

DATA SCIENCE 2

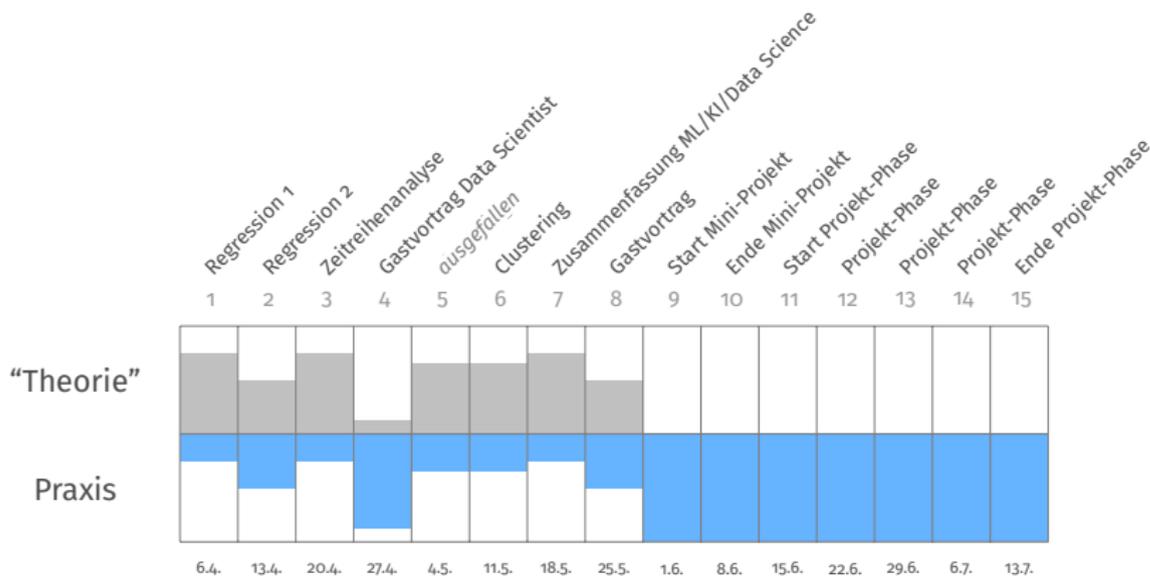
PROJEKTPHASE

PROF. DR. CHRISTIAN BOCKERMANN

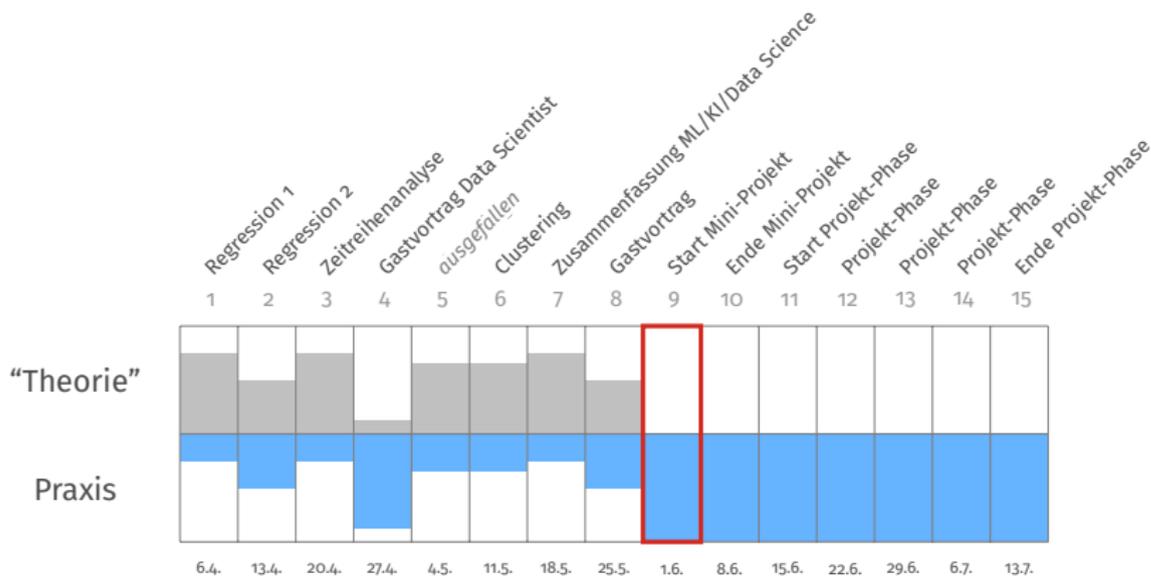
HOCHSCHULE BOCHUM

SOMMERSEMESTER 2021

Themen der Vorlesung



Themen der Vorlesung



- 1 Ziele und Organisatorisches
- 2 Der Weg zum Data Scientist
- 3 Die Kaggle Plattform
- 4 Projektphase

