ECPR Methods Summer School: Big Data Analysis in the Social Sciences

Pablo Barberá

London School of Economics
pablobarbera.com

Course website:
pablobarbera.com/ECPR-SC105


Data is everywhere


243.26
TOMACCO
JUICE


Rx

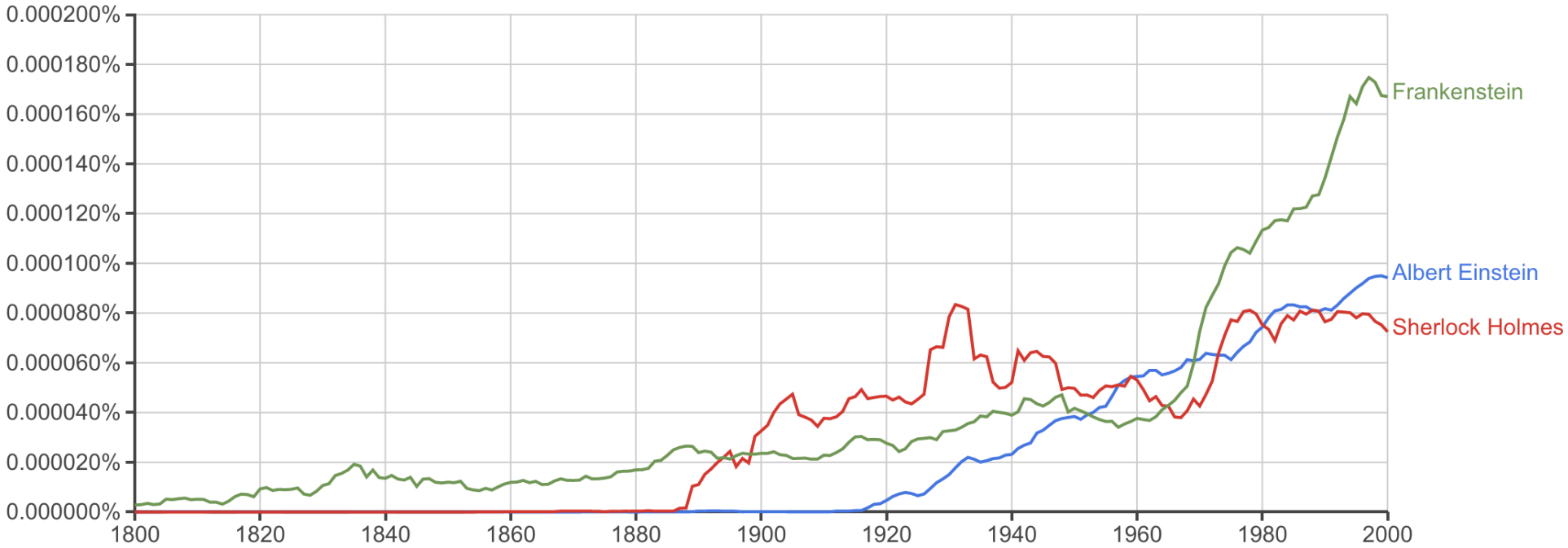
NBC/Instagram


Gefällt mir
Instagram
Pinterest
twitter
LinkedIn
facebook
flickr
skype
YouTube
XING

Google Books Ngram Viewer

Graph these comma-separated phrases: Albert Einstein,Sherlock Holmes,Frankenstein case-insensitive

between 1800 and 2000 from the corpus English with smoothing of 3. Search lots of books




Strongly agree
Agree
Disagree
ly disagree

The Data revolution in election campaigns

CNN Tech » Gadgets | Cyber Security | Innovation Nation Live TV U.S. Edition + menu

How Obama's data crunchers helped him win

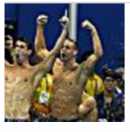
By Michael Scherer

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 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.

