Data Science 1

Vorlesung 6 - Intro

Prof. Dr. Christian Bockermann

Hochschule Bochum

Sommersemester 2023

Was geschah zuletzt?

- Einfache Klassifikation
- Entscheidungsbäume, Lernen aus Daten

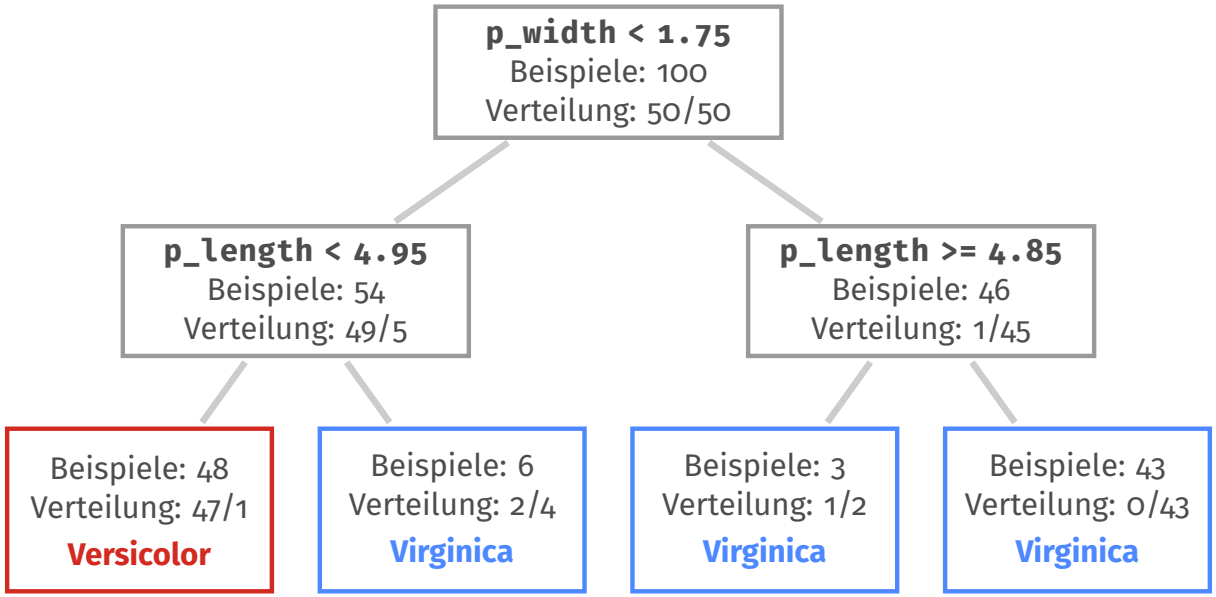
Was geschah zuletzt?

- Einfache Klassifikation
- Entscheidungsbäume, Lernen aus Daten

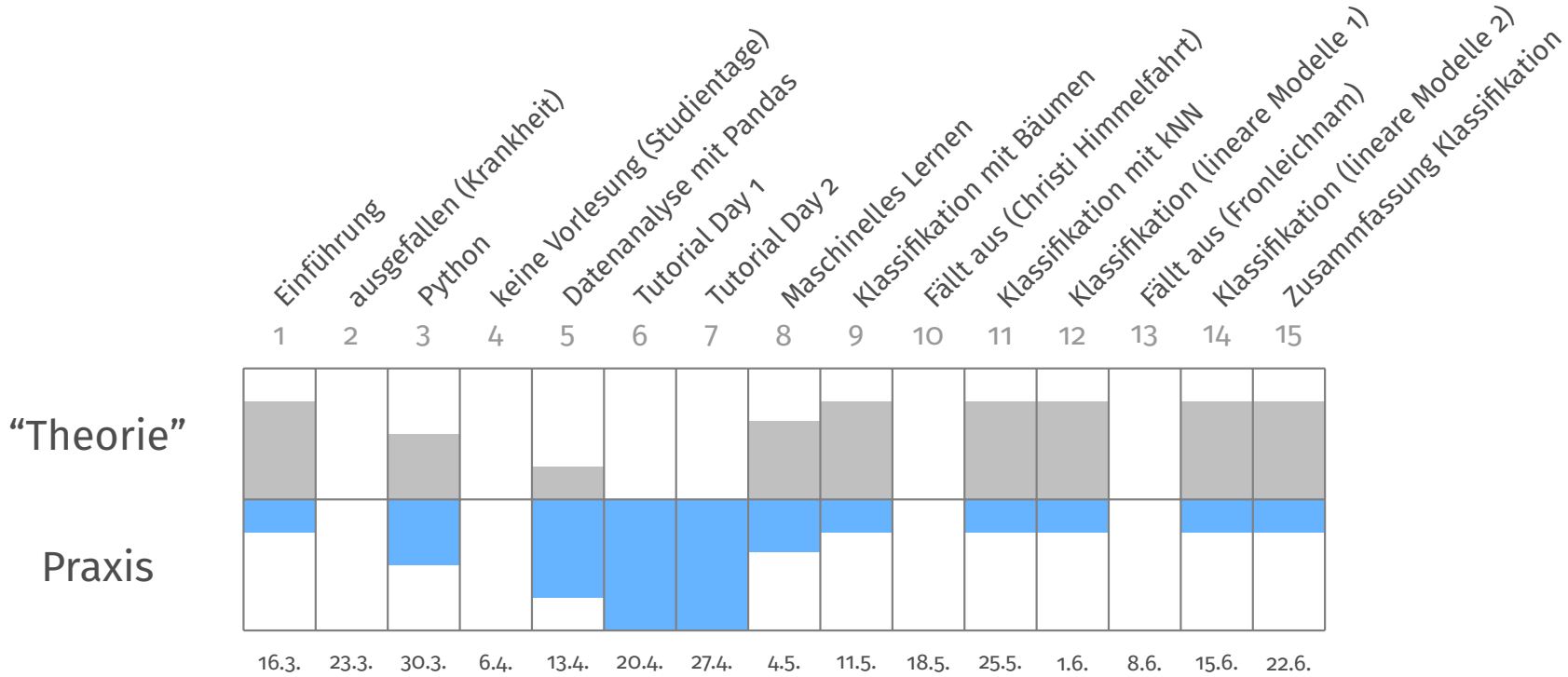
1. Model `m` (Baum) auf Trainingsdaten gelernt
2. Dann mit `m` Testdaten vorhergesagt
3. Vorhersage-Fehler bestimmt

