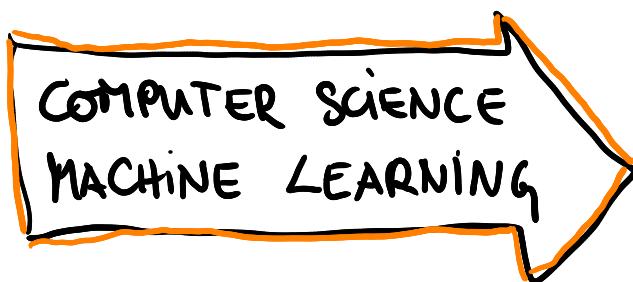
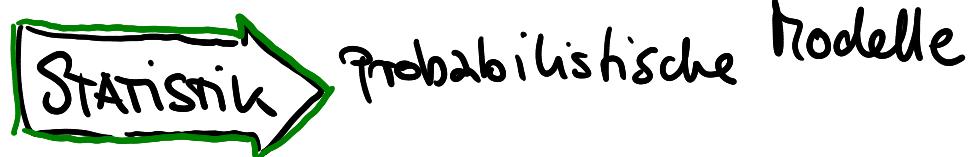


## 1.3) Konzepte & Grundlagen

### 1.3.1) Was ist ein "Modell"?

Die fundamentale Frage heißt:

Wie extrahieren wir optimal Information aus Daten und wie machen wir Vorhersagen basierend auf dieser Information?



vereinigt Methoden aus der Statistik, Mathematik, Informationswissenschaften, ...

→ eher algorithmisch als analytisch!

Grundlegende  
Vorstellung:

Wir beobachten Daten  $D = \{x_1, \dots, x_n\}$  und nehmen an, dass diese von einem darunterliegenden WÄHREN Modell  $M_{\text{true}}$  generiert wurden ...

→ Wir wollen die Daten → erklären  
→ Voraussagen treffen

→ wir machen Hypothesen in der Form  $M_1, M_2, \dots$  (vermutete Modelle)  
mit Parametern  $\Theta_1, \Theta_2, \dots$  (Bem.: es gibt auch Modelle ohne Parameter - mehr  
dazu später...)

⚠ Das wahre Modell  $M_{\text{true}}$  ist unbekannt!

Alles, was wir tun können ist Daten  $\mathcal{D} = \{x_1, x_2, \dots\}$  beobachten, die  
vom Modell abstammen!

Bemerkung: Die angenommenen Modelle  $M_1, M_2, \dots$  sind NIE perfekt.

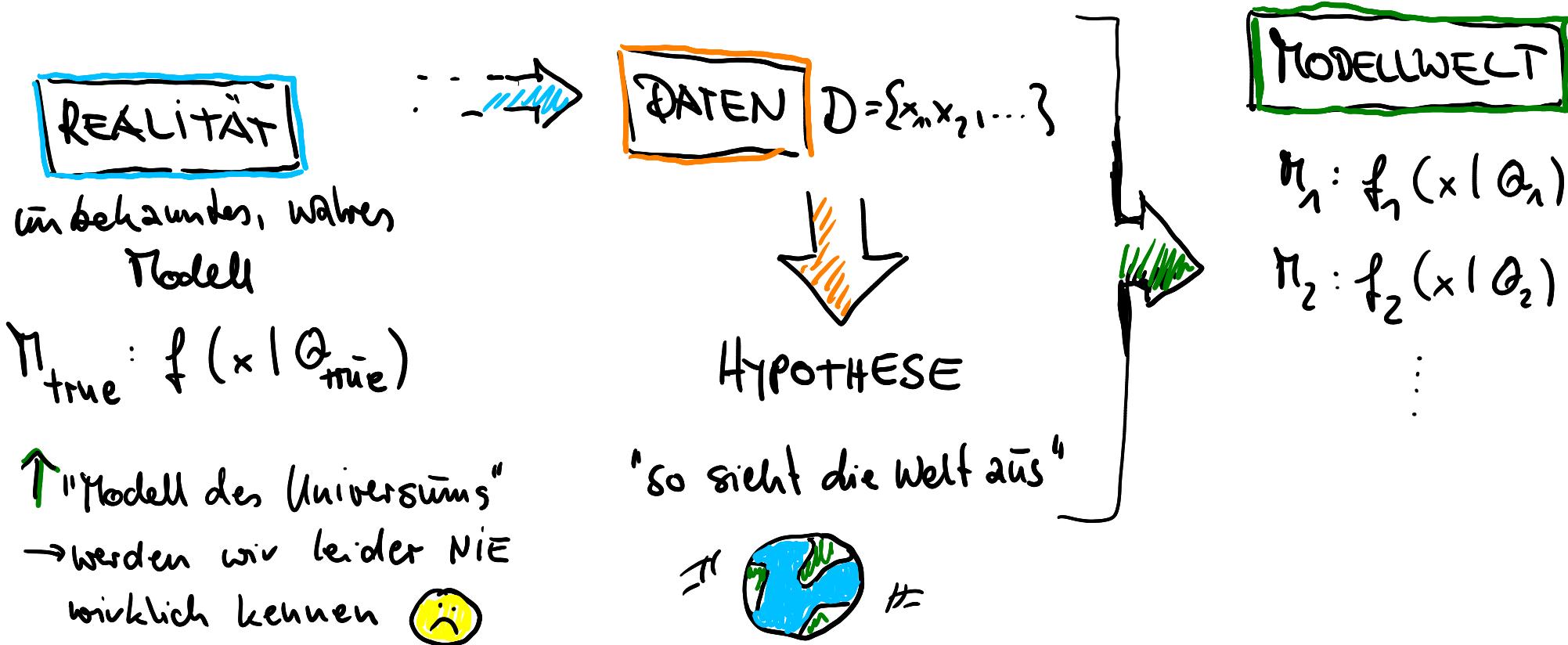
Es ist nicht mal sicher, dass  $M_{\text{true}}$  unter den Kandidaten ist 😊

Das ist nur in einer sehr idealisierten Situation der Fall.

