

Data Science

01: Organizational and Introduction

Data Science

Organizational: About me



`klaus.kaiser@fh-dortmund.de`

Room: B.2.04

Prof. Dr. Klaus Kaiser

- Data science, computer science
- Professor since May 2024
- Worked as data scientist for different companies
 - Insurance, IoT, consulting
- Other lectures:
 - "Adaptive Systeme"
 - "Data Science Datenbanken"
 - "Data Science Projekt"

1 Introduction

- Example data science project
- Some real world examples

2 Overview on the lecture

- Content
- Dates and exam
- Some notes

3 How to continue?

4 Summary & References

Introduction

Data Science

Introduction: What do you think?

we
focus
on
students



Data Science

Introduction: What is data science?

”Data science turns raw data into meaningful insights, transforming information into impact.”

Data Science

Introduction: What is data science?

”Data science turns raw data into meaningful insights, transforming information into impact.”

ChatGPT

Data Science

Introduction: What is data science?

Unfortunately, there is no consent on the definition of "data science" ...

Data Science

Introduction: The term data science

- 1962 • John Tukey describes a similar field as "data analysis".
- 1985 • Jeff Wu uses the term "data science" as a description of statistics.
- 1996 • Peter Naur proposes the term "data science" as an alternative for computer science.
- 1990s • The terms "data mining" and "knowledge discovery" are used for typical data science processes.
- 2001 • "Data Science: An Action Plan for Expanding the Technical Areas of the field of Statistics." is released by William S. Cleveland - often seen as the reference for the modern concept of data science.

Table: History of the term "data science", adapted from ^[1] and ^[2]

Data Science

Introduction: What is data science?

What is data science?

Data science is an **interdisciplinary** academic field that uses statistics, scientific computing, scientific methods, processes, scientific visualization, algorithms and systems to **extract or extrapolate knowledge and insights** from potentially noisy, structured, or unstructured **data**.^[3]

Data Science

Introduction: What is data science?

we
focus
on
students

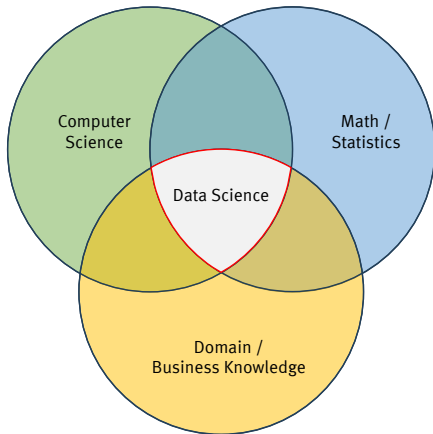


Figure: Data science Venn diagram.

Data Science

Introduction: What is data science?

In short?

Creating knowledge from data.

Data science can be used: If ...

... there is (a lot of) data, which is needed for something ...

Data Science

Introduction: Let's try

we
focus
on
students



Data Science

Introduction: Let's try

How did we got to our decision?

- Analyzing data and domain:
 - It's about driving a bicycle - linear relation

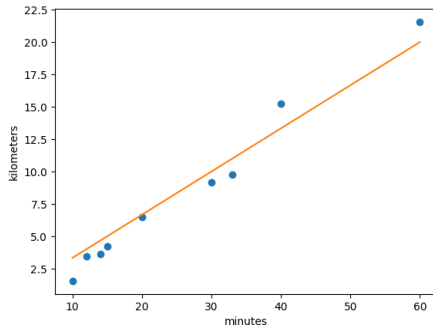
$$\text{distance} = \text{velocity} \times \text{time}$$

- Fitting a model:

$$\text{distance} \approx \frac{20}{60} \text{ km/min} \times \text{time}$$

- Making a prediction:

$$\frac{20}{60} \text{ km/min} \times 50 \text{ min} = 16\frac{2}{3}$$



Data Science

Introduction: Let's try

we
focus
on
students

Let's check the results

Data Science

Introduction: Related fields?

Data engineering

Data engineering refers to the building of systems to enable the collection and usage of data. This data is usually used to enable subsequent analysis and data science, which often involves machine learning. Making the data usable usually involves substantial compute and storage, as well as data processing.^[4]

Data analysis

Data analysis is the process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.^[5]

Data Science

Introduction: Tasks

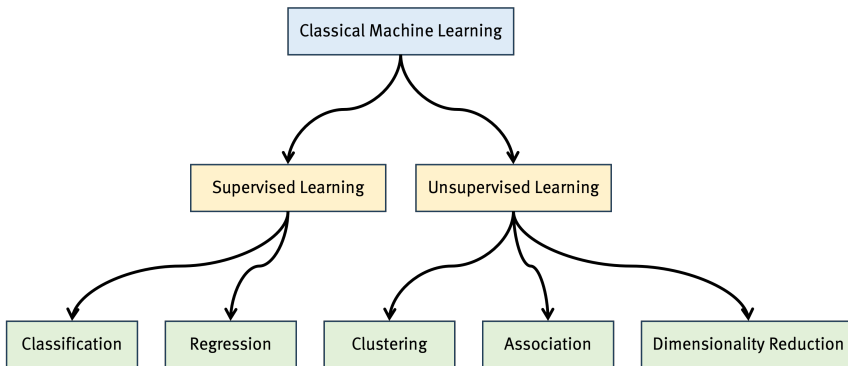


Figure: Overview on different tasks in classical machine learning.

