

DATA SCIENCE 2

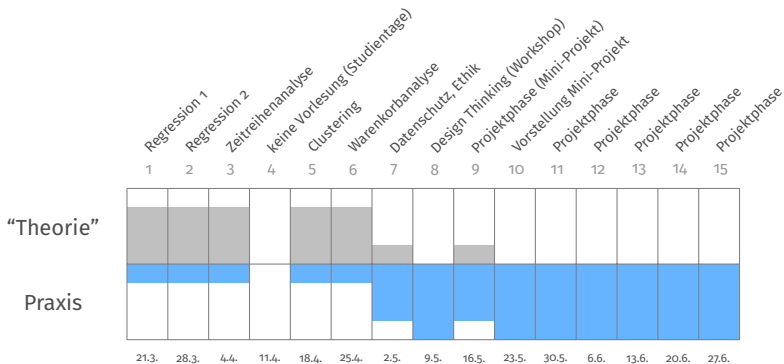
PROJEKTPHASE

PROF. DR. CHRISTIAN BOCKERMANN

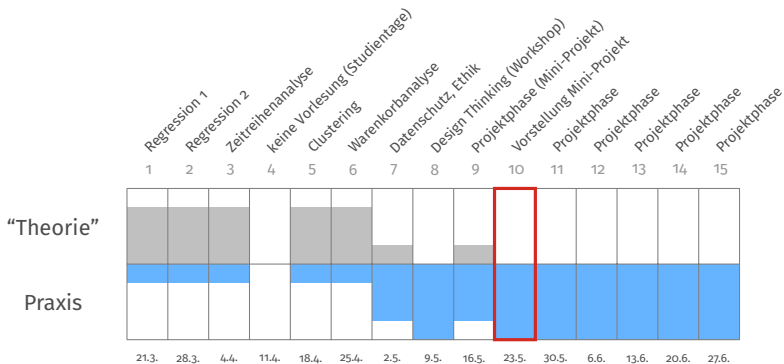
HOCHSCHULE BOCHUM

SOMMERSEMESTER 2023

Zeitplan



Zeitplan



Heute:

- Präsentation der Titanic Challenge
- Feedback-Runde - Was lief gut? Was nicht?
- Vorstellung möglicher Datensätze

Danach:

- Themenauswahl
- Gruppenfindung

Projektphase – Teil 2

Was ist das **Ziel**?

- Eigenständig Datenanalyse “erarbeiten”
- Wirtschaftliche Fragestellung überlegen
- Wissenschaftliche Fragestellung ableiten, Lernaufgabe(n) formulieren/anpassen
- Datensatz explorieren
- Daten analysieren und eigene Analyse bewerten

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

Modus A

- Gemeinsame Kaggle Challenge
- Jede Gruppe nimmt als Team teil

Modus A

- Gemeinsame Kaggle Challenge
- Jede Gruppe nimmt als Team teil
- Problem: **Individuelle Prüfungsleistung?**

Modus A

- Gemeinsame Kaggle Challenge
- Jede Gruppe nimmt als Team teil
- Problem: **Individuelle Prüfungsleistung?**

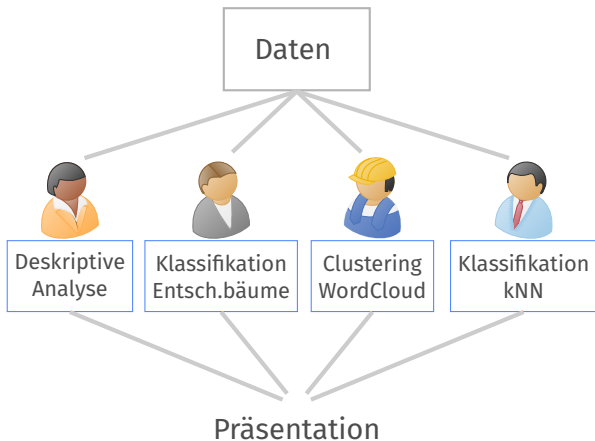
Modus B

- Freie Wahl des Datensatzes
- Je Gruppe unterschiedliche Fragen/Aspekte
- Jeder Teilnehmer bearbeitet u.a. eigenständigen Aspekt
- Gemeinsame Präsentation der Analyse des Datensatzes

