University of St.Gallen

Course and Examination Fact Sheet: Autumn Semester 2018

5,244: Data Handling: Import, Cleaning and Visualisation

ECTS credits: 4

Overview examination/s

(binding regulations see below)
Central - Written examination (100%, 90 mins.)

Attached courses

Timetable -- Language -- Lecturer
5,244,1.00 Data Handling: Import, Cleaning and Visualisation -- Englisch -- Matter Ulrich

Course information

Course 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 for the social sciences at large. Data used in economic research (as well as in market research, business analytics, etc.) comes 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 aspects of what digital data are and how to handle them, 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 (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.

Part I: Data fundamentals

1. Introduction: Big Data/Data Science, course overview
2. An introduction to data and data processing
3. *Exercises/Workshop 1: Tools, working with text files*
4. Data storage and data structures
5. 'Big Data' from the Web
6. *Exercises/Workshop 2: Computer code and data storage*

Part II: Data gathering and data preparation

1. Programming with data
2. Data sources, data gathering, data import
3. *Exercises/Workshop 3: Programming with data*
4. Working with semi-structured and unstructured data
5. Data preparation and manipulation
6. *Exercises/Workshop 4: Data import and data preparation/manipulation*
7. Case Study: The Programmable Web, Big Public Data, and Political Economics

Part III: Analysis and visualization

1. Understanding basic statistics with R
2. *Exercises/Workshop 5: Applied data analysis with R*
3. Visualization, dynamic documents
4. *Exercises/Workshop 6: Visualization, dynamic documents*
5. Wrap-Up, Q&A

Course literature

The course's main textbooks are "Introduction to Data Technologies" by Paul Murrel (https://www.crcpress.com/Introduction-to-Data-Technologies/Murrell/p/book/9781420065176) (more about the book and a free pdf version can be found here: https://www.stat.auckland.ac.nz/~paul/ItDT/), and the book "R for Data Science" by Hadley Wickham and Garred Grolemund (http://r4ds.had.co.nz/). Current versions of these books as well as additional material like data examples and R-scripts are freely available online.

University of St.Gallen

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.

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

Additional course information

--

Examination information

Examination sub part/s

1. Examination sub part (1/1)

Examination time and form

Central - Written examination (100%, 90 mins.)

Remark

--

Examination-aid rule

Extended Closed Book
The use of aids is limited; any additional aids permitted are exhaustively listed under "Supplementary aids". Basically, the following is applicable:

- At such examinations, all the pocket calculators of the Texas Instruments TI-30 series and mono- or bilingual dictionaries (no subject-specific dictionaries) without hand-written notes are admissible. Any other pocket calculator models and any electronic dictionaries are inadmissible.
- In addition, any type of communication, as well as any electronic devices that can be programmed and are capable of communication such as notebooks, tablets, PDAs, mobile telephones and others, are inadmissible.
- Students are themselves responsible for the procurement of examination aids.

Supplementary aids

None

Examination languages

Question language: English
Answer language: English

Examination 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).

Examination relevant literature

University of St.Gallen

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.

Please note

We would like to point out to you that this fact sheet has absolute priority over other information such as StudyNet, faculty members' personal databases, information provided in lectures, etc.When will the fact sheets become binding?

- Information about courses and examination time (central/decentral and grading form): from the start of the bidding process on 23 August 2018
- Information about decentral examinations (examination-aid rule, examination content, examination relevant literature): after the 4th semester week on 15 October 2018
- Information about central examinations (examination-aid rule, examination content, examination relevant literature): from the start of the enrolment period for the examinations on 05 November 2018

Please look at the fact sheet once more after these deadlines have expired.