## ANNAHME

Selbst ein nicht perfektes Modell liefert eine mögliche mathematische Approximation der Realität.

Es fängt charakteristische Eigenschaften des wahren Modells  $M_{\text{true}}$  ein und hilft uns, die Daten zu interpretieren.



Bemerkung: was bedeutet "Modell"?

Ausgangssituation: (supervised)

- Input Raum  $X \subset \mathbb{R}^n$ 
  - Punktmenge  $x \in X$
- Output Raum  $Y \subset \mathbb{R}^n$ 
  - Labels (metrisch, kategorisch)  $y \in Y$

(TRAININGS) DATEN

$$\mathcal{D}: (x_1, y_1), \dots, (x_n, y_n) \in X \times Y$$

ZIEL: Finde Transferfunktion  $f: X \rightarrow Y$   
die den Zusammenhang zwischen  $X$  und  $Y$   
optimal beschreibt

Def.: Eine Repräsentation von  $f$  (mathematisch, als Datenstruktur, ...)  
heißt Modell

Dimensionen von  $X$ : Inputs, Prädiktoren, Features, ...

zufälliger Fehler (ZV)

$Y$ : Targets, Responses, ...

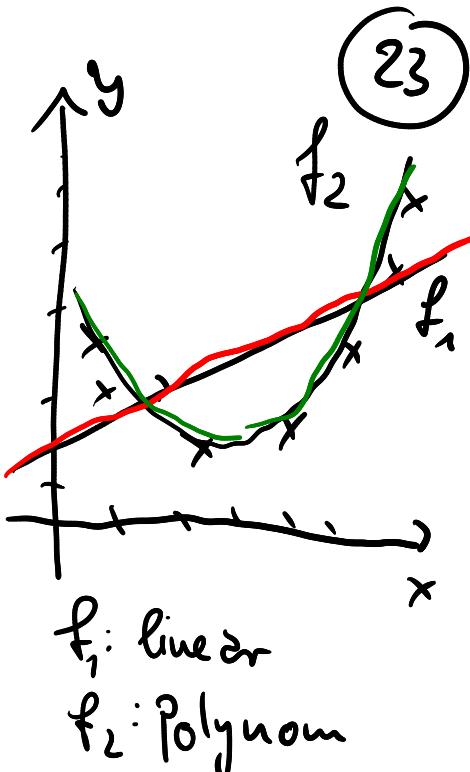
Allgemeine Form:

$$Y = f(X) + \varepsilon$$

$$\leftarrow E(\varepsilon) = 0$$

Die Transferfunktion  $f$  beinhaltet die Systematik der Zusammensetzungstruktur zwischen  $X$  und  $Y$ . Die Funktion ist zunächst unbekannt und wird aus den Daten geschätzt werden. Die Struktur von  $f$  ist eine a priori Hypothese!

(23)



Aus den Daten bekommen wir

$$\hat{f}(x) \leftarrow \text{Schätzung für } f$$

und damit auch eine "Prognose" - Schätzung für  $Y$

$$\hat{Y} = \hat{f}(X)$$

# ANFORDERUNGEN ANS MODELL

24

## INFERNZ (PROZESSVERSTÄNDNIS)

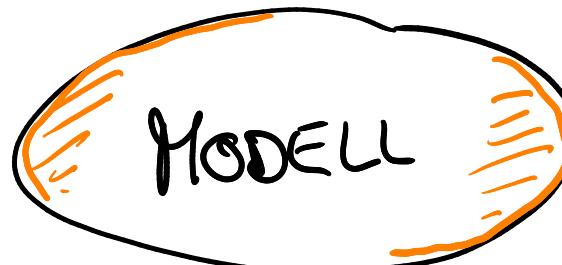
das Modell dient nicht  
mir zur (optimalen)  
Prognose von  $y$

Interpretierbarkeit ist relevant



Modell liefert Einsichten über den daten-  
erzeugenden Prozess

- welche Features sind relevant?
- funktionaler Zusammenhang?
- Reihung der Inputs ...



## PRÄDIKTION (VORHERSAGE)

$y$  ist unbekannt  
Prädiktion  $y = f(x)$   
so gut wie möglich!



- nur die Qualität der Vorhersage zählt!
- Modell darf nicht sich eine "Black-Box" sein

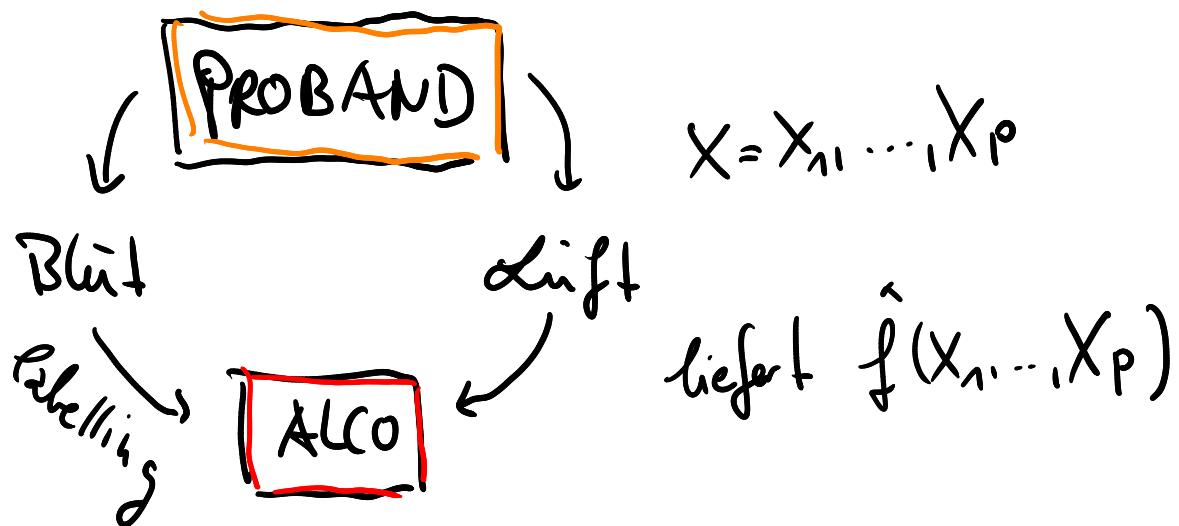
## Anwendungen: PROGNOSÉ

Beobachtungen für  $X$  sind leicht zu kriegen, für  $Y$  aber nicht ...

