# Feature-Extraction

## Theoretische Grundlagen, Methodik und Umsetzung

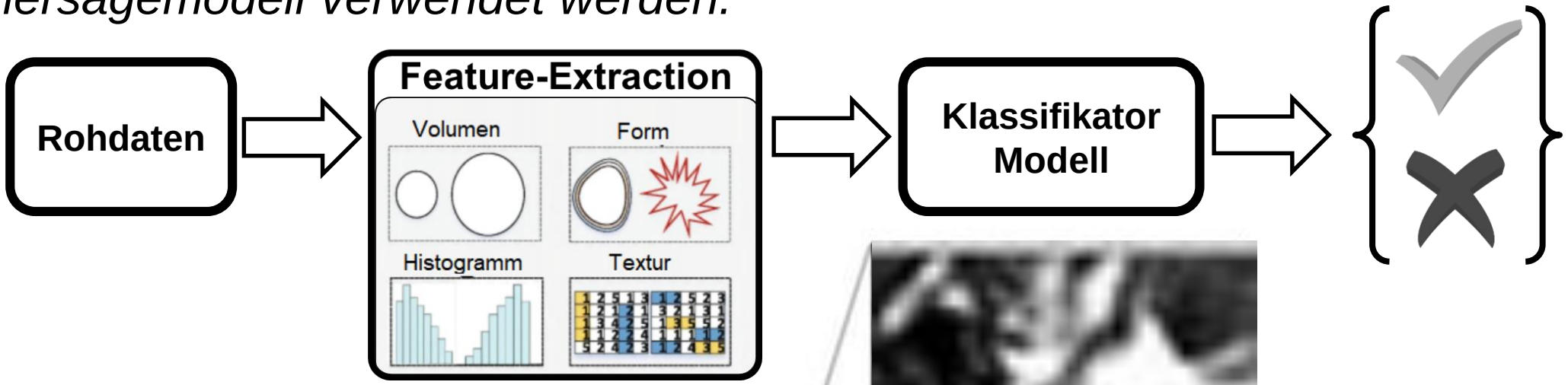
*am Beispiel von medizinischen Diagnosen*

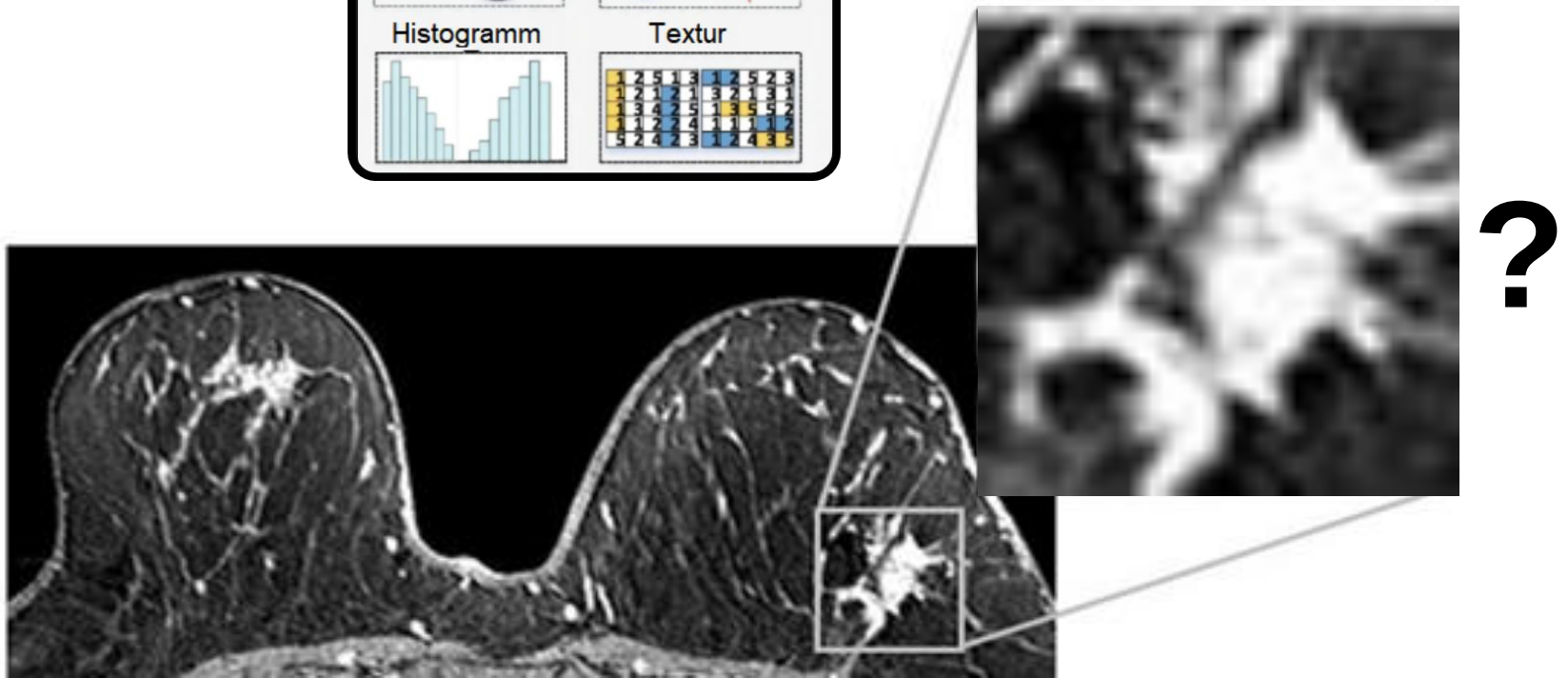
# Inhalt

- **Grundlagen** der Feature Extraction
- **Methodik**: Manuelle vs. Automatische Feature-Extraction
- **Umsetzung in der medizinischen Diagnostik**: die Krebsklassifizierung
    - Manuelle Feature Extraction
    - Automatische Feature-Extraction
    - Data-Engineering-Pipelines
    - Analyse und Auswertung
- **Fazit**

