Data Science

Vorlesung 6 - Intro

Prof. Dr. Christian Bockermann

Hochschule Bochum

Sommersemester 2021

Wiederholung

Was geschah zuletzt? Tutorial Day!

- Wiederholung von Python Grundlagen
- Pandas Series + DataFrame

Was geschah zuletzt? Tutorial Day!

- Wiederholung von Python Grundlagen
- Pandas Series + DataFrame

Tutorial Sessions?

- Extra Übungsaufgaben mit Fokus auf Python
- 1 - 2 Stunden zum Üben mit Live-Unterstützung (BBB)

Wiederholung

Was geschah zuletzt? Tutorial Day!

- Wiederholung von Python Grundlagen
- Pandas Series + DataFrame

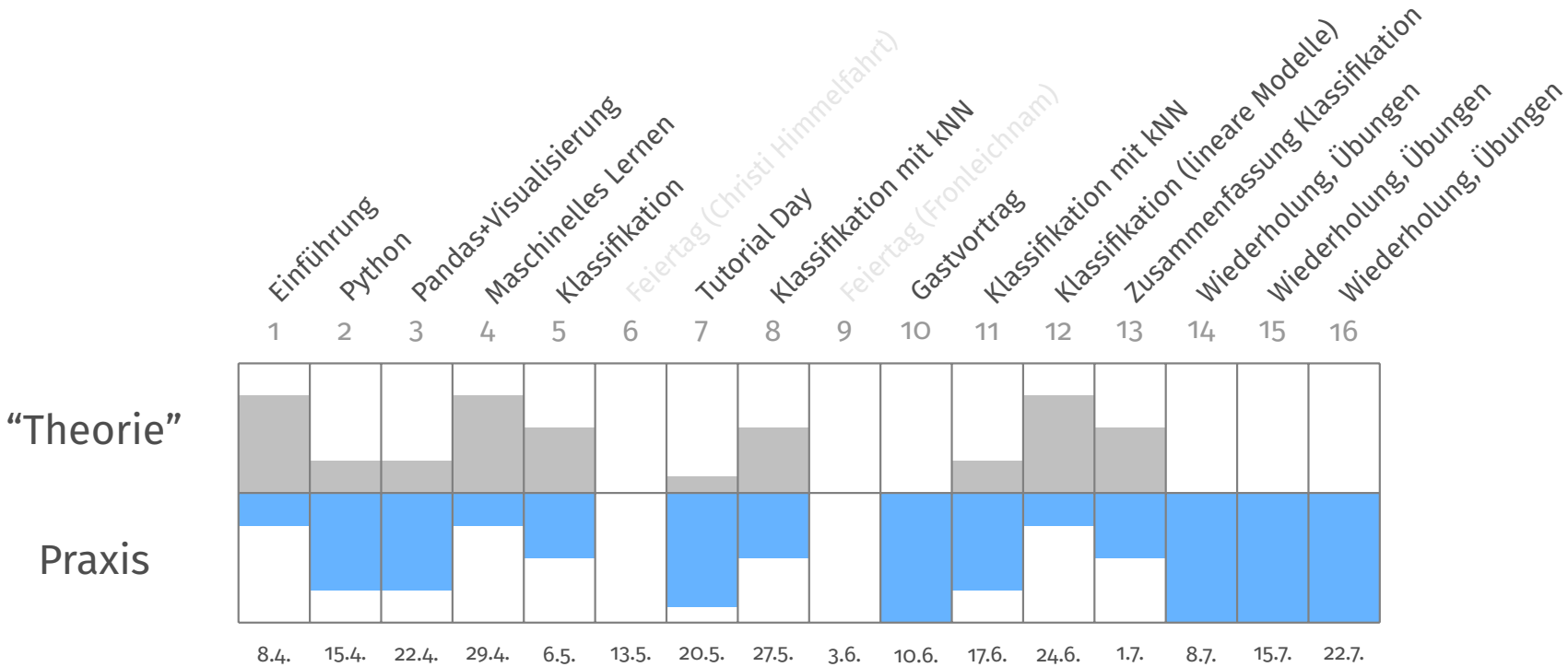
Tutorial Sessions?

