Data Science 2

Vorlesung 1 - Intro

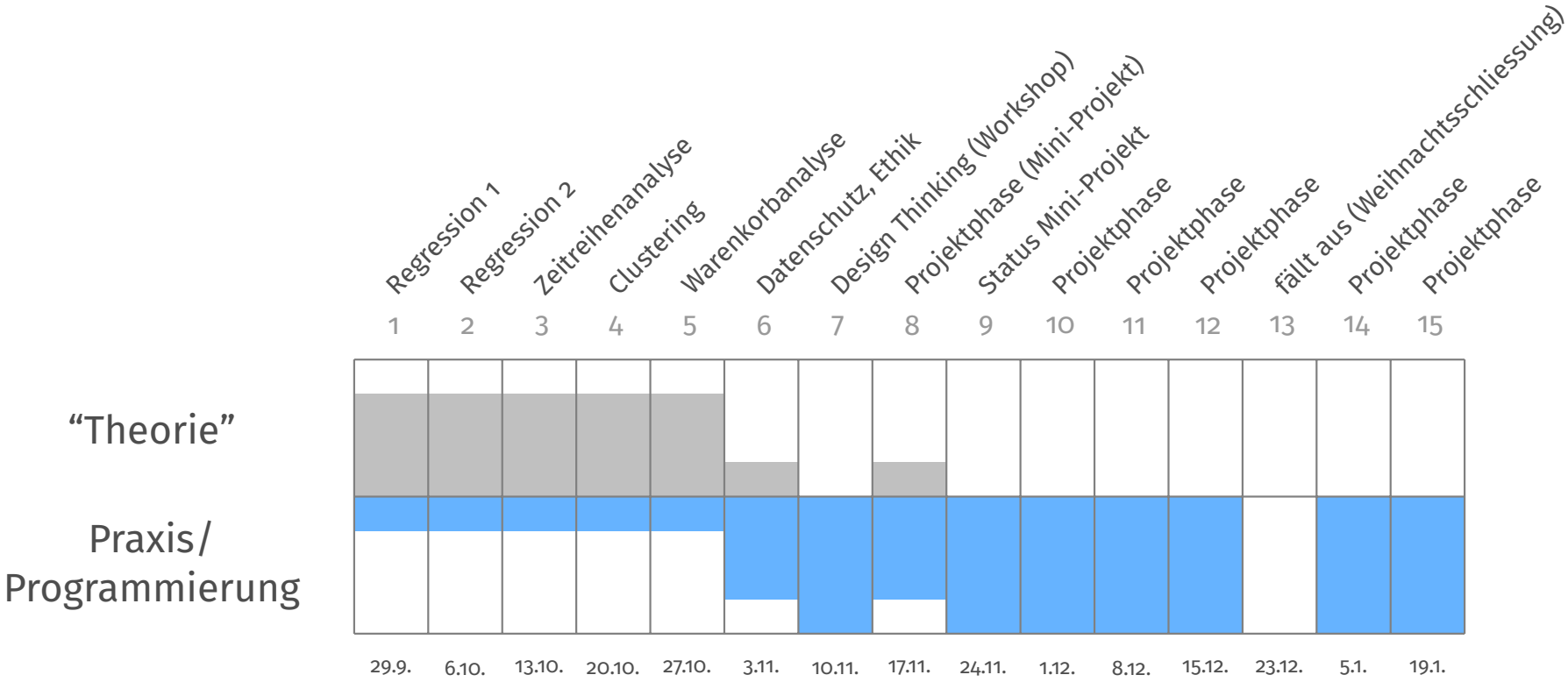
Prof. Dr. Christian Bockermann

Hochschule Bochum

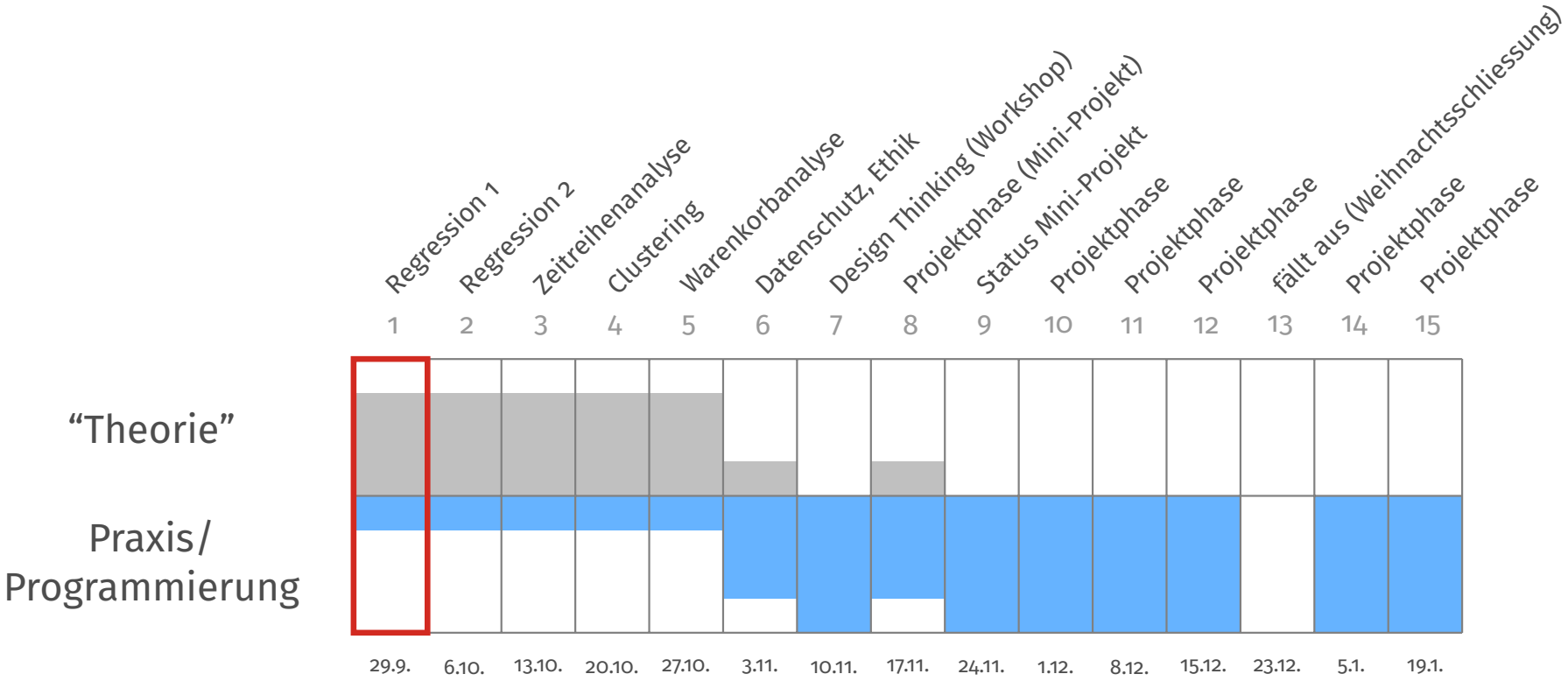
Wintersemester 2022 / 2023

Inhalt

Aufbau der Vorlesung



Aufbau der Vorlesung



Ziel des Kurses/der Projektphase

- Datengetriebenes Denken fördern
- Lern-Probleme in Anwendungen identifizieren
- Ideen für Data Science Lösungen entwickeln
- Exploration+Prototyping von Daten/Modellen