Einfaches Modell: Entscheidungsbäume

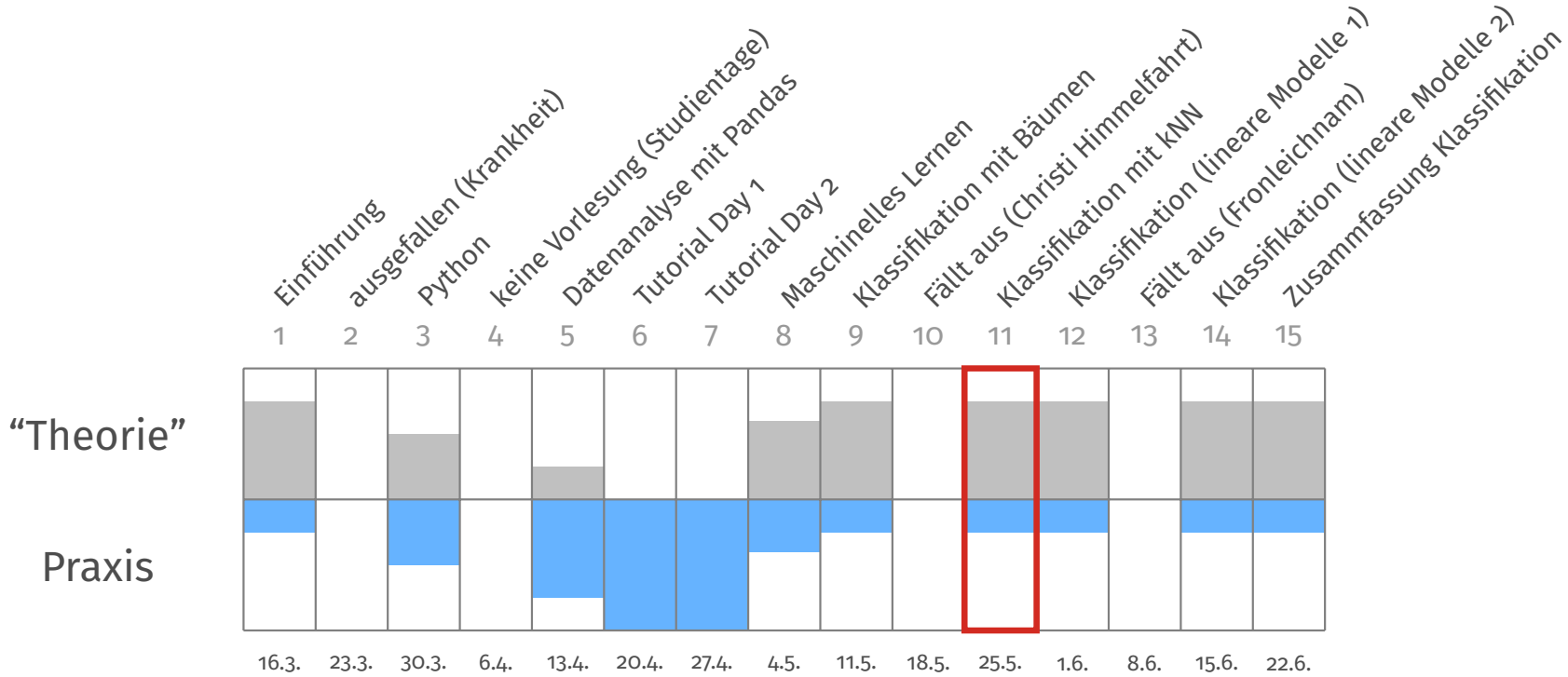
- Innere Knoten sind Entscheidungsknoten
- Blätter stellen Vorhersage dar



Wo sind wir heute (Vorlesung 6) ?

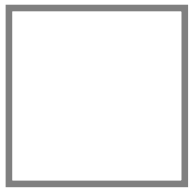


Wo sind wir heute (Vorlesung 6) ?