Bsp.: Alco-Test (Atemluft), char. Features  $X_1, \dots, X_p$

Die Atemluft ist leicht auszuwerten, jedenfalls einfacher als eine Blutprobe

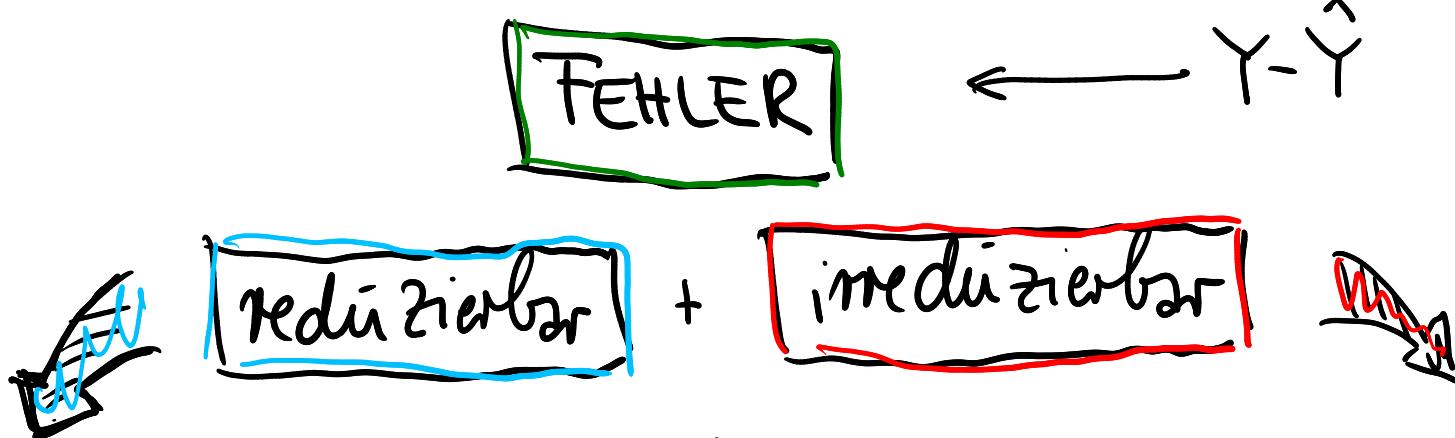
TRAINING



FRAGE

Wie genau ist  $\hat{f}(X) = \hat{Y}$  als Prognose?

... perfekt wirds ja nie 



$\hat{f}$  ist für einen Teil des Fehlers verantwortlich!

→ lässt sich verkleinern durch

- bessere Modellwahl
- Lern-Methodik
- mehr Daten

⋮

wegen  $Y = f(X) + \varepsilon$

trägt auch die Varianz von  $\varepsilon$  zum Fehler bei

- $\varepsilon$  setzt sich zusammen aus
  - unbeobachteten (wichtigen)  $X_i$
  - nicht beobachtbaren Einflüssen

⋮

Überlegung: Erwartungswert des Fehlers (Residuum s) (27)

$$E(Y - \hat{Y})^2 = E(f(X) + \varepsilon - \hat{f}(X))^2 = \dots$$

$$\begin{aligned} &= (f(X) - \hat{f}(X))^2 + \text{Var}(\varepsilon) \\ &\quad \downarrow \quad \downarrow \\ \text{auch als} &\quad \text{reduzierbarer} & \text{immedizierbarer} \\ \text{"Streuungszerlegung"} &\quad \text{Fehler} & \text{Fehler} \\ (\text{z.B. Regressionsanalyse}) & & \end{aligned}$$

bekannt



Stellt immer die Grenze  
für die Güte der Prognose  
dar (Abschätzung !!)

## Anwendungen: INFERENZ

Manchmal reicht eine ("gute") Prognose nicht aus.

Man will wissen, wie Änderungen der  $X_1, \dots, X_p$  die Zielgröße beeinflussen → ZUSCHAENHANGSSTRUKTUR

⇒  $\hat{f}$  kann keine reine "BLACK BOX" mehr sein!



- ① welche Features sind relevant?
- ② wie genau sieht die Beziehung zw.  $X_1, \dots, X_p$  und  $Y$  aus?
- ③ sind die Beziehungen linear \ nicht-linear?

Bemerkung: nicht jede Modellklasse kann beides gleich gut! Bsp.: NEURO

## 1.3.2) Daten & Aufgaben

29

$$X, Y \subset \mathbb{R}^n$$



$$\mathcal{D} = \{(x_i, y_i) \mid x_i \in X, y_i \in Y\}$$

Transferfunktion

$$f(X) = Y \quad (\text{unbekannt})$$

↑      ↑  
Feature Zielgröße (bekannt)  
↓      ↓  
supervised

X = Feature Space

Y = Target Space

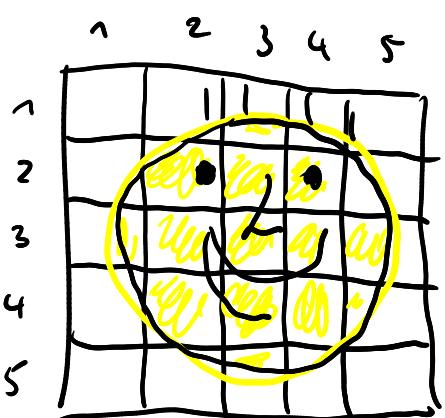
Bsp.:  $X \hat{=} \text{Temperatur in } {}^\circ\text{C}$ , Werte in  $\mathbb{I} \subset \mathbb{R}$

$X = (x_1, x_2) \hat{=} (\text{Temp}, \text{wind speed})$ , Werte  $(x_1, x_2) \in \mathbb{R}^2$

$X \hat{=} \text{Bild mit } 128 \times 128 \text{ Pixeln}$ , Werte  $\in \mathbb{R}^{16384}$

(Intensität Graustufen von 0 bis 1, oder

RGB  $\rightarrow$  dann wirds noch mehr 😊)



$$\{x_{11}, x_{12}, \dots, x_{51}, \dots, x_{55}\}$$

Die Struktur des Target Space  $Y$  definiert die Art der Aufgabe!

