Einführung in Data Science

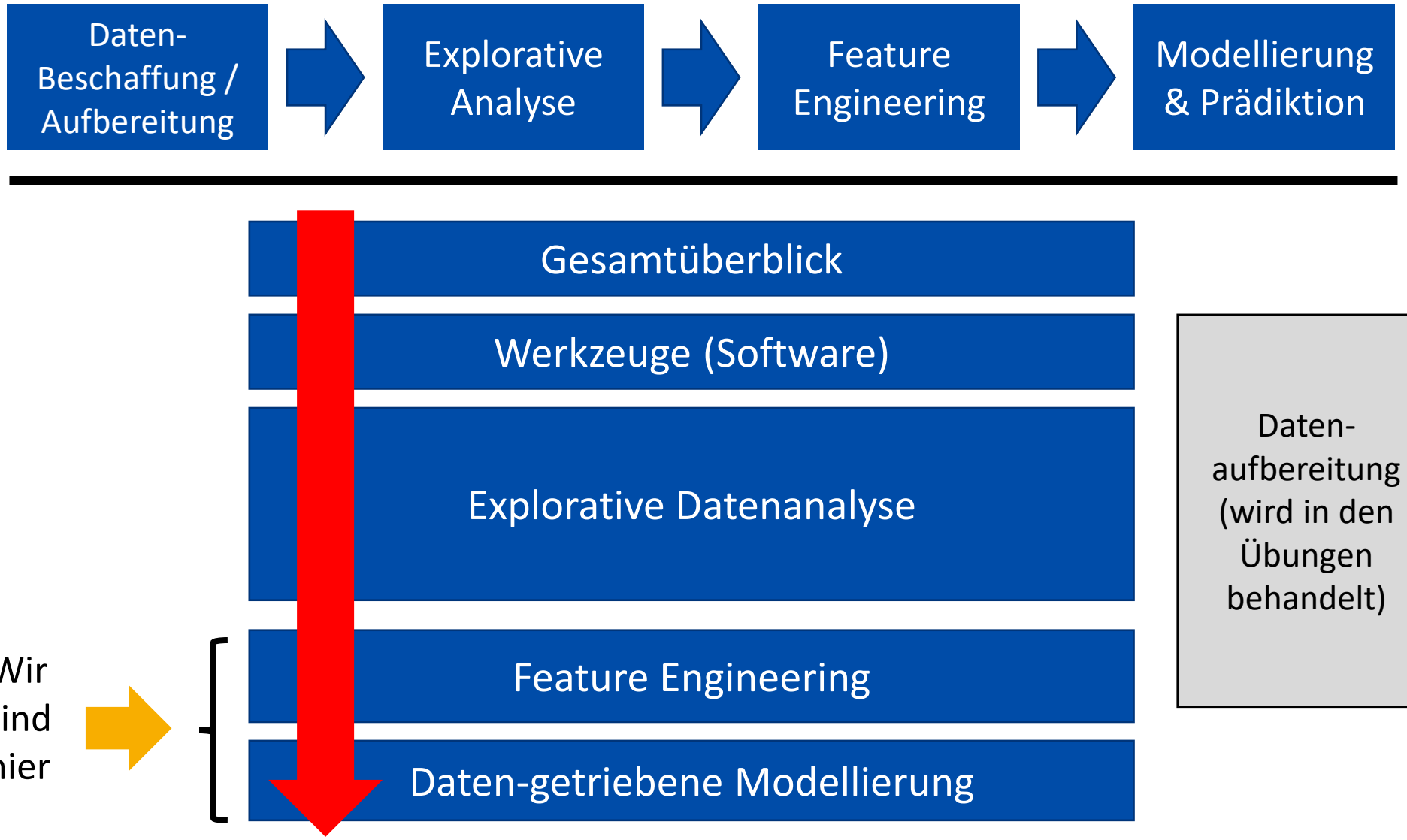
Unser Plan für heute:

1. Feature Engineering

2. Daten-getriebene Modellierung
    1. Schwellwert-basierte Klassifikation (binäre Klassifikation)
    2. Gütemaße für binäre Klassifikation
    3. Fallbeispiel Anfallsvorhersage
    4. Gütemaße für Multiklassen-Klassifikation
    5. NN Modell

Klausurinformation. Falls Sie einen FH-Laptop zum Schreiben Ihrer Klausur benötigen, melden Sie sich bei Niklas Grieger (grieger@fh-aachen.de). **Nennen Sie den Standort, an dem Sie Ihre Klausur schreiben!**

Data Science



Curriculum

1. Grundbegriffe / Überblick
2. Zentrale Softwarebibliotheken

3. Univariate explorative Analyse (EDA) Visualisierung (Teil 1)
4. Visualisierung (Teil 2), Multivariate explorative Analyse (Teil 1)
5. Multivariate explorative Analyse (Teil 2)
6. Dimensionsreduktion (Teil 1): PCA
7. Dimensionsreduktion (Teil 2): MDS, Isomap
8. Clustering: K-Means, HCA
9. Clustervalidierung

10. Probeklausur

11. Feature Engineering, Datengetriebene Modellierung (Teil 1)
12. Datengetriebene Modellierung (Teil 2)

Überblick / Begriffe

Explorative Analyse (EDA)

Feature Engineering & Modellierung

Feature Engineering

auf deutsch: Merkmalskonstruktion

- der Prozess, mithilfe von **Domänenwissen** Merkmale aus Daten zu erzeugen, um die daten-getriebene Modellierung (und damit Vorhersagen) zu ermöglichen.

> Feature Engineering vermittelt zwischen Daten und Modellen.

Feature Engineering ist

- meist zeitaufwändig
- oft entscheidend für den Erfolg eines Machine Learning Projektes
- domänenspezifisch
  (statt eines eigenen Wissensgebiets „Feature Engineering“ gibt es Gebiete wie z.B. „Natural Language Processing“ oder „Bildverarbeitung“ oder „Zeitreihenanalyse“)
  unterschiedliche Datenarten → unterschiedliche Features

Feature Engineering

Typisches Vorgehen (eines Data Scientist) beim Feature Engineering:

a) Klassische Methoden der Explorativen Analyse (EDA) werden genutzt
(z.B. deskriptive Statistik, Dimensionsreduktion, Cluster)

Dies machen wir in der verbleibenden Vorlesung.

b) Domänenexperten aufsuchen und über Daten und ihre wichtigsten Eigenschaften befragen

Dies ist meine Empfehlung an Sie:
Suchen Sie immer das Gespräch mit den Domänenexperten, sofern möglich!

c) Selbst zum Domänenexperten für bestimmte Datenarten werden (typischerweise während der Berufsausübung oder wissenschaftlichen Ausbildung)

Stichwort: Spezialisierung.

Feature Engineering vs Explorative Datenanalyse

	Zielsetzungen
Explorative Datenanalyse:	Kennenlernen der Daten und Entwicklung von Fragen
Feature Engineering:	Merkmale für die *Daten-getriebene Modellierung* erstellen, die wiederum Vorhersagen ermöglicht

---

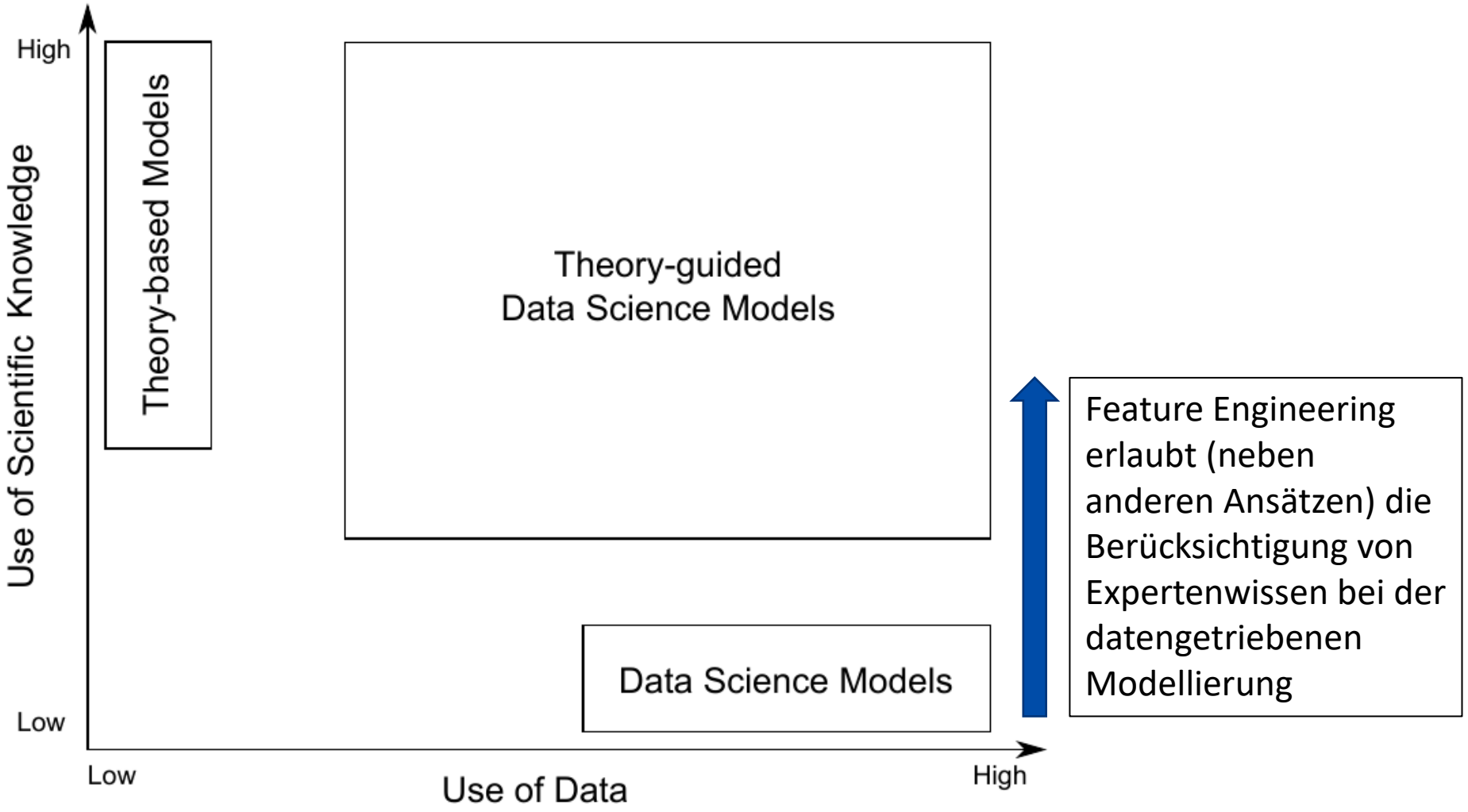
Daten-getriebene Modellierung
Erstellung einer mathematischen Beschreibung mithilfe von Daten

