Einführung in maschinellem Lernen (Teil 1)

Carl Herrmann
Health Data Science Unit
Medizinische Fakultät & BioQuant
www.hdsu.org

carl.herrmann@bioquant.uni-heiderlberg.de

Medizinische Fakultät Heidelberg

Inhalt

Medizinische Fakultät Heidelberg

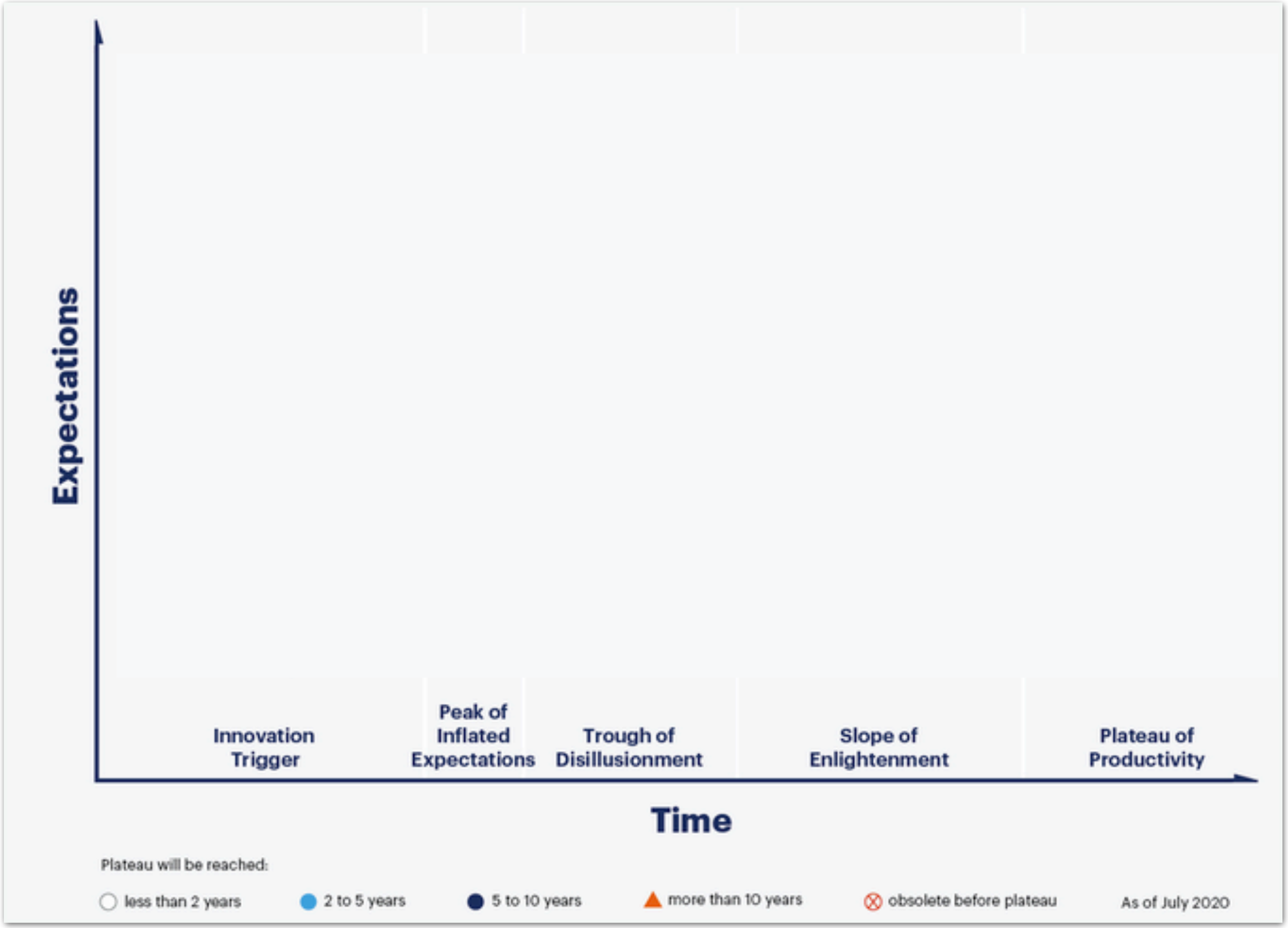
1. Was ist maschinelles Lernen?
2. Datentypen
3. Grundkonzepte in ML
4. Modelle lernen
5. Anwendung : Regressionsmodelle (lineare Regression)
6. Anwendung : Klassifizierungsmodelle (k-NN, Entscheidungsbäume)
7. Zusammenfassung

Medizinische Fakultät Heidelberg

1. Was ist maschinelles Lernen (ML)?

Trends in künstlicher Intelligenz

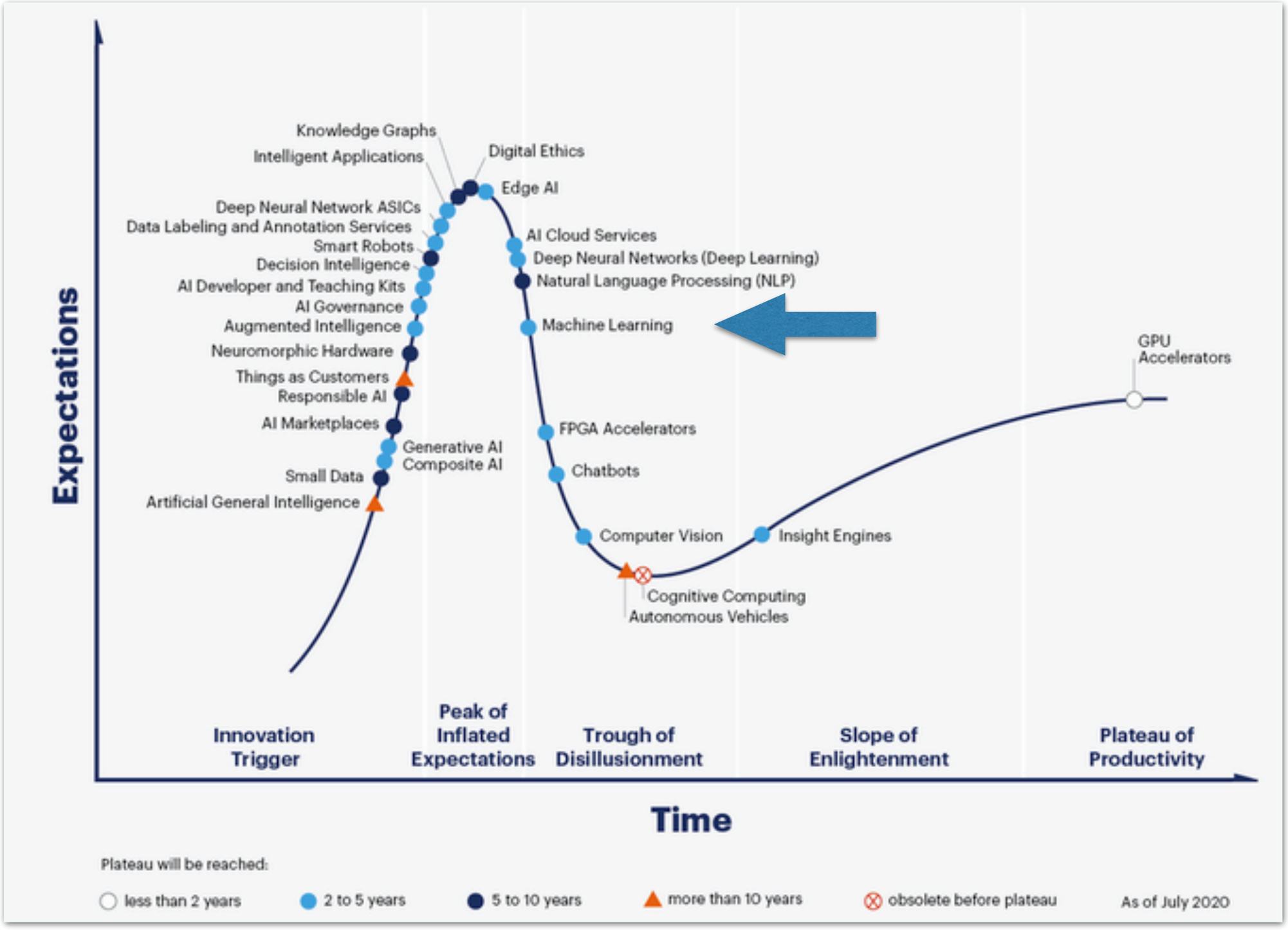
Medizinische Fakultät Heidelberg



[www.gartner.com]

Trends in künstlicher Intelligenz

Medizinische Fakultät Heidelberg



[www.gartner.com]

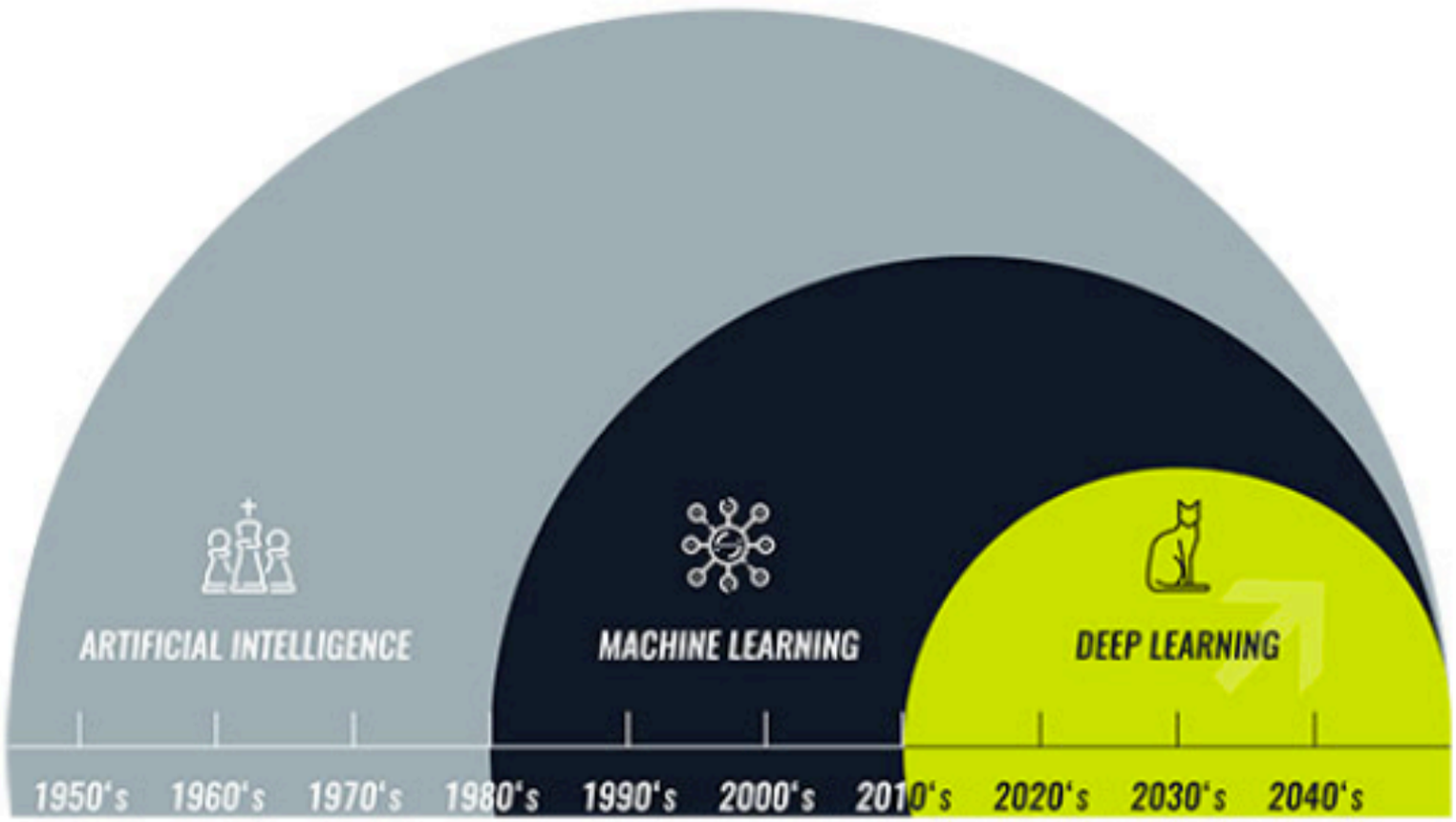
Künstliche Intelligenz? maschinelles Lernen?

Medizinische Fakultät Heidelberg

Künstliche Intelligenz
Konzept: Maschinen bauen,
die in der Lage sind, humane Aufgaben zu erfüllen
Beispiel: Schachkomputer, Expertensystem

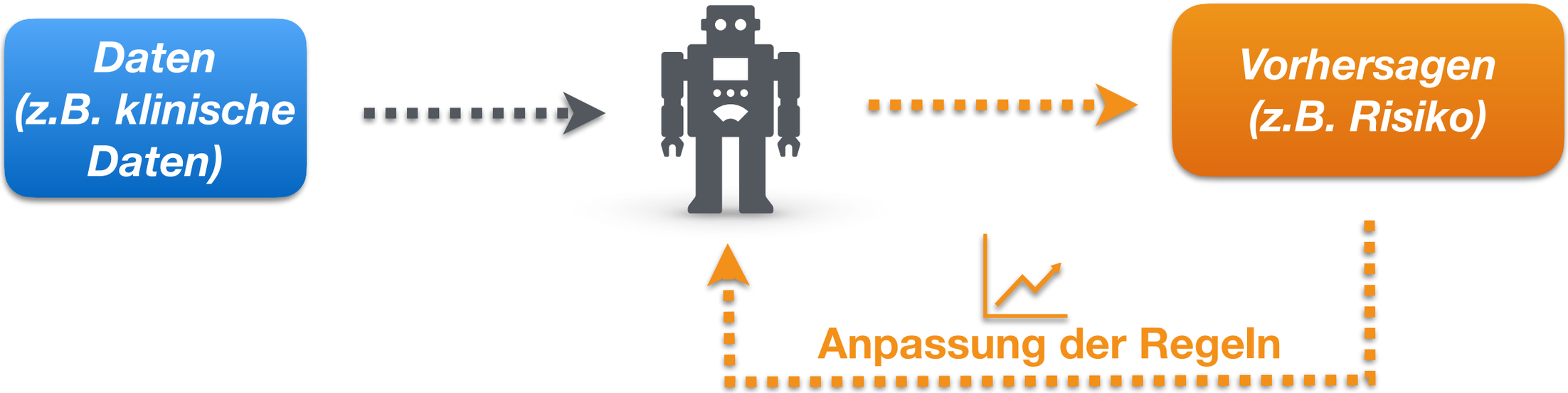
Maschinelles Lernen
konkrete Verfahren, um die Ziele der KI zu implementieren
"adaptive Algorithmen", die aus Daten lernen

Deep Learning
eine Form von ML, die auf neuronalen Netzwerken basiert



https://www.twt.de

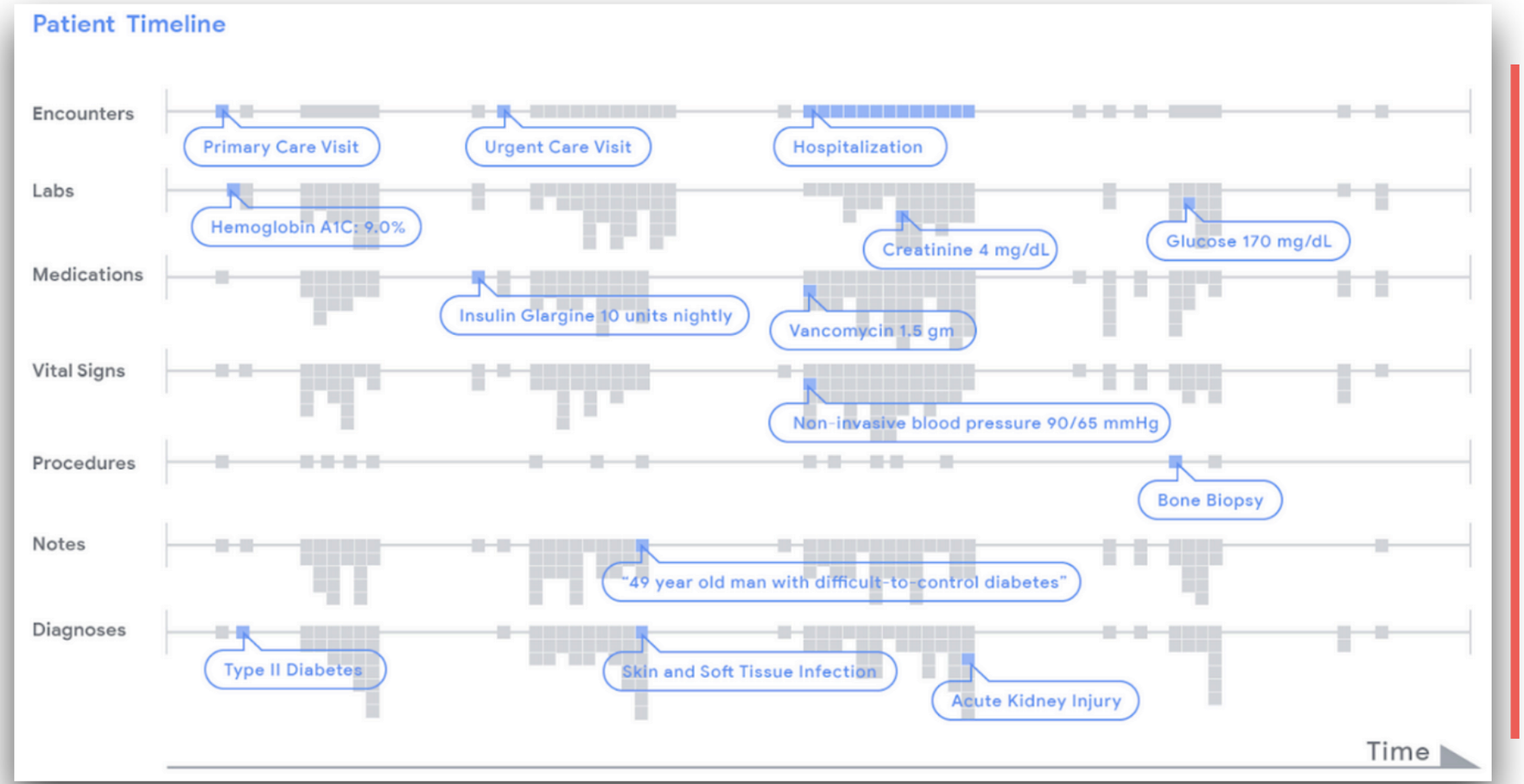
Maschinelles Lernen



- ML = iteratives Verfahren
- Kostenfunktion wird dabei minimiert (oder Qualitätsfunktion maximiert)
- Dazu werden die Vorhersageregeln angepasst

Beispiel

- Vorhersage von *Patient-outcome* aus elektronischer Patientenakte



Merkmale

Patient outcome?
(duration of stay readmission,...)

https://ai.googleblog.com/2018/05/deep-learning-for-electronic-health.html

Medizinische Fakultät Heidelberg

2. Datentypen

Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Beispiel: klinische Daten, Diabetes Patienten

Medizinische Fakultät Heidelberg

Merkmale ("features") (klinische Daten)

Beobachtungen (Patienten)

id	chol	stab.glu	hdl	ratio	glyhb	location	age	gender	height	weight	frame	bp.1s	bp.1d	bp.2s	bp.2d	waist	hip	time.ppn
1000	203	82	56	3.60	4.31	Buckingham	46	female	62	121	medium	118	59	NA	NA	29	38	720
1001	165	97	24	6.90	4.44	Buckingham	29	female	64	218	large	112	68	NA	NA	46	48	360
1002	228	92	37	6.20	4.64	Buckingham	58	female	61	256	large	190	92	185	92	49	57	180
1003	78	93	12	6.50	4.63	Buckingham	67	male	67	119	large	110	50	NA	NA	33	38	480
1005	249	90	28	8.90	7.72	Buckingham	64	male	68	183	medium	138	80	NA	NA	44	41	300
1008	248	94	69	3.60	4.81	Buckingham	34	male	71	190	large	132	86	NA	NA	36	42	195
1011	195	92	41	4.80	4.84	Buckingham	30	male	69	191	medium	161	112	161	112	46	49	720
1015	227	75	44	5.20	3.94	Buckingham	37	male	59	170	medium	NA	NA	NA	NA	34	39	1020
1016	177	87	49	3.60	4.84	Buckingham	45	male	69	166	large	160	80	128	86	34	40	300
1022	263	89	40	6.60	5.78	Buckingham	55	female	63	202	small	108	72	NA	NA	45	50	240
1024	242	82	54	4.50	4.77	Louisa	60	female	65	156	medium	130	90	130	90	39	45	300
1029	215	128	34	6.30	4.97	Louisa	38	female	58	195	medium	102	68	NA	NA	42	50	90
1030	238	75	36	6.60	4.47	Louisa	27	female	60	170	medium	130	80	NA	NA	35	41	720
1031	183	79	46	4.00	4.59	Louisa	40	female	59	165	medium	NA	NA	NA	NA	37	43	60
1035	191	76	30	6.40	4.67	Louisa	36	male	69	183	medium	100	66	NA	NA	36	40	225
1036	213	83	47	4.50	3.41	Louisa	33	female	65	157	medium	130	90	120	96	37	41	240
1037	255	78	38	6.70	4.33	Louisa	50	female	65	183	medium	130	100	NA	NA	37	43	180

