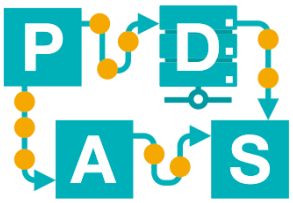


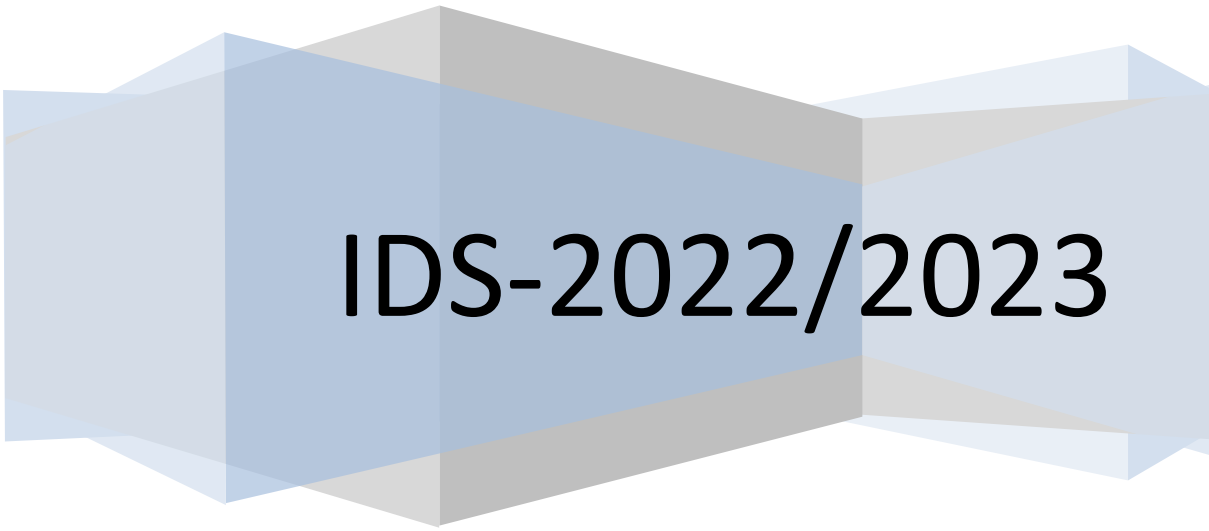
Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

Study Guide

Introduction to Data Science

Wil van der Aalst, Niklas Adams, Bianka Bakullari, Harry Beyel, Tobias Brockhoff, Tsunghao Huang, Benedikt Knopp, Viki Peeva.



IDS-2022/2023

Table of Contents

Study Guide - Introduction to Data Science (WS 2022/2023)

Lecturers and Instructors

Lecturer

- Prof.dr.ir. Wil van der Aalst

Instructors

- Niklas Adams
- Bianka Bakullari
- Harry Beyel
- Tobias Brockhoff
- Tsung-Hao Huang
- Benedikt Knopp
- Viki Peeva

All the questions related to the course lectures and instructions should be asked via Moodle. In case of urgent personal questions, please contact *ids@pads.rwth-aachen.de*. To allow for fast processing, always include your immatriculation number in the email. Please avoid sending e-mails to individuals or even multiple lecturers. Moreover, if you have problems with RWTH Online or RWTH Moodle that are not specifically related to this course, please contact the persons responsible for these systems and not the lecturer or the instructors.

Data Science and This Course

Recently, data science has emerged as a new and important discipline. Data science can be viewed as an amalgamation of classical disciplines, such as statistics, data mining, databases, and distributed systems. This combination helps to turn data into value for the profit of individuals and society. In addition, new challenges are constantly emerging and making this field highly dynamic and appealing. These challenges are not just in terms of size ("Big data"), but also regarding complexity of the questions to be answered. Data science provides numerous opportunities to develop exciting products and services. With the technological evolution, the boundaries of what algorithms can perform are pushed even further. This development raises significant questions that will be addressed in this course.

The practical relevance of data science is evident when we consider how the most successful companies (e.g., Amazon, Google, Facebook, Spotify, Netflix, Salesforce, Uber, eBay, Airbnb, Zalando, Celonis, Signavio, SAP, Siemens, DHL, etc.) use their data. Also in academic research, data science is becoming a key component, influencing other fields while continuously developing itself.

With growing importance of data and digitalization, organizations are looking for data scientists. One could even argue that in the future we might need more data scientists than computer scientists. There is an urgent need for data science experts and it is the right time to become one. This course therefore aims to provide the basics for becoming an all-round skilled "data wizard."

The course is mainly focused on data analysis and discusses a substantial range of analytical approaches and tools. All in all, the course aims to provide a *comprehensive overview* of data science using analytical tools applied to real-life as well as synthetic datasets. The course covers three main parts of data science:

1. Data science *infrastructure* concerned with volume and velocity of data. The topics include instrumentation, big data infrastructures and distributed systems, databases and data management, and programming. The main challenges are scalability and responsiveness.

