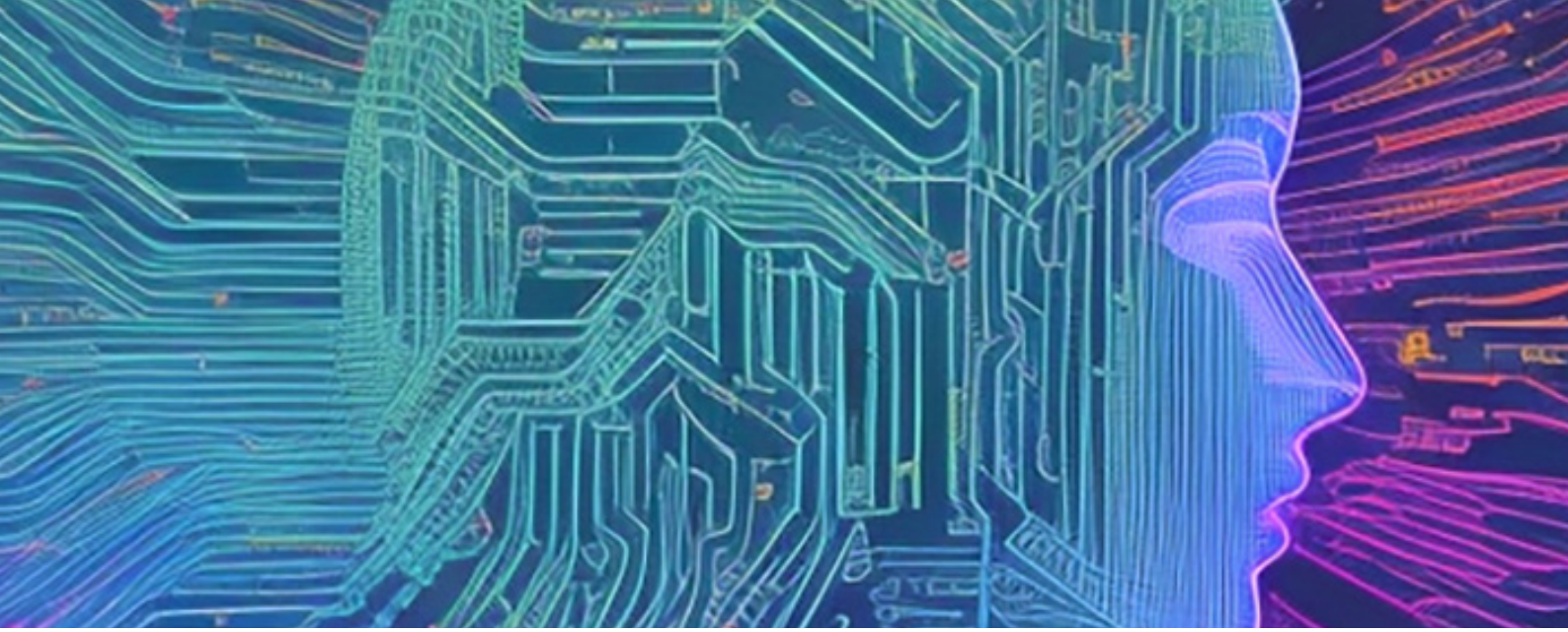

BASICS OF DATA SCIENCE

ABOUT THIS COURSE

„Basics of Data Science“ gives a comprehensible overview of many fundamental concepts and tools of data science, including data quality assessment and data preprocessing, supervised and unsupervised learning techniques including their evaluation, frequent itemsets and association rules, sequence mining, process mining, text mining, and responsible data science.

PREREQUISITES

Everyone from any discipline with an interest in data science can start this course. We expect this course to be useful for everyone. Prior knowledge in maths is of advantage (i.e., mathematical notations, linear algebra, stochastics, and statistics), but not mandatory.

