Data Science

Vorlesung 7 - Intro

Prof. Dr. Christian Bockermann

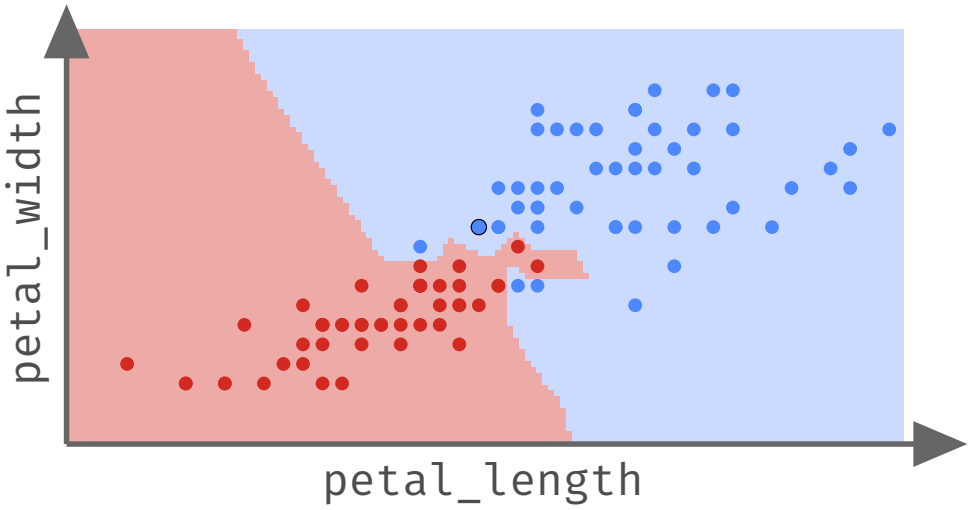
Hochschule Bochum

Sommersemester 2021

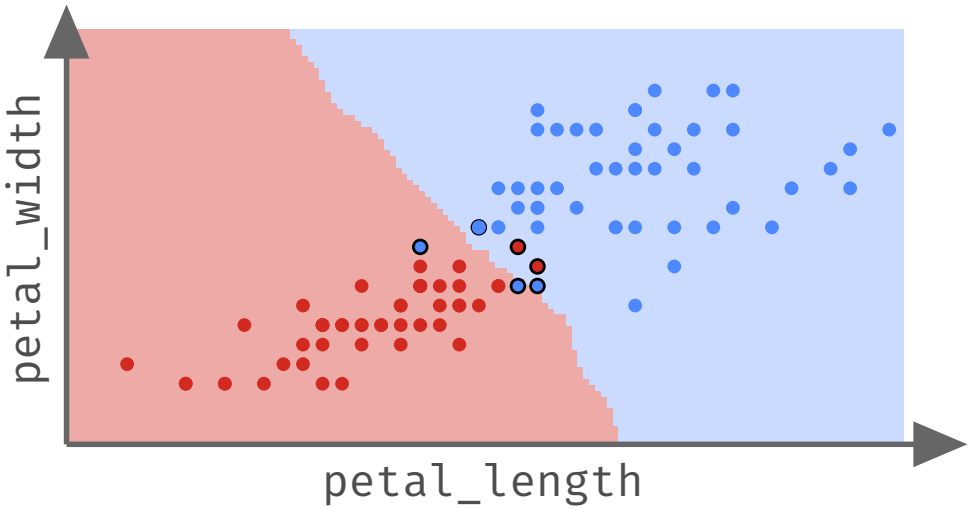
Was geschah zuletzt?

- Instanzbasiertes Lernen über Ähnlichkeit
- Distanz-Funktion auf Beispielen (eukl. Distanz)
- Normalisierung von Daten (Min/Max-, z-Normalisierung)
- k-NN als Vorhersagemodell

