

DATA SCIENCE 2

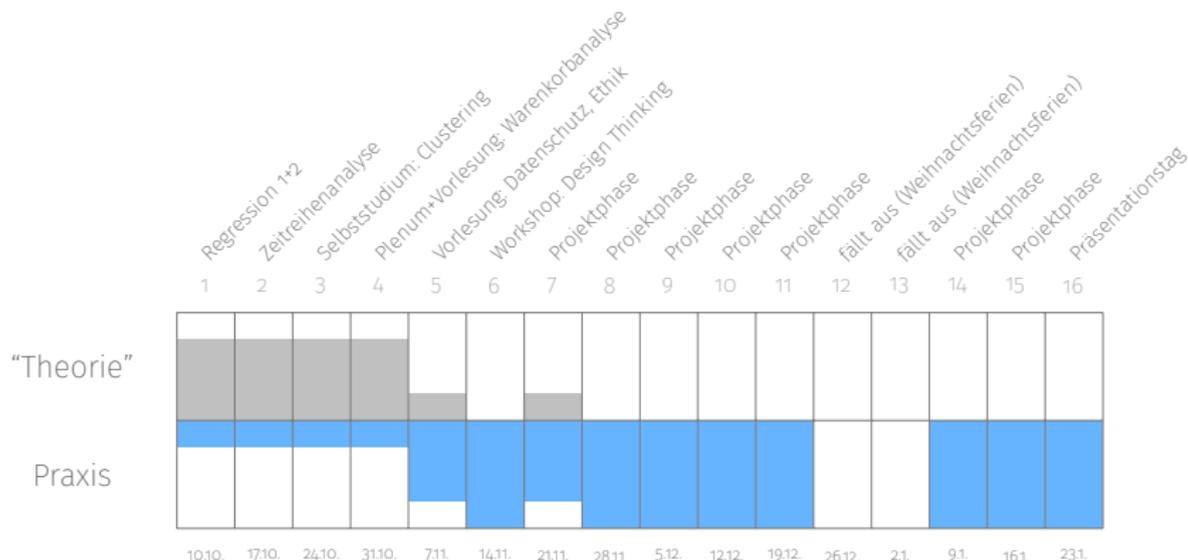
PROJEKTPHASE

PROF. DR. CHRISTIAN BOCKERMANN

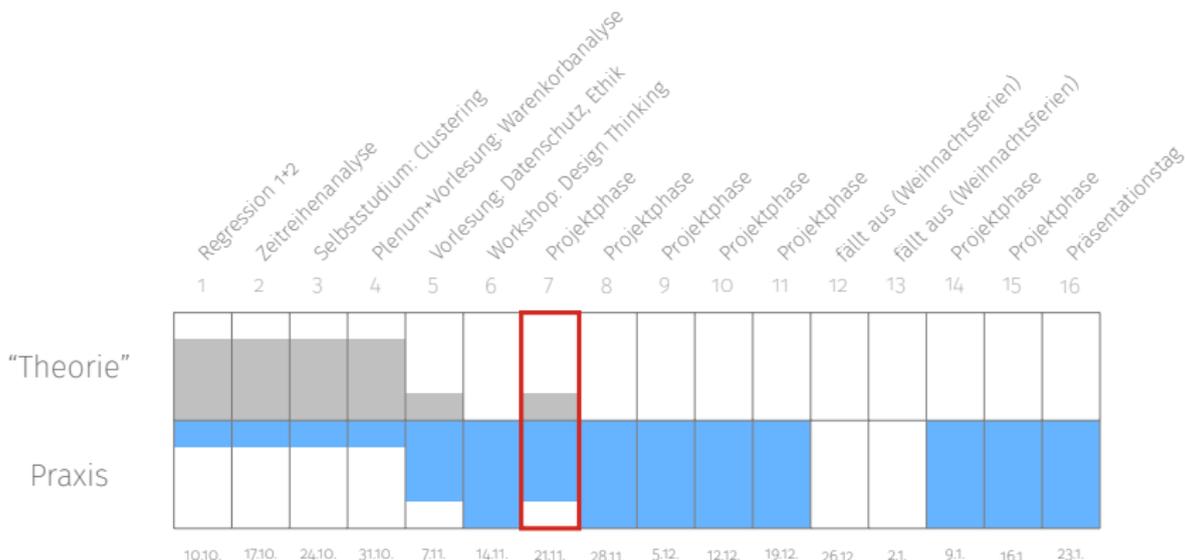
HOCHSCHULE BOCHUM

WINTERSEMESTER 2024/2025

Zeitplan



Zeitplan



Heute:

