

# ORGANISATIONAL MATTERS

## LECTURE: UNSUPERVISED LEARNING AND EVOLUTIONARY COMPUTATION USING R

**Jakob Bossek**

MALEO Group, Department of Computer Science, Paderborn University, Germany

21<sup>st</sup> Oct, 2024

## Dr. Jakob Bossek - Lecturer

### Akademischer Rat (Assistant Professor)

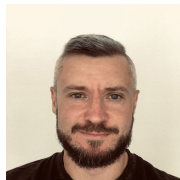
Machine Learning and Optimisation group (MALEO)

Department of Computer Science

Paderborn University

**E-Mail:** [jakob.bossek@uni-paderborn.de](mailto:jakob.bossek@uni-paderborn.de)

**Website:** <http://www.jakobbossek.de/>



### Research interests

- ▶ Heuristic Optimisation (in particular Evolutionary Algorithms)
- ▶ Theory of randomised search heuristics
- ▶ Combinatorial (Multi-Objective) Optimisation
- ▶ Evolutionary Diversity Optimisation (EDO) and Quality Diversity (QD)
- ▶ Automated Artificial Intelligence (AutoAI)

## About me ii

### Academic footprint

- ▶ 2005-2014: Studying Computer Science with minor Statistics at the TU-Dortmund University (degree: Diploma; M.Sc. equivalent)

## About me ii

### Academic footprint

- ▶ 2005-2014: Studying Computer Science with minor Statistics at the TU-Dortmund University (degree: Diploma; M.Sc. equivalent)
- ▶ 09/2008-05/2013: Studying Statistics with minor Computer Science at the TU-Dortmund University (degree: B.Sc.)

## About me ii

### Academic footprint

- ▶ 2005-2014: Studying Computer Science with minor Statistics at the TU-Dortmund University (degree: Diploma; M.Sc. equivalent)
- ▶ 09/2008-05/2013: Studying Statistics with minor Computer Science at the TU-Dortmund University (degree: B.Sc.)
- ▶ 02/2015-09/2019: Research Associate at the Department of Information Systems (Chair for Statistics and Optimization), University of Münster, Germany

## About me ii

### Academic footprint

- ▶ 2005-2014: Studying Computer Science with minor Statistics at the TU-Dortmund University (degree: Diploma; M.Sc. equivalent)
- ▶ 09/2008-05/2013: Studying Statistics with minor Computer Science at the TU-Dortmund University (degree: B.Sc.)
- ▶ 02/2015-09/2019: Research Associate at the Department of Information Systems (Chair for Statistics and Optimization), University of Münster, Germany
- ▶ 11/2018: Doctoral degree in Information Systems (Dr. rer. pol.) at the University of Münster, Germany (supervisors: Prof. Dr. Heike Trautmann, Prof. Dr. Frank Neumann)

## About me ii

### Academic footprint

- ▶ 2005-2014: Studying Computer Science with minor Statistics at the TU-Dortmund University (degree: Diploma; M.Sc. equivalent)
- ▶ 09/2008-05/2013: Studying Statistics with minor Computer Science at the TU-Dortmund University (degree: B.Sc.)
- ▶ 02/2015-09/2019: Research Associate at the Department of Information Systems (Chair for Statistics and Optimization), University of Münster, Germany
- ▶ 11/2018: Doctoral degree in Information Systems (Dr. rer. pol.) at the University of Münster, Germany (supervisors: Prof. Dr. Heike Trautmann, Prof. Dr. Frank Neumann)
- ▶ 10/2019-08/2020: PostDoc Researcher at the School of Computer Science, The University of Adelaide, Australia in the Optimisation and Logistics group of Prof. Dr. Frank Neumann

## About me ii

### Academic footprint

- ▶ 2005-2014: Studying Computer Science with minor Statistics at the TU-Dortmund University (degree: Diploma; M.Sc. equivalent)
- ▶ 09/2008-05/2013: Studying Statistics with minor Computer Science at the TU-Dortmund University (degree: B.Sc.)
- ▶ 02/2015-09/2019: Research Associate at the Department of Information Systems (Chair for Statistics and Optimization), University of Münster, Germany
- ▶ 11/2018: Doctoral degree in Information Systems (Dr. rer. pol.) at the University of Münster, Germany (supervisors: Prof. Dr. Heike Trautmann, Prof. Dr. Frank Neumann)
- ▶ 10/2019-08/2020: PostDoc Researcher at the School of Computer Science, The University of Adelaide, Australia in the Optimisation and Logistics group of Prof. Dr. Frank Neumann
- ▶ 09/2020-04/2022: PostDoc (Akademischer Rat auf Zeit) at the Department of Information Systems (Chair for Statistics and Optimization), University of Münster, Germany



