University of St.Gallen

Course and Examination Fact Sheet: Autumn Semester 2020

3,230: Data Handling: Import, Cleaning and Visualisation

ECTS credits: 4

Overview examination/s

(binding regulations see below)
Central - Written examination (100%, 90 mins.)
Examination time: inter-term break

Attached courses

Timetable -- Language -- Lecturer
3,230,1.00 Data Handling: Import, Cleaning and Visualisation -- Englisch -- Matter Ulrich
3,230,2.00 Data Handling: Import, Cleaning and Visualisation: Exercises -- Englisch -- Widmer Philine

Course information

Course prerequisites

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

Learning 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 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 concepts taught in this course 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

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 raw data sets from various sources in various formats). Students get familiar with best practices to gather, clean, and store digital data for research purposes. They understand the concept of a data pipeline in the context of academic research and 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 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. *Exercises/Workshop 3: Programming with data*
3. Data sources, data gathering, data import
4. Working with semi-structured and unstructured data
5. Data preparation and manipulation
6. *Exercises/Workshop 4: Data import and data preparation/manipulation*
7. Guest lecture or research insights

Part III: Analysis and visualization

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

NOTE ON LECTURES: Given the current COVID-19 situation, lectures will take place in person but the number of students attending the lectures in person might be restricted (division into alternating groups). All students accepted to the course will be informed about the specifics of the attendance in due time. In any case, the lectures will also be broadcasted via StudyNet/Zoom.

NOTE ON EXERCISE/WORKSHOP SESSIONS: Due to the current COVID-19 situation, Exercise/Workshop sessions will be held virtually via Zoom (Detailed step-by-step explanations of the solutions, as well as Q&A sessions will be offerend).

GENERAL NOTE ON HYBRID-FORMAT OF THIS COURSE: Detailed lecture notes for each lecture as well as code and data examples will be made available throughout the course. Our aim is to facilitate a well-guided learning experience with online learning material closely fitting the structure of this course.

University of St.Gallen

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.

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

In case of a new COVID-related lockdown, all the lectures will be given online via StudyNet/Zoom. Given the planed hybrid-format (see notes on structure above), we expect this to be a rather smooth transition. Online lectures will include additional time for Q&A at the end of each session (in order to compensate for the exchanges with students during lecture breaks/at the end of lectures).

Examination information

Examination sub part/s

1. Examination sub part (1/1)

Examination time and form
Central - Written examination (100%, 90 mins.)
Examination time: inter-term break

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, mobile telephones and others, are inadmissible.

- Students are themselves responsible for the procurement of examination aids.

Supplementary aids
No additional aids are permitted.

Examination languages
Question language: English
Answer language: English

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

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

Please note

Please note that only this fact sheet and the examination schedule published at the time of bidding are is binding and takes precedence over other information, such as information on StudyNet (Canvas), on lecturers' websites and information in lectures etc.

Any references and links to third-party content within the fact sheet are only of a supplementary, informative nature and lie outside the area of responsibility of the University of St.Gallen.

Documents and materials are only relevant for central examinations if they are available by the end of the lecture period (CW51) at the latest. In the case of centrally organised mid-term examinations, the documents and materials up to CW 42 are relevant for testing.

Binding nature of the fact sheets:

- Course information as well as examination date (organised centrally/decentrally) and form of examination: from bidding start in CW 34 (Thursday, 20 August 2020);
- Examination information (regulations on aids, examination contents, examination literature) for decentralised examinations: in CW 42 (Monday, 12 October 2020);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised mid-term examinations: in CW 42 (Monday, 12 October 2020);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised examinations: two weeks before the end of the registration period in CW 44 (Thursday, 29 October 2020).