Ziele und Organisatorisches

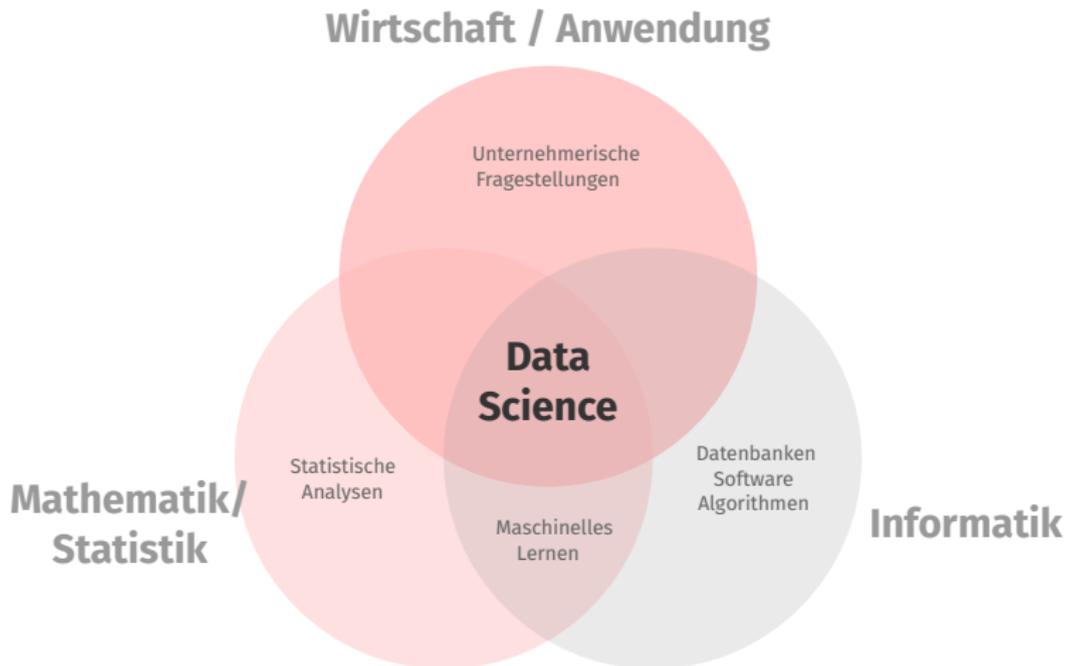
Worum geht es im Kurs Data Science?

- Etablieren von datengetriebener Denk-/Arbeitsweise
- Kennenlernen von Methoden (ML)
- Wissenschaftliche Arbeitsweise
- Vorbereitung auf BA, Beruf

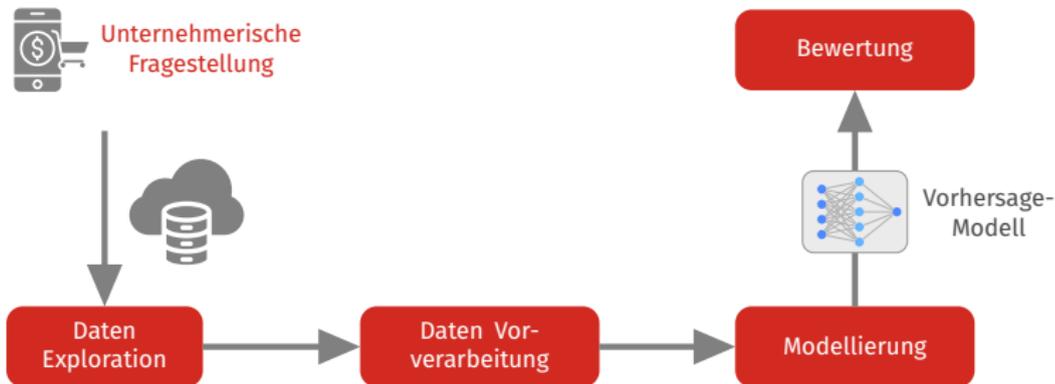
Worum geht es im Kurs Data Science?

