Data Science 2

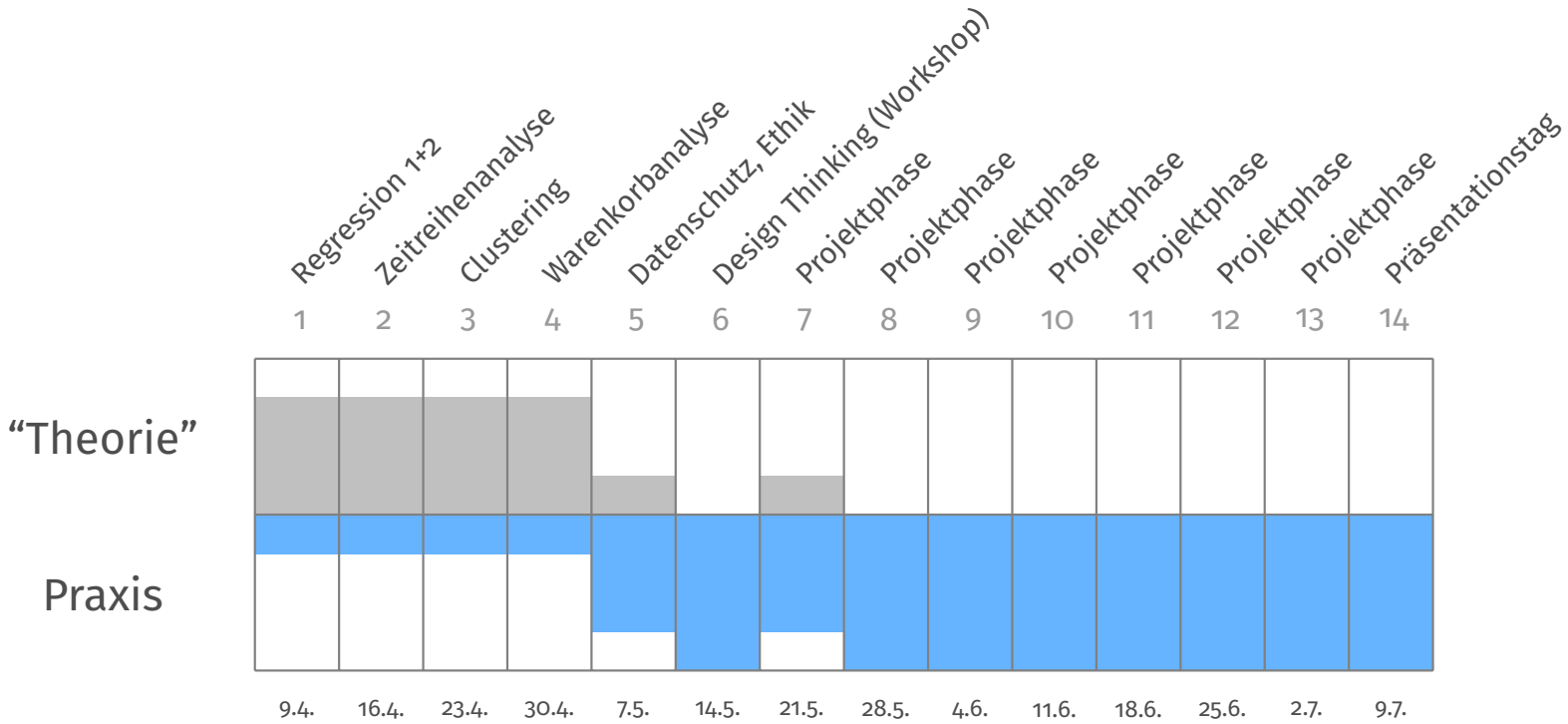
Projektphase

Prof. Dr. Christian Bockermann

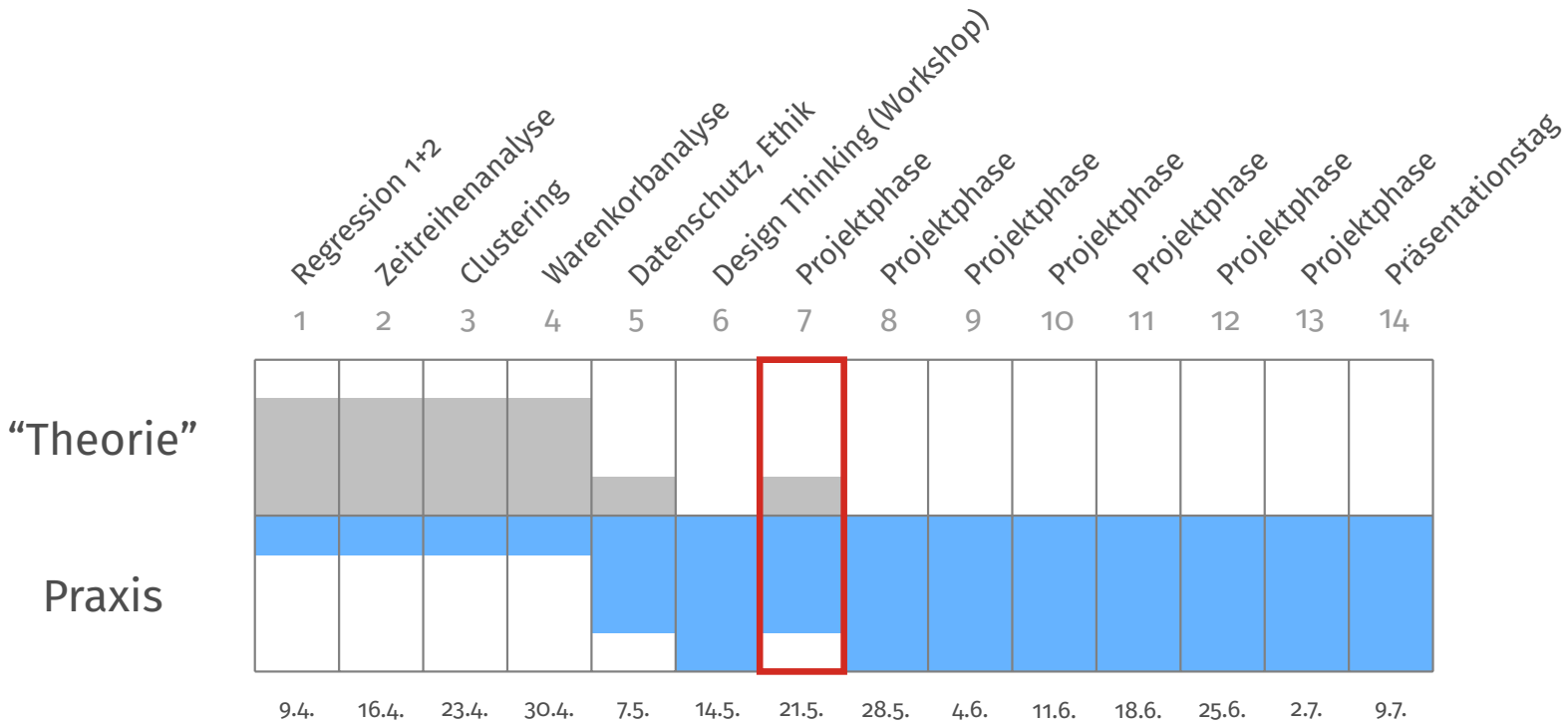
Hochschule Bochum

Sommersemester 2024

Zeitplan


Regression 1+2
Zeitreihenanalyse
Clustering
Warenkorbanalyse
Datenschutz, Ethik
Design Thinking (Workshop)
Projektphase
Projektphase
Projektphase
Projektphase
Projektphase
Projektphase
Projektphase
Präsentationstag
1
2
3
4
5
6
7
8
9
10
11
12
13
14
“Theorie”
Praxis
9.4.
16.4.
23.4.
30.4.
7.5.
14.5.
21.5.
28.5.
4.6.
11.6.
18.6.
25.6.
2.7.
9.7.

Zeitplan


Regression 1+2
Zeitreihenanalyse
Clustering
Warenkorbanalyse
Datenschutz, Ethik
Design Thinking (Workshop)
Projektphase
Projektphase
Projektphase
Projektphase
Projektphase
Projektphase
Projektphase
Präsentationstag
1
2
3
4
5
6
7
8
9
10
11
12
13
14
“Theorie”
Praxis
9.4.
16.4.
23.4.
30.4.
7.5.
14.5.
21.5.
28.5.
4.6.
11.6.
18.6.
25.6.
2.7.
9.7.