- Extra Übungsaufgaben mit Fokus auf Python
- 1 - 2 Stunden zum Üben mit Live-Unterstützung (BBB)
- Mögliche Zeiten:

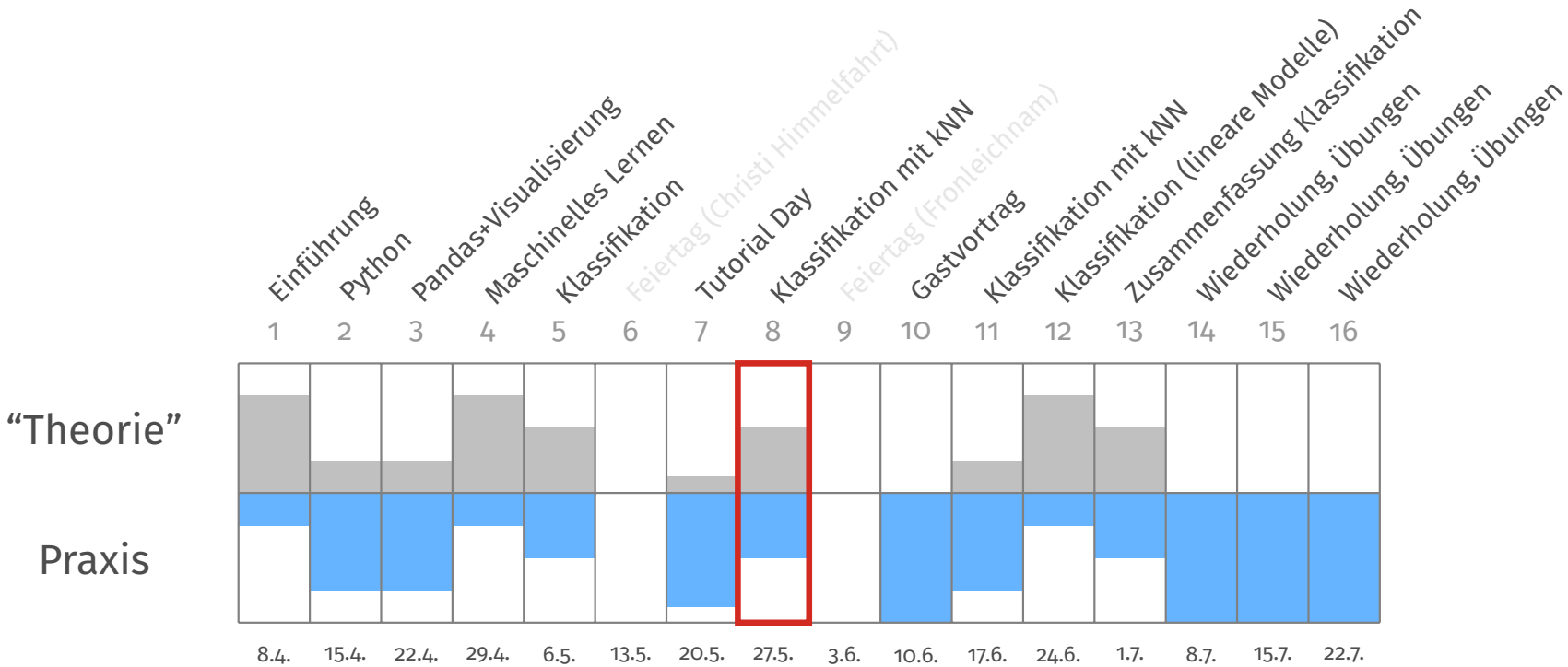
A: Mittwochs, 12-13 Uhr

B: Donnerstags, 14-15 Uhr

Wo sind wir heute (Vorlesung 6) ?



Wo sind wir heute (Vorlesung 6) ?