## About me iii

### Academic footprint

- ▶ 04/2022-11/2023: Assistant Professor (Akademischer Rat auf Zeit) at the Department of Computer Science (Chair for AI Methodology), RWTH Aachen, Germany

## About me iii

### Academic footprint

- ▶ 04/2022-11/2023: Assistant Professor (Akademischer Rat auf Zeit) at the Department of Computer Science (Chair for AI Methodology), RWTH Aachen, Germany
- ▶ **Since 11/2023:** Assistant Professor (Akademischer Rat) at the Department of Computer Science (Machine Learning and Optimisation Group; short MALEO), Paderborn University, Germany

## Dr. Urban Škvorc - Main Responsible for Tutorials

### Postdoctoral Researcher

Machine Learning and Optimisation group (MALEO)

Department of Computer Science

Paderborn University

**E-Mail:** urban.skvorc@uni-paderborn.de



### Research interests

- ▶ Benchmarking
- ▶ Exploratory Landscape Analysis (ELA)
- ▶ Black-box optimisation
- ▶ Automated Artificial Intelligence (AutoAI)

What is it all about?

## The Epoch of Data

Undeniably we are living in the epoch of (*big*) *data*:

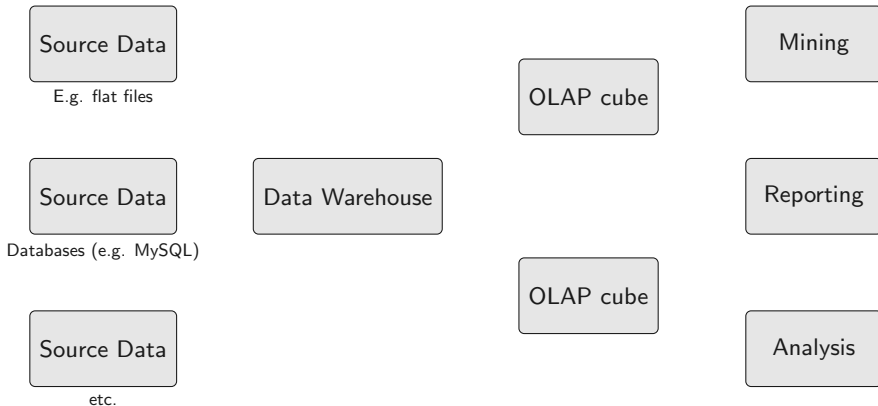
- ▶ Massive amounts of data is being collected ubiquitously: online transactions, online footprints, sensor data, any kind of measurements etc.
- ▶ According to IBM: by 2020, the overall amount of data is about 40 **zetabytes**.<sup>1</sup>
- ▶ Data is basically a set of observations (numbers, text etc.)
- ▶ Need for sophisticated methods to make sense of it!

---

<sup>1</sup> One zetabyte = 1 trillion gigabytes!

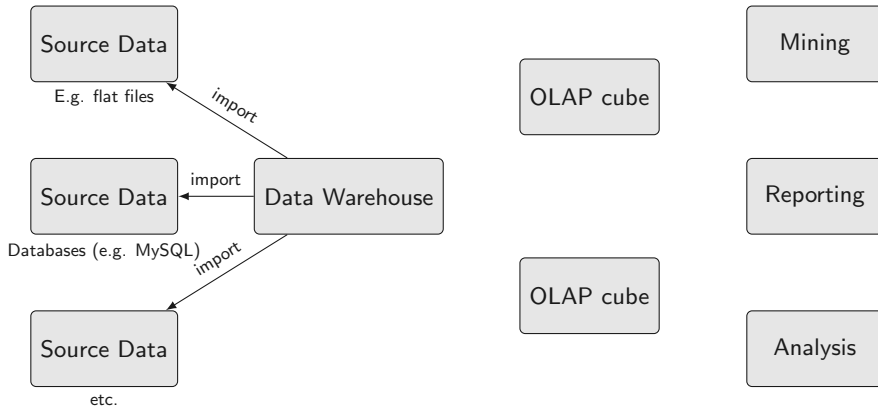
# OLAP (Online Analytical Processing)