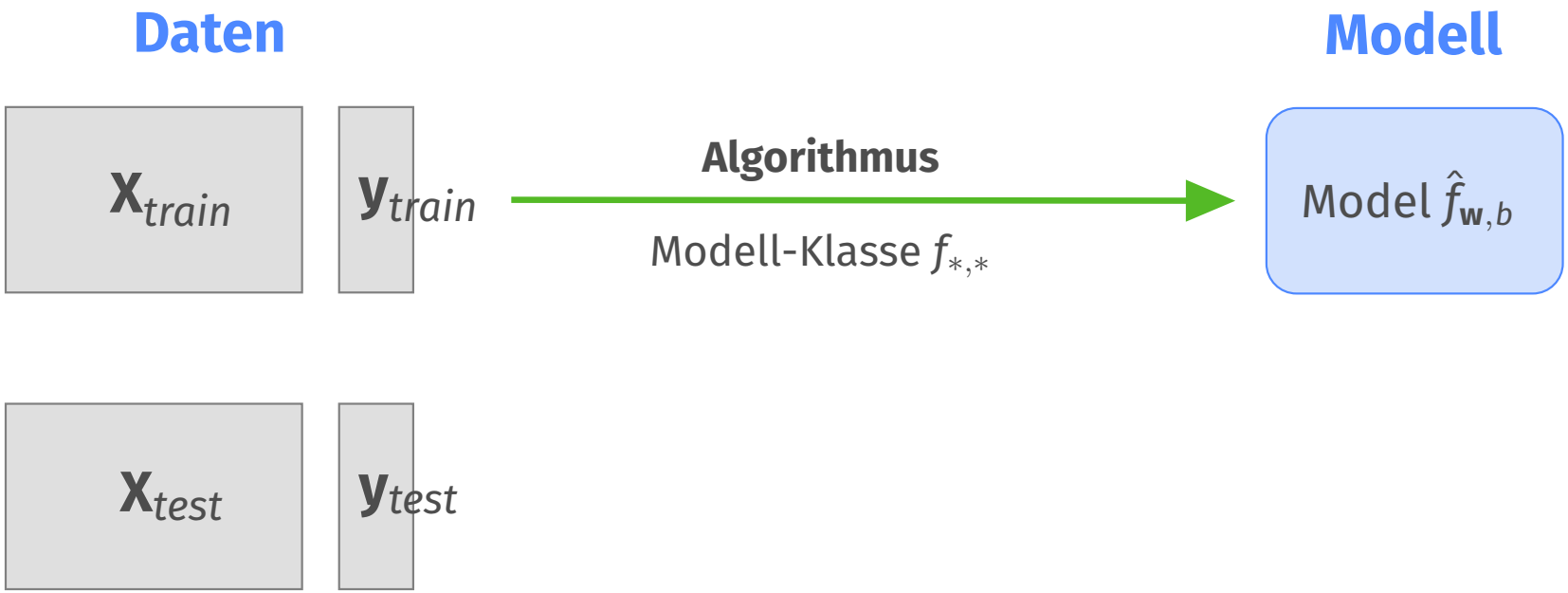
k=1



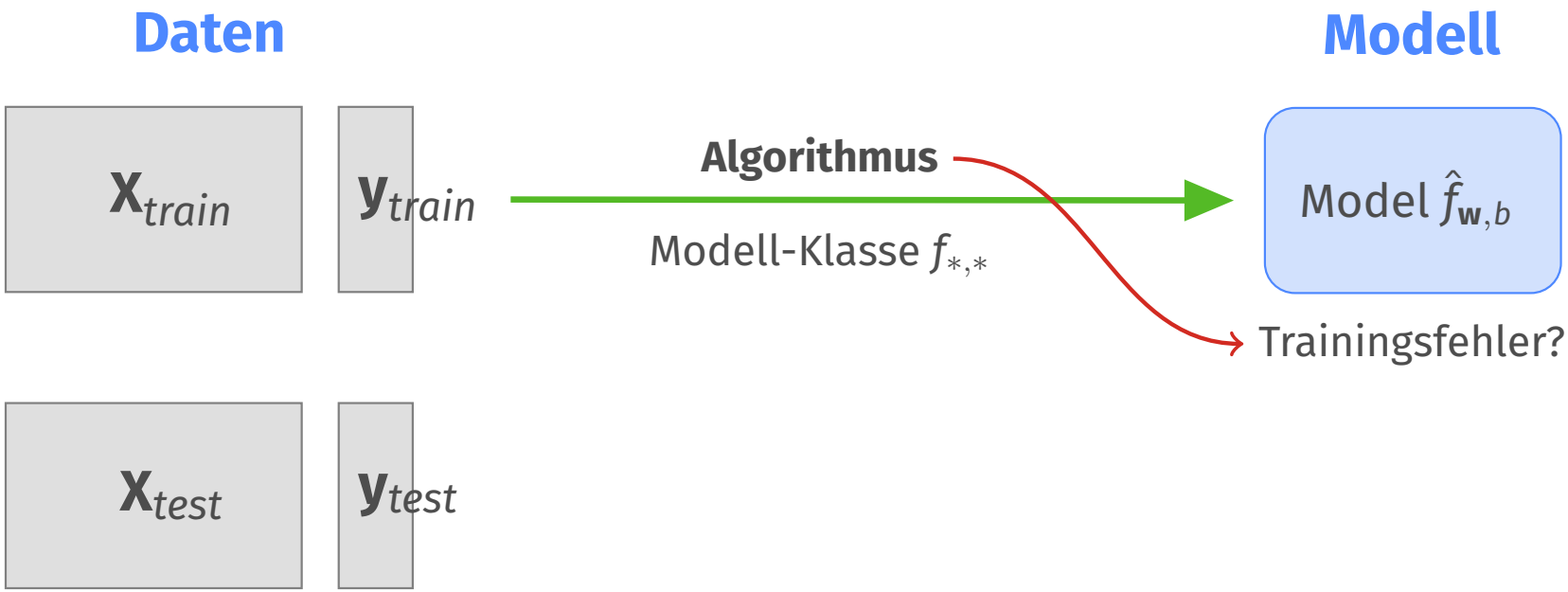
k=16



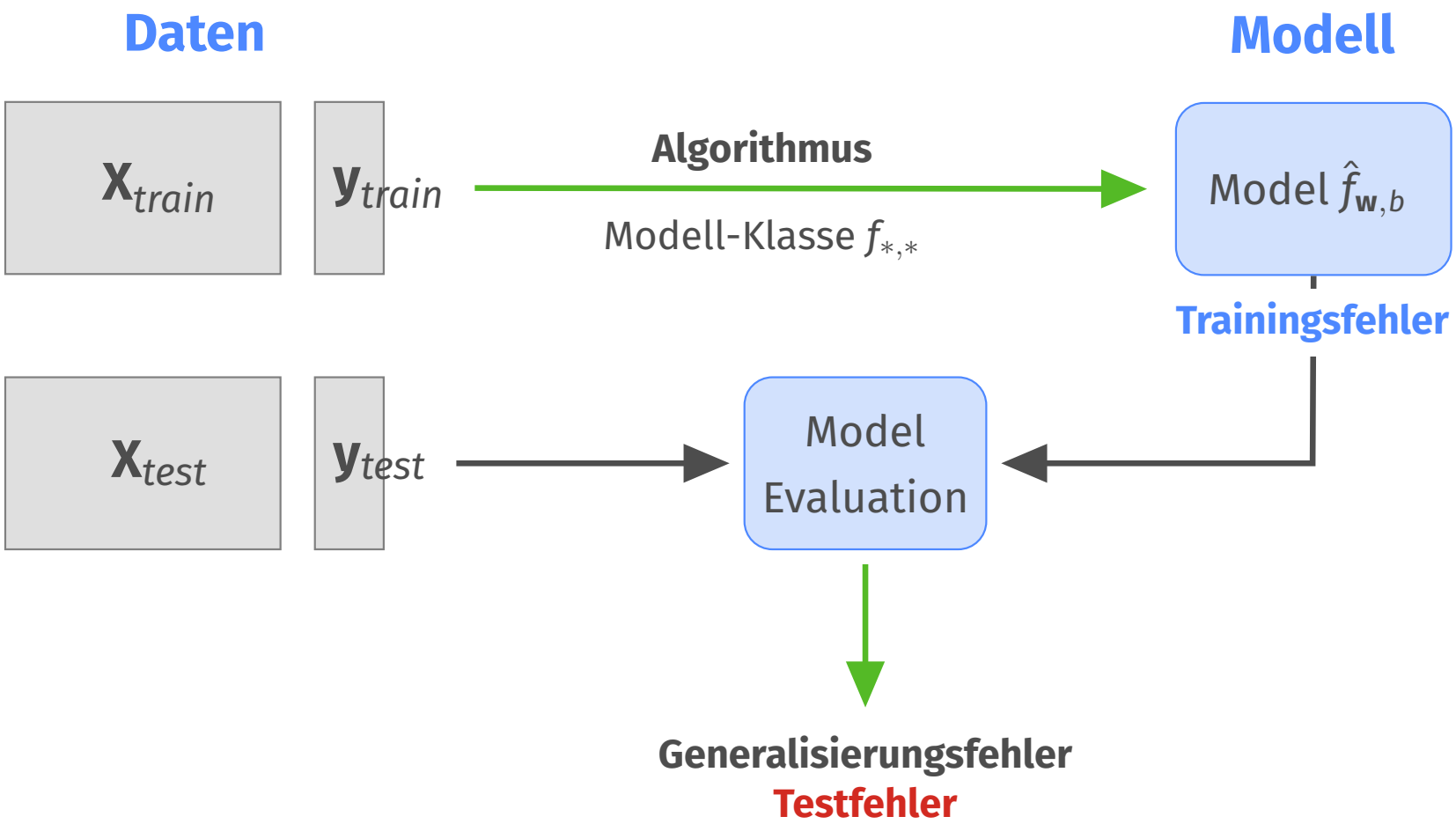
Bewertung der Modell-Güte – Generalisierungsfehler



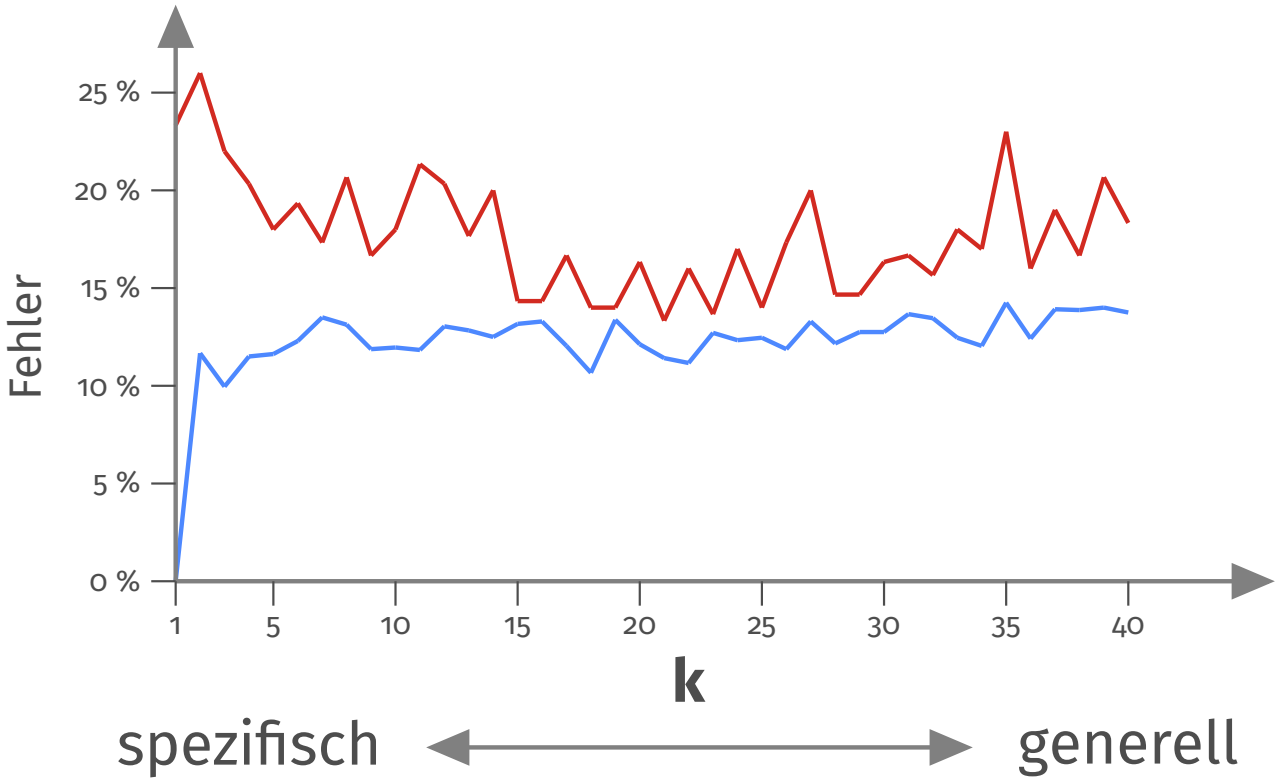
Bewertung der Modell-Güte – Generalisierungsfehler



Bewertung der Modell-Güte – Generalisierungsfehler



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



Overfitting

“Das Modell passt nur zu den Trainingsdaten.”