Beispiele: Machine Learning Modelle (z.B. Bild-/ Spracherkennung)

Theorie-basierte Modellierung
Erstellung einer mathematischen Beschreibung mithilfe von Grundprinzipien (*first principles*)

Beispiele: Physikalische Modelle (z.B. Wettervorhersage)

Daten- und Theorie-basierte Modellierung



Daten-getriebene Modellierung | Vorhersagen

- Modelle ermöglichen Vorhersagen

Typen von Vorhersagen daten-getriebener Modelle:

	Klassifikation	**Regression**
Ergebnis der Vorhersage:	Kategorie / Klasse	numerischer Größe
Modell wird auch genannt:	Klassifikator (*classifier)*	Regressor (*regressor)*

Aktivität

Nennen Sie den Typ der Vorhersage für folgende Fälle:

- Vorhersage von Nettomieten
- Vorhersage von epileptischen Anfällen
- Erkennung (Vorhersage) handgeschriebener Zahlen
- Vorhersage von Aktienkursen

Daten-getriebene Modellierung | Klassifikation

binäre Klassifikation

- Modell unterscheidet zwischen *zwei* Klassen
- einfachstes Modell: Schwellwert-basierte Klassifikation

Beispiel

Vorhersage des Geschlechts „Mann“ anhand von Daten des *National Health and Nutrition Examination Survey (NHANES, USA)* aus den Jahren 2009-2010. Kennzahlen[1] (summary statistics):

	Min	25%	Median	75%	Max
Age	241	418	584	748	959
Weight	32.4	67.2	78.8	92.6	218.2
Height	140	160	167	175	204
Leg Length	23.7	35.7	38.4	41	55.5
Arm Length	29.5	35.5	37.4	39.4	47.7
Arm Circumference	19.5	29.7	32.8	36.1	141.1
Waist	59.1	87.5	97.95	108.3	172

1) Kennen Sie aus Vorlesung 3.

Schwellwert-basierte Klassifikation

Beispiel (Fortsetzung)

Feature Engineering: Suche nach Merkmalen, die sich zwischen Frauen und Männern unterscheiden

Merkmal: Körpergrößen

Modell:

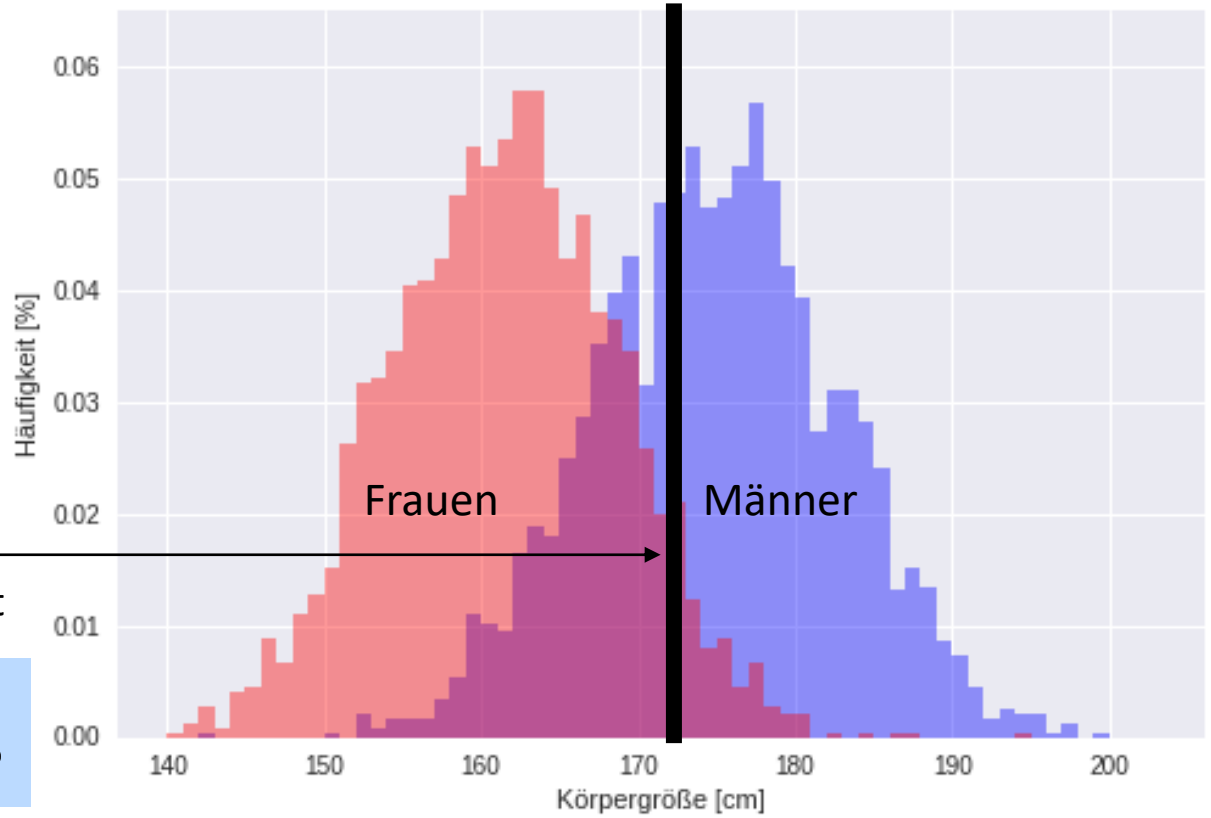
$$f : X \to \{0, 1\}$$

Frau Mann

$$f(x) = \begin{cases} 1 & \text{falls } x > \theta \\ 0 & \text{sonst} \end{cases}$$

Schwellwert

4797 Menschen zwischen 20-80 Jahren



Frage:
Wie wählen Sie den Schwellwert?

Beurteilung eines Klassifikators | Grundkonzepte

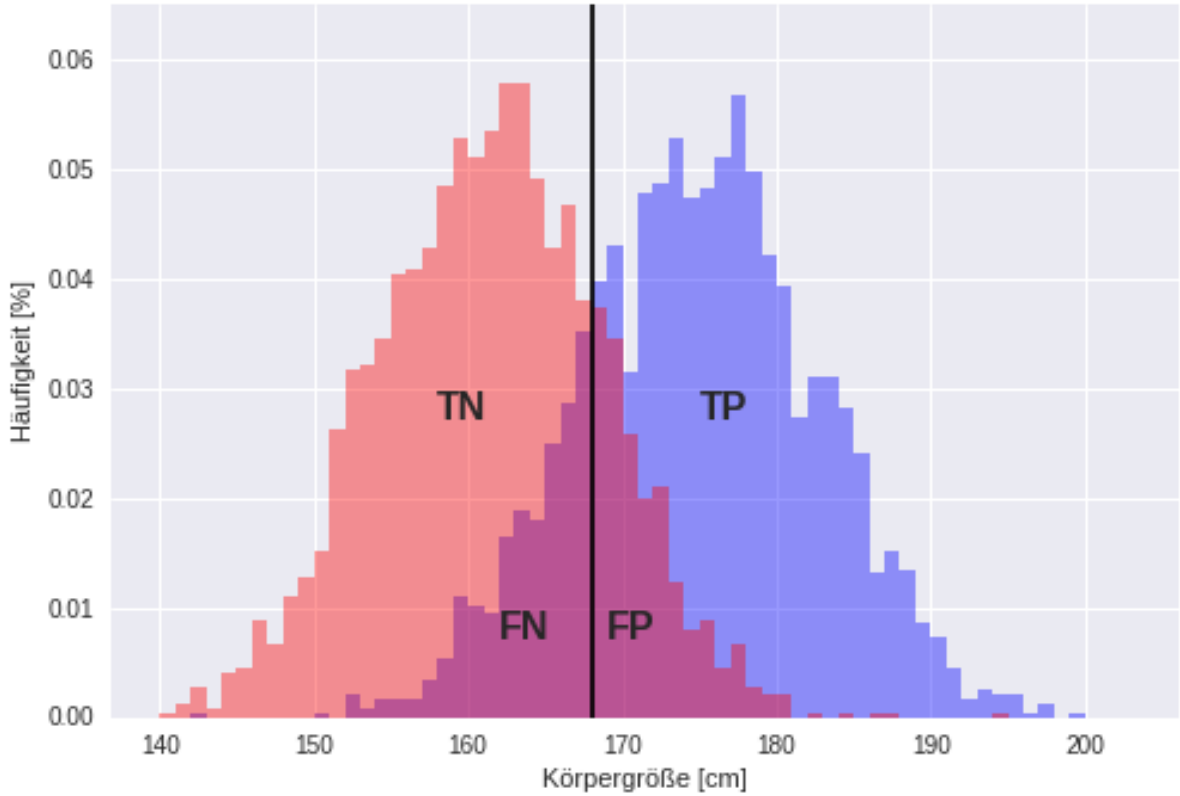
P: Anzahl der Datenpunkte der vorherzusagenden Klasse (*Positiv*)
(in unserem Beispiel: Klasse „Mann“)
N: Anzahl Datenpunkte, die *nicht* der vorherzusagenden Klasse entsprechen (*Negativ*)