# Grundlagen der Feature-Extraction

*Feature-Extraction: Umwandlung von Rohdaten in numerische Features unter Beibehaltung der Informationen im Originaldatensatz die später in einem Vorhersagemodell verwendet werden.*

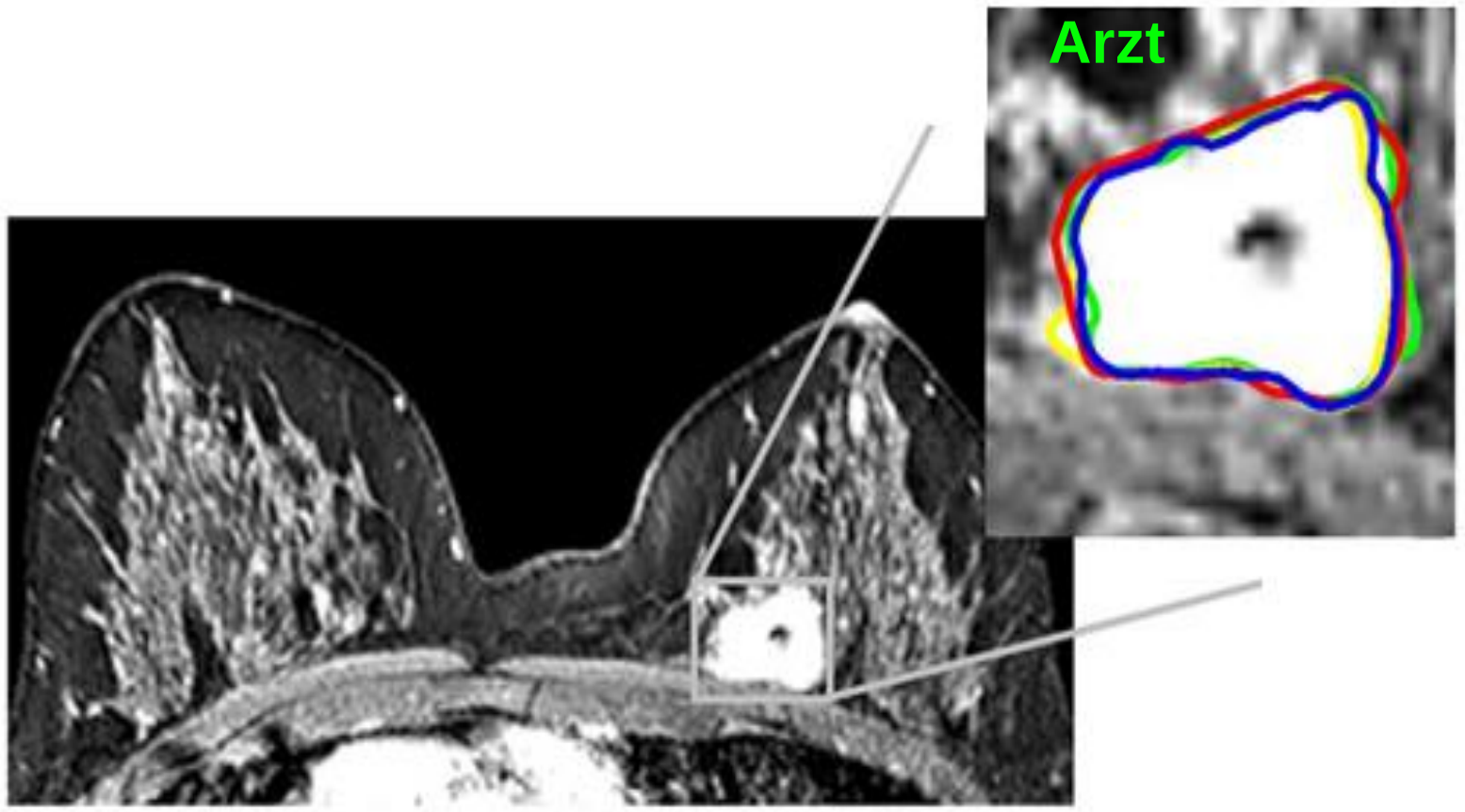


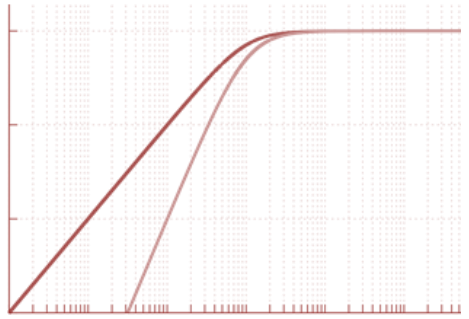
?