Introduction

Example data science project

Understanding the setting

Scientists want to analyze the movement of monkeys in the jungle. For this, an automated tool to detect monkeys is needed. Discussion with a domain expert:

- Tool should be used in the forest to investigate monkeys.
- Tool should distinguish monkeys and other types of animals.
- Images are made with a camera, which is located at the top of a tree.

Goal

Our goal is to build a tool, which can detect monkeys.

Data Science

Example project: Monkey detection

Getting data

The first step is to collect data corresponding to the described setting. In this example we need images of monkeys in similar positions, made with a similar camera. Furthermore, negative examples are also needed.



Data Science

Example project: Monkey detection

Labeling / cleaning / preparation and analyzing of data

Next step is to take the collected data and label it, i.e. categorizing all images into classes. Afterwards, unneeded data is filtered, the images are prepared for the AI model and analyzed concerning the desired setting.



monkey



monkey



no_monkey

class	num. images
monkey	32 847
no_monkey	124 3213

Training a model

When the data is prepared, a model is chosen. For image recognition a neural network, especially a convolutional neural network, might be the model of choice. The model is then trained with the given data. This means the model optimizes parameters by making a prediction on a training image and then comparing the prediction with the desired value.

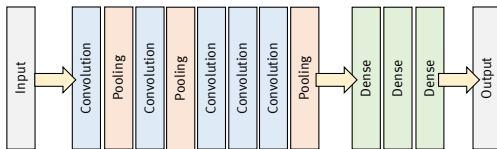


Figure: Structure of a convolutional neural network.

Testing

The trained model can now be tested. Therefore, an independent test dataset is used and the predictions of the model are compared with the exact values.

class	accuracy
monkey	98 %
no_monkey	99.8 %

Data Science

Example project: Monkey detection

we
focus
on
students

Let's take a look

Finalizing

The model is trained and tested, therefore it can now be added to the camera system. This can be done for example on device or images are sent to the cloud.

Going to the forest

The tool is finally done and can be used in the forest!



Data Science

CRISP-DM

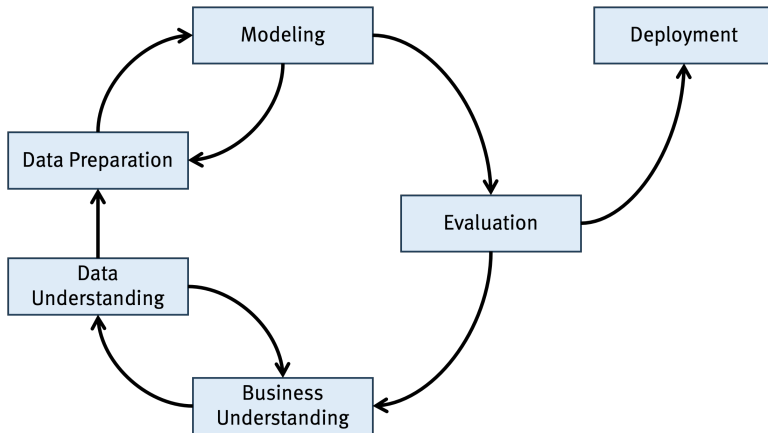


Figure: Cross-industry standard process for data mining^[6]

Introduction

Some real world examples

Data Science

Real world examples: Autonomous driving

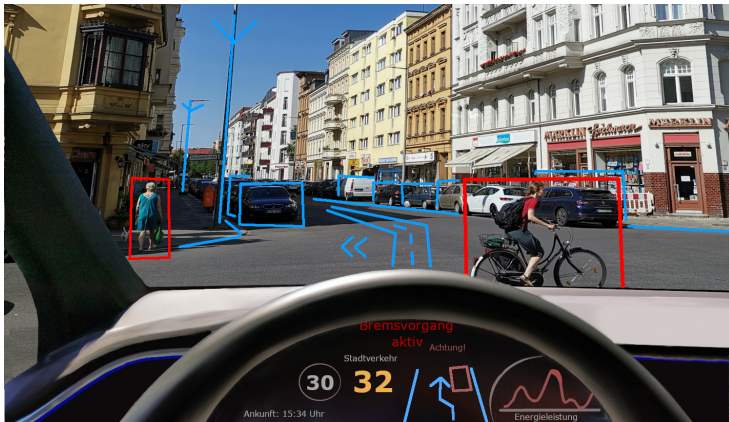


Figure: Example of the detection a autonomous car needs to perform on a street scene.