TP: Anzahl der korrekt vorhergesagten Positiven (***T**rue **P**ositives*)

TN: Anzahl der korrekt vorhergesagten Negativen (***T**rue **N**egatives*)

FP: Anzahl der Falsch-Positiven (***F**alse **P**ositives*)

FN: Anzahl der Falsch-Negativen (***F**alse **N**egatives*)



Beurteilung eines Klassifikators | Gütemaße

Gütemaße zur Beurteilung von Klassifikatoren

- … setzen sich aus Grundkonzepten der vorherigen Folie zusammen.

Gütemaß

Accuracy
(dt: Genauigkeit)

$$\mathrm{ACC} = \frac{TP+TN}{TP+TN+FN+FP}$$

Anzahl der richtig klassifizierten Datenpunkte dividiert durch Gesamtzahl aller Datenpunkte

Aktivität F

Nehmen Sie an, dass 50% Ihrer Daten der Klasse „Frauen“ und die restlichen 50% der Klasse Männer angehören. Zusammen mit Ihrem Banknachbarn diskutieren Sie folgende ***Baseline Models*** (Basismodelle):

1. Welche *Accuracy* erreichen Sie mit einem Modell, das für jeden Datenpunkt zufällig eine Klasse vorhersagt? → 0.5 (50%)
2. Welche Accuracy erreichen Sie mit einem Modell, das immer nur eine Klasse (z.B. Frau) vorhersagt? → 0.5 (50%)

Beurteilung eines Klassifikators | Gütemaße

Baseline Models:

- einfache Modelle für die Einschätzung von Beurteilungsmetriken
- helfen bei der Frage: „Wie gut ist mein (oft mühsam konstruiertes) Modell gegenüber einem einfachen, schnell erzeugten *Baseline Model*?“

Aktivität **F**

Ihr Datensatz bestehe aus Daten von Krebspatienten. Vorhergesagt werden soll die Krebserkrankung (Klasse *P*). 95% der Patienten im Datensatz haben kein Krebs (Klasse *N*); 5% sind an Krebs erkrankt. Diskutieren Sie folgende *Baseline Models* (Basismodelle):

1. Welche *Accuracy* erreichen Sie mit einem Modell, das für jeden Datenpunkt zufällig eine Klasse vorhersagt? → 0.5 (50%)
2. Welche Accuracy erreichen Sie mit einem Modell, das immer nur Klasse *N* (kein Krebs) vorhersagt? → 0.95 (95%)

Beurteilung eines Klassifikators | Gütemaße

Accuracy: einfaches Gütemaß

problematisch: ungleich großen Klassen (*class imbalance)*

(Beispiel: Krebsdetektion auf vorheriger Folie)

→ Wir benötigen weitere Maße!

Precision: $\mathrm{PPV} = \frac{TP}{TP+FP}$

(auch bekannt als
***P**ositive **P**redictive **V**alue*,
positiver prädiktiver Wert)

Anzahl der Richtig-Positiven dividiert durch Anzahl aller als positiv deklarierten Punkte.

Dieses Maß „bestraft“ Falsch-Positive.

Frage F

Der Datensatz enthalte 5% Krebskranke (P) und 95% Gesunde (N).

- Welche PPVs erhalten Sie mit einem Modell, das alle Punkte als Positiv (Krebs) oder Negativ (Nicht Krebs) vorher sagt?

→ Fall P: 0.05, Fall N: nicht definiert (Division durch 0)

Beurteilung eines Klassifikators | Gütemaße

Recall:

$$\mathrm{TPR} = \frac{TP}{P} = \frac{TP}{TP+FN}$$

Anzahl der Richtig-Positiven dividiert durch alle Positiven.

auch bekannt als
***T**rue **P**ositive **R**ate*
(Richtig-Positiv-Rate)
bzw. *Sensitivity*

Dieses Maß „bestraft“ Falsch-Negative.

Oft werden *Precision* und *Recall* in einem Maß zusammengefasst:

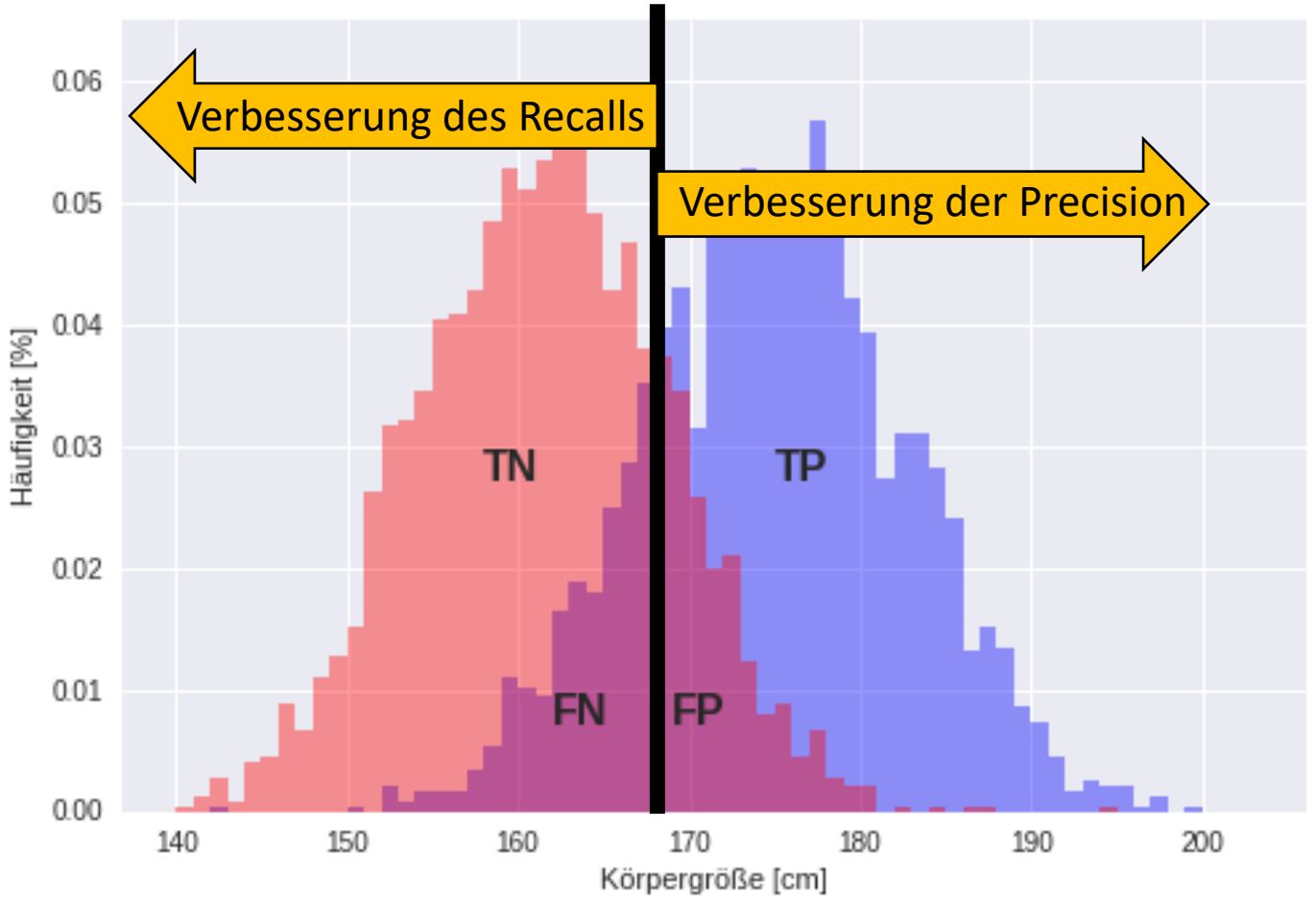
F1-Score: harmonisches Mittel aus *Precision* und *Recall*

$$F_1 = \left(\frac{\mathrm{recall}^{-1}+\mathrm{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precision}\cdot\mathrm{recall}}{\mathrm{precision}+\mathrm{recall}} = 2\frac{\mathrm{PPV}\cdot\mathrm{TPR}}{\mathrm{PPV}+\mathrm{TPR}} \in [0,1]$$

Je größer der F1-Score, desto besser ist der Klassifikator.

Beurteilung eines Klassifikators | Gütemaße

Beispiel (Fortsetzung)



Recall: $\mathrm{TPR} = \frac{TP}{P} = \frac{TP}{TP+FN}$

Precision: $\mathrm{PPV} = \frac{TP}{TP+FP}$

Beurteilung eines Klassifikators | Gütemaße

Anwendungsfall entscheidet über Relevanz eines Maßes.