Agenda

Heute:

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

Danach:

- Themenauswahl
- Gruppenfindung

Einstieg in ein Forschungsgebiet

Problem: Automatisierte Bauanträge mit KI

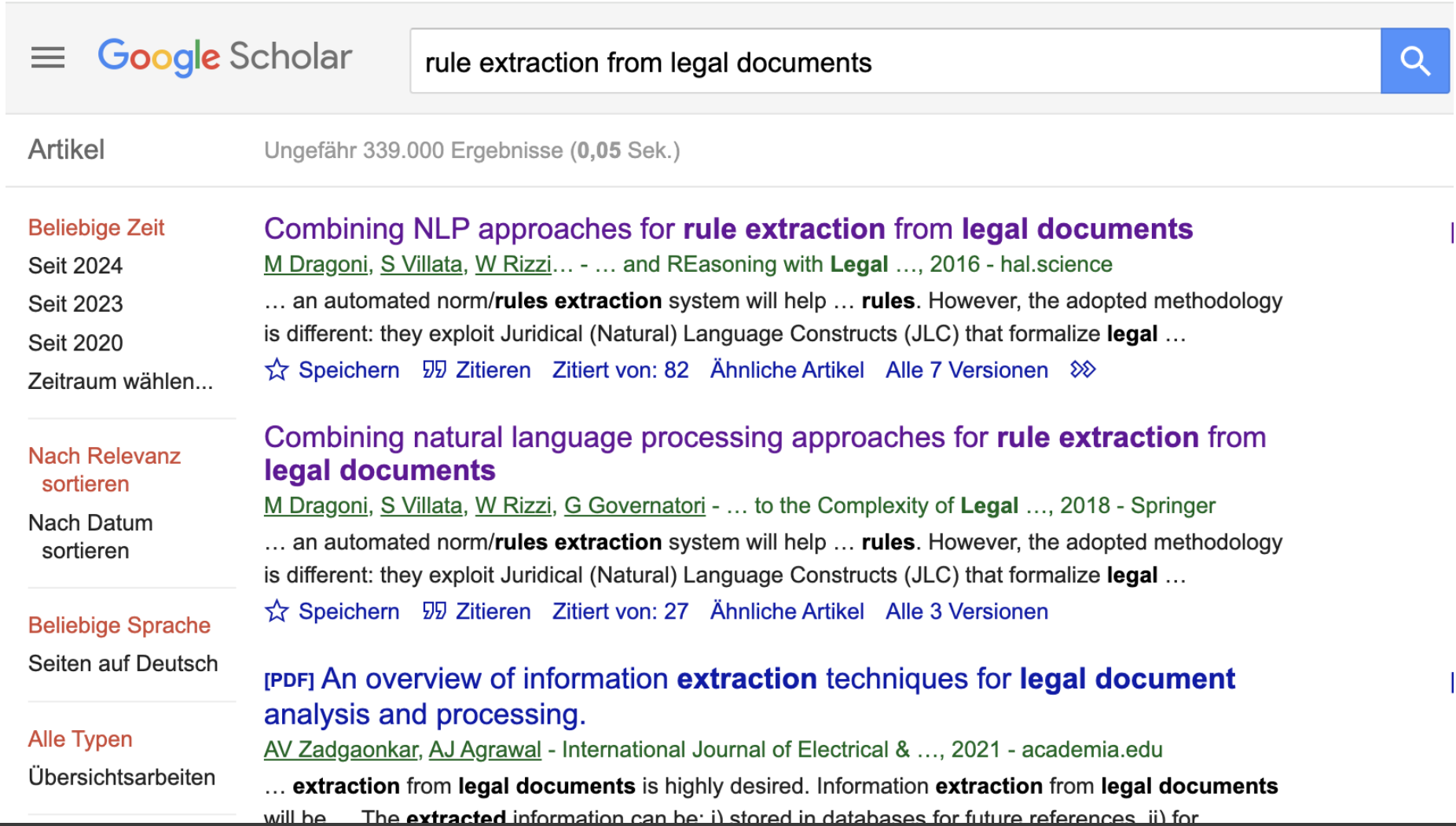
- Was brauchen wir? Wo beginnt man?

Problem: Automatisierte Bauanträge mit KI

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

`https://scholar.google.de`

Google Scholar



Was ist eine gute Arbeit?

