University of St.Gallen

Course and Examination Fact Sheet: Spring Semester 2022

8,272: Big Data Analytics

ECTS credits: 4

Overview examination/s

(binding regulations see below)
Decentral - examination paper written at home (in groups - all given the same grades) (60%)
Examination time: term time
Decentral - Presentation (in groups - all given the same grades) (40%)
Examination time: term time

Attached courses

Timetable -- Language -- Lecturer
8,272,1.00 Big Data Analytics -- Englisch -- Matter Ulrich

Course information

Course prerequisites

- 'A Brief Introduction to Programming with R' in the integration week (or an equivalent basic R programming course).
- 7,310: 'Data Analytics I: Predictive Econometrics'

Learning objectives

- Students will know the concept of Big Data in the context of empirical economic research.
- Students will understand the technical challenges of Big Data Analytics and how to practically deal with them.
- Students will know how to apply the relevant R packages and programming practices to effectively and efficiently handle large data sets.

Course content

Short summary

This course introduces students to the concept of Big Data in the context of empirical economic research. Students learn about the computational constraints underlying Big Data Analytics and how to handle them in the statistical computing environment R (local and in the cloud). Revisiting basic statistical/econometric concepts, we look at each step of dealing with large data sets in empirical economic research (storage/import, transformation, visualization, aggregation).

Description

The increasing size of datasets in empirical economic research (both in number of observations and number of variables) offers new opportunities and poses new challenges for economists. 'Big Data' is discussed as the new 'most valuable' resource in highly developed economies, driving the development of new products and services in various industries. Extracting knowledge from large data sets is increasingly seen as a strategic asset for firms, governments, and NGOs. Successfully navigating the data-driven economy presupposes a certain understanding of the technologies and methods used to gain insights from Big Data.

This course introduces students to the basic concepts of Big Data Analytics to gain insights from large and complex data sets. Thereby, the focus of the course is on the practical application of econometrics/machine learning, given large/complex datasets. The course does NOT (or only to a very limeted degree) introduce basic econometric/machine learning concepts/models. It is, therefore, crucial that students taking this course are already equiped with solid knowledge in statistics/econometrics (and basic knowledge in machine learning). The course combines conceptual/theoretical material with the practical application of the concepts with the open source programming language R. Thereby, students will acquire the basic skillset of analysing large data sets both locally and in the cloud. The practical applications of the learned techniques are focused on empirical research in economics and the social sciences.

The first part of the course covers the basics of computation (in an applied econometrics context). Students learn about the physical constraints of standard computers used for data analytics and learn how to identify bottle-necks in data analysis tasks and how to identify them within the R environment. Students then learn how to handle the identified computational constraints with R (and related tools such as Keras and Spark), first locally and then in the cloud. Thereby, the course covers each step of the data pipeline in economic research (storage/import, transformation, visualization, aggregation, model estimation).

Course structure and indications of the learning and teaching design

Due to the ongoing Sars-Cov-2 pandemic, this course is planned to be taught **fully online**.

The course will be conducted in form of online lectures (including extended Q&A): 2 hours per week throughout the spring semester.

During the course, students will apply the relevant R packages and programming practices to effectively and efficiently handle large data sets. After guided application of concepts in R, students are required to apply concepts in R independently in groupsas well as presenting their approach/strategy and results in class.

- The course proceeds as follows:

Part I: The Basics

- Introduction: Big Data, Data Economy, Course Overview Walkowiak (2016): Chapter 1
- Computation and Memory in Applied Econometrics
- Advanced R Programming (Concepts/Applied) Wickham (2019): Chapters 2, 3, 17,23, 24.

Part II: Local Big Data Analytics

- Import, Cleaning and Transformation of Big Data (Applied) Walkowiak (2016): Chapter 3: p. 74-118.
- Aggregation and Visualization (Applied: data tables, ggplot) Walkowiak (2016): Chapter 3: p. 118-127; Wickham et al. (2015); Schwabish (2014).
- Data Storage, Databases Interaction with R Walkowiak (2016): Chapter 5.

Part III: Advanced Topics

- Cloud Computing: Introduction/Overview (Concepts)
- Machine Learning and GPUs
- Distributed Systems, MapReduce/Hadoop with R (Concepts/Applied)Walkowiak (2016): Chapter 4.
- Applied Econometrics with Spark

Course literature

- **Main textbooks**

  Walkowiak, Simon (2016): *Big Data Analytics with R.* Birmingham, UK: Packt Publishing.

