Data Science 2

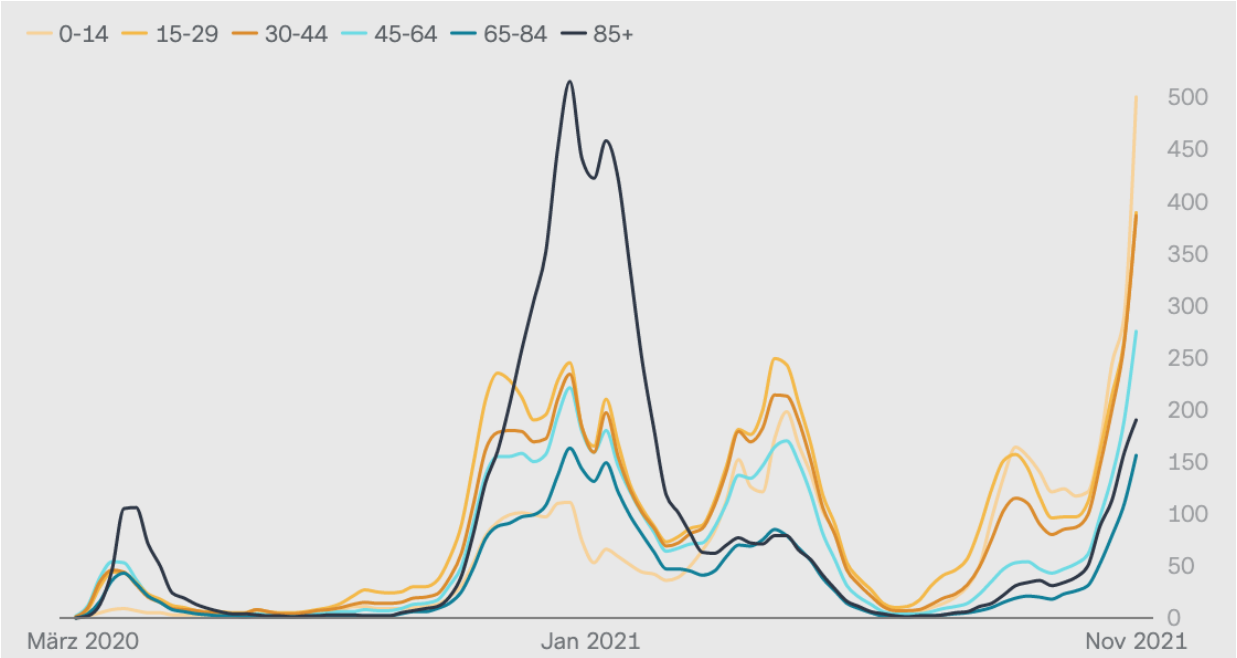
Projektphase

Prof. Dr. Christian Bockermann

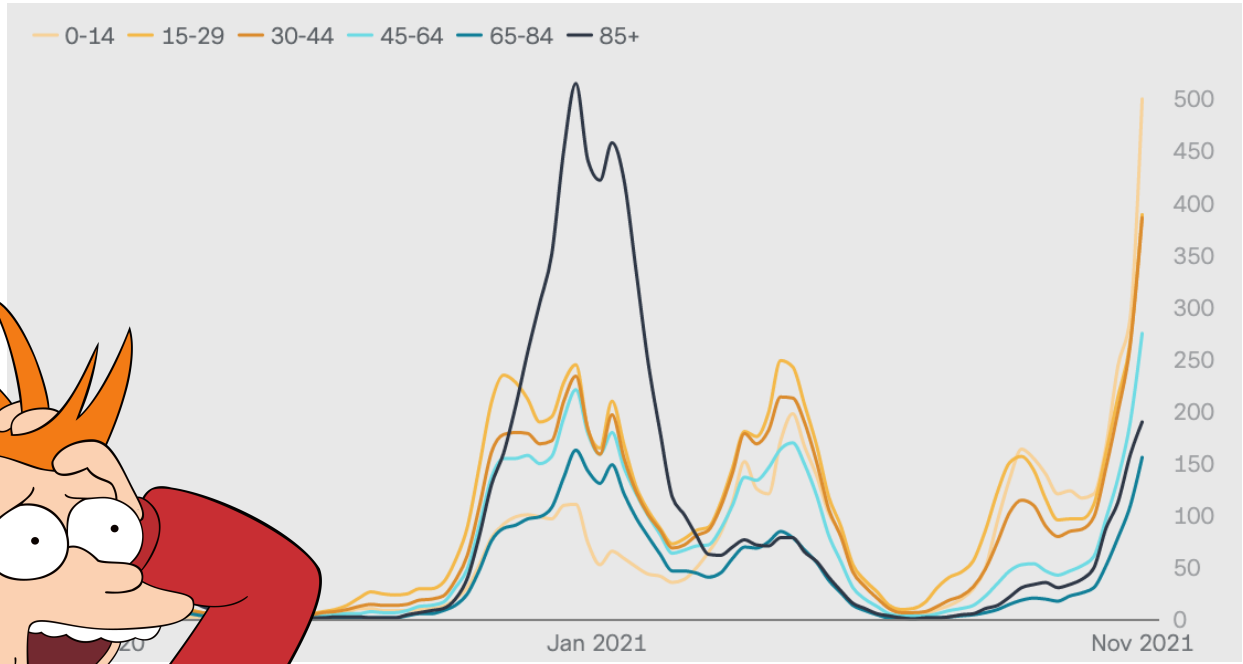
Hochschule Bochum

Wintersemester 2021/2022

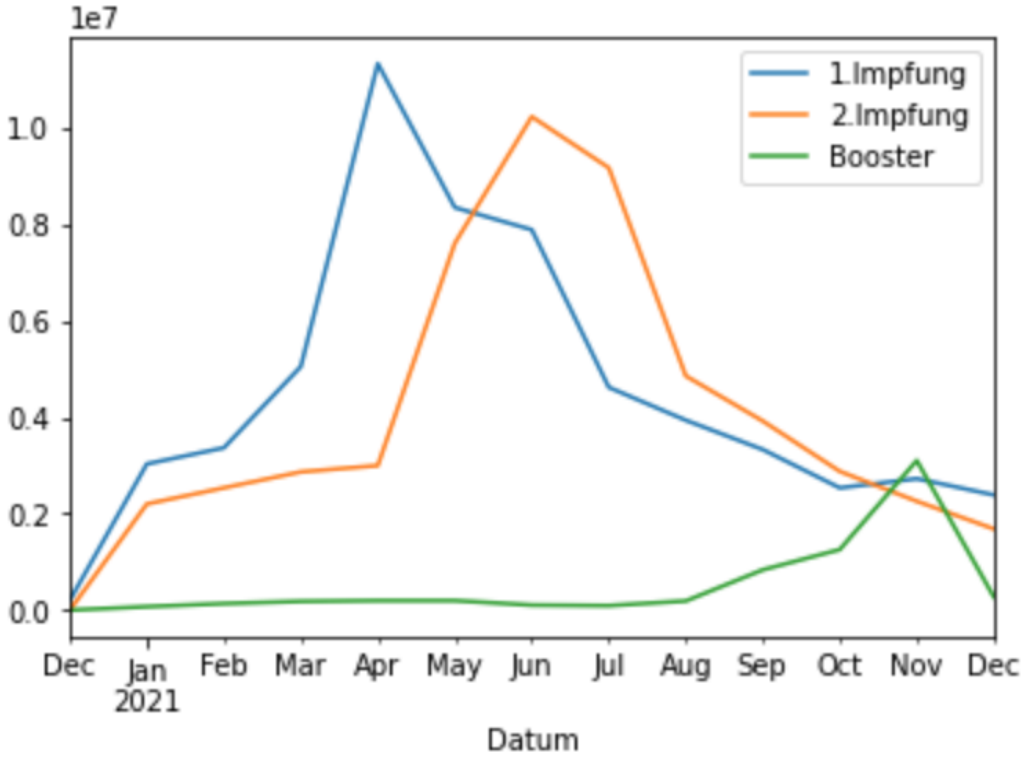
Corona – Inzidenzen nach Altersgruppe



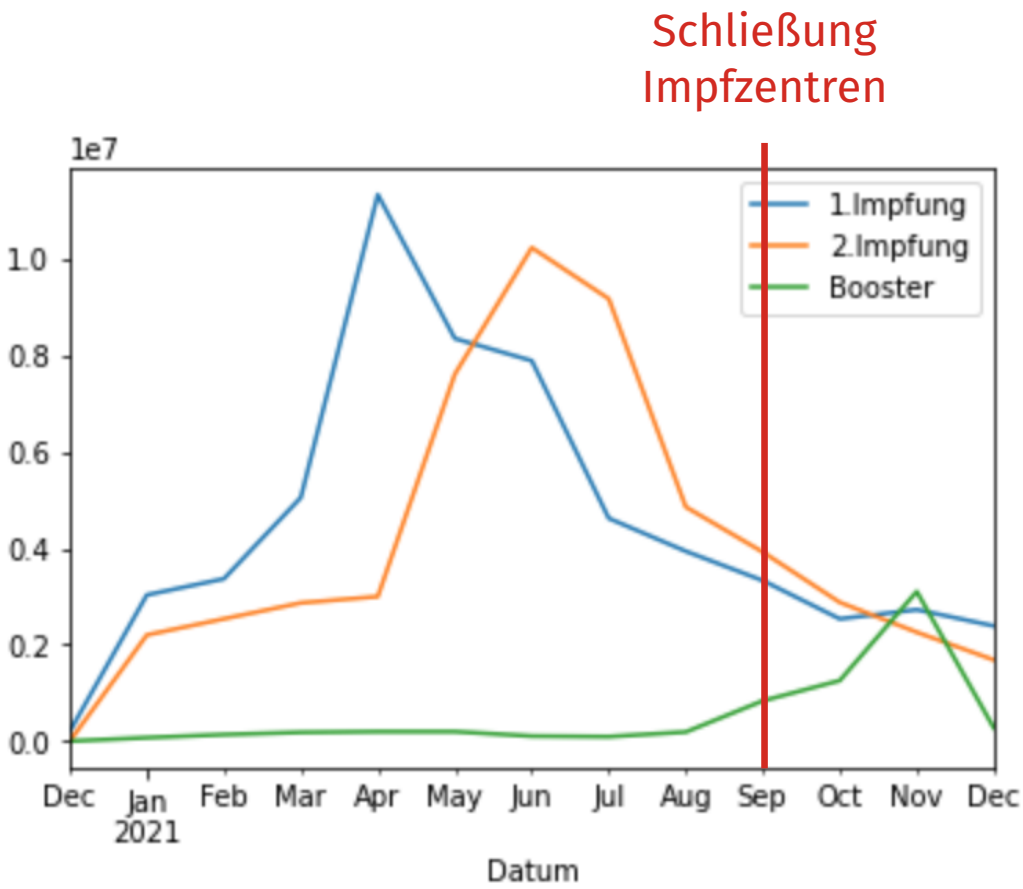
Corona – Inzidenzen nach Altersgruppe



Aktuelle Lage: Wie konnte es so weit kommen?



Aktuelle Lage: Wie konnte es so weit kommen?



War das ein gut gewählter Zeitpunkt?

Corona Situation

Informationsquellen

- RKI Wochenbericht (jeden Donnerstag)
- RKI Infektionsdaten (u.a. als CSV-Datei)
- Impfquotenmonitoring

Informationsquellen

- RKI Wochenbericht (jeden Donnerstag)