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

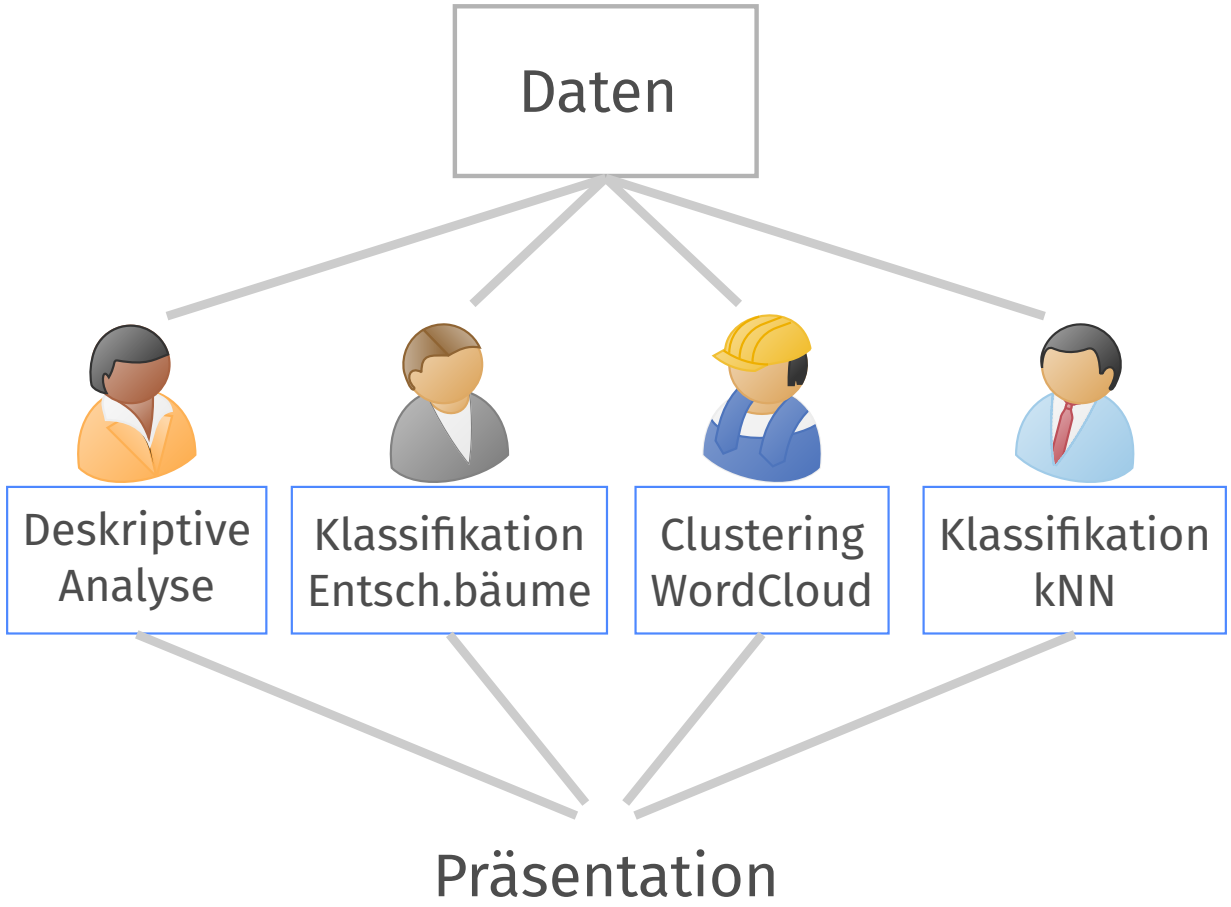
Was ist das Ziel?