- Etablieren von datengetriebener Denk-/Arbeitsweise
- Kennenlernen von Methoden (ML)
- Wissenschaftliche Arbeitsweise
- Vorbereitung auf BA, Beruf

- Praktische Umsetzung erlernen
- Datenbasiertes *Story-Telling*



Vorgehen bei der Datenanalyse



Data Science 1

Beispiel für wirtschaftlichen Kontext, Anwendungsfall gegeben;
vorgegebene Aufgaben:

1. Datenvorverarbeitung
2. Statistiken / Exploration
3. Modellierung / Evaluation

Data Science 2

Sie bekommen einen Datensatz.

Data Science 2

Sie bekommen einen Datensatz.

- Bewusst offene Aufgabenstellung

Data Science 2

Sie bekommen einen Datensatz.

- Bewusst offene Aufgabenstellung
- Eigenständig Datenanalyse “erarbeiten”
- Bericht/Dossier mit Beschreibung und Analyse erstellen
- Ergebnisse präsentieren
- Eigene Erfahrungen berichten

Sie haben einen Datensatz bekommen.

Und nun?

- Schauen Sie sich die Daten an
- Beschreiben, Fragestellung finden, Analysieren

Datensatz:

BestellNr	Datum	Kunde	AnzahlArtikel	Brand:A	Brand:B	Brand:C	Brand:D	Brand:E
239821	2019-12-02	3913	1	49.99	0	0	0	0
239822	2019-12-02	1024	1	0	0	21.99	0	0
239823	2019-12-03	0232	2	0	8.99	2.99	0	0
239824	2019-12-03	9218	1	0	0	0	0	13.49
239825	2019-12-04	4120	3	0	0	23.98	0	13.49

Datensatz:

BestellNr	Datum	Kunde	AnzahlArtikel	Brand:A	Brand:B	Brand:C	Brand:D	Brand:E
239821	2019-12-02	3913	1	49.99	0	0	0	0
239822	2019-12-02	1024	1	0	0	21.99	0	0
239823	2019-12-03	0232	2	0	8.99	2.99	0	0
239824	2019-12-03	9218	1	0	0	0	0	13.49
239825	2019-12-04	4120	3	0	0	23.98	0	13.49

Exploration – nach Kunde

- Durchschnittliche Kauf-Häufigkeit?
- Gibt es viele Wiederholungskäufer? In bestimmten Gruppen? (z.B. männlich/weiblich/Altersgruppe?)
- Durchschnittliche Zeit zwischen Käufen eines Kunden?

Angenommen, wir haben für jeden Kunden noch

- Alter, Geschlecht, ...
- Wohnort/Stadt

Modellierung – nach Kunde

- Welche Kunden haben länger nicht gekauft?
- Ab wann gilt ein Kunde als “verloren”?
- Können wir Kunden erkennen, die nicht mehr kaufen?

Angenommen, wir haben für jeden Kunden noch

- Alter, Geschlecht, ...
- Wohnort/Stadt

Modellierung – nach Kunde

- Welche Kunden haben länger nicht gekauft?
- Ab wann gilt ein Kunde als “verloren”?
- Können wir Kunden erkennen, die nicht mehr kaufen?

Churn Prediction!

Formalisieren des Problems

- Sei T die durchschnittliche Zeit zwischen 2 Käufen
- Kunde ist “verloren”, wenn er $2 \cdot T$ lang nicht kauft

Formalisieren des Problems

- Sei T die durchschnittliche Zeit zwischen 2 Käufen
- Kunde ist “verloren”, wenn er $2 \cdot T$ lang nicht kauft

Modellierung / Lernaufgabe:

- Merkmal verloren berechnen (0 / 1)
- Aufgabe: Binäre Klassifikation, verloren vorhersagen

Organisation

- selbstständiges Arbeiten in Kleingruppen (3-4)
- möglichst WiInf'ler und BWL/VWLER gemischt
- Abgabe als Jupyter-Notebook/Blog-Eintrag pro Person
- Präsentation als Gruppe

“The only way to learn data science, data analysis, machine learning, or artificial intelligence topics is by practicing or doing projects.

There is no other alternative to that.”

