Data Science 1

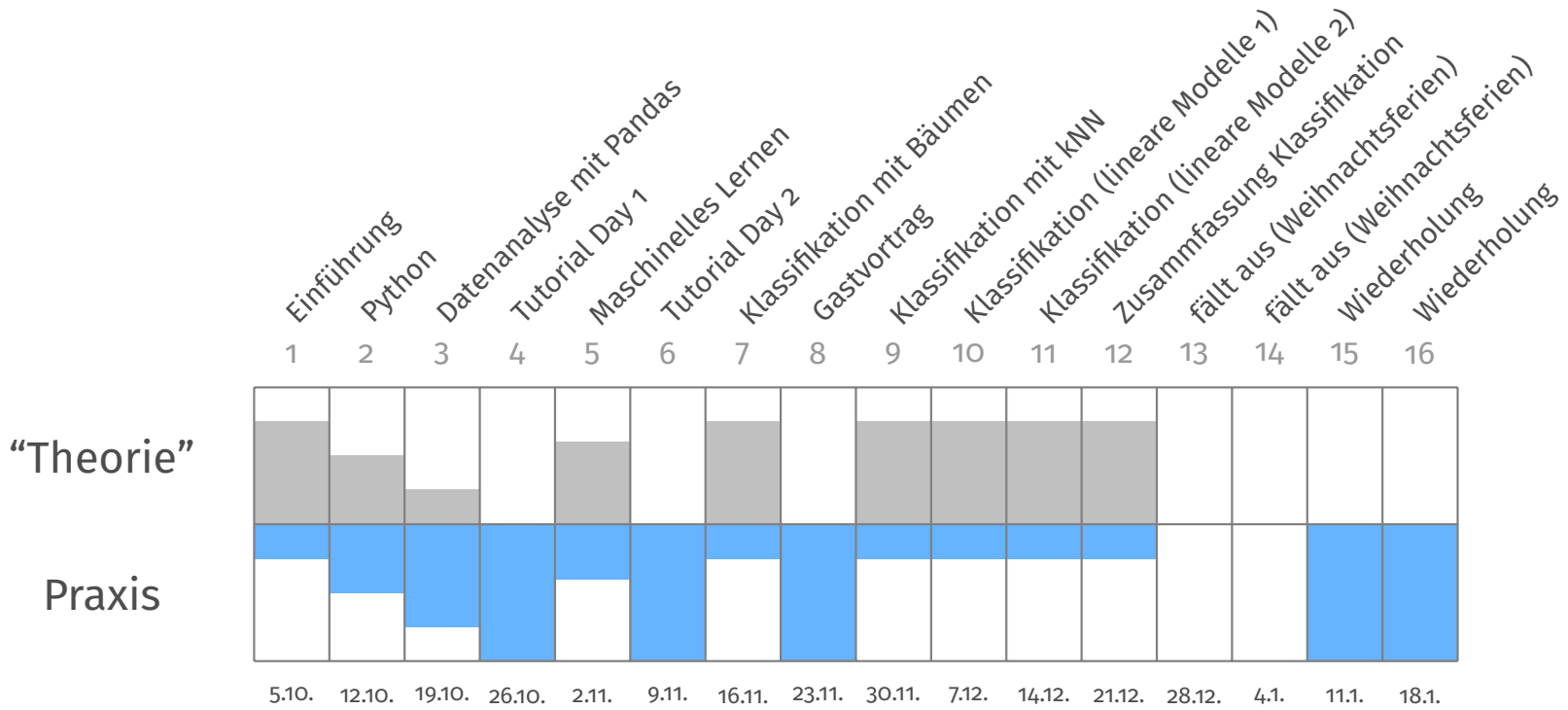
Vorlesung 10 – Offene Fragerunde

Prof. Dr. Christian Bockermann

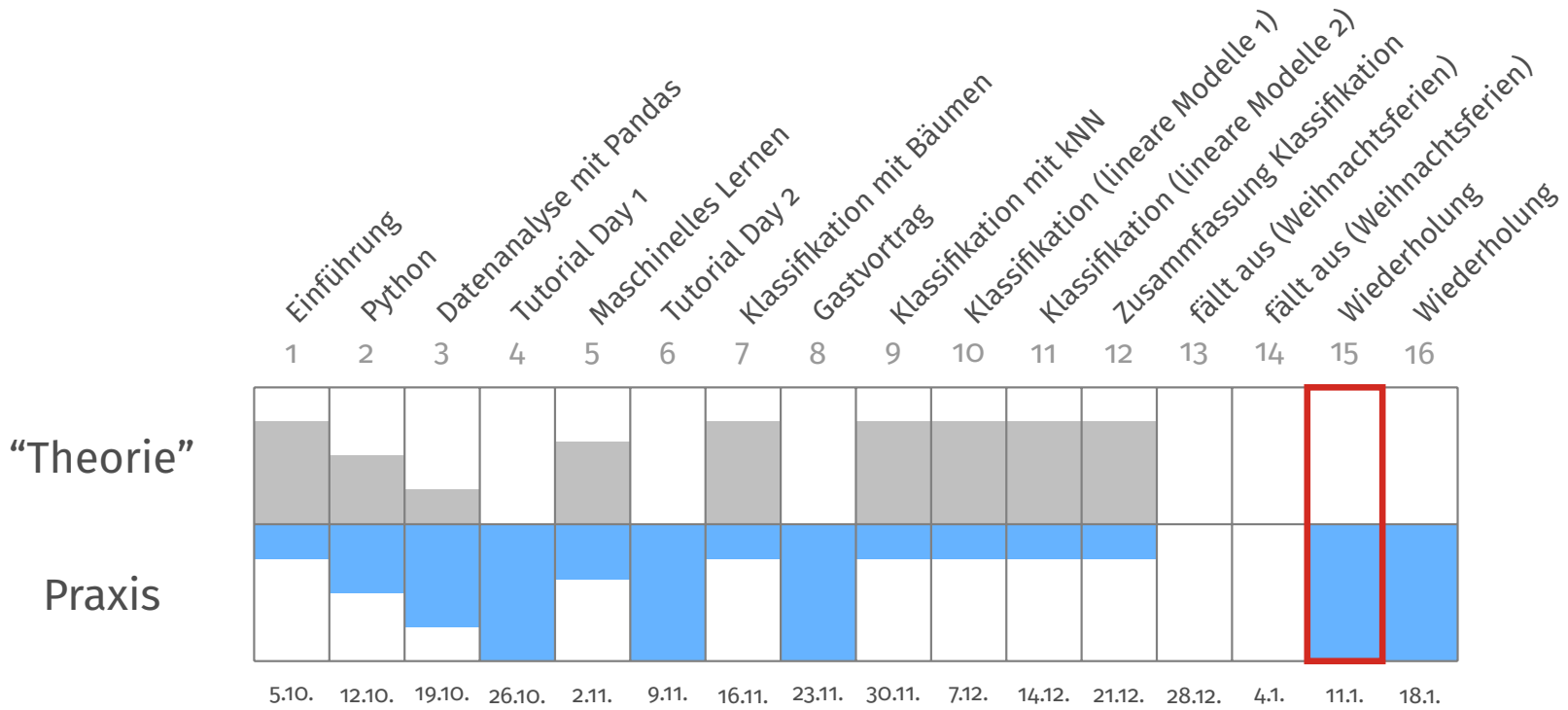
Hochschule Bochum

Wintersemester 2021/2022

Wo sind wir heute?



Wo sind wir heute?



Prüfungsleistung - Hausarbeit

- Prüfungsleistung ist Hausarbeit
- Aufgabenstellung am 9.2.2022 um **8:00 Uhr** im BBB
- Bearbeitung bis 18.2.2022
  (Abgabe: PDF des Notebooks per Mail an mich)

- Gruppenarbeit mit max. 3 Personen möglich
- Selbstorganisation der Gruppen, bitte bis 9.2. verbindlich mitteilen

Lehre-Evaluation

https://befragung.hs-bochum.de/evasys/online.php?p=GCMMY

Bitte bis spätestens heute (11.1.2022) ausfüllen (freiwillig)!

Wiederholung, Fragen

Warum brauchen wir die Normalisierung?