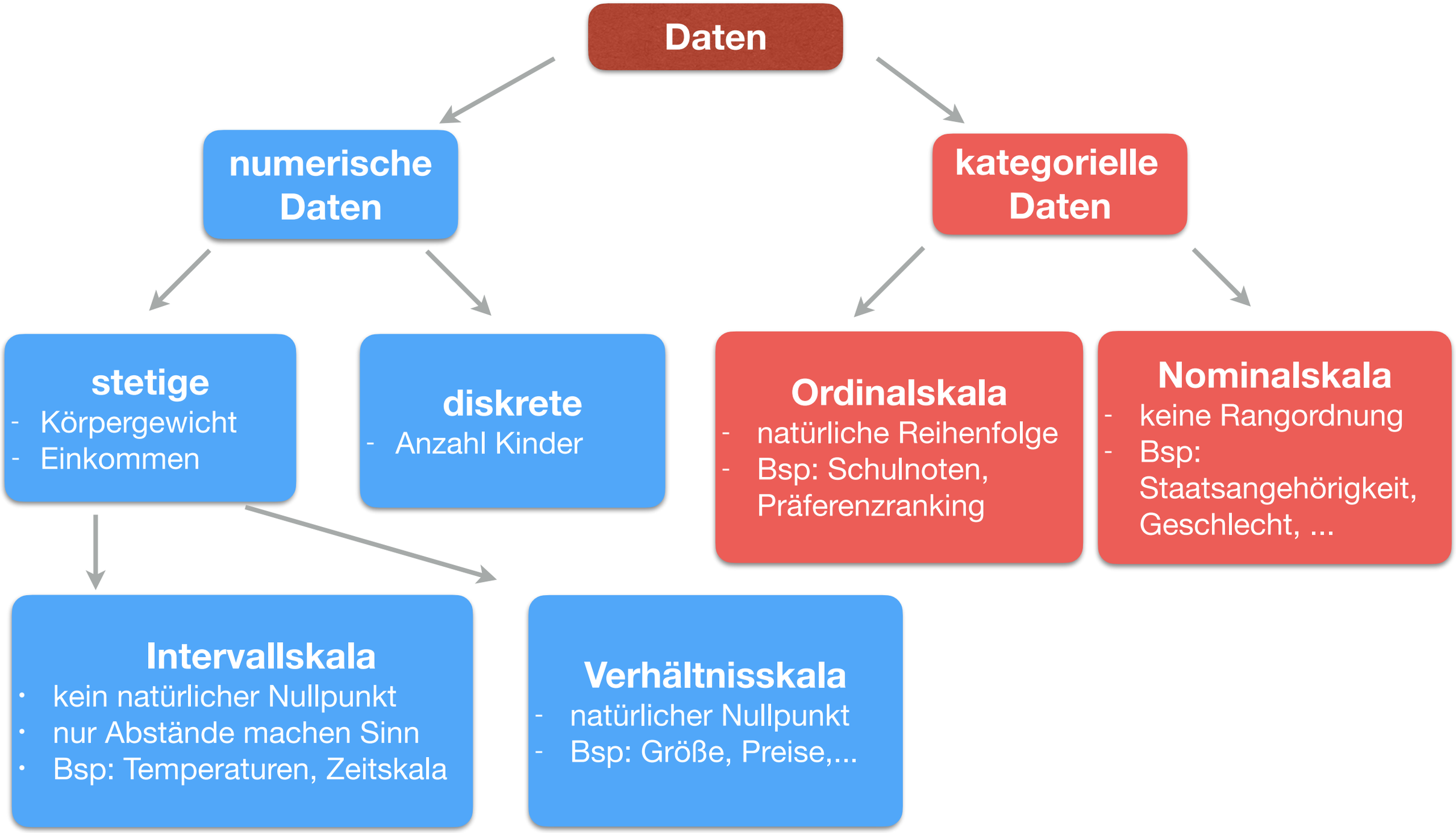
Unterschiedliche Datentypen

Medizinische Fakultät Heidelberg

Variable	Explanation	Unit	Type
chol	total cholesterol		stetig
stab.glu	Stabilized Glucose		stetig
hdl	High Density Lipoprotein		stetig
ratio	Cholesterol/HDL Ratio		stetig
glyhb	Glycosolated Hemoglobin		stetig
location			**kategoriell**
age			numerisch, diskret
gender			**kategoriell**
height		inches	stetig
weight		pounds	stetig
frame			**kategoriell**
bp.1s	systolic blood pressure		stetig
bp.1d	diastolic blood pressure		stetig
bp.2s	systolic blood pressure		stetig
bp.2d	diastolic blood pressure		stetig
waist		inches	stetig
hip		inches	stetig
time.ppn	Time since last meal	minutes	stetig

Datentypen

Medizinische Fakultät Heidelberg



Datenumwandlung

Medizinische Fakultät Heidelberg

- Einige Algorithmen können nur mit **numerischen** (diskreten/stetigen) Variablen arbeiten; kategorische Variablen müssen daher **umgewandelt** werden
- **Ordinale Daten:** haben wie Zahlen eine natürliche Rangordnung
  - "erster", 'zweiter',''dritter' → 1, 2, 3
  - 'sehr gut', 'gut', 'befriedigend', ... → 1, 2, 3, ...
- **Nominale Daten:** Nominale Daten haben keine natürliche Rangordnung
  - 'ja','nein' → 1, 0 oder 0, 1
  - 'rot','grün', 'blau' → 1, 2, 3 aber warum nicht 2, 3, 1 ?
  - Problem: rot=1, grün=2, blau=3: sind rot und grün näher als rot und blau?
- Lösung: "***one-hot encoding***"

one-hot encoding

Medizinische Fakultät Heidelberg

	Ethnicity
Patient 1	Caucasian
Patient 2	AfricanAmerican
Patient 3	Hispanic
Patient 4	Caucasian

	Caucasian	AfricanAmerican	Hispanic
Patient 1	1	0	0
Patient 2	0	1	0
Patient 3	0	0	1
Patient 4	1	0	0

- Nominale Variablen werden durch **binäre Variablen** ersetzt *("dummy variable")*
- Anzahl der binären Variablen = Anzahl der Ausprägungen (hier 3: Caucasian/AA/Hispanic)

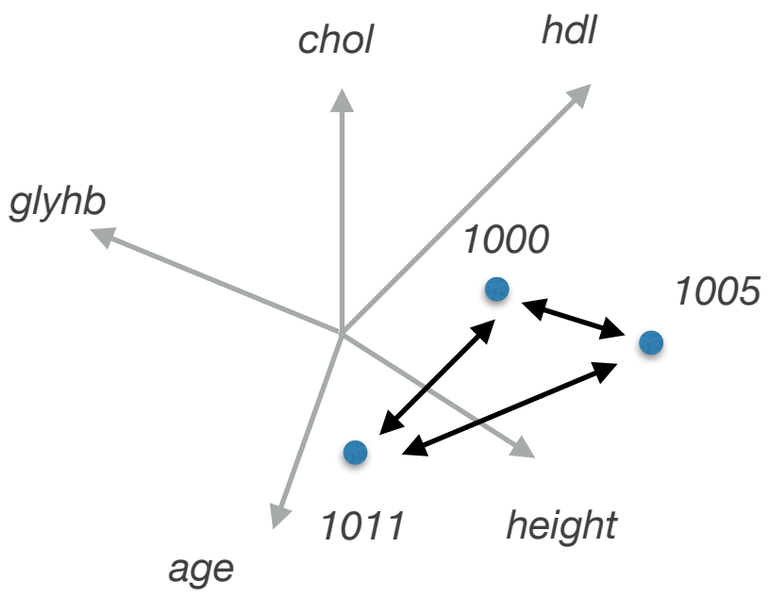
Beispiel: klinische Daten, Diabetes Patienten

Medizinische Fakultät Heidelberg

Merkmale (klinische Daten)

Beobachtungen (Patienten)

id	chol	stab.glu	hdl	ratio	glyhb	age	height	weight	bp.1s	bp.1d	waist	hip	time.ppn
1000	203	82	56	3.60	4.31	46	62	121	118	59	29	38	720
1001	165	97	24	6.90	4.44	29	64	218	112	68	46	48	360
1002	228	92	37	6.20	4.64	58	61	256	190	92	49	57	180
1003	78	93	12	6.50	4.63	67	67	119	110	50	33	38	480
1005	249	90	28	8.90	7.72	64	68	183	138	80	44	41	300
1008	248	94	69	3.60	4.81	34	71	190	132	86	36	42	195
1011	195	92	41	4.80	4.84	30	69	191	161	112	46	49	720



jede **Beobachtung** (= Patient)
kann als ein Punkt in einem mehr-dimensionalen Raum dargestellt werden
Koordinaten = **Merkmale**

Wie können Abstände/Ähnlichkeiten gemessen werden?

- **Distanzen** (je **kleiner** desto näher)
- **Ähnlichkeitsmaße** (je **größer** desto ähnlicher)

Distanzen

- Mögliche Distanzen:
  - **Euklidische Distanz**

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$d^2 = (9-2)^2+(7-4)^2 = 58$

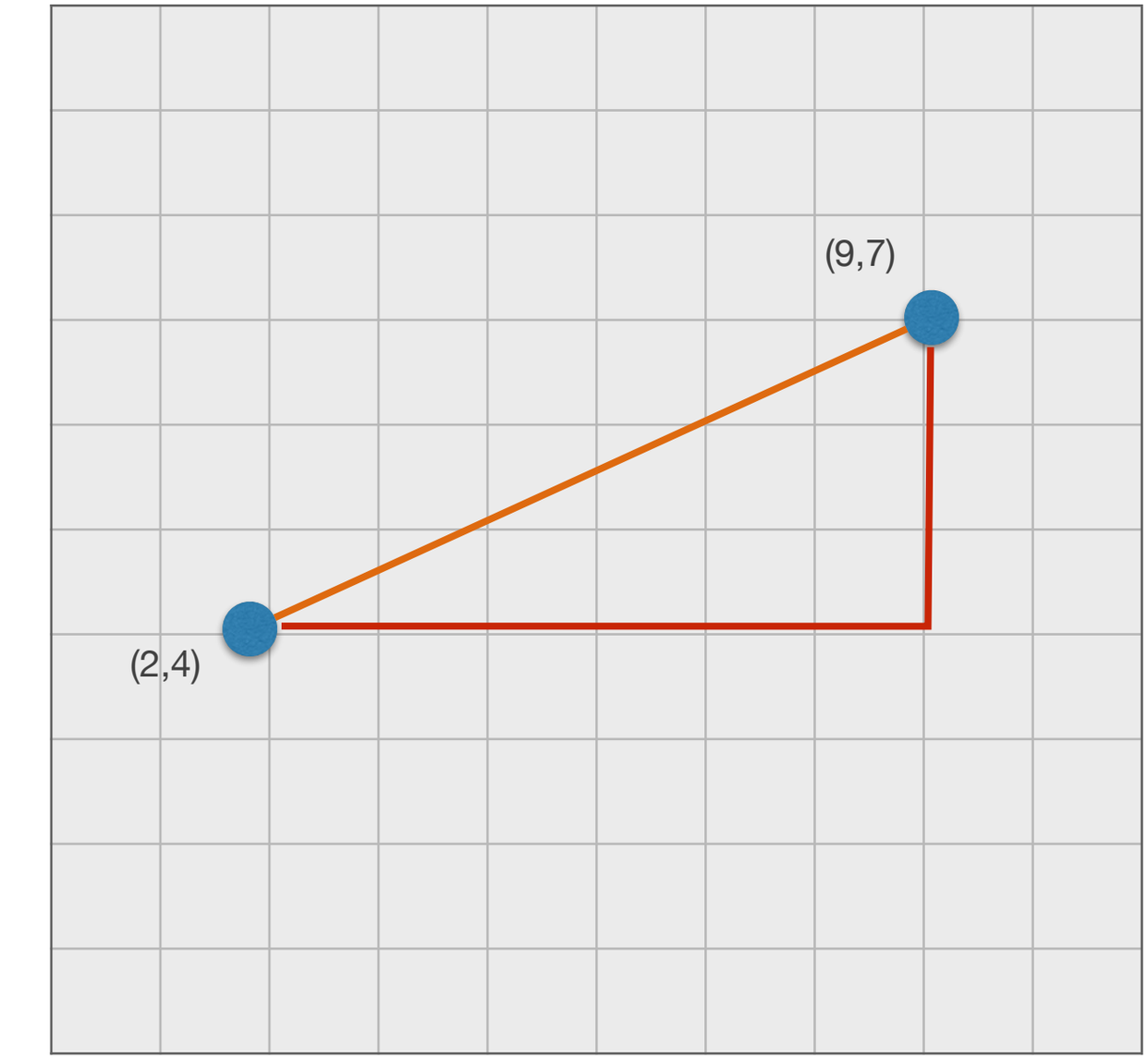
  - **Manhattan-Distanz**

$$d_{Manhattan}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

$d = 7 + 3 = 10$

  - **Korrelationsdistanz**
  - **...**

Diese Distanzen lassen sich auf mehrdimensionale Räume verallgemeinern!



Datenskalierung

Medizinische Fakultät Heidelberg

- Stetige Daten sind in bestimmten **Einheiten** dargestellt (Meter, Kg, Anzahl pro Minute,...)
- Daher sind die Intervalle, in denen sich die Daten bewegen, meistens unterschiedlich
  - Größe (m) : [1.50 - 2.00]
  - Größe (cm) : [150 - 200]
  - Herzschläge/Minute : [30 - 60]
- Distanzen sind meistens empfindlich gegenüber solchen Unterschieden
- Beispiel: Größe *h*, Gewicht *w*

$$d_{Euclidean}(x,y) = \sqrt{(h_x - h_y)^2 + (w_x - w_y)^2}$$

	Größe (m)	Gewicht (kg)
A	1.62	75
B	1.65	87
C	1.92	91

euklidische Distanz

	A	B	C
A	0	12	16
B		0	4
C			0

B/C am nächsten

	Größe (cm)	Gewicht (kg)
A	162	75
B	165	87
C	192	91

euklidische Distanz

	A	B	C
A	0	12.4	34
B		0	27.3
C			0

A/B am nächsten

Datenskalierung

- Lösung: Daten können durch eine Z-Transformation **zentriert** und **skaliert** werden:

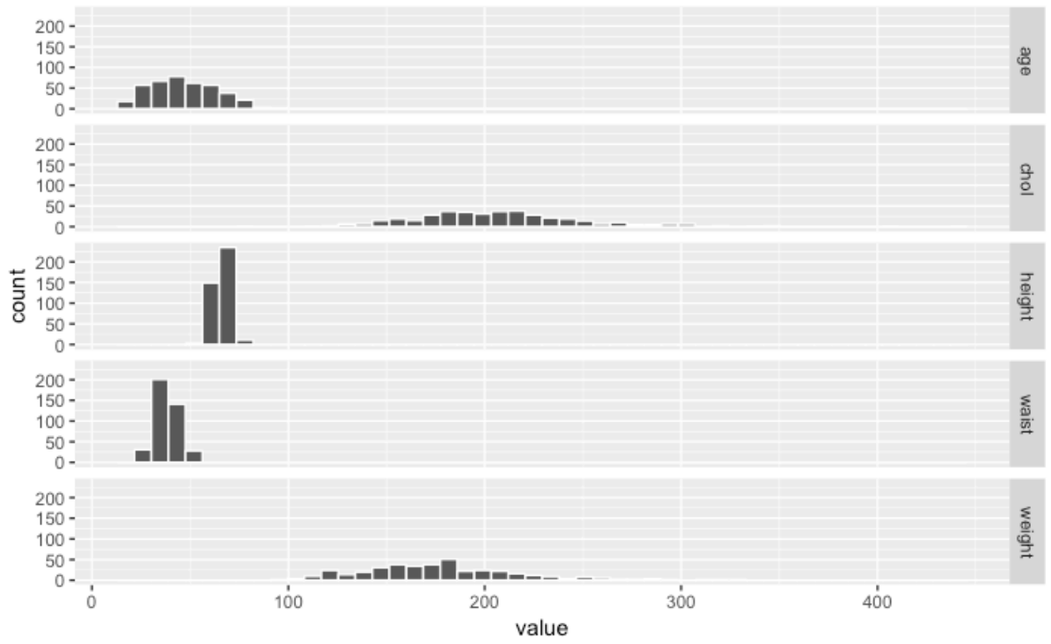
zentriert: Mittelwert wird abgezogen

$$\hat{x} = \frac{x - \bar{x}}{\sigma_x}$$

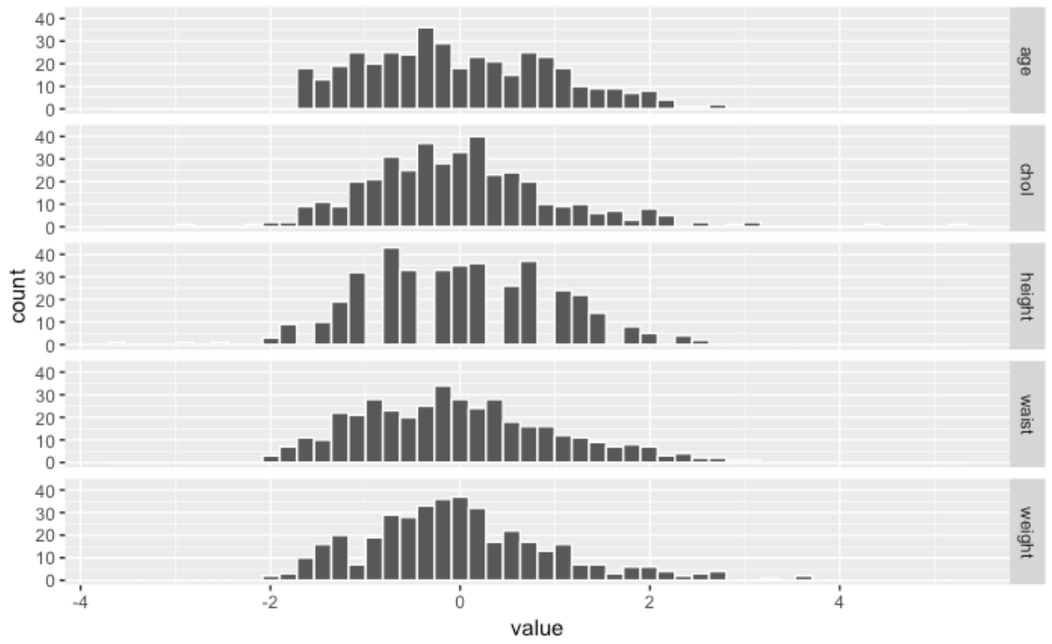
skaliert: geteilt durch Standardabweichung

- Normierte Daten haben Mittelwert 0 und Standardabweichung 1

ursprüngliche Daten



normierte Daten



Fragen ?

Medizinische Fakultät Heidelberg

3. Grundkonzepte in ML

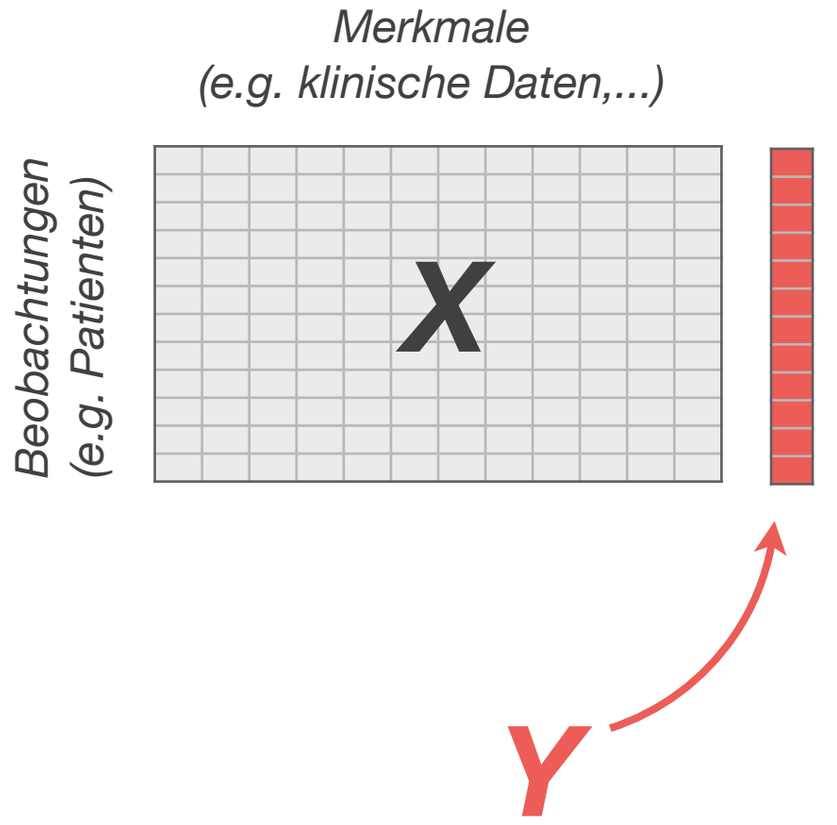
Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Aufgabe bei ML

- ***X*** : Beobachtete **Merkmale** (stetig, diskret, kategorielle Merkmale)
  - *klinische Parameter (Temperatur, Blutwerte, Bilddaten, ...)*
  - *Allgemeine Merkmale (Geschlecht, Alter,...)*
  - *Vorgeschichte (vorherige Behandlungen,...)*

- ***Y*** : **Klassen** oder **Zielwerte**, die *teilweise bekannt*, oder *nicht bekannt* sind
  - *Rückfallwahrscheinlichkeit (nur retrospektiv bekannt)*
  - *Versicherungsrisiko (nur retrospektiv bekannt)*
  - *Tumor-subtyp (Expertenmeinung = aufwendig)*
  - *Diagnose zu diabetischer Retinopathie (Expertenmeinung = nicht-vorhanden)*



- Überlebenszeit (stetig)
- Stärke der Nebenwirkungen (ordinal)
- Risikopatient? (binär)
- Subtyp (nominal)

Aufgabe bei ML

$$f : X \longrightarrow Y = f(X)$$

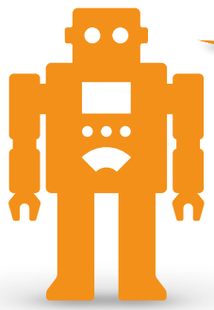
***Y* teilweise bekannt**
(supervidiertes Lernen)
oder **unbekannt**
(nicht-supervidiertes Lernen)

Merkmale X

Wahrer Zustand Y

?