[<https://towardsdatascience.com/all-the-datasets-you-need-to-practice-data-science-skills-and-make-a-great-portfolio-857a348883b5>]

Bleiben Sie neugierig!

- Stellen Sie Fragen und versuchen Sie, diese mit Daten zu beantworten!

Bleiben Sie neugierig!

- Stellen Sie Fragen und versuchen Sie, diese mit Daten zu beantworten!

Bleiben Sie präzise!

- Welche Merkmale sind für eine Frage überhaupt vorhanden?
- Was sagt ihr mögliches Modell genau aus?

Bleiben Sie neugierig!

- Stellen Sie Fragen und versuchen Sie, diese mit Daten zu beantworten!

Bleiben Sie präzise!

- Welche Merkmale sind für eine Frage überhaupt vorhanden?
- Was sagt ihr mögliches Modell genau aus?

Bleiben Sie kritisch!

- Vorhersagefehler 0% – glauben Sie Ihrer Analyse?

Bleiben Sie neugierig!

- Stellen Sie Fragen und versuchen Sie, diese mit Daten zu beantworten!

Bleiben Sie präzise!

- Welche Merkmale sind für eine Frage überhaupt vorhanden?
- Was sagt ihr mögliches Modell genau aus?

Bleiben Sie kritisch!

- Vorhersagefehler 0% – glauben Sie Ihrer Analyse?

Lernen Sie von anderen!

- Viele Blogs/Beispiel Projekte zum Thema DataScience!
- <https://towardsdatascience.com>

Kaggle – DataScience als **Wettbewerb**



- Portal mit Data Science Challenges
- Jupyter Notebooks, Forum zum Austausch
- Für Challenges existiert Leaderboard
- Oft *accuracy* als Maß für Gewinner

Kaggle Einstieg: Titanic Challenge

Getting Started Prediction Competition

Titanic - Machine Learning from Disaster

Start here! Predict survival on the Titanic and get familiar with ML basics

Kaggle · 43,384 teams · Ongoing

[Overview](#)

[Data](#)

[Code](#)

[Discussion](#)

[Leaderboard](#)

[Rules](#)

[Team](#)

[My Submissions](#)

[Submit Predictions](#)

Overview

Description

Evaluation

Frequently Asked
Questions

Ahoy, welcome to Kaggle! You're in the right place.

This is the legendary Titanic ML competition – the best, first challenge for you to dive into ML competitions and familiarize yourself with how the Kaggle platform works.

The competition is simple: use machine learning to create a model that predicts which passengers survived the Titanic shipwreck.

Mini Projekt – 1.6. bis 15.6.

- Registrieren Sie sich bei [kaggle.com](https://www.kaggle.com)
- Melden Sie ein Team für ihre Gruppe an
- Probieren Sie die Titanic Challenge aus!
- Präsentieren Sie ihre Ergebnisse **am 15.6. um 12 Uhr** in der Vorlesung!

Mini Projekt – 1.6. bis 15.6.

- Registrieren Sie sich bei [kaggle.com](https://www.kaggle.com)
- Melden Sie ein Team für ihre Gruppe an
- Probieren Sie die Titanic Challenge aus!
- Präsentieren Sie ihre Ergebnisse **am 15.6. um 12 Uhr** in der Vorlesung!

Präsentation

- 1-2 Folien mit Ergebnis + Vorgehen

Mini Projekt – 1.6. bis 15.6.

- Registrieren Sie sich bei [kaggle.com](https://www.kaggle.com)
- Melden Sie ein Team für ihre Gruppe an
- Probieren Sie die Titanic Challenge aus!
- Präsentieren Sie ihre Ergebnisse **am 15.6. um 12 Uhr** in der Vorlesung!