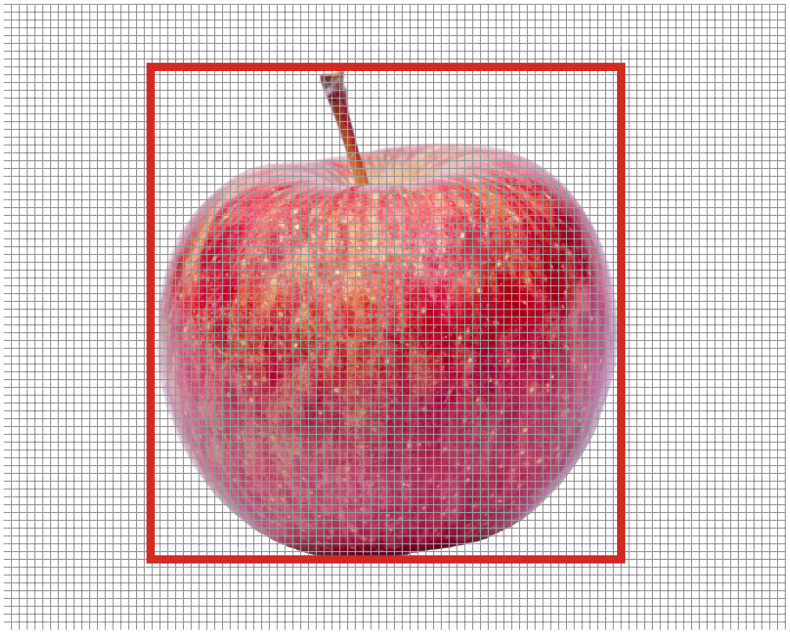
Beispiel: Obstwaage

Supermarkt-Innovation um 2008: Intelligente Waagen

- Die Waage erkennt das Obst/Gemüse per Kamera

https://www.spiegel.de/netzwelt/tech/supermarkt-technik-waage-erkennt-aufgelegtes-gemuese-a-569740.html

Wieder: Bild-Daten



- Objekt Pixel erkennen, Schwerpunkt berechnen
- Maße bestimmen, Formen ausprobieren?
- Mittleren Farb-Index berechnen

Bilder Datensatz: https://www.kaggle.com/moltean/fruits

Einfache Daten aus der Waage

Name	Gewicht	Breite	Höhe	Farbe
apple	192	8.4	7.3	0.55
apple	180	8.0	6.8	0.59
apple	176	7.4	7.2	0.6
mandarin	86	6.2	4.7	0.8
mandarin	84	6.0	4.6	0.79
mandarin	80	5.8	4.3	0.77

Einfache Daten aus der Waage

Name	Gewicht	Breite	Höhe	Farbe
apple	192	8.4	7.3	0.55
apple	180	8.0	6.8	0.59
apple	176	7.4	7.2	0.6
mandarin	86	6.2	4.7	0.8
mandarin	84	6.0	4.6	0.79
mandarin	80	5.8	4.3	0.77

Ähnlichkeit über Abstandsmaß - unterschiedliche Skalen!

Datensatz (48 Früchte): Vorlesung/data/fruits_with_colors.csv

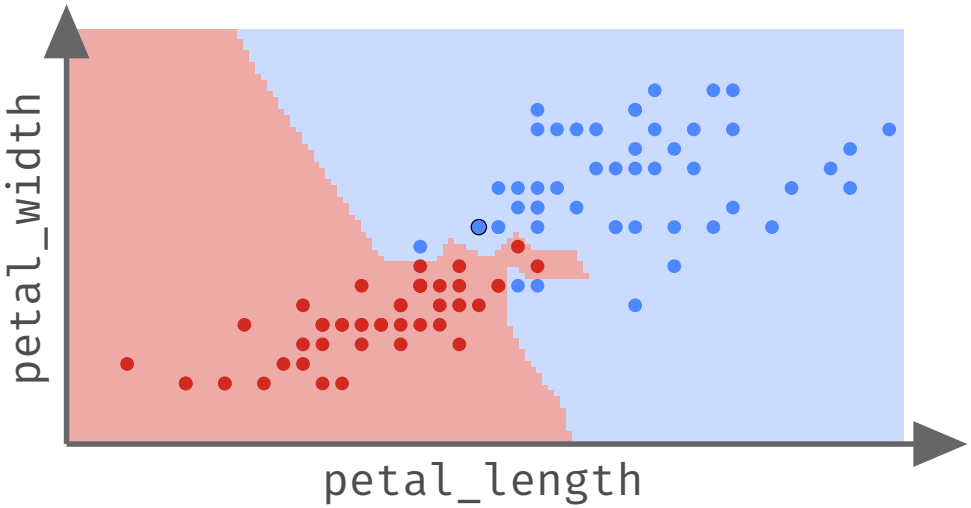
Diskussion: Wie genau müssen wir sein?

Diskussion: Wie genau müssen wir sein?