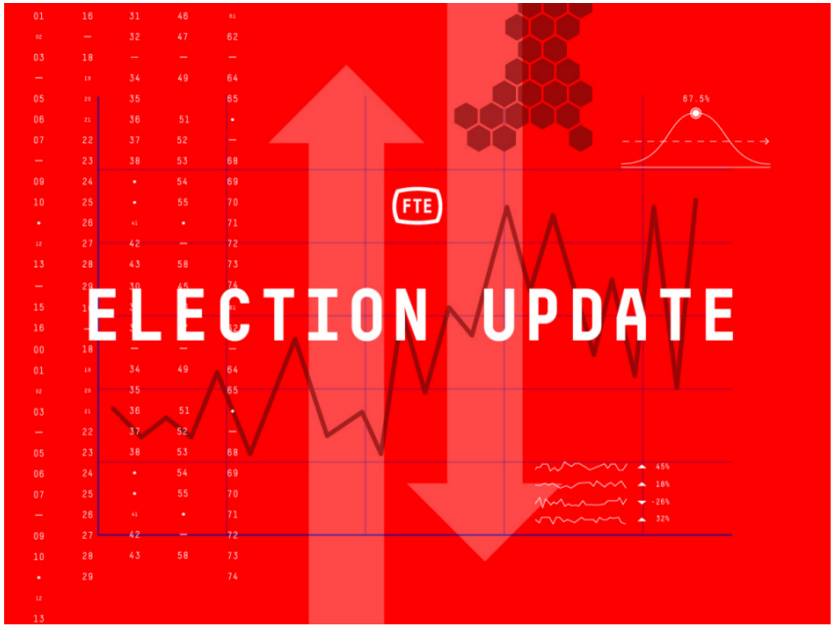
Data Journalism

FiveThirtyEight

ESPN

Politics Sports Science & Health Economics Culture

All week: Rio 2016 coverage



2016 ELECTION

National Polls Show The Race Tightening — But State Polls Don't

By Nate Silver

THE LATEST

8:55 AM
Significant Digits For Monday, Aug. 22, 2016

AUG 21
Election Update: National Polls Show The Race Tightening — But State Polls Don't

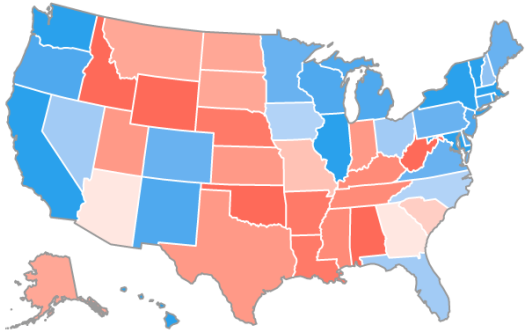
AUG 19
Winning An Olympic Gold Medal Hasn't Been This Difficult Since 1896