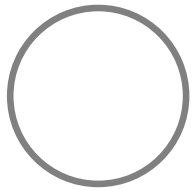
Menschliches Lernen nutzt Ähnlichkeiten aus

Zum Beispiel über Formen:



Quadrat

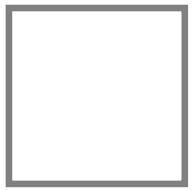
Rechteck



Kreis

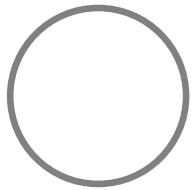
Menschliches Lernen nutzt Ähnlichkeiten aus

Zum Beispiel über Formen:



Quadrat

Rechteck



Kreis

Oder andere Eigenschaften (Größe, Gewicht):

Fussball

Handball

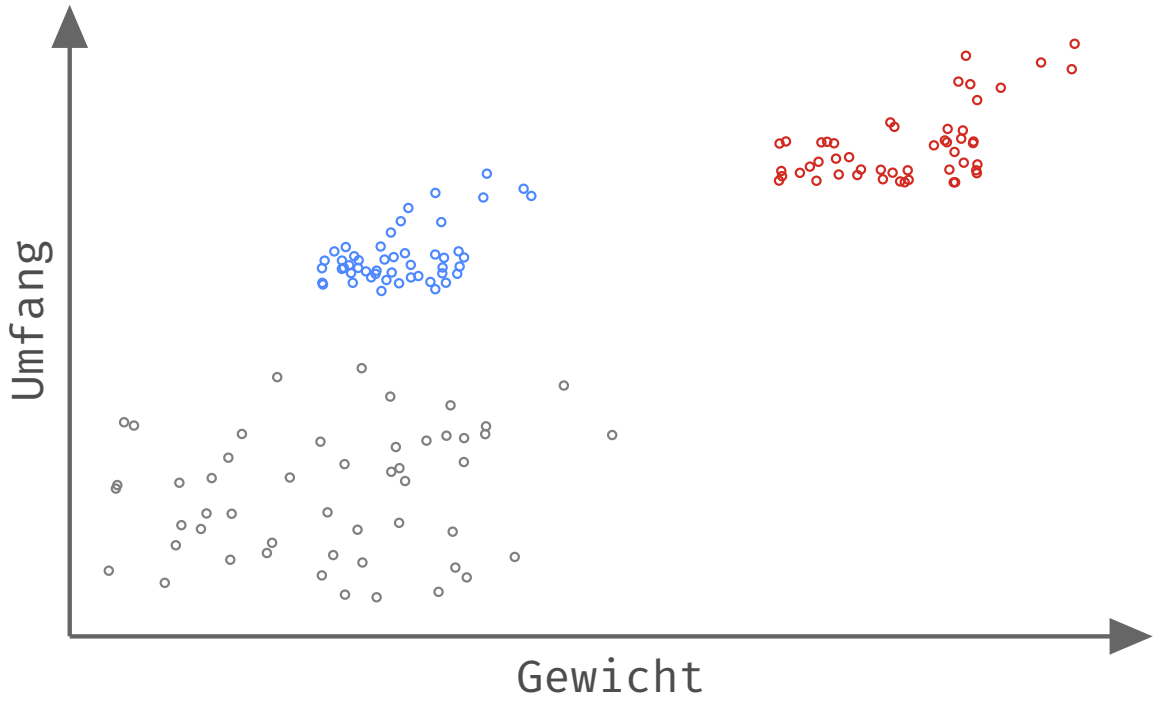
Basketball

Beispiel: Klassifikation von Bällen

Wir wollen Bälle ihrer Sportart zuordnen (Klassifikationsaufgabe)

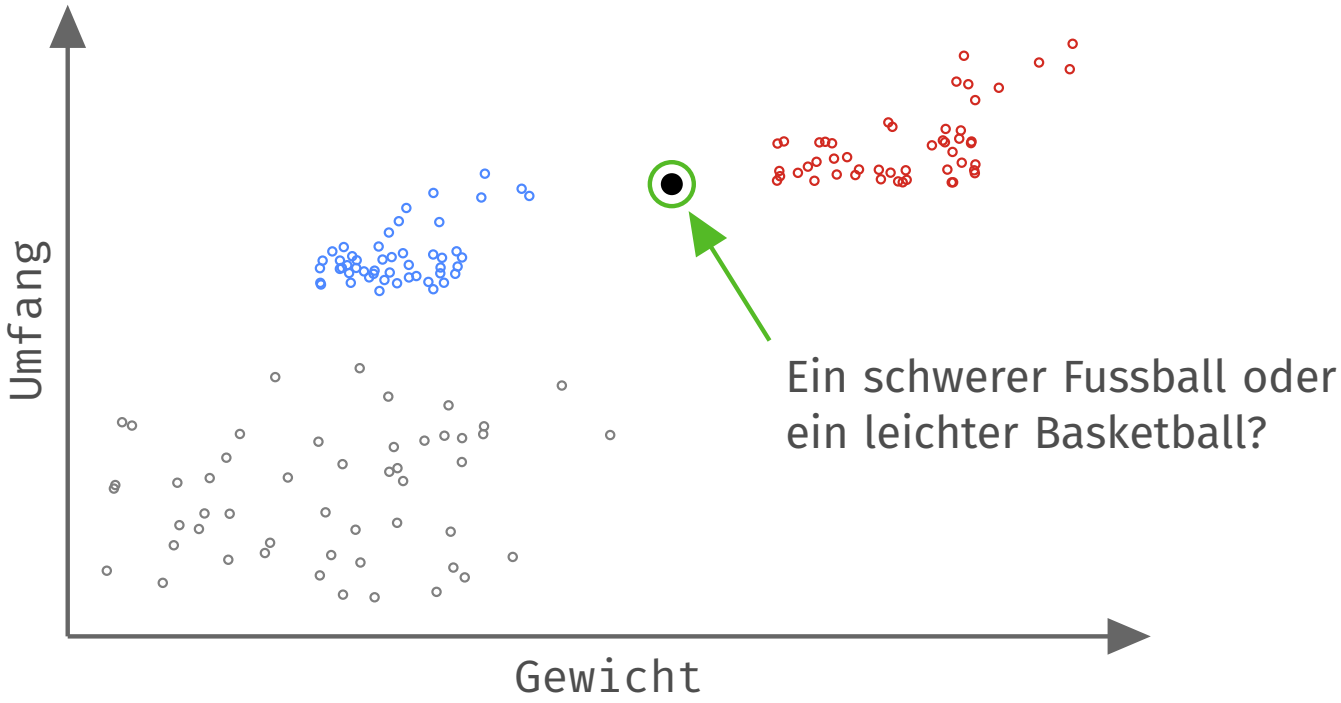
Umfang (cm)	**Gewicht** (g)	**Sportart**
70.29	444.30	Fussball
77.73	647.53	Basketball
53.34	427.07	Handball
57.09	406.12	Handball
68.28	440.96	Fussball
80.38	648.94	Basketball