- RKI Infektionsdaten (u.a. als CSV-Datei)
- Impfquotenmonitoring

Beispiel Impfquotenmonitoring:

`https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Daten/Impfquotenmonitoring.html`

RKI Empfehlungen

- Impfen! (Effekt erst in 2+ Wochen)
- Hygiene-Maßnahmen (Effekt sofort!)
- **Kontaktbeschränkungen** (Effekt sofort!)

RKI Empfehlungen

- Impfen! (Effekt erst in 2+ Wochen)
- Hygiene-Maßnahmen (Effekt sofort!)
- **Kontaktbeschränkungen** (Effekt sofort!)
- Was ist mit Schnelltests?

Corona Situation

RKI Empfehlungen

- Impfen! (Effekt erst in 2+ Wochen)
- Hygiene-Maßnahmen (Effekt sofort!)
- **Kontaktbeschränkungen** (Effekt sofort!)
- Was ist mit Schnelltests?

Hochschule Bochum

- Strikte Maskenpflicht in Innenräumen
- Einhalten von Mindestabständen (1,5 m)

RKI Empfehlungen

- Impfen! (Effekt erst in 2+ Wochen)
- Hygiene-Maßnahmen (Effekt sofort!)
- **Kontaktbeschränkungen** (Effekt sofort!)
- Was ist mit Schnelltests?

Hochschule Bochum

- Strikte Maskenpflicht in Innenräumen
- Einhalten von Mindestabständen (1,5 m)

Fachbereich Wirtschaft

- Sämtliche Veranstaltungen im Bachelor-Bereich ab Montag, 29.11. wieder online

DataScience 2 Kurs

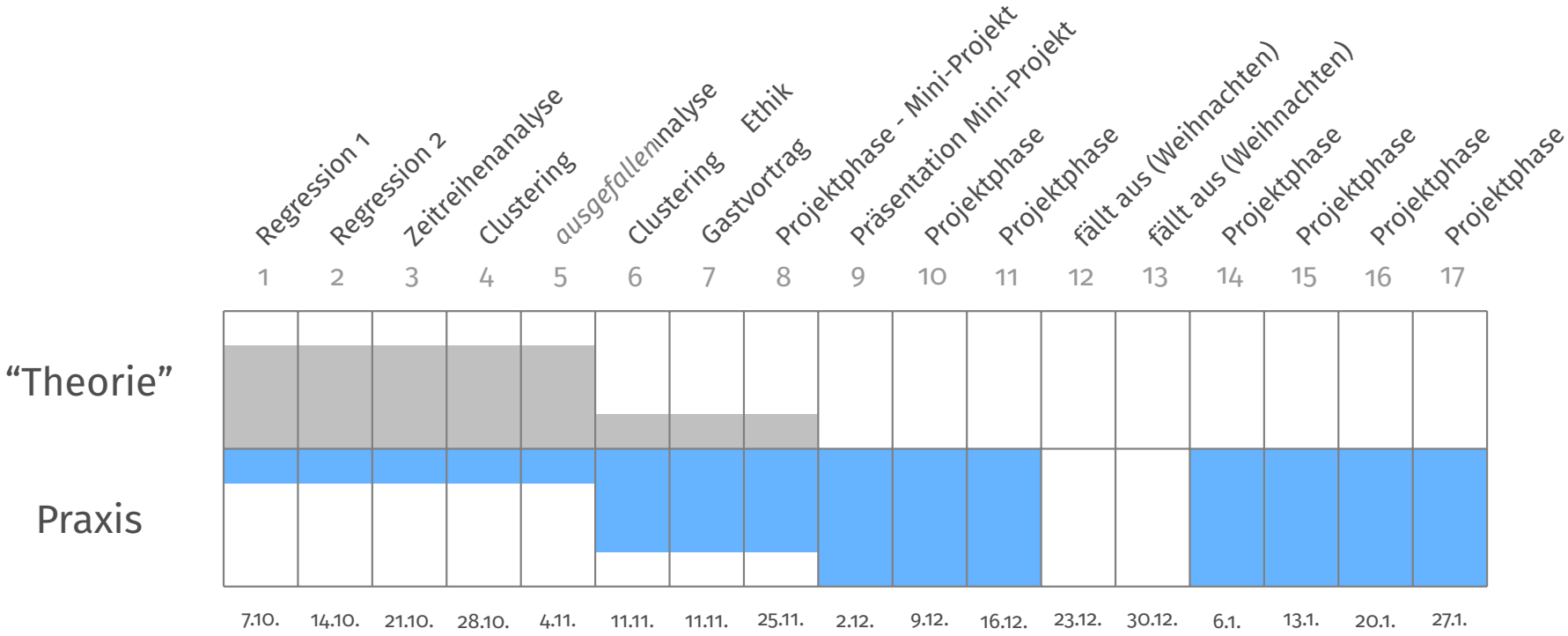
- Als nächstes Projektarbeit
- weitgehend selbstständiges Arbeiten
- wöchentlich Status-Berichte in Vorlesungszeit