WHAT YOU WILL LEARN

After taking this course, participants will have gained:

- An understanding of the role of data science in today's society and businesses, including challenges and opportunities,
- A good general overview of a broad range of data science techniques,
- The ability to conceptualize and implement basic data analysis and to accurately evaluate and interpret the outcomes,
- An understanding the challenges of responsible data science (fairness, accuracy, confidentiality, transparency) and possible solutions,
- An understanding of the limitations of machine learning, data mining and AI techniques, and
- The ability to write short Python programs and to use mainstream Python libraries.
- In particular, understanding of and ability to apply the following data analysis concepts and techniques:
- Data visualization and exploration techniques,
- Decision trees,
- Linear and logistic regression (basic overview),
- Support vector machines (basic overview),
- Neural networks (basic overview),
- Naïve Bayesian classification (basic overview),
- Evaluation and interpretation of the results obtained using supervised learning,
- Clustering techniques,
- Frequent item sets,
- Association rules,
- Sequence mining,
- Process mining,
- Text mining,
- Data preprocessing, data transformation, spotting, and handling of data quality problems, and
- Application of data analysis techniques without violating confidentiality and fairness.

INSTRUCTOR

Prof.dr.ir. Wil van der Aalst
Wil van der Aalst is a full professor at RWTH Aachen University leading the Process and Data Science (PADS) group and Chief Scientist at Celonis. He is an IFIP Fellow, IEEE Fellow, ACM Fellow, and one of the most cited computer scientists in the world, also known as the „Godfather of Process Mining". In 2018, he was awarded a Humboldt Professorship, Germany's most valuable research award (five million euros).

Lisa Luise Mannel
Lisa Luise Mannel is a doctoral student at the Process and Data Science (PADS) group led by Wil van der Aalst. She received her master's and bachelor's degree in computer science from RWTH Aachen University. Her main research interests are in the area of process mining with a focus on process discovery. In particular, she develops the process discovery algorithm ‚eST-Miner' in the context of her Ph.D. project.

TIME COMMITMENT

The course is self-paced and designed in such a way that it can be completed within 9 weeks. After finishing the course material, you must pass a final exam. The expected time to complete each week is approximately 3-4 hours for the audit track and 4-5 hours for the verified track. This includes:

- Watching lecture videos
- Completing the recap quizzes
- Completing the Python lab
- Completing the graded exercise questions for each week (verified track)
- Finishing a final exam (verified track)

COURSE OUTLINE

WEEK 1

Introduction, Data Exploration & Visualization

In the first half of the week, we will provide an overview of the course and illustrate the advantages and challenges when applying data science techniques. Students will get an overview of the data science pipeline, data sources and data types, data analysis techniques and challenges related to their application.
The second half of the week focuses on basic data exploration, visualization, and transformation techniques.

WEEK 2

Supervised Learning Techniques

In the first half of this week, students will delve into data analysis using decision trees. We introduce the basic ID3 Algorithm and its extension to different notions of information gain, as well as pruning techniques, random forests, and the applicability of decision trees to continuous data.
The second half of the week is dedicated to a brief overview of other supervised learning techniques (students interested in detail are referred to the „Basics of Machine Learning“ course which is also part of the BridgingAI course series). These techniques include linear regression, logistic regression, support vector machines (SVMs), neural networks and naive Bayesian classification.

WEEK 3

Evaluation of Supervised Learning, Data Quality & Preprocessing

The first half of this week is dedicated to the evaluation of supervised learning techniques and the models they produce. We introduce the confusion matrix, ROC curve, R2 Coefficient and cross validation including their extension and adaptation to specific goals or contexts. Furthermore, challenges and pitfalls regarding the evaluation and interpretation of supervised learning techniques are highlighted.
In the second half of the week, students will learn about data quality issues, their causes and avoidance strategies as well as possible approaches to dealing with outliers or missing values. Furthermore, an overview of data transformation, data reduction and normalization techniques are given.