**In a nutshell:** the process of gathering, storing and analysing data.



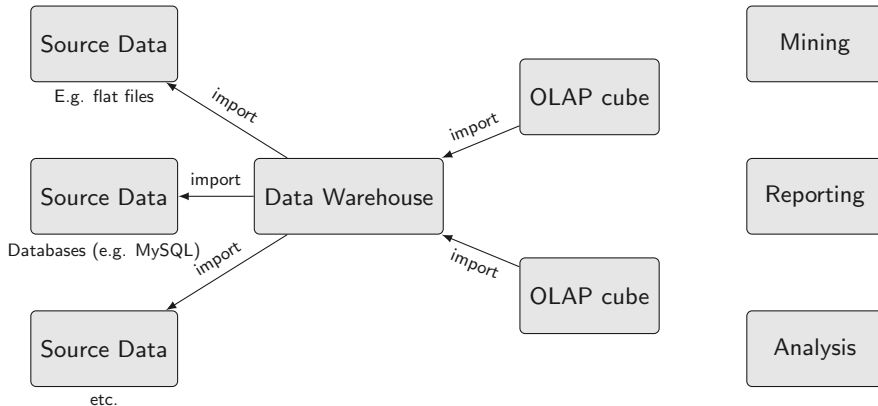
# OLAP (Online Analytical Processing)

**In a nutshell:** the process of gathering, storing and analysing data.



# OLAP (Online Analytical Processing)

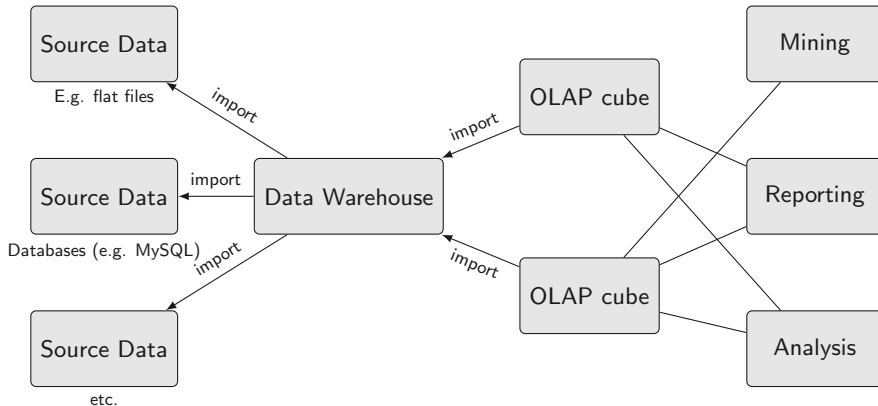
**In a nutshell:** the process of gathering, storing and analysing data.





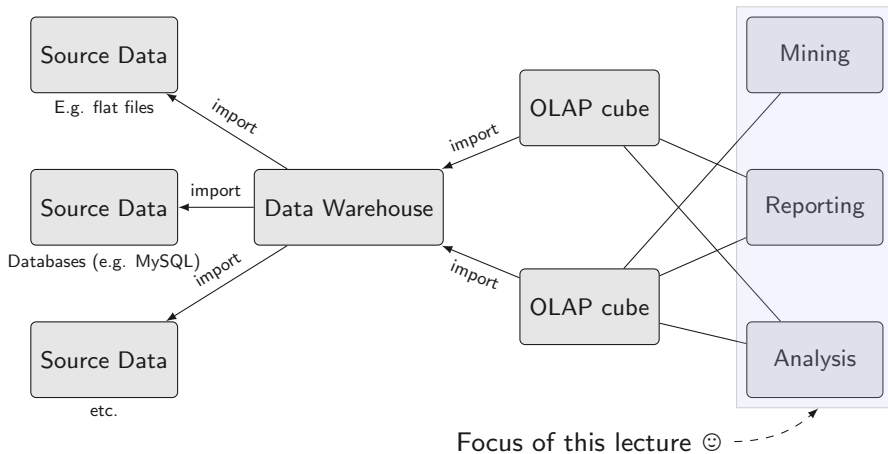
# OLAP (Online Analytical Processing)

**In a nutshell:** the process of gathering, storing and analysing data.



# OLAP (Online Analytical Processing)

**In a nutshell:** the process of gathering, storing and analysing data.



## Example (Data Analytics)

Starting point:  $N$  observations of  $p$  features or variables.