# Grundlagen der Feature-Extraction

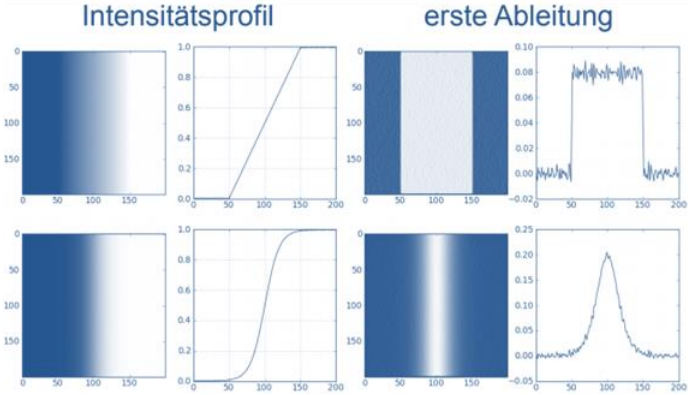
MRT-Brustkrebserkennung



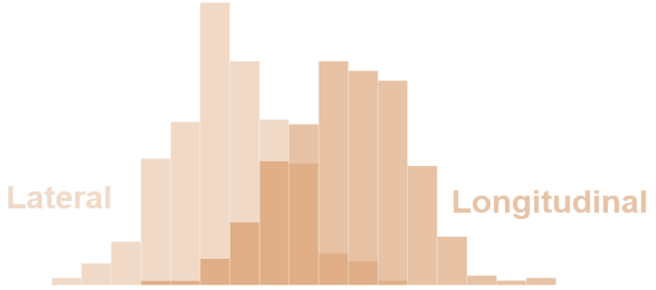

*Hochpassfilter*



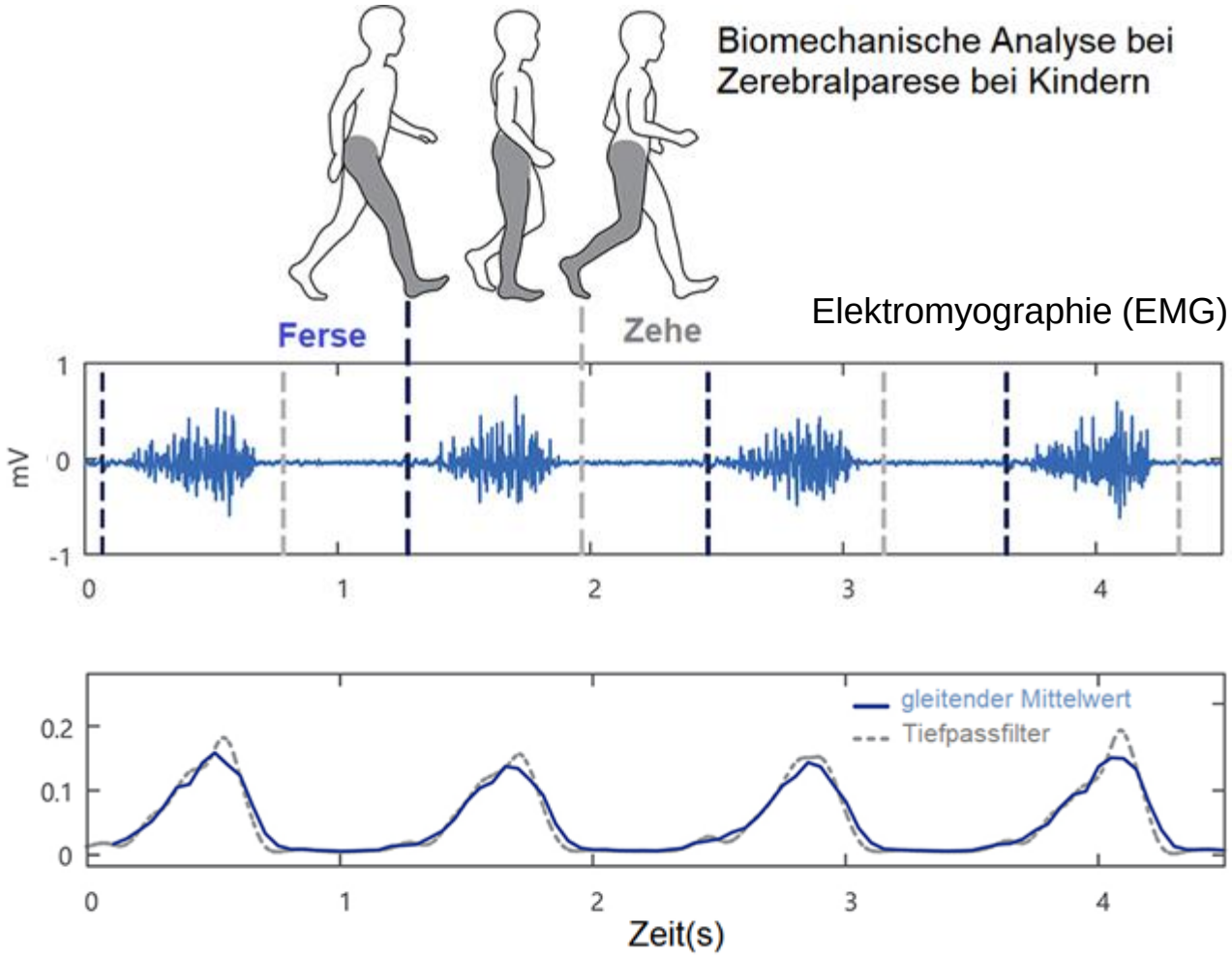
*Sobel-Operator*



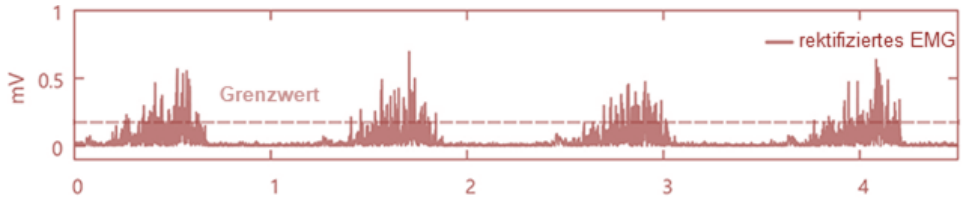

Histogramm



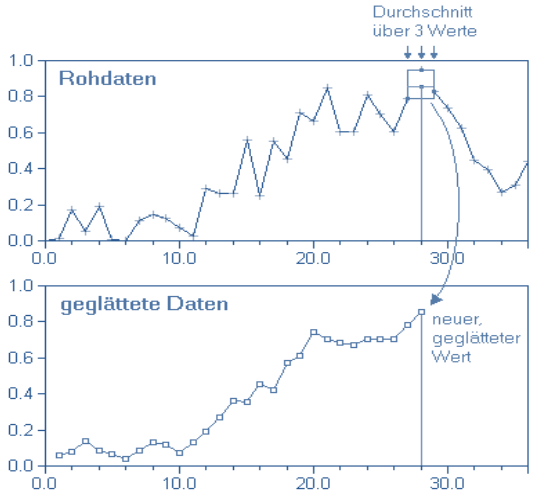
# Grundlagen der Feature-Extraction



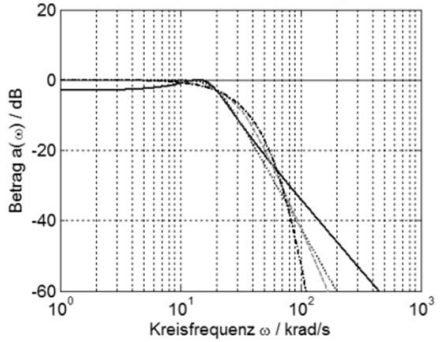

Signal-Gleichrichtung



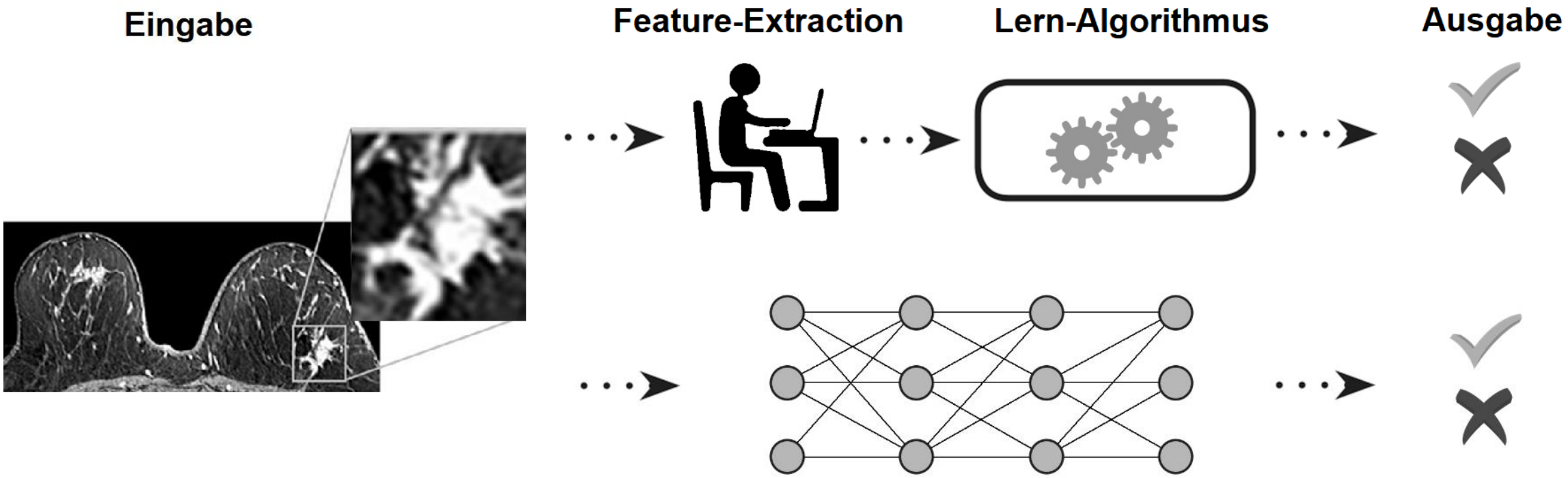

Gleitender Mittelwert



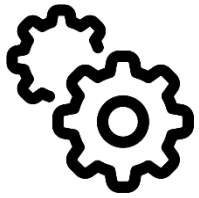

Tiefpassfilterung



