GHENT UNIVERSITY

Course Specifications

Valid in the academic year 2024-2025

Big Data Science (C003802)

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)	English	Gent	lecture
			independent work
			seminar

Lecturers in academic year 2024-2025

De Witte, Dieter	TWO6	lecturer-in-charge
De Bie, Tijl	TWO6	co-lecturer
Lijffijt, Jefrey	TWO6	co-lecturer
Pizurica, Aleksandra	TWO7	co-lecturer

Offered in the following programmes in 2024-2025	**crdts**	**offering**
Master of Science in Statistical Data Analysis	5	A

Teaching languages

English

Keywords

Artificial Intelligence, Machine learning, regression, classification, clustering, neural networks, deep learning, Ethical dimension of AI Hashing, Sketching, Algorithms for large-scale ML and Data Mining, Big Data frameworks, Graph embedding, Graph, analytics, Pattern mining

Position of the course

In this course, we cover both the theoretical foundations and practical applications of advanced data processing and Artificial Intelligence techniques, with a particular focus on Machine Learning and data-driven model building.
Advanced, scalable algorithms are essential tools for modern data engineers, extending beyond typical deep learning techniques to address the challenges of handling data with complex and large-scale dimensions. This data, often referred to as Big Data, can reside in centralized data warehouses, decentralized systems, or even in continuous data streams.
A key component of the course is understanding the strengths and limitations of these approaches, as well as addressing critical issues such as data quality and data privacy, which present significant challenges in working with large and sensitive datasets. Additionally, we explore techniques for extracting insights from (knowledge) graphs, including the use of graph embeddings.
Students will gain both theoretical knowledge and practical experience through a series of hands-on assignments, allowing them to apply advanced AI and scalable data processing techniques to real-world problems.

Contents

The course consists of a combination of lecturers from the course Big Data Algorithms (E018250) and Artificial Intelligence (E01635).
The Big Data Algorithms takes place on Tuesdays 14h30-17h30 in weeks 1-12
The Artificial Intelligence takes place on Mondays 11h30-13h and Thursdays 8h30-10h in weeks 1-7, 12

- Fundamental machine learning concepts (dataset, training set, validation set, dimensionality, overfitting, bias and variance, cross validation);

- Supervised learning: construction of data driven models; regression and classification; white-box ML (logistic regression, decision trees, nearest neighbour models, kernel method); black-box ML (neural networks, basics of deep learning)
- Unsupervised learning: dimensionality reduction, clustering
- Scalable data mining: algorithms for clustering, dimensionality reduction, and machine learning fit for a distributed context. (for example: with MapReduce)
- Decentralized Data mining: federated querying and federated learning in decentralized data sources such as data vaults (SOLID) and in data spaces
- Processing data streams: sketching algorithms, probabilistic counting, and online algorithms
- Data Quality: generic techniques to improve data quality, hashing techniques for detecting (near-)duplicates.
- Mining of (knowledge) graphs: algorithms for detecting structures in (social) networks (triangles, communities, centrality, k-cores, etc.), embeddings.
- Data Privacy and Security: algorithms for anonymization and pseudonymization, the European AI Act and GDPR, risks of de-anonymization. Privacy-preserving machine learning (differential privacy, homomorphic encryption, etc).

Initial competences

- Basic knowledge of Statistics
- Basic knowledge of Linear Algebra
- Programming experience in Python

Final competences

1 Handle datasets with multiple challenging dimensions (size, format, quality, ...).
2 Setting up a (cloud) environment for scalable data processing.
3 Applying machine learning / data mining algorithms to Big Data.
4 Applying sketching techniques to solve challenging Big Data problems.
5 Have an in-depth understanding on how to transform graphs to be used in machine learning setups.
6 Being able to conduct an analysis on a large relational graph.
7 Detect near-duplicates in large datasets using hashing techniques.
8 Having an overall view on the different generic problem classes in the AI discipline.
9 Having insight in the fundamentals and concepts underlying commonly used solution techniques in this discipline, especially focussing on data driven model construction (white box as well as black box).
10 Solving specific problems in AI using the methods of this course (and extending these methods as needed in terms of applicability and context), as well on paper as in Python.
11 Being able to assess the limitations and ethical consequences of AI-techniques.

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, Practical, Independent work

Study material

Type: Handbook

Name: Artificial Intelligence: A Modern Approach (Global Edition), 4th Edition
Indicative price: € 70
Optional: yes
Language : English
Author : Stuart Russell and Peter Norvig
ISBN : 1-292-40113-3
Number of Pages : 1115
Oldest Usable Edition : 3rd 3dition

Online Available : Yes
Available in the Library : Yes
Available through Student Association : Yes
Usability and Lifetime within the Course Unit : regularly
Usability and Lifetime within the Study Programme : regularly
Usability and Lifetime after the Study Programme : regularly

Type: Handbook

Name: Deep Learning: Foundations and Concepts
Indicative price: Free or paid by faculty
Optional: yes
Language : English
Author : Christopher M. Bishop and Hugh Bishop
ISBN : 3-031-45467-7
Number of Pages : 650
Online Available : Yes
Available in the Library : Yes
Available through Student Association : Yes
Usability and Lifetime within the Course Unit : regularly
Usability and Lifetime within the Study Programme : regularly
Usability and Lifetime after the Study Programme : regularly

Type: Handbook

Name: Mining of Massive Datasets (3rd edition)
Indicative price: Free or paid by faculty
Optional: no
Language : English
Author : Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman
ISBN : 978-1-10847-634-8
Number of Pages : 315
Online Available : Yes
Available in the Library : No
Usability and Lifetime within the Course Unit : regularly
Usability and Lifetime within the Study Programme : one-time
Usability and Lifetime after the Study Programme : not

Type: Syllabus

Name: Lecture Notes: Artificial Intelligence
Indicative price: € 5
Optional: yes
Language : English
Available on Ufora : No
Online Available : No
Available in the Library : No
Available through Student Association : Yes

Type: Slides

Name: Slides for the course Artificial Intelligence
Indicative price: Free or paid by faculty
Optional: no
Language : English
Available on Ufora : Yes
Available in the Library : No
Available through Student Association : Yes

References

S. Russel and P. Norvig: Artificial Intelligence: A Modern Approach. Fourth Edition, Prentice Hall, 2020.
A. Lindholm, N. Wahlström, F. Lindsten, and T.B. Schön. Machine Learning: A First Course for Engineers and Scientists. Cambridge University Press, 2022.
C. Bishop and H. Bishop: Deep Learning - Foundations and Concepts, Springer, 2024.

Course content-related study coaching

Assessment moments

end-of-term and continuous assessment

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

Oral assessment, Written assessment

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

Oral assessment, Written assessment

Examination methods in case of permanent assessment

Skills test, Presentation, Assignment

Possibilities of retake in case of permanent assessment

examination during the second examination period is possible in modified form

Extra information on the examination methods

- periodic evaluation: Written exam (closed book, closed personal notes except summary on 1 page, A4 format) on the AI part and oral exam with limited preparation time on the Big Data part.
- permanent evaluation: Evaluation of practical work in groups, spread over the semester, as well as individual home work assignments.

Calculation of the examination mark

50% permanent evaluation and 50% periodic evaluation