Chair of Process
and Data Science

RWTH AACHEN
UNIVERSITY

# Study Guide

## Introduction to Data Science

Wil van der Aalst, Mahsa Bafrani, Marco Pegoraro, Majid Rafiei, Yaguang Sun, and Anja Syring

IDS-2018/2019

# Study Guide Introduction to Data Science (WS 2018/2019)

## Lecturers

- Wil van der Aalst (lectures)
- Mahsa Bafrani (instructions and assignments)
- Marco Pegoraro (instructions and assignments)
- Majid Rafiei (instructions and assignments)
- Yaguang Sun (instructions and assignments)
- Anja Syring (instructions and assignments)

## Course Contents and Motivation

In recent years, data science emerged as a new and important discipline. It can be viewed as an amalgamation of classical disciplines like statistics, data mining, databases, and distributed systems. Existing approaches need to be combined to turn abundantly available data into value for individuals, organizations, and society. Moreover, new challenges have emerged, not just in terms of size ("Big Data") but also in terms of the questions to be answered.

The course aims to provide a *comprehensive overview* of data science and expose students to real-life data sets and tools. The course provides three angles on data science:

1. Data science infrastructure concerned with volume and velocity. Topics include instrumentation, big data infrastructures and distributed systems, databases and data management, and programming, and the main challenge is to make making things scalable and instant.
2. Data science analysis concerned with extracting knowledge from data. Topics include statistics, data/process mining, machine learning/artificial intelligence, operations research, algorithms, and visualization, and the main challenge is to provide answers to known and unknown unknowns.
3. Data science effects concerned with people, organizations, and society. Topics include ethics & privacy, IT law, human-technology interaction, operations management, business models, entrepreneurship, and the main challenge is to do all of the above in a responsible manner.

The main focus will be on the second angle, and a broad range of data science approaches will be presented. There is an urgent need for data science experts in all economic sectors. Organizations are desperately seeking data scientists and with the growing importance of data and digitalization, this will continue to be the case for many years to come. One could even argue that in the future we need more data scientists than computer scientists. Data will always need to be interpreted in context and as technology progresses the boundary between what people do and algorithms do will be pushed further. This generates a continuous stream of interesting, and highly relevant, questions. Also in research, we see that data science has become the key differentiator. Data science is changing other sciences and at the same time is developing itself.

Hence, this is the right time to become a data scientist. However, it is not easy to combine the different disciplines. Therefore, a data scientist can be seen a sheep with five legs. This course lays the basis for becoming such an all-round "data wizard."

## 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/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, and
- have hands-on experience using a variety of data sets provided.

## Organization & Lecture Material

The course starts on 10-10-2018. Lectures are on normally Wednesdays and Thursdays from 8.30 to 10.00 in Aula 2 (2352|021). Instructions are on Fridays from 8.30 to 10.00 also in Aula 2 (2352|021).

Since this course provides a broad overview of data science, different sources of lecture material will be used. The slides will be self-contained assuming that students attend the lectures and instructions. The following two books are highly recommended as a basis (different parts will be used):

- **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 and www.processmining.org:

- Exercises,
- Data sets, and
- Assignments.

## Software

During the first instruction of the course, we will give detailed installation procedures for all the software. You will also find the procedures on RWTHmoodle. Here we will just give the list of the required software. For the course, you will need a working installation of Python 3.6.x with some common Data Science software packages like

- numpy
- scipy
- pandas
- scikit-learn
- jupyter

More packages will need to be installed later in the course.
Anaconda Python is the advised Python distribution. Pip is the advised package manager. JetBrains Pycharm Community Edition is the advised IDE.

## Examination

The exam consists of three parts: two assignments (Schriftliche Hausarbeit) each counting for 20% of the final result, and the final written test which counts for remaining 60% of the final result. Participation in the assignments is required for participation in the final test. Only the final test can be retaken in this semester (there will be one re-exam). Assignments can only be redone in the next academic year.

- **Final written test** (60%) Questions to test the theoretical knowledge of the algorithms and techniques learned:
  - First option (PT1): 25/02/2019 09:00 – 11:00 in Aula 2
  - Second option (PT2): 25/03/2019 09:00 – 11:00 in Aula 2
- **Schriftliche Hausarbeit / DS Assignment 1** (20%): Analysis of a real-life and/or synthetic data sets using the techniques and tools provided in the course. This assignment is used to test the understanding of the material in lectures 1-10. Deadline Sunday 09/12/2018 23:59
- **Schriftliche Hausarbeit / DS Assignment 2** (20%): Analysis of more complex data sets using various data science techniques. This includes the interpretation of the results and creatively using multiple views on the data. The focus is on the lectures 11-21. Deadline Sunday 20/01/2019 23:59

Important: participation in both Schriftliche Hausarbeiten / Assignments is a prerequisite for taking the written exam. The three parts form a whole and it is not possible to retake parts of the course, i.e., the results of the assignments expire after the exam.