  Wickham, Hadley (2019): *Advanced R*. Second Edition, CRC Press, FL: Boca Raton.

  Journal articles and additional books

  Wickham, Hadley and Dianne Cook and Heike Hofmann (2015): Visualizing statistical models: Removing the blindfold. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 8(4):203-225.

  Schwabish, Jonathan A. (2014): An Economist's Guide to Visualizing Data. *Journal of Economic Perspectives*. 28(1):209-234.

Additional course information

In the case of the President's Board having to implement new directives due to the SARS-CoV-2 pandemic in SpS2022, the course information listed above will be changed as follows:

- The course is conducted online via the platform Zoom;
- The recordings of lectures in this course are available for 30 days;
- The lecturer informs via StudyNet on the changed implementation modalities of the course.

The examination information listed below would be changed as follows:

- There are no changes necessary to the examination information.

Examination information

Examination sub part/s

1. Examination sub part (1/2)

Examination time and form

Decentral - examination paper written at home (in groups - all given the same grades) (60%)
Examination time: term time

Remark

Take-home exercise-set. Groups of 2-3.

Examination-aid rule

Term papers

Written work must be written without outside help according to the known citation standards, and a declaration of authorship must be attached, which is available as a template on the StudentWeb.

Documentation (quotations, bibliography, etc.) must be carried out universally and consistently according to the requirements of the chosen/specified citation standard such as e.g. APA or MLA.

The legal standard is recommended for legal work (cf. by way of example: FORSTMOSER, P., OGOREK R., SCHINDLER B., Juristisches Arbeiten: Eine Anleitung für Studierende (the latest edition in each case), or according to the recommendations of the Law School).

The reference sources of information (paraphrases, quotations, etc.) that has been taken over literally or in the sense of the original text must be integrated into the text in accordance with the requirements of the citation standard used. Informative and bibliographical notes must be included as footnotes (recommendations and standards e.g. in METZGER, C., Lern- und Arbeitsstrategien (latest edition)).

For all written work at the University of St.Gallen, the indication of page numbers is mandatory, regardless of the standard chosen. Where page numbers are missing in sources, the precise designation must be made differently: chapter or section title, section number, article, etc.

Supplementary aids

--

Examination languages

Question language: English
Answer language: English

---

2. Examination sub part (2/2)

Examination time and form

Decentral - Presentation (in groups - all given the same grades) (40%)
Examination time: term time

Remark
Presentation of analytics project. Teams of 2-3.

Examination-aid rule
Practical examination
No examination-aid rule is necessary for such examination types. The rules and regulations of the University of St. Gallen apply in a subsidiary fashion.

Supplementary aids
--

Examination languages
Question language: English
Answer language: English

Examination content

- 60% - Take-home exercises as part of the students project, solved in teams of 3-4 students: Technical report on the project. Conceptual questions related to the application.
- 40% - Students project presentation (same teams of 3-4 students): Own application of concepts in R, approach/strategy and results presented in class (or online via Zoom, if the pandemic circumstances require it).

Examination relevant literature

Main textbooks

Walkowiak, Simon (2016): *Big Data Analytics with R.* Birmingham, UK: Packt Publishing.

Wickham, Hadley (2019): *Advanced R*. Second Edition, CRC Press, FL: Boca Raton.

Journal articles and additional books

Wickham, Hadley and Dianne Cook and Heike Hofmann (2015): Visualizing statistical models: Removing the blindfold. *Statistical Analysis and Data Mining: The ASA Data Science Journal*. 8(4):203-225.

Schwabish, Jonathan A. (2014): An Economist's Guide to Visualizing Data. *Journal of Economic Perspectives*. 28(1):209-234.

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 (CW21) at the latest. In the case of centrally organised mid-term examinations, the documents and materials up to CW 12 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 04 (Thursday, 27 January 2022);
- Examination information (regulations on aids, examination contents, examination literature) for decentralised examinations: in CW 12 (Monday, 21 March 2022);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised mid-term examinations: in CW 12 (Monday, 21 March 2022);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised examinations: two weeks before the end of the registration period in CW 15 (Monday, 11 April 2022).