

## Introduction to Data Science (1002892)

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

**Credits** 4.0 **Study time** 120 h

**Course offerings in academic year 2023-2024**

### Lecturers in academic year 2023-2024

Verwaeren, Jan LA26 lecturer-in-charge

**Offered in the following programmes in 2023-2024**

	<b>crdts</b>	<b>offering</b>
<a href="#">Master of Science in Pharmaceutical Engineering</a>	4	A

### Teaching languages

English

### Keywords

Data wrangling, data exploration, data visualization, Python, machine learning

### Position of the course

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. The process of gathering, cleaning and wrangling these data and using them to solve a (business) question or problem is generally understood as *Data Science*. This process requires a specific set of skills from a data analyst, including the knowledge of a (data-oriented) programming language that allows to clean, wrangle and explore large amounts of data in an efficient and reproducible manner. Moreover, a lot of data science projects require both theoretical and hands-on knowledge and skills on machine learning. In this course, students are introduced to the methodological and practical aspects of data science.

### Contents

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

**Part 1:** Introduction to data science and Python as an environment for data science

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

**Part 2:** 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 packages Numpy and Pandas are used for this purpose.
- Data visualization: properties of good data-graphics, the layered grammar of graphics. In the PC-labs, Matplotlib and Seaborn are used for this purpose

**Part 3:** Introduction to machine learning

- Recognizing types of machine learning problems
- kNN as a prototype machine learning technique
- Supervised machine learning: Linear methods for regression and classification, Basis expansions and regularization, Performance evaluation, Tree-based methods and neural networks
- Unsupervised machine learning: dimensionality reduction and clustering

#### **Part 4: Project: analysing unstructured data**

Students work on a data science project that involves the analysis of unstructured data.

Important aspects are: feature engineering, applying machine learning models on real-life data and transparent reporting on a data analysis project.

#### **Initial competences**

The students have programming experience (basic level) in at least one scientific programming language (R, Matlab, Python, etc.)

The students have a basic understanding of statistics (including descriptive statistics, simple linear regression)

#### **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 Python 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 applies dimensionality reduction techniques to gain insight into datasets.
- 5 The student selects and applies a predictive modeling method for solving a prediction problem.

#### **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, lecture, seminar

#### **Extra information on the teaching methods**

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

#### **Learning materials and price**

Lecture slides and handbooks are digitally available (free of charge)

#### **References**

Jake Vanderplas (2016). Python Data Science Handbook, O'Reilly Media, Inc., 548p

Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (2021). An Introduction to Statistical Learning - 2nd Edition, Springer, 597p.

#### **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 (20% of total) evaluates the theoretical competences. Practical competences are evaluated during an open book PC-exam (50% of total). The written report (report of group-work) is evaluated and contributes to the final grade (30% of total).

#### **Calculation of the examination mark**

The final score is the weighted sum of: theoretical exam (20%), practical exam (50%) and group-work (30%).