## Who can take the course?

The course is a master course. It is mandatory for students taking the Data Science master (it is listed as a Wahlpflichtfach, but a requirement for doing a master thesis). It is a Wahlpflichtfach for Informatik, Media Informatics, and Software Engineering. Other students are free 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 RWTHonline, RWTHmoodle, etc., that 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/or supported in a data-driven manner. Process-centricity is combined with an array of Data Science techniques. The group's research and teaching activities can be characterized by the keywords: Data Science, Process Science, Process Mining, Business Process Management, Data Mining, Process Discovery, Conformance Checking, and Simulation.

The group has been established in the context of the Alexander von Humboldt Professorship awarded to prof.dr.ir Wil van der Aalst in 2017. The award is Germany's most prestigious and valuable prize for international researchers. The PADS group supports RWTH's strategy to further strengthen its Data Science capabilities. The group also closely collaborates with the Fraunhofer Institute for Applied Information Technology (FIT).

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.

## Planning of lectures and instructions

| # | Lecture | date | day |
|---|---|---|---|
| **Lecture 1** | Introduction | 10/10/2018 | Wednesday |
| **Lecture 2** | Crash Course in Python | 11/10/2018 | Thursday |
| *Instruction 1* | *Python* | *12/10/2018* | *Friday* |
| **Lecture 3** | Basic data visualisation/exploration | 17/10/2018 | Wednesday |
| **Lecture 4** | Decision trees | 18/10/2018 | Thursday |
| *Instruction 2* | *Decision trees  and data visualization/explora* | *19/10/2018* | *Friday* |
| **Lecture 5** | Regression | 24/10/2018 | Wednesday |
| **Lecture 6** | Support vector machines | 25/10/2018 | Thursday |
| *Instruction 3* | *Regression and support vector machines* | *26/10/2018* | *Friday* |
| **Lecture 7** | Neural networks (1/2) | 31/10/2018 | Wednesday |
| *Instruction 4* | *Neural networks and supervised learning* | *02/11/2018* | *Friday* |
| **Lecture 8** | Neural networks (2/2) | 07/11/2018 | Wednesday |
| **Lecture 9** | Evaluation of supervised learning problems | 08/11/2018 | Thursday |
| *Instruction 5* | *Neural networks and supervised learning* | *09/11/2018* | *Friday* |
| **Lecture 10** | Clustering | 14/11/2018 | Wednesday |
| **Lecture 11** | Frequent items sets | 15/11/2018 | Thursday |
| **Lecture 12** | Association rules | 21/11/2018 | Wednesday |
| **Lecture 13** | Sequence mining | 22/11/2018 | Thursday |
| *Instruction 6* | *Clustering, frequent items sets, association ru* | *23/11/2018* | *Friday* |
| **Lecture 14** | Process mining (unsupervised) | 28/11/2018 | Wednesday |
| **Lecture 15** | Process mining (supervised) | 29/11/2018 | Thursday |
| *Instruction 7* | *Process mining and sequence mining* | *30/11/2018* | *Friday* |
| **Lecture 16** | Text mining (1/2) | 05/12/2018 | Wednesday |
| *Instruction 8* | *Text mining and process mining* | *06/12/2018* | *Thursday !!* |
| **Lecture 17** | Text mining (2/2) | 12/12/2018 | Wednesday |
| **Lecture 18** | Data preprocessing, data quality, binning, etc | 13/12/2018 | Thursday |
| **Lecture 19** | Visual analytics & information visualization | 19/12/2018 | Wednesday |
| backup |  | 20/12/2018 | Thursday |
| *Instruction 9* | *Text mining, preprocessing and visualization* | *21/12/2018* | *Friday* |
| **Lecture 20** | Responsible data science (1/2) | 09/01/2019 | Wednesday |
| **Lecture 21** | Responsible data science (2/2) | 10/01/2019 | Thursday |
| *Instruction 10* | *Responsible data science* | *11/01/2019* | *Friday* |
| **Lecture 22** | Big data (1/2) | 16/01/2019 | Wednesday |
| **Lecture 23** | Big data (2/2) | 17/01/2019 | Thursday |
| *Instruction 11* | *Big data* | *18/01/2019* | *Friday* |
| **Lecture 24** | Closing | 23/01/2019 | Wednesday |
| backup |  | 24/01/2019 | Thursday |
| *Instruction 12* | *Example exam questions* | *25/01/2018* | *Friday* |
| backup |  | 30/01/2019 | Wednesday |
| backup |  | 31/01/2019 | Thursday |
| extra | *Question hour* | *01/02/2019* | *Friday* |