2. Data science *analysis* concerned with extracting knowledge from data. The topics cover statistics, data and process mining, machine learning and artificial intelligence, operational research, algorithms, and data visualization. In this part, the main challenge is to provide answers to known and unknown unknowns.
3. Data science *effects* concerned with people, organizations, and society. We discuss ethics and privacy, IT laws, human-technology interaction, operations management, business models, and entrepreneurship. Here, the main challenge is to implement data practices in a responsible manner.

Course Objectives

After taking this course, students should:

- have a good understanding of a broad range of data science techniques,
- be able to apply the mainstream data science techniques and corresponding tools,
- understand the role of big data and data science in today's society,
- understand the limitations of machine learning, data mining and process mining techniques,
- be able to write short Python programs and apply existing programs,
- understand data visualization and exploration techniques,
- be able to construct decision trees from any data set,
- understand and apply regression techniques,
- understand and apply support vector machines,
- understand and apply basic neural networks,
- be able to evaluate the results obtained using supervised learning,
- understand and apply clustering techniques,
- be able to construct frequent item sets,
- understand and discover association rules,
- understand and apply sequence mining,
- understand and apply process mining,
- understand and apply text mining,
- be able to do data preprocessing and spot data quality problems,
- understand visual analytics and advanced information visualization approaches,
- understand the four elements of responsible data science (fairness, accuracy, confidentiality, and transparency),
- know the big data challenges and technological approaches
- have hands-on experience using a variety of data sets provided.

Organization and Recommended Literature

This year both lectures and instructions will take place in presence. The **lectures** take place **on Mondays from 12:30 to 14:00 at PPS H1 (2315|101)** and **Tuesdays from 12:30 to 14:00 at FT (2090|120)**. The **instructions** take place on **Fridays at AH IV (2354|030) from 14:30 to 16:00**. There are deviations from this rule, so check the schedule at the end of this document and follow the announcements on Moodle for updates.

The lecture material will be made available via RWTH Moodle. If there are problems, you can alternatively watch the videos via

- Video AG https://video.fsmpi.rwth-aachen.de/21ws-ids

- YouTube
  https://www.youtube.com/watch?v=czc07quxyEw&list=PLG_1ZxIPXO0vTTfheRNDhq4vNYAdJBUyC

Note that the online lectures were recorded last year. The content did not change, but we suggest you use the current materials provided in RWTHMoodle to study.

The course provides a broad overview of data science and engages with different sources. The slides should be self-contained assuming that the students attend both lectures and instructions. The following two books are highly recommended for the course:

- **Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies** by John D. Kelleher, Brian Mac Namee and Aoife D'Arcy. MIT Press. (both the 2015 and 2020 version can be used) (http://machinelearningbook.com/).
- **Data Mining: Concepts and Techniques** (3rd edition) by Jiawei Han, Micheline Kamber, Jian Pei. The Morgan Kaufmann Series in Data Management Systems, Elsevier. ISBN: 9780123814807, 744 pages, 2011 (http://hanj.cs.illinois.edu/book).

Next to the slides and books, the following material will be distributed via the RWTH Moodle platform:

- Exercises
- Datasets
- Assignment (two parts).

Please note that due to the uncertainty related to the pandemic, new regulations and rules appointed from the university and/or government may affect the lecture plan and schedule.

Software

In this course, we use a single, unified virtual environment for Python programming. It is highly recommended for you to practice and work on the assignment using a virtual environment. This will make your package management easier, while allowing us to avoid incompatibilities among different versions of Python and packages when grading the assignment. Please see the installation guide and the first instruction for details.

Examination

The module comprises two parts: one **assignment** delivered in two parts (each accounting for 20%) and the **written exam** that accounts for the remaining 60% of the final grade. The assignment must be submitted in groups of 2-3 students. To pass the course, it is required to pass both the assignment and the written exam, that is, you should achieve at least 50% in the assignment and at least 50% in the final written exam. Please note that the dates and times are tentative and may be subject to changes.

- **The written exam** (60%): questions to test theoretical knowledge of the algorithms and techniques learned in the course:
  - First option (PT1): 14/02/2023 (tentative, check later)
  - Second option (PT2): 22/03/2023 (tentative, check later)
- **IDS Assignment Part 1** (20%): analysis of real-life and synthetic datasets using techniques and tools discussed in the course. This part tests your understanding of the material given in lectures 1-9. The deadline is on Sunday 18/12/2022 23:59.

- **IDS Assignment Part 2** (20%): analysis of complex datasets using various data science techniques. This includes interpretation of the results and using multiple views on the data. This part focuses on the material from lectures 10-21. The deadline is on Sunday 22/01/2023 23:59.

Important: Only the written exam can be retaken in this semester. It is not possible to retake parts of the course; therefore, the results of the assignment expire after the end of the semester.

Plagiarism: We systematically check for plagiarism. All group members are responsible to submit an individual piece of work and should avoid unfair academic practices. We will report cases of proven plagiarism to the examination board.

Who can take the course?