# Manuelle vs. Automatische Feature-Extraction



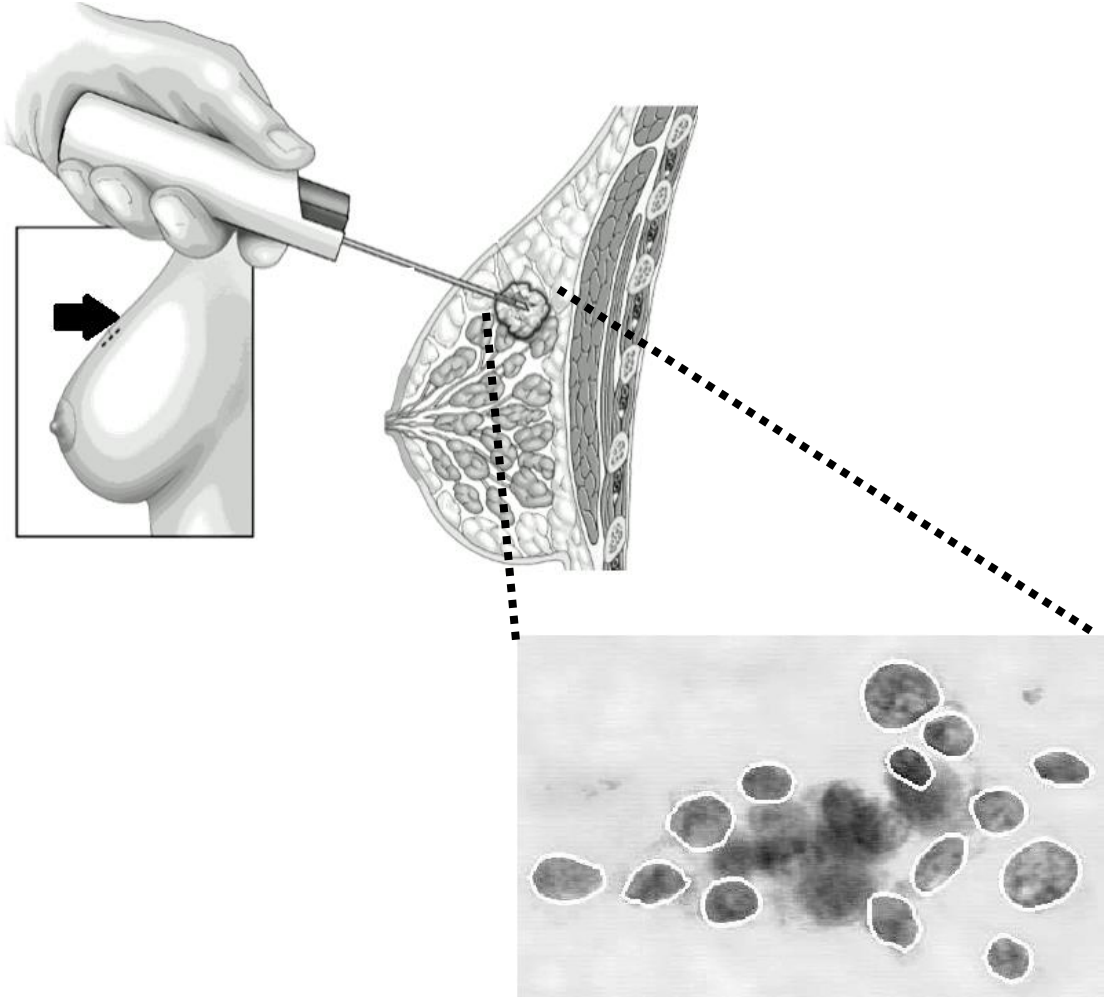

: Filterung, Stichprobe Statistik(z.B. Korrelation), Signalverarbeitung...



: PCA, Autoencoders, Convolutional Neuronale Netze…

# Umsetzung in der medizinischen Diagnostik

## Brustkrebs-Diagnose



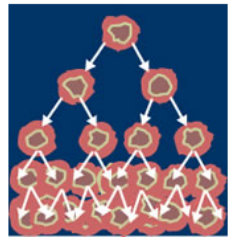
**Aufgabe**: Die Brustmasse als **gut-** oder **bösartig** zu klassifizieren

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

**Breast Cancer Wisconsin (Diagnostic) Data Set**

*Download*: Data Folder, Data Set Description

**Abstract**: Diagnostic Wisconsin Breast Cancer Database



**569 Patientinnen, 30 Features**

10 Charakteristika des Brustmassenzellkerns wurden gemessen:

- Radius (Mittelwert aller Abstände vom Zentrum zu Punkten auf dem Perimeter)
- Textur (Standardabweichung der Grauskala-Werte)
- Umfang
- Fläche
- Glattheit (lokale Variation der Radiuslängen)
- Kompaktheit (Umfang^2 / Fläche - 1,0)
- Konkavität (Stärke der konkaven Teile der Kontur)
- Konkavitätspunkte (Anzahl der konkaven Teile der Kontur)
- Symmetrie
- Fraktale Dimension ("Küstenlinienapproximation" - 1)

Für jedes Feature werden 3 Maße angegeben:

- Kleinste
- Standardfehler
- Größte/"schlechteste

https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

# Umsetzung in der medizinischen Diagnostik

## Manuelle Feature-Extraction

*Theorie, Methodik und Umsetzung von Korrelation*

Gegeben $(x_i, y_i)^\top,\ i = 1, \ldots, n$ eine zweidimensionale Stichprobe

mit den empirischen Mitteln

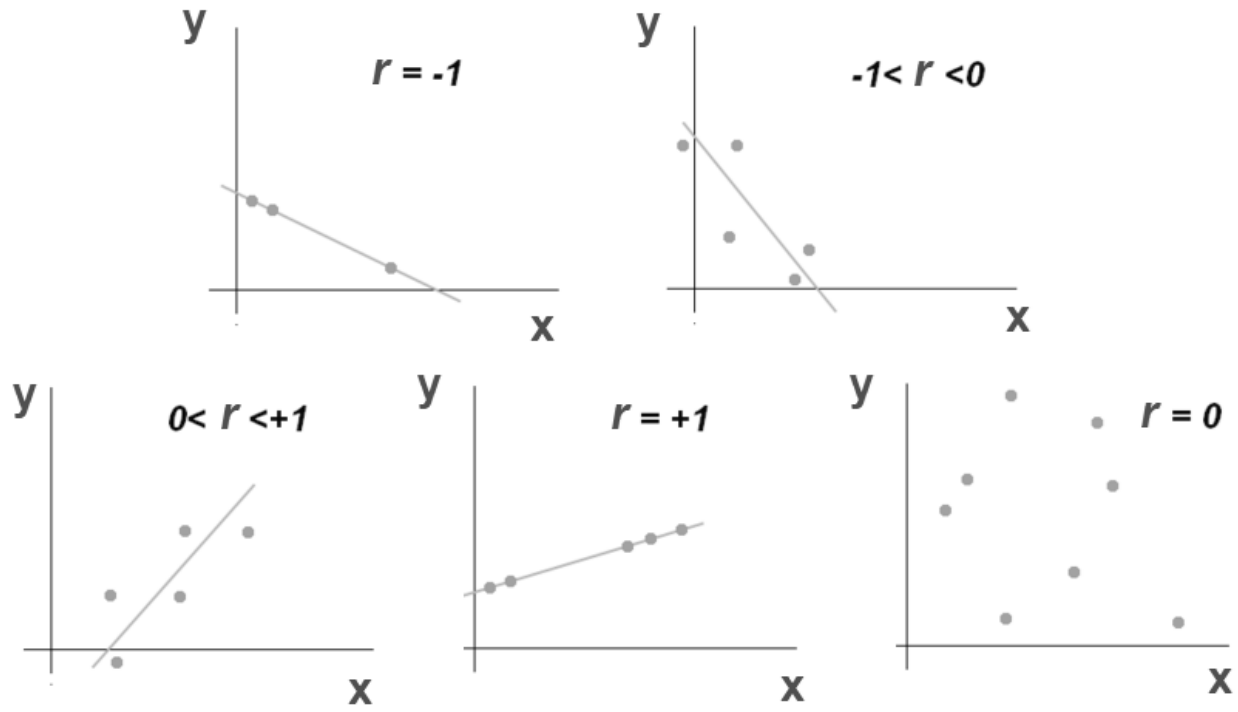
$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$y = (y_1, \ldots, y_n)^\top \qquad x = (x_1, \ldots, x_n)^\top$$

der *empirische Korrelationskoeffizient ist*

$$r_{x,y} := \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

*Beispiel Korrelationskoeffizient Scatter-Diagramme*



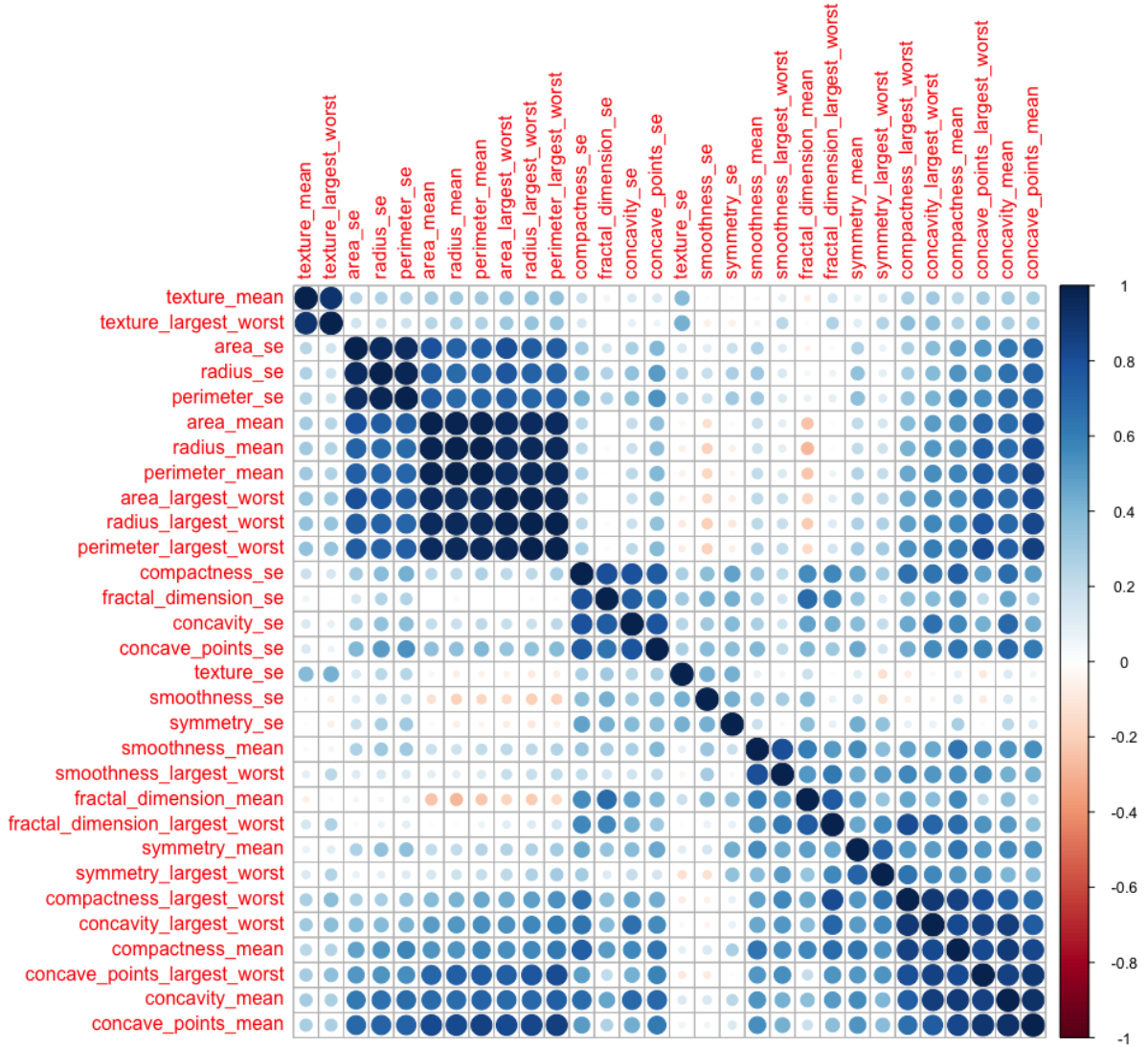
# Umsetzung in der medizinischen Diagnostik