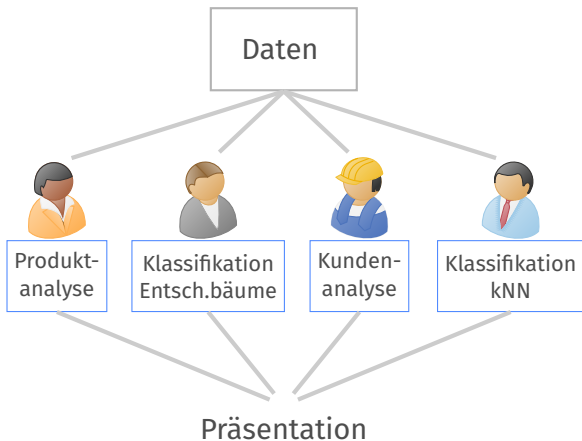
Abschlussprojekte

- Daten-/Themenvergabe heute
- Mögliche Themen/Datensätze stelle ich gleich vor
- Am Ende 1 Präsentation je Gruppe
- Jeder Teilnehmer lädt Notebook mit Analyse (ausführlich!)
bis TBD 23:59 Uhr in Moodle-Kurs hoch
- Präsentationen am **TBD**

Abschlussprojekte



Abschlussprojekte



Abschlusspräsentation

- Termin: **TBD**

Abschlusspräsentation

- Termin: **TBD**

Bewertung

- Verschiedene Aspekte je Teilnehmer
- Schlüssige Analyse wichtig
- Ordentliches Notebook (Visualisierungen!)
- Auch erfolglose Modelle (mit Begründung!) gut
- Bewertet wird Präsentation + Hausarbeit
- “Leitfaden” für Projektphase 2:

[https:](https://datascience.hs-bochum.de/vorlesungen/ss2023/datascience2)

[//datascience.hs-bochum.de/vorlesungen/ss2023/datascience2](https://datascience.hs-bochum.de/vorlesungen/ss2023/datascience2)

Mögliche Datensätze

- California House Pricing, House Prices Advances

<https://www.kaggle.com/c/california-house-prices/overview>

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

- Predict Future Sales for Store/Product

<https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview>

<https://www.kaggle.com/c/instacart-market-basket-analysis/data>

- Natural Language Processing with Disaster Tweets

<https://www.kaggle.com/c/nlp-getting-started>

- Inside AirBnB

<http://insideairbnb.com/get-the-data.html>

- RKI Covid19

[https:](https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0)

[//npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0](https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0)

Mögliche Datensätze

- Sprint-Preise bei Tankerkönig

<https://tankerkoenig.de>

- Immobilienpreise von Immoscout 24

<https://www.rwi-essen.de/forschung-beratung/weitere/forschungsdatenzentrum-ruhr/datenangebot/rwi-geo-red-real-estate-data>

Vorhersage von Immobilienpreisen

Getting Started Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 11,285 teams · Ongoing

Overview Data Code Discussion Leaderboard Rules [Join Competition](#)

Overview

Description

Evaluation

Tutorials

Frequently Asked Questions

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

Competition Description

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Vorhersage von Verkäufen

Featured Prediction Competition

Rossmann Store Sales

Forecast sales using store, promotion, and competitor data

3,298 teams · 6 years ago

\$35,000
Prize Money

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Team](#) [My Submissions](#) [Late Submission](#)

Overview

Description

Evaluation

Prizes

Timeline

Rossmann operates over 3,000 drug stores in 7 European countries. Currently, Rossmann store managers are tasked with predicting their daily sales for up to six weeks in advance. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. With thousands of individual managers predicting sales based on their unique circumstances, the accuracy of results can be quite varied.