Overfitting

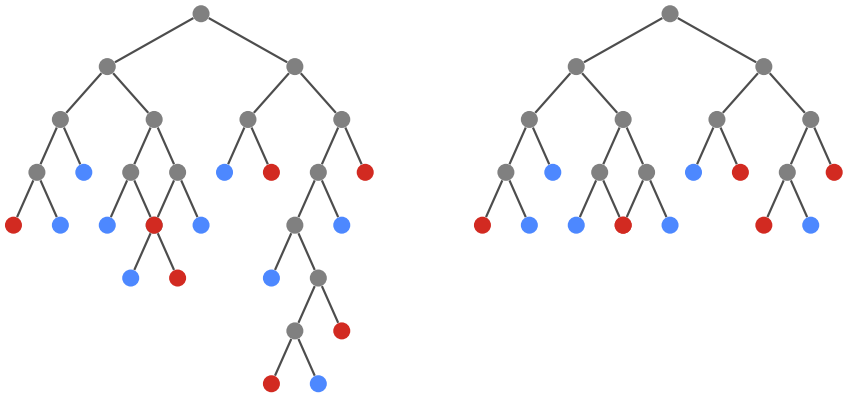
"Das Modell passt nur zu den Trainingsdaten."

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

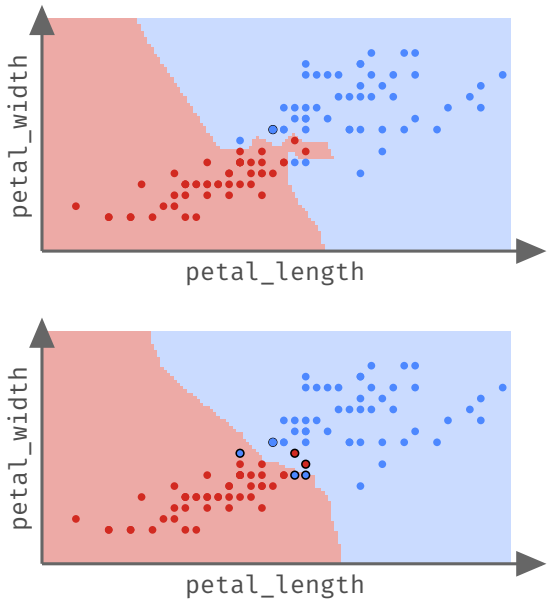
Overfitting - zu spezifisches Modell

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

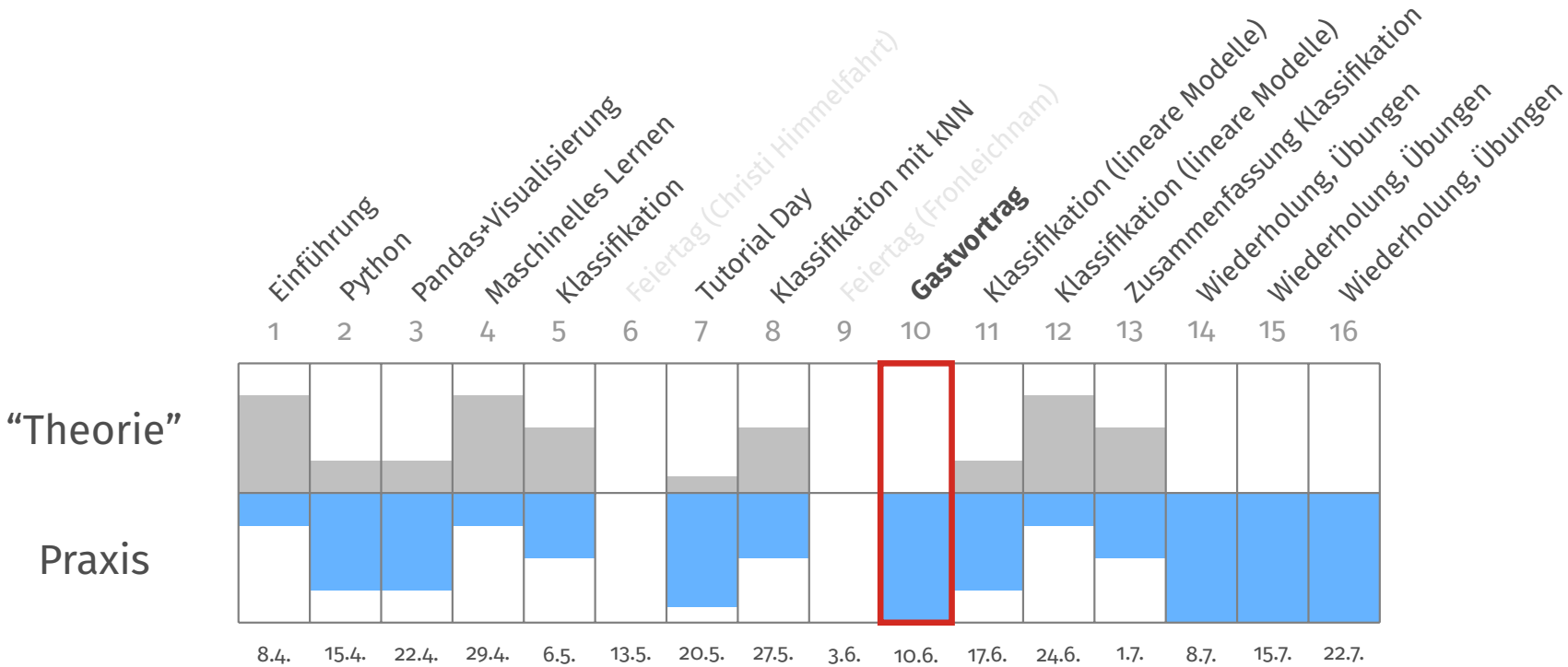
Tiefe bei Bäumen beschränken



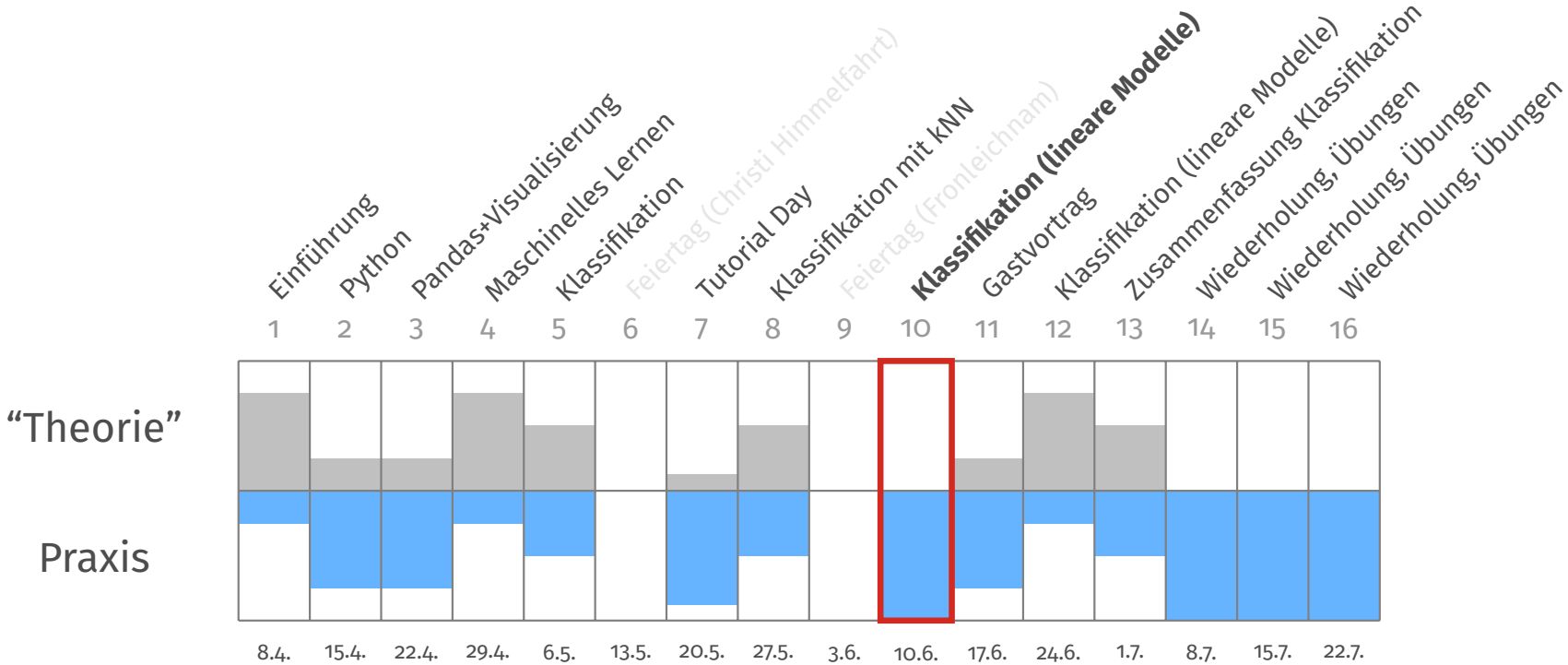
k bei k-NN erhöhen



Wo sind wir heute (Vorlesung 7) ?



Wo sind wir heute (Vorlesung 7) ?



Gastvortrag auf 17.6. verschoben

- Leider kurzfristige Verschiebung des Gastvortrags
- Bereitstellung von Foliensatz 7 (lineare Modelle)
- Heute kurzer **Überblick über lineare Modelle**