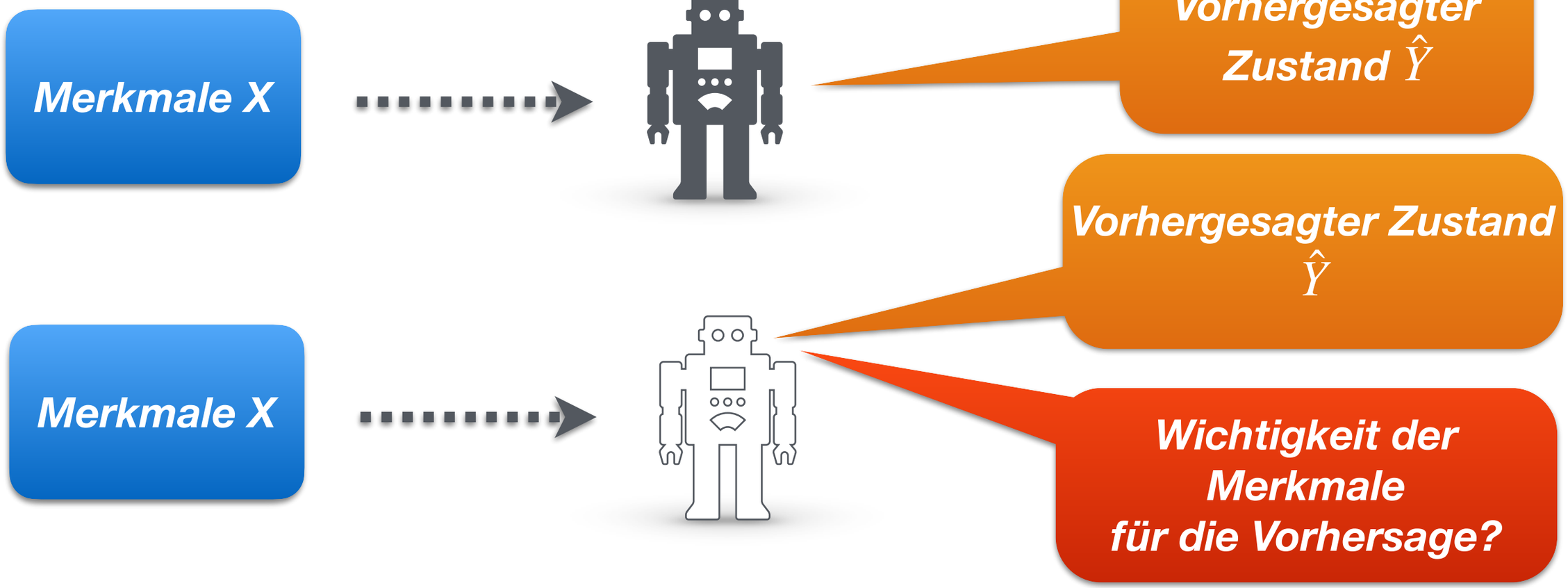
Modell



Vorhergesagter Zustand $\hat{Y}$

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

"Black-box" / "White-box"



- Bei Black-box Verfahren sind die Merkmale meistens nicht interpretierbar (e.g. SVM, deep neural networks,...)
- Bei **"White-box"** Verfahren wird die Wichtigkeit der Merkmale *("feature importance")* vorhergesagt (lineare Regression, Entscheidungsbäume,...)
- Dadurch können unwichtige Merkmale eliminiert werden *("feature selection")*

Medizinische Fakultät Heidelberg

Klassen / Zielwerte *Y* bekannt?

nein

Verfahren sollte ...

... explorativ sein

... Gruppen vorhersagen

Dimensions-Reduktion

- Hauptkomponenten Analyse
- Manifold learning
- t-SNE
- ...

Clustering

- hierarchisches Clustering
- k-means
- Gaussian mixture model

Nicht-supervidierte Verfahren

ja

Y beschreibt ...

... Klassen (z.B. Tumortypen)

... stetige Werte (z.B. Cholesterin)

Klassifizierung

- k-nearest neighbors
- Entscheidungsbäume
- Random Forest
- Support Vector Machine (SVM
- ...

Regression

- (nicht) lineare Regression
- Random Forest Regression
- SVM
- ...

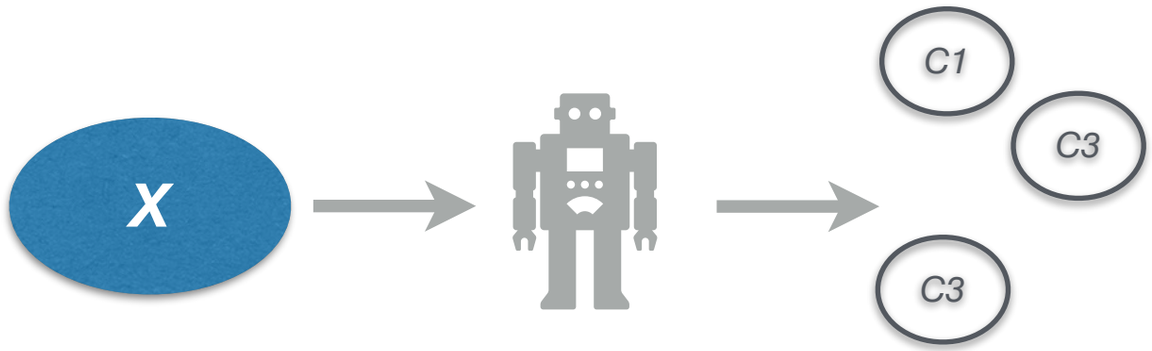
Supervidierte Verfahren

Nicht-supervidiertes Lernen

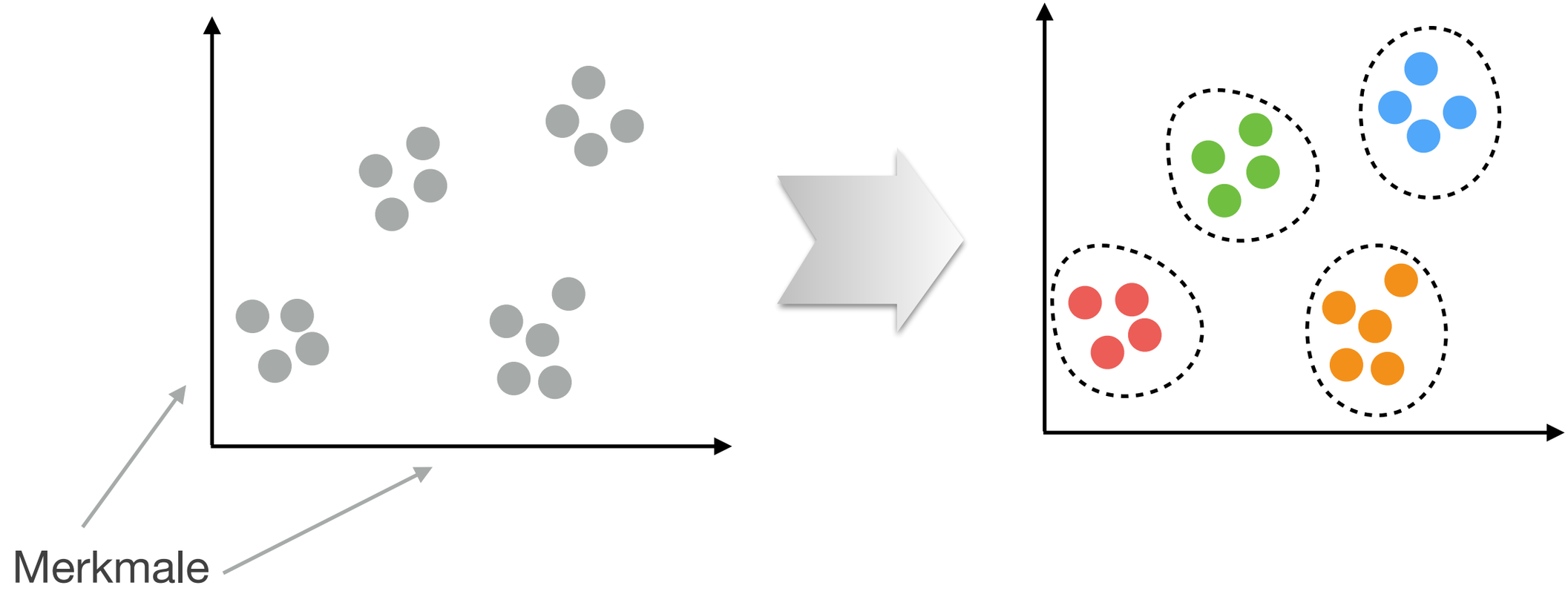
Medizinische Fakultät Heidelberg

- Klasse/Zielwert ***Y*** nicht-bekannt:

  nicht-supervidiertes Lernen
  - Clustering
  - Bildsegmentierung



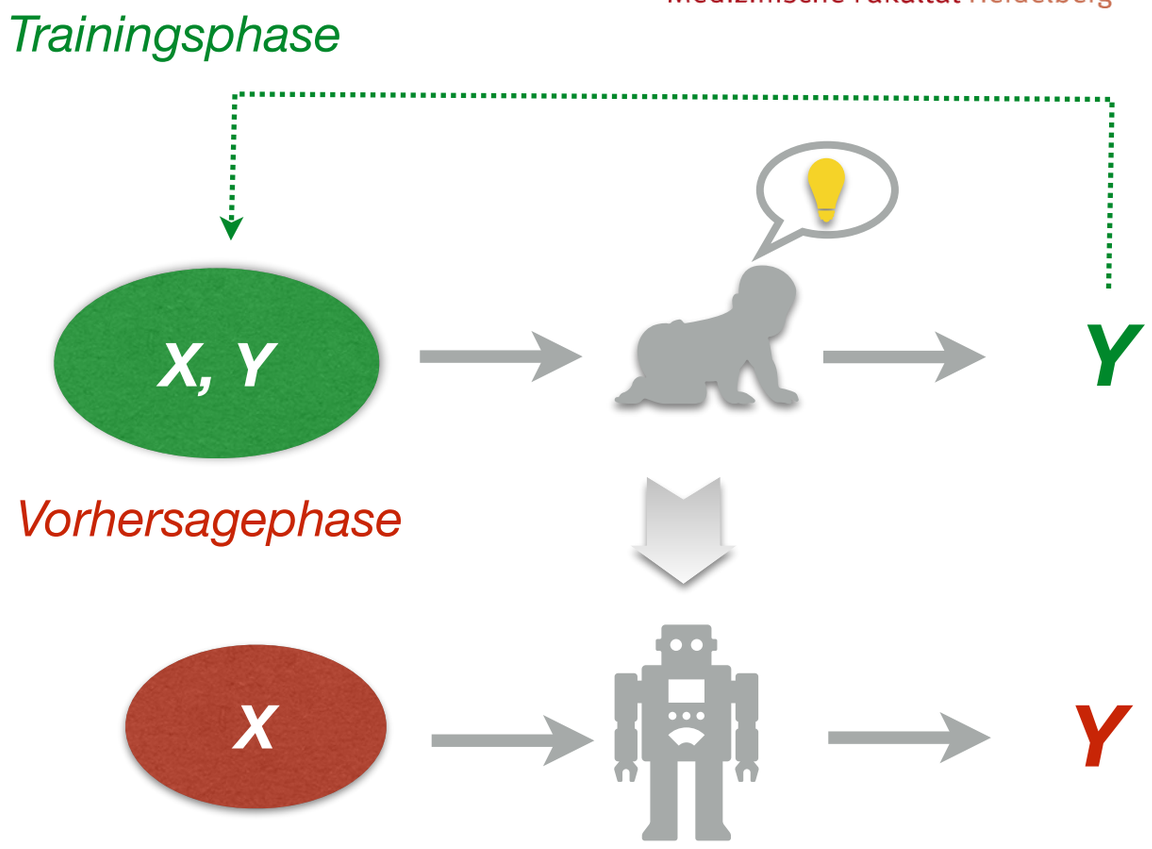
Vorhersage von Clustern



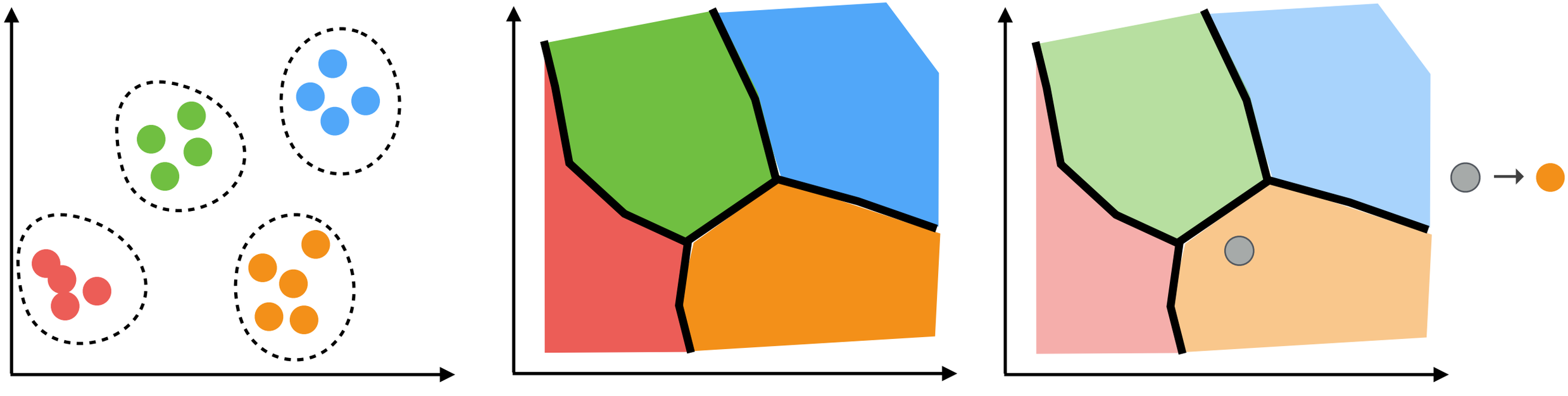
Supervidiertes Lernen

Medizinische Fakultät Heidelberg

- Klasse/Zielwert Y für eine Untermenge bekannt: **supervidiertes Lernen**
  - Trainingsphase anhand des Trainingsdatensatzes
  - Evaluation auf dem Testdatensatz
  - Vorhersage auf neuem Datensatz
  - Klassifizierung / Regression



Beispiel: k-nearest neighbors



Aufgabe bei ML

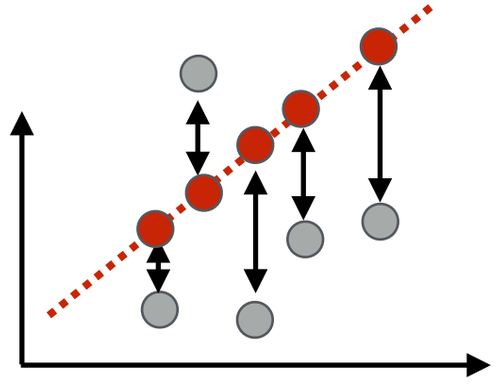
Medizinische Fakultät Heidelberg

- Wir suchen eine **Abbildung** $\hat{f}$:

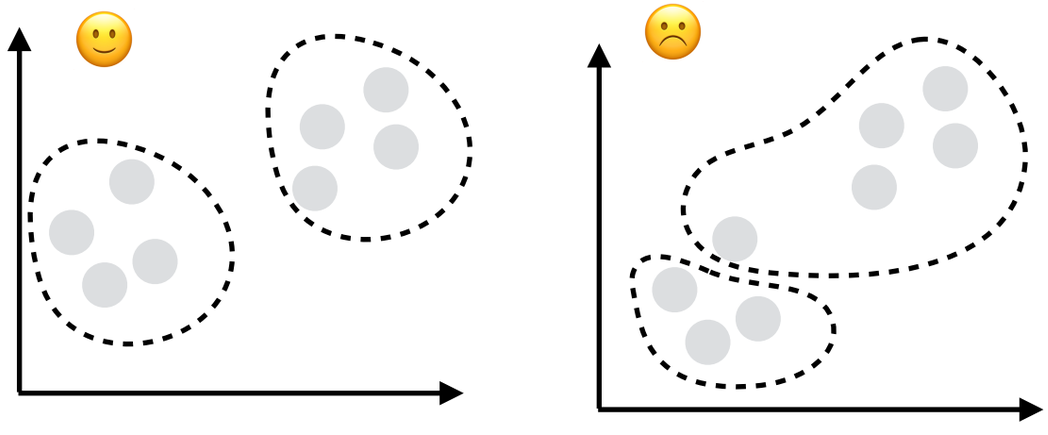
$$\hat{f} : X \longrightarrow \hat{Y} = \hat{f}(X)$$

- Sie sollte eine gewisse **Kostenfunktion** minimieren

  - *supervidiert*: **Vorhersagen** $\hat{Y}$ sollten so nah wie moglich an den **wahren Werten** sein

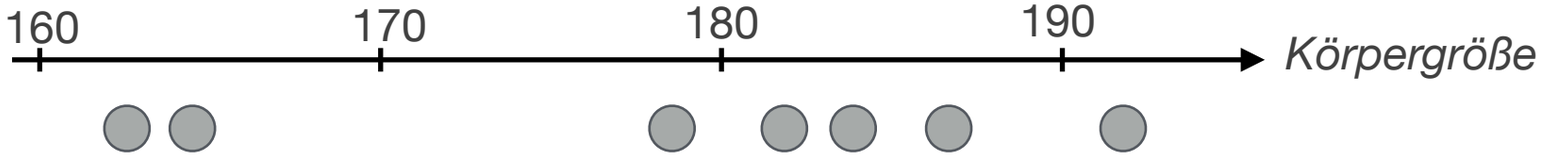


  - *nicht-supervidiert*: vorhergesagte Klassen $\hat{Y}$ sollten **konsistent** sein



- **nominal** bei nicht-supervidiertem Verfahren
- **nominal/ordinal** bei supervidierter *Klassifizierung*
- **stetig** bei supervidierter *Regression*

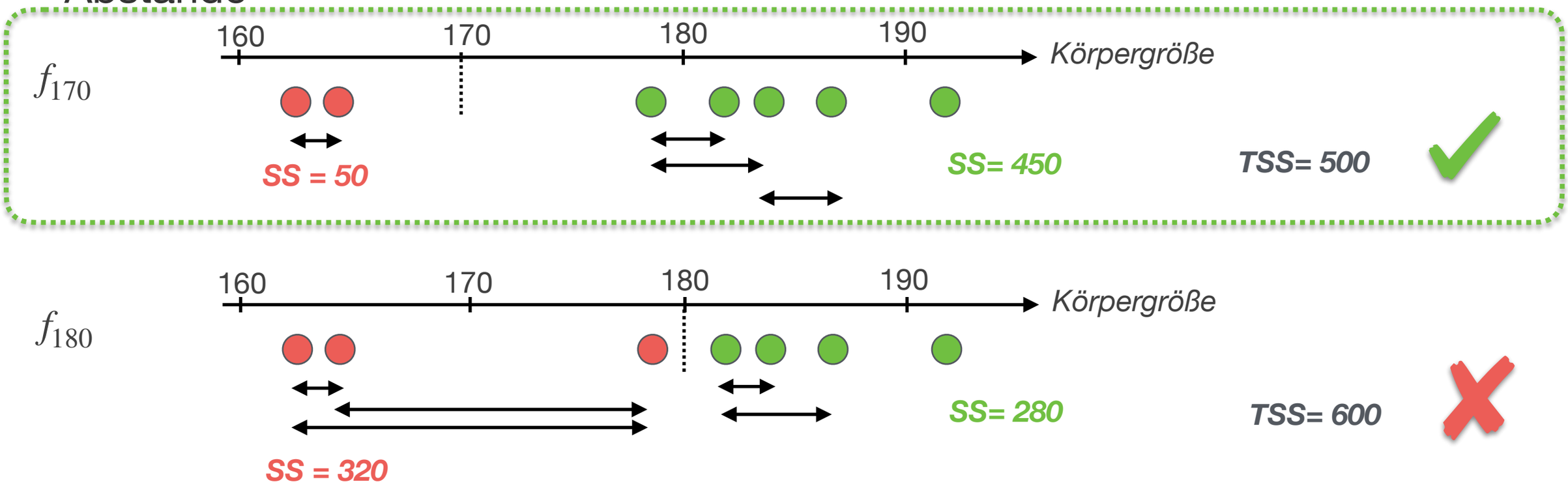
Beispiel: nicht-supervidiert



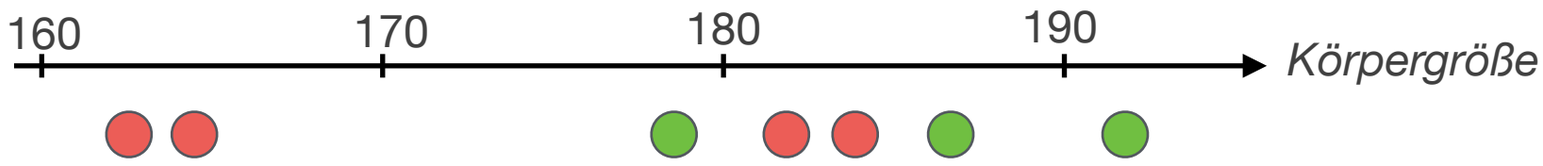
- **Ziel**: definiere 2 Gruppen
- Mögliche **Abbildung**: Stufenfunktion f_k

h >= k : ● → ●
h < k : ● → ●

- Mögliche **Kostenfunktion**: Summe der quadrate der paarweisen Abstände



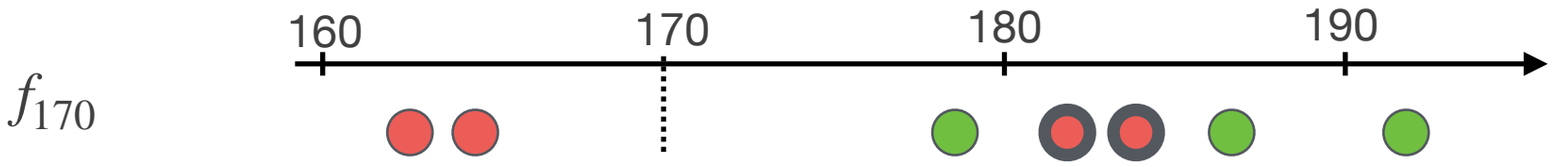
Beispiel: supervidiert



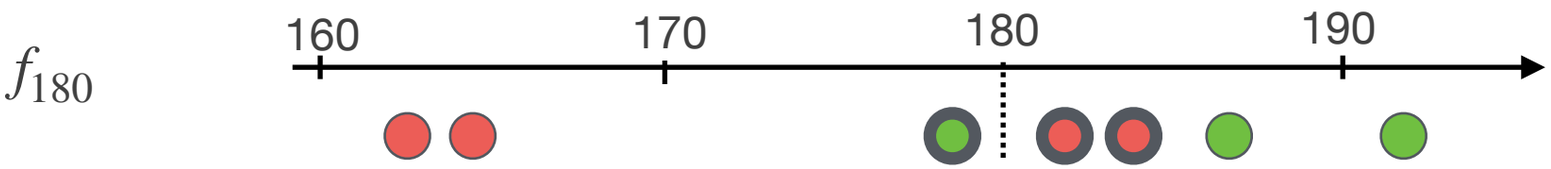
- **Ziel**: Lerne Regel für Entscheidung
- Mögliche **Abbildung**: Stufenfunktion f_k

h >= k : →
h < k : →

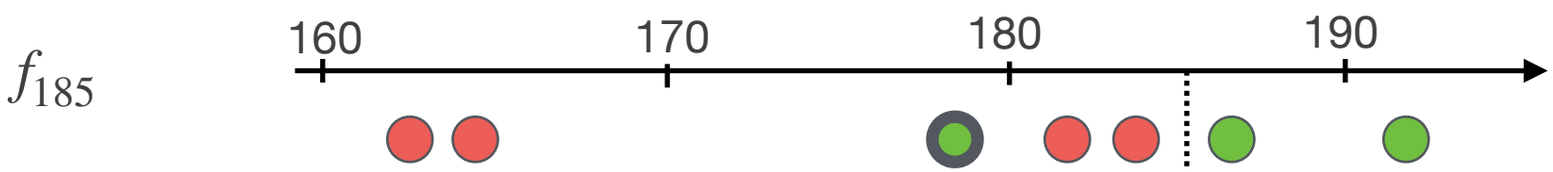
- Mögliche **Kostenfunktion**: Anteil der falsch Klassifizierten Punkte



