Using Big Data for Public Policy

Mihály Fazekas
Department of Public Policy
Central European University
Winter semester 2025-26 (2 credits)
Class times: 15.40-17.20, Tuesday (weekly)
Office hours: By appointment
Teaching Assistant: Zdravko Veljanov

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, one each week. Each session lasts for 100 minutes.

Pre-requisites

- Elementary proficiency in quantitative methods and familiarity with R. Please send an example R script to demonstrate meeting this requirement at least one week before the first session.
- 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.

Assessment

Criteria	Weight
Attendance and class-room participation	10%
Data abstract	15%
In-class presentation-1 (data and codes also to be submitted)	20%
In-class presentation-2 (data and codes also to be submitted)	20%
Final paper* (data and codes also to be submitted)	35%

* example paper and grading criteria will be shared

All coursework that you submit during your studies has to be in line with common standards of academic integrity (as outlined here and here). You may use AI tools, such as ChatGPT, to support you in achieving the learning goals defined for the course, for example assisting in code writing, bug fixing, or checking grammatical errors. However, maintaining academic integrity is essential, which is why submitting text generated by an AI is prohibited.

Deadlines

Data abstract due on the 18th of January 2025

Final papers due on the 7th of April 2025 (Each week of delay will result in a reduction of the final grade by one 'step', for example from a B+ to a B, then from a B to a B- etc.)

Core reading

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2021) An Introduction to Statistical Learning: With Applications in R. 2nd edition, Springer, London. For data and R codes see: https://www.statlearning.com/ [**ISL henceforth**]

Optional introductory readings to R

- Robert I. Kabacoff (2022) R in Action. Data Analysis and Graphics with R. Manning Publications, New York (3rd edition): https://www.manning.com/books/r-in-action-third-edition
- Garrett Grolemund and Hadley Wickham (2026) R for Data Science. O'Reilly Media, Sebastopol, CA. See: https://r4ds.hadley.nz/ (2nd edition)
- Garrett Grolemund (2026) Hands-On Programming with R. See: https://rstudio-education.github.io/hopr/

Optional advanced readings

- Felix Chan and László Mátyás (2022) Econometrics with Machine Learning. Springer, London.
- Martin Huber (2023) Causal Analysis: Impact Evaluation and Causal Machine Learning with Applications in R. MIT Press.
- Mutlu Yuksel and Yigit Aydede (2026) Causal Inference and Machine Learning: In Economics, Social, and Health Sciences. Routledge, See: https://www.causalmlbook.com/
- Christoph Molnar (2024) Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. LeanPub Books, see: https://christophm.github.io/interpretable-ml-book/
- Christen, Peter (2012) Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, London.
- Van Atteveldt, W., Trilling, D., & Calderón, C. A. (2022). Computational Analysis of Communication. Wiley Blackwell. See: https://cssbook.net/

Course Schedule

#	Date	Topic
1	9/1/2025	Introduction
2	13/1/2025	Potential data sources and how to assess them
3	20/1/2025	Visual arguments
4	27/1/2025	Web scraping, APIs, and parsing I
5	30/1/2025	Web scraping, APIs, and parsing II
6	10/2/2025	Model evaluation and significance testing
7	17/2/2025	Student presentations: data collection
8	24/2/2025	Unsupervised learning: introduction to clustering and text mining
9	3/3/2025	Text mining continued, data matching, and deduplication
10	10/3/2025	Supervised learning: overview and tree-based methods
11	17/3/2025	Supervised learning: SVM and Neural Networks
12	24/3/2025	Student presentations: analytical results

Detailed course program

Session 1: Introduction

Session 1: 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:*
  - *ISL. Chapter 2.3*

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

Session 2: 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.*
  Short videos on the paper:
  http://methods.sagepub.com/video/srmpromo/0Vt2p3/introduction-to-big-data-for-social-science-research and
  http://methods.sagepub.com/video/srmpromo/WHQehe/using-big-data-to-measure-formidable-concepts-the-case-of-government-contra

Further readings

- *Advanced introduction to R:*
  - *Robert Kabacoff (2022) R in Action, Third Edition: Data Analysis and Graphics with R and Tidyverse*

Session 3: Visual arguments: principles of good data visualisation, data visualisation practice using R and Tableau

- *Edward Tufte (2001) The Visual Display of Quantitative Information. 2nd edition, Graphics Press. Chapter 2.*
- *R Shiny introduction (Start Your first Shiny app): http://shiny.rstudio.com/articles/#first-app*
- *Tableau introductory video (1. Tableau Public Overview): https://public.tableau.com/en-us/s/resources*