## Example (Data Analytics)

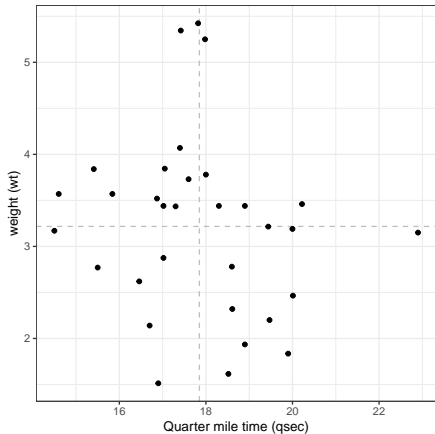
Starting point:  $N$  observations of  $p$  features or variables.

mtcars data: collected for 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for different automobiles.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
Duster 360	14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

## Example (Data Analytics) II

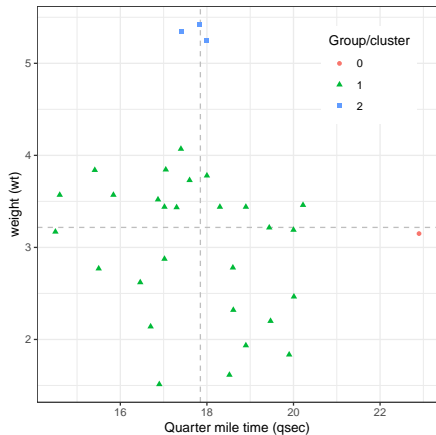
**Goal:** "Make sense" of the raw data! For example:



► Basic visualization

## Example (Data Analytics) II

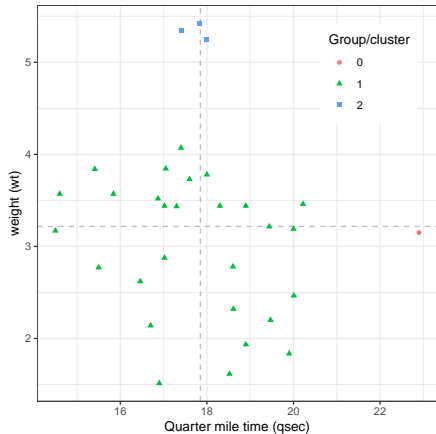
**Goal:** "Make sense" of the raw data! For example:



- Basic visualization
- Group identification

## Example (Data Analytics) II

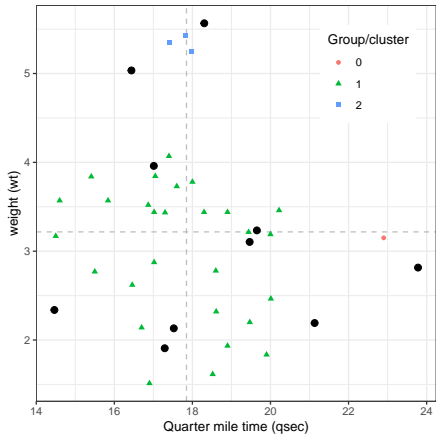
**Goal:** "Make sense" of the raw data! For example:



- ▶ Basic visualization
- ▶ Group identification
- ▶ Outlier detection

## Example (Data Analytics) II

**Goal:** "Make sense" of the raw data! For example:

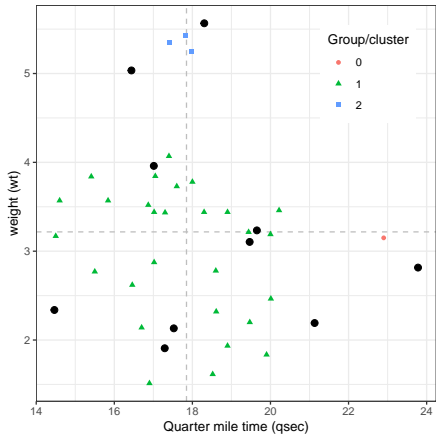


- ▶ Basic visualization
- ▶ Group identification
- ▶ Outlier detection
- ▶ Handling stream data