- Diskussion Hausarbeit (Vorlage, Kriterien,...)
- Zeitplanung? Abgabe, Präsentationstag
- Vorstellung möglicher Datensätze

Danach:

- Themenauswahl
- Gruppenfindung

Einstieg in ein Forschungsgebiet

Problem: Automatisierte Bauanträge mit KI

- Was brauchen wir? Wo beginnt man?

Problem: Automatisierte Bauanträge mit KI

- Was brauchen wir? Wo beginnt man?
- Recherche, z.B.

<https://scholar.google.de>



Artikel

Ungefähr 339.000 Ergebnisse (0,05 Sek.)

Beliebige Zeit

Seit 2024

Seit 2023

Seit 2020

Zeitraum wählen...

Combining NLP approaches for rule extraction from legal documents

[M Dragoni](#), [S Villata](#), [W Rizzi](#)... - ... and REasoning with Legal ..., 2016 - hal.science

... an automated norm/**rules extraction** system will help ... **rules**. However, the adopted methodology is different: they exploit Juridical (Natural) Language Constructs (JLC) that formalize **legal** ...

☆ Speichern Zitieren Zitiert von: 82 Ähnliche Artikel Alle 7 Versionen

Nach Relevanz
sortieren

Nach Datum
sortieren

Combining natural language processing approaches for rule extraction from legal documents

[M Dragoni](#), [S Villata](#), [W Rizzi](#), [G Governatori](#) - ... to the Complexity of Legal ..., 2018 - Springer

... an automated norm/**rules extraction** system will help ... **rules**. However, the adopted methodology is different: they exploit Juridical (Natural) Language Constructs (JLC) that formalize **legal** ...

☆ Speichern Zitieren Zitiert von: 27 Ähnliche Artikel Alle 3 Versionen

Beliebige Sprache

Seiten auf Deutsch

[PDF] An overview of information extraction techniques for legal document analysis and processing.

[AV Zadgaonkar](#), [AJ Agrawal](#) - International Journal of Electrical & ..., 2021 - academia.edu

... **extraction** from **legal documents** is highly desired. Information **extraction** from **legal documents** will be ... The **extracted** information can be: i) stored in databases for future references. ii) for

Alle Typen

Übersichtsarbeiten

Wie sieht eine wissenschaftliche Arbeit aus?

1. Einleitung / Motivation + Fragestellung
2. Stand der Forschung (*related work*)
3. Theoretische Grundlagen
4. Eigene Entwicklung
5. Experimente
6. Schlussfolgerung

Vorlagen, Review-Prozess

- Vorlagen von Verlag/Konferenz/etc
- Häufig *blind-review*, Begutachtung durch Experten
- Annahme/Ablehnung zu Konferenz/Journal/..

Was macht eine gute Arbeit aus?

- Darstellung des Themas (Innovation, Relevanz)
- klare Fragestellung + Untersuchungsmethode
- verständlich geschrieben
- nachvollziehbare Schlussfolgerungen

Links:

- <https://www1.aucegypt.edu/faculty/kseddik/Homepage/doing-research.html>
- https://www.cs.ucr.edu/~eamonn/Keogh_SIGKDD09_tutorial.pdf

