



Course and Examination Fact Sheet: Spring Semester 2020

8,272: Big Data Analytics

ECTS credits: 4

Overview examination/s

(binding regulations see below)

Decentral - Group examination paper (all given the same grades) (60%)

Decentral - Presentation (in groups - all given the same grades) (40%)

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 limited degree) introduce basic econometric/machine learning concepts/models. It is, therefore, crucial that students taking this course are already equipped 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

Lectures: 2 hours per week throughout the spring semester.

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

--

Examination information

Examination sub part/s



1. Examination sub part (1/2)

Examination time and form

Decentral - Group examination paper (all given the same grades) (60%)

Remark

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

Examination-aid rule

Term papers

- Term papers must be written without anyone else's help and in accordance with the known quotation standards, and they must contain a declaration of authorship.
- The documentation of sources (quotations, bibliography) has to be done throughout and consistently in accordance with the APA or MLA standards. The indications of the sources of information taken over verbatim or in paraphrase (quotations) must be integrated into the text in accordance with the precepts of the applicable quotation standard, while informative and bibliographical notes must be added as footnotes (recommendations and standards can be found, for example, in METZGER, C. (2017), Lern- und Arbeitsstrategien (12th ed., Cornelsen Schweiz).
- For any work written at the HSG, the indication of the page numbers both according to the MLA and the APA standard is never optional.
- Where there are no page numbers in sources, precise references must be provided in a different way: titles of chapters or sections, section numbers, acts, scenes, verses, etc.
- For papers in law, the legal standard is recommended (by way of example, cf. FORSTMOSER, P., OGOREK R. et SCHINDLER B. (2018, Juristisches Arbeiten: Eine Anleitung für Studierende (6. Auflage), Zürich: Schulthess, or the recommendations of the Law School).

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%)

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

- Take-home exercises solved in teams of 2-3 students: Guided application of concepts in R. Conceptual questions related to the application.
- Analytics project (same teams of 2-3 students): Own application of concepts in R, approach/strategy and results presented



in class.

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 this fact sheet alone is binding and has priority over any other information such as StudyNet (Canvas), personal databases or faculty members' websites and information provided in their lectures, etc.

Any possible references and links within the fact sheet to information provided by third parties are merely supplementary and informative in nature and are outside the University of St.Gallen's scope of responsibility and guarantee.

Documents and materials that have been submitted no later than the end of term time (CW21) are relevant to central examinations.

Binding nature of the fact sheet:

- Information about courses and examination time (central/decentral) and examination type starting from the beginning of the bidding on 23 January 2020
- Information about examinations (examination aid regulations, examination content, examination-relevant literature) for decentral examinations after the 4th semester week on 16 March 2020
- Information about examinations (examination aid regulations, examination content, examination-relevant literature) for central examinations as from the starting date for examination registration on 6 April 2020

Please consult the fact sheet again after these deadlines have expired.