# Social Media and Big Data Research

Pablo Barberá
University of Southern California

## Short bio – Pablo Barberá

Pablo Barberá (PhD in Politics, New York University) is an Assistant Professor of International Relations at the University of California, and a former Moore-Sloan Fellow at the Center for Data Science in New York University. His primary research interests include quantitative political methodology and computational social science applied to the study of political and social behavior. He is an active contributor to the open source community and has authored several R packages to mine social media data. His research has been published in high-impact journals such as Political Analysis, Journal of Communication, PLOS ONE, Psychological Science, Political Science Research and Methods, the Journal of Computer-Mediated Communication, and Social Science Computer Review, among others. More information is available at www.pablobarbera.com

## Course description

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 web data present unprecedented methodological challenges. The goal of this course is to introduce participants to new computational methods and tools required to explore and analyze Big Data from online sources using the R programming language. We will focus in particular on data collected from social networking sites, such as Facebook and Twitter, whose use is becoming widespread in the social sciences.

Each session will provide an overview of the literature and research methods on one of three main themes of the course -- big data, network data, and text data -- to then dive into a specific application, documenting each step from data collection to the analysis required to test hypotheses related to core social science questions. Code and data for all the applications will be provided, and students are encouraged to bring their own laptops to follow along.

The first session will begin with a discussion of the definition of "Big Data" and the research opportunities and challenges of the use of massive-scale datasets in the social sciences. We will then focus on how social media sites represent a new source of data to study human behavior, and also how its use raises a whole new series of questions that are relevant to social scientists. The applied part of this session will demonstrate different methods to "scrape" data from the web, with sentiment analysis of newspaper stories as a running example.

Social network analysis applied to social media data will be the main theme of the second session. After a short introduction to the main concepts and methods in network analysis, we will go over two applications: the online diffusion of information in the context of a social protest, and the debate on whether social media represents an "echo chamber" where individuals are only exposed to information that aligns with their previous political beliefs. Both examples rely on Twitter data -- this session will also explain the different types of data available and how they can be collected.

The last session of the course will focus on automated text analysis. I will introduce the two main techniques currently used by social scientists: large-scale classification of documents into categories (supervised learning) and discovery of "topics" or classes of documents in a corpus (unsupervised learning). These methods will be demonstrated in the context of two applications -- one related to the automated detection of hate speech in tweets, and another showing how Facebook posts by politicians can be classified into relevant political issues. As with the previous sessions, here we will cover the entire research process, from the collection of data through the Facebook API to the estimation of machine learning methods and the interpretation of the output.

After the course, students will have an advanced understanding of the opportunities of big data and social media mining for social science studies, and will be equipped with the technical skills necessary to conduct their own research.

## Software

The course will use the open-source software R, which is freely available for download at https://www.r-project.org/ . We will interact with R through RStudio, which can be downloaded at https://www.rstudio.com/products/rstudio/download/ Please download the most recent version at the time of the workshop (currently R 3.3.2 and RStudio 1.0.136). We will also utilize the following R packages: rvest, jsonlite, igraph, rtweet, streamR, quanteda, Rfacebook.

## Prerequisites

The course will assume familiarity with the R statistical programming language. Participants should be able to know how to read datasets in R, work with vectors and data frames, and run basic statistical analyses, such as linear regression. More advanced knowledge of statistical computing, such as writing functions and loops, is helpful but not required.

**Schedule**

**June 26, 2016**

| Time | Topic |
|---|---|
| 9.00-9.30 | Introductions and course overview |
| 9.30-10.00 | What is Big Data? The 4 V's of Big Data. Research opportunities and challenges in the social sciences. |
| 10.00-10.45 | What can we learn from web and social media data? Overview of social media research: theories, methods, and data. |
| 10.45-11.15 | Break |
| 11.15-11.45 | Accessing online data: introduction to webscraping. Extracting web data in table format and unstructured format. |
| 11.45-12.30 | Working with APIs (Application Programming Interfaces) |
| 12.30-13.00 | Application: Measuring the tone of media coverage of the economy. |