Gastvortrag auf 17.6. verschoben

- Leider kurzfristige Verschiebung des Gastvortrags
- Bereitstellung von Foliensatz 7 (lineare Modelle)
- Heute kurzer **Überblick über lineare Modelle**

Hausarbeit

- **Vorstellung Test-Hausarbeit**
- Planung zum Zeitraum der Hausarbeit

Gastvortrag auf 17.6. verschoben

- Leider kurzfristige Verschiebung des Gastvortrags
- Bereitstellung von Foliensatz 7 (lineare Modelle)
- Heute kurzer **Überblick über lineare Modelle**

Hausarbeit

- **Vorstellung Test-Hausarbeit**
- Planung zum Zeitraum der Hausarbeit

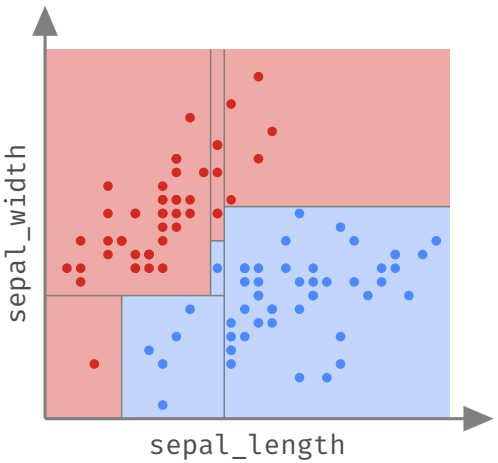
Heutige Übung:

- Tutorial Blatt 2 (Python Listen)
- Falls gewünscht: Test-Hausarbeit Aufgabe 1

Worum geht es im Foliensatz 7?