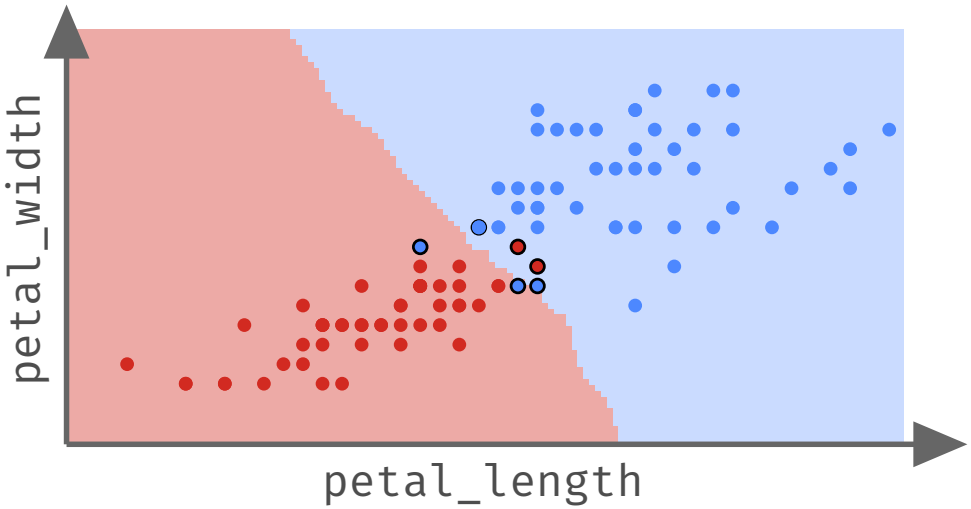
Diskussion: Wie genau müssen wir sein?

Diskussion: Wie genau müssen wir sein?

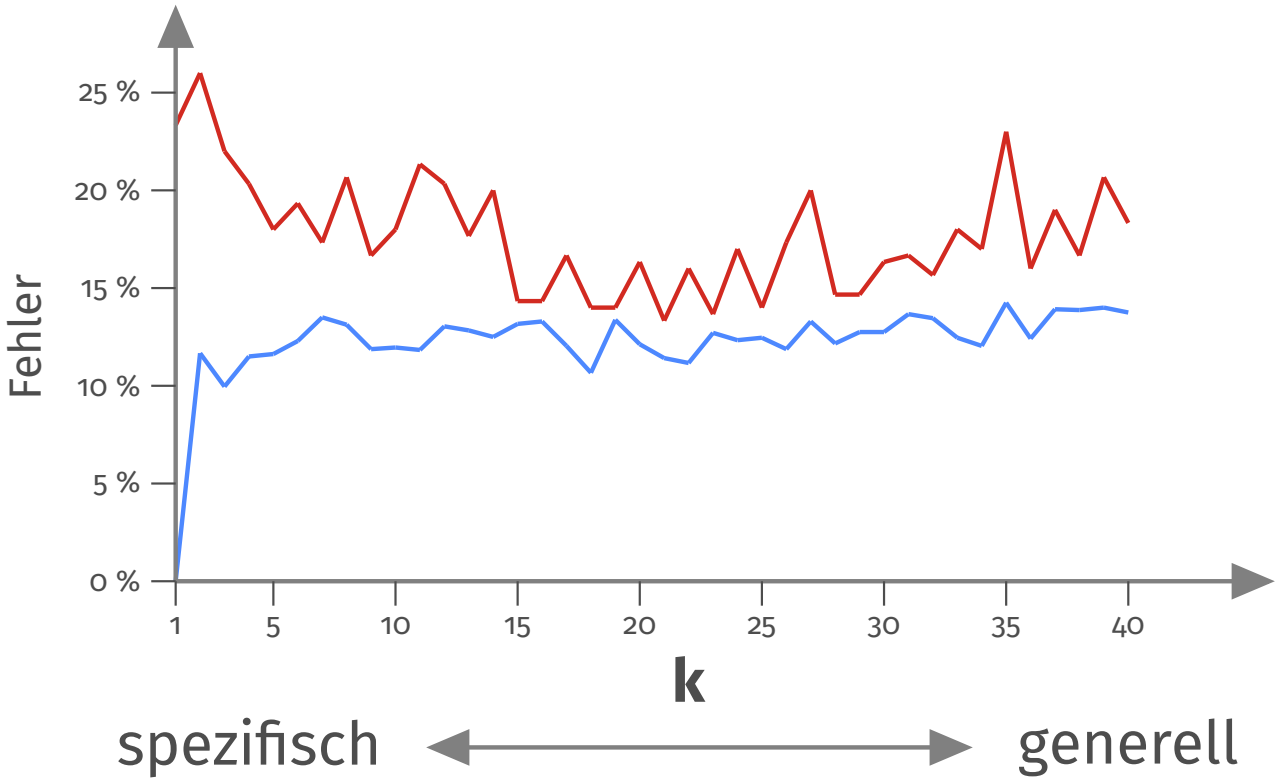
k=1



k=16



Training und Test-Fehler auf generiertem Datensatz (k-NN)



Overfitting

“Das Modell passt nur zu den Trainingsdaten.”