Projektphase

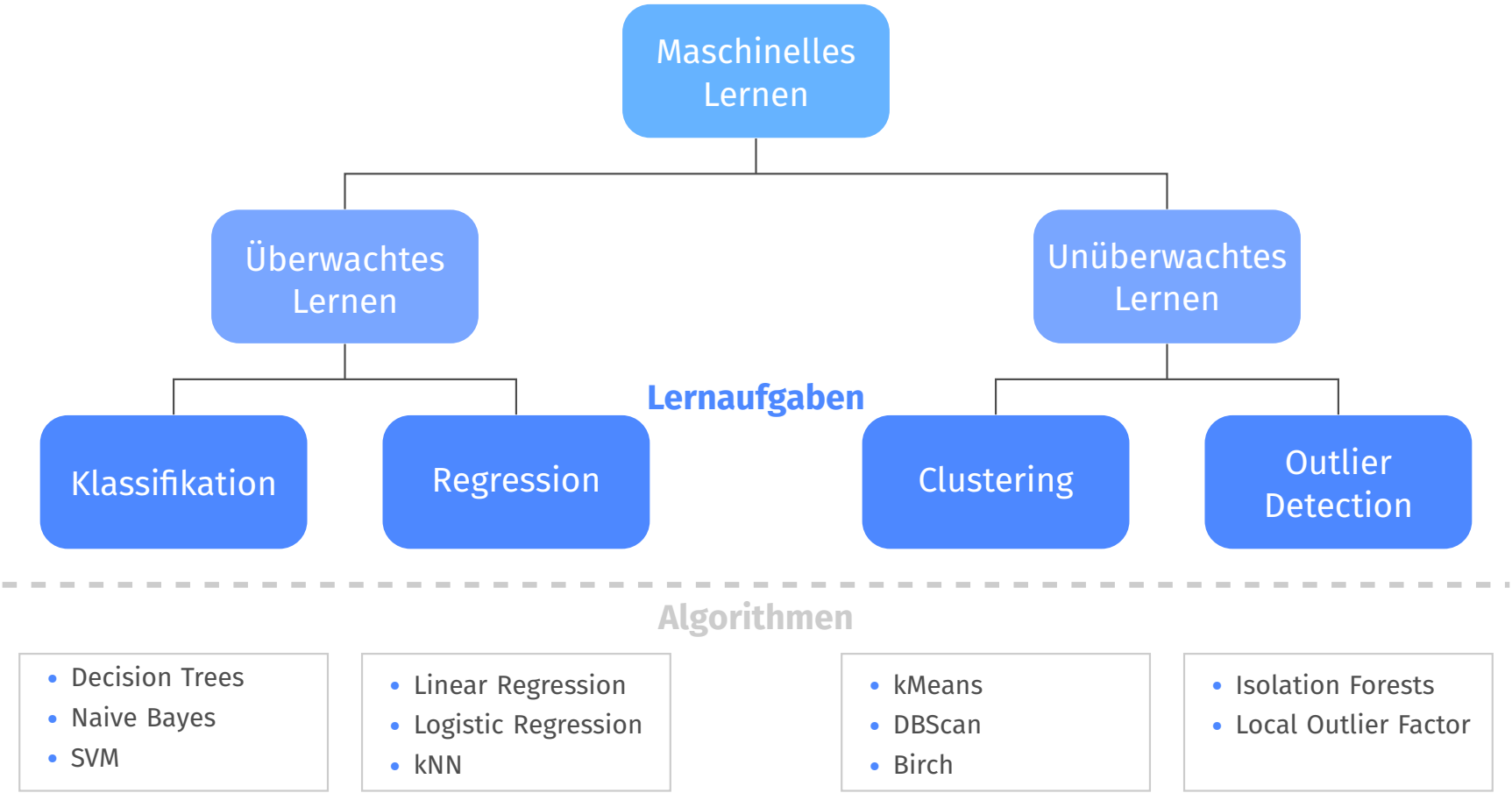
- Zusammenhängenden Anwendungsfall bearbeiten
- Unterschiedliche Aufgaben in gleichem Fallbeispiel
- Gruppenarbeit, gemischte Studiengänge (!?)

Prüfungsleistung

- Präsentation des Projektes (Vortrag)
- Hausarbeit über das Projekt ca. 8-10 Seiten
- Abgabe + Präsentation in Gruppen

Wiederholung – ML, Klassifikation

Maschinelles Lernen - Überblick



Lernaufgaben definieren Ein- und Ausgabe, sowie das Ziel der Modellierung, z.B.

> "Entscheide für einen Text **x** ob er zur Klasse *Spam* oder zur Klasse *KeinSpam* gehört."

Lernaufgaben definieren Ein- und Ausgabe, sowie das Ziel der Modellierung, z.B.

> "Entscheide für einen Text **x** ob er zur Klasse *Spam* oder zur Klasse *KeinSpam* gehört."