Projektphase

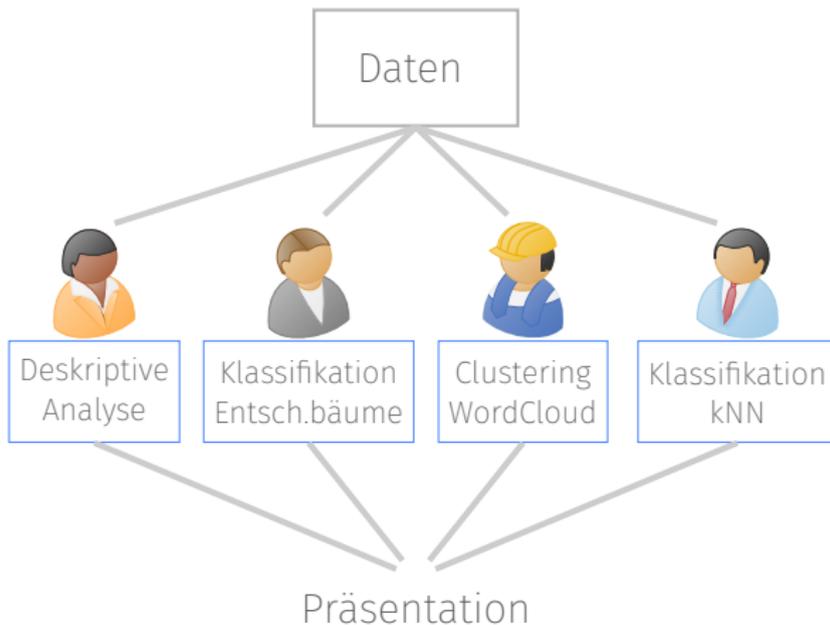
Was ist das **Ziel**?

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

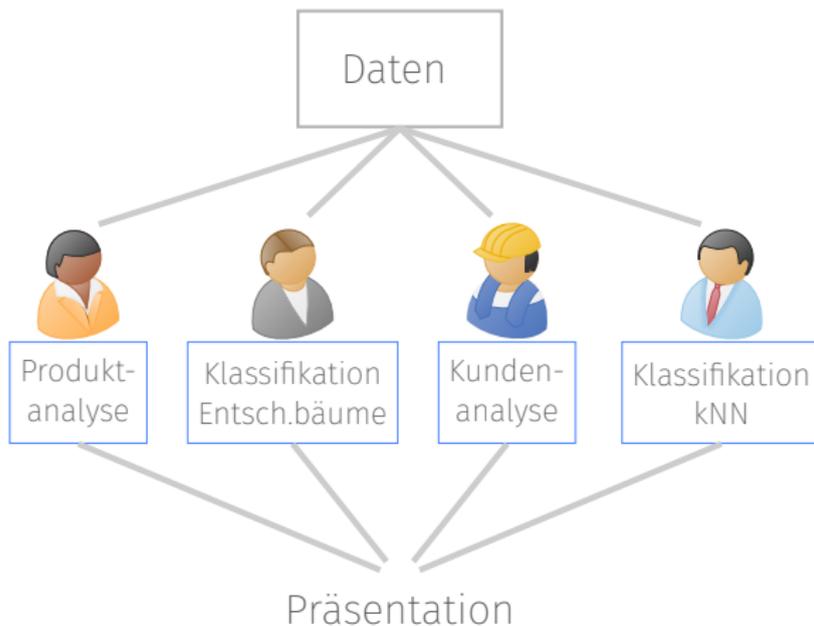
Organisation

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

Abschlussprojekte



Abschlussprojekte



Abschlusspräsentation

- Termin: **23.1.2025, 9 Uhr**

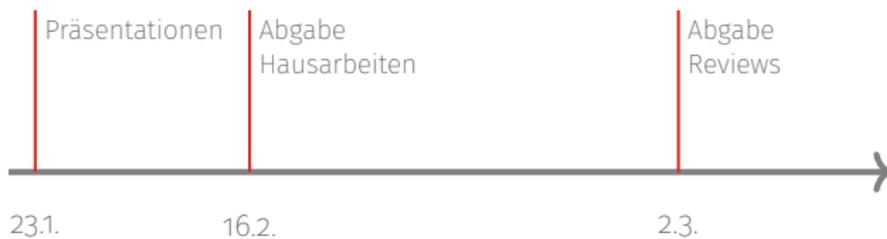
Abschlusspräsentation

- Termin: **23.1.2025, 9 Uhr**

Bewertung

- Schlüssige Analyse wichtig
- Auch erfolgreiche Modelle (mit Begründung!) gut
- Vorlage LNCS (LaTeX/Word-Vorlagen verfügbar)
- 6-8 Seiten Hausarbeit
- Review von 2 fremden Hausarbeiten
- Bewertet wird Hausarbeit + Reviews

Zeitplan



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://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/dd4580c810204019a7b8eb3e0b329dd6_0