**June 27, 2016**

| Time | Topic |
|---|---|
| 9.00-10.00 | Introduction to network analysis: basic definitions, network types, centrality measures. |
| 10.00-10.45 | Application: Analyzing the structure of online protest networks |
| 10.45-11.15 | Break |
| 11.15-12.00 | Collecting Twitter data with R: keyword- and location-based filters. |
| 12.00-12.30 | Collecting Twitter data with R: user profiles, network connections. |
| 12.30-13.00 | Application: Studying modularity and polarization of network interaction patterns on Twitter |

**June 28, 2016**

| Time | Topic |
|---|---|
| 9.00-9.30 | Introduction to automated text analysis |
| 9.30-10.15 | Supervised machine learning: large-scale classification of text data |
| 10.15-10.45 | Application: Measuring incivility and hate speech in online communication through social media. |
| 10.45-11.15 | Break |
| 11.15-11.45 | Collecting Facebook data with R: public pages |
| 11.45-12.15 | Unsupervised machine learning: topic discovery in text data |
| 12.15-13.00 | Application: Testing theories of agenda-setting issue ownership using legislators' Facebook posts |

## References

Course largely based on:

### Day 1

Main readings:

Lazer, D., & Radford, J. (2016). Introduction to Big Data. *Annual Review of Sociology*, *43*(1).

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063-1064.

Munzert, S., Rubba, C., Meißner, P., & Nyhuis, D. (2014). *Automated data collection with R: A practical guide to web scraping and text mining*. John Wiley & Sons.

Recommended readings:

Golder, S. A., & Macy, M. W. (2014). Digital footprints: Opportunities and challenges for online social research. *Sociology*, *40*(1), 129.

Nagler, J., & Tucker, J. A. (2015). Drawing inferences and testing theories with big data. *PS: Political Science & Politics*, *48*(01), 84-88.

Grimmer, J. (2015). We are all social scientists now: how big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, *48*(01), 80-83.

Monroe, B. L., Pan, J., Roberts, M. E., Sen, M., & Sinclair, B. (2015). No! Formal theory, causal inference, and big data are not contradictory trends in political science. *PS: Political Science & Politics*, *48*(01), 71-74.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... & Jebara, T. (2009). Life in the network: the coming age of computational social science. *Science (New York, NY)*, *323*(5915), 721.

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063-1064.

Barberá, P., & Rivero, G. (2014). Understanding the political representativeness of Twitter users. *Social Science Computer Review*, 0894439314558836.

### Day 2

Main readings:

Barabási, A. L. (2016). *Network science*. Cambridge University Press.

González-Bailón, S., Borge-Holthoefer, J., & Moreno, Y. (2013). Broadcasters and hidden influentials in online protest diffusion. *American Behavioral Scientist*.

Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting From Left to Right Is Online Political Communication More Than an Echo Chamber?. *Psychological science*.

### Recommended readings:

Newman, M. (2010). *Networks: an introduction*. Oxford University Press.

Barberá, P., Wang, N., Bonneau, R., Jost, J. T., Nagler, J., Tucker, J., & González-Bailón, S. (2015). The critical periphery in the growth of social protests. *PloS one*, *10*(11), e0143611.

Steinert-Threlkeld, Z. C., Mocanu, D., Vespignani, A., & Fowler, J. (2015). Online social networks and offline protest. *EPJ Data Science*, *4*(1), 1.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, *23*(1), 76-91.

Conover, M. D., Gonçalves, B., Flammini, A., & Menczer, F. (2012). Partisan asymmetries in online political activity. *EPJ Data Science*, *1*(1), 1.

## Day 3

### Main readings:

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, mps028.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 6). New York: springer.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

### Recommended readings:

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review*, *97*(02), 311-331.

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2015). Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review*.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S., & Parnet, O. (2016) A Bad Workman Blames His Tweets? The Consequences of Citizens' Uncivil Twitter Use When Interacting with Party Candidates. *Journal of Communication*, forthcoming.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Albertson, B., ... & Rand, D. (2014). Topic models for open ended survey responses with applications to experiments. *American Journal of Political Science*, *58*, 1064-1082.