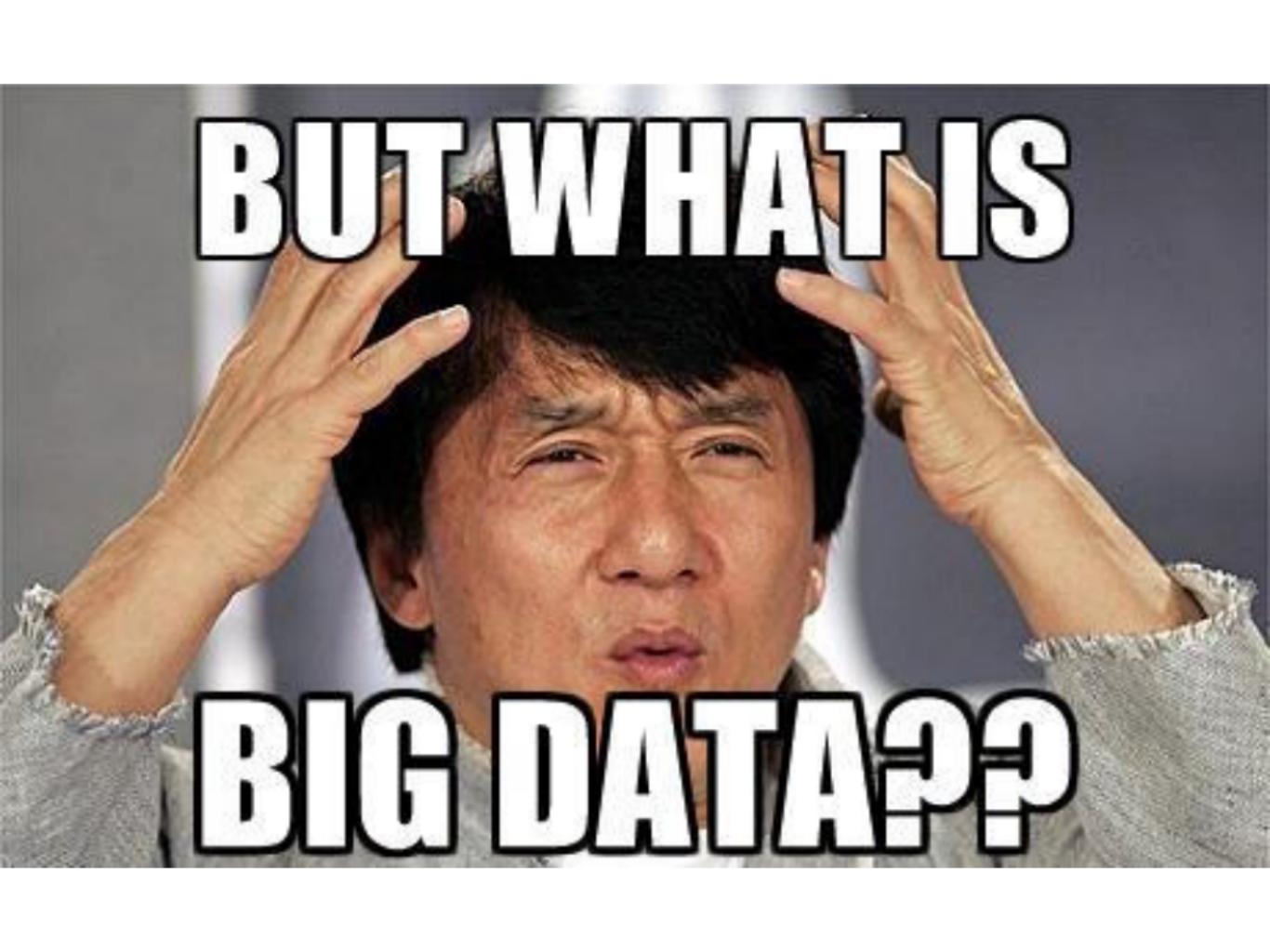
Zoom Export Clear All

Diamond Pricing

Clarity

- H1
- SI2
- SI1
- VS2
- VS1
- VVS2
- VVS1
- IF

Big Data: Opportunities and Challenges

A photograph of Jackie Chan from the chest up. He has dark hair and is looking directly at the camera with a confused expression. His hands are raised to his head, with his fingers pointing upwards. He is wearing a light-colored, possibly white, collared shirt.