Beispiel: Krebsdiagnostik.
Was ist wichtiger? Große Zahl Richtig-Positiver (TP)
oder kleine Anzahl Falsch-Positiver (FP)?
→ Abwägungsfrage

Viele Klassifikationsprobleme verfügen über einstellbare (freie) Parameter, mit denen sich Abwägungen zwischen Richtig-Positiv-Rate und Falsch-Positiv-Raten einstellen lassen.

Typisches Gütemaß: **ROC** (Receiver Operating Characteristic)

basiert auf:

True Positive Rate $\mathrm{TPR} = \frac{TP}{P}$

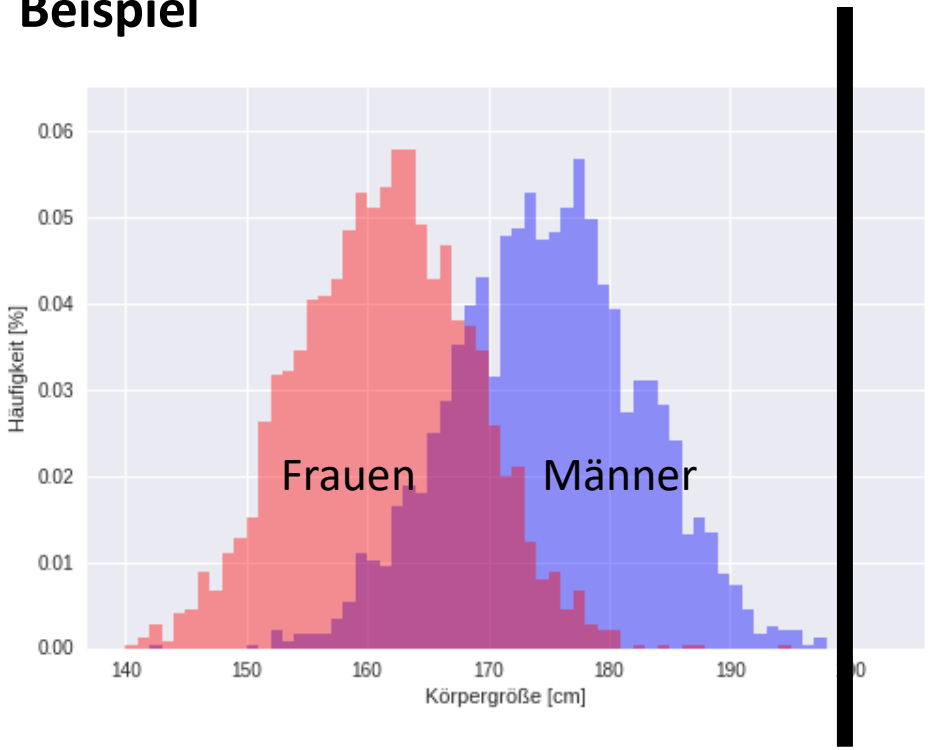
False Positive Rate $\mathrm{FPR} = \frac{FP}{N}$

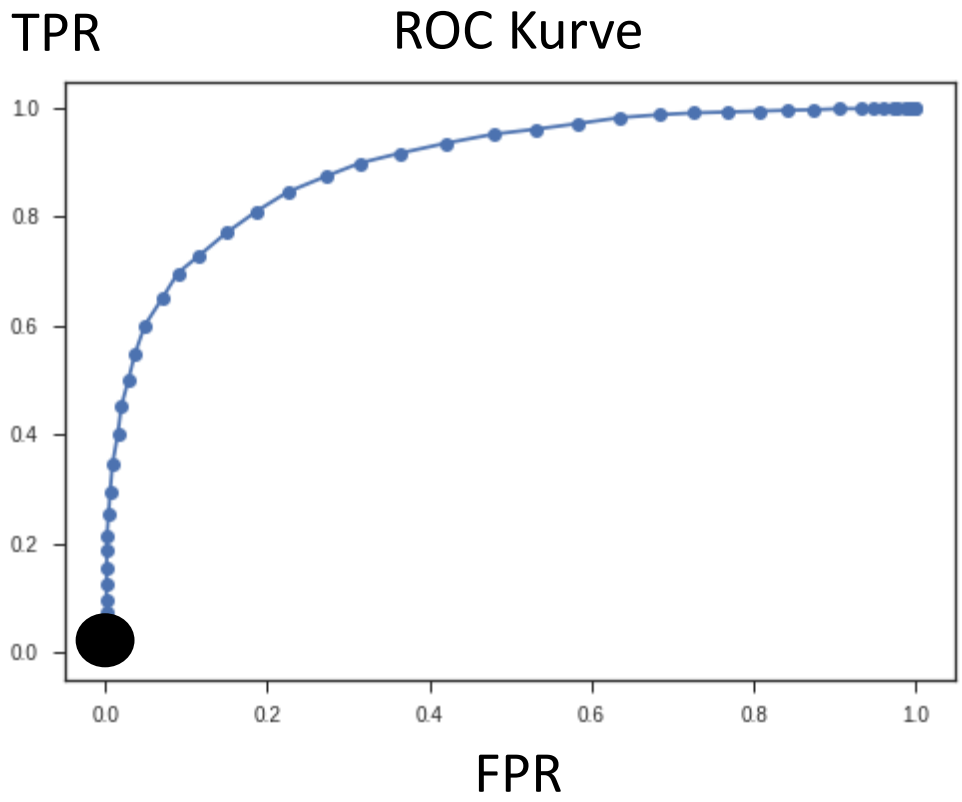
Anzahl Falsch-Positiver dividiert durch Anzahl aller Negativer

Gütemaße | Receiver Operating Characteristic (ROC)

ROC Kurve: TPR gegen FPR auftragen, während freier Parameter (Schwellwert) variiert wird

Beispiel





ROC Kurve charakterisiert, wie gut beide Verteilungen durch Schwellwert trennbar sind.

Gütemaße | Receiver Operating Characteristic (ROC)

Frage
Welche Kurve erhalten wir, wenn beide Verteilungen identisch sind (und daher nicht durch Schwellwert trennbar)?

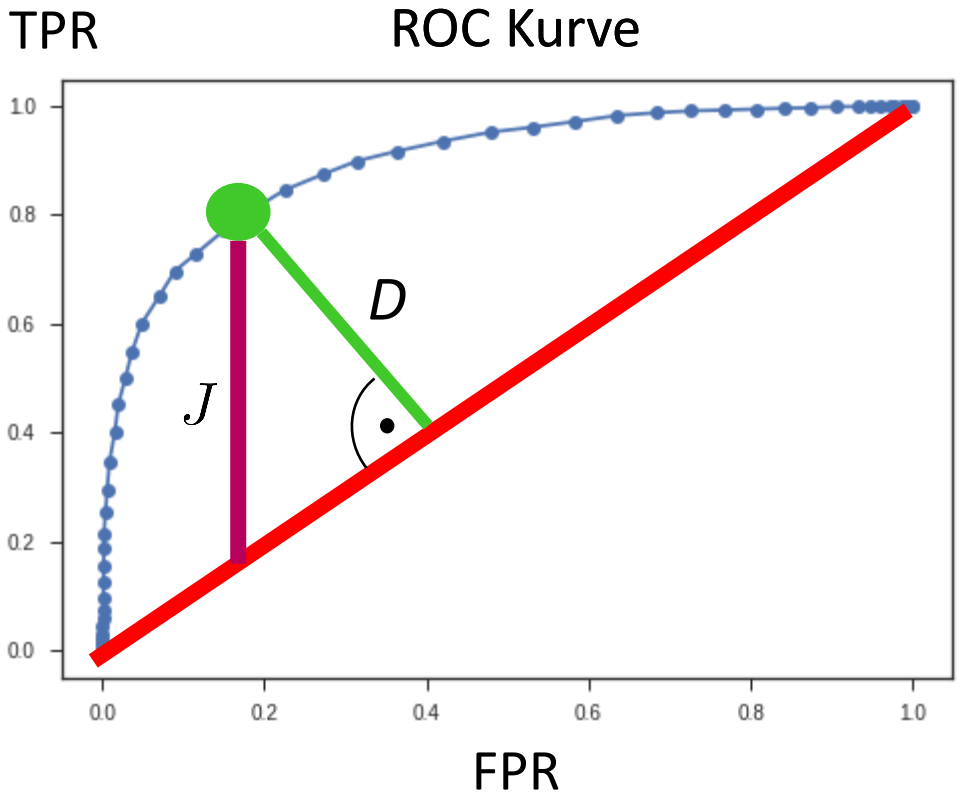
F

Raumdiagonale – entspricht **Zufallsprädiktor** bzw. nicht-trennbaren Verteilungen.

Frage
Welcher Punkt auf der ROC Kurve entspricht dem besten Klassifikator?

F

Punkt, der die größte Distanz *D* zur Diagonalen hat.



Youden Index $J = \mathrm{TPR} - \mathrm{FPR}$
wird an diesem Punkt maximal.