Beispiel: Maße unterschiedlicher Bälle



Verschiedene Bälle nachgemessen: Fussball, Handball und Basketball

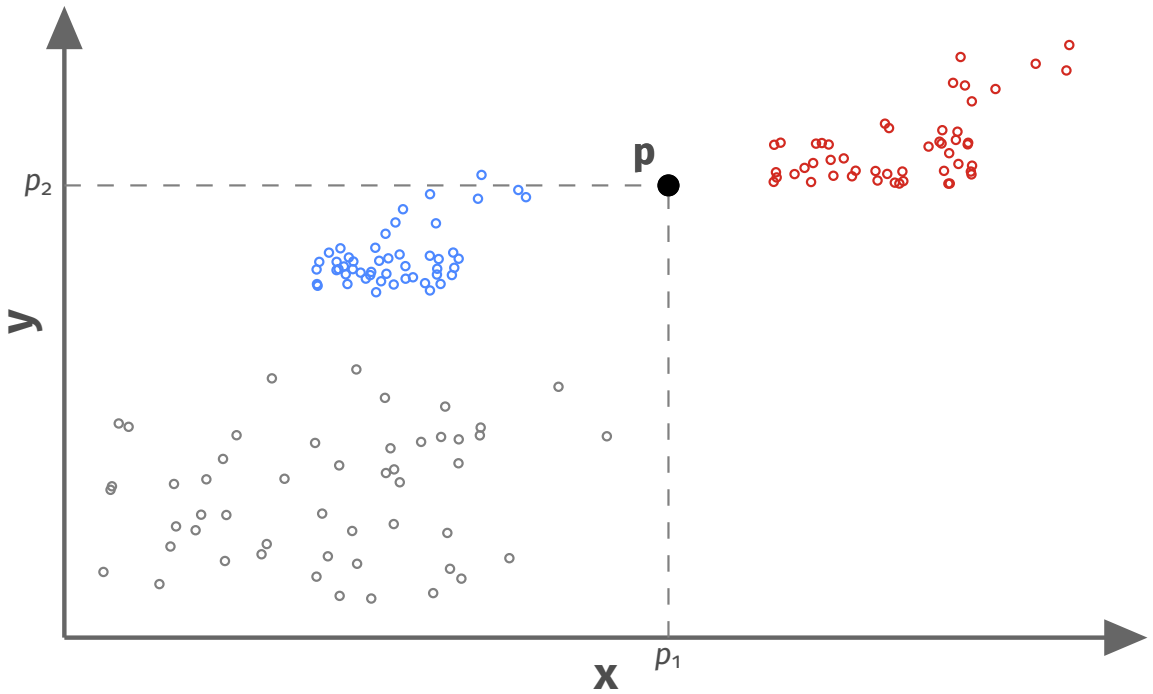
Beispiel: Maße unterschiedlicher Bälle



Verschiedene Bälle nachgemessen: Fussball, Handball und Basketball

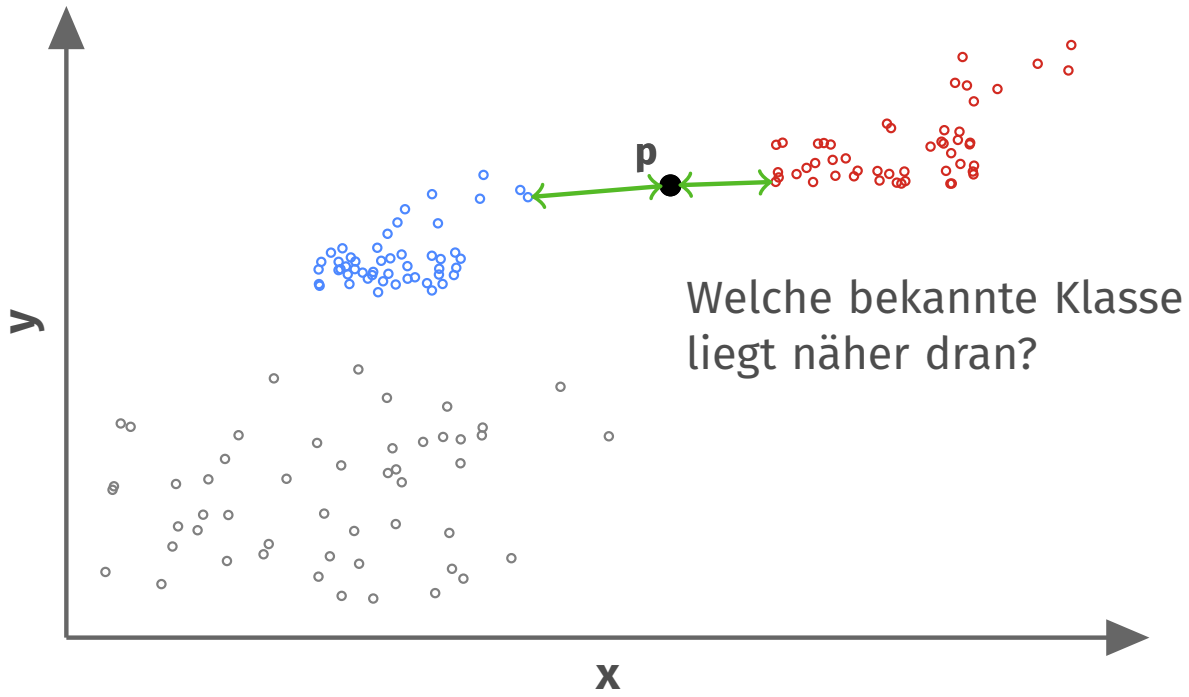
Beispiel: Ball-Arten

Betrachte 2-dimensionalen Raum: $\mathbb{R}^2$