## Manuelle Feature-Extraction

*Theorie, Methodik und Umsetzung von Korrelation*

$$r_{x,y} := \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

```
import seaborn as sns
from sklearn import preprocessing
(X, y) = load_breast_cancer(return_X_y=True, as_frame=True)
X.columns
X = X.iloc[:,1:-1]
label_encoder = preprocessing.LabelEncoder()
X.iloc[:,0] = label_encoder.fit_transform(X.iloc[:,0]).astype('float64')
corr = X.corr()
sns.heatmap(corr)
columns = np.full((corr.shape[0],), True, dtype=bool)
for i in range(corr.shape[0]):
    for j in range(i+1, corr.shape[0]):
        if corr.iloc[i,j] >= 0.7:
            if columns[j]:
                columns[j] = False
selected_columns = X.columns[columns]
X = X[selected_columns]
```




**Wir entfernen alle Features mit einer Korrelation von mehr als 0,7 (16/30 Features)!**

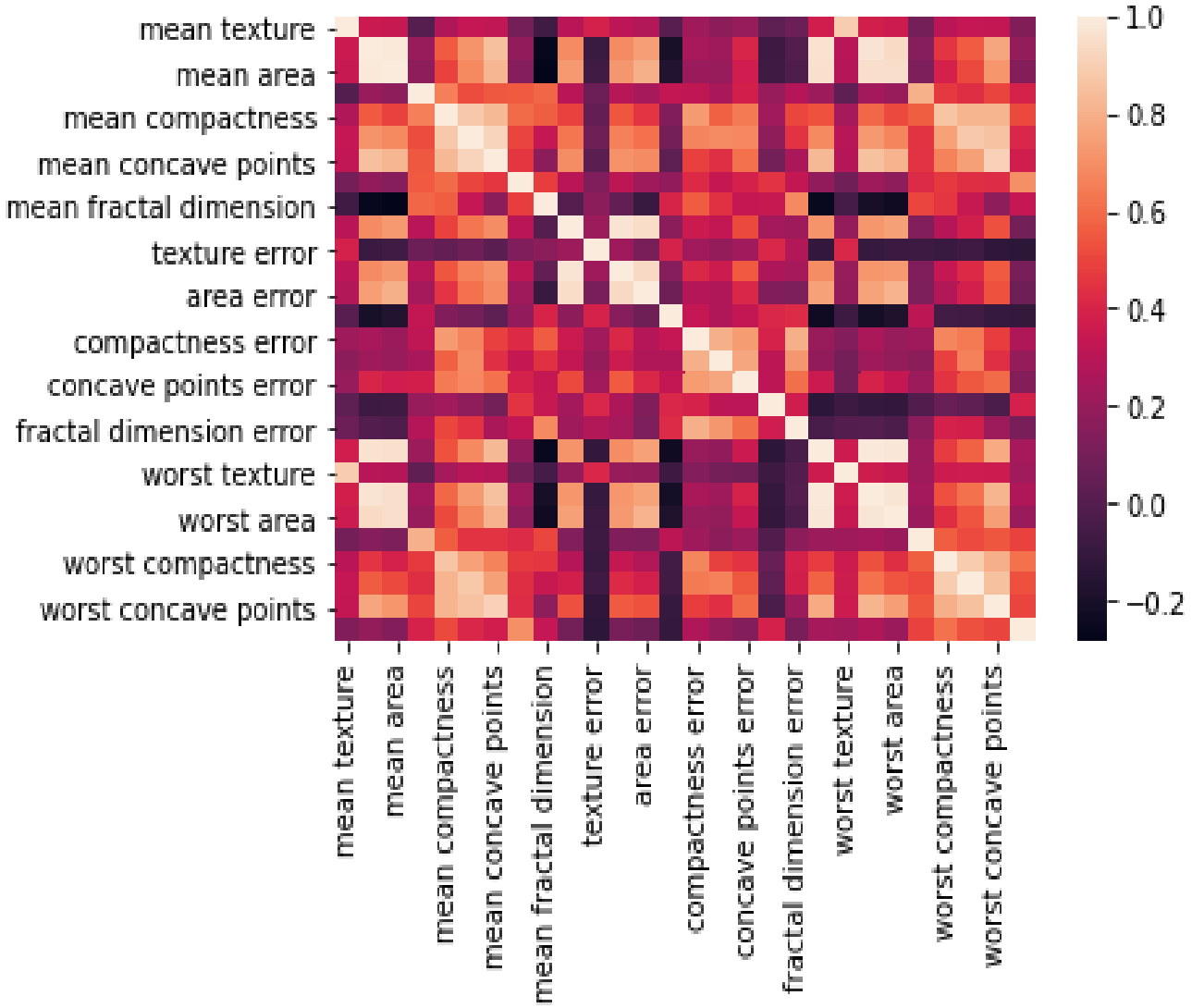
# Umsetzung in der medizinischen Diagnostik

## Manuelle Feature-Extraction

*Theorie, Methodik und Umsetzung von Korrelation*

$$r_{x,y} := \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

```
import seaborn as sns
from sklearn import preprocessing
(X, y) = load_breast_cancer(return_X_y=True, as_frame=True)
X.columns
X = X.iloc[:,1:-1]
label_encoder = preprocessing.LabelEncoder()
X.iloc[:,0] = label_encoder.fit_transform(X.iloc[:,0]).astype('float64')
corr = X.corr()
sns.heatmap(corr)
columns = np.full((corr.shape[0],), True, dtype=bool)
for i in range(corr.shape[0]):
    for j in range(i+1, corr.shape[0]):
        if corr.iloc[i,j] >= 0.7:
            if columns[j]:
                columns[j] = False
selected_columns = X.columns[columns]
X = X[selected columns]
```