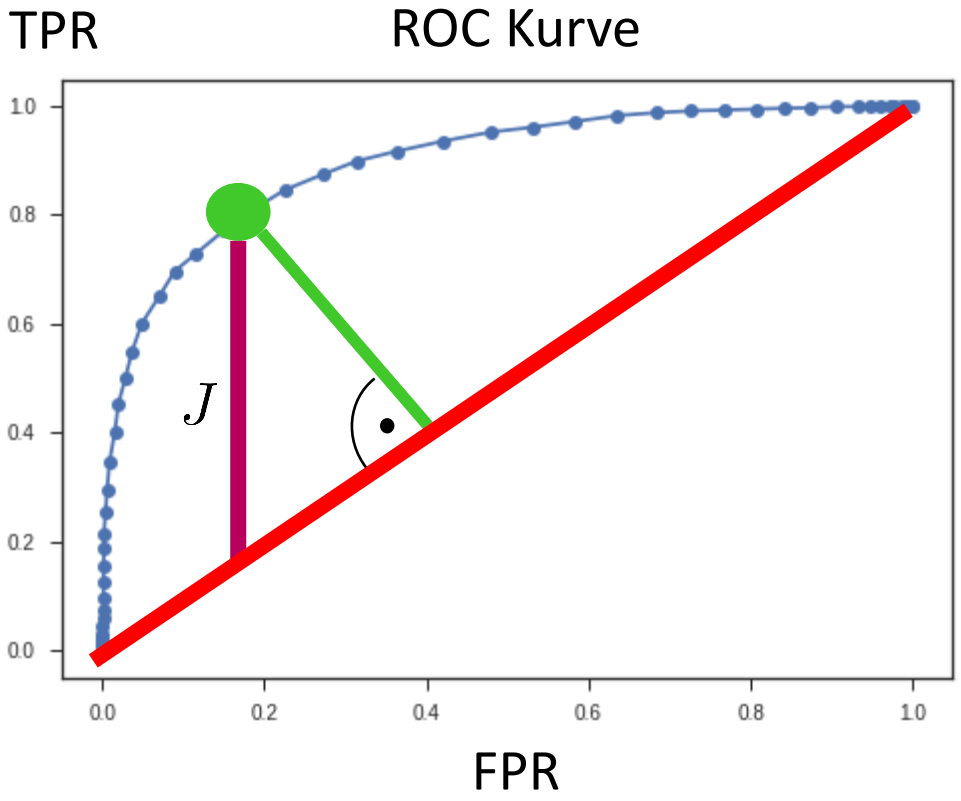
Gütemaße | Receiver Operating Characteristic (ROC)

Wahl des Schwellwerts über Youdens *J*:

$$\tilde{\theta} = \arg\max_{\theta}(J(\theta))$$

Je nach Anwendungsfall Wahl anderer Schwellwerte wenn

- hohe TPR wichtiger als niedrige FPR
- niedrige FPR wichtiger als hohe TPR

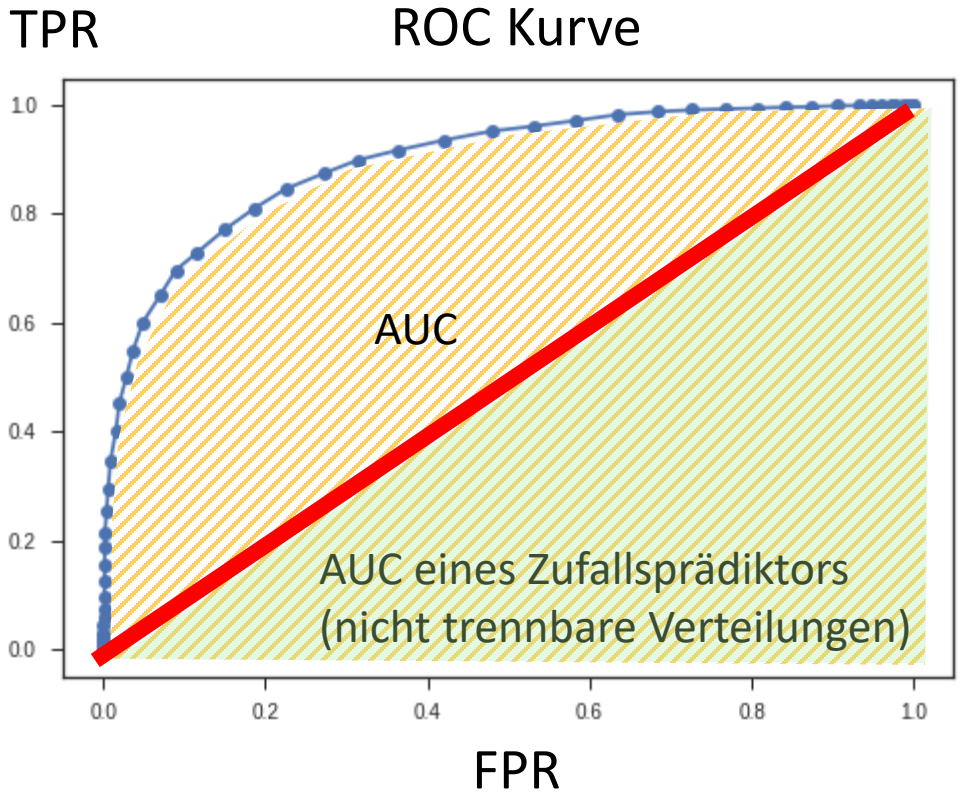


1) benannt nach William John Youden, amerikanischer Statistiker , 1900—1970

ROC | Area Under the ROC Curve (AUC)

AUC: Fläche unter der ROC Kurve
(**A**rea **U**nder the ROC **C**urve)

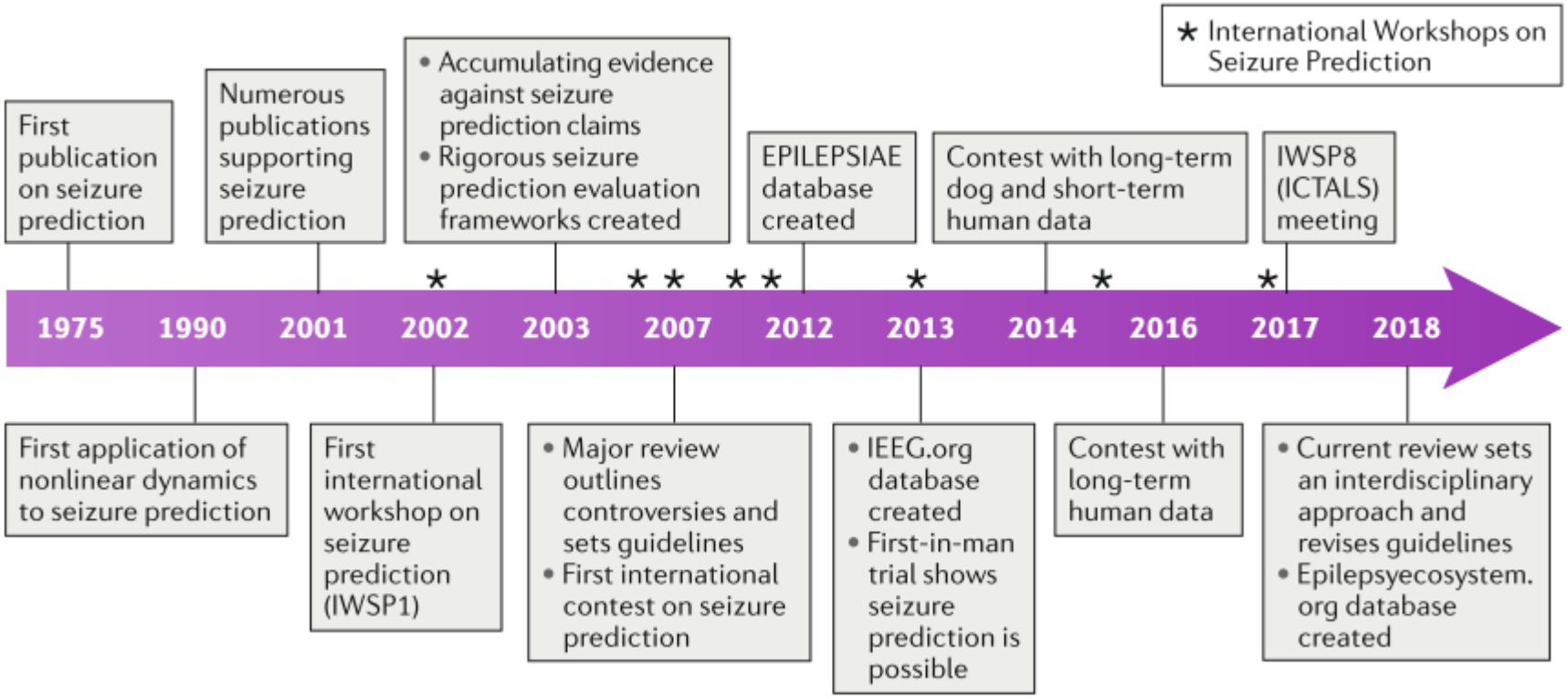
- Maß für die Güte eines Klassifikators
- Oft genutzt, um verschiedene binäre Klassifikatoren zu vergleichen
- Variiert zwischen 0.5 (Zufallsprädiktor) und 1 (perfekter Prädiktor/Klassifikator).