Fehlerrate: 2/7 = 28.6%

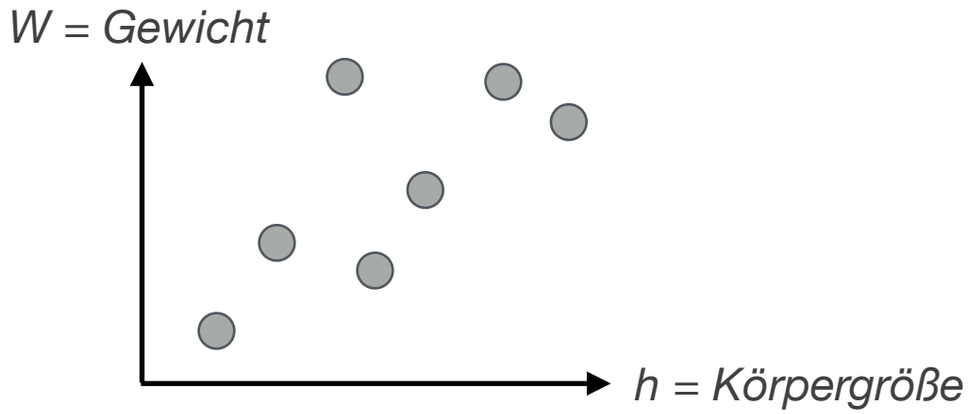


Fehlerrate: 3/7 = 42.8%

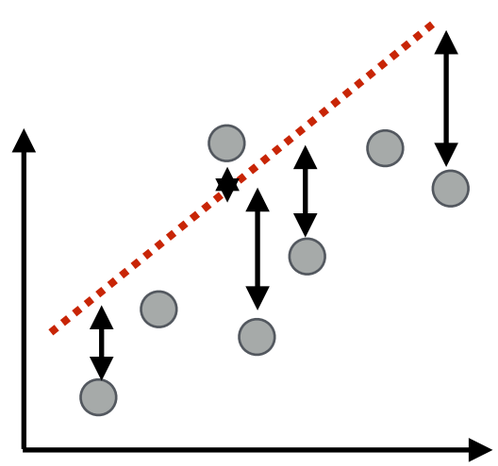


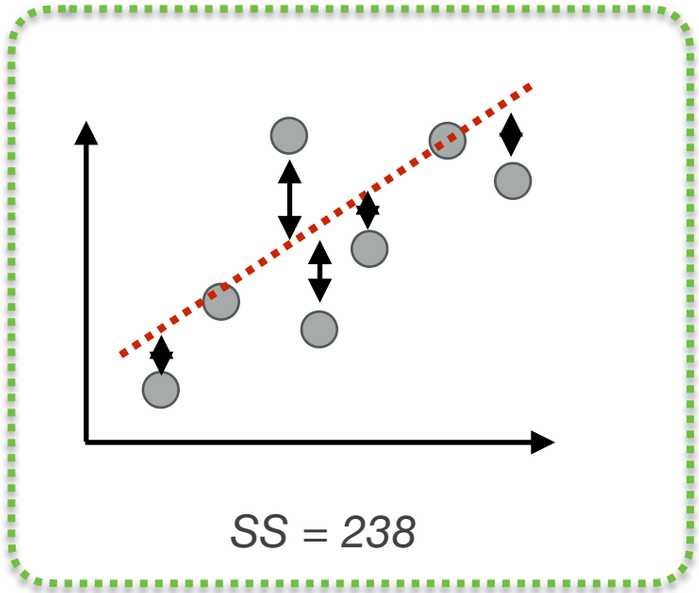
Fehlerrate: 1/7 = 14.3%

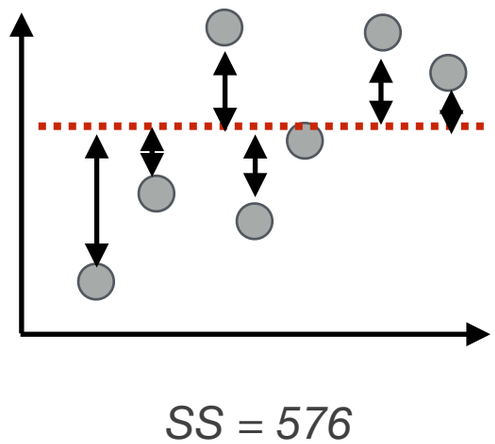
Beispiel: supervidiert



- **Ziel**: Lerne Regel *Größe* → *Gewicht*
- Mögliche **Abbildung**: lineare Funktion $W = f(h) = \theta_0 + \theta_1 \cdot h$
- Mögliche **Kostenfunktion**: Summe der quadratischen Abstände







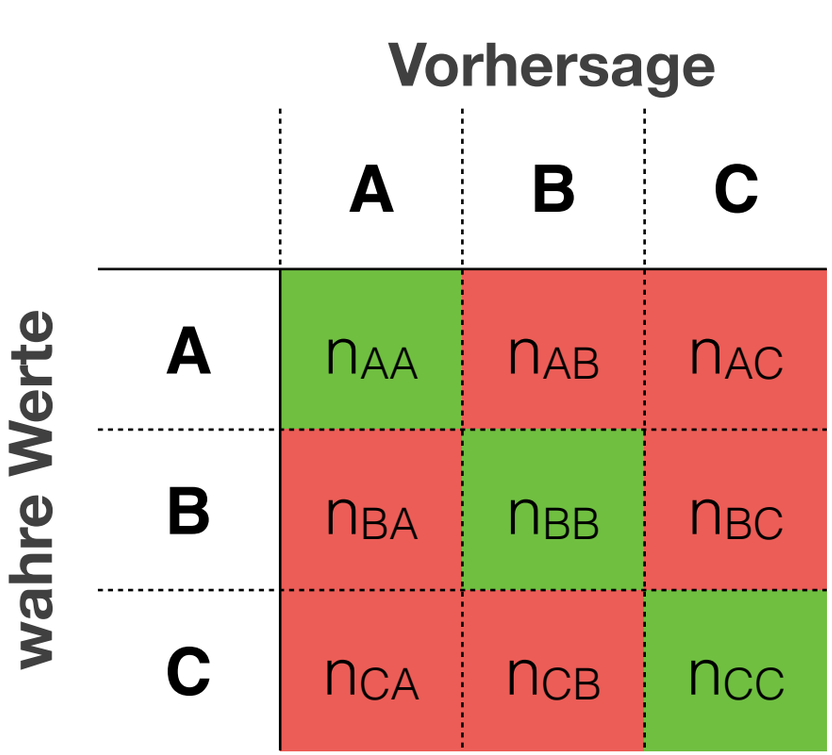
Kostenfunktion

Medizinische Fakultät Heidelberg

- bei *supervidierten Verfahren* werden die tatsächlichen Werte Y mit den vorhergesagten Werten Y verglichen → **Kostenfunktion**
- bei **Regressionsverfahren** sind Y und $\hat{Y}$ *stetige* Werte
- Mögliche Kostenfunktionen:
  - **(root) mean square error (RMSE)** $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$
  - **mean absolute error (MAE)** $$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$
- RMSE ist anfälliger für Ausreißer (durch die Potenz 2)!

Kostenfunktion

- bei *supervidierten Verfahren* werden die tatsächlichen Werte Y mit den vorhergesagten Werten Y verglichen → **Kostenfunktion**
- bei **Klassifizierungsverfahren** sind Y und $\hat{Y}$ *nominale/ordinale* Werte
- Vergleich durch *Konfusionsmatrix*



Korrektklassifizierungsrate (*Accuracy*)

$$KKR = \frac{\sum_i n_{ii}}{\sum_{i,j} n_{ij}} \in [0,1]$$

Falschklassifizierungsrate (*False prediction error*)

$$FKR = 1 - Accuracy = \frac{\sum_{i \neq j} n_{ij}}{\sum_{i,j} n_{ij}} \in [0,1]$$

Kostenfunktion

- **Trefferquote** (*recall*) für Klasse A:

$$Rec = \frac{WP}{WP + FN} = \frac{n_{AA}}{n_{AA} + n_{AB} + n_{AC}}$$

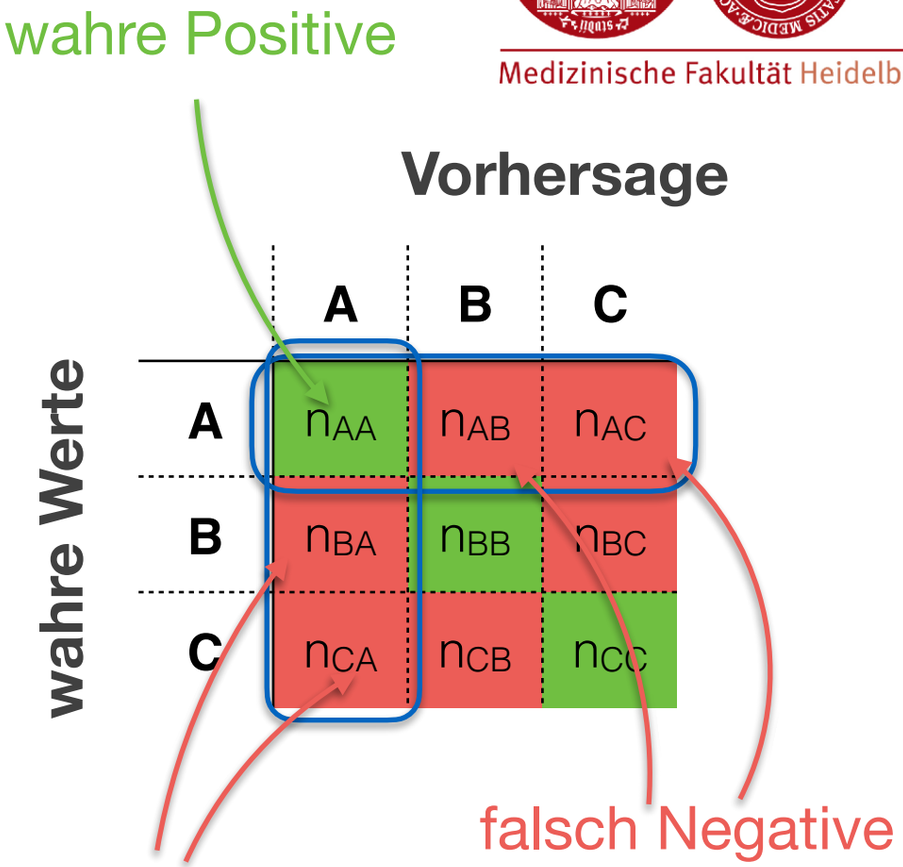
- **Genauigkeit** (*precision*) für Klasse A:

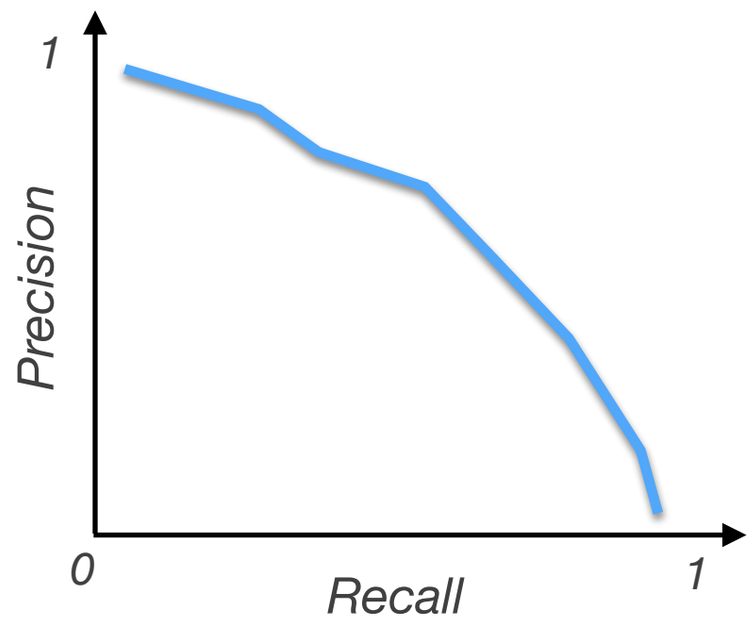
$$Prec = \frac{WP}{WP + FP} = \frac{n_{AA}}{n_{AA} + n_{BA} + n_{CA}}$$

- **F1-score**: kombiniert *Prec* und *Rec*

$$F1 = 2\,\frac{Prec \cdot Rec}{Prec + Rec}$$

- *Prec* und *Rec* hängen von den Parametern θ des Verfahrens ab; Verhalten kann in einem *Precision/Recall Diagramm* dargestellt werden





Fragen ?

Medizinische Fakultät Heidelberg

4. Modelle lernen

Carl Herrmann

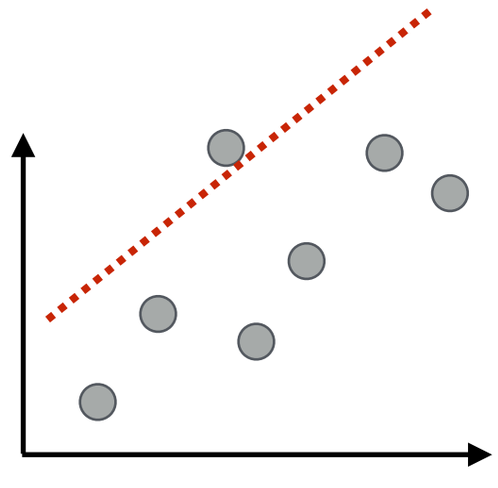
Health Data Science Unit - Medizinische Fakultät Heidelberg

Wie gut ist das Modell?

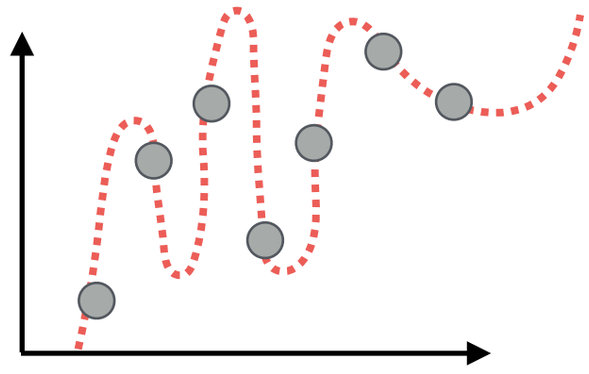
Medizinische Fakultät Heidelberg

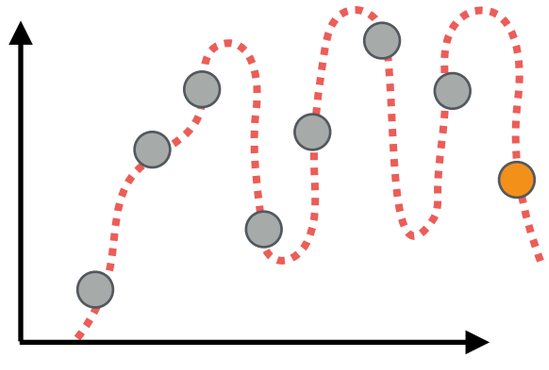
Modell: $\hat{Y} \equiv \hat{f}(X)$

- Es soll die vorhandenen Daten gut erklären
  - kein systematischer Fehler
  - geringe **Verzerrung** der Daten



- Es soll sich auch gut auf andere Daten verallgemeinern lassen
  - Robustheit gegenüber kleinen Änderungen in den Daten
  - geringe **Varianz** des Modells



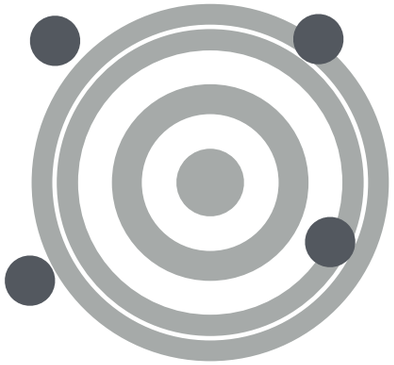


Varianz/Verzerrung Dilemma

Medizinische Fakultät Heidelberg

Fehler =

+



+

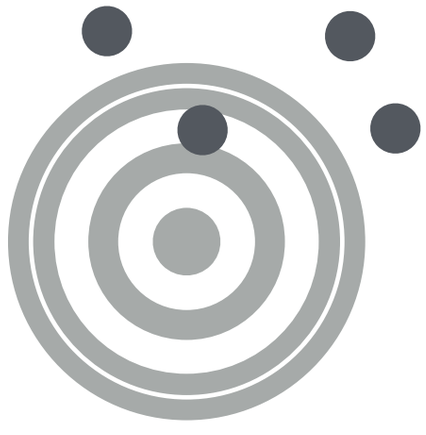
Verzerrung: wie weit vom Zielwert Y?

Varianz: wie variabel sind die Vorhersagen?

Rauschen in den Daten

Varianz/Verzerrung Dilemma

Medizinische Fakultät Heidelberg

	Hohe Varianz	Geringe Varianz
Hohe Verzerrung		
Geringe Verzerrung		

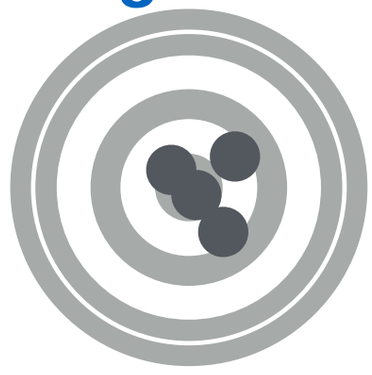
Varianz/Verzerrung Dilemma

Medizinische Fakultät Heidelberg

Hohe Varianz

Geringe Varianz

Geringe Verzerrung



- Optimales Modell: geringe Varianz, geringe Verzerrung
- Geringe Verzerrung mit hoher Varianz: Modell ist auf die wahren Werte zentriert, aber fluktuiert start und ist nicht stabil
  → lässt sich nicht auf neue Daten verallgemeinern!
- Effekt von **overfitting**:
  - Modell ist gut auf Trainingsdaten angepasst
  - lässt sich nicht verallgemeinern
- Ein komplexeres Modell (z.B. mit weiteren Merkmalen) führt zu **geringerer Verzerrung** aber **höherer Varianz**!

