Data Handling: Import, Cleaning and Visualisation

Prof. Dr. Ulrich Matter

20/09/2018

Prerequisites

None. This course should be taken in parallel with Statistics (3,222).

Course Content

Short summary

This course introduces students to the fundamental practices of Data Science in the context of economic research. The course covers basic theoretical concepts and practical skills in gathering, preparing/cleaning, visualizing, storing, and analyzing digital data for research purposes.

Description

The increasing abundance of digital data covering every-day human activities offers opportunities and poses challenges for empirical research in economics and more broadly in the social sciences at large. Data used in economics come more and more often from novel digital sources (e.g, social media, web applications, or sensors), in diverse formats (e.g., JSON, unstructured text), and in large quantities. In order to effectively and efficiently engage with these developments, economists need a basic understanding of data technologies and practical skills in working with digital data.

This course covers basic theoretical concepts and practical skills in (automatically) gathering, preparing, visualizing, and storing digital data for research purposes. It thus covers the crucial first steps underlying empirical research projects. These steps are often rather neglected in traditional social science methodology but are of great relevance in the age of Big Data; this course aims to fill this gap and thereby aims to exploit synergies with other methodology courses such as: Statistics and Empirical Economic Research. Hands-on exercises and case studies from current real-world research projects are meant to deepen the taught concepts and train students in the basics of programming with data.

The course covers both theoretical concepts in handling digital data as well as practical hands-on exercises focusing on different data structures and data formats (CSV, HTML, JSON). All exercises are based on freely available open-source-tools (R, RStudio, Atom). Students are expected to install these tools and work with them on their own machines. In the first part of the course, students learn about the relevance and challenges of Big Data for research in economics and related fields, by introducing students to basic data formats and how their use in every-day life has evolved in recent years (with a particular focus on the spread of the Internet and online data). Based on this, the second part of the course introduces concepts and practices to gather and prepare digital data from various sources. In this part, students acquire basic programming skills with R in order to apply these practices with real-world datasets. The last part of the course focuses on analysis and visualization as well as storage and documentation of (relatively) large data sets and discusses the implications of the contents covered in the course for econometric research and applied data science.

The structure of the course offers the opportunity to invite guest speakers (in the second and third part of the course) who can give insights into social science research with Big Data and/or applied Data Science in the industry.

Course Goals

The main goal of the course is to enable students to handle digital data for analysis/research purposes in economics (with a particular focus on unusual and large data sets from various sources). Students get familiar with best practices to gather, clean, and store digital data for research purposes. They are capable of planning and managing the first steps of an empirical research project based on digital data, preceding the actual econometric analyses. Finally, students acquire basic programming skills with R in the context of real-world data sets.

Course Objectives

- Students will know the basic concepts of data technologies/data structures.
- Students will understand the basics of computer code and data storage.
- Students will know how to apply the relevant R packages and programming practices to effectively and efficiently parse, filter, clean, and store digital data from various sources.

Course Structure

Lectures: 2-4 hours per week throughout the autumn semester; 4 credits. Notes on the literature: mandatory texts (M) as well as compulsory texts (C) are indicated for each lecture. Mandatory texts are mainly covered in Murrell (2009) and Wickham and Grolemund (2017) and indicated with "Mu" or "WiGr", respectively. Additional literature is indicated with references to the bibliography. The course is structured as follows:

Part I: Data fundamentals

Date	From	To	Room	Topic	Literature
20.09.2018	10:15	12:00	01-U203	Introduction: Big Data/Data Science, course overview	M: Mu Preface. C: Einav and Liran (2014), Lazar et al. (2009)
27.09.2018	10:15	12:00	01-U203	An introduction to data and data processing	M: Mu Chapter 5.1-5.4
27.09.2018	16:15	18:00	01-U121	Exercises/Workshop 1: Tools, working with text files	NA
04.10.2018	10:15	12:00	01-U203	Data storage and data structures	M: Mu Chapter 2; WiGr chapter 1
11.10.2018	10:15	12:00	01-U203	'Big Data' from the Web	M: Mu Chapter 5.5
11.10.2018	16:15	18:00	01-U121	Exercises/Workshop 2: Computer code and data storage	NA

Part II: Data gathering and data preparation

Date	From	To	Room	Topic	Literature
18.10.2018	10:15	12:00	01-U203	Programming with data	M: Mu Chapter 9.11; WiGr chapter 4 and 6
25.10.2018	10:15	12:00	01-U203	Data sources, data gathering, data import	M: Mu Chapter 9.7; WiGr Chapter 11
25.10.2018	16:15	18:00	01-U121	Working with semi-structured and unstructured data	M: Mu Chapter 9.9
15.11.2018	10:15	12:00	01-U203	Guest Lecture: Dr. Michael Zehnder (Swiss Data Labs, gateB)	NA
22.11.2018	10:15	12:00	01-U203	Data preparation and manipulation	M: Mu Chapter 9.8; WiGr Chapter 5
22.11.2018	16:15	18:00	01-U121	Exercises/Workshop 4: Data import and data preparation/manipulation	NA
29.11.2018	10:15	12:00	01-U203	Case Study: The Programmable Web, Big Public Data, and Political economics	M: Matter and Stutzer (2015)

Part III: Visualization, and storage of large datasets

Date	From	To	Room	Topic	Literature
06.12.2018	10:15	12:00	01-U203	Basic statistics with R	NA
06.12.2018	16:15	18:00	01-U121	Exercises/Workshop 5: Applied data analysis with R	NA
13.12.2018	10:15	12:00	01-U203	Visualization, dynamic documents	M: Mu Chapter 9.10; C: Wickham (2016)
20.12.2018	10:15	12:00	01-U123	Exercises/Workshop 6: Visualization, dynamic documents	NA
20.12.2018	16:15	18:00	01-U203	Wrap-Up, Q&A	NA

Exam Information

Central - Written examination

Exam Information: Content

The written examination consists of different types of multiple-choice questions, covering both the theoretical concepts and practical applications in R (questions based on code examples).

Literature

The course's main textbooks are „Introduction to Data Technologies“ by Paul Murrel (more about the book and a free pdf version can be found here), and the book “R for Data Science” by Hadley Wickham and Garred Grolemund. Current versions of these books as well as additional material like data examples and R-scripts are freely available online.

Main textbooks

Murrell, Paul (2009). *Introduction to Data Technologies*, London: Chapman & Hall/CRC.

Wickham, Hadley and Garred Grolemund (2017). *R for Data Science*, 1st Edition. Sebastopol, CA: O'Reilly.

Journal articles

Lazer, David, Pentland, Alex, Adamic, Lada, Aral, Sinan, Barabási, Albert-László, Brewer, Devon and Christakis, Nicholas, Contractor, Noshir, Fowler, James, Gutmann, Myron, Jebara, Tony, King, Gary, Macy, Michael, Roy, Deb and Van Alstyne, Marshall. (2009). Computational Social Science. *Science*, 323(5915):721-723.

Einav, Lirand and Levin Jonathan. (2014). Economics in the Age of Big Data. *Science*. 346(6210).

Matter, Ulrich and Stutzer, Alois (2015). pvsR: An Open Source Interface to Big Data on the American Political Sphere. *PLoS ONE* 10(7): e0130501.