Beispiel: Ball-Arten

Betrachte 2-dimensionalen Raum: $\mathbb{R}^2$



Idee: Wir nutzen den Abstand als Ähnlichkeit und sagen die Klasse vorher, die am nächsten ist!

Folien: DataScience1-06-Folien.pdf

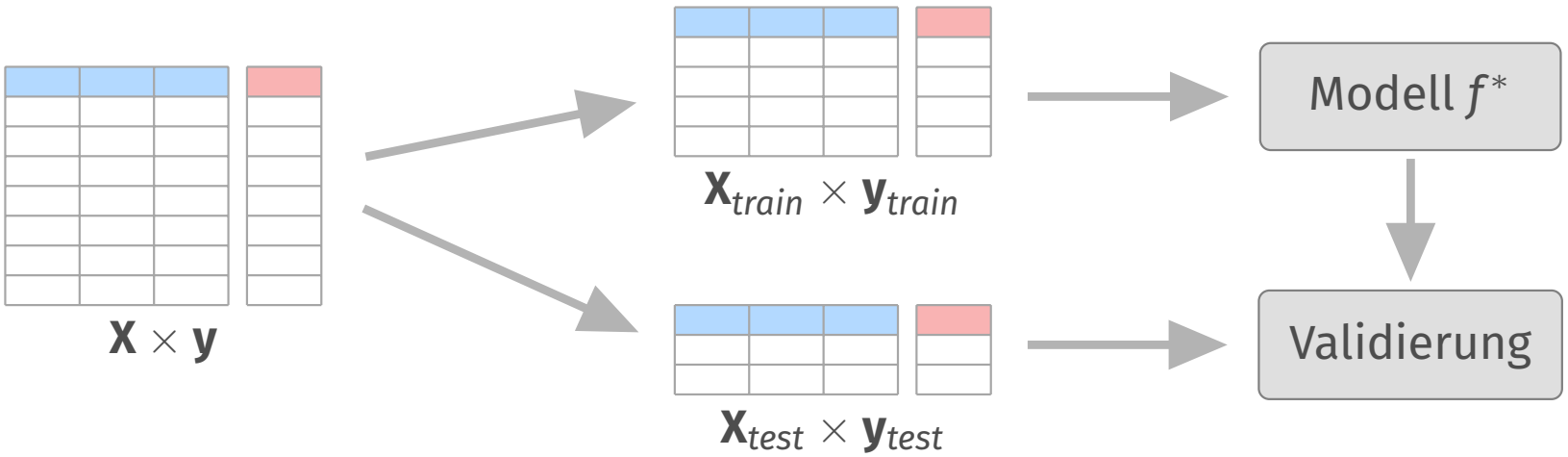
Diskussion: Wie genau müssen wir sein?

Diskussion: Wie genau müssen wir sein?