This course is a master program course and is mandatory for students taking the Data Science master (it is listed as an elective (Wahlpflichtfach) but required for doing a master thesis). It is an elective course (Wahlpflichtfach) for Computer Science (Informatik) master, Media Informatics, Software Engineering, and Physics. Other students are welcome to participate, but it is up to the management and rules of the corresponding study programs to decide whether the credits of this course will count. If you have problems with RWTH Online or RWTH Moodle which are not specifically related to this course, please contact the persons responsible for these systems and not the lecturer(s).

About the Process and Data Science (PADS) Group @ RWTH

The Process and Data Science (PADS) group, headed by prof.dr.ir. Wil van der Aalst, is one of the research units in the Department of Computer Science. The scope of PADS includes all activities where discrete processes are analyzed, re-engineered, and supported in a data-driven manner. Process-centricity is combined with an array of data science techniques. The group's research and teaching activities cover the following areas: data science, process mining, business process management, data mining, process discovery, conformance checking, and simulation.

The group closely collaborates with the Fraunhofer Institute for Applied Information Technology (FIT), software firms developing process mining software (e.g., Celonis, Fluxicon, and UiPath), larger organizations using process mining (e.g., Siemens, BMW, Philips, and Vanderlande), and consultancy firms (KPMG, Deloitte, EY, and PwC).

Currently, the main research focus is on process mining (including process discovery, conformance checking, performance analysis, predictive analytics, operational support, and process improvement). This is combined with neighboring disciplines such as operations research, algorithms, discrete event simulation, business process management, and workflow automation.

Visit http://www.pads.rwth-aachen.de/ to learn more about our chair and to find out about possible bachelor and master theses.

Timetable of Lectures and Instructions

	Lecture	date	day	place
Lecture 1	Introduction	10/10/2022	Monday	PPS H1
Lecture 2	Basic data visualization/exploration	11/10/2022	Tuesday	FT
Instruction 1	Python installation	14/10/2022	Friday	AH IV
Lecture 3	Decision trees	17/10/2022	Monday	PPS H1
Lecture 4	Regression	18/10/2022	Tuesday	FT
Instruction 2	Crash Course in Python	21/10/2022	Friday	AH IV
Lecture 5	Support vector machines	31/10/2022	Monday	PPS H1
Instruction 3	Decision trees and data visualization/exploration	04/11/2022	Friday	AH IV
Lecture 6	Neural networks (1/2)	07/11/2022	Monday	PPS H1
Lecture 7	Neural Networks (2/2)	08/11/2022	Tuesday	FT
Instruction 4	Regression and support vector machines	11/11/2022	Friday	AH IV
Lecture 8	Evaluation of supervised learning problems	14/11/2022	Monday	PPS H1
Lecture 9	Clustering	15/11/2022	Tuesday	FT
Instruction 5	Neural networks	18/11/2022	Friday	AH IV
Lecture 10	Frequent item sets	21/11/2022	Monday	PPS H1
Lecture 11	Association rules	22/11/2022	Tuesday	FT
Instruction 6	Evaluation	25/11/2022	Friday	AH IV
Lecture 12	Sequence mining	28/11/2022	Monday	PPS H1
Lecture 13	Process mining (unsupervised)	29/11/2022	Tuesday	FT
Instruction 7	Clustering and frequent item sets	02/12/2022	Friday	AH IV
Lecture 14	Process mining (supervised)	05/12/2022	Monday	PPS H1
Lecture 15	Text mining (1/2)	06/12/2022	Tuesday	FT
Instruction 8	Association rules and sequence mining	09/12/2022	Friday	AH IV
Lecture 16	Text mining (2/2)	12/12/2022	Monday	PPS H1
Instruction 9	Q&A Assignment Part 1	13/12/2022	Tuesday	FT
Instruction 10	Process mining	16/12/2022	Friday	AH IV
	Deadline Assignment Part 1 (18.12.2022 23:59)			
Lecture 17	Data preprocessing	19/12/2022	Monday	PPS H1
Instruction 11	Text mining	20/12/2022	Tuesday	FT
Lecture 18	Visual analytics & information visualization	09/01/2023	Monday	PPS H1
Lecture 19	Big data	10/01/2023	Tuesday	FT
Instruction 12	Preprocessing and visualization	13/01/2023	Friday	AH IV
Instruction 13	Big data (1/2)	16/01/2023	Monday	PPS H1
Instruction 14	Big data (2/2)	17/01/2023	Tuesday	FT
Instruction 15	Q&A Assignment Part 2	20/01/2023	Friday	AH IV
	Deadline Assignment Part 2 (22.01.2023 23:59)			
Lecture 20	Responsible data science (1/2)	23/01/2023	Monday	PPS H1
Lecture 21	Responsible data science (2/2)	24/01/2023	Tuesday	FT
Instruction 16	Responsible data science	27/01/2023	Friday	AH IV
Lecture 22	Closing	30/01/2023	Monday	PPS H1
Instruction 17	Example Exam Questions	03/02/2023	Friday	AH IV