Overfitting

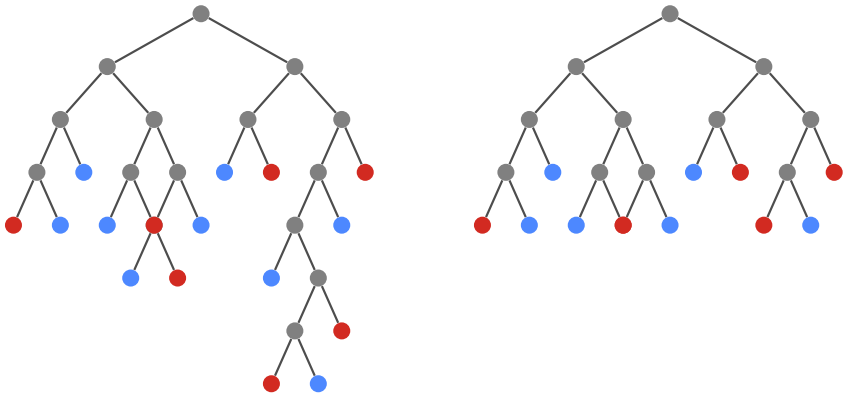
"Das Modell passt nur zu den Trainingsdaten."

	Trainingsfehler **klein**	Trainingsfehler **groß**
Testfehler **klein**	Das sieht gut aus!	
Testfehler **groß**	**Overfitting!**	Das Modell lernt nicht!?

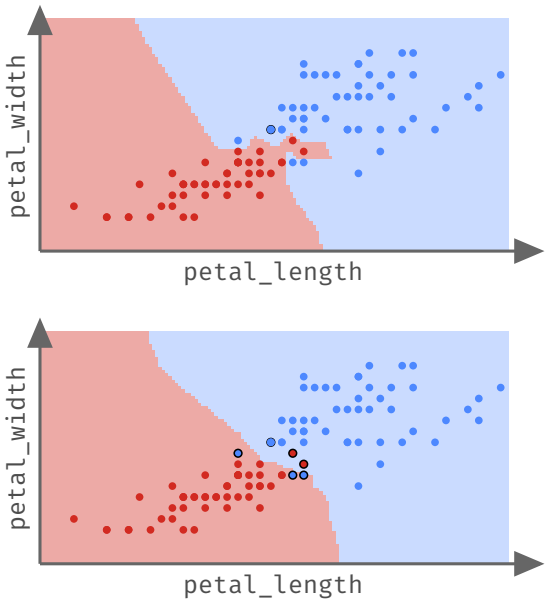
Overfitting - zu spezifisches Modell

- Modell zu sehr an die Trainingsdaten angepasst
- Vorhersage auf unbekannten Daten schlechter
- Modellkomplexität begrenzen (generelleres Modell)

Tiefe bei Bäumen beschränken



k bei k-NN erhöhen



A capella Song zum Thema Overfitting

https://youtu.be/DQWI1kvmwRg

Wie geht's weiter?

Tutorials:

- heute, 27.5.2021, 14:00 Uhr bis 15:00 Uhr

weitere Tutorial-Termine:

- Mittwoch, 2.6.2021, 12:00 Uhr bis 13:00 Uhr
- Donnerstag, 3.6.2021 – entfällt wegen Feiertag

- Mittwoch, 9.6.2021, 12 Uhr bis 13 Uhr
- Donnerstag, 10.6.2021, 14 Uhr bis 15 Uhr

Wie geht's weiter?

Vorschau auf Vorlesung 7:

- **Gastvortrag** von **Jonas Rashedi**,
  Parfümerie Douglas (eCom)

Virtueller Raum für die Vorlesung 7:

http://glight.hs-bochum.de/b/chr-nti-sjt