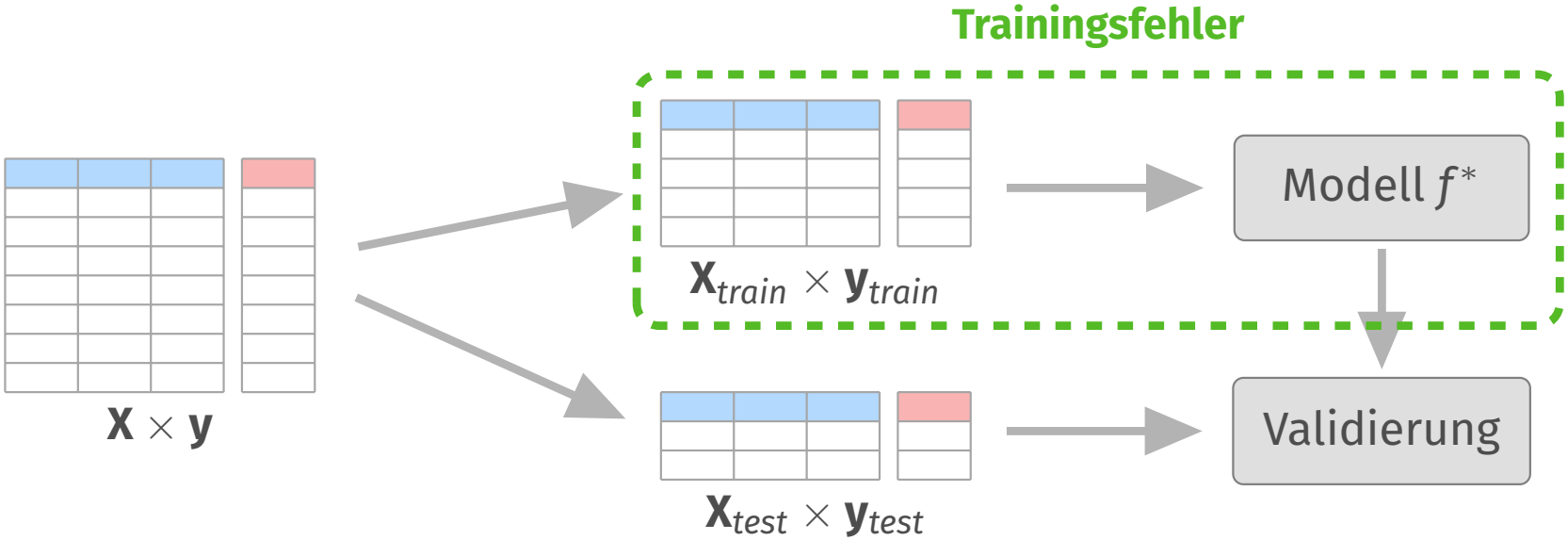
Diskussion: Wie genau müssen wir sein?

Diskussion: Wie genau müssen wir sein?

Diskussion: Wie genau müssen wir sein?

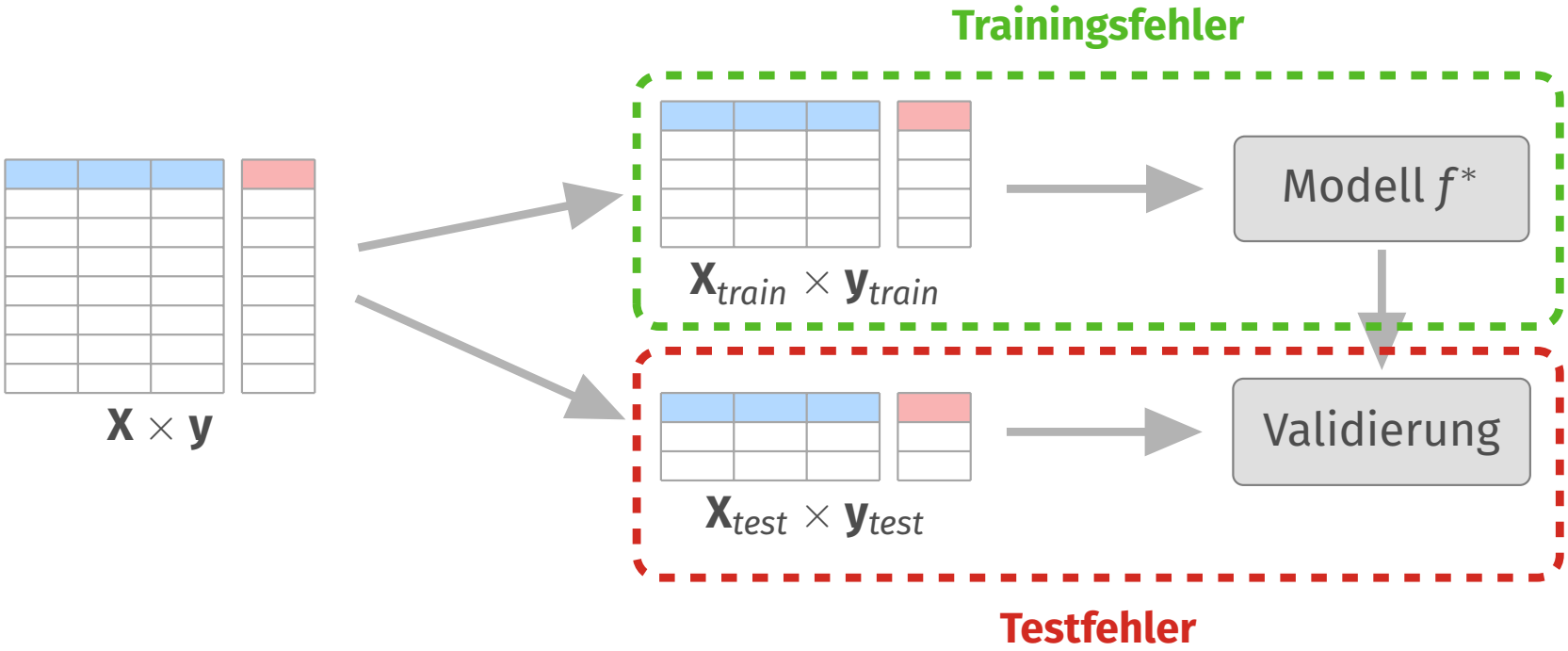


Diskussion: Wie genau müssen wir sein?