DataScience 2 Kurs

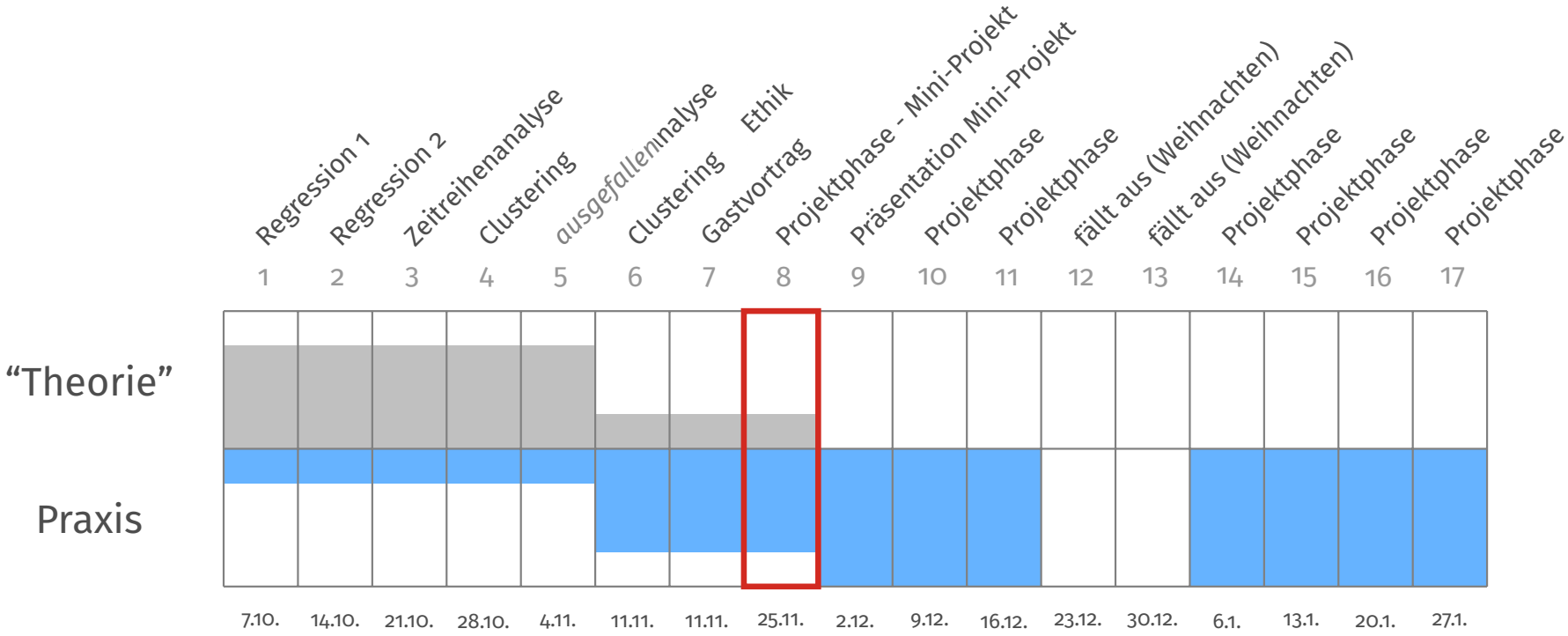
- Als nächstes Projektarbeit
- weitgehend selbstständiges Arbeiten
- wöchentlich Status-Berichte in Vorlesungszeit

- Unterstützung/Hilfestellungen per Mail, Video-Konferenz, Discord

Themen der Vorlesung



Themen der Vorlesung



Inhalt

Ziele und Organisatorisches

Worum geht es im Kurs Data Science?