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

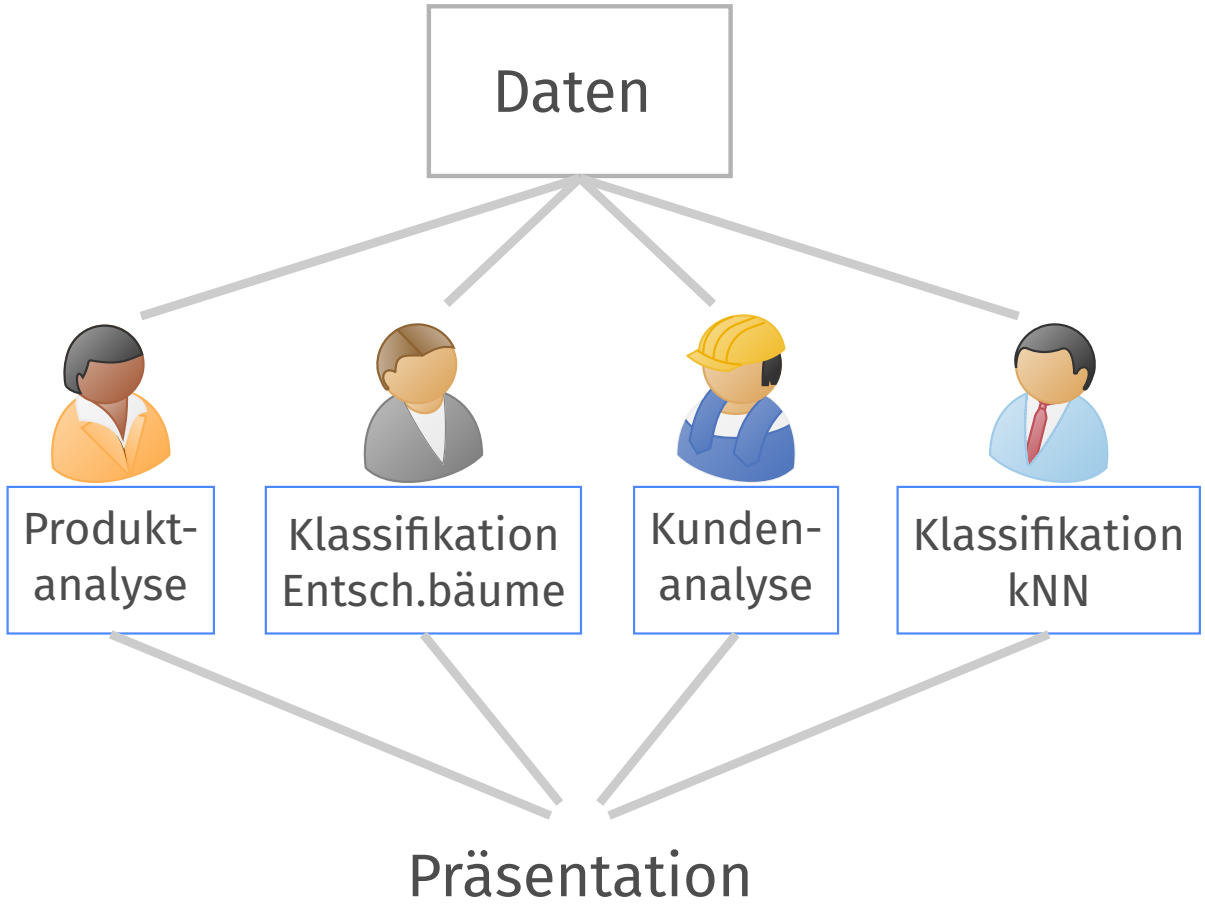
Organisation

- 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: **9.7.2024, 9 Uhr**

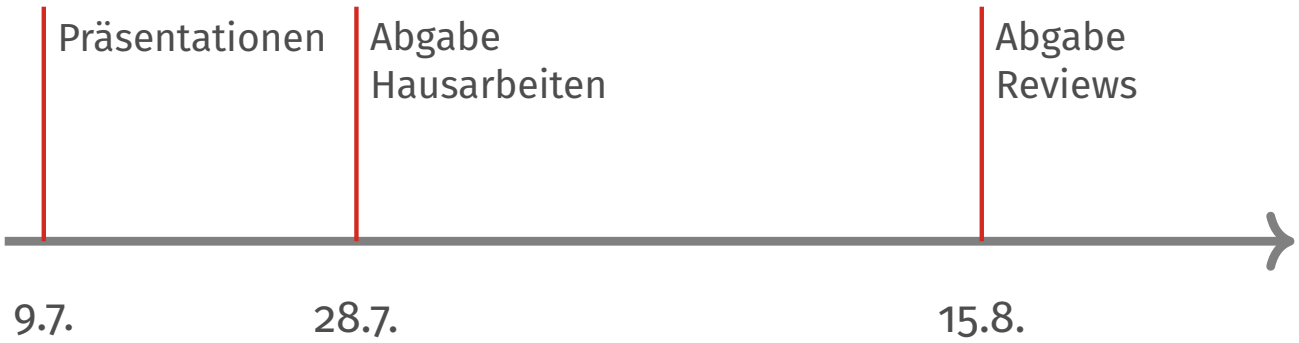
Abschlusspräsentation

- Termin: **9.7.2024, 9 Uhr**

Bewertung

- Verschiedene Aspekte je Teilnehmer
- Schlüssige Analyse wichtig
- Auch erfolglose Modelle (mit Begründung!) gut
- Vorlage LNCS (LaTex/Word-Vorlagen verfügbar)
- 6-8 Seiten Hausarbeit
- Review von 2 fremden Hausarbeiten
- Bewertet wird Präsentation, Hausarbeit + Reviews

Zeitplan


Präsentationen
Abgabe
Hausarbeiten
Abgabe
Reviews
9.7.
28.7.
15.8.