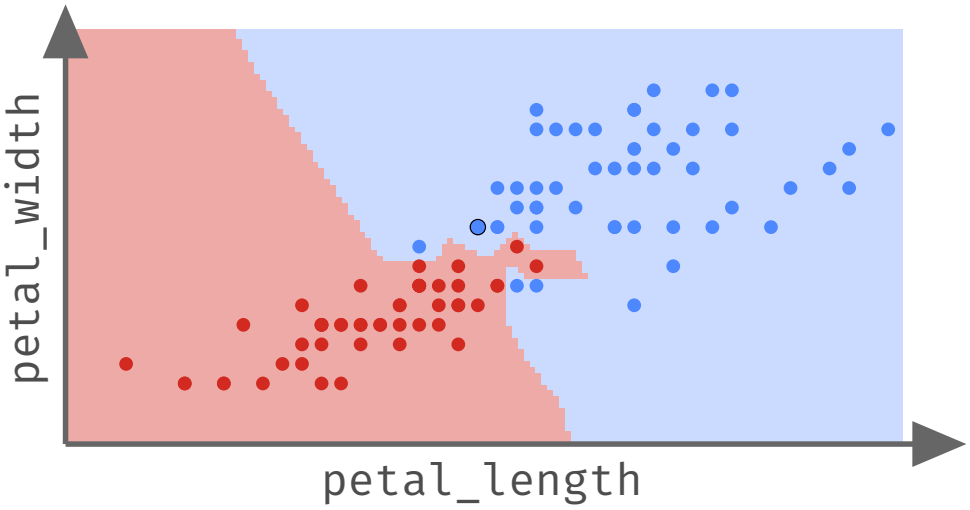
- Trainingsfehler: Modell an Trainingsdaten anpassen

Diskussion: Wie genau müssen wir sein?

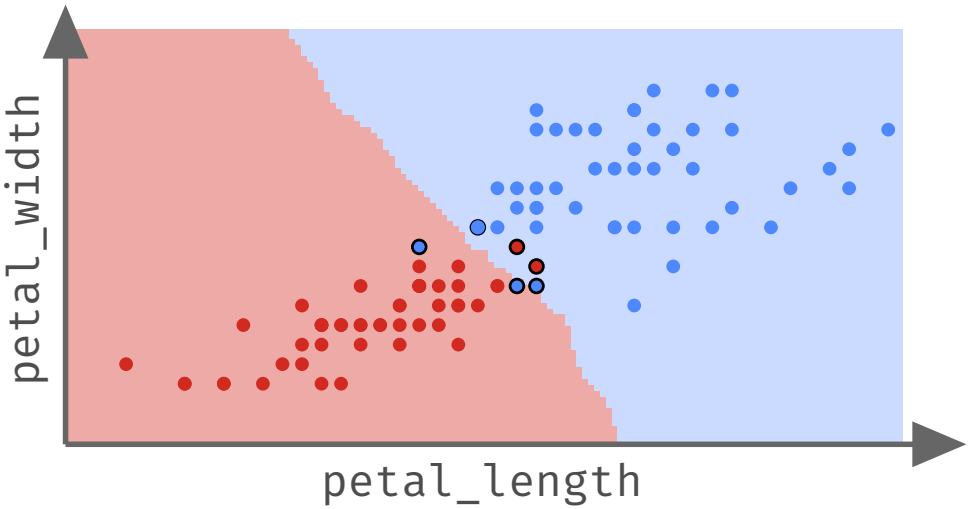


- Trainingsfehler: Modell an Trainingsdaten anpassen
- Testfehler: Modell auf unbekannten Daten (auch: Generalisierungsfehler)