## Example (Data Analytics) II

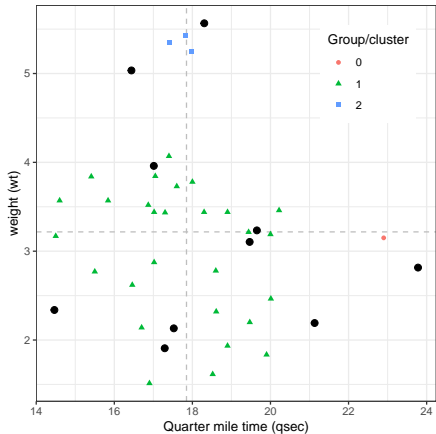
**Goal:** "Make sense" of the raw data! For example:



- ▶ Basic visualization
- ▶ Group identification
- ▶ Outlier detection
- ▶ Handling stream data
- ▶ Awareness that basically everything is problem-specific! 😊

## Example (Data Analytics) II

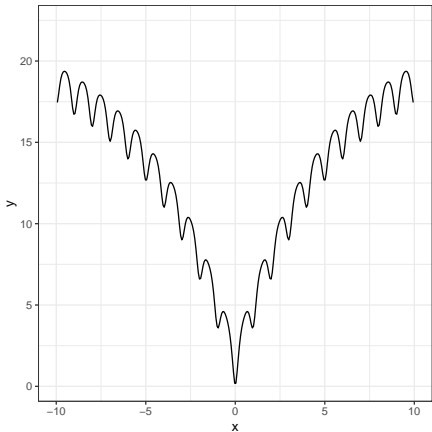
**Goal:** "Make sense" of the raw data! For example:



- ▶ Basic visualization
- ▶ Group identification
- ▶ Outlier detection
- ▶ Handling stream data
- ▶ Awareness that basically everything is problem-specific! 😊
- ▶ etc.

## Example (Optimization)

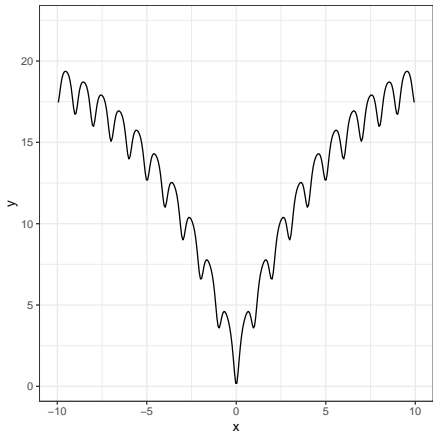
**Goal:** Find optimal/“good” value(s) that optimize some given target(s).



- Optimization basics: classes of problems, hardness

## Example (Optimization)

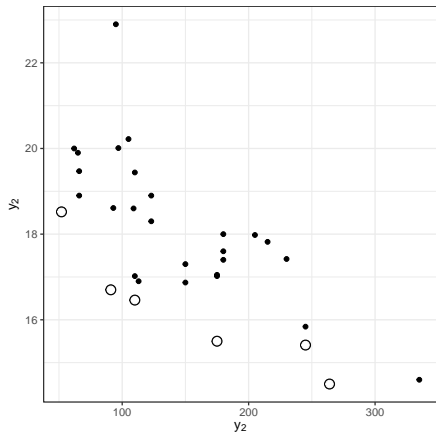
**Goal:** Find optimal/“good” value(s) that optimize some given target(s).



- ▶ Optimization basics: classes of problems, hardness
- ▶ Single-objective problems

## Example (Optimization)

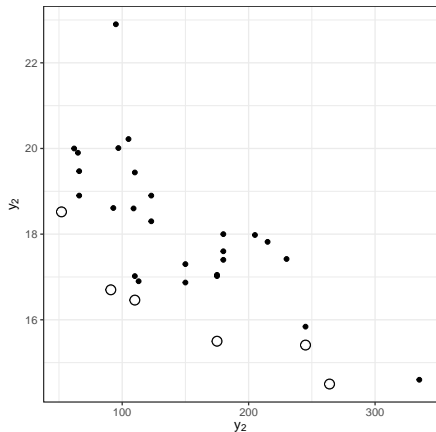
**Goal:** Find optimal/“good” value(s) that optimize some given target(s).