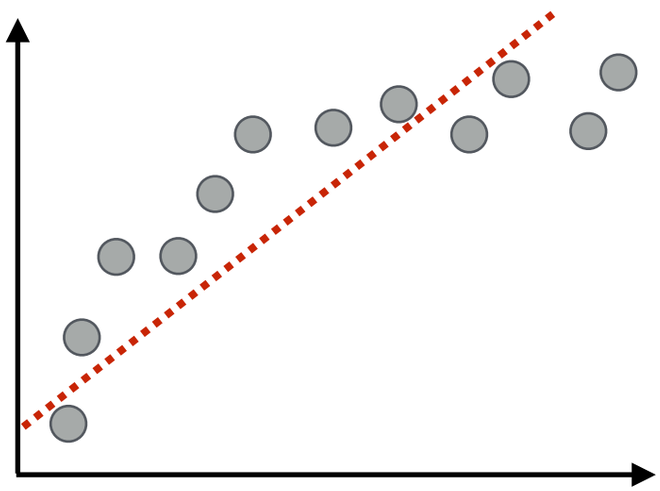
Varianz/Verzerrung Dilemma

	Hohe Varianz	***Geringe Varianz***
Hohe Verzerrung		

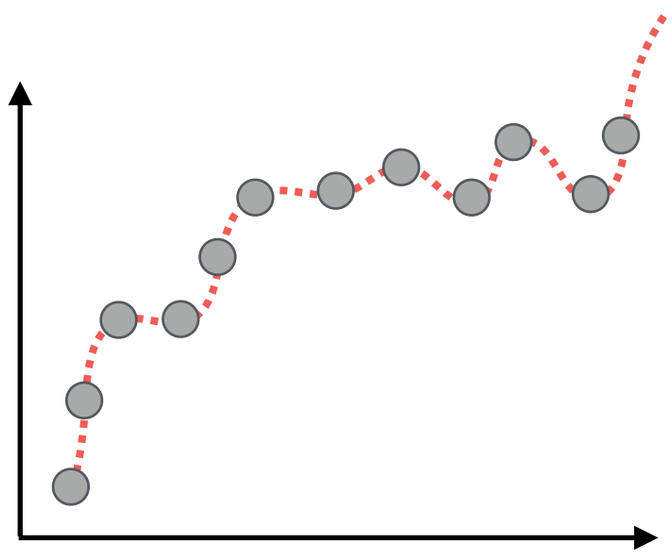
- Optimales Modell: geringe Varianz, geringe Verzerrung
- Geringe Varianz mit hoher Verzerrung: Modell hat einen systematischen Fehler

  → vermutlich ist das Modell zu einfach, um die Daten zu erklären
- Effekt von **underfitting**:
  - Modell hat eine zu geringe Komplexität, oder hat falsche Annahmen!

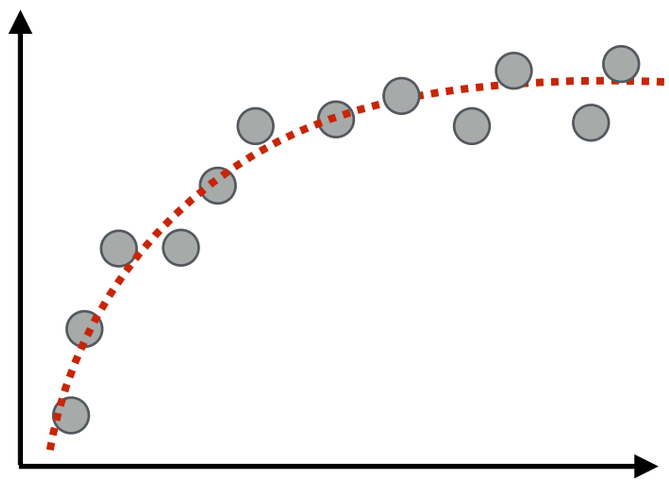
Overfitting / underfitting



- **Underfitting**:
  Modell ist zu einfach, um die Daten zu beschreiben
  → systematischer Fehler, **hohe Verzerrung**



- **Overfitting**:
  Modell ist perfekt an die Trainingsdaten angepasst, lässt sich aber schlecht auf neue Daten verallgemeinern!
  → hohe **Varianz**



- **Gutes Modell**:
  Modell ist gut an die Trainingsdaten angepasst, und lässt sich auf neue Daten verallgemeinern!
  → geringe **Varianz**
  → geringer **Verzerrung**

Modell trainieren

Medizinische Fakultät Heidelberg

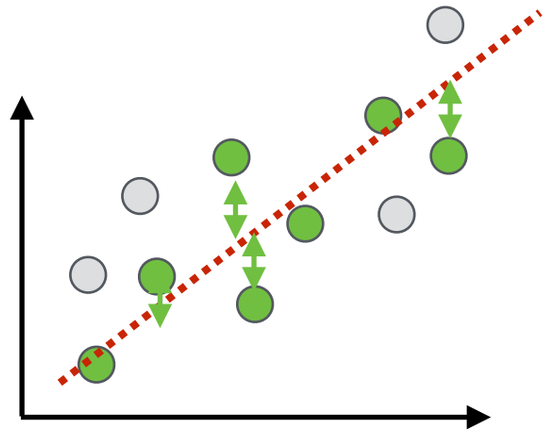
Klassen bekannt!

Trainings-Datensatz (~80%)

Test-Datensatz (~20%)

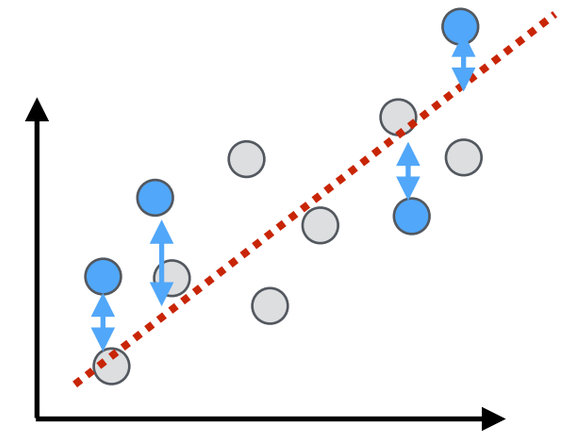
Modell wird auf dem Trainingsdatensatz erstellt...

... und auf dem Test-Datensatz bewertet



$\sum \updownarrow$ = Trainingsfehler

- **Trainingsfehler** hoch, **Testfehler** hoch
  → **under-fitting** (hoher Verzerrung)
- **Trainingsfehler** niedrig, **Testfehler** hoch
  → **over-fitting** (hohe Varianz)
- **Trainingsfehler** niedrig, **Testfehler** niedrig
  → **gutes Modell**



$\sum \updownarrow$ = Testfehler

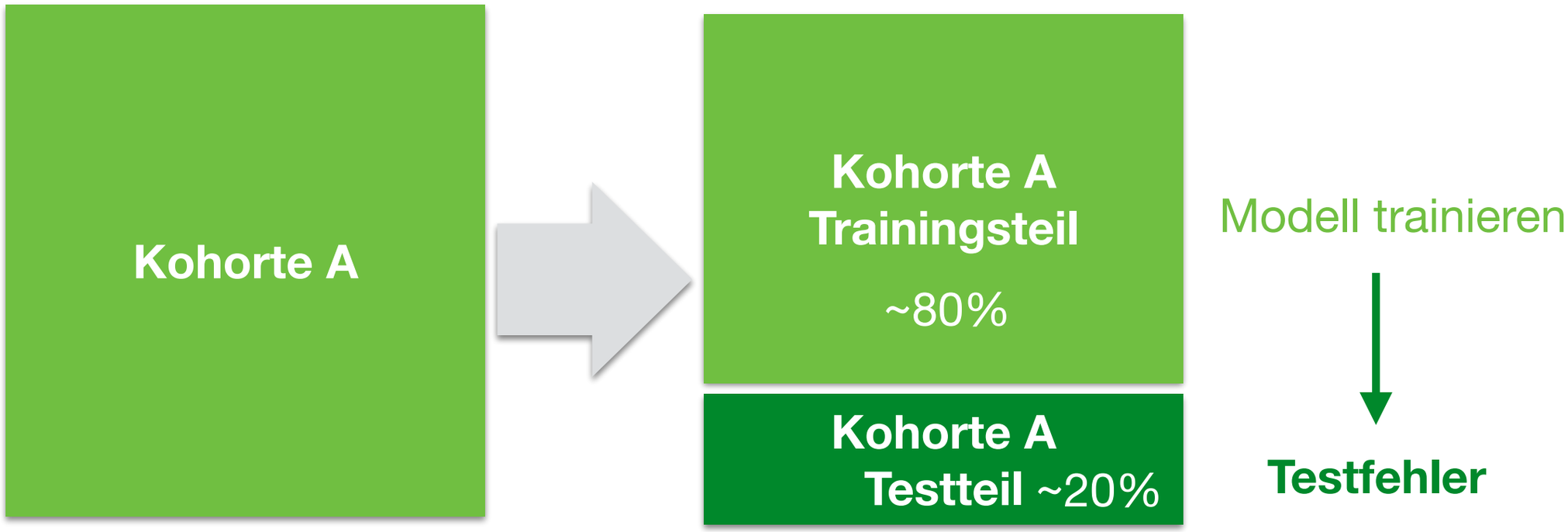
Modell optimieren

Medizinische Fakultät Heidelberg

- **Option 1:**
  es gibt einen großen Trainingsdatensatz, und einen unabhängigen Testdatensatz

Modell optimieren

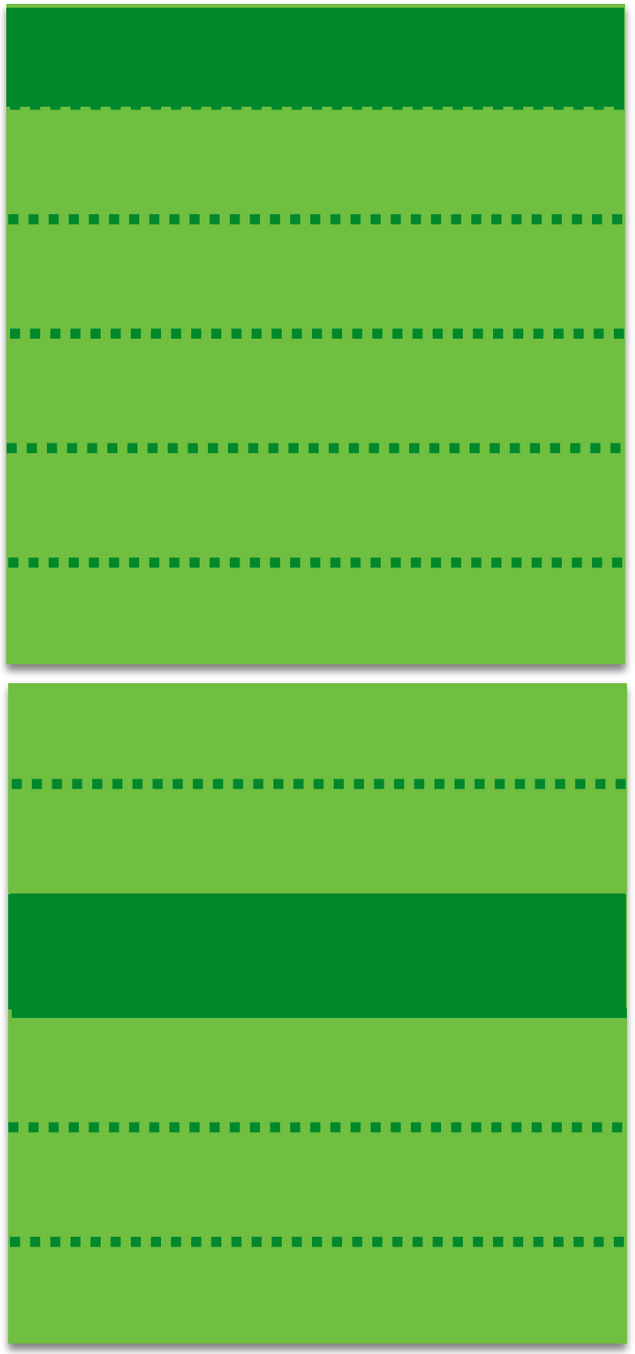
- **Option 2:**
  nur eine Kohorte → Aufteilung in Training / Testset



- **Nachteil**
  - 20% der Daten werden nicht zum Training benutzt
  - Wie soll die Aufteilung erfolgen?

Modell optimieren

- **Option 3: k-Fold cross-Validierung**
  - Aufteilung in *k* gleiche Teile
  - Lernen auf *k*-1, Testfehler auf übriggelassenem Teil
  - Wiederholung *k*-Mal
  - Mittlerer Testfehler + Standardabweichung



Testfehler 1

Testfehler 3

Modell optimieren

- **Option 4: Leave-one-our-cross validation (LOOCV)**
  - Eine der *n* Beobachtungen wird ausgelassen
  - Modell wir auf den *n-1* Datenpunkten trainiert
  - Anwendung auf übriggelassenem Punkt
  - *n*-Mal wiederholt!
  - Mittlerer Testfehler + Standardabweichung

Overfitting vermeiden

Medizinische Fakultät Heidelberg

- ***Over-fitting sollte vermieden werden!***
- Verschiedene Möglichkeiten

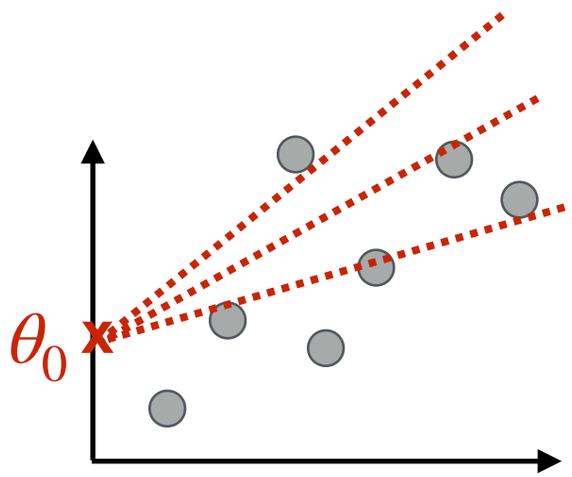
  - Modell **vereinfachen** (weniger Parameter)
    z.B. Grad des Regressions-Polynoms, Tiefe des Entscheidungsbaums, maximale Anzahl von Clustern, Wahl der Merkmale, ...

  - Modell **regularisieren**
    zusätzliche Bedingungen auf die Parameter des Modells

Regularisierung

- Beispiel: lineare Regression $\hat{Y} = \theta_0 + \theta_1 \cdot X$ 2 Parameter: θ_0, θ_1

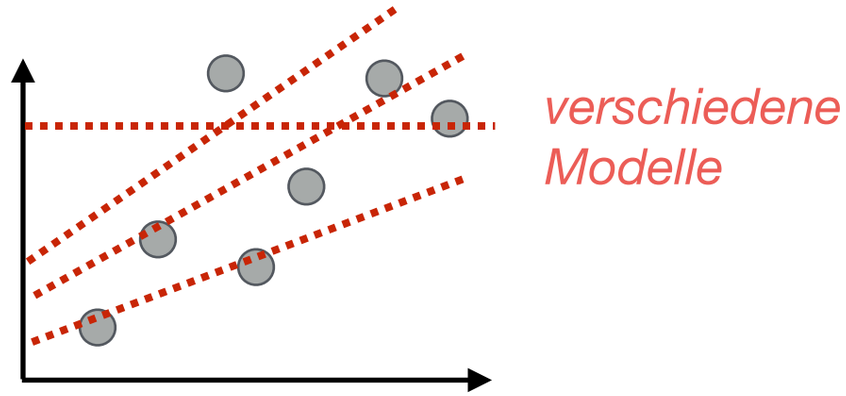
Parameter festlegen (hier θ_0)



In diesen Beispielen sind θ_0 oder λ **Hyper-parameter**, die vor Begin des Trainings festgelegt werden

→ ***wie werden diese Werte bestimmt?***

Zusätzliche Bedingungen auf Parameter



Ziel: bestimme θ_0, θ_1 die folgende Kostenfunktion minimieren

$$C = \sum_{i=1}^{n} (\theta_0 + \theta_1 \cdot x_i - y_i)^2 + \lambda (\theta_0^2 + \theta_1^2)$$

quadrierte Abweichungen *Regularisierung*

[Lasso Regularisierung]

Modell optimieren

Medizinische Fakultät Heidelberg

Modell wird auf dem Trainingsdatensatz für bestimmte Hyperparameter erstellt...

... und auf dem Validierungs-Datensatz bewertet

... bevor neue Werte der Hyperparameter ausprobiert werden

Test-
Datensatz
(~20%)

Finales Modell wird auf dem Test-Datensatz bewertet...

neuer
Datensatz

... bevor es auf einem neuen Datensatz angewendet wird!

Fragen ?

Medizinische Fakultät Heidelberg

5. supervidiertes Lernen: Regression

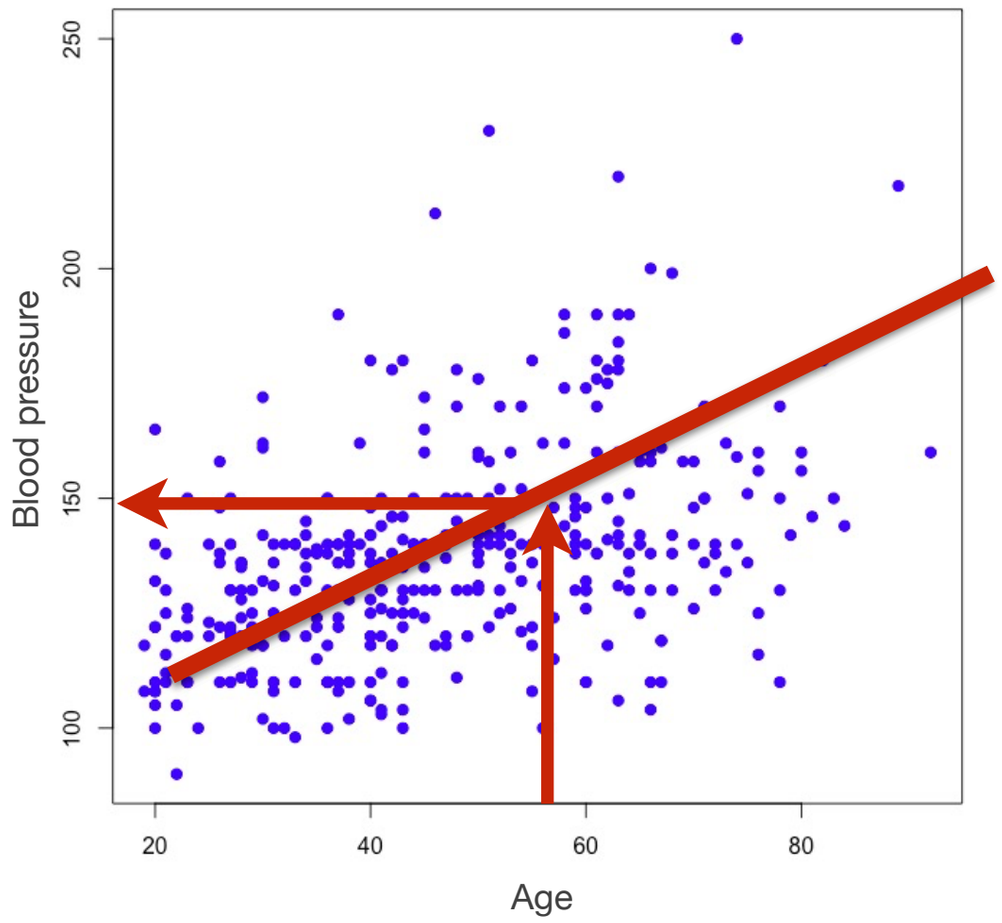
Carl Herrmann

Health Data Science Unit - Medizinische Fakultät Heidelberg

Regressions-Verfahren

Medizinische Fakultät Heidelberg

- **Ziel: Vorhersage eines stetigen Merkmals (z.B. Blutdruck) anhand anderer Merkmale**
- Mögliche Fragestellungen:
  - *Gibt es einen Zusammenhang zwischen Alter und BD?* → **Korrelation**
  - *Kann ich BD anhand des Alters vorhersagen?*→ **Regressions-Modell**



- Verfahren
  - **Lerne** Regressionsmodell aus den Daten
  - **Teste** auf nicht gesehenen Daten
  - **Vorhersagen auf neuen Daten**

$$\hat{Y}_i = \theta_0 + \theta_1 X_i$$

Merkmale

- **Stetige Merkmale** (Größe, Gewicht,...); Variablen müssen nicht unbedingt skaliert werden!

$$\hat{Y}_i = \theta_0 + \theta_1 X_i = \theta_0 + \theta_1 s \cdot \frac{X_i}{s}$$

- **Ordinale Merkmale**: small / medium / large ; können durch dargestellt werden 1 / 2 / 3, da es eine natürliche Reihenfolge gibt!
- **Nominale Merkmale**: Frau / Mann ; "one-hot encoding": Benutzung von Dummy Variablen (istFrau = 0/1 ; istMann = 0/1)

Training

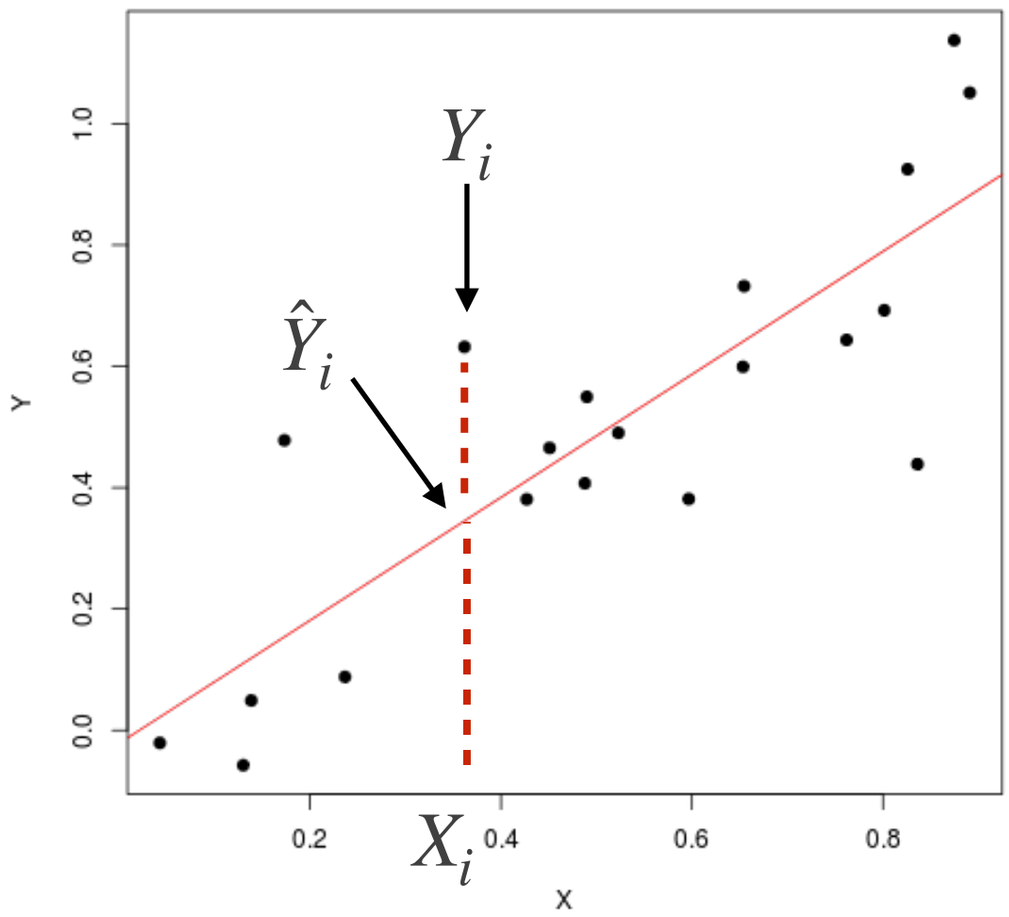
- Wir nehmen an, dass es einen linearen Zusammenhang gibt!