Präsentation

- 1-2 Folien mit Ergebnis + Vorgehen

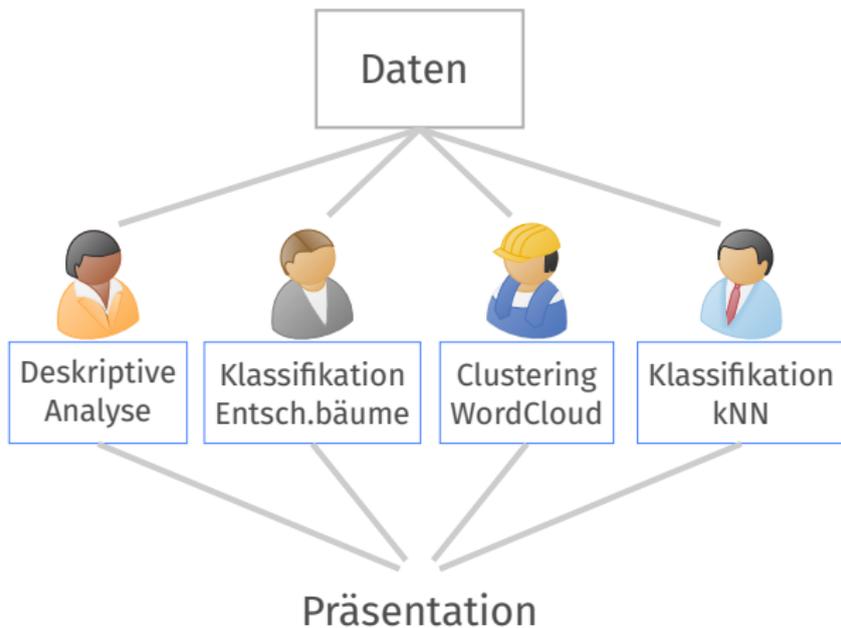
Nächste Woche:

- Keine reguläre Vorlesung
- Bei Bedarf Hilfestellung im BBB Raum von 12-16 Uhr

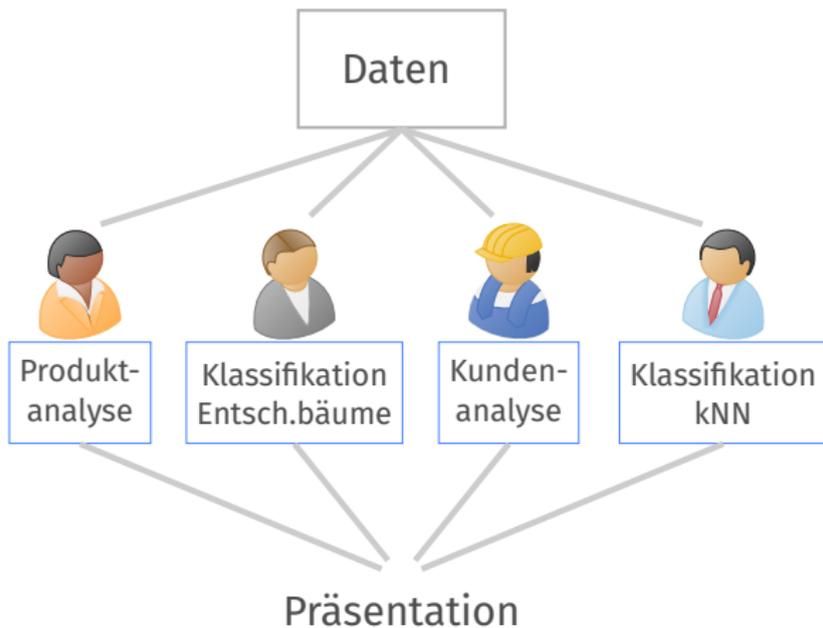
Abschlussprojekte

- Themenvergabe am 15.6. nach Präsentation der Miniprojekte
- Mögliche Themen:
 - Kaggle: Disaster-Tweets?
 - Kaggle: Indian Diabetes?
 - Kaggle: CommonLit Readability?
 - *Inside AirBnB Datensatz*
 - *Rossmann Daten*
- Am Ende 1 Präsentation je Gruppe
- Jeder Teilnehmer lädt Notebook mit Analyse (ausführlich!)
bis 20.7. 23:59 Uhr in Moodle-Kurs hoch

Abschlussprojekte



Abschlussprojekte



Abschlusspräsentation

- Terminfindung?

Abschlusspräsentation

- Terminfindung?

Bewertung

- Verschiedene Aspekte je Teilnehmer
- Schlüssige Analyse wichtig
- Ordentliches Notebook (Visualisierungen!)
- Auch erfolglose Modelle (mit Begründung!) gut
- Bewertet wird Präsentation + Hausarbeit

Abschlusspräsentation

- Terminfindung?

Bewertung

- Verschiedene Aspekte je Teilnehmer
- Schlüssige Analyse wichtig
- Ordentliches Notebook (Visualisierungen!)
- Auch erfolglose Modelle (mit Begründung!) gut
- Bewertet wird Präsentation + Hausarbeit

Details klären wir am 15.6.