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.

Computational Social Science

*"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

Good (enough) practices in scientific computing

Based on Nagler (1995) “Coding Style and Good Computing Practices” (PS) and Wilson *et al* (2017) “Good Enough Practices in Scientific Computing” (PLOS Comput Biol)

Good practices in scientific computing

Why should I waste my time?

- ▶ Replication is a key part of science:
 - ▶ Keep good records of what you did so that others can understand it
 - ▶ “Yourself from 3 months ago doesn’t answer emails”
 - ▶ More efficient research: avoid retracing own steps
 - ▶ Your future self will be grateful

General principles:

1. Good documentation: README and comments
2. Modularity with structure
3. Parsimony (without being too smart)
4. Track changes

Summary of good practices

1. Safe and efficient data management
2. Well-documented code
3. Organized collaboration
4. One project = one folder
5. Track changes
6. Manuscripts as part of the analysis

1. Data management

- ▶ Save raw data as originally generated
- ▶ Create the data you wish to see in the world:
 - ▶ Open, non-proprietary formats: e.g. `.csv`
 - ▶ Informative variable names that indicate direction:
`is_political` instead of `topic` or `V322`; `voted` vs `turnout`
 - ▶ Recode missing values to `NA`
 - ▶ File names that contain metadata: e.g. `05-alaska.csv` instead of `state5.csv`
- ▶ Record all steps used to process data and store intermediate data files if computationally intensive (easier to rerun parts of a data analysis pipeline)
- ▶ Separate data manipulation from data analysis
- ▶ Prepare README with codebook of all variables
- ▶ Periodic backups (or Dropbox, Google Drive, etc.)
- ▶ Sanity checks: summary statistics after data manipulation

2. Well-documented code

- ▶ Number scripts based on execution order:
 - e.g. 01-clean-data.r, 02-recode-variables.r, 03-run-regression.r, 04-produce-figures.R...
- ▶ Write an explanatory note at the start of each script:
 - Author, date of last update, purpose, inputs and outputs, other relevant notes
- ▶ Rules of thumb for modular code:
 1. Any task you run more than once should be a function (with a meaningful name!)
 2. Functions should not be more than 20 lines long
 3. Separate functions from execution (e.g. in functions.r file and then use source(functions.r) to load functions to current environment)
 4. Errors should be corrected when/where they occur
- ▶ Keep it simple and don't get too clever
- ▶ Add informative comments before blocks of code

3. Organized collaboration

- ▶ Create a `README` file with an overview of the project: title, brief description, contact information, structure of folder
- ▶ Shared to-do list with tasks and deadlines
- ▶ Choose one person as corresponding author / point of contact / note taker
- ▶ Split code into multiple scripts to avoid simultaneous edits
- ▶ ShareLatex, Overleaf, Google Docs to collaborate in writing of manuscript

4. One project = one folder

Logical and consistent folder structure:

- ▶ `code` or `src` for all scripts
- ▶ `data` for raw data
- ▶ `temp` for temporary data files
- ▶ `output` or `results` for final data files and tables
- ▶ `figures` or `plots` for figures produced by scripts
- ▶ `manuscript` for text of paper
- ▶ `docs` for any additional documentation

5 & 6. Track changes; producing manuscript

- ▶ Ideally: use version control (e.g. GitHub)
- ▶ Manual approach: keep dates versions of code & manuscript, and a `CHANGELOG` file with list of changes
- ▶ Dropbox also has some basic version control built-in
- ▶ Avoid typos and copy&paste errors: tables and figures are produced in scripts and compiled directly into manuscript with `LATEX`

Examples

Replication materials for some of my published articles:

- ▶ 2019 APSR
- ▶ 2017 ISQ

John Myles White's [ProjectTemplate R package](#).

Replication materials for Leeper 2017:

- ▶ Code and data

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