Fallbeispiel epileptische Anfallsvorhersage

Aus der ersten Vorlesungsstunde:

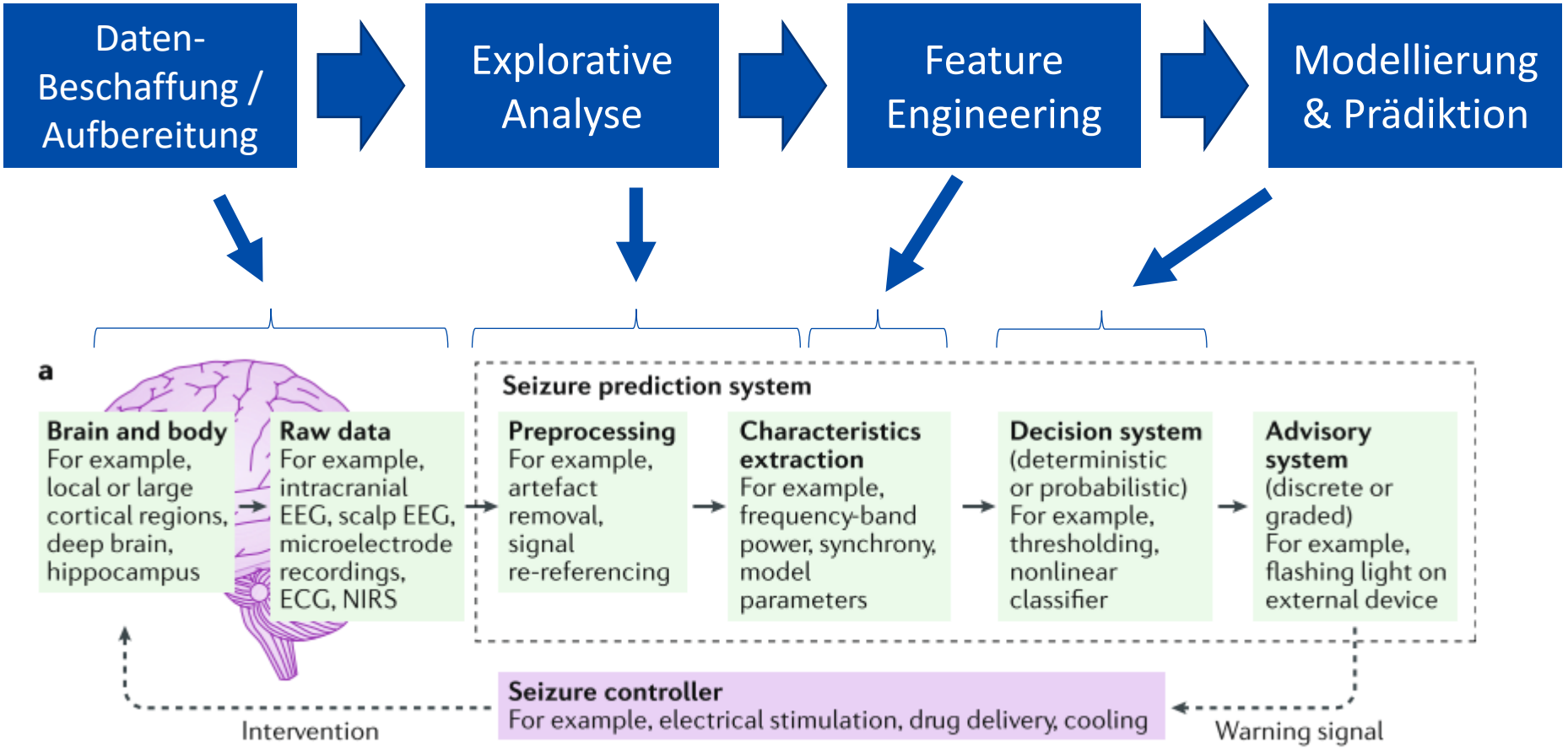
- 1% der Weltbevölkerung leidet unter epileptische Anfälle
- **Fernziel**: Vorhersage epileptischer Anfälle (*seizure prediction*) (mittels implantierten Elektroden und Datenanalyse)



Quelle: Kuhlmann et al, Nature Reviews Neurology 14, 618-630, 2018

Fallbeispiel epileptische Anfallsvorhersage

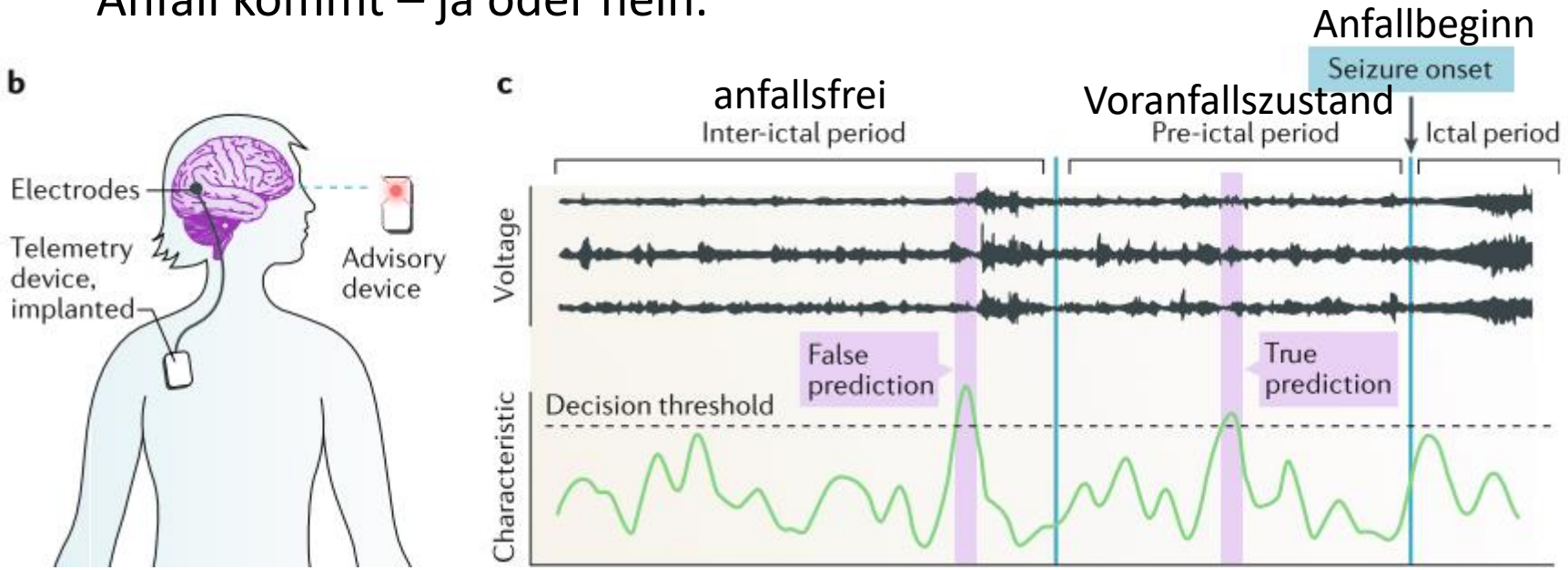
Die Data Science Schritte finden Sie in der Analysekette wieder:



Quelle: Kuhlmann et al, Nature Reviews Neurology 14, 618-630, 2018

Fallbeispiel epileptische Anfallsvorhersage

- EEG-Zeitreihen werden in Fenster (windows) aufgeteilt.
- Für jedes Fenster sagt ein Modell die **Wahrscheinlichkeit** für einen Anfall voraus.
- **Binärer Klassifikator**: Wahrscheinlichkeiten werden mit Schwellwert-basiertem Modell in eine Vorhersage umgewandelt: Anfall kommt – ja oder nein.



Quelle: Kuhlmann et al, Nature Reviews Neurology 14, 618-630, 2018

Fallbeispiel epileptische Anfallsvorhersage

Aktivität

Sie erhalten einen Auszug aus einem Paper mit Ergebnisse eines Anfallvorhersage-Wettbewerbs (mit internationalen Teams).

- Finden Sie heraus, was die besten AUC-Werte waren, die die Teams erreichen konnten.

F

Table 2 **AUC scores for the held-out data experiment compared to scores on the public and private leader boards**

Team	Window (overlap, s)	Features	Machine learning algorithm	Ensemble method	Public leader board	Private leader board	Held-out data	Per cent change	Sensitivity at 75% specificity
Team A notsorandomanymore (1st place)	20, 30, 50 (0)	Spectral power, distribution statistics, AR error, fractal dimensions, Hurst exponent, Riemannian autocorrelation[a,b], cross-frequency coherence, correlation, other features[c,d]	Extreme gradient boosting[e], k-nearest neighbours, generalized linear model, linear SVM[f]	Ranked average	0.85276	0.80701	0.75275	−6.7234	0.58322
Team B Arete Associates (2nd)	60 (30), 600 (0)	Correlation, distribution statistics, zero crossings, complexity, mobility, maximum frequency, total summed energy, entropy, normalized summed spectral energy	Extremely randomized trees[g]	n/a	0.78328	0.79898	0.73364	−8.1773	0.56306
Team C GarethJones (3rd)	80, 160, 240 (0)	Spectral power, distribution statistics, RMS of signal, first and second derivatives, correlation, spectral edge	Polynomial SVM, random under-sampling boosted tree ensemble	Weighted average	0.81524	0.79652	0.65523	−17.7388	0.41632
Team D QingnanTang (4th)	75 (0)	Spectral power, correlation, spectral entropy, spectral edge power, square of features	Gradient boosting[h], extreme gradient boosting, radial basis function SVM	Weighted average	0.7965	0.79458	0.71805	−9.6319	0.52086
Team E Nullset (5th)	30 (0)	Spectral power, correlation (and eigenvalues), spectral entropy, Shannon entropy, spectral edge frequency, Hjorth parameters, fractal dimensions	Adaptive boosting, gradient boosting, random forest, extreme gradient boosting, gridsearch	Voting classifier	0.81423	0.79363	0.62929	−20.7074	0.46132
Team F tralala boum boum	60 (0)	Spectral power, spectral entropy, time/spectral correlation (and	SVM, random forest, extreme gradient	Weighted average	0.80493	0.79197	0.71822	−9.3118	0.49742