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

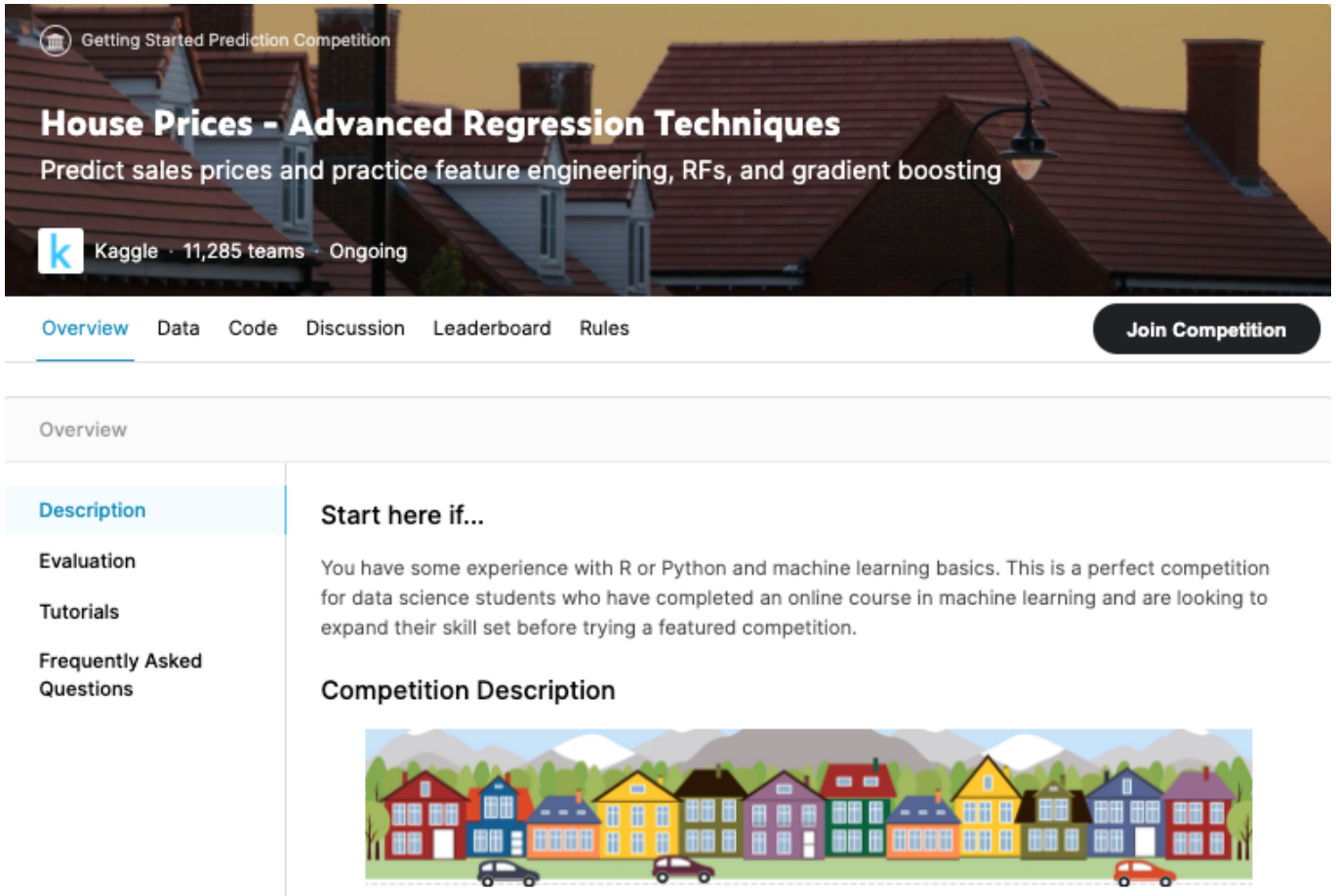
- Sprint-Preise bei Tankerkönig

  `https://tankerkoenig.de`

- Immobilienpreise von Immoscout 24