①  $Y$  ist endlich, diskret

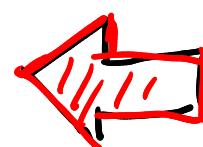
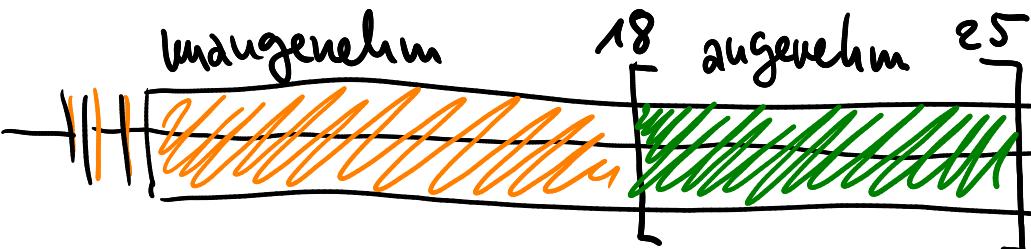
$$Y = \{C_1, \dots, C_k\} \quad k \in \mathbb{N}$$

$C_k$  heißt Klasse (Kategorie)

→ für  $k=2$ : binäre Klassifikation

Bsp. 1)  $X \hat{=} \text{Temperatur}$

$$Y \hat{=} \{\text{zugelassen}, \text{n zugelassen}\}$$



## Klassifikationsproblem

$$f: X \rightarrow \{C_1, \dots, C_k\} \quad k \in \mathbb{N}$$

$$f: \mathbb{R} \rightarrow \{\text{zugelassen}, \text{n zugelassen}\}$$

$$x \mapsto \begin{cases} \text{zugelassen} & \text{für } x \in [18, 25] \\ \text{n zugelassen} & \text{sonst} \end{cases}$$

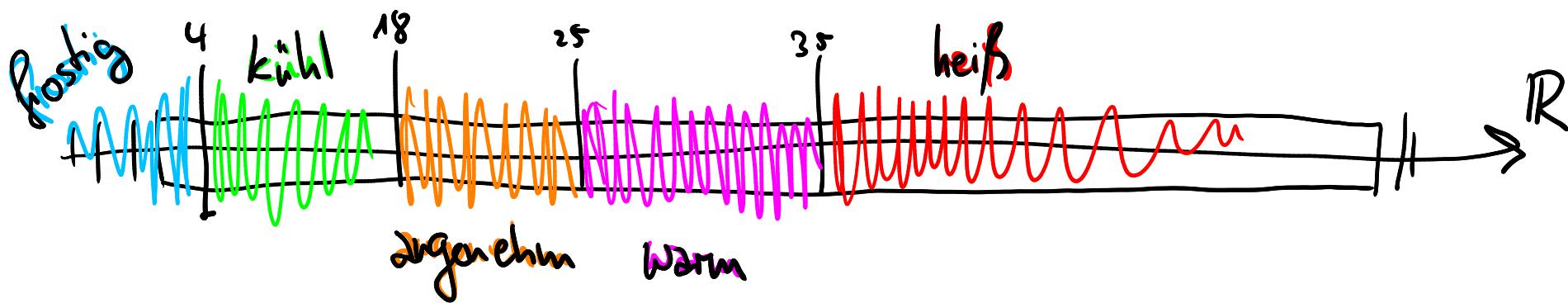


Bsp. 2) $X \triangleq \text{Temperatur}$  $Y \triangleq \{\text{frostig, k\"uhl, angenehm, warm, hei\ss}\}$ 

mehr als 2 Klassen  
"multiclass" Klassifikation

 $f: \mathbb{R} \rightarrow \{\text{frostig, \dots, hei\ss}\}$ 

$$x \mapsto \begin{cases} \text{frostig} & x \in (-\infty, 4) \\ \text{k\"uhl} & \text{Intervall } x \in [4; 18) \\ \vdots & \vdots \\ \text{he\ss} & x \in [35; \infty) \end{cases}$$



(2)

$Y$  unendlich, kontinuierlich

$$Y = I \subset \mathbb{R}$$

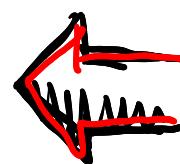
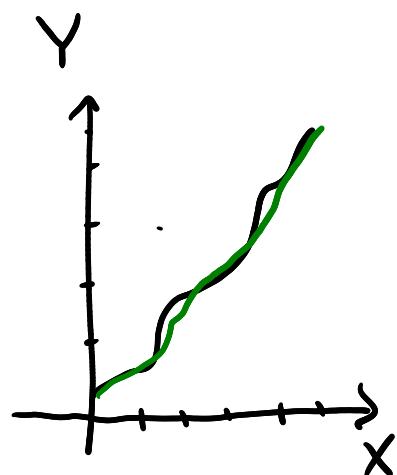
Bsp. 1)

$X \hat{=} \text{Niederschlag}$

$Y \hat{=} \text{Flussspeigel}$

$$f: \mathbb{R} \rightarrow \mathbb{R}$$

$$x \mapsto f(x) = y$$



## REGRESSIONS - PROBLEM

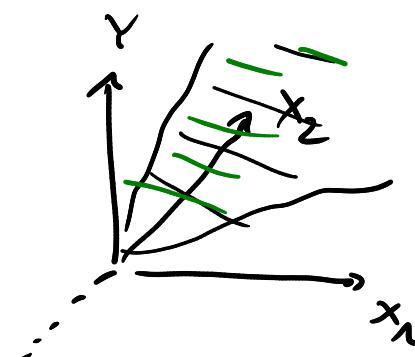
~ Wir wollen quantitative Aussagen treffen!