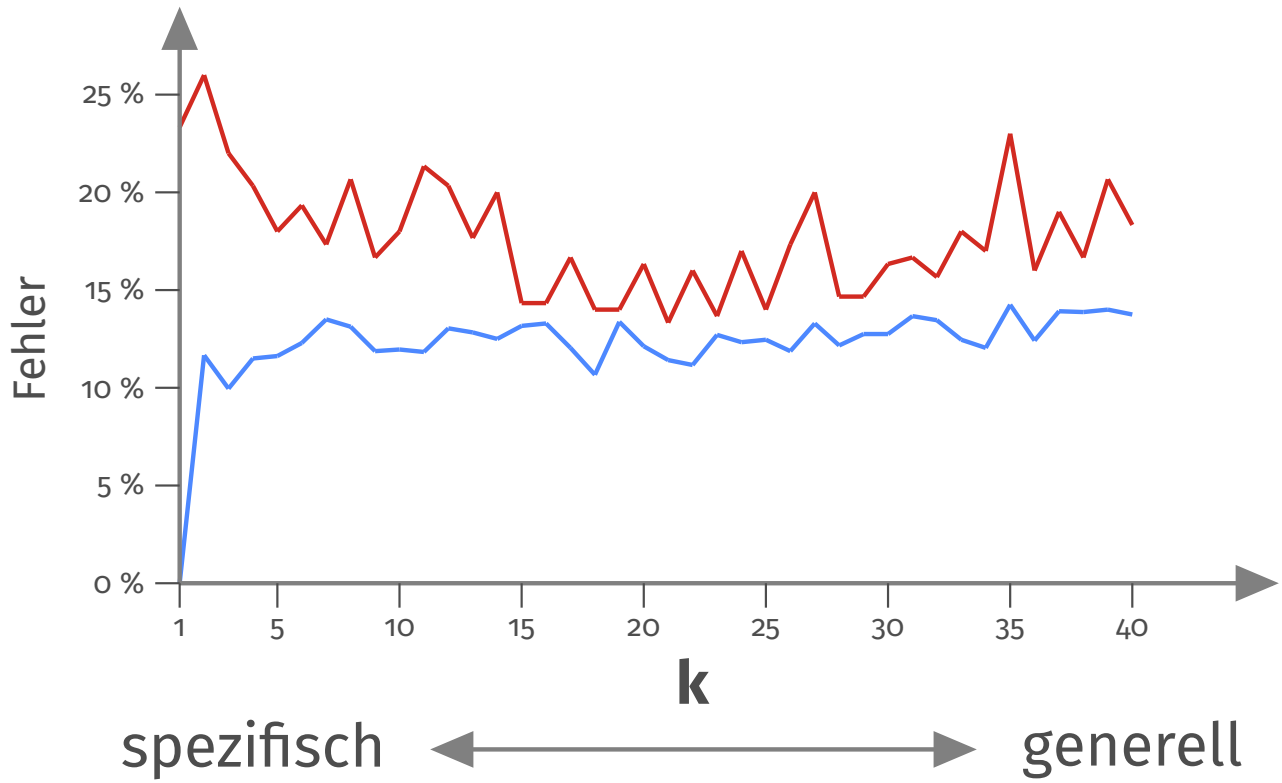
k=1



k=16



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



Overfitting

“Das Modell passt nur zu den Trainingsdaten.”

Overfitting

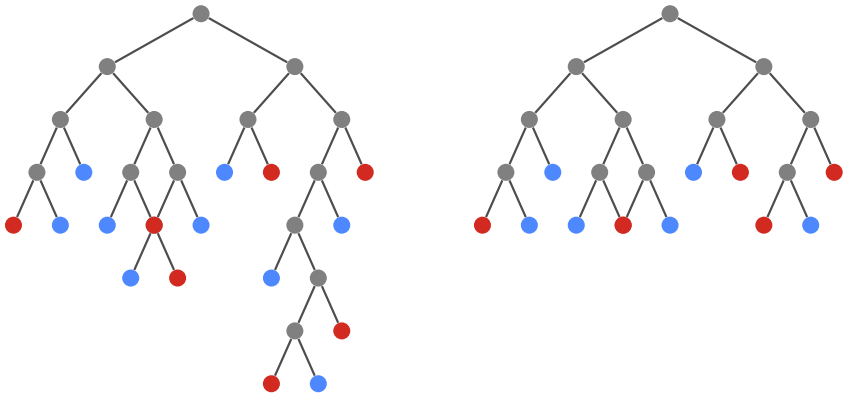
"Das Modell passt nur zu den Trainingsdaten."

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

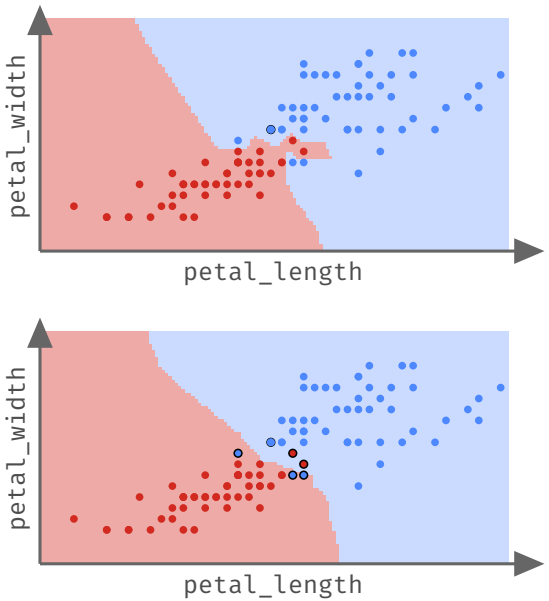
Overfitting - zu spezifisches Modell

- 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

Weitere Klassifikationsverfahren

- Klassifikation mit linearen Funktionen
- Vertiefung der Python-Kenntnisse

Weitere Klassifikationsverfahren

- Klassifikation mit linearen Funktionen
- Vertiefung der Python-Kenntnisse

Hausarbeit - Termine

- Bekanntgabe des Themas: 22.6. nach der Vorlesung (Übung)
- Abgabe: 31.7.2023 bis 23:59 Uhr per Mail

Terminplanung Hausarbeit