Data Science

Real world examples: Face recognition

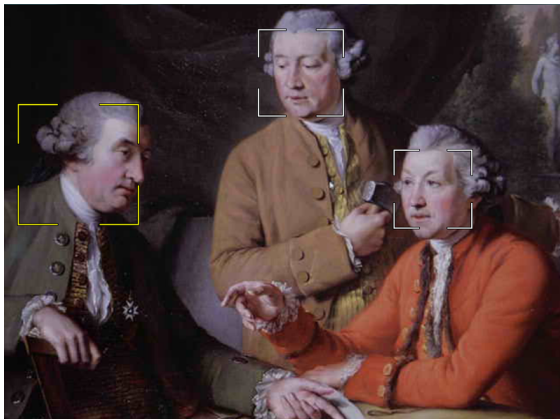


Figure: Detection of faces on a classical painting.

Data Science

Real world examples: Predictive maintenance

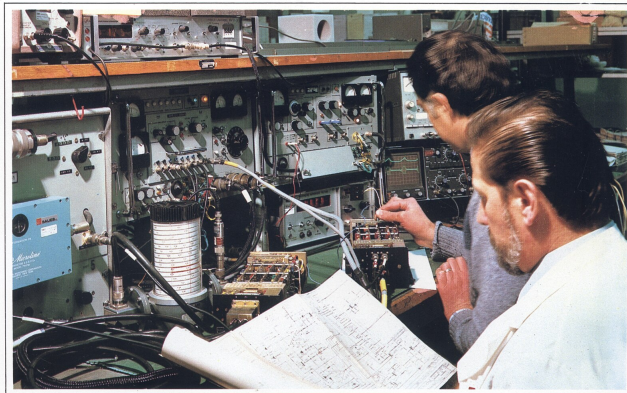


Figure: Maintenance of a device.

Data Science

Real world examples: Fraud detection



Figure: Logo of the movie "Catch me if you can" by Steven Spielberg. The movie is about an imposter performing some scam.

Data Science

Real world examples: Recommendation systems

we
focus
on
students

Related to items you've viewed [See more](#)

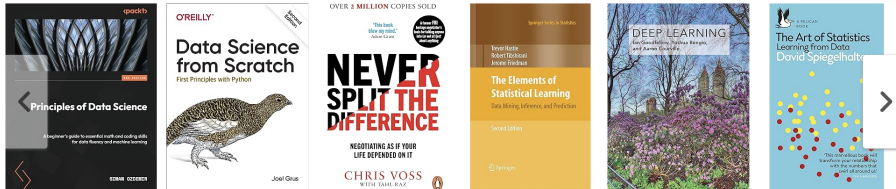


Figure: Example of recommendations, Amazon gives based on articles viewed.

Data Science

Real world examples: Detection of cancer



Figure: X-Ray image of a chest with lung cancer.

Data Science

Real world examples: ChatGPT



Figure: Answer of ChatGPT on the question: "What do we need data science for?"

Overview on the lecture

Overview on the lecture

Content

Data Science

Introduction: Content of the lecture

Data basics

First look on data, ways to extract data and steps to prepare.

Statistics basics

Analyzing data and extracting some first knowledge from it.

Presentation basics

Evaluating and presenting results / data.

Machine learning

Some first machine learning models.

Attention!

Work in Progress: Concrete topics are currently work in progress!

Information

There are a lot of books concerning data science and surrounding topics. The following books are examples:

- some of them are used for the lecture
- most of them are freely available at FH Dortmund Bibliothek



Figure: Screenshot of thalia.de: Number of results for the search term "Data Science".

Overview on the lecture

Dates and exam

Data Science

Organizational: Schedule

Lecture

Every Monday at 10:15 Room: A.E.02

- Presentation and discussion of new topics
- Language: course material in English, lecture in German

Exercise and practical training

Every Monday at 12:00 Room: B.2.20 **or** at 14:15 Room: B.2.20

- Deepening and application of lectures content
- Language: exercise material in English, exercise in German

First exercise: **Today** - crash course Python!

Data Science

Organizational: Exam

Written exam

- **100 points**
- 90 minutes
- **Language:**
 - Exercise in English
 - Answers in English or German

Bonus points: data analytics

- **10 points**
- More information coming soon!

Data Science

Organizational: Exam

we
focus
on
students



Ilias course room password

#DS-2024!