- ▶ Optimization basics: classes of problems, hardness
- ▶ Single-objective problems
- ▶ Multi-objective problems

## Example (Optimization)

**Goal:** Find optimal/“good” value(s) that optimize some given target(s).



- ▶ Optimization basics: classes of problems, hardness
- ▶ Single-objective problems
- ▶ Multi-objective problems
- ▶ Important class of algorithms: (bio-inspired) randomized search heuristics

## Overview of Topics Covered in ULEOR

- ▶ Data preparation: collect, merge and transform data (make it "tidy").
- ▶ Exploratory data analysis (EDA): basic (multi-variate) visualizations and summary statistics.
- ▶ Data imputation: How to deal with missing data?
- ▶ Dimensionality reduction: How to reduce high-dimensional data to a representative low-dimensional representation?
- ▶ Unsupervised learning: How do detect groups/clusters in data? I.e. similar observations.
- ▶ (Multi-objective) optimization: How to find optimal or at least good enough parameter values?
- ▶ Main tool: statistical programming language R.

## Overview of Topics not Covered in ULEOR

- ▶ Further unsupervised learning: e.g. reinforcement learning
- ▶ Supervised learning: Regression and classification. How do we make predictions for labelled data?
- ▶ Deep Learning: How do (deep) neural networks work?
- ▶ Parameter tuning/configuration: how to set the parameters of my machine learning algorithm to obtain the best possible performance?
- ▶ Feature selection: How to select a subset of my variables for learning to make my models simple yet high-performing?



## Further Topics not Covered in ULEOR

- ▶ Data Warehousing (DWH) concepts
- ▶ Databases (relational algebra, SQL<sup>2</sup>, modern architectures concepts)
- ▶ The Online Analytical Processing (OLAP) pipeline as a whole
- ▶ Business Intelligence (BI) tools like: Microsoft PowerBI, Oracle Business Intelligence Enterprise Edition etc.

---

<sup>2</sup> Structured Query Language

## General Information

# General Information I

## Lecture

- ▶ In-person lectures/tutorials in F1.101.
- ▶ Lecture: Mondays, 1:15pm - 3:45pm
- ▶ Please (at least try to) arrive here in time!!!

## General Information II

### Participation

- ▶ Do not just "consume" my presentations: do not hesitate to ask questions.<sup>3</sup>
- ▶ Use the discussion forum in PANDA  
(ask and respond, search for learning groups etc.)
- ▶ Search for more information online
- ▶ Invest time on a weekly basis!
- ▶ Aim to understand the tools and methods. I.e., do not just memorize "recipes"

---

<sup>3</sup> I am sure many other students will have the same questions, but won't ask (for some reason).

## General Information III

### Get in touch

- ▶ Use the PANDA discussion forum to ask questions that might be of interest for all or at least some other students
- ▶ Via mail for individual questions

## General Information IV

### Tutorials

- ▶ Exercise sheets will be uploaded weekly on Tuesday (first one October 22<sup>nd</sup>)
- ▶ Mixture of applied and "theoretical" exercises.
- ▶ Discussion of (hopefully *your*) results the week after
- ▶ Exemplary solutions will be made available

## General Information IV

### Tutorials

- ▶ Exercise sheets will be uploaded weekly on Tuesday (first one October 22<sup>nd</sup>)
- ▶ Mixture of applied and "theoretical" exercises.
- ▶ Discussion of (hopefully *your*) results the week after
- ▶ Exemplary solutions will be made available
- ▶ The good thing: working on exercise sheets is optional 😊

## General Information IV

### Tutorials

- ▶ Exercise sheets will be uploaded weekly on Tuesday (first one October 22<sup>nd</sup>)
- ▶ Mixture of applied and "theoretical" exercises.
- ▶ Discussion of (hopefully *your*) results the week after
- ▶ Exemplary solutions will be made available
- ▶ The good thing: working on exercise sheets is optional 😊
- ▶ But the bad thing is: you will fail to pass the exam with overwhelming probability if you skip working actively on the tutorials ☹



## General Information V

### Exam

- ▶ Written exam at the end of the lecture period (100% of the final grade)
- ▶ Duration: 90 minutes
- ▶ Content: Mix-up of applied exercises (e.g. apply algorithm  $A$  on some toy problem), reproducible (e.g. explain how clustering algorithm  $A$  works in your own words), multiple choice, maybe a R related exercise and maybe some theoretical stuff (show that something holds)
- ▶ All covered topics are relevant!
- ▶ Best way to prepare: learn actively, do the exercises etc.