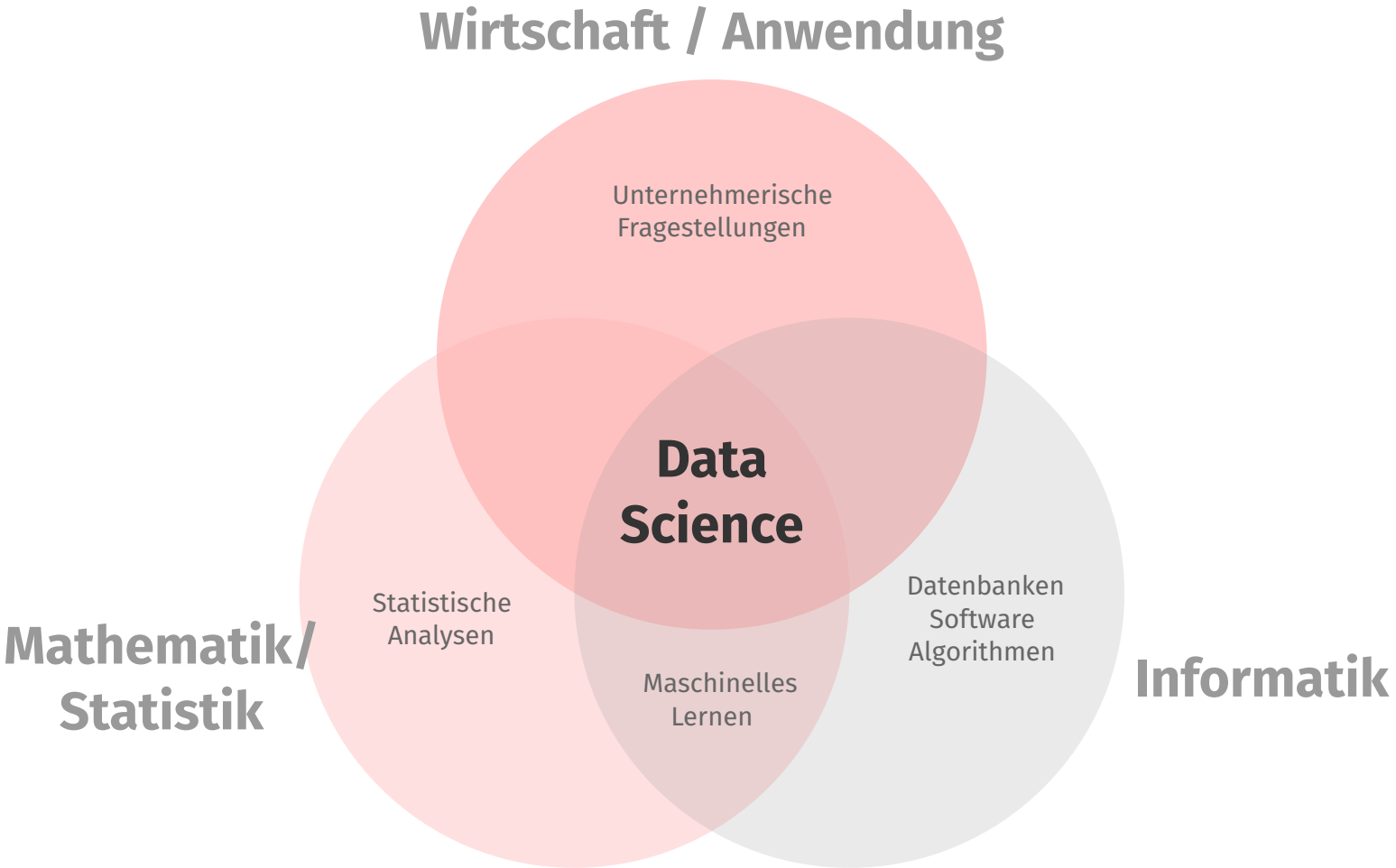
- Etablieren von datengetriebener Denk-/Arbeitsweise
- Kennenlernen von Methoden (ML)
- Wissenschaftliche Arbeitsweise
- Vorbereitung auf BA, Beruf

Worum geht es im Kurs Data Science?