<https://www.kaggle.com/c/rossmann-store-sales/overview>

Meint der Tweet eine Katastrophe? Ja/Nein

Getting Started Prediction Competition

Natural Language Processing with Disaster Tweets

Predict which Tweets are about real disasters and which ones are not

Kaggle · 3,114 teams · Ongoing

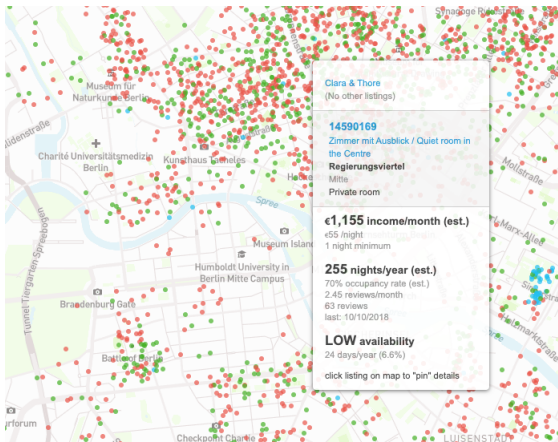
Overview Data Code Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

Overview

Description	Welcome to one of our "Getting Started" competitions 🙌
Evaluation	This particular challenge is perfect for data scientists looking to get started with Natural Language Processing. The competition dataset is not too big, and even if you don't have much personal computing power, you can do all of the work in our free, no-setup, Jupyter Notebooks environment called Kaggle Notebooks .
FAQ	Competition Description Twitter has become an important communication channel in times of emergency.

<https://www.kaggle.com/c/nlp-getting-started>

Analyse von AirBnB-Daten (z.B. Berlin)



<http://insideairbnb.com/get-the-data.html>

Mögliche Fragestellungen

- Welche Gegenden von z.B. Berlin sind teuer/günstig?
- Welche Eigenschaften von Wohnungen führen zu hohen Mietpreisen?
- Wie gut läßt sich der *est. income* vorhersagen?
- Wo befinden Sie die Hotspots *professioneller* Vermieter?
(Hosts mit mehr als 2 oder 3 Angeboten)

Mögliche Fragestellungen

- Welche Gegenden von z.B. Berlin sind teuer/günstig?
- Welche Eigenschaften von Wohnungen führen zu hohen Mietpreisen?
- Wie gut läßt sich der *est. income* vorhersagen?
- Wo befinden Sie die Hotspots *professioneller* Vermieter? (Hosts mit mehr als 2 oder 3 Angeboten)

Advanced

- Wie unterscheiden sich Preise/Angebote von Stadt A und Stadt B?

Analyse der Spritpreise

- Veröffentlichungspflicht der Tankstellen
- Spritpreise seit 2014



Spritpreise für alle!

Analyse der Spritpreise

- Steigen/sinken die Preise in allen Regionen gleich?
- Wie verläuft der Spritpreis zum Ölpreis?
- Gibt es Marken, die stärker an den Ölpreis gebunden sind?
- Hat sich das Verhalten seit Kriegsbeginn (Ukraine) geändert?

Analyse von Daten von ImmoScout24

- Leibniz-Institut für Wirtschaftsforschung
- Forschungsdatenzentrum Ruhr (FDZ)

<https://www.rwi-essen.de/forschung-beratung/weitere/forschungsdatenzentrum-ruhr/datenangebot/rwi-geo-red-real-estate-data>

Fussball Bundesliga

- <https://datahub.io/sports-data/german-bundesliga>
- <http://www.bulibox.de/statistik/1-Bundesliga.html>
- <https://www.keinemathematik.de/datenquellen/>

Wie geht's weiter?

Zeitplan Projektphase – 2023

30.5.2023, 9 Uhr – Visualisierung

- Status: Bericht aus den Gruppen
- Vortrag: Visualisierung

6.6.2023, 9 Uhr – Big Data

- Status: Bericht aus den Gruppen
- Vortrag: Big Data

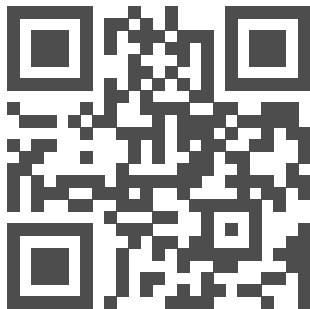
13.6.2023, 9 Uhr – No-Code Datenanalyse

- Status: Bericht aus den Gruppen
- Vorstellung: No-Code Datenanalyse

20.6.2023, 9 Uhr – Deep Learning/ChatGPT?

- Status: Bericht aus den Gruppen

Lehre-Evaluation (Data Science 2, Sommersemester 2023)



<https://hsbo.de/ds2ev>