Social Media Research
EITM Europe Summer Institute
July 5 – July 7, 2018
Course website: www.pablobarbera.com/EITM

Prof. Pablo Barberá
P.Barbera@lse.ac.uk
www.pablobarbera.com

Course description:

Citizens across the globe spend an increasing proportion of their daily lives on social media websites, such as Twitter and Facebook. 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 social media data present unprecedented methodological challenges. The goal of this course is to gain the skills necessary to automate the process of downloading, cleaning, and analyzing social media data using the R programming language for statistical computing.

Logistics:

We will follow a "learning-by-doing" approach, with short guided coding sessions followed by data challenges that will prompt participants to practice what they just learned. Given the applied nature of the course, there will be no required readings, but students are expected to complete and submit the data challenges before the beginning of the second and third sessions.

Note that our first Thursday session will start at 10.30am.

Key background readings:

Klašnja, M., Barberá, P., Beauchamp, N., Nagler, J., & Tucker, J. (2016). Measuring public opinion with social media data. In *The Oxford Handbook of Polling and Survey Methods.*

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

Tucker, J. A., Theocharis, Y., Roberts, M. E., & Barberá, P. (2017). From Liberation to Turmoil: Social Media And Democracy. *Journal of Democracy*, 28(4), 46-59.

Other recommended readings:

Barberá, P. (2014). Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1), 76-91.

Beauchamp, N. (2017). Predicting and Interpolating State‐Level Polls Using Twitter Textual Data. *American Journal of Political Science*, 61(2), 490-503.

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

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24), 8788-8790.

Jäger, K. (2017). The potential of online sampling for studying political activists around the world and across time. *Political Analysis*, 1-15.

King, G., Pan, J., & Roberts, M. E. (2014). Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, 345(6199), 1251722.
Lazer, D. & and Radford, J. (2017). Data ex Machina: Introduction to Big Data. *Annual Review of Sociology.*

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*, 323(5915), 721-3.

Salganik, M. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton, NJ: Princeton University Press.

Steinert-Threlkeld, Z. (2018) *Twitter as Data.* Cambridge University Press.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A. and Parnet, O. (2016), A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates. *Journal of Communication*, 66: 1007–1031.

Tufekci, Z. (2014). Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls. *ICWSM*, 14, 505-514.

Course schedule:

Thursday July 5th, 10:30am-1pm: Social media research. Opportunities and challenges.

The course begins with a discussion of how social media sites represent a new source of data to study human behavior, and an overview of the research opportunities and challenges of using social media data in the social sciences, using examples from published research.

Thursday July 5th, 2.30pm-5pm: Scraping the web.

This session covers basics of webscraping; that is, how to automatically collect data from the web. We will explore the two most common scenarios for webscraping: when data is in table format (e.g. Wikipedia tables or election results) and when it is in an unstructured format (e.g. across multiple parts of a website). The tools available in R to achieve these goals – the *rvest* and *httr* packages – will be introduced in the context of applied examples in the social sciences. As we study different methods to scrape data, we will also learn how to efficiently parallelize loops and work with lists, as two of the most important building blocks of a scalable data collection process.

Friday July 6th, 9.30am-12pm: Collecting data from social media

During this session, we will discuss the data available through Twitter's REST and Streaming API. We will learn how to collect tweets filtered by keywords, location, and language in real time; and how to analyze the data to find the most mentioned hashtags and users and to map the location of the tweets. This session will demonstrate how to collect data from Twitter's REST API, including user profiles and tweets, user networks, recent tweets filtered using keywords, and user lists. We will also go over other social media datasets (from Facebook, reddit, etc.) that are currently available for social science research.

Friday July 6th, 1.30pm-4pm: Topic discovery in social media datasets.

Exploratory data analysis can be a powerful tool for social scientists when they are interested in analyzing a new dataset. This session will cover the existing tools for large-scale discovery in social media data using R, applied to textual and network datasets. We will cover different techniques that allow us to identify salient themes and ideas across documents. Then, we will move to topic models, which allow researchers to automatically identify latent classes of documents in a corpus.

Saturday July 7th, 9.30am-12pm: Querying large-scale datasets using SQL

The volume and heterogeneity of the new datasets available in the digital age present unprecedented opportunities for social scientists, but also new methodological challenges. Computing a simple average for a variable across groups can take minutes when a researcher is working with government records, large-scale survey studies or social media datasets with millions of rows. The goal of this session is to learn how to overcome challenges associated to massive-scale online databases. We will learn the basics of SQL, a language designed to query relational databases that is currently employed by most tech companies; and how to use it from R using the DBI package.

Friday July 6th, 1.30pm-4pm: Big Data analysis using Google BigQuery

To conclude the course, we will learn how to work with large-scale online datasets. From all the available options, we will focus on BigQuery, which relies on Google's infrastructure to efficiently store and query databases at scale. We will learn how to

process, upload, and query databases of up to a billion rows in a matter of seconds, and how to export the results of our queries.