AUG 19
Let Caster Run! We Should Celebrate Semenya's Extraordinary Talent

INTERACTIVES

2016 Election Forecast

UPDATED 4 HOURS AGO



See polls and forecasts

MLB Predictions

UPDATED 15 HOURS AGO

Upcoming games

Nationals def. Orioles	50%
Pirates def. Astros	56%

Non-profit sector

Development data
Datablog

Data without borders: why I want to change the world

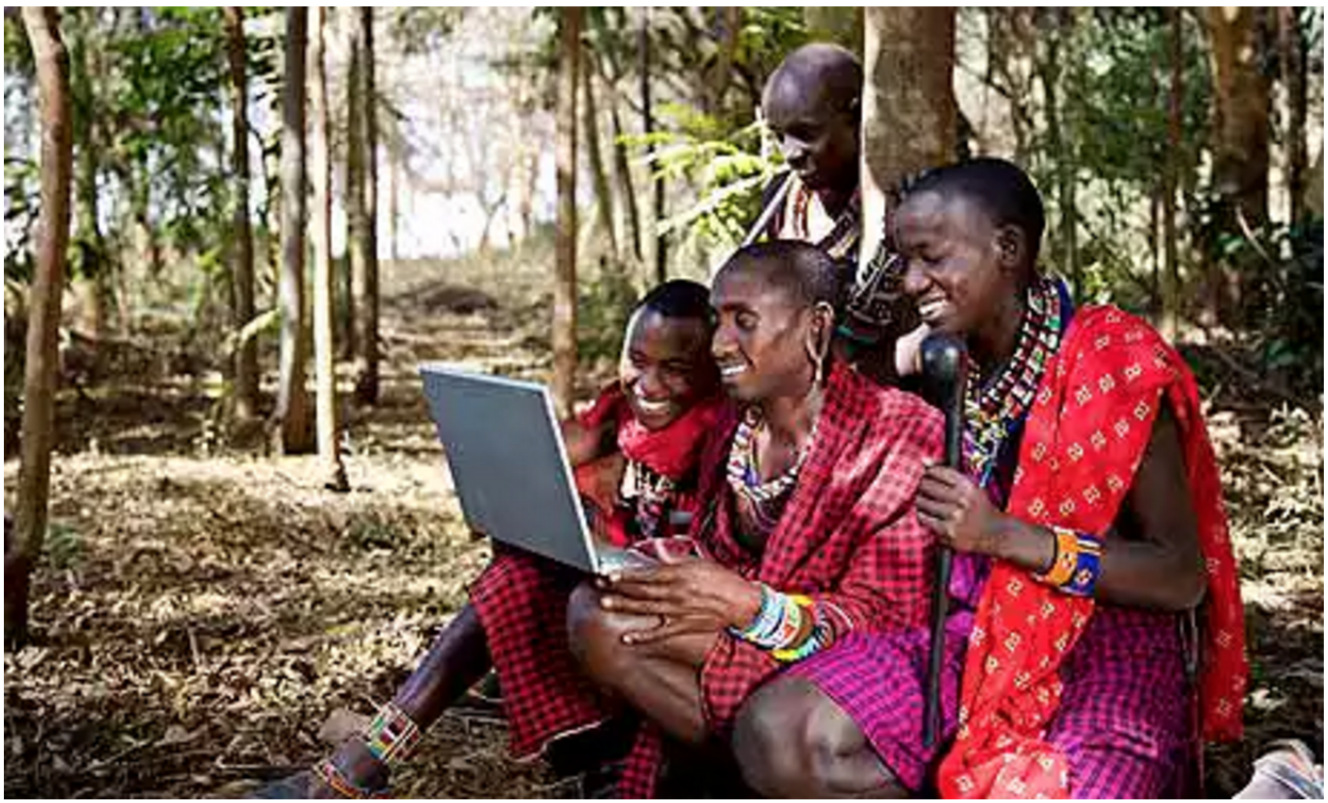
Data scientist **Jake Porway** wants to hook up developers with charities and the developing world. Here he explains why

Jake Porway

Thursday 23 June 2011
11.10 EDT

Shares 8 Comments 0

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

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 social science questions?

Course outline

1. Efficient data analysis in R
   - Good coding practices
   - Parallel computing
2. Cloud computing
   - SQL for data manipulation
   - Large-scale data processing in the cloud
3. Large-scale discovery in networks
   - Community detection
   - Latent space network models
4. Text classification at scale
   - Supervised machine learning
   - Large-scale text classification
5. Topic discovery in text
   - Exploratory analysis of textual datasets
   - Topic models

Hello!

About me: Pablo Barberá

- Assistant Professor of Computational Social Science at the London School of Economics
  - Previously Assistant Prof. at Univ. of Southern California
  - PhD in Politics, New York University (2015)
  - Data Science Fellow at NYU, 2015–2016
- My research:
  - Social media and politics, comparative electoral behavior
  - Text as data methods, social network analysis, Bayesian statistics
  - Author of R packages to analyze data from social media
- Contact:
  - P.Barbera@lse.ac.uk
  - www.pablobarbera.com
  - @p_barbera

About me: Tom Paskhalis

- PhD candidate in Social Research Methods at the London School of Economics
- My research:
  - Interest groups and political parties
  - Text as data, record linkage, Bayesian statistics
  - Author/contributor to R packages to scrape websites and PDF documents
- Contact:
  - `T.G.Paskhalis@lse.ac.uk`
  - `tom.paskhal.is`
  - `@tpaskhalis`

Your turn!

1. Name?
2. Affiliation?
3. Research interests?
4. Previous experience with R?
5. Why are you interested in this course?

Course philosophy

How to learn the techniques in this course?

- Lecture approach: not ideal for learning how to code
- You can only learn by doing.

→ We will cover each concept three times during each session

1. Introduction to the topic (20-30 minutes)
2. Guided coding session (30-40 minutes)
3. Coding challenges (30 minutes)

- You're encouraged to continue working on the coding challenges after class. Solutions will be posted the following day.
- Warning! We will move fast.