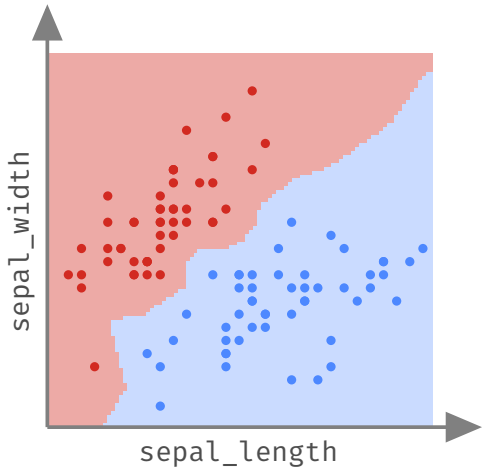
1. Daten im Vektorraum
2. Lineare Modelle
3. Hausarbeit (Test)

Entscheidungsbäume, nächste Nachbarn



Entscheidungsbaum

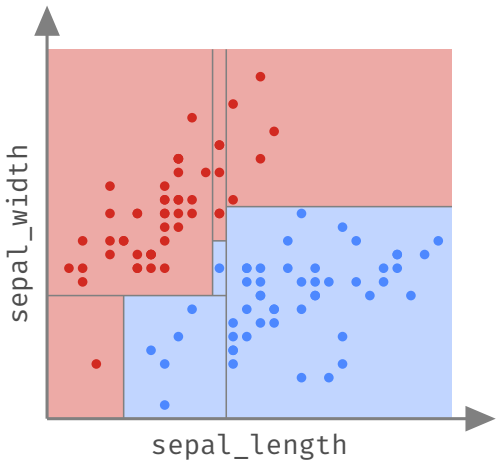
Trennung nach einzelnen Attributen, achsenparallel



k-nächste Nachbarn

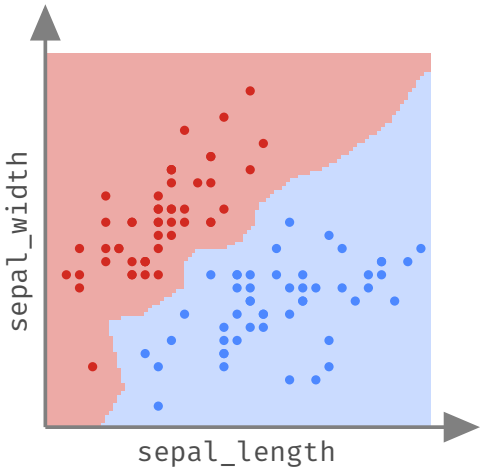
Trennung in Regionen, nach Distanz (Berechnung über alle Attribute)

Entscheidungsbäume, nächste Nachbarn und lineare Modelle



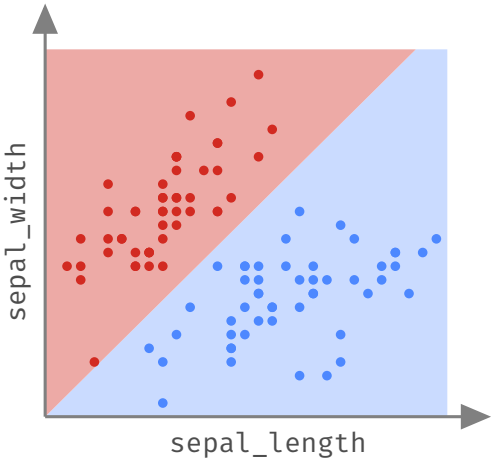
Entscheidungsbaum

Trennung nach einzelnen Attributen, achsenparallel



k-nächste Nachbarn

Trennung in Regionen, nach Distanz (Berechnung über alle Attribute)



Lineare Modelle

Trennung mit linearer Funktion über alle Attribute

Vektorraum: Iris-Daten im Vektorraum $\mathbb{R}^4$

sepal_length	sepal_width	petal_length	petal_width
4.700	3.200	1.300	0.200
6	2.200	4	1
4.600	3.100	1.500	0.200
7.600	3	6.600	2.100
6.300	2.900	5.600	1.800
5.400	3.900	1.700	0.400

$$\mathbf{x}_3 = \begin{pmatrix} 4.6 \\ 3.1 \\ 1.5 \\ 0.2 \end{pmatrix}$$

Vektor-Darstellung der
Zeile 3 aus dem Datensatz

Vektorraum: Iris-Daten im Vektorraum $\mathbb{R}^2$

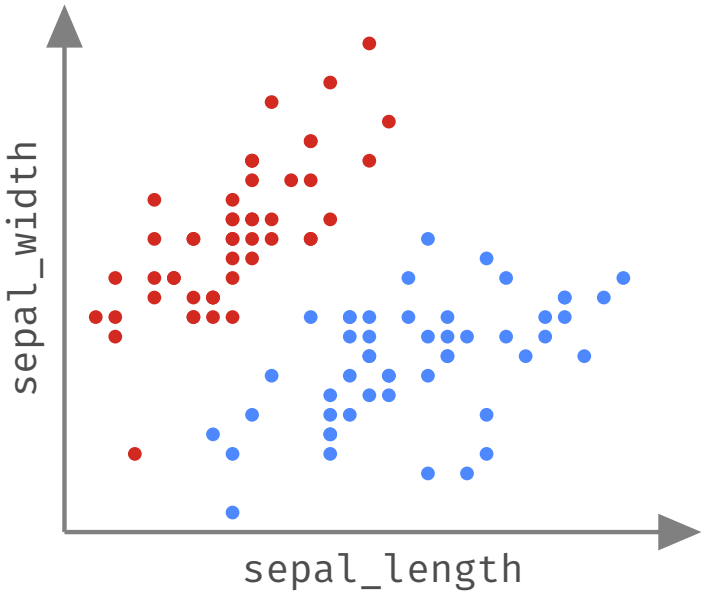
sepal_length	sepal_width
4.700	3.200
6	2.200
4.600	3.100
7.600	3
6.300	2.900
5.400	3.900