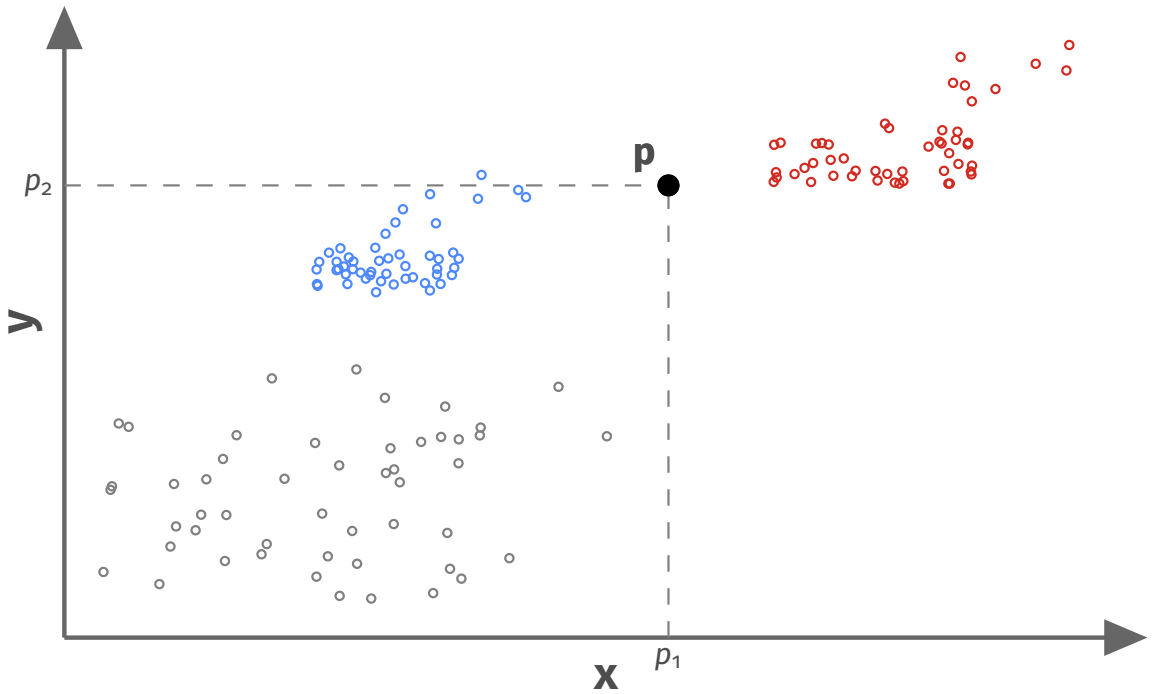
Beispiel: Klassifikation von Bällen

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: Ball-Arten

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

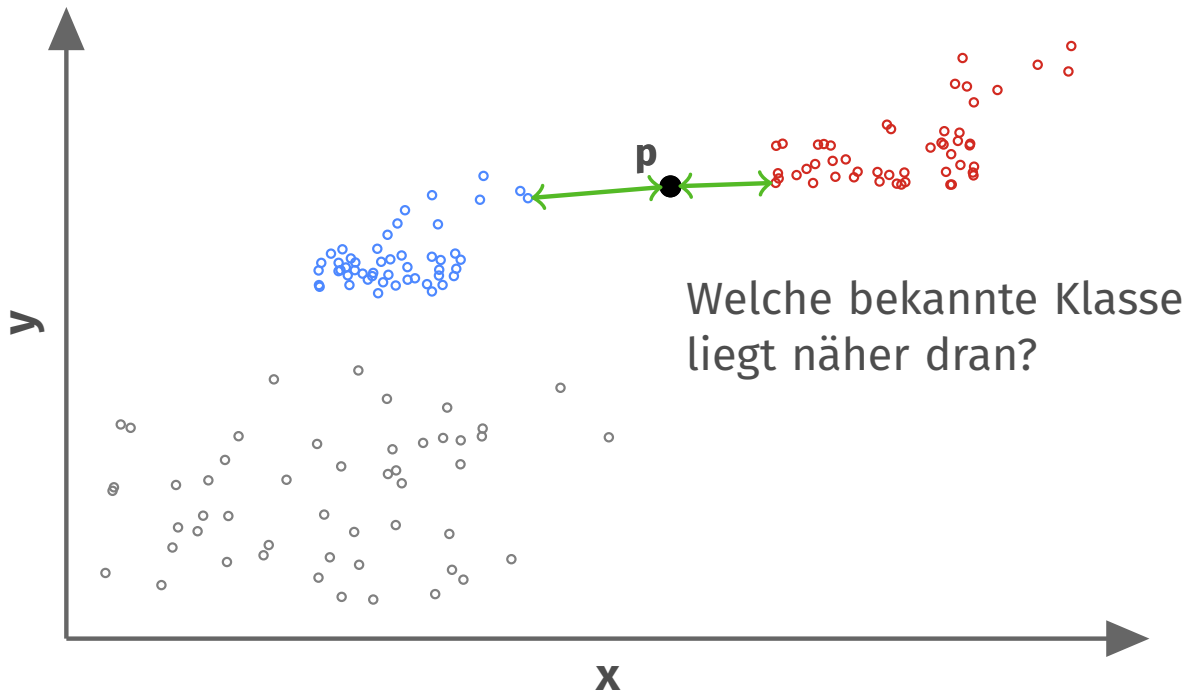


2-dimensionaler Raum: Jeder Punkt **p** besteht aus 2 Koordinaten:

$$\mathbf{p} = (p_1, p_2)$$

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!