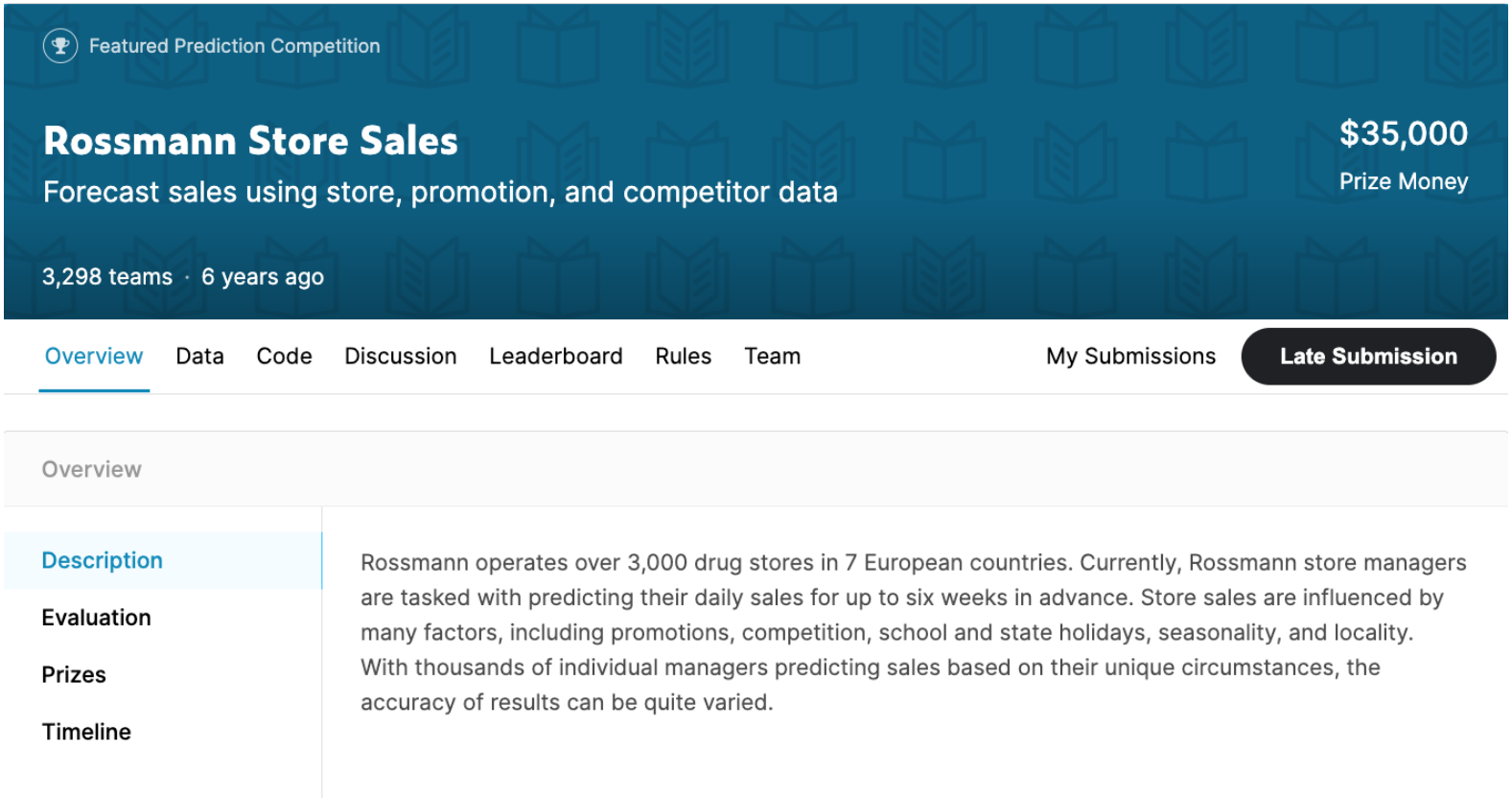
  `https://www.rwi-essen.de/forschung-beratung/weitere/forschungsdatenzentrum-ruhr/datenangebot/rwi-geo-red-real-estate-data`

