GHENT UNIVERSITY

Course Specifications

Valid in the academic year 2023-2024

Data Quality (E018700)

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

Credits 3.0 **Study time 90 h**

Course offerings in academic year 2023-2024

A (semester 1) English Gent

Lecturers in academic year 2023-2024

Bronselaer, Antoon TW07 lecturer-in-charge

Offered in the following programmes in 2023-2024	crdts	offering
Bridging Programme Master of Science in Bioinformatics(main subject Engineering)	3	A
Master of Science in Business Engineering(main subject Data Analytics)	3	A
Master of Science in Bioinformatics(main subject Engineering)	3	A
Master of Science in Business Engineering (Double Degree)(main subject Operations Management)	3	A
Master of Science in Business Engineering(main subject Operations Management)	3	A
Master of Science in Computer Science	3	A
Master of Science in Computer Science Engineering	3	A
Master of Science in Information Engineering Technology	3	A
Exchange Programme in Computer Science (master's level)	3	A
Exchange Programme Information Engineering Technology	3	A

Teaching languages

English

Keywords

Measurement of data quality, consistency, detection and repair of errors, outlier detection

Position of the course

This course is a specialization course in which mechanisms for safeguarding data quality are thought. A first part of the course deals with different methods to measure data quality. A second part studies algorithms that allow to detect and systematically repair errors in data. A third part deals with the problem of deduplication. Next, in a fourth part, the problem of outlier detection is treated and finally, a fifth part will focus on specific quality problems with temporal data.

Contents

- Introduction to data quality
- Measurement of data quality: ordinal systems, uncertainty models and cost-based measurement
- Basics of constraint-based formalisms
- Control digits
- Edit rules: error localization, the Fellegi-Holt model, FCF algorithm
- The Chase algorithm for functional dependencies
- Data deduplication: Fellegi-Sunter model, string comparison, merging of duplicate data.
- Outlier detection: distance-based models, pivot index, spatial partitioning, isolation forests
- Data quality in temporal databases: trend decomposition, change detection, currency

Initial competences

Basic principles of data structures and relational databases. Basic knowledge of programming.

Final competences

1 Knowing and understanding the basic techniques for measurement of data quality

2 Understanding how consistency can be enforced and being able to apply this
3 Understanding how a dataset can be deduplicated
4 Understanding how outliers can be found
5 Knowing and understanding specific quality problems with temporal data.

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

Seminar, Lecture

Learning materials and price

- Slides
- Additional study material available via Ufora (short videoclips and articles)

References

- Ton De Waal, Jeroen Pannekoek en Sander Scholtus (2011). Handbook of Statistical Data Editing and Imputation, Wiley
- Wenfei Fan en Floris Geerts (2012). Foundations of Data Quality Management. Morgan & Claypool Publishers.

Course content-related study coaching

Exercise classes will be supervised by assistents

Assessment moments

end-of-term assessment

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

Written assessment

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

Written assessment

Examination methods in case of permanent assessment

Possibilities of retake in case of permanent assessment

not applicable

Extra information on the examination methods

Periodic evaluation: written open book exam with open questions that measure insights in the concepts of the course

Calculation of the examination mark

100% written, open book exam