Bsp. 2)  $X = (X_1, X_2) \hat{=} (\text{Gewicht, PS Auto})_{\text{Auto}}$

$Y \hat{=} \text{Verbrauch}$

$$f: \mathbb{R}^2 \rightarrow \mathbb{R}$$

$$(x_1, x_2) \mapsto f(x_1, x_2) = Y$$



### 1.3.3) Los gehts ...

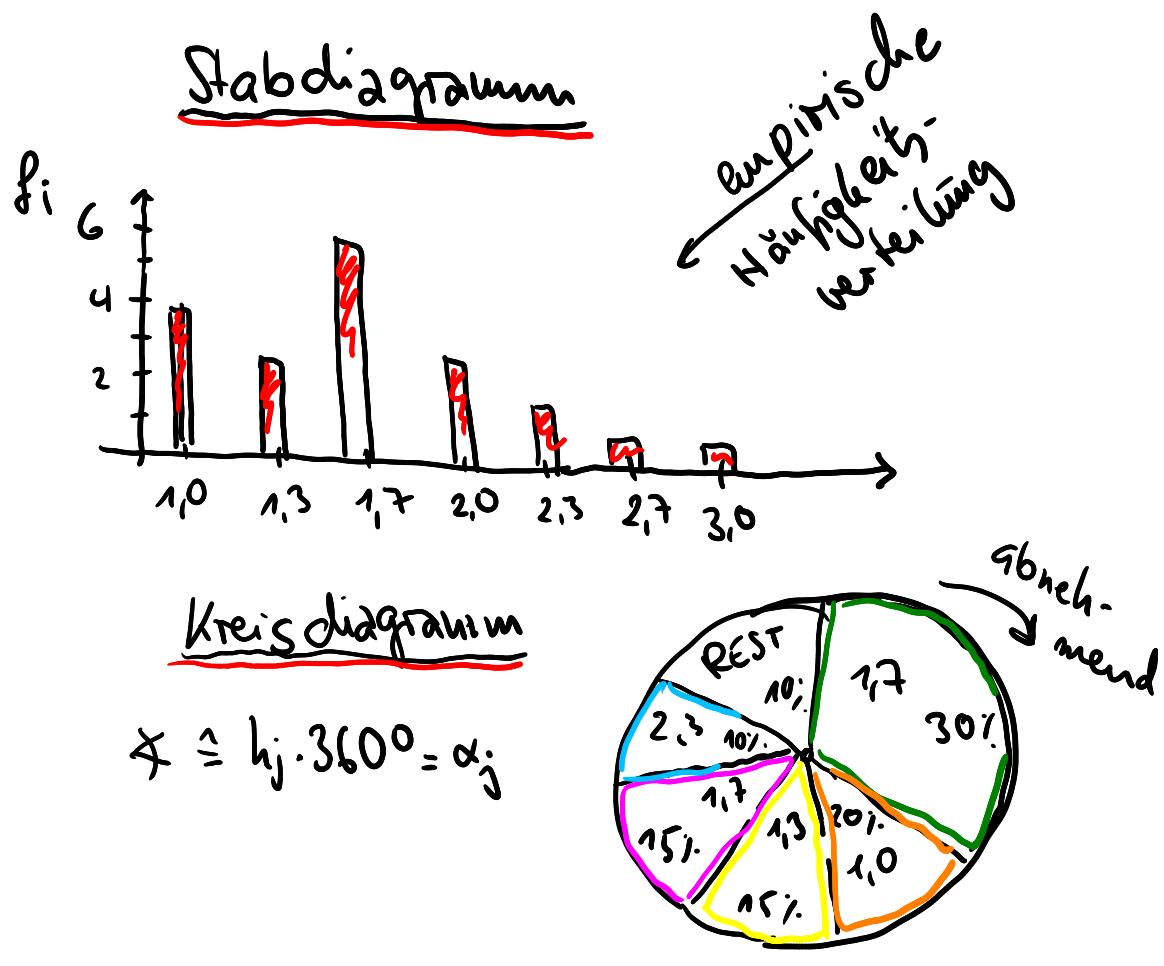
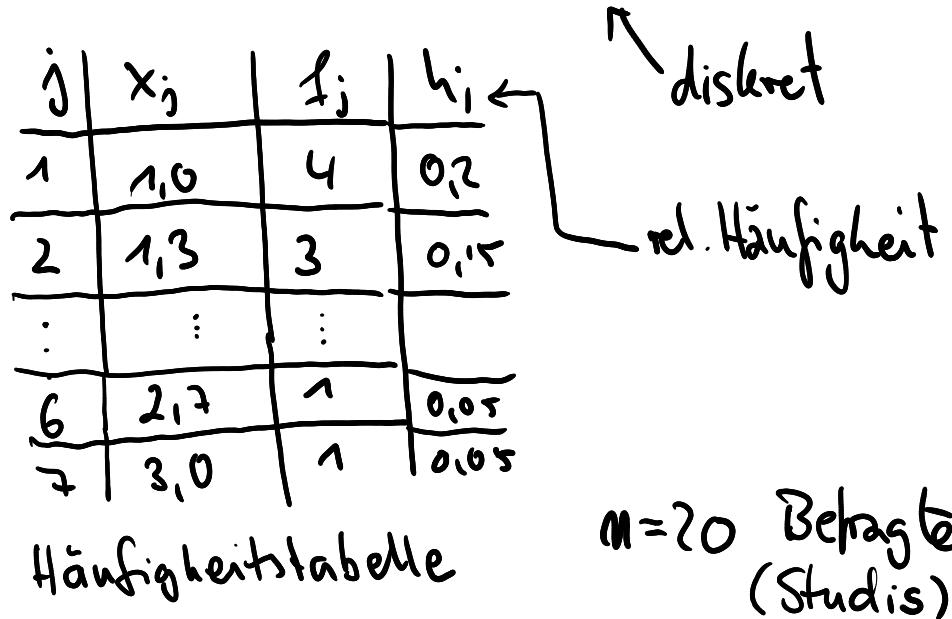
Zu Beginn eines ML-Projektes sollte man sich auf jeden Fall einen Überblick über die Daten verschaffen.