$$X = \{X_1, X_2, \ldots, X_n\} \quad Y = \{Y_1, Y_2, \ldots, Y_n\}$$

$$\hat{Y}_i = \theta_0 + \theta_1 X_i$$

- Für jedes X_i kann ein $\hat{Y}_i$ bestimmt werden!
- θ_0 = Schnittpunkt

  θ_1 = Steigung



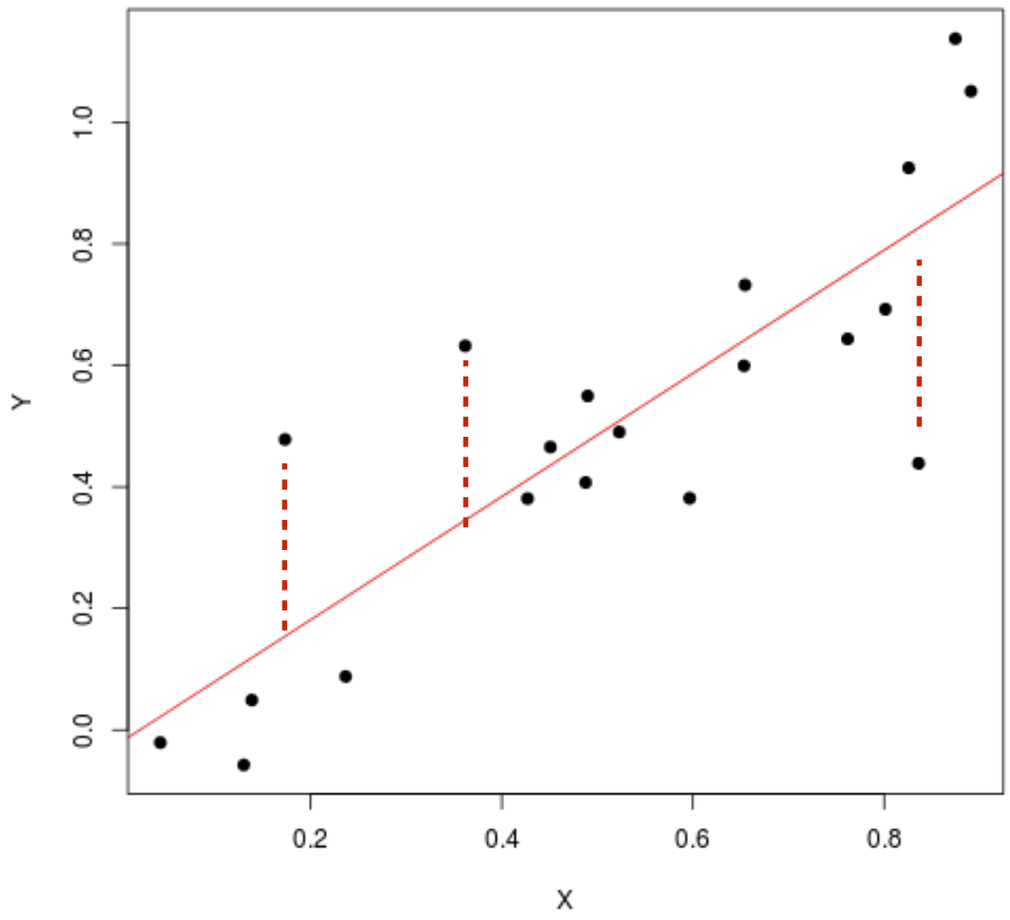
Training

- Parameter werden durch **Minimierung** der Kostenfunktion bestimmt:

$$\min_{\theta_0, \theta_1} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\theta_0 = \bar{Y} - \theta_1 \cdot \bar{X}$$

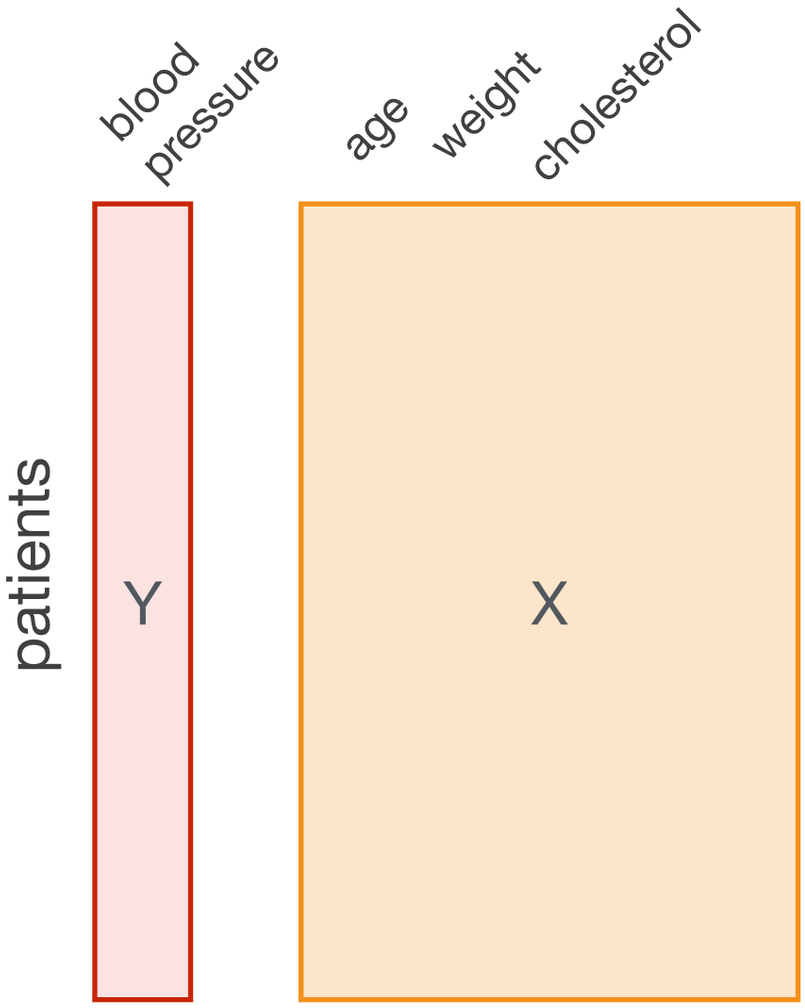
$$\theta_1 = corr(X, Y)\frac{s_Y}{s_X}$$



Multiple Regression

$$\hat{Y}_i = \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + \cdots$$

- Y ist das zu vorhersagende merkmal (z.B. Blutdruck)
- i = 1,…, n sind die **Beobachtungen** (z.B. Patienten)
- X_k (k = 1,…, r) sind die **Merkmale** (z.B. Alter, Cholesterin, Gewicht, …)
- Achtung!
  - M**ultiple Regression**: ein vorhersagtes Merkmal Y
  - **Multivariate Regression**: mehrere vorhergesagte Merkmale (Y,Z,...)



Medizinische Fakultät Heidelberg

6. supervidiertes Lernen: Klassifizierung

Carl Herrmann

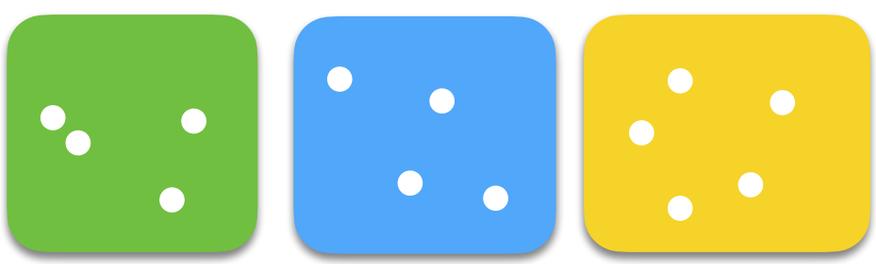
Health Data Science Unit - Medizinische Fakultät Heidelberg

Klassifizierungsalgorithmen

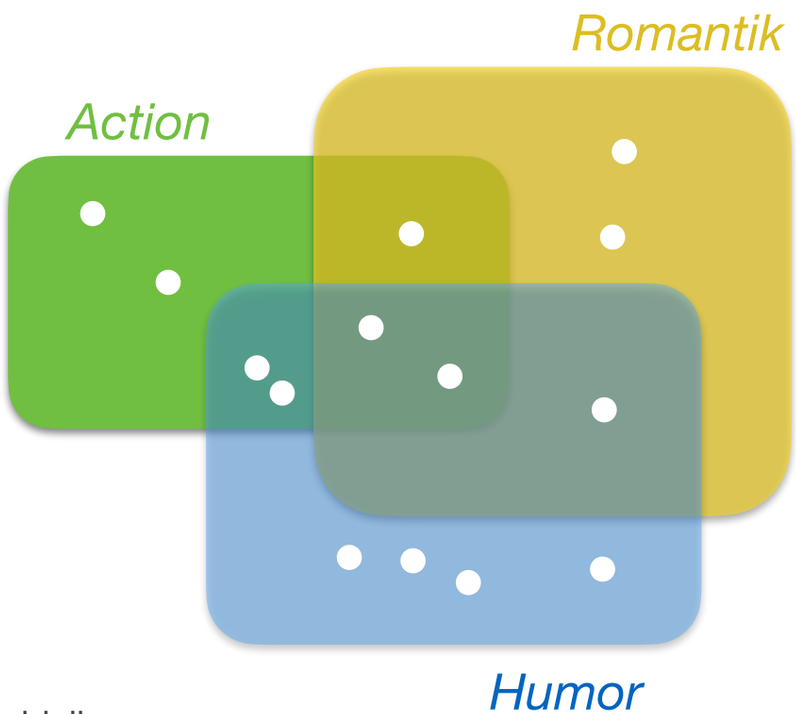
Medizinische Fakultät Heidelberg

- **Binäre Klassifizierung**: nur 2 mögliche Zustände
  Bestanden / nicht-bestanden ; Risikopatient ja / nein

- **Mehrklassen-Klassifizierung**: mehrere **exklusive** Zustände
  Glioblastom: proneural / mesenchymal / IDH
  Risiko: gering / mittel / hoch / sehr hoch



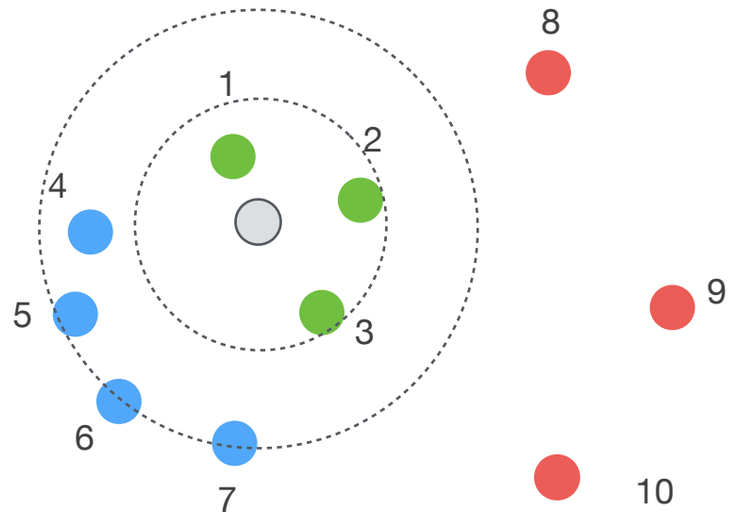
- **Multilabel-Klassifizierung**: jede Beobachtung kann mehrere Zustände gleichzeitig annehmen
  Klassifizerung von Filmen
  "Fluch der Karibik" → Action, Romantik, Humor



k-nearest neighbors

Medizinische Fakultät Heidelberg

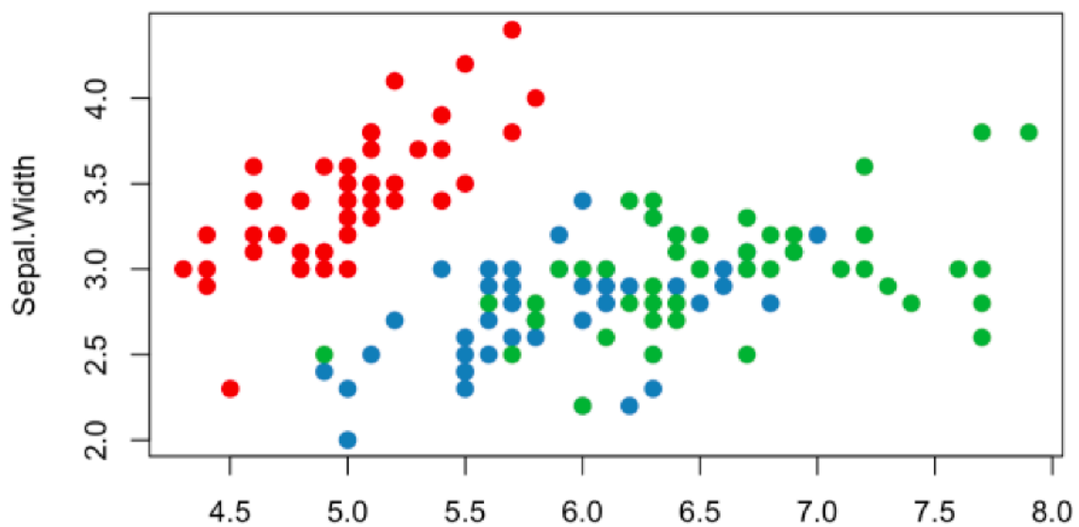
- **Mehrklassen**-Klassifizierung
- Verfahren
  1. Abstände zwischen den Beobachtungen werden anhand einer Metrik bestimmt (euklidische Distanz, Manhattan Distanz,...)
  2. Für jede Beobachtung i, bestimme die k-nächsten Nachbarn
  3. Bestimme die Klasse von i anhand einer **Mehrheitsregel**



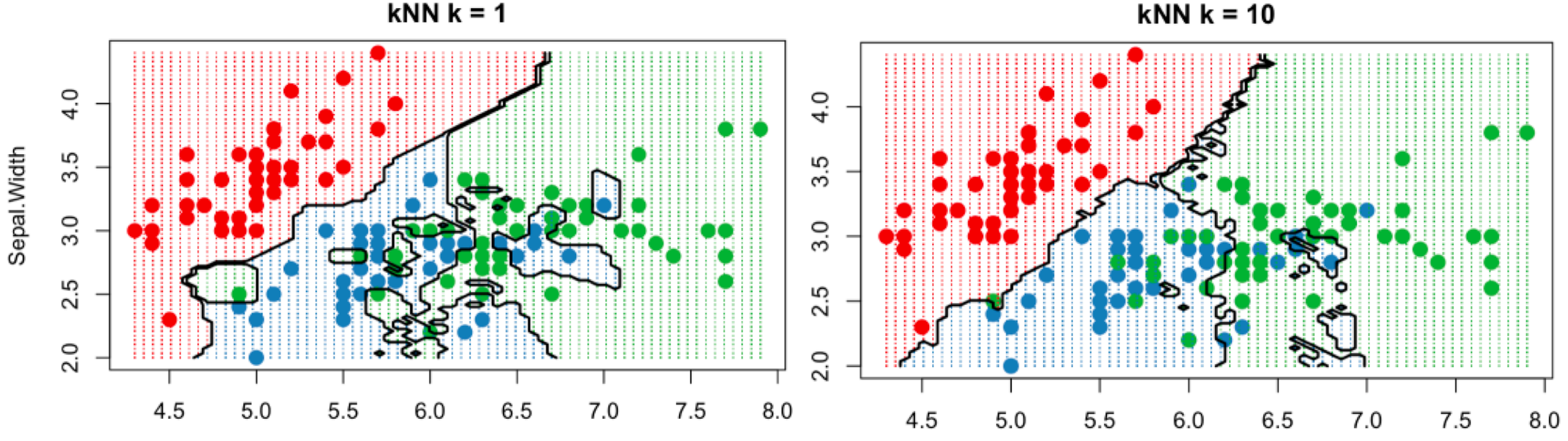
$k = 3$: ○ → ●

$k = 7$: ○ → ●

k-nearest neighbors



Entscheidungsgrenzen



Entscheidungsbäume

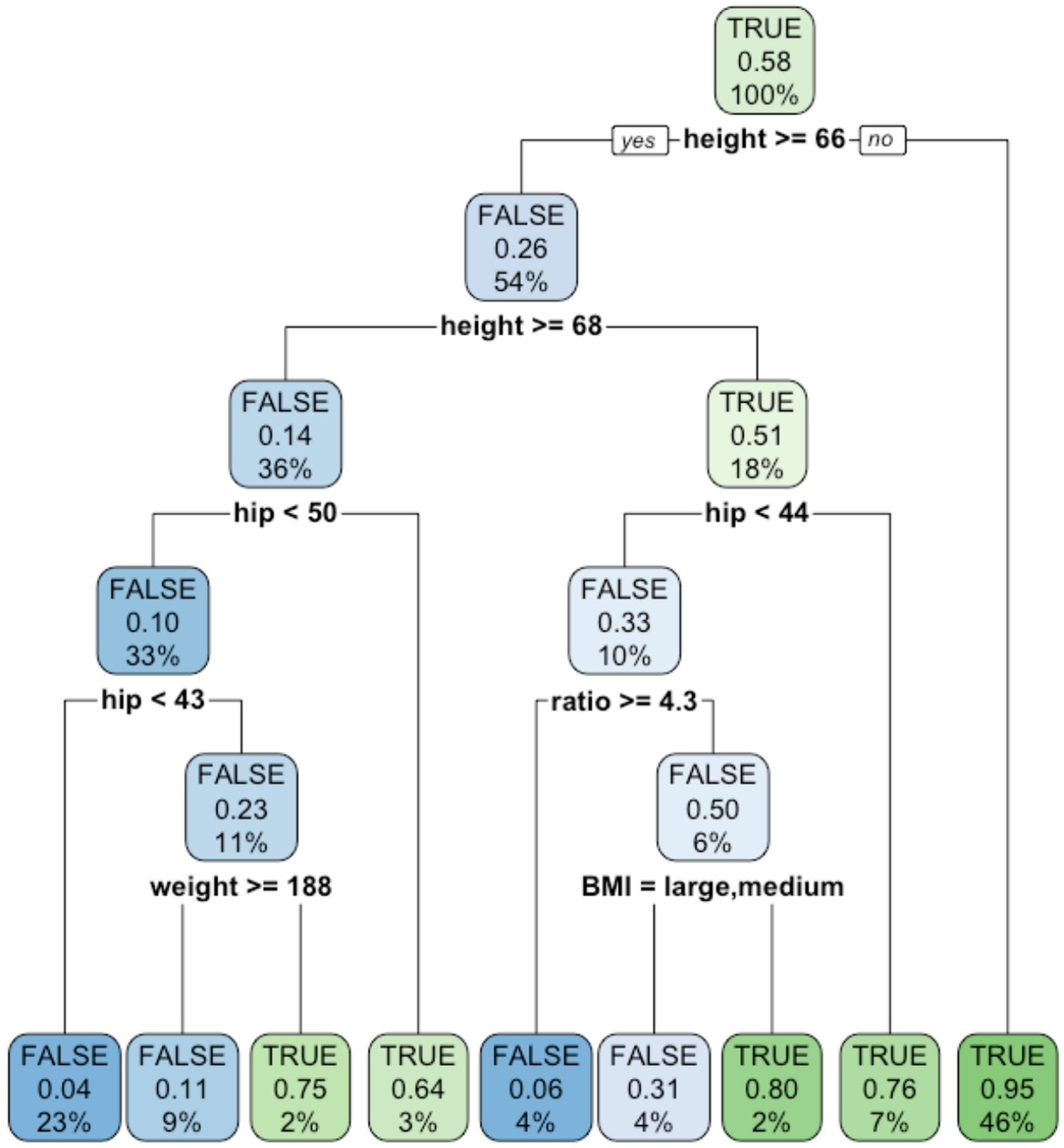
Medizinische Fakultät Heidelberg

- **Entscheidungsbäume** spiegeln das menschliche Verfahren für Klassifizierungsprobleme ab
- **Iteratives Verfahren**, bei dem die Gesamtmenge der Beobachtungen in immer kleinere Menge entsprechend bestimmten Kriterien aufgeteilt wird
- **Vorteile**:
  - sehr einfaches und intuitives Verfahren
  - funktioniert für numerische und kategorische Daten
  - Daten müssen nicht skaliert werden
- **Nachteile**
  - Modelle sind sehr empfindlich gegenüber Änderungen in den Daten: leicht veränderte Daten führen oft zu ganz unterschiedlichen Modellen!
  - Lassen sich daher schwer verallgemeinert ("overfitting"!)

