GHENT UNIVERSITY

Course Specifications

Valid as from the academic year 2024-2025

Data Science (I002440)

Course size *(nominal values; actual values may depend on programme)*

Credits 5.0 **Study time 150 h**

Course offerings and teaching methods in academic year 2024-2025

A (semester 2)	Dutch	Gent	group work
			seminar
			lecture

Lecturers in academic year 2024-2025

Meys, Joris	LA26	staff member
Verwaeren, Jan	LA26	lecturer-in-charge

Offered in the following programmes in 2024-2025	**crdts**	**offering**
Bachelor of Science in Bioscience Engineering	5	A
Master of Science in Bioinformatics(main subject Bioscience Engineering)	5	A

Teaching languages

Dutch

Keywords

Data wrangling, data exploration, data visualization, R, explorative high-dimensional data analysis

Position of the course

This course is an introductory course into data science. In this course, students are introduced into the world of data analysis, both from a theoretical and an applied perspective (over two thirds of this course consists of hands-on data analysis with R). Special attention is given to the fact that, in a modern-day data analysis project, data can emerge in multiple forms and formats ranging from small structured flat-file datasets, that collect observations of a small dedicated research project to online sources that are less structured and potentially contain less reliable data, missing values, etc. In this course, the insight and the tools are introduced that are needed to extract information from these data sources in a reliable, reproducible and transparent way. In a first, more descriptive part of this course, the students are introduced to the life-cycle of data in a scientific setting, the composition and exploration of low-dimensional data tables with special attention to data visualization, data cleaning and data wrangling. In a second, more technical/mathematical part of this course, the focus is oriented to the theory and application of resampling methods, methods for exploring high-dimensional datasets and (an introduction to) predictive modeling. In a final (smaller) part, this use of relational databases for data science is introduced.

Contents

This course is subdivided into five parts. Each part involves both theory and practical skills (PC-labs in R).

Part 1: Introduction to data science and R

- Introduction to data and the data analysis cycle
- Introduction to R (important data types for data analysis, control flow, built-in and user-defined functions, vectorization, data intake)

Part 2: Univariate and bivariate explorative data analysis

- Data tables: scales of measurement, data collection principles and terminology, data storage and data access.
- Univariate explorative analysis: frequency distributions, sample percentiles, measures of central tendency and measures of scale, standardization, resampling methods (resampling from a population and bootstrapping), missing values

- Bivariate explorative analysis: bivariate frequency distributions, simple linear regression, R^2, local linear regression and scatterplot smoothing, Spearman correlation, Pearson correlation, Cramér's V, bivariate outliers

Part 3: Data manipulation and visualization

- Principles of reproducible and transparent data analysis and communication
- Data manipulation (for data-preparation): selection, filtering, aggregation, join-operations, wide and long formats, tidy data. In the PC-labs, the implementation of data-wrangling in tidyverse is used.
- Data visualization: properties of good data-graphics, the layered grammar of graphics. In the PC-labs, the implementation of the grammar of graphics in ggplot2 is used.

Part 4: Model-based and multivariate data analysis

- Curve fitting: loss functions, Newton's method and gradient descent, the algorithms of Gauss-Newton and Levenberg-Marquardt
- Dimensionality reduction methods: principal component analysis and multidimensional scaling
- Introduction to predictive modelling: k nearest neighbors and performance evaluation, linear methods for regression and classification

Part 5: Relational databases and using personal data

A brief introduction to relational databases and SQL is given (with a focus on applications in data science) and privacy-issues with respect to the use of personal data are discussed.

Initial competences

Data science builds on certain learning outcomes of course units 'Calculus, 'Linear algebra', and 'Scientific computing'; or the learning outcomes have been achieved differently.

Final competences

1 The student is aware of the different forms in which data appear, are capable of performing basic integrity checks for the most important data forms/types and can select and apply a proper visualization method.
2 The student can use R as a programming environment for data analysis.
3 The student performs data loading tasks for data that are available in a variety of text-based data formats, merges these data and transform it into a shape that allows further processing.
4 The student performs an explorative univariate and bivariate analysis.
5 The student can compose and format summarizing figures and tables that meet the quality standards required for scientific communication.
6 The student selects a useful loss function and applies an optimization alogirthm for solving a curve fitting problem.
7 The student applies dimesionality reduction techniques to gain insight into datasets.
8 The student selects a predictive modeling method for solving a prediction problem.
9 The student queries a simple relational database.

Conditions for credit contract

Access to this course unit via a credit contract is determined after successful competences assessment

Conditions for exam contract

This course unit cannot be taken via an exam contract

Teaching methods

Group work, Seminar, Lecture

Extra information on the teaching methods

During the theoretical lectures, the fundamental concepts are discussed. The practical PC room classes exist of 10 hands-on practical sessions. In the group work, the students have to complete a real-life data collection and synthesis task.

Study material

Type: Syllabus

Name: Course notes data science
Indicative price: Free or paid by faculty
Optional: no
Language : Dutch
Number of Pages : 300
Oldest Usable Edition : 2024
Available on Ufora : Yes
Online Available : Yes
Available in the Library : No

Available through Student Association : No

References

Benjamin S. Baumer, Daniel T. Kaplan and Nicholas J. Horton. (2017) Modern Data Science with R. CRC Press, 578p.
John L. Faundeen, Thomas E. Burley, et al. (2013) The United States Geological Survey Science Data Lifecycle Model, USGS Open-File Report 2013-1265.

Course content-related study coaching

Students can make an appointment with the lecturer for asking questions related to the theoretical classes throughout the entire semester. Teaching assistants address questions w.r. t. the PC-labs and Ufora is used to provide on-line feedback if needed.

Assessment moments

end-of-term and continuous assessment

Examination methods in case of periodic assessment during the first examination period

Written assessment with open-ended questions

Examination methods in case of periodic assessment during the second examination period

Written assessment with open-ended questions

Examination methods in case of permanent assessment

Assignment

Possibilities of retake in case of permanent assessment

examination during the second examination period is possible

Extra information on the examination methods

The written exam (25% of total) evaluates the theoretical competences. Practical competences are evaluated during an open book PC-exam (65% of total). The written report (report of group-work) is evaluated and contributes to the final grade (10% of total).
Students are allowed to rework the group assignment form NPE for the second examination period. The reworked version is re-evaluated.

Calculation of the examination mark

The final grade is the result of: a written exam (25%), an open-book PC-exam (65%) and a written report (10%)