Overview on the lecture

Some notes

Data Science

Organizational: Expectations

on me

Respect, transparency & fairness
Helpfulness, openness & availability
Professional competence

on the students

Respect & consideration
Willingness to participate & openness
Personal responsibility & autonomy

Data Science

Organizational: important note 1

we
focus
on
students

Ask questions!

Data Science

Organizational: important note 2

Do not hesitate to contact me! (best is via mail or teams!)

Data Science

Organizational: important note 3

This is a new lecture: If you see any errors or ways of improvement, please contact me!

How to continue?

Data Science

Introduction: Literature

- *Data Analytics - Models and Algorithms for Intelligent Data Analysis*
T. A. Runkler - 978-3-658-29778-7
- *Data Science - Create Teams That Ask the Right Questions and Deliver Real Value* D.
Rose - 978-1-484-22252-2
- *Data Science Fundamentals*
O. Campesato - 978-1-683-92733-4
- *Predicting the Unknown*
S. Kampakis - 978-1-484-29504-5
- *Statistik - Der Weg zur Datenanalyse*
L. Fahrmeir, C. Heumann, R. Künstler, I. Pigeot, G. Tutz - 978-3-662-67525-0
- *SQL- & NoSQL-Datenbanken*
M. Kaufmann, A. Meier - 978-3-662-67091-0

Data Science

Introduction: Literature

- *The Data Science Design Manual*
S. S. Skiena - 978-3-319-55443-3
- *The Elements of Statistical Learning*
T. Hastie, R. Tibshirani, J. Friedman - 978-0-387-84857-0
- *Designing Data-Intensive Applications*
M. Kleppmann - 978-1-449-37332-0
- *Hands-On Machine Learning with Scikit-Learn, Keras, and Tensorflow: Concepts, Tools, and Techniques to Build Intelligent Systems*
A. Géron - 978-1-098-12597-4
- *Statistik: Eine Einführung mit interaktiven Elementen*
H.-J. Mittag - 978-3-662-55320-6

Data Science

Introduction: How to continue?

Title	Lecturer	Semester
Adaptive Systeme	Prof. Kaiser	8
Künstliche Intelligenz	Prof. Bab	8
Datenbanken 2	Prof. Saatz	8
Moderne Datenbanken	Prof. Saatz	8
Digitale Bildverarbeitung	Prof. Friedrich	5 / 9
Data Mining in Industrie und Wirtschaft	Prof. Kuhnt	8
Machine Learning	Prof. Friedrich	Master

Table: List of lectures in curriculum, which are related to data science.

Summary & References

Data Science

Summary

- You can explain what data science is
- You remember different data science applications

Data Science

List of images

- [https://commons.wikimedia.org/wiki/File:Golden_monkey_\(Cercopithecus_kandti\)_eating.jpg](https://commons.wikimedia.org/wiki/File:Golden_monkey_(Cercopithecus_kandti)_eating.jpg)
- https://commons.wikimedia.org/wiki/File:MonkeyJungle_03.JPG
- [https://commons.wikimedia.org/wiki/File:Jungle_at_New_York_2023_\(Jon_Vasquez\).jpg](https://commons.wikimedia.org/wiki/File:Jungle_at_New_York_2023_(Jon_Vasquez).jpg)
- https://commons.wikimedia.org/wiki/File:Shaheed_Island,_Andamans,_Interior_jungle.jpg
- [https://commons.wikimedia.org/wiki/File:Golden_monkey_\(Cercopithecus_kandti\)_eating.jpg](https://commons.wikimedia.org/wiki/File:Golden_monkey_(Cercopithecus_kandti)_eating.jpg)
- https://commons.wikimedia.org/wiki/File:MonkeyJungle_03.JPG
- https://commons.wikimedia.org/wiki/File:Shaheed_Island,_Andamans,_Interior_jungle.jpg
- <https://commons.wikimedia.org/wiki/File:Autonomous-driving-Barcelona.jpg>
- <https://commons.wikimedia.org/wiki/File:Gesichtserkennung.Gruppe.png>
- https://commons.wikimedia.org/wiki/File:TRT_Philips_avionics_equipment_maintenance_1980.jpg
- <https://de.wikipedia.org/wiki/Datei:Catchmeifyoucan-logo.svg>
- <https://commons.wikimedia.org/wiki/File:LungCancer-Xray-01.jpg>

Data Science

Endnotes

[1] https://en.wikipedia.org/wiki/Data_science

[2] <https://www.dataversity.net/brief-history-data-science/>

[3] https://en.wikipedia.org/wiki/Data_science

[4] https://en.wikipedia.org/wiki/Data_engineering

[5] https://en.wikipedia.org/wiki/Data_analysis

[6] Shearer C., The CRISP-DM model: the new blueprint for data mining, J Data Warehousing (2000); 5:13–22.

“Data science is 80% preparing data, and 20% complaining about preparing data.”

unknown