Vorhersage von Immobilienpreisen



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

Vorhersage von Verkäufen



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

Meint der Tweet eine Katastrophe? Ja/Nein

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

Overview

Description
Evaluation
FAQ

Welcome to one of our "Getting Started" competitions 👋

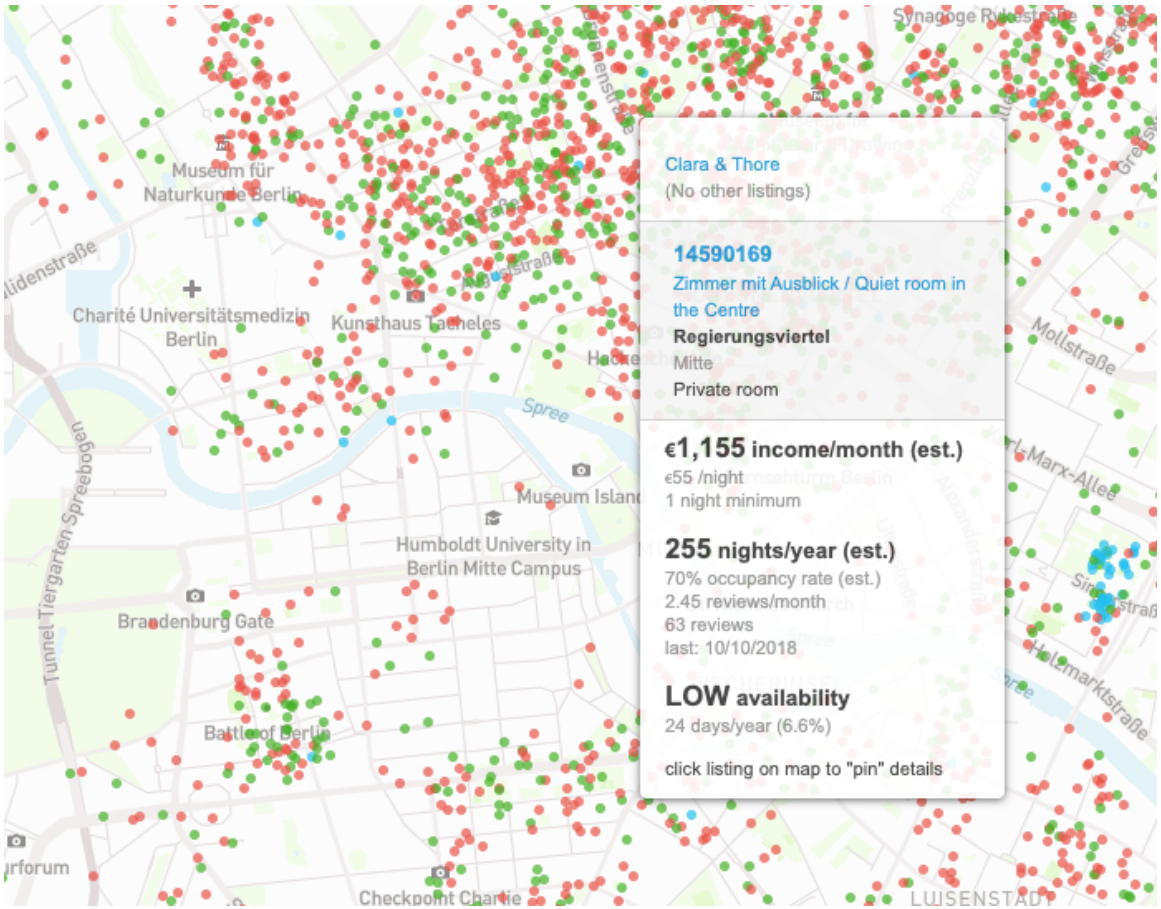
This particular challenge is perfect for data scientists looking to get started with Natural Language Processing. The competition dataset is not too big, and even if you don't have much personal computing power, you can do all of the work in our free, no-setup, Jupyter Notebooks environment called Kaggle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

28.5.2024, 9 Uhr

- Freie Gruppenarbeit

4.6.2024, 9 Uhr – Visualisierung

- Status: Bericht aus den Gruppen
- Vortrag: Visualisierung

11.6.2024, 9 Uhr – Big Data

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

18.6.2024, 9 Uhr – No-Code Datenanalyse

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

25.6.2024, 9 Uhr – Deep Learning/ChatGPT?

- Status: Bericht aus den Gruppen