Course logistics

ECTS credits:

- Attendance: 2 credits (pass/fail grade)
- Submission of at least 3 coding challenges: +1 credit
  - Due before beginning of following class via email to Tom or Alberto
  - Only applies to challenge 2 of the day
  - Graded on a 100-point scale
- Submission of class project: +1 credit
  - Due by August 20th
  - Goal: collect and analyze data from the web or social media
  - 5 pages max (including code) in Rmarkdown format
  - Graded on a 100-point scale

If you wish to obtain more than 2 credits, please indicate so in the attendance sheet

Social event

Save the date:

Wednesday Aug. 8, 6.30pm

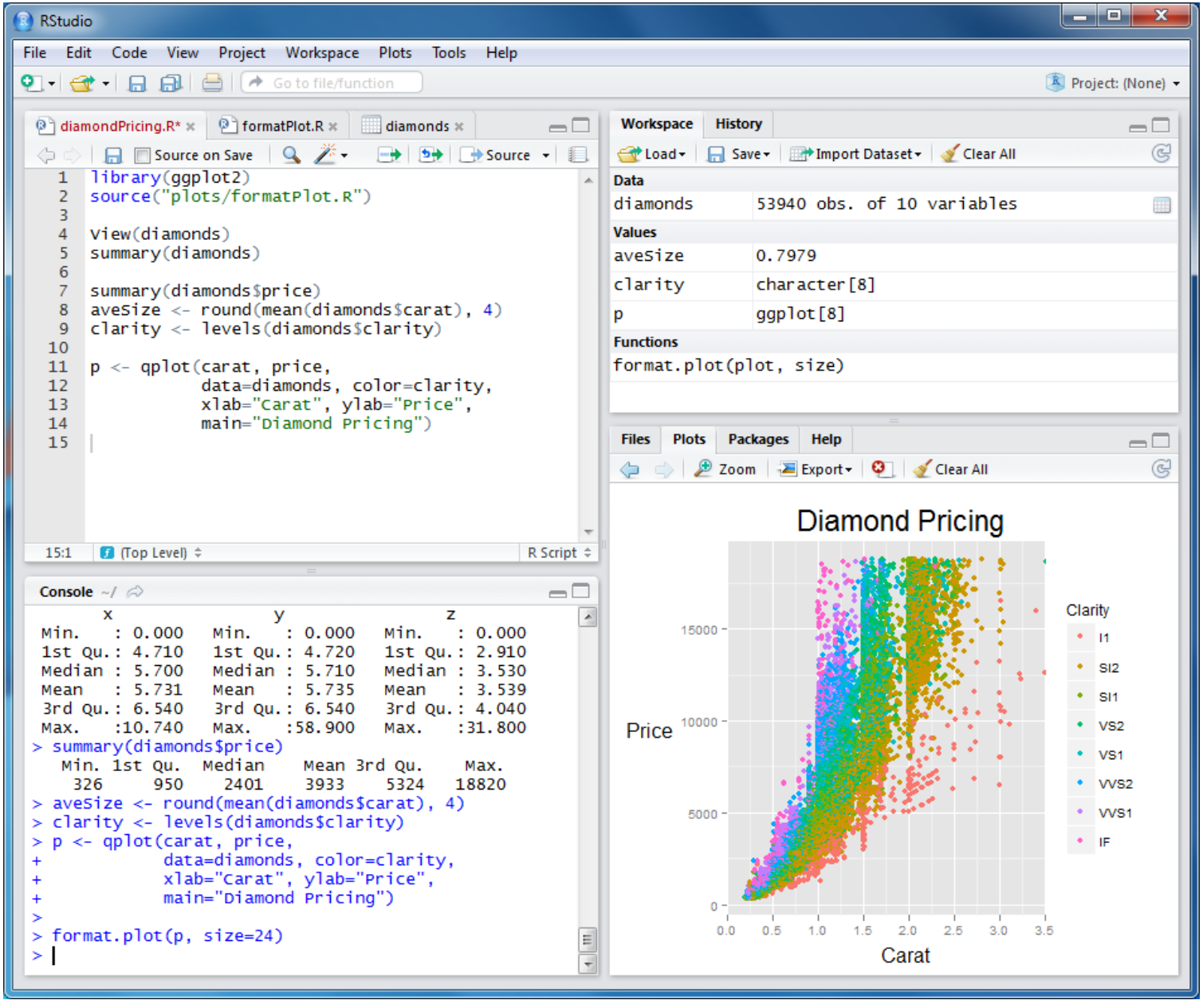
Location TBA

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.