Quelle: Kuhlmann et al, Brain 141, 2619-2630, 2018

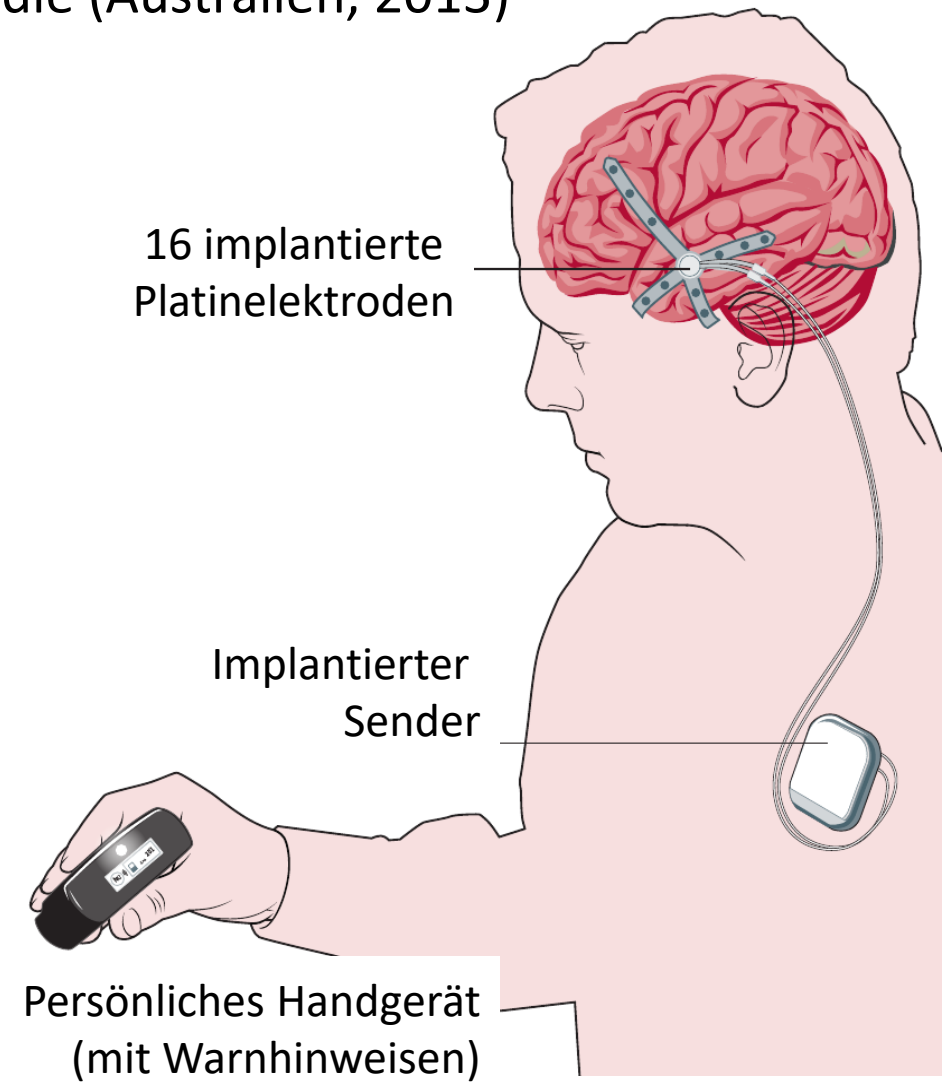
Fallbeispiel epileptische Anfallsvorhersage

Erste Machbarkeitsstudie (Australien, 2013)

≈ „FPR“

TPR

	High likelihood performance		
	Time in advisory (%)	Seizures (n)	Sensitivity
Patient 1	27	7 (13)	86% (77%)
Patient 2	31	3	100%
Patient 3	29	58 (106)	56% (45%)
Patient 4†	..	..	..
Patient 8	28	36 (86)	63% (62%)
Patient 9	11	49 (52)	18%§ (17%)
Patient 10	17	109 (164)	54% (51%)
Patient 11	15	11 (39)	56% (39%)
Patient 13	28	26 (113)	57% (50%)
Patient 14	3	3	100%
Patient 15	41	21 (24)	71%



Cook et al, Lancet Neurol. 12, 563-571, 2013

Fallbeispiel epileptische Anfallsvorhersage

Erste Machbarkeitsstudie (Australien, 2013)

… und ihre Auswirkungen 2023

Warum einer Frau ein lebensveränderndes Hirnimplantat wieder entnommen wurde

Der Australierin Rita Leggett half ein experimentelles Neuro-Implantat im Gehirn, ein besseres Leben zu führen. Doch behalten durfte sie es nicht.

Lesezeit: **12 Min.** In Pocket speichern 269

01.06.2023 07:00 Uhr | MIT Technology Review

Von Jessica Hamzelou

Rita Leggett stammt aus Australien – und sie trug über längere Zeit ein experimentelles Hirnimplantat, das ihr in ihrem Alltag enorm half. Es habe ihr neues Selbstwertgefühl gegeben, sagt sie, und sei "eins mit ihr geworden". Entsprechend war sie am Boden zerstört, als man ihr zwei Jahre später mitteilte, dass das Implantat wieder entfernt werden müsse. Das Unternehmen, das es hergestellt hatte, war pleitegegangen.

Quelle: https://heise.de/-9067490

Daten-getriebene Modellierung | Klassifikation

Bisher: binäre Klassifikation

- Modell unterscheidet zwischen *zwei* Klassen
- einfachstes Modell: Schwellwert-basierte Klassifikation

Multiklassen Klassifikation

- Modell unterscheidet zwischen mehr als zwei Klassen
- einfachstes Modell: Nächste Nachbarn

Nächste Nachbarn Modell (NN Modell)

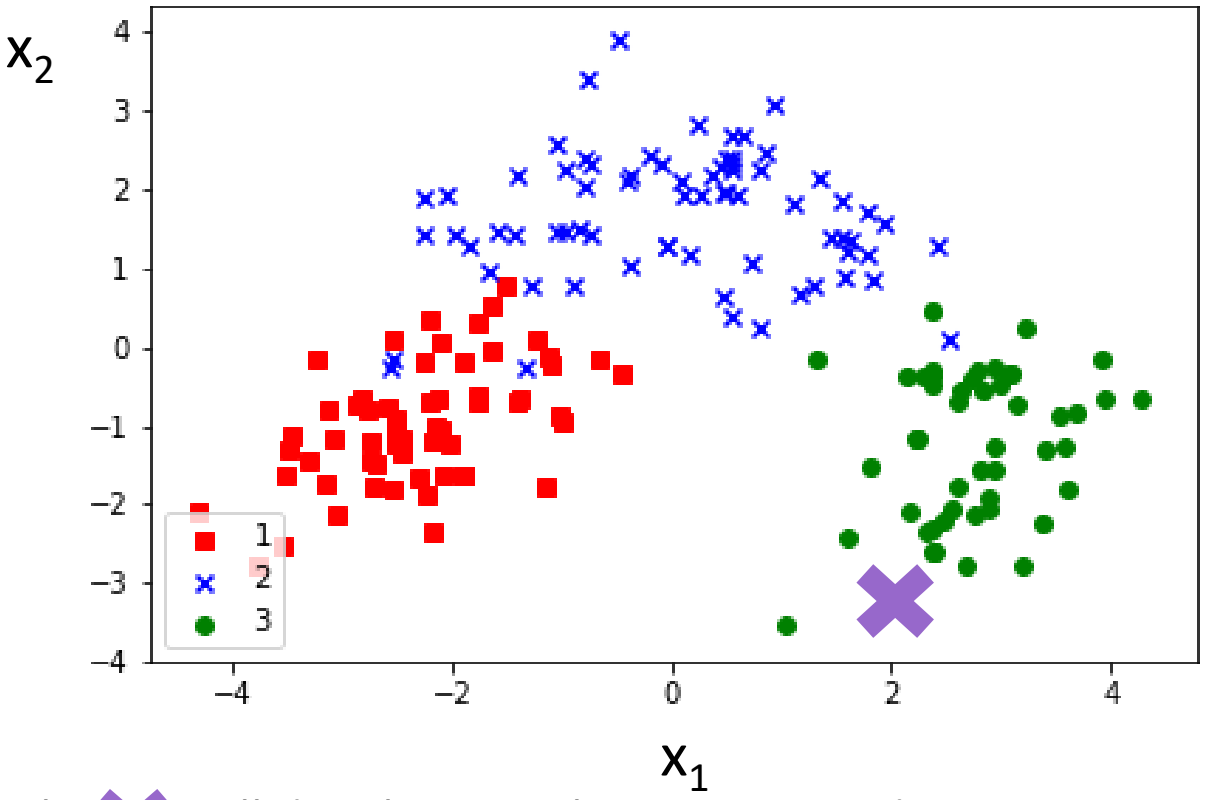
- Nähe (bzw. Distanz) im Merkmalsraum bestimmt vorherzusagende Klasse.

Daten-getriebene Modellierung | NN Modell

Wie lässt sich eine Multiklassen-Klassifikator erstellen?

Beispiel:

- Daten mit drei Klassen
- Generieren eines Nächste Nachbarn Modells (kurz: NN Modell)



Für einen *neuen* Datenpunkt ✖ soll die Klasse vorhergesagt werden.

- Gebe als Vorhersage die Klasse des nächsten Nachbarn von ✖ zurück. (Nähe wird hier durch euklidische Distanz definiert.)