#### ① Visualisierung

##### • Ein einzelnes Feature:

###### Bsp. 1) Abi-Note

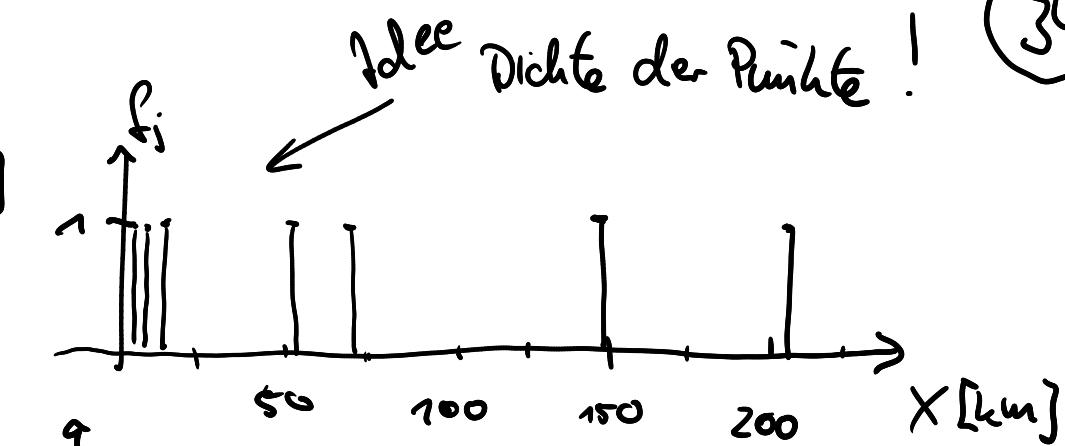
$X \hat{=} \text{Abi-Note } (1,0; 1,3; \dots; 3,0)$



## 3sp. 2) Entfernung Heimatort $\leftrightarrow$ OT H

+ Befragte	$j$	$x_j$ [km]
1	1	20
2	2	210
3	3	57
4	4	70
5	5	150
6	6	10
7	7	15

$X \hat{=} \text{Entf. [km]}$



falle Dichte der Punkte!

bringt nichts bei stetigen Größen  
→ jeder Wert taucht nur 1x auf

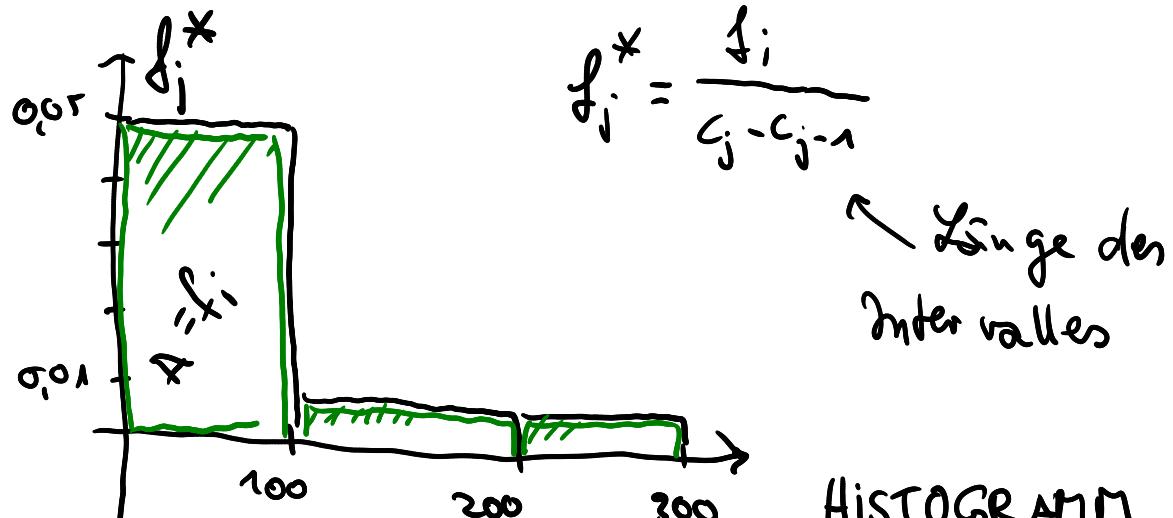
V4

$j$	$[c_{j-1}; c_j)$	$f_i$	$h_i$
1	$[0; 100)$	5	0,72
2	$[100; 200)$	1	0,14
3	$[200; 300)$	1	0,14

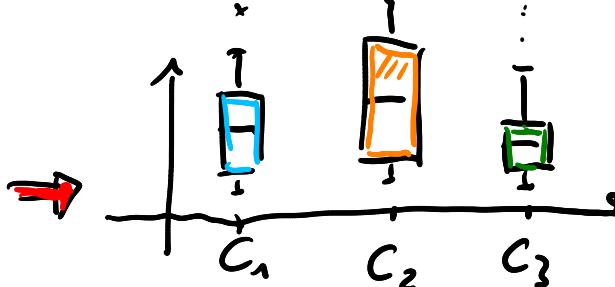
Gruppierung

Weitere Möglichkeiten z.B.