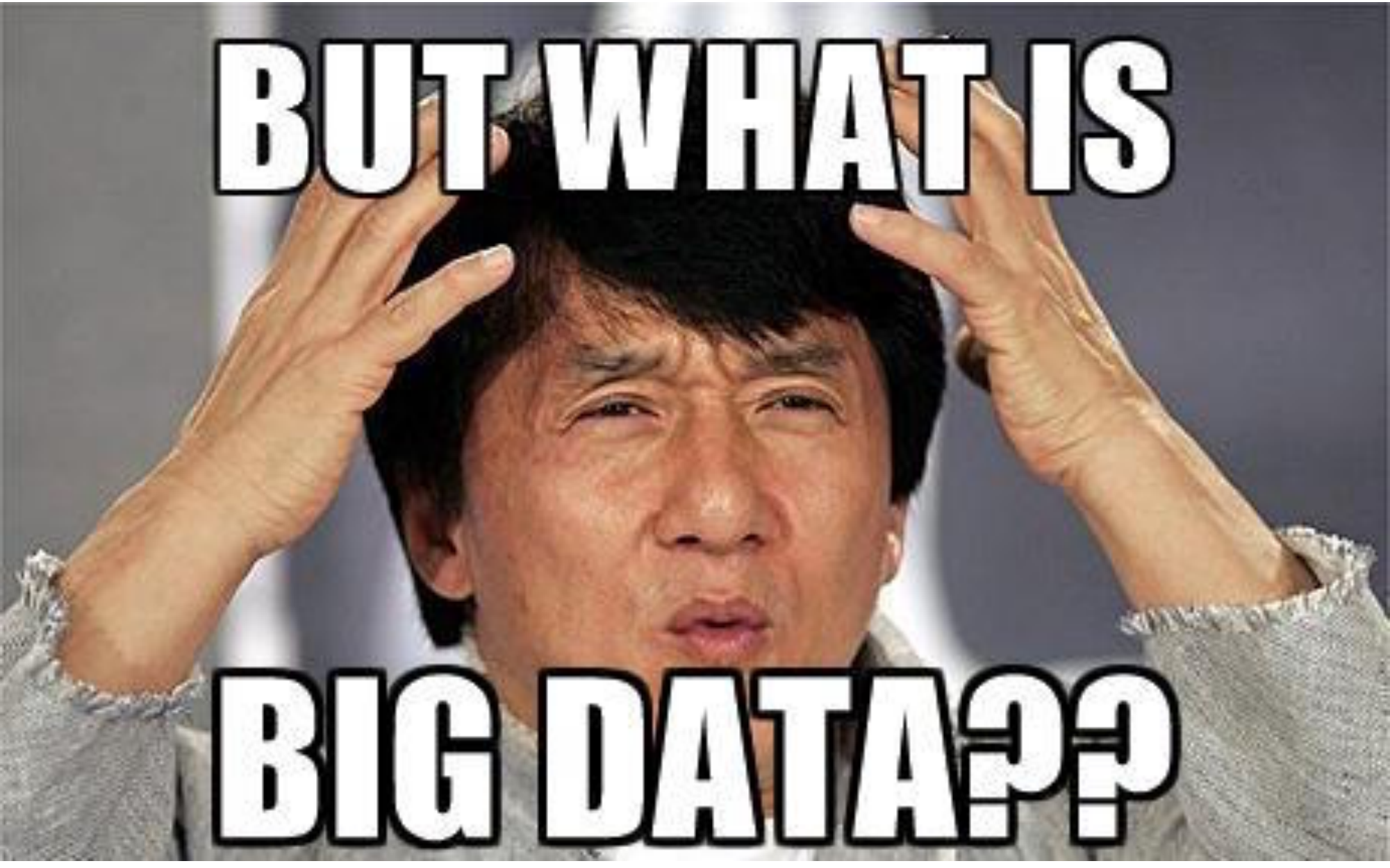
RStudio Server



Course website

pablobarbera.com/ECPR-SC105

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.

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

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: `female` instead of `gender` 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 my 2014 PA paper:

- Code on GitHub
- Code and Data

John Myles White's ProjectTemplate R package.

Replication materials for Leeper 2017:

- Code and data