Using Big Data for social science research

Mihály Fazekas
Department of Political Science
Central European University

MA Programme in Political Science

Spring semester 2016-17 (2 credits)
Class meetings : April 10-24
Office hours: By appointment (Október 6. Street 12. building: room 402/a)
Version: 29/3/2017

Introduction

The course is an introduction to state-of-the-art methods to use Big Data in social sciences research. It is a hands-on course requiring students to bring their own research problems and ideas for independent research. The course will review three main topics making Big Data research unique:

1. New and emerging data sources such social media or government administrative data;
2. Innovative data collection techniques such as web scraping; and
3. Data analysis techniques typical of Big Data analysis such as machine learning.

Big Data means that both the speed and frequency of data created are increasing at an accelerating pace virtually covering the full spectrum of social life in ever greater detail. Moreover, much of this data is more and more readily available making real-time data analysis feasible.

During the course students will acquaint themselves with different concepts, methodological approaches, and empirical results revolving around the use of Big Data in social sciences. As this domain of knowledge is rapidly evolving and already vast, the course can only engender basic literacy skills for understanding Big Data and its novel uses. Students will be encouraged to use acquired skills in their own research throughout the course and continue engaging with new methods.

Learning outcomes

Students will be acquainted with basic concepts and methods of Big Data and their use for social sciences research. They will gain first-hand experience with applying such methods to real-life research problems. The acquired knowledge will enable students to use Big Data methods in their individual research on various topics of political science, economics, and sociology.

Teaching format

The course consists of 12 sessions concentrated in a 2-week period between April 10-24. Each session lasts for 100 minutes.

Pre-requisites

Elementary proficiency in quantitative methods and familiarity with statistical softwares, in particular R. Enrolment in MA or PhD course.

Requirements

- Students are required to attend classes regularly, familiarize themselves with each session's reading list and to participate actively in course discussions, in particular providing constructive feedback on other students' presentations.
- Students will pick a data source and research question at the beginning of the course which they will have to regularly work on and report to the class. The methods and approaches learnt in each session will have to be applied to the selected source and research question.
- Students will have to write individual final papers and submit their database and codes which they produced throughout the whole course. The final paper will be short, not longer than 3000 words, describing and critically assessing the data source, data collection method, and analytical tools used in light of the selected research question and relevant prior literature. Great emphasis will be given to the submitted database and annotated codes. Final student project delivery is due 2 weeks after the last session.

Assessment

Attendance and class-room participation 15 %
In-class presentations 40 %
Individual student project & final paper 45%
Final papers will be due on the 30th of April (a week after teaching ends).

Core readings

- Mihály Fazekas (2014), The Use of 'Big Data' for Social Sciences Research: An Application to Corruption Research. SAGE Research Methods Case, see: http://srmo.sagepub.com/view/methods-case-studies-2014/n283.xml?rskey=eFkV0g&row=12
- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer, London. For data and R codes see: http://www-bcf.usc.edu/~gareth/ISL/book.html

Optional introductory reading to R

- Alain F. Zuur, Elena N. Ieno, and Erik H.W.G. Meesters (2009) A Beginner's Guide to R. Springer, London.
- Garrett Grolemund and Hadley Wickham (2016) R for Data Science. O'Reilly Media, Sebastopol, CA. See: http://r4ds.had.co.nz/

Optional advanced reading

- Trevor Hastie, Robert Tibshirani, Jerome Friedman (2013), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, Springer. For data and R codes see: http://statweb.stanford.edu/~tibs/ElemStatLearn/

Course program

Session 1: Introduction

10th of April: Course overview, planning student projects (scoping student interest, selection of topics), introduction to what Big Data means and getting started with R

Easy introductory readings:

- *Dutcher, Jenna. (2014). What is Big Data? UC Berkeley Data Science Blog. See: https://datascience.berkeley.edu/what-is-big-data/*
- *Chris Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine, vol 16 no 7. June 2008. See: https://www.wired.com/2008/06/pb-theory/*
- *Introduction to R:*
  - *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer. Chapter 2.3*