Box Plot



HISTOGRAMM



← Eigenschaften der Verteilung

## Zeitreihe $x(t)$

ist eigentlich eine 2D-Zufallsvariable

$$X = (X, T)$$

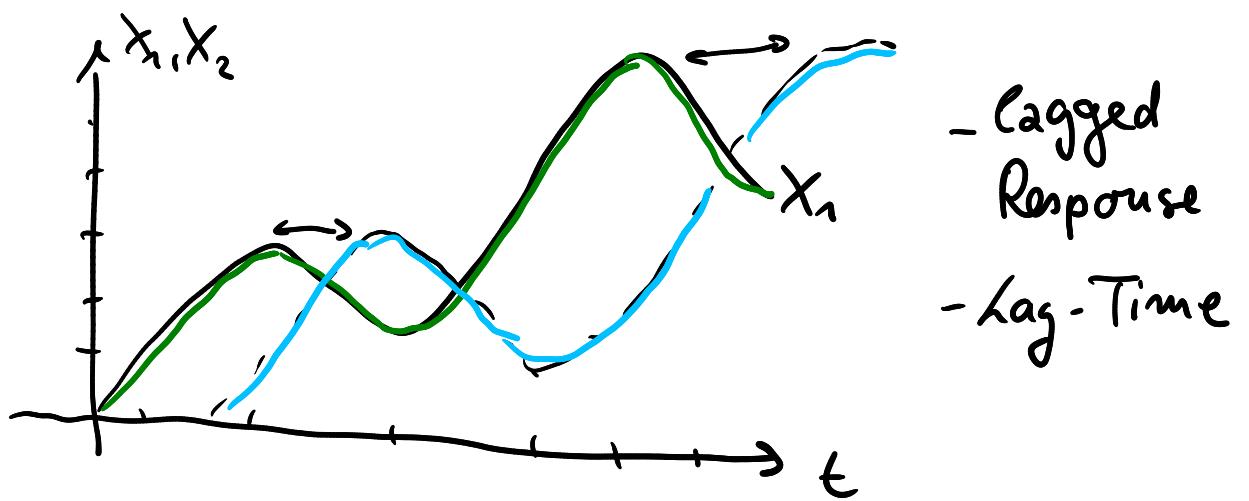
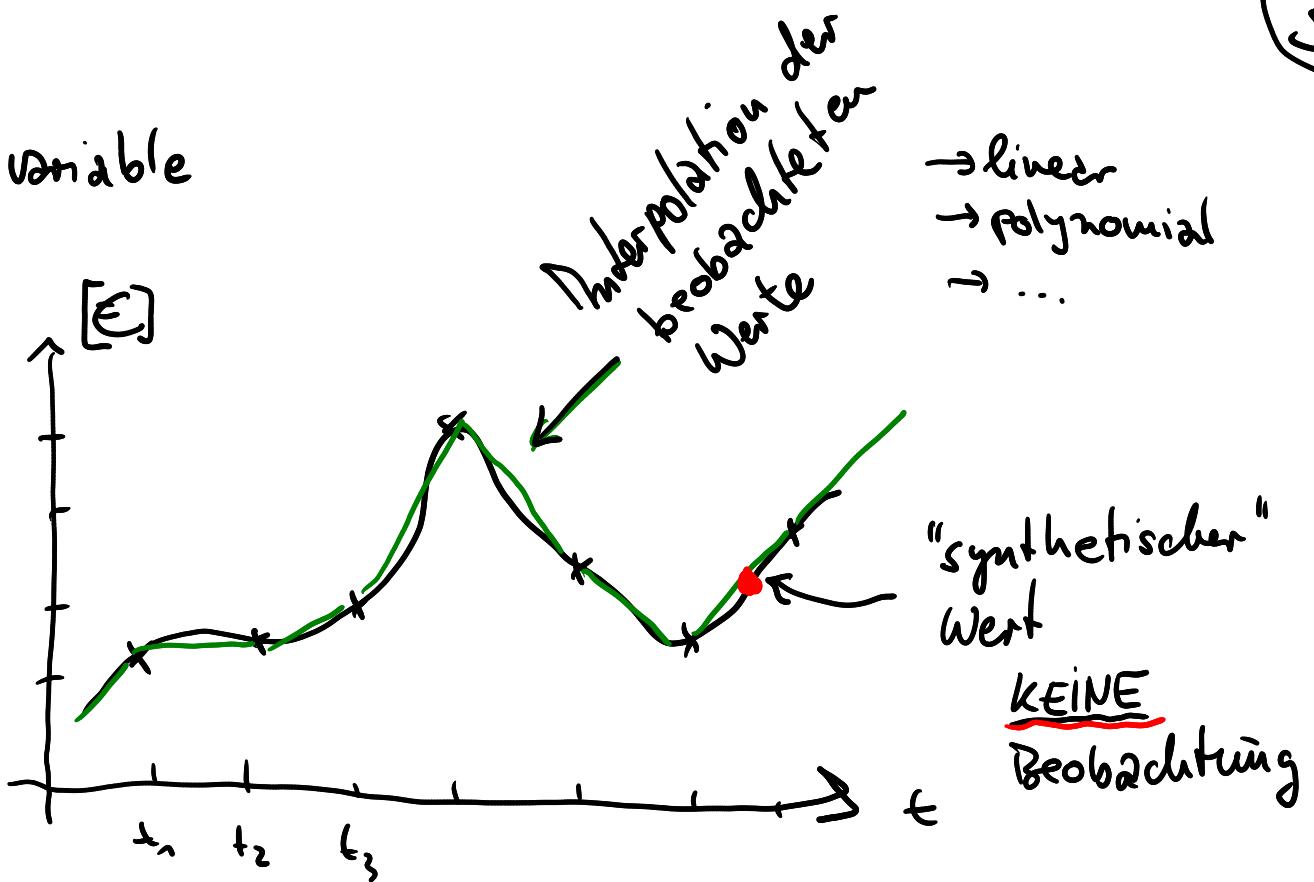
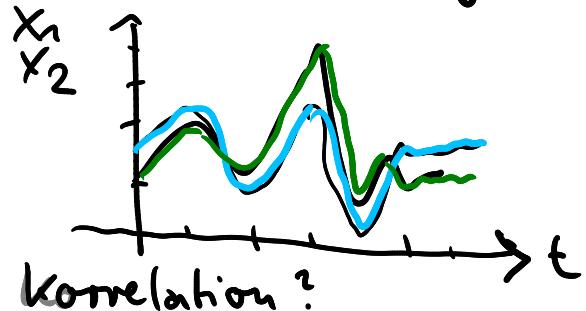
$$\approx (x_1, t_1), (x_2, t_2), \dots$$