Distanzen funktionieren nur auf metrischen, skalierten Variablen

Datensätze bisher:

- Iris-Daten, Attribute waren Maße in Zentimetern
- Ball-Daten, Attribute in Gramm und Zentimetern

Frage: Was bedeutet $dist(p, q) = 10$ bei den Ball-Daten?

$$\sqrt{\underbrace{(p_1 - q_1)}_{\textit{Umfang}}^2 + \underbrace{(p_2 - q_2)}_{\textit{Gewicht}}^2} = 10$$

0 cm	10 g	Gleicher Umfang, 10g schwerer
10 cm	0 g	10cm größer, gleiches Gewicht

Distanzen funktionieren nur auf metrischen, skalierten Variablen

Datensätze bisher:

- Iris-Daten, Attribute waren Maße in Zentimetern
- Ball-Daten, Attribute in Gramm und Zentimetern

Frage: Was bedeutet $dist(p, q) = 10$ bei den Ball-Daten?

$$\sqrt{\underbrace{(p_1 - q_1)}_{Umfang}^2 + \underbrace{(p_2 - q_2)}_{Gewicht}^2} = 10$$

0 cm	10 g	Gleicher Umfang, 10g schwerer
10 cm	0 g	10cm größer, gleiches Gewicht

Wertebereich *Umfang*: 48,57 cm bis 83,97 cm

Wertebereich *Gewicht*: **315,64** g bis **686,33** g

Frage: Was bedeutet $dist(p, q) = 10$ bei den Ball-Daten?

$$\sqrt{\underbrace{(p_1 - q_1)}_{Umfang}^2 + \underbrace{(p_2 - q_2)}_{Gewicht}^2} = 10$$

0 cm	10 g	Gleicher Umfang, 10g schwerer
10 cm	0 g	10cm größer, gleiches Gewicht

Frage: Was bedeutet $dist(p, q) = 10$ bei den Ball-Daten?

$$\sqrt{\underbrace{(p_1 - q_1)}_{Umfang}^2 + \underbrace{(p_2 - q_2)}_{Gewicht}^2} = 10$$

0 cm	10 g	Gleicher Umfang, 10g schwerer
10 cm	0 g	10cm größer, gleiches Gewicht

Wertebereich *Umfang*: 48,57 cm bis 83,97 cm

Wertebereich *Gewicht*: **315,64** g bis **686,33** g

Frage: Was bedeutet $dist(p, q) = 10$ bei den Ball-Daten?

$$\sqrt{\underbrace{(p_1 - q_1)}_{\textit{Umfang}}^2 + \underbrace{(p_2 - q_2)}_{\textit{Gewicht}}^2} = 10$$

0 cm	10 g	Gleicher Umfang, 10g schwerer
10 cm	0 g	10cm größer, gleiches Gewicht

Wertebereich *Umfang*: 48,57 cm bis 83,97 cm **35.40**

Wertebereich *Gewicht*: **315,64** g bis **686,33** g **370.69**

Die Metrik behandelt beide Variablen gleich.
Das macht eventuell nicht immer so viel Sinn!

Bezug z.B. zur Wirtschaftsstatistik

Charakterisierung von Attributen/Merkmalen/Variablen durch

- Minimum, Maximum
- Mittelwert, Standardabweichung

In welcher Relation steht dist(p,q) = 10 zu Umfang/Gewicht?

Idee: Normalisierung der Attribute/Variablen

- Anpassung der Werte auf gleichen Wertebereich
- z.B. Skalierung jeder Spalte auf [0,1]

Min-Max-Normalisierung einer Variablen X

$$X' = \frac{X - min(X)}{max(X) - min(X)}$$

```
zaehler = df['Umfang'] - min(df['Umfang'])
nenner = max(df['Umfang']) - min(df['Umfang'])

df['Umfang'] = zaehler / nenner
```

Frage: Was ist mit der Verteilung der Attribute?

- Wertebereiche mit Min/Max auf [0,1] normalisiert
- Variablen haben aber ggf. unterschiedliche Mittelwerte/Std-Abweichung?

Z-Normalisierung einer Variablen X

$$X'' = \frac{X - \mu(X)}{\sigma(X)}$$

ergibt eine Variable X'' mit Mittelwert 0 und Standardabweichung etwa bei 1.