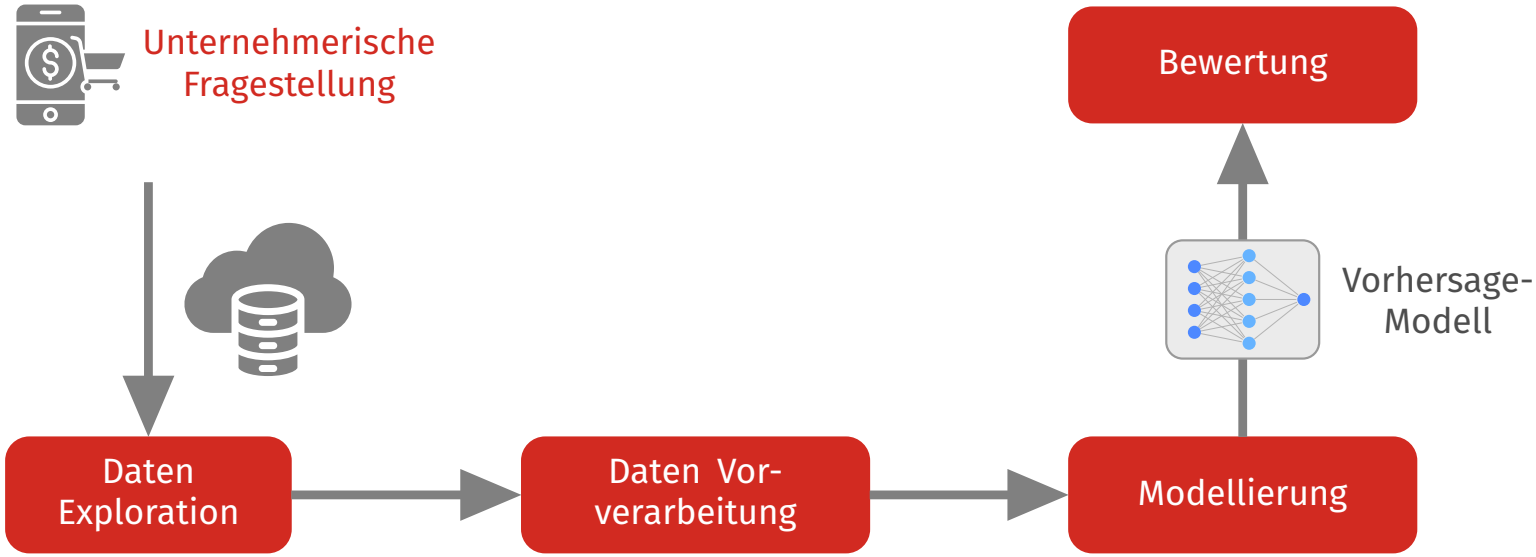
- Etablieren von datengetriebener Denk-/Arbeitsweise
- Kennenlernen von Methoden (ML)
- Wissenschaftliche Arbeitsweise
- Vorbereitung auf BA, Beruf

- Praktische Umsetzung erlernen
- Datenbasiertes *Story-Telling*

Worum geht es?



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?

Beispiel Cont'd

Angenommen, wir haben für jeden Kunden noch

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

Modelllierung – nach Kunde

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

Beispiel Cont'd

Angenommen, wir haben für jeden Kunden noch

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

Modelllierung – 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

Beispiel Cont'd

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

Der Weg zum Data Scientist

"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!

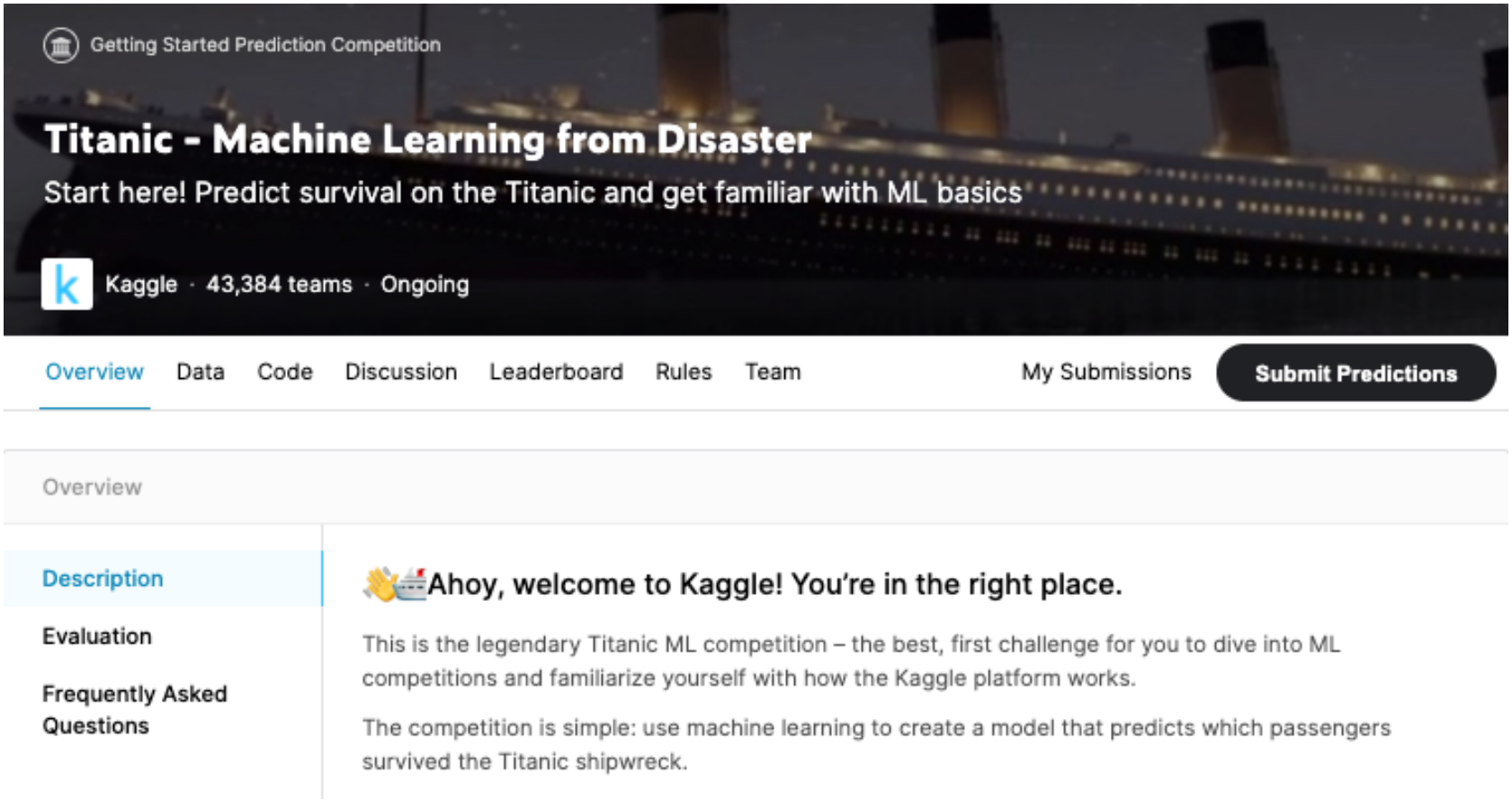
- Viele Blogs/Beispiel Projekte zum Thema DataScience!
- `https://towardsdatascience.com`