Beurteilung eines Klassifikators | Grundkonzepte

Multiklassen-Klassifkation

→ Vorhersagen über mehr als zwei Klassen

Beispiel: Detektion der drei Klassen „Covid-19“, „Influenza“, „Nicht erkrankt“

Verschiedene Ansätze:

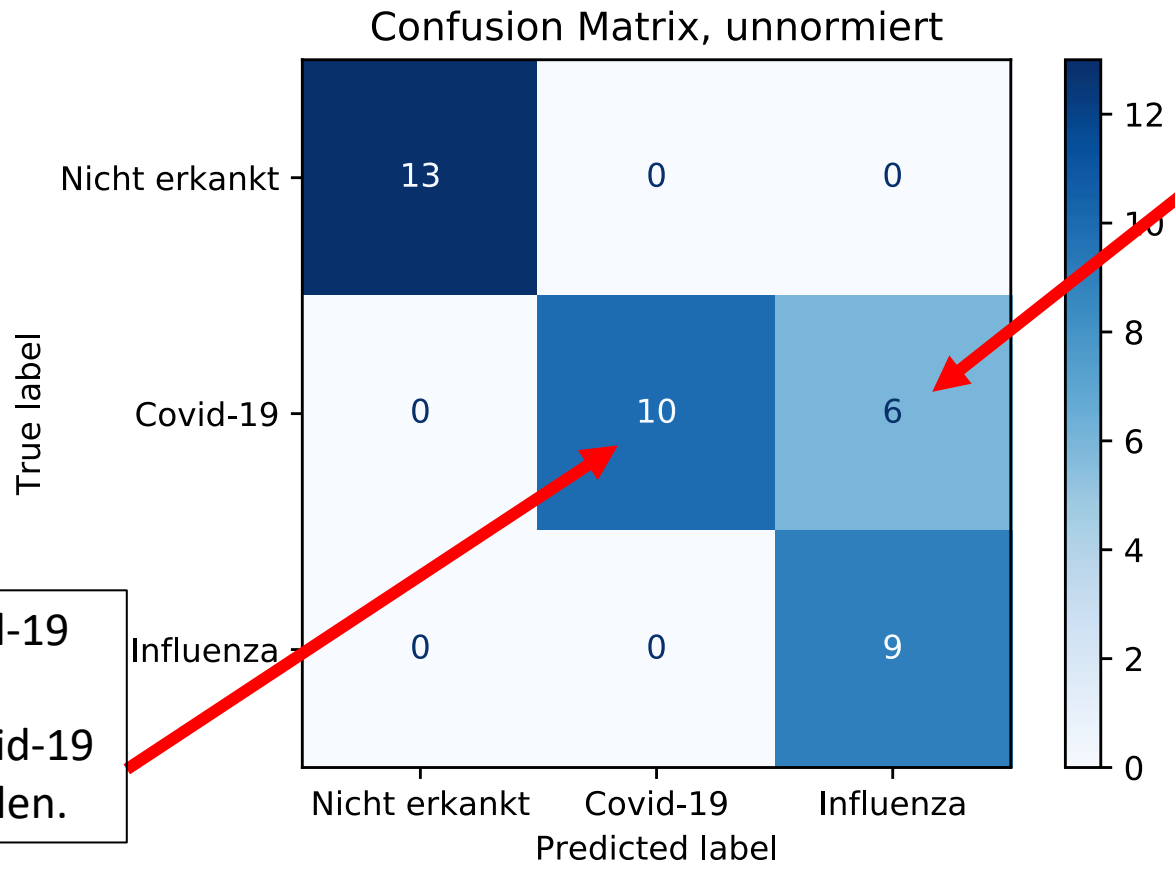
a) Top-K Accuracy (oft: K=5)
Vorhersage wird als True Positive *TP* gewertet, sofern die wahre Klasse unter den Top-K der wahrscheinlichsten Klasse liegt.
(Grenzfall K=1 entspricht der normalen *Accuracy*)

b) One-vs-All Ansatz
Ermitteln des F1-Scores für jede Klasse (z.B. Covid-19 vs „Nicht-covid-19“
Mittelwert über die F1-Scores aller Klassen bilden.

c) Confusion Matrix (Wahrheitsmatrix) → auf der nächsten Folie

Beurteilung eines Klassifikators | Confusion Matrix

Confusion Matrix C_{ij} : Anzahl Elemente der Klasse i die als Klasse j vorhergesagt werden.

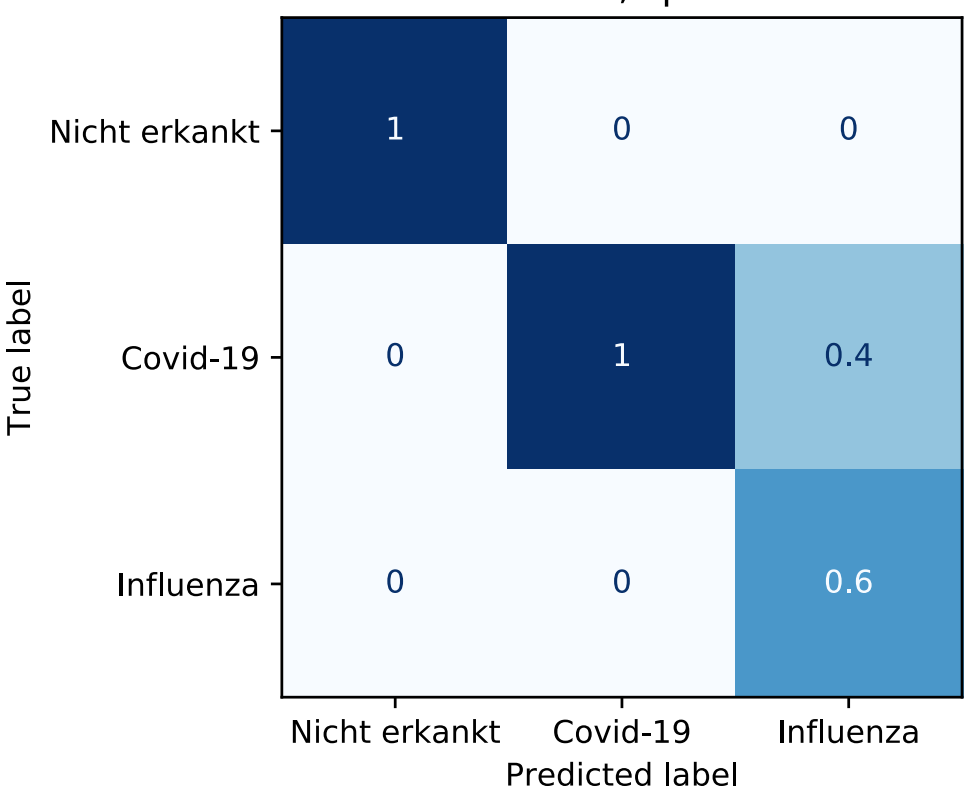


Anzahl der Covid-19 Patienten, die als Influenza-Fälle fehlklassifiziert wurden

Anzahl der Covid-19 Patienten, die (korrekt) als Covid-19 klassifiziert wurden.

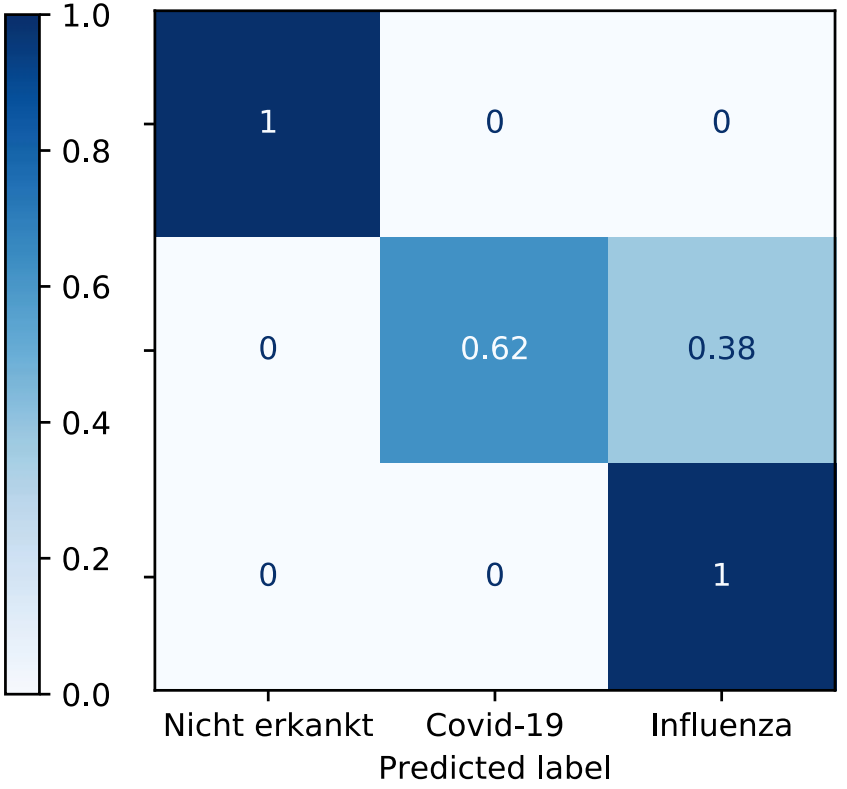
Normierte Confusion Matrizen

$$\tilde{C}_{ij} = C_{ij} / (\sum_k C_{kj})$$



Diagonaleinträge entsprechen PPV (Precision) der jeweiligen Klasse!

$$\tilde{C}_{ij} = C_{ij} / (\sum_k C_{ik})$$



Diagonaleinträge entsprechen TPR (Recall) der jeweiligen Klasse!