Mögliche Datensätze

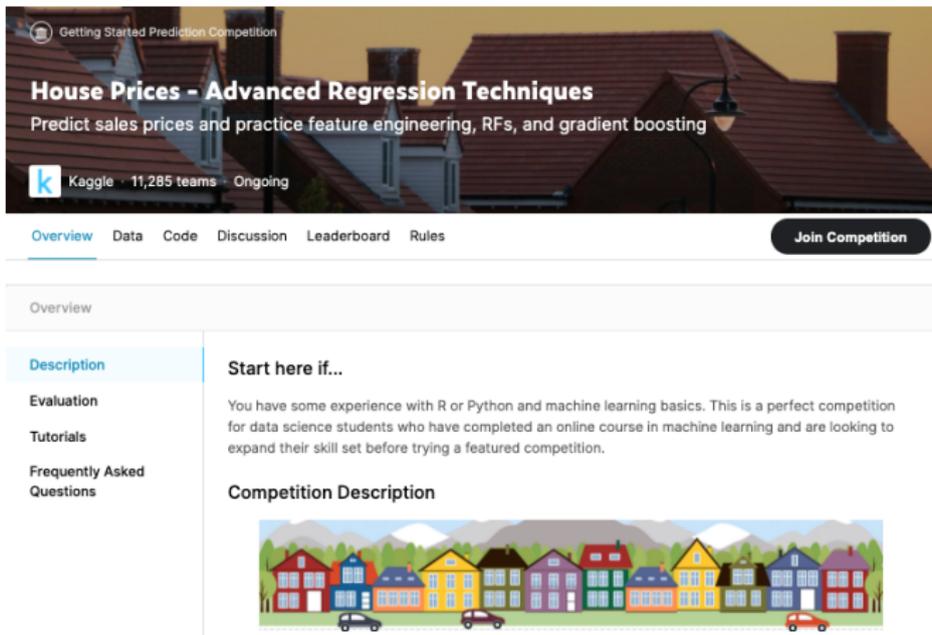
- Sprint-Preise bei Tankerkönig

<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

Evaluation

FAQ

Welcome to one of our "Getting Started" competitions 🙌

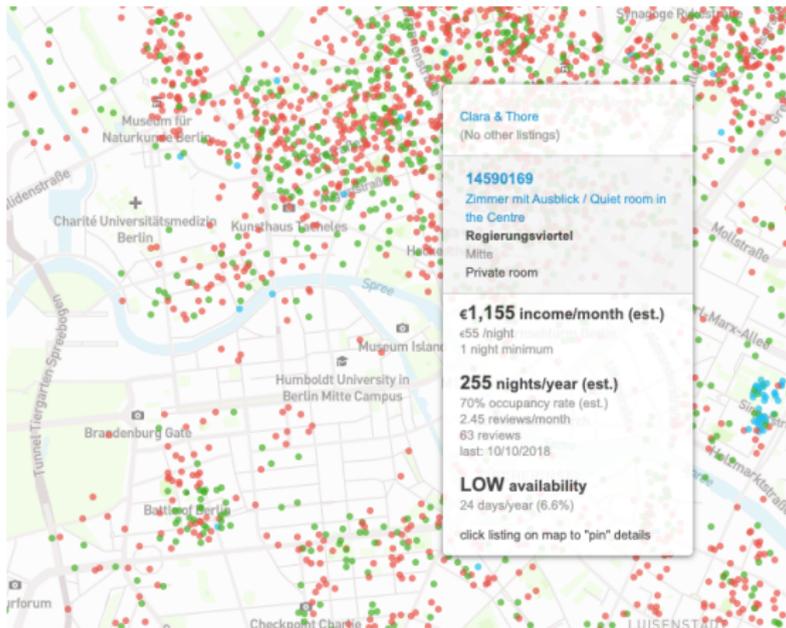
This particular challenge is perfect for data scientists looking to get started with Natural Language Processing. The competition dataset is not too big, and even if you don't have much personal computing power, you can do all of the work in our free, no-setup, Jupyter Notebooks environment called [Kaggle Notebooks](#).

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?

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 – SoSe 2024

28.11.2024, 9 Uhr – Visualisierung

- Status: Bericht aus den Gruppen
- Vortrag: Visualisierung

5.12.2024, 9 Uhr – Big Data

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

12.12.2024, 9 Uhr – Big Data

- Status: Bericht aus den Gruppen
- Vortrag: Deep-Learning und ChatGPT

19.12.2024, 9 Uhr – No-Code Datenanalyse

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