Sessions 2-3: Identifying, understanding, structuring, and critically assessing new data sources

11th of April: Potential data sources and how to assess them (e.g. social media data, government administrative data, internet analytics (e.g. google trends), smartphone data) and getting started with R

- *Mihály Fazekas (2014), The Use of 'Big Data' for Social Sciences Research: An Application to Corruption Research. SAGE Research Methods Case.*
- *Advanced introduction to R:*
  - *Luis Torgo (2011) Data Mining with R: Learning with Case Studies. CRC Press. Chapter 1.*
  - *Atz, Ulrich. (2013). 11 Tips on How to Handle Big Data in R. Open Data Institute Blog. See: https://theodi.org/blog/fig-data-11-tips-how-handle-big-data-r-and-1-bad-pun*

12th of April: Student presentations of selected data sources and research designs

Sessions 4-6: Understanding and using new data collection techniques and assessing their strengths and weaknesses

13th of April: Web scraping, APIs, and parsing I

13th of April: Web scraping, APIs, and parsing II

Combined readings for sessions 4-5:

- *Conceptual overview: Simon Munzert, Christian Rubba, Peter Meissner, Dominic Nyhuis (2015) Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. Wiley. Chapter 9.*
- *Documented practical examples:*

- *http://stat4701.github.io/edav/2015/04/02/rvest_tutorial/ (scraping and parsing)*
- *http://www.columbia.edu/~cjd11/charles_dimaggio/DIRE/styled-4/styled-6/code-13/ (scraping and API)*
- *https://sites.google.com/a/stanford.edu/rcpedia/screen-scraping/web-scraping-with-r (scraping and parsing)*
- *http://bogdanrau.com/blog/collecting-tweets-using-r-and-the-twitter-search-api/ (API)*

Further readings for sessions 4-5:

- *Challenges of "found data" – methods to process data originally collected for other purposes:*
  - *Karimi, Fariba, et al. "Inferring gender from names on the web: A comparative evaluation of gender detection methods." Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016.*
  - *Inferring gender and race from facial image data: Face++. https://github.com/FacePlusPlus/detect-demo*

18th of April: Student presentation of data collection results and data clinic

Sessions 7-10: Data analytic techniques

18th of April: Evaluation and validation: predictive power, cross-validation, assessing statistical significance and resampling methods

- *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer. Chapter 2.2 & 5.1*
- *Phillip I. Good (2006) Resampling Methods. A Practical Guide to Data Analysis. 3rd edition, Birkhauser, Boston. Chapter 3.*

19th of April: Supervised learning: regression methods and their variants

- *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer.*
  - *Introduction: Ch 2.1*
  - *Technical details and codes: Ch. 6-7 (in particular ridge, lasso, regression splines)*

20th of April: Supervised learning: decision trees and random forests

- *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer. Chapter 8.*

20th of April: Unsupervised learning: Introduction to clustering and text mining (main empirical examples from text mining)

- *Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2015) An Introduction to Statistical Learning: With Applications in R. 6th edition, Springer. Chapter 10.*

- *Justin Grimmer and Brandon M. Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, Vol. 21, No. 3, pp 267-297.*

Sessions 11-12: Innovative applications and discussion of student projects

21st of April: Inspiring and cautionary examples (e.g. Google flu prediction and its failure)

Diverse inspiring readings(subject to change depending on student interest):

- *Ginsberg et al. 2009. Detecting influenza epidemics using search engine query data. Nature.*
- *Lazer et al. 2014. The Parable of Google Flu: Traps in Big Data Analysis. Science*
- *Roberts et al. 2014. Structural topic models for open-ended survey responses. American Journal of Political Science*
- *Toole, Jameson L., et al. "Tracking employment shocks using mobile phone data." Journal of The Royal Society Interface 12.107 (2015): 20150185.*
- *Hannak, Aniko, et al. "Measuring price discrimination and steering on e-commerce web sites."* Proceedings of the 2014 conference on internet measurement conference. *ACM, 2014.*

24th of April: Student projects' final presentation and discussion