Eingabedaten werden typischerweise in einen Merkmalsraum $\mathcal{X}$ der Dimension d abgebildet

$$\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^d$$

Die Ausgabemenge $\mathcal{Y}$ kann eine Menge von Klassen oder eine reelle Zahl sein, z.B.

$$\mathcal{Y} = \{\text{Spam}, \text{KeinSpam}\}$$

Das Ziel besteht darin, eine Funktion (Modell) $f : \mathcal{X} \to \mathcal{Y}$ zu lernen, mit

$$f(\mathbf{x}) = \begin{cases} +1, & \text{falls } \mathbf{x} \text{ Spam Nachricht} \\ -1, & \text{sonst} \end{cases}$$

Das Ziel besteht darin, eine Funktion (Modell) $f : \mathcal{X} \to \mathcal{Y}$ zu lernen, mit

$$f(\mathbf{x}) = \begin{cases} +1, & \text{falls } \mathbf{x} \text{ Spam Nachricht} \\ -1, & \text{sonst} \end{cases}$$

Bei der binären Klassifikation wird häufig $\mathcal{Y} = \{-1, +1\}$ gewählt.

Lern-Algorithmen erwarten Daten häufig in Form einer Tabelle:

	d Merkmale				
ID	a_1	a_2	$\cdots$	a_d	y
1	0	0	$\cdots$	1	**-1**
2	0	1	$\cdots$	1	**+1**
3	1	0	$\cdots$	1	**-1**

Beispiel $\mathbf{x}_2 = (x_{a_1}, x_{a_2}, \ldots, x_{a_4}, y)$
$= (0, 1, \ldots, 1, \mathbf{+1})$

- Beispiele werden auch *examples* oder *instances* genannt
- Merkmale (engl. *features*) werden auch *attributes* oder *Variablen* (Statistik) bezeichnet

ML - Modell Training

a_1	a_2	$\cdots$	a_d	y
0	0	$\cdots$	1	-1
0	1	$\cdots$	1	+1
1	0	$\cdots$	1	-1

Trainingsdaten **X**, **y**

Algorithmus/
Optimierung

$$f : \mathcal{X} \to \mathcal{Y}$$

Modell

a_1	a_2	$\cdots$	a_d	y
1	1	$\cdots$	0	?

Neue Daten x',
y unbekannt

Vorhersage

$$\hat{y} = f(x')$$

Regression – Motivation/Beispiele

Hausarbeit: Kundendaten aus Online-Shop

id	discounts	sport	beauty	luxury	fancy	age	category
1	0.140	0.040	0.330	0.210	0.420	50	HIGH
2	0.360	0.140	0.310	0.210	0.340	19	LOW
3	0.270	0.060	0.310	0.210	0.420	73	HIGH
4	0.330	0.150	0.390	0.180	0.280	49	MID
5	0.360	0.160	0.330	0.200	0.310	53	MID

- Kategorisierung der Kunden in Klassen low/mid/high
- Grundlage: Umsatz des Kunden in letzten 2 Jahren
- Klassifikator zur Vorhersage

Hausarbeit: Kundendaten aus Online-Shop

id	discounts	sport	beauty	luxury	fancy	age	category
1	0.140	0.040	0.330	0.210	0.420	50	HIGH
2	0.360	0.140	0.310	0.210	0.340	19	LOW
3	0.270	0.060	0.310	0.210	0.420	73	HIGH
4	0.330	0.150	0.390	0.180	0.280	49	MID
5	0.360	0.160	0.330	0.200	0.310	53	MID

Hausarbeit: Kundendaten aus Online-Shop

id	discounts	sport	beauty	luxury	fancy	age	sales
1	0.140	0.040	0.330	0.210	0.420	50	6404
2	0.360	0.140	0.310	0.210	0.340	19	1200
3	0.270	0.060	0.310	0.210	0.420	73	3802
4	0.330	0.150	0.390	0.180	0.280	49	2098
5	0.360	0.160	0.330	0.200	0.310	53	3763

Was, wenn wir den Kundenumsatz direkt vorhersagen?

- Wert eines Kunden: *Customer Lifetime Value*
- Wieviel Umsatz wird mit Kunde erzeugt?

Weitere Beispiele

- Lieferzeit von Paketen
- Fahrtzeiten
- Umsatzprognosen
- und vieles mehr...