Further readings

- *Alberto Cairo (2019) How Charts Lie: Getting Smarter About Visual Information. W W Norton*

- *Edward Tufte* (2006), Beautiful Evidence, Cheshire, CT: Graphics Press
- Cole Nussbaumer Knaflic (2015), Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley

Sessions 4-5: Understanding and using new data collection and management techniques and assessing their strengths and weaknesses

Session 4: Web scraping, APIs, and parsing I

Session 5: 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/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)*
  - *https://blog.predictiveheuristics.com/2014/10/28/a-primer-on-web-scraping-with-r/*
  - *https://github.com/pablobarbera/social-media-workshop*

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.*
  - *Rieder, B. (2013, May). Studying Facebook via data extraction: the Netvizz application. In Proceedings of the 5th annual ACM web science conference (pp. 346-355). ACM.*
  - *Easy intro to constructing APIs: https://www.storybench.org/how-to-access-apis-in-r/*

Sessions 6-12: Data analytic techniques

Session 6: Model evaluation and significance testing

- *ISL, Chapter 2, 5.1*
- *Phillip I. Good (2006) Resampling Methods. A Practical Guide to Data Analysis. 3rd edition, Birkhauser, Boston. Chapter 3.*

Session 7: Student presentation of data collection results and data clinic

Session 8: Unsupervised learning: Introduction to clustering and text mining (main empirical examples from text mining)

- *ISL. 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.*

Further useful readings, compendium of data and software packages

- *Gary E. Hollibaugh (2019) The Use of Text as Data Methods in Public Administration: A Review and an Application to Agency Priorities. JPART, Vol. 29, No. 3. Pp. 474–490*
- *Justin Grimmer, Margaret E. Roberts, Brandon M. Stewart (2022) Text as Data: A New Framework for Machine Learning and the Social Sciences. Princeton University Press.*
- *https://quanttext.com/*

Session 9: Text mining continued, data matching, and deduplication

- *Roberts et al. 2014. Structural topic models for open-ended survey responses. American Journal of Political Science*
- *Dusetzina SB, Tyree S, Meyer AM, et al. (2014) Linking Data for Health Services Research. A Framework and Instructional Guide. Rockville (MD): Agency for Healthcare Research and Quality (US). Ch. 4.: https://www.ncbi.nlm.nih.gov/books/NBK253312/*

Further readings

- *Eshima, S., Imai, K. and Sasaki, T. (2024), Keyword-Assisted Topic Models. American Journal of Political Science, 68: 730-750.*
- *John Wilkerson and Andreu Casas (2017), Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges. Annu. Rev. Polit. Sci. 2017. 20:529–44*
- *https://cran.r-project.org/web/packages/RecordLinkage/RecordLinkage.pdf*
- *https://www.r-bloggers.com/fuzzy-string-matching-a-survival-skill-to-tackle-unstructured-information/*
- *Christen, Peter (2012) Data Matching. Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, London.*

Session 10: Supervised learning: Overview and tree-based methods

- *ISL. Chapter 2.*
- *ISL. Chapter 8.*

Further readings

- *Ginsberg et al. 2009. Detecting influenza epidemics using search engine query data. Nature.*

Session 11: Supervised learning: SVM and Neural Networks

- *ISL. Chapter 9 – SVM*
- *ISL. Chapter 10 – Neural networks*

Further readings:

SVM

- *Schölkopf, Bernhard, and Alex Smola. "Support vector machines." Encyclopedia of Biostatistics (1998): 978-0471975762. http://users.ece.northwestern.edu/~yingwu/teaching/EECS432/Reading/heaschdumosupla98.pdf*
- *Tian, Y., Shi, Y., & Liu, X. (2012). Recent advances on support vector machines research. Technological and Economic Development of Economy, 18(1), 5–33. https://www.tandfonline.com/doi/epdf/10.3846/20294913.2012.661205?needAccess=true*

Neural networks with R

- *Collection of resources: https://github.com/rdrr1990/kerasformula/blob/master/short_course/APSA_readme.md*
- *On neural networks - Wordliczek, Lukasz. "Neural networks and political science: Testing the methodological frontiers." EMPIRIA. Revista de Metodología de las Ciencias Sociales 57 (2023): 37-62. https://www.redalyc.org/journal/2971/297176235002/html/*
- *Trevor Hastie, Robert Tibshirani, Jerome Friedman (2013), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, Springer. Chapter 11*
- *François Chollet and JJ Allaire. Deep Learning with R. Manning Publications Co., 2018*

Session 12: Student projects' final presentation and discussion