Chair of Process and Data Science

Study guide

Introduction to Data Science

Wil van der Aalst, Anahita Farhang, Lisa Mannel, Gyunam Park, Mahnaz Qafari, Miriam Wagner, Iana Gein, and Mahnaz Mirbolouki

IDS-2020/2021

Table of contents

Lecturers and instructors 2
Data science and this course 2
Course objectives 3
Organization and recommended literature 3
Software 4
Examination 4
Who can take the course? 5
About the Process and Data Science (PADS) group @ RWTH 5
Timetable of lectures and instructions 6

Lecturers and instructors

Lecturer

- Prof.dr.ir. Wil van der Aalst

Instructors

- Anahita Farhang
- Gyunam Park
- Iana Gein
- Lisa Mannel
- Mahnaz Mirbolouki
- Mahnaz Qafari
- Miriam Wagner

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*. Please avoid sending e-mails to individuals or even multiple lecturers.

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 make this field highly dynamic and appealing. These 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 technological evolution, the boundaries of what algorithms can perform will be 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 (Amazon, Google, Facebook, Spotify, Netflix, Salesforce, Uber, eBay, Airbnb, Zalando, Celonis, Signavio, SAP, Siemens, DHL, etc.) use their data. Also in an academic research, we observe that data science has become the key component. Data science is influencing other sciences while continuously developing itself.

With growing importance of data and digitalization, organizations are looking for data scientists which can be a long-term trend. 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 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 and synthetic datasets. The course discusses three main parts of data science:

1. Data science *infrastructure* concerned with volume and velocity. The topics include instrumentation, big data infrastructures and distributed systems, databases and data management, and programming. The main challenge is to make making things scalable and instant.

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. The topics 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 and data and process mining techniques,
- able to write small 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 neural networks,
- be able to evaluate of results obtained using supervised learning,
- understand and apply clustering techniques,
- construct frequent items sets,
- understand and discover association rules,
- understand and apply sequence mining,
- understand and apply process mining (both unsupervised and supervised),
- understand and apply text mining,
- 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

The course starts on 28-10-2020 and **is held online**. The online **lectures** are on Wednesdays and Thursdays from 12:30 to 14:00 and 08:30 to 10:00 and they will be prerecorded and made available before the scheduled lecture times. There will be a few Q&A sessions to explain things which are not clear.

The online **instructions** are on Fridays from 8:30 to 10:00. The instructions will also be recoded excluding the Q&A part at the end. Please use these opportunities to ask your questions! You are assumed to follow the pace of the course and prepare for the instructions. We cannot answer questions about random parts of the course in the instructions and Q&A sessions (this would be annoying for the other students).

Study guide - Introduction to Data Science (WS 2020/2021)

The course provides a broad overview of data science and engages with different sources. The slides should be self-explanatory assuming that a student attends 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. ISBN: 9780262029445, 624 pages, July 2015 (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 RWTHmoodle platform:

- Exercises
- Datasets
- Assignment (two parts).

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 your assignment with the 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.
The environment is built with Python 3.7.9 and has following software packages:

- numpy
- scipy
- pandas
- matplotlib
- scikit-learn
- nltk

Examination

The exam consists of two parts: one **assignment** delivered in two parts (each counting 20%) and the **written exam** which counts for the remaining 60% of the final result. The assignment has to be done in groups of 2-3 students. Successful participation in the assignment is required for participation in the final exam. To pass the course it is required to pass both the assignment and the test. It means that you should obtain at least 50% for the assignment and at least 50% for the final written exam.

- **The written exam** (60%): questions to test theoretical knowledge of the algorithms and techniques learned in the course:
  - First option (PT1): 01/03/2021 at 08:00 (tentative)
  - Second option (PT2): 27/03/2021 at 16:00 (tentative)
    Please note, that the dates and times are tentative and can be changed.
- **IDS Assignment Part 1** (20%): analysis of real-life and synthetic datasets using techniques and tools discussed in the course. This part tests an understanding of the material given in lectures 1-8. The deadline is on Wednesday 23/12/2020 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. The part is focused is on the material from lectures 9-21. The deadline is on Friday 12/02/2021 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. In case of proven plagiarism, all group members fail the assignment. The case will be reported and this may lead to your expulsion from the university.

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 a Wahlpflichtfach for Informatik, Media Informatics and Software Engineering. Other students are welcome to participate, but it is up to the management and rules of the corresponding programs to decide whether the course "counts". If you have problems with the RWTHonline or the RWTHmoodle, which are not specific for this course, please contact the persons responsible for these systems and not the lecturer.

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 possible bachelor and master theses. Use an opportunity to ask questions during the instructions and lectures.

Timetable of lectures and instructions

#	Lecture	date	day
Lecture 1	Introduction	28/10/2020	Wednesday
Instruction 1	Python	29/10/2020	Thursday
Instruction 2	Crash Course in Python	30/10/2020	Friday
Lecture 2	Basic data visualization/exploration	04/11/2020	*Wednesday*
Lecture 3	Decision trees	05/11/2020	Thursday
Instruction 3	Decision trees and data visualization/exploration	06/11/2020	Friday
Lecture 4	Regression	11/11/2020	Wednesday
Lecture 5	Support vector machines	12/11/2020	Thursday
Instruction 4	Regression and support vector machines	13/11/2020	Friday
Lecture 6	Neural networks (1/2)	18/11/2020	Wednesday
Lecture 7	Neural networks (2/2)	19/11/2020	Thursday
Instruction 5	Neural networks	20/11/2020	Friday
Lecture 8	Evaluation of supervised learning problems	25/11/2020	Wednesday
Instruction 6	Neural networks and evaluation	27/11/2020	Friday
Lecture 9	Clustering	02/12/2020	Wednesday
Lecture 10	Frequent items sets	03/12/2020	Thursday
Instruction 7	Clustering and frequent item sets	04/12/2020	Friday
Lecture 11	Association rules	09/12/2020	Wednesday
Lecture 12	Sequence mining	10/12/2020	Thursday
Instruction 8	Association rules and sequence mining	11/12/2020	Friday
Lecture 13	Process mining (unsupervised)	16/12/2020	Wednesday
Lecture 14	Process mining (supervised)	17/12/2020	Thursday
Instruction 9	Q&A assignment 1	18/12/2020	Friday
Lecture 15	Text mining (1/2)	06/01/2021	Wednesday
Lecture 16	Text mining (2/2)	07/01/2021	Thursday
Instruction 10	Process mining	08/01/2021	Friday
Lecture 17	Data preprocessing, data quality, binning, etc.	13/01/2021	Wednesday
Lecture 18	Visual analytics & information visualization	14/01/2021	Thursday
Instruction 11	Text mining	15/01/2021	Friday
Lecture 19	Responsible data science (1/2)	20/01/2021	Wednesday
Lecture 20	Responsible data science (2/2)	21/01/2021	Thursday
Instruction 12	Preprocessing and visualization	22/01/2021	Friday
Lecture 21	Big data	27/01/2021	Wednesday
Instruction 13	Responsible data science	28/01/2021	Thursday
Instruction 14	Big data (1/2)	29/01/2021	Friday
Lecture 22	Closing	03/02/2021	Wednesday
Instruction 15	Big data (2/2)	04/02/2021	Thursday
Instruction 16	Q&A assignment 2	05/02/2021	Friday
Instruction 17	Example exam questions	10/02/2021	Wednesday
Instruction 18	Questions	11/02/2021	Thursday