## General Information VI

### Do I need programming skills?

- ▶ This is an applied course: we will use R
- ▶ Short introduction<sup>4</sup> into R
- ▶ In the final exam: 1-2 R exercise(s)  $\leadsto$  you can pass without knowledge of R
- ▶ However: **you should teach yourself the basics!** Why?
  - ▶ R gains increasing interest in industry (pimp your CV 😊)
  - ▶ R really shines when it comes to data analysis (very active community)
  - ▶ It is always good to learn a new programming language

---

<sup>4</sup> This is not a course on programming language basics.

How to prepare/postprocess the lectures/tutorials

# Lecture Notes I

Lecture Notes for Master Degree

## Data Analytics I

Winter Term 2021/2022

Dr. Jakob Bossek

Last changed: October 4, 2021

### Abstract

Lecture notes for the Information Systems Master degree lecture *Data Analytics I* in winter term 2021/2022.

Deliberately plagiarized from module description ©: “This course focuses on multivariate statistical methods in the context of data mining.

- ▶ Kind of a lecture-specific book with additional information and exercises.
- ▶ Not finished! Will evolve in the course of the winter term.
- ▶ Not perfect! Read with care. Please report any typos, unclear formulations, ideas for improvement etc.

Call for participation!

# Call for participation!

- ▶ Report typos
- ▶ Give constructive feedback on how to improve the manuscript
- ▶ Discuss stuff with others in the Learnweb
- ▶ There are many additional exercises with no sample solutions: send me your solutions
- ▶ **Benefit for you:** fame and honor, my gratitude, no price or whatever (sorry! 😊)

## How to post-process?

### Simple fact

You have no chance to pass the module without intense and active learning!

### How to avoid failing

- ▶ Ideally before and after a lecture: read the corresponding lecture notes chapter(s) / book chapters

## How to post-process?

### Simple fact

You have no chance to pass the module without intense and active learning!

### How to avoid failing

- ▶ Ideally before and after a lecture: read the corresponding lecture notes chapter(s) / book chapters
- ▶ Try to solve the exercises! (this can take a while; do not give up quickly and google the solution  $\leadsto$  you cheat yourself this way)



## How to post-process?

### Simple fact

You have no chance to pass the module without intense and active learning!

### How to avoid failing

- ▶ Ideally before and after a lecture: read the corresponding lecture notes chapter(s) / book chapters
- ▶ Try to solve the exercises! (this can take a while; do not give up quickly and google the solution  $\leadsto$  you cheat yourself this way)
- ▶ Naturally, some things will be hard to understand:

## How to post-process?

### Simple fact

You have no chance to pass the module without intense and active learning!

### How to avoid failing

- ▶ Ideally before and after a lecture: read the corresponding lecture notes chapter(s) / book chapters
- ▶ Try to solve the exercises! (this can take a while; do not give up quickly and google the solution  $\leadsto$  you cheat yourself this way)
- ▶ Naturally, some things will be hard to understand:
  - ▶ Ask questions: lecturer, learnweb, stackoverflow etc.

## How to post-process?

### Simple fact

You have no chance to pass the module without intense and active learning!

### How to avoid failing

- ▶ Ideally before and after a lecture: read the corresponding lecture notes chapter(s) / book chapters
- ▶ Try to solve the exercises! (this can take a while; do not give up quickly and google the solution  $\leadsto$  you cheat yourself this way)
- ▶ Naturally, some things will be hard to understand:
  - ▶ Ask questions: lecturer, learnweb, stackoverflow etc.
  - ▶ Actively search for further sources online (e.g., on YouTube)

## How to post-process?

### Simple fact

You have no chance to pass the module without intense and active learning!

### How to avoid failing

- ▶ Ideally before and after a lecture: read the corresponding lecture notes chapter(s) / book chapters
- ▶ Try to solve the exercises! (this can take a while; do not give up quickly and google the solution  $\leadsto$  you cheat yourself this way)
- ▶ Naturally, some things will be hard to understand:
  - ▶ Ask questions: lecturer, learnweb, stackoverflow etc.
  - ▶ Actively search for further sources online (e.g., on YouTube)
  - ▶ In R: make mistakes, google error messages etc.