Die Kaggle Plattform

Kaggle – DataScience als Wettbewerb

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

Kaggle Einstieg: Titanic Challenge



Projektphase

Mini Projekt – 25.11. bis 9.12.

- Registrieren Sie sich bei kaggle.com
- Melden Sie ein Team für ihre Gruppe an
- Probieren Sie die Titanic Challenge aus!
- Präsentieren Sie ihre Ergebnisse **am 9.12. um 9 Uhr** in der Vorlesung!

Mini Projekt – 25.11. bis 9.12.

- Registrieren Sie sich bei kaggle.com
- Melden Sie ein Team für ihre Gruppe an
- Probieren Sie die Titanic Challenge aus!
- Präsentieren Sie ihre Ergebnisse **am 9.12. um 9 Uhr** in der Vorlesung!

Präsentation

- 1-2 Folien mit Ergebnis + Vorgehen

Mini Projekt – 25.11. bis 9.12.

- Registrieren Sie sich bei kaggle.com
- Melden Sie ein Team für ihre Gruppe an
- Probieren Sie die Titanic Challenge aus!
- Präsentieren Sie ihre Ergebnisse **am 9.12. um 9 Uhr** in der Vorlesung!

Präsentation

- 1-2 Folien mit Ergebnis + Vorgehen

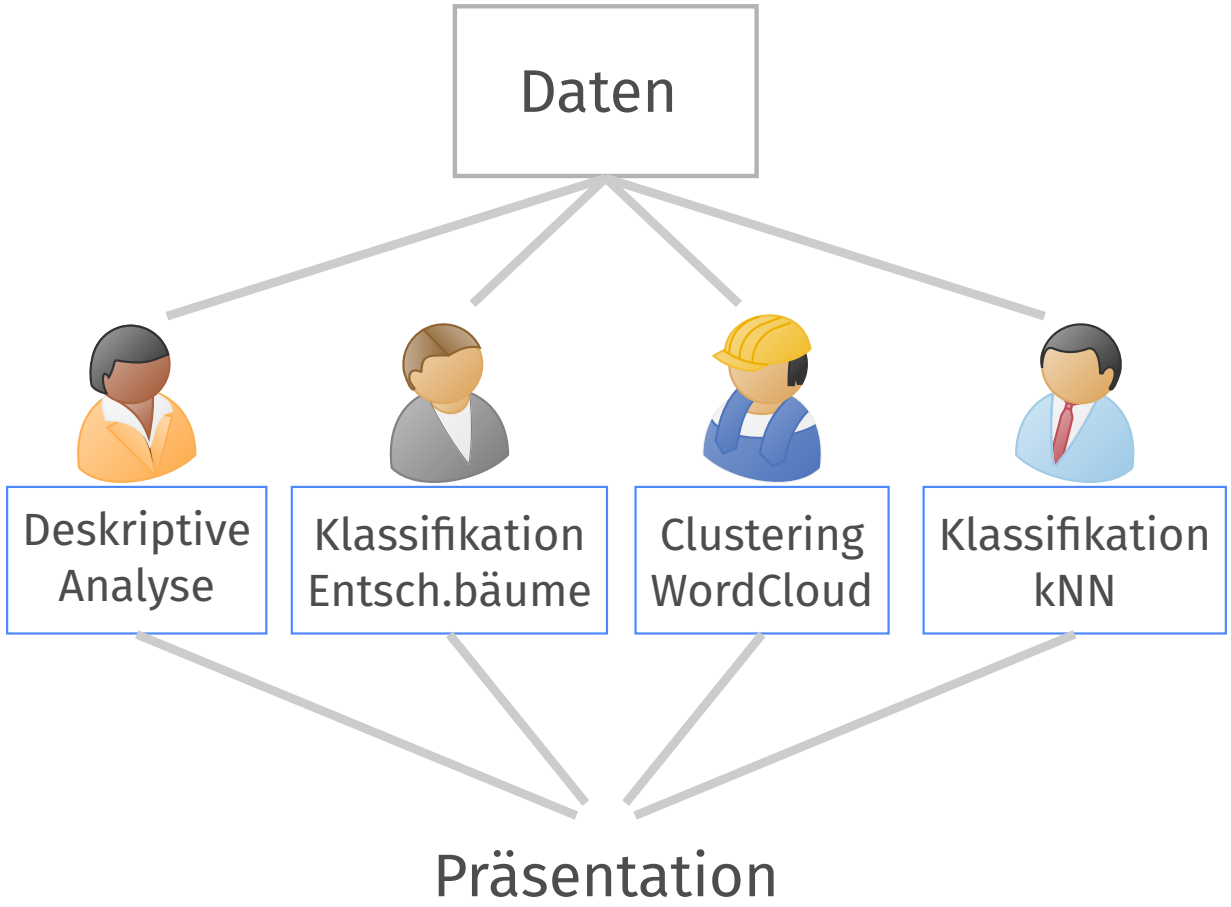
Nächste Woche:

- Keine reguläre Vorlesung
- Bei Bedarf Hilfestellung im BBB Raum von 10-12 Uhr

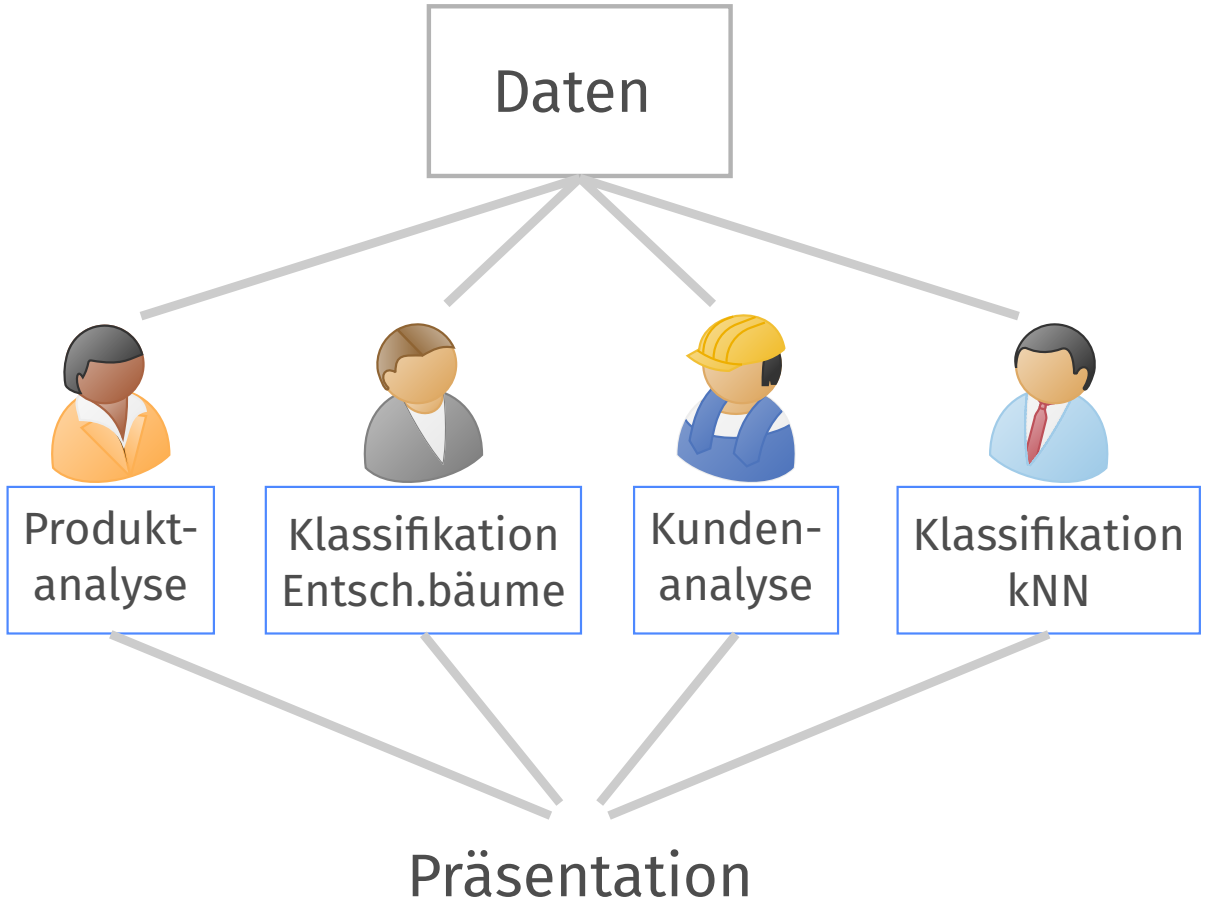
Abschlussprojekte

- Themenvergabe am 9.12. nach Präsentation der Miniprojekte
- Mögliche Themen:
  - Kaggle: Desaster-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 27.1. 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 9.12.