$$\mathbf{x}_3 = \begin{pmatrix} 4.6 \\ 3.1 \end{pmatrix}$$

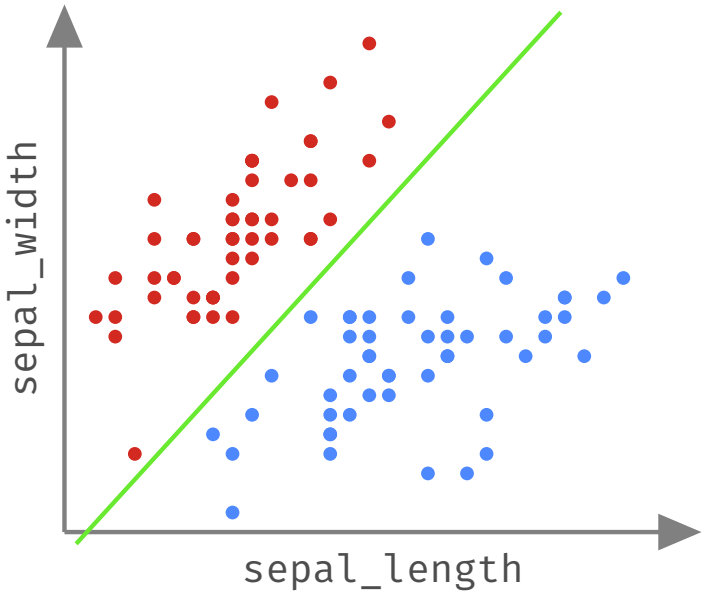
Vektor-Darstellung der
Zeile 3 aus dem Datensatz

Betrachten wir vereinfacht nur 2 Attribute!

Plot der beiden Attribute der Iris-Daten:

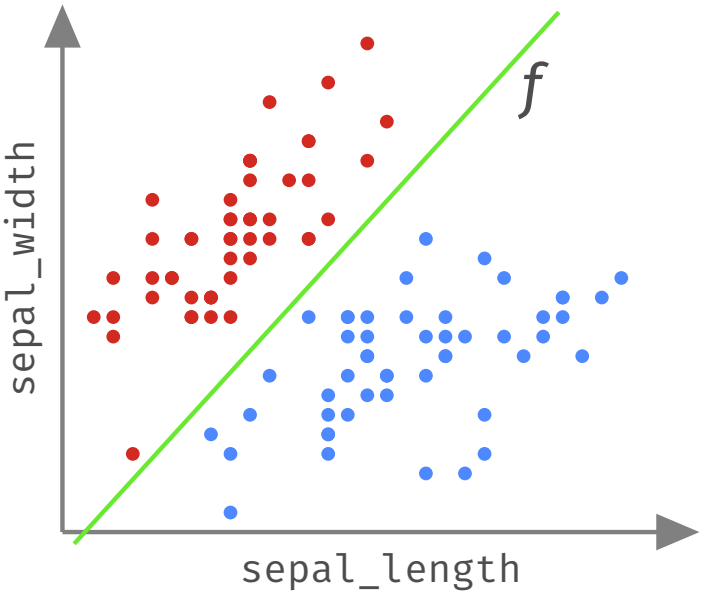


Plot der beiden Attribute der Iris-Daten:



Intuitiv lassen sich die beiden Klassen trennen.

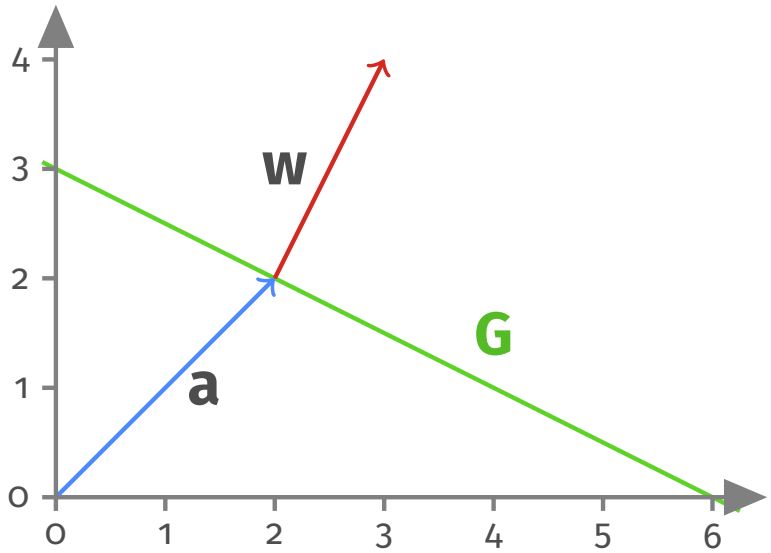
Idee: Daten mit einer Geraden trennen



Geradengleichung (Schule) im 2-dimensionalen Raum ($\mathbb{R}^2$):

$$f(x) = b \cdot x + c$$

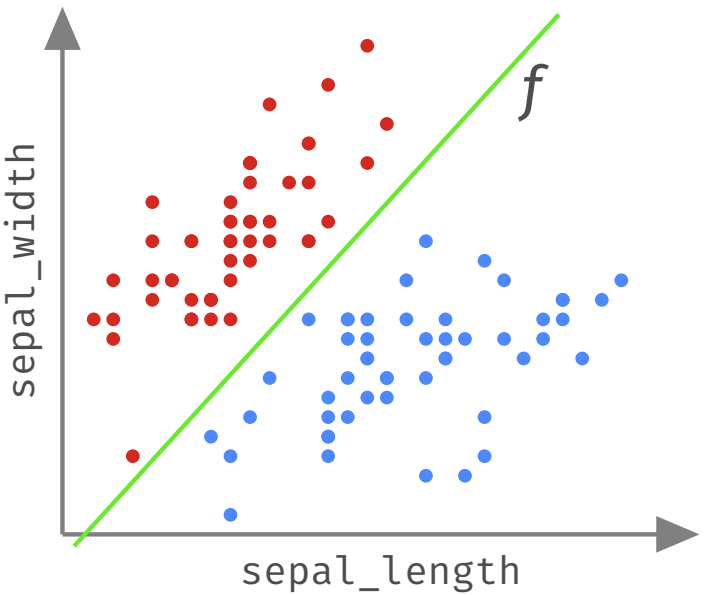
Beispiel: Gerade im $\mathbb{R}^2$



Gerade ist definiert durch Stützvektor $\vec{a} = \binom{2}{2}$ und Normalenvektor $\vec{w} = \binom{1}{2}$