COURSE OUTLINE

WEEK 4

Clustering, Frequent Itemsets

In the first half of this week clustering is introduced as the first unsupervised learning technique. In particular, we present various similarity measures, the k-means and k-medoids algorithms, density-based clustering (DBSCAN) and give an overview of agglomerative clustering techniques and self-organizing maps (SOM).
The second half of the week focuses on the introduction of frequent itemsets. Two algorithms to compute such itemsets are explained: the straightforward Apriori approach as well as the more efficient FP-Growth algorithm.

WEEK 5

Association Rule Mining, Sequence Mining

In this week, we build upon the concepts of frequent itemsets to generate and evaluate association rules. Furthermore, we use association rules to illustrate Simpson‘s paradox.
The second half of the week revolves around sequence mining, particularly the AprioriAll algorithm. The relationships between frequent itemsets, association rules, sequence mining and process mining (introduced in Week 6) are clarified

WEEK 6

Process Mining

The whole week is dedicated to various aspects of process mining. We start out with an extensive introduction to the topic, including various types of models, tools, and applications. Next, various approaches to process discovery are presented as the most prominent example of unsupervised learning in the context of process mining. Finally, supervised problems in process mining are discussed with the main focus on conformance checking techniques.

WEEK 7

Text Mining

In this week we explore the topic of text mining. Various approaches to text preprocessing are discussed, including corpus annotation, tokenization, stop word removal, token normalization, stemming and lemmatization, followed by an overview of modelling techniques, i.e., BoW, document-term matrix and TF-IDF scoring. We briefly discuss the inclusion of semantics using public databases (Linked Open Data) before proceeding with a detailed introduction to N-grams and their application to word prediction and text generation. These concepts are extended in the following when discussing word embeddings, particularly the concepts of autoencoders, word2vec, CBoW and doc2vec.

COURSE OUTLINE

WEEK 8

Responsible Data Science

In this week we discuss challenges and solution approaches to confidentiality and fairness in data science. The first half of the week is dedicated to confidentiality. We give a brief overview of data encryption before introducing various techniques to anonymize data while maintaining its usefulness for analysis and to objectively evaluate the level of anonymization.
The second part of the week, focusing on fairness, introduces various metrics to objectively measure fairness and explores approaches to decrease discrimination of data science models and techniques. We conclude with a discussion of the potential trade-offs between model performance and model fairness.

WEEK 9

The Bigger Picture

In the final week, we briefly recap the contents of the course and discuss connections, trade-offs, conflicts, and interactions between the various topics as well as their context and impact within the bigger picture of data science. An outlook on further perspectives and topics omitted in this introductory course is given.

ASSESSMENTS & GRADING

Recap questions: After watching lecture videos, you will be asked to answer some questions relevant to the lecture which aims to help you practice and improve your knowledge. These questions are not graded.

Weekly exercise questions: After watching the lecture videos in the respective weeks and completing the weekly programming exercise you will be asked to answer a series of questions. The weekly exercise questions are contributing with 40% to your final grade.

Final Exam: In Week 9, your knowledge will be evaluated in a final exam that covers the material taught within the whole course. The final exam will be evaluated as 60% of your final grade.

DISCUSSION FORUM

If you have general questions, ask them in the general discussion forum. At the end of each week, discussion forums are located. These forums are supposed to be an interactive environment in which you can ask your questions regarding each week and share your ideas with other students and instructors. You should not post the answers to graded questions in the forum. A moderator from the course team will administer and monitor the forum. A document explaining the rules and guidelines on how to use the discussion forum is provided in the handout section.

ACADEMIC HONOR CODE

By participating in this course, you pledge to follow the edX honor code (https://www.edx.org/edx-terms-service). Explicitly, we expect you to be a diligent student and contribute to the course.

We believe it is not too hard to achieve a good grade when participating regularly and you will learn a lot about the topic at hand. We put a lot of effort in creating a great course for you and highly appreciate your feedback and suggestions!