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 as 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


Optional introductory reading to R


Optional advanced reading

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:
- Introduction to R:

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

- Advanced introduction to R:

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:
- Documented practical examples:
Further readings for sessions 4-5:

- Challenges of “found data” – methods to process data originally collected for other purposes:
  - 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


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

  - 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


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

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):


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