ECPR Methods Summer School: Automated Collection of Web and Social Data

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:
github.com/pablobarbera/ECPR-SC103

George Takei
March 28 at 10:10pm ·

Who's with me.

Like · Comment · Share

408,735 people like this.

66,990 shares

Bon Alimagno
@karma_thief

Follow

I need a hug. I have never been so traumatized by a television show. #gameofthrones

Reply Retweet Favorite ••• More

RETWEETS	FAVORITES
356	110

10:06 PM - 2 Jun 2013

Google

how do I convert to

how do i convert to **judaism**
how do i convert to **islam**
how do i convert to **catholicism**
how do i convert to **pdf**

Press Enter to search.

Justin Bieber
@justinbieber

Follow

I make music. I love music.

Reply Retweet Favorite ••• More

RETWEETS	FAVORITES
54,213	59,205

10:09 PM - 7 Apr 2014

dustin curtis
@dcurtis

Follow

"At any moment, Justin Bieber uses 3% of our infrastructure. Racks of servers are dedicated to him. - A guy who works at Twitter

RETWEETS	FAVORITES
1,528	267

8:56 PM - 6 Sep 2010

Dmitry Medvedev
@MedvedevRussiaE

Follow

The harmonious development of Crimea and Sevastopol as part of our state is one of the main objectives of the Russian Government

Reply Retweet Favorite More

RETWEETS	FAVORITES
144	57

10:39 AM - 21 Mar 2014

The New York Times
April 2

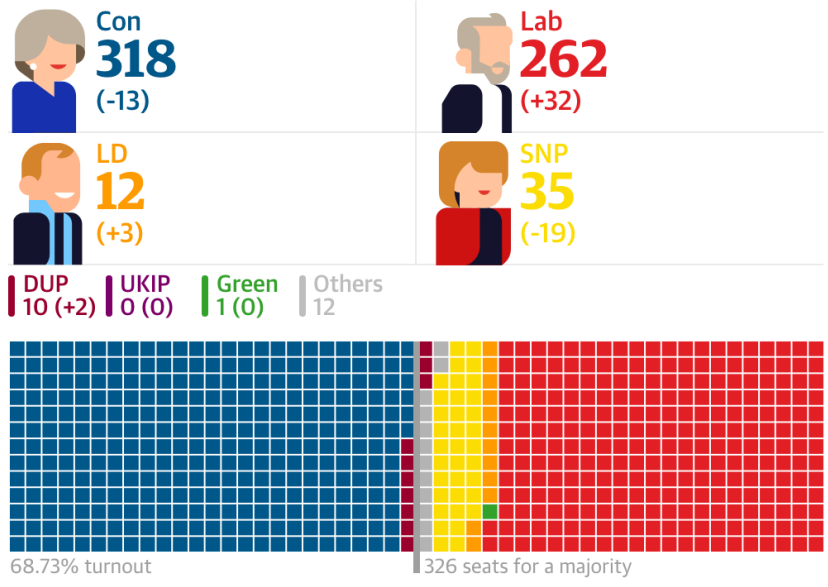
"Much of the foreign media coverage has distorted the reality of my country and the facts surrounding the events," writes Nicolás Maduro, the president of Venezuela, in Opinion: http://nyti.ms/1gP5o2I

Like · Comment · Share 57

262 people like this. Top Comments

Total seats won

650 of 650 seats declared at 9 Jun 21.06 BST



Donald J. Trump
@realDonaldTrump

Folgen

Are you allowed to impeach a president for gross incompetence?

Original (Englisch) übersetzen

RETWEETS	GEFÄLLT
195.387	161.489

03:23 - 4. Juni 2014

15 Tsd. 195 Tsd. 161 Tsd.

Shift in communication patterns

How can we collect web and social data to answer social science questions?

Course outline

1. Scraping data from the web
   - Key tools for webscraping
   - Tables; web data in unstructured format
2. Working with APIs
   - How to build an http request
   - Interacting with Newspaper APIs
3. Collecting social media data
   - Twitter: streaming and static data
   - Facebook: posts on public pages
4. New types of data
   - Text as data methods
   - Social network analysis
5. Advanced topics
   - Parsing data in PDF format
   - Data manipulation

Hello!

About me

- Assistant Professor in Computational Social Science at the London School of Economics as of January 2018
- Currently Assistant Prof. at Univ. of Southern California
- Director, Networked Democracy Lab, `netdem.org`
- 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`

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?

- Traditional lecture approach is not ideal for learning how to code
- You can only learn by doing.

→ In this course, we will cover each concept three times:

1. Introduction to the topic (20-30 minutes)
2. Guided coding session (30 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.
- Additional questions? We can arrange one-on-one meetings after class

Why we're using R

- Becoming *lingua franca* of statistical analysis in academia
- It's what employers in private sector demand
- It's FREE and open-source
- Flexible and extensible through *packages*, with new statistical methods being implemented first in R
- Command-line interface favors reproducibility
- Great for data visualization

R is also a full programming language; once you understand how to use it, you can learn other languages too.