Bsp.: Aktienkurs

$X \approx \text{Kurs [€]}$

$T \approx \text{Zeit}$

Vergleich von Features auf Zusammenhang

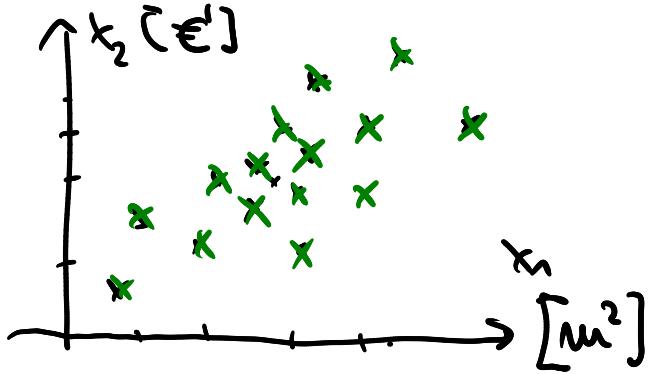


## • 2 (3) Features

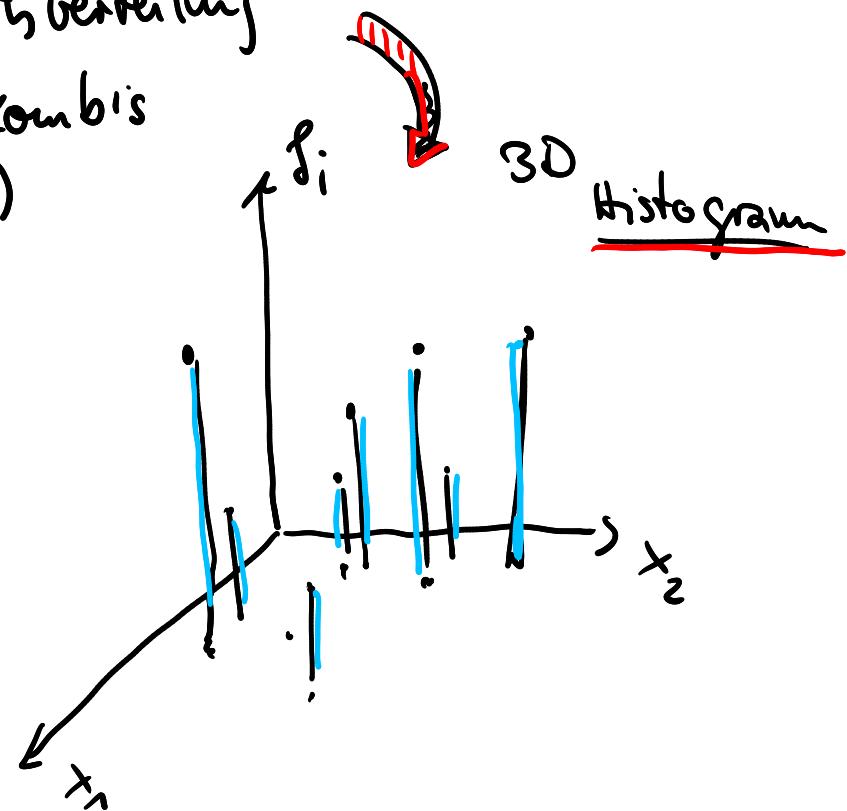
$$X = (x_1, x_2)$$

Bsp.: Große Wohnung  $\leftrightarrow$  Miete

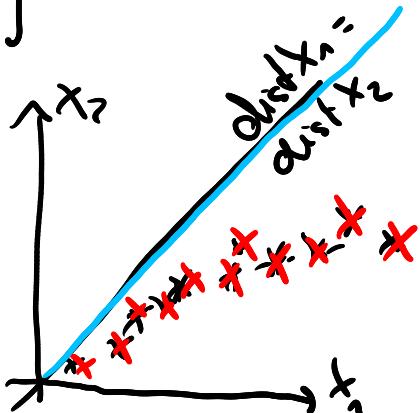
### ⇒ SCATTER Plot



auch Häufigkeitsverteilung  
beobachteter Kombis  
(entl. gruppiert)



### Probability Plot Q-Q Plot



Quantile von  $x_1$  gegen

Quantile von  $x_2$  auftragen

⇒ Vergleich, ob Verteilungen ähnlich sind

~ theoretische, gemeinsame  
W-Dichte

## ② Data Preparation

37

→ beobachtete Daten sind in der Realität nie ideal!

Bevor man ein ML-Modell trainieren kann müssen sie vorbereitet werden.

Dazu gehört:

+ Selektion

Auswahl relevanter Features, Zeiträume etc...

+ Formatierung

Wählte geeignete Datenformat  
bereinige die Daten  
(Ausreißer, NaNs,...)  
→ Sampling (Zeitreihen)

+ Transformation

bringe alle Features in vergleichbaren Range  
 $[0; 1]$   
 $[-1; 1]$   
Aggregation (Zeitreihen)  
Upscaling

### ③ TRAIN-TEST Split

In der Praxis weiß man vorher nicht, mit welchem Modell oder Verfahren man für einen gegebenen Datensatz am besten abschneidet.

Frische Erfahrungen aus anderen Projekten sind nur bedingt übertragbar.

⇒ VALIDIERUNG der Modelle mit geeigneten Metriken!

wichtig dafür: Aufteilung der verfügbaren Daten

