

## Big Data Algorithms (E018250)

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

**Credits 3.0**                      **Study time 90 h**

**Course offerings in academic year 2024-2025**

A (semester 2)                      English                      Gent

**Lecturers in academic year 2024-2025**

De Witte, Dieter	TW06	lecturer-in-charge
De Bie, Tijl	TW06	co-lecturer
Lijffijt, Jeffrey	TW06	co-lecturer

**Offered in the following programmes in 2024-2025**

	<b>crdts</b>	<b>offering</b>
<a href="#">Bridging Programme Master of Science in Bioinformatics(main subject Engineering)</a>	3	A
<a href="#">Master of Science in Bioinformatics(main subject Engineering)</a>	3	A
<a href="#">Master of Science in Industrial Engineering and Operations Research(main subject Manufacturing and Supply Chain Engineering)</a>	3	A
<a href="#">Master of Science in Industrial Engineering and Operations Research(main subject Transport and Mobility Engineering)</a>	3	A
<a href="#">Master of Science in Computer Science Engineering</a>	3	A
<a href="#">Master of Science in Industrial Engineering and Operations Research</a>	3	A

**Teaching languages**

English

**Keywords**

Hashing, Sketching, Algorithms for large-scale ML and Data Mining, Big Data frameworks, Graph embedding, Graph, analytics, Pattern mining

**Position of the course**

Advanced, scalable algorithms are indispensable in the toolkit of modern data engineers. Beyond the typical deep learning tools, these algorithms often play a significant role in their daily practice where data can have 'challenging dimensions'. This 'Big Data' can be found in both centralized data warehouses as well as decentralized systems or even data streams.

We also cover data quality and data privacy, as they are some of the biggest hurdles in daily handling of (sensitive) datasets.

Finally, we explore how insights can be extracted from (knowledge) graphs using graph embeddings.

A number of hands-on assignments are provided, allowing students to explore theoretical topics.

**Contents**

- **Processing data streams:** sketching algorithms, probabilistic counting, and online algorithms
- **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
- **Data Quality:** generic techniques to improve data quality, hashing techniques for detecting (near-)duplicates.
- **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).

- **Mining of (knowledge) graphs:** algorithms for detecting structures in (social) networks (triangles, communities, centrality, k-cores, etc.), embeddings.

#### **Initial competences**

- basic programming skills
- Experience with Python (passed the course Informatics E015041 or an equivalent course)
- Experience with Object Oriented Programming (passed the course Computer Programming E017210 or an equivalent course)
- Experience with data structures and algorithms
- elementary understanding about basic data formats (CSV, TSV, etc.)
- linear algebra
- introductory course on statistics
- ML and AI basics

#### **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

#### **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**

Lecture, Practical

#### **Study material**

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

Available through Student Association : 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

#### **References**

#### **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

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

Oral assessment

#### **Examination methods in case of permanent assessment**

Skills test, Presentation

**Possibilities of retake in case of permanent assessment**

examination during the second examination period is not possible

**Extra information on the examination methods**

- Periodical evaluation
- Oral exam with limited preparation time

**Calculation of the examination mark**

50% lab reports, 50% oral exam