Entscheidungsbäume

Medizinische Fakultät Heidelberg

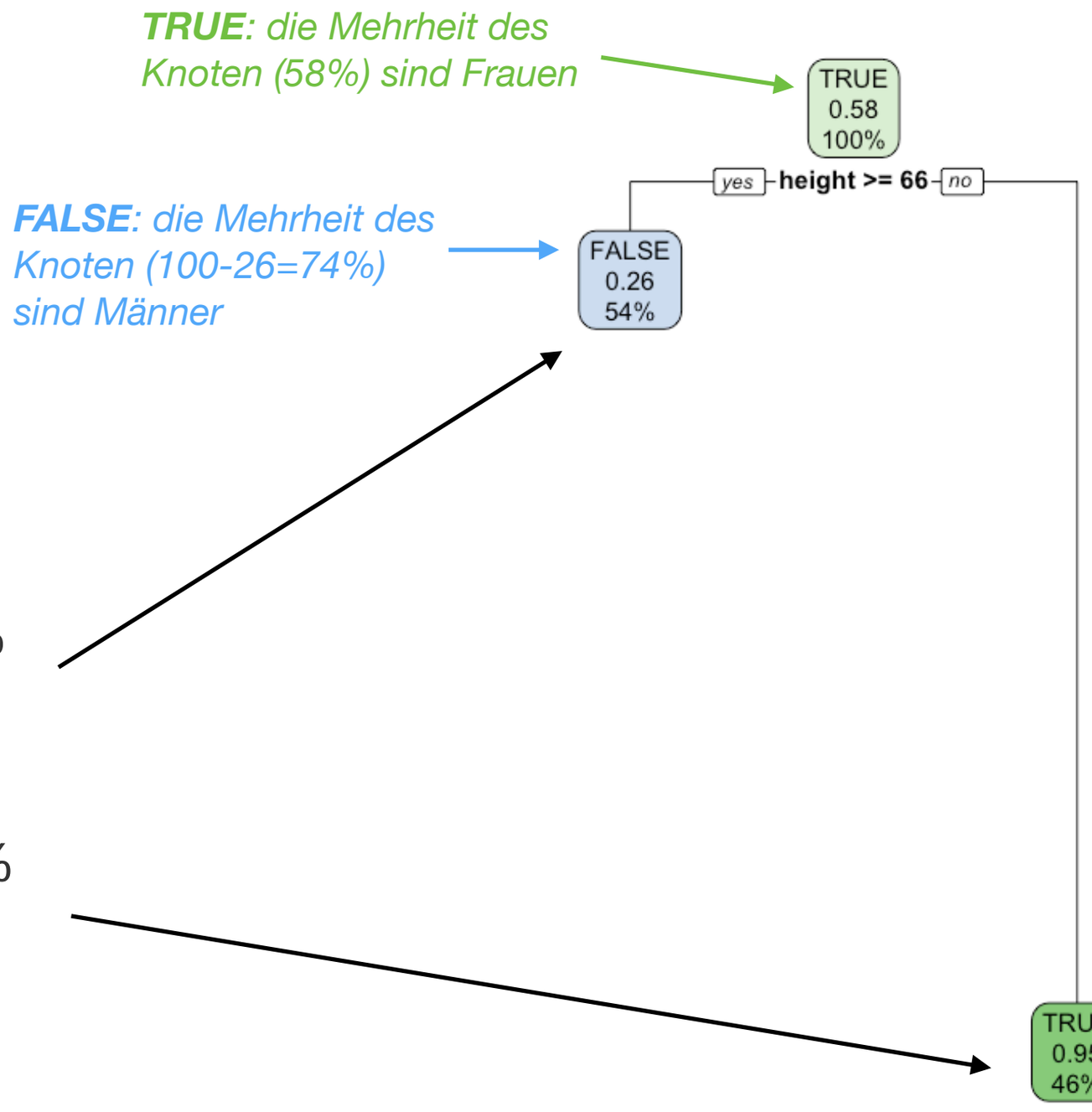
- Bestimmung des Geschlechts anhand klinischer Parameter
  - **Größe**
  - **Hüftumfang**
  - **Gewicht**
  - **Cholesterin/HDL ratio**
  - **BMI (small/medium/large)**
  - Blutdruck
  - Alter
  - Stabilized Glucose
  - Herkunftsort
  - ....
- Prinzip: Gesamtkohorte wird entsprechend der Variablen aufgeteilt



Entscheidungsbäume

Medizinische Fakultät Heidelberg

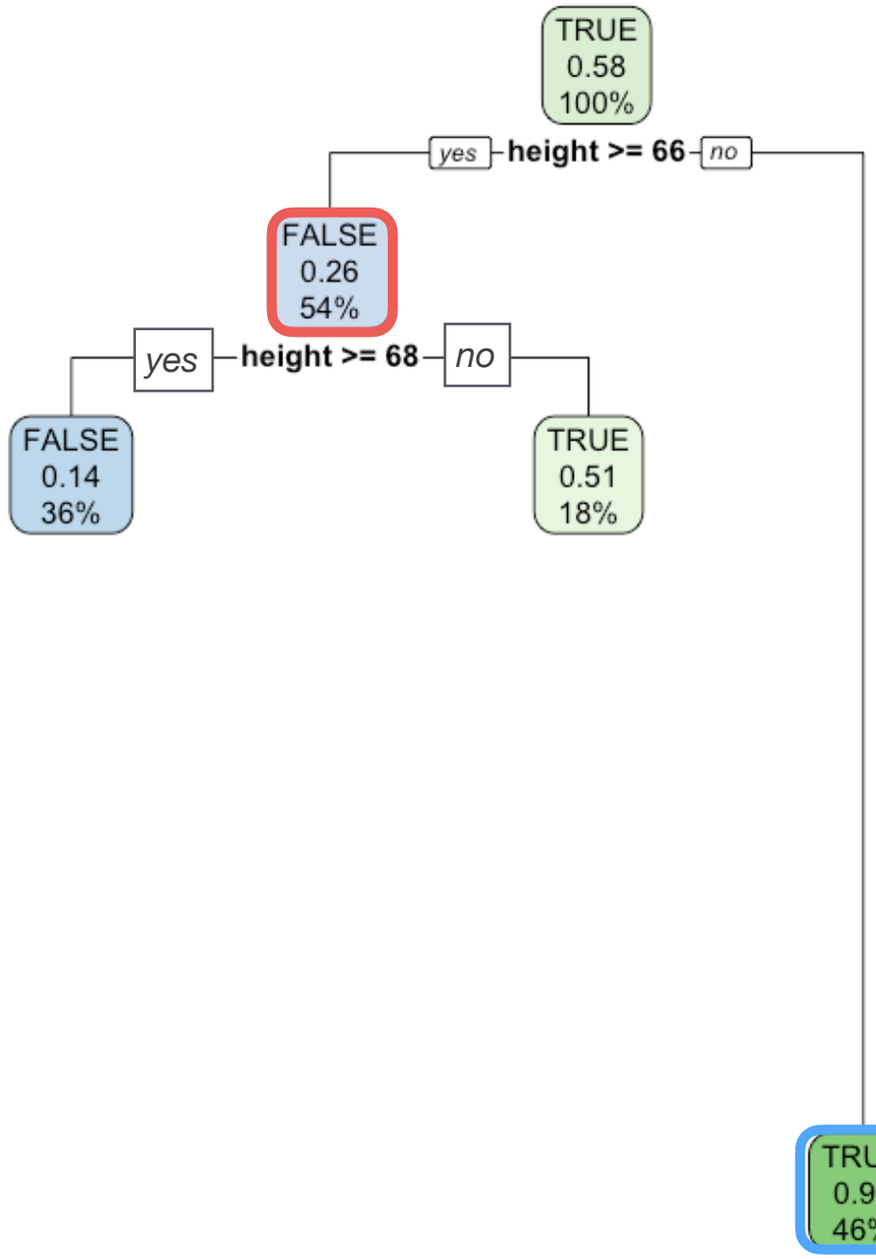
- **Schritt 1**: Aufteilung der Kohorte ensprechend dem Kriterium "Körpergröße >= 66 inches"
- Ursprüngliche Aufteilung
  - 58% Frauen
  - 42% Männer
- Danach:
  - linker Knoten (">=66") : 54% der Patienten, davon 26% Frauen
  - rechter Knoten ("<66") : 46% der Patienten, davon 95% Frauen



Entscheidungsbäume

Medizinische Fakultät Heidelberg

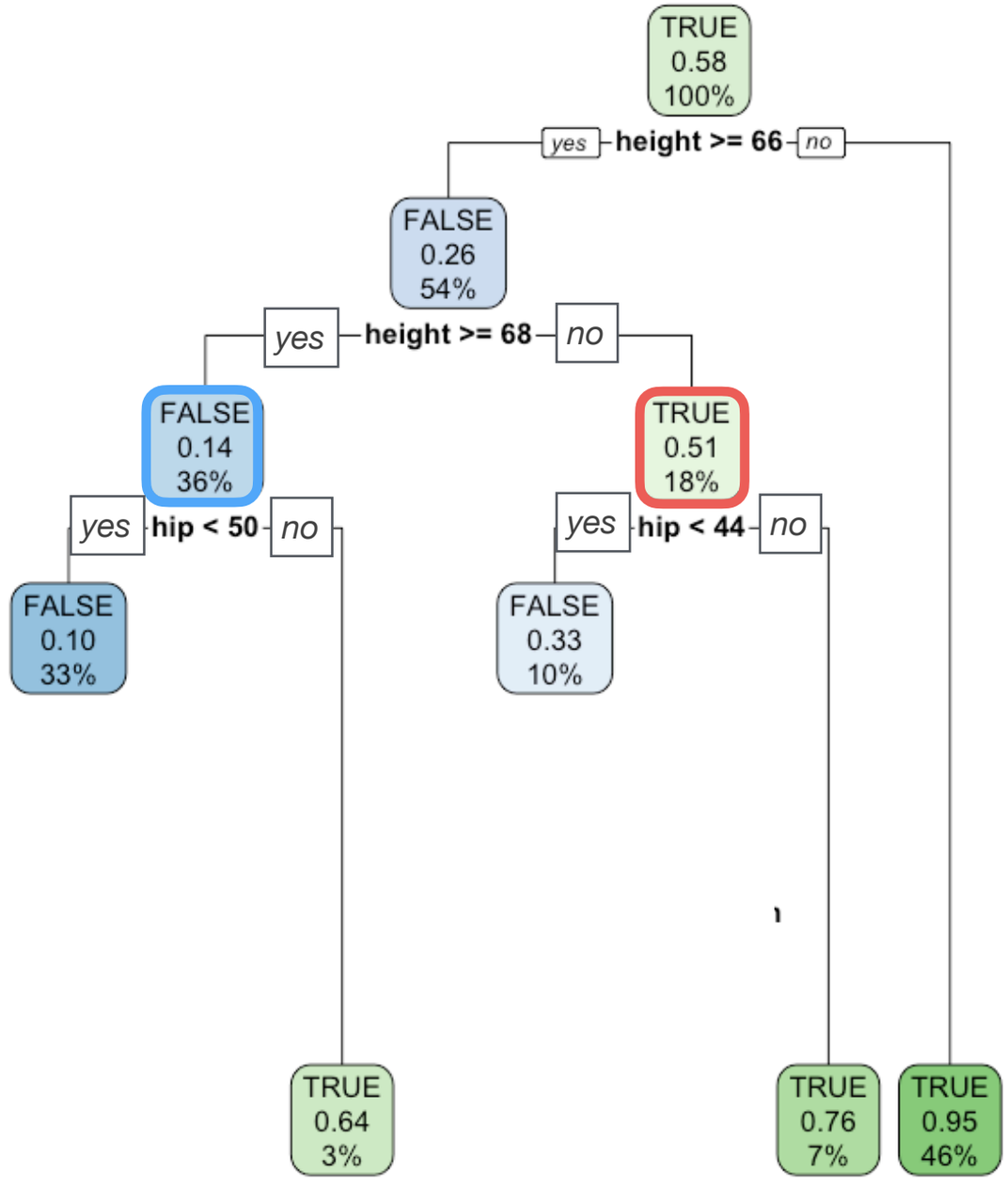
- **Schritt 2**
- **rechter Knoten** kann nicht besser aufgeteilt werden
  → "*leaf node*"
- **linker Knoten** wird aufgeteilt
  "Körpergröße >= 68 inches"
  - ja (linker Unterknoten): 36% der Patienten, davon 14% Frauen
  - nein (rechter Unterknoten): 18% der Patienten, davon 51% Frauen



Entscheidungsbäume

Medizinische Fakultät Heidelberg

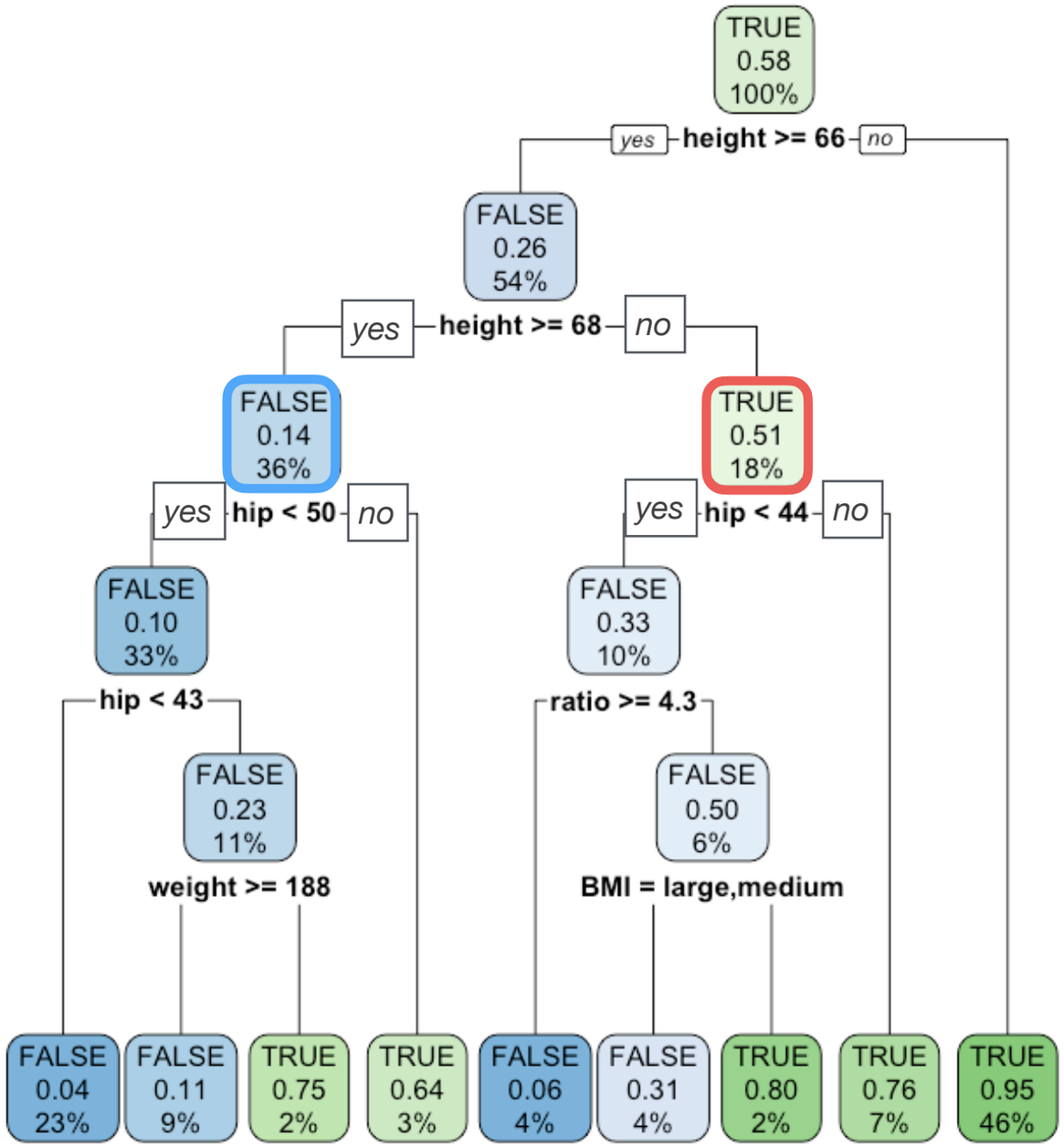
- **Schritt 3**
- **rechter Knoten** wir aufgeteilt
  "Hüftumfang < 44 inches"
  - nein (rechter Unterknoten): 7% der Patienten, davon 76% Frauen
  - ja (linker Unterknoten): 10% der Patienten, davon 33% Frauen
- **linker Knoten** wird aufgeteilt
  "Hüftumfang < 50 inches"
  - ja (linker Knoten): 33% der Patienten, davon 10% Frauen
  - nein (rechter Knoten): 3% der Patienten, davon 64% Frauen



Entscheidungsbäume

Medizinische Fakultät Heidelberg

- **kompletter Durchgang**
- folgende Variablen wurden benutzt
  - Größe
  - Hüftumfang
  - Gewicht
  - Cholesterin/HDL ratio
  - BMI (small/medium/large)
- Einige davon wurden mehrfach benutzt
  - Größe
  - Hüftumfang



Frage 1
wie wird das Teil-Kriterium augewählt?

..ledizinische Fakultät Heidelberg

- Variable (und Schwellenwert), die die "**Reinheit**" der Unterknoten am meisten erhöht
  - maximale Reinheit: nur eine Klasse vorhanden ("Frauen")
  - maximale Unreinheit: perfekte Mischung (50% Männer, 50% Frauen)
- Maß (je kleiner, desto reiner)
  - Gini-Index:

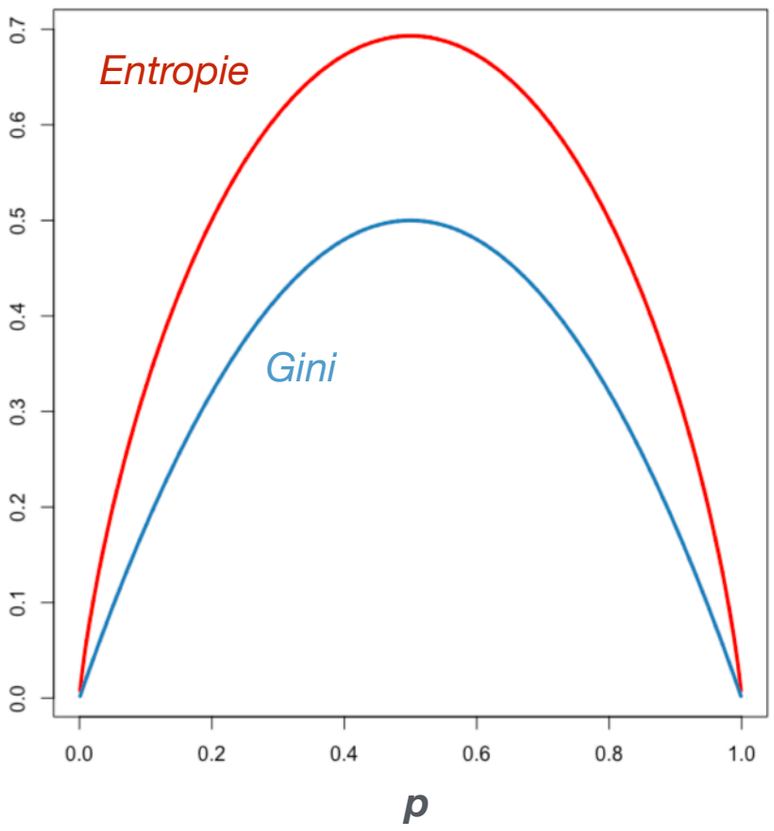
$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2$$

  - Entropie:

$$H_i = -\sum_{k=1}^{n} p_{i,k} \log(p_{i,k})$$

Anteil der Klasse *k* am Knoten *i*

Verhalten für 2 Klassen

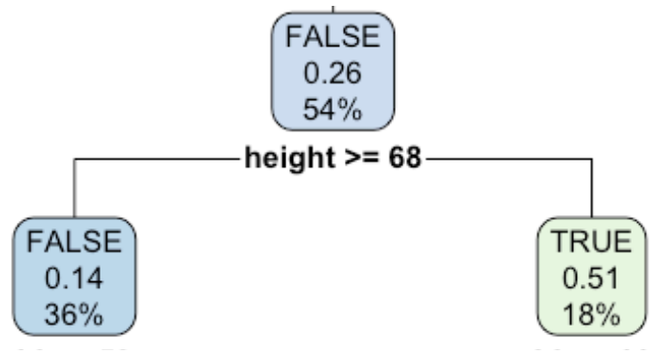


Gini-Koeffizient: Beispiel

Medizinische Fakultät Heidelberg

- Elternknoten
  - 26% Frauen, 74% Männer
  - Gini: $G = 1 - (0.26)^2 - (0.74)^2 = 0.3848$
- Linker Unterknoten
  - 14% Frauen, 86% Männer
  - Gini: $G_L = 1 - (0.14)^2 - (0.86)^2 = 0.2408$
- Rechter Unterknoten
  - 51% Frauen, 49% Männer
  - Gini: $G_R = 1 - (0.51)^2 - (0.49)^2 = 0.4998$
- Gewichteter Gini decrease (je kleiner, desto besser):