$$\mathbf{G} = \left\{(x, y) \in \mathbb{R}^2 \mid 1 \cdot x + 2 \cdot y = 6\right\}$$

Beispiel: Gerade im $\mathbb{R}^2$



Wir müssen also nur die richtigen **a** und **w** finden!

Allgemeine Form für Hyperebenen im $\mathbb{R}^d$

Hyperebene H definiert durch

$$H = \left\{ \mathbf{x} \in \mathbb{R}^d \mid \mathbf{w}^T\mathbf{x} + b = 0 \right\}$$

Allgemeine Form für Hyperebenen im $\mathbb{R}^d$

Hyperebene H definiert durch

$$H = \left\{ \mathbf{x} \in \mathbb{R}^d \mid \mathbf{w}^T\mathbf{x} + b = 0 \right\}$$

Parametrisierung von f durch $\mathbf{x}$ und b:

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$$

Modell-Training ist damit die Auswahl von w und b!

Foliensatz 7 (zum weiteren Selbststudium)

- Geht auf Vektoren und Vektorräume ein
- Stellt einfaches Verfahren für lineares Modell vor
- Skizziert die Idee der Stützvektor-Methode (SVM)

Foliensatz 7 (zum weiteren Selbststudium)

- Geht auf Vektoren und Vektorräume ein
- Stellt einfaches Verfahren für lineares Modell vor
- Skizziert die Idee der Stützvektor-Methode (SVM)

Schauen Sie sich den Foliensatz an!

Diskussion in der nächsten Woche (nach dem Gastvortrag)

Hausarbeit (Test)

Die Hausarbeit

- Datensatz zu (fiktivem) Anwendungsfall
- Wird als Jupyter-Notebook bearbeitet
- Besteht aus 3 Aufgaben

Die Hausarbeit

- Datensatz zu (fiktivem) Anwendungsfall
- Wird als Jupyter-Notebook bearbeitet
- Besteht aus 3 Aufgaben

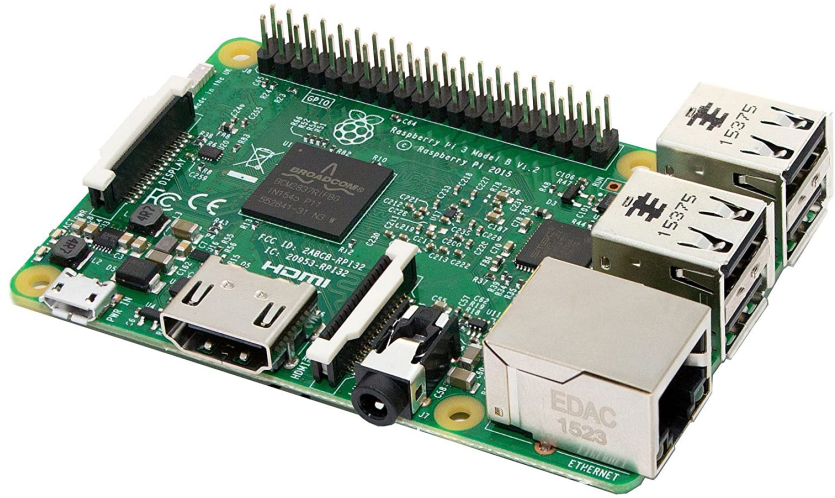
Die Arten der Aufgaben:

1. Python-Funktionen schreiben
2. Daten lesen + erkunden (Statistiken $\rightarrow$ Pandas)
3. Einfaches Modell für Vorhersage berechnen


tonieboxⓇ

Verschiedene Bastelprojekte

- Python-Tools für RFID-Karten
- Open-Source MP3-Player Software
- Phoniebox: `http://phoniebox.de/`



Wie ist das Hör-Verhalten der Kinder?

- Welche Titel/Themen sind beliebt?
- Gibt es Zielgruppen mit bestimmten Vorlieben?

Wie ist das Hör-Verhalten der Kinder?

- Welche Titel/Themen sind beliebt?
- Gibt es Zielgruppen mit bestimmten Vorlieben?

Andere machen das auch:

Vorstellung:
Test-Hausarbeit zu Kinder-Musikboxen

Terminfindung Hausarbeit

- 1 Woche Bearbeitungszeit für die Hausarbeit
- Klausurphase W: 10.7. - 31.7.
- Vorschlag 0: 5.7. bis 12.7. (Mo bis Mo)
- Vorschlag 1: 23.7. bis 30.7. (Fr bis Fr)
- Vorschlag 2: 30.7. bis 6.8. (Fr bis Fr)
- Diskussion!

Terminfindung Hausarbeit

- 1 Woche Bearbeitungszeit für die Hausarbeit
- Klausurphase W: 10.7. - 31.7.
- Vorschlag 0: 5.7. bis 12.7. (Mo bis Mo)
- Vorschlag 1: 23.7. bis 30.7. (Fr bis Fr)
- Vorschlag 2: 30.7. bis 6.8. (Fr bis Fr)
- Diskussion!

Vorschlag 2 nur mit verzögerter Korrektur möglich.