## How to post-process?

### Simple fact

You have no chance to pass the module without intense and active learning!

### How to avoid failing

- ▶ Ideally before and after a lecture: read the corresponding lecture notes chapter(s) / book chapters
- ▶ Try to solve the exercises! (this can take a while; do not give up quickly and google the solution  $\leadsto$  you cheat yourself this way)
- ▶ Naturally, some things will be hard to understand:
  - ▶ Ask questions: lecturer, learnweb, stackoverflow etc.
  - ▶ Actively search for further sources online (e.g., on YouTube)
  - ▶ In R: make mistakes, google error messages etc.

Course achievement and exam

# Exam Modalities

## Course achievement

- ▶ Case study in R
- ▶ Mid december; to be submitted mid January
- ▶ Submission of running and documented R-code
- ▶ Submission of summarising report in PDF format (using our  $\text{\LaTeX}$ -template)
- ▶ **Has to be passed to be admitted to exam**

# PAUL (de)registrations

## Course achievement

- ▶ If you did not pass the course achievement, you will be de-registered from the exam automatically

## Exam

- ▶ 1<sup>st</sup>-phase registration: 21 Oct until 21 Nov 2024
- ▶ 2<sup>nd</sup>-phase registration: 24 Feb until 7 Mar 2025
- ▶ De-registration from exam is possible until two days before the examination date!



# Literature Recommendations

## Literature: Machine Learning

1. Gareth James, Daniela Witten, Trevor John Hastie & Robert Tibshirani (2021). An Introduction to Statistical Learning. 2nd Edition. Springer. (James et al. 2013)★  
~> Focus on understanding, excellent introductory text to grasp the working principles.
2. Trevor John Hastie, Robert Tibshirani & Jerome Harold Friedman (2009). The Elements of Statistical Learning. 2<sup>nd</sup> Edition, Springer. (Hastie, Tibshirani, and Friedman 2009)  
~> Advanced computer science perspective, focus on mathematics.
3. Ethem Alpaydin (2014). Introduction to Machine Learning. MIT press. (Alpaydin 2014)  
~> Rather advanced.

## Literature: Data Analysis with R

1. Hadley Wickham & Garrett Golemund (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. O'Reilly. (Wickham and Golemund 2017)
2. Hadley Wickham (2016). ggplot2: Elegant Graphics for Data Analysis. Springer. (Wickham 2009)
3. Hadley Wickham (2014). Advanced R. Chapman & Hall/CRC The R Series. Taylor & Francis. (Wickham 2014)
4. Hadley Wickham (2015). R Packages. 1<sup>st</sup> Edition. O'Reilly Media. (Wickham 2015)

Wrap-Up

# Wrap-Up

## Today's content

Mainly organizational stuff ☺

---

<sup>5</sup> URL: <https://www.r-project.org>

<sup>6</sup> URL: <https://www.rstudio.com>

# Wrap-Up

## Today's content

Mainly organizational stuff ☺

## Your task(s)

- ▶ Download and install both R<sup>5</sup> and R Studio<sup>6</sup>.
- ▶ Optional: Check the introduced literature (recall that most of this stuff is free).
- ▶ Optional: Read the introduction of An Introduction to Statistical Learning (James et al. 2013).

---

<sup>5</sup> URL: <https://www.r-project.org>

<sup>6</sup> URL: <https://www.rstudio.com>

## References I

- James, Gareth et al. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer.
- Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer. ISBN: 9780387848570.
- Alpaydin, Ethem (2014). *Introduction to Machine Learning*. 3rd ed. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press. ISBN: 978-0-262-02818-9.
- Wickham, Hadley and Garrett Golemund (Jan. 2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1st ed. O'Reilly Media. ISBN: 1491910399.
- Wickham, Hadley (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN: 978-0-387-98140-6.
- (2014). *Advanced R*. Chapman & Hall/CRC The R Series. Taylor & Francis. ISBN: 9781466586963. URL: <https://books.google.de/books?id=PFHFNAEACAAJ>.
- (2015). *R Packages*. 1st. O'Reilly Media, Inc. ISBN: 1491910593.