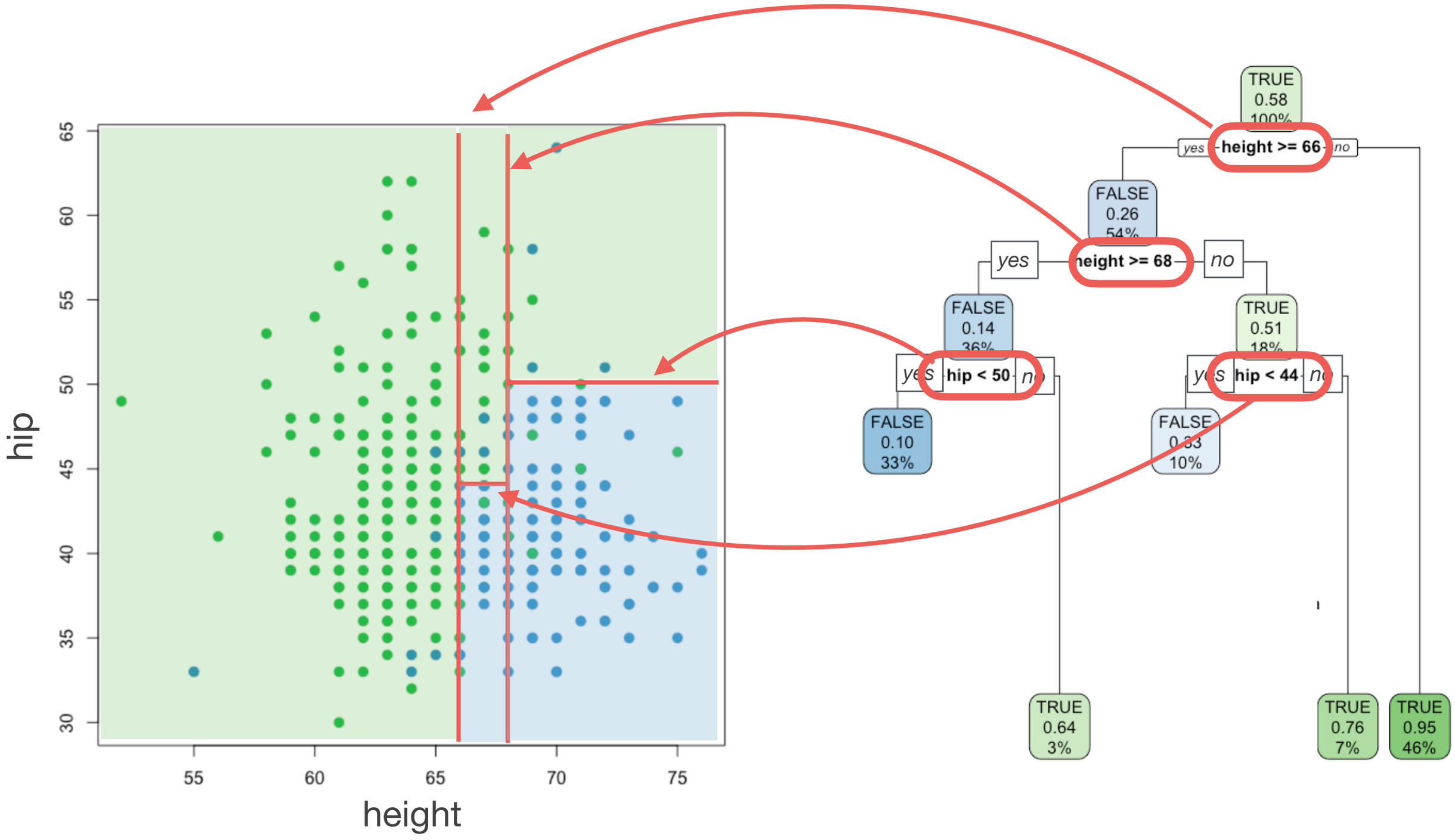
$$\Delta G = G - \frac{2}{3} G_L - \frac{1}{3} G_R = 0.0577$$



2/3 der Patienten des Elternknoten

1/3 der Patienten des Elternknoten

kein anderes Kriterium würde zu einer so starken Verbesserung der purity führen!


hip
height
65
60
55
50
45
40
35
30
55
60
65
70
75
TRUE
0.58
100%
yes
height >= 66
no
FALSE
0.26
54%
yes
height >= 68
no
FALSE
0.14
36%
yes
hip < 50
no
TRUE
0.51
18%
yes
hip < 44
no
FALSE
0.10
33%
FALSE
0.33
10%
TRUE
0.64
3%
TRUE
0.76
7%
TRUE
0.95
46%

Frage 2:
wie lange wird aufgeteilt?

Medizinische Fakultät Heidelberg

- Das Teilen der Knoten könnte so lange weitergeführt werden, **bis jeder Knoten perfekt rein ist** (evt. nur noch 1 Datenpunkt enthält):
  - perfekte Klassifizierung auf dem Trainingsdatensatz
  - aber: würde sich schlecht auf weitere Datensätze verallgemeinern lassen! ("overfitting")
- Daher werden Bedingungen gesetzt, wann das Teilen gestoppt wird:
  - **maximale Tiefe** des Baumes erreicht; oder
  - Jeder Knoten erreicht eine **minimale Anzahl an Datenpunkten**; oder
  - Teilung eines Knotens würde zu **keiner signifikanten Verbesserung** der Reinheit führen.
- Diese zusätzlichen Parameter nennt man **Hyperparameter**; sie müssen während des Trainings optimiert werden

Frage 3:
wie kann das Modell benutzt werden?

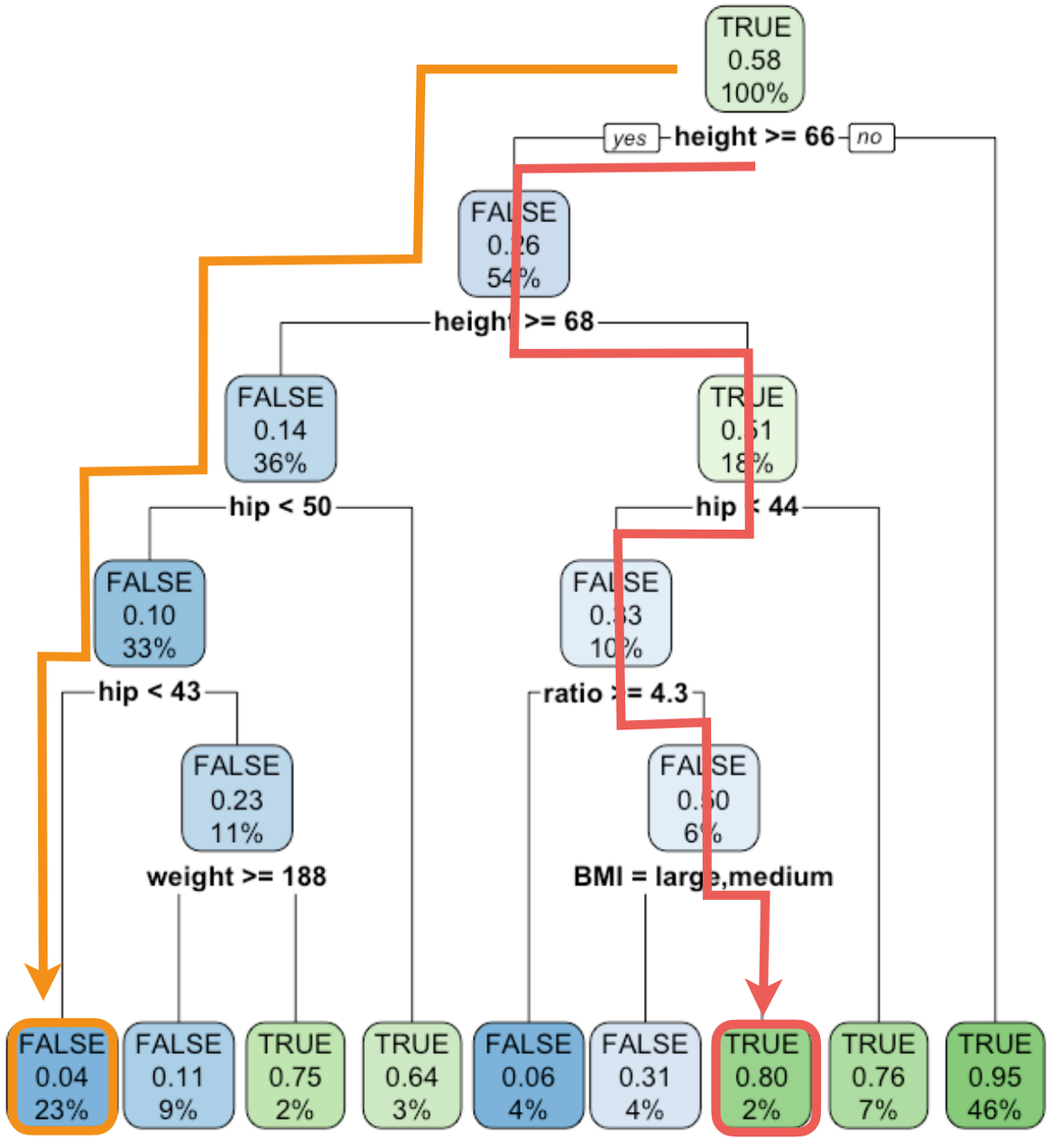
Medizinische Fakultät Heidelberg

- **Mehrheitsregel!**
- **Patient 423:**
  - Größe = 67
  - Hüftumfang = 43
  - Gewicht = 180
  - ratio = 3.8
  - BMI = small

  → **Frau**
- **Patient 677:**
  - Größe = 69
  - Hüftumfang = 51
  - Gewicht = 180
  - ratio = 3.8
  - BMI = small

  → **Mann**

diese Informationen wurden in diesem Fall nicht benutzt!

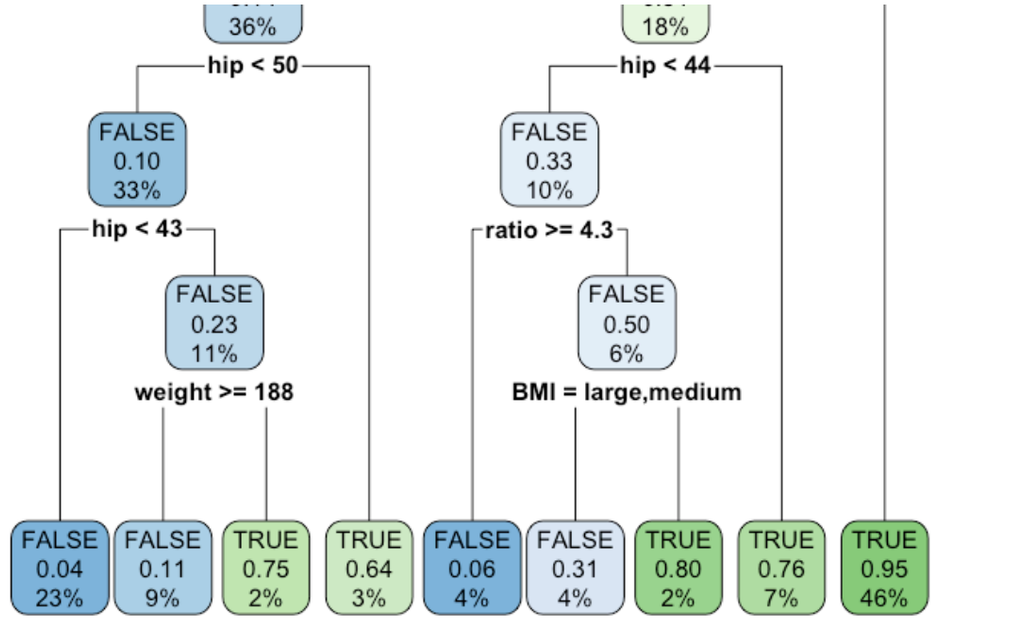


Frage 4:
wie gut funktioniert das Modell?

Medizinische Fakultät Heidelberg

- Wie gut funktioniert es auf dem **Trainingsdatensatz**?

Vorhersage	Wahrer Wert: Frau	Wahrer Wert: Mann
Frau	220	14
Mann	25	144



4% der Patienten in diesem "Männer-Knoten" sind Frauen

20% der Patienten in diesen "Frauen-Knoten" sind Männer

Klassifizierungsfehler: $\frac{14 + 25}{403} = 0.0968$

Genauigkeit: $\frac{220 + 144}{403} = 0.9032$

Einige der Datenpunkte aus dem Trainingsdatensatz werden falsch klassifiziert!

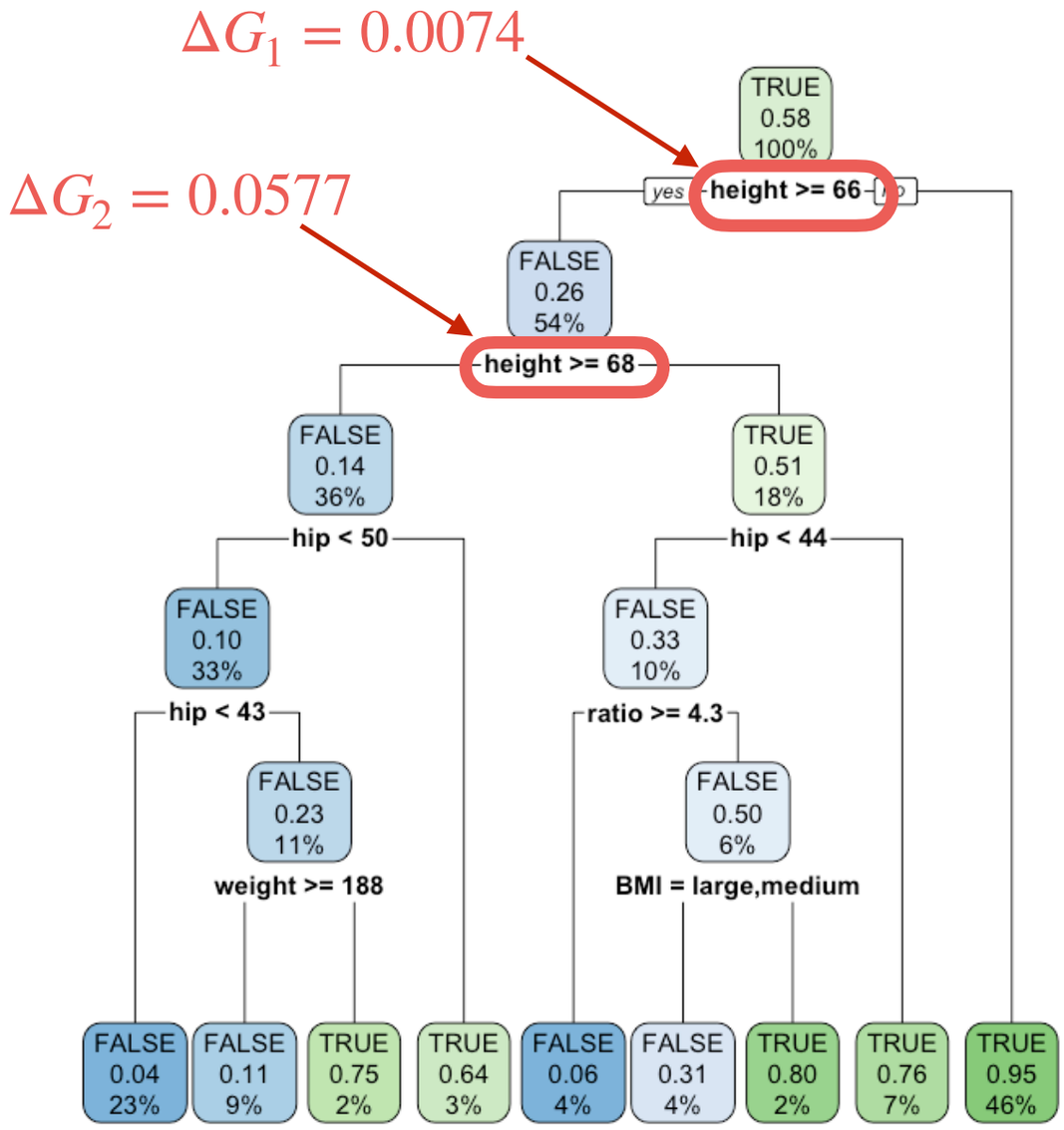
Frage 5:
was sind die wichtigen Variablen?

Medizinische Fakultät Heidelberg

- Entscheidungsbäume sind **"white-box"** Modelle
- Die Vorhersagekraft der Merkmale kann bestimmt werden
- Durschnittliche Verbesserung der Purity an den Knoten, an denen ein Merkmal verwendet wurde:

$$\Delta G_{height} = 1 \cdot \Delta G_1 + 0.54 \cdot \Delta G_2$$

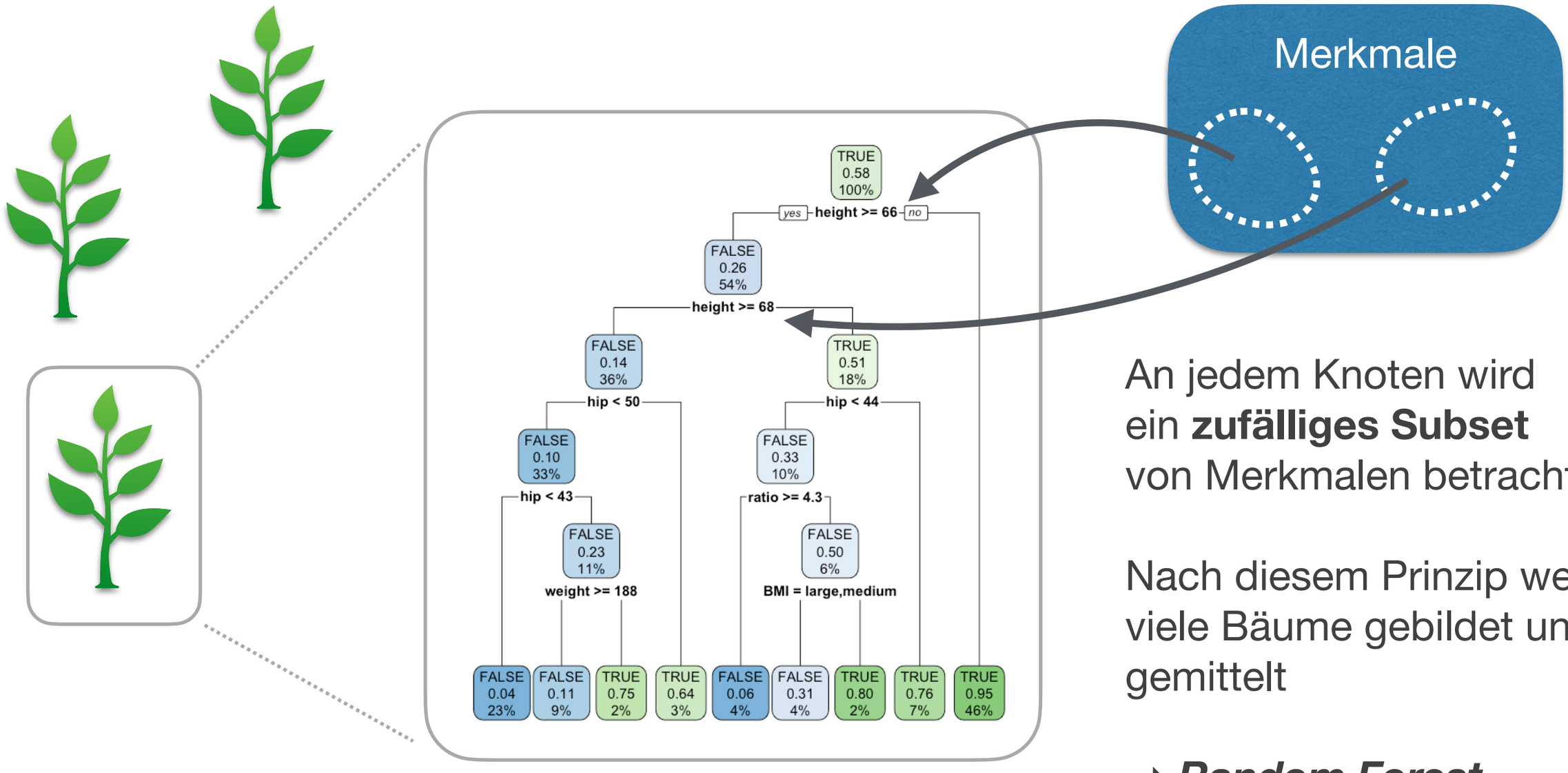
Gewichtung nach Knotengröße



Alle benutzten Merkmale
können so untersucht werden → Ranking

Emsemble Methoden

- Entscheidungsbäume haben eine Tendenz zum Overfitting
- Lösung: über viele Bäume mitteln → **Random Forest**



An jedem Knoten wird ein **zufälliges Subset** von Merkmalen betrachtet

Nach diesem Prinzip werden viele Bäume gebildet und gemittelt

→ ***Random Forest***

Medizinische Fakultät Heidelberg

7. Zusammenfassung

Take-home messages

Medizinische Fakultät Heidelberg

- Unterscheidung supervidiertes/nicht-supervidiertes Lernen
  - **supervidiert**: Labels teilweise bekannt
  - **nicht-supervidiert**: Labels nicht-bekannt
- Lernverfahren sollen eine **Kostenfunktion minimieren** (oder Scorefunktion maximieren)
  - root-mean-square error / mean average error bei Regression
  - accuracy / precision / recall bei Klassifizierung
- **Varianz/Verzerrung** Dilemma
  - **underfitting** (zu einfaches Modell) → hohe Verzerrung
  - **overfitting** (zu komplexes Modell) → hohe Varianz; Modell lässt sich nicht verallgemeinern
- Benutzung von Trainings- / Validierungs- / Test-Datensatz
  - k-fold cross-Validierung
  - LOOCV

Medizinische Fakultät Heidelberg

Klassen / Zielwerte *Y* bekannt?

nein

Verfahren sollte ...

... explorativ sein

... Gruppen vorhersagen

Dimensions-Reduktion

- Hauptkomponenten Analyse
- Manifold learning
- t-SNE
- ...

Clustering

- hierarchisches Clustering
- k-means
- Gaussian mixture model

Nicht-supervidierte Verfahren

ja

Y beschreibt ...

... Klassen (z.B. Tumortypen)

... stetige Werte (z.B. Cholesterin)

Klassifizierung

- k-nearest neighbors
- Entscheidungsbäume
- Random Forest
- Support Vector Machine (SVM
- ...

Regression

- (nicht) lineare Regression
- Random Forest Regression
- SVM
- ...

Supervidierte Verfahren