**Wir entfernen alle Features mit einer Korrelation von mehr als 0,7 (16/30 Features)!**

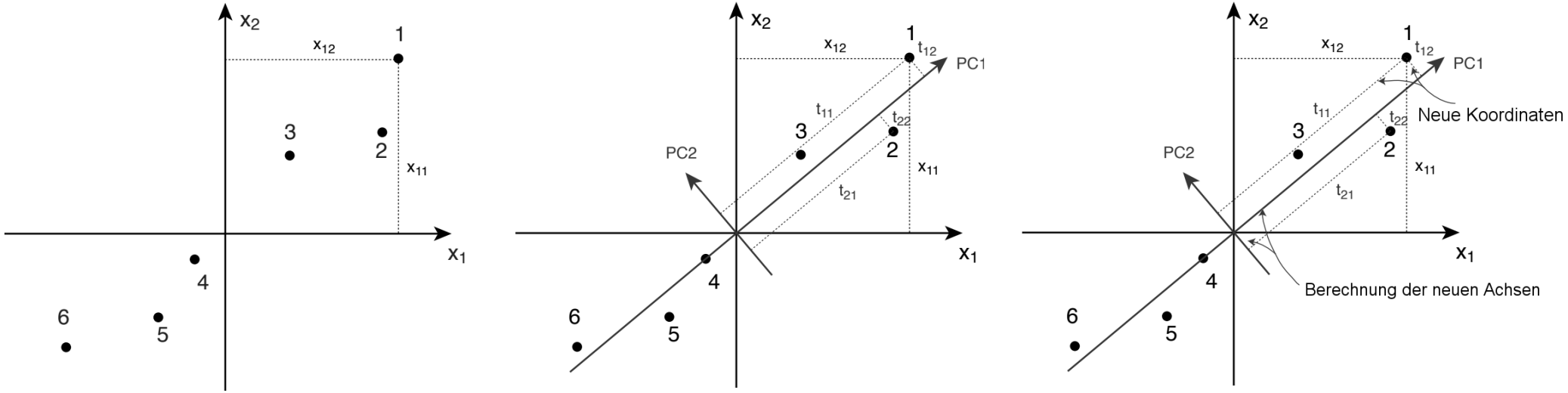
# Umsetzung in der medizinischen Diagnostik

## Automatische Feature-Extraction

*Theorie, Methodik und Umsetzung von PCA*

PCA

- **Unüberwachtes Lernen** zur **Dimensions-Reduzierung** genutzt (**Komprimierung**)
- "**Transformiere**" eine Menge von Beobachtungen in ein anderes **Koordinatensystem**, in dem die Werte der **ersten Koordinate (Komponente)** die **größtmögliche Varianz** aufweisen
- **Lineare Transformation** bei dem die Rekonstruktion der Beobachtungen aus den führenden Hauptkomponenten hat den **niedrigsten quadratischen Fehler**

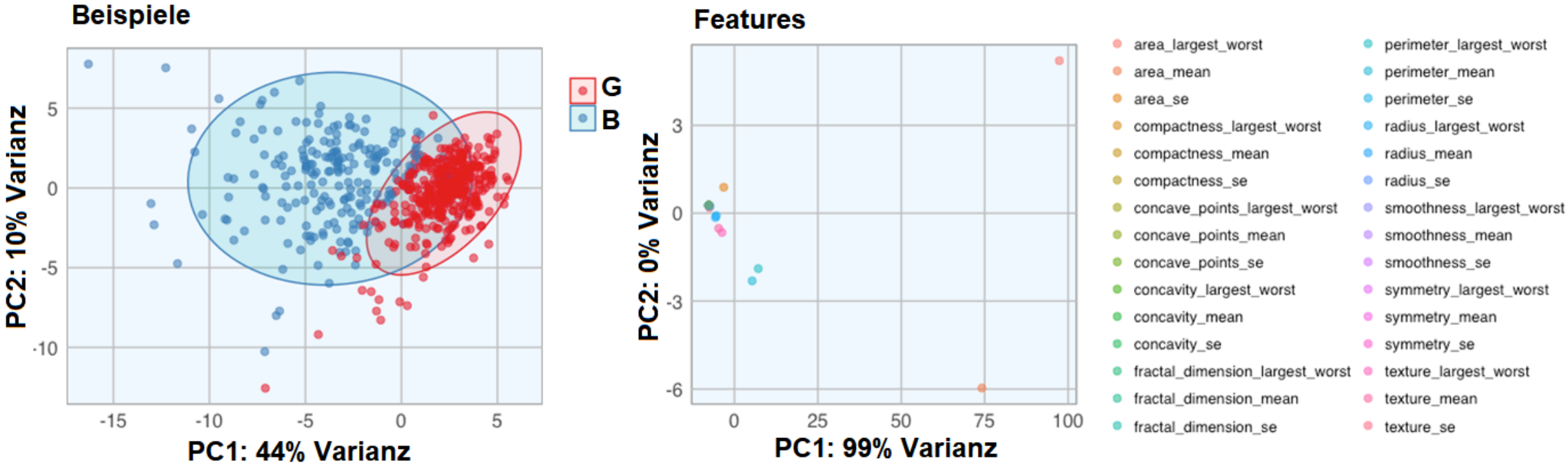


# Umsetzung in der medizinischen Diagnostik

## Automatische Feature-Extraction

*Theorie, Methodik und Umsetzung von PCA*

Die ersten zwei Hauptkomponenten (PC) erklären den Großteil der Variation in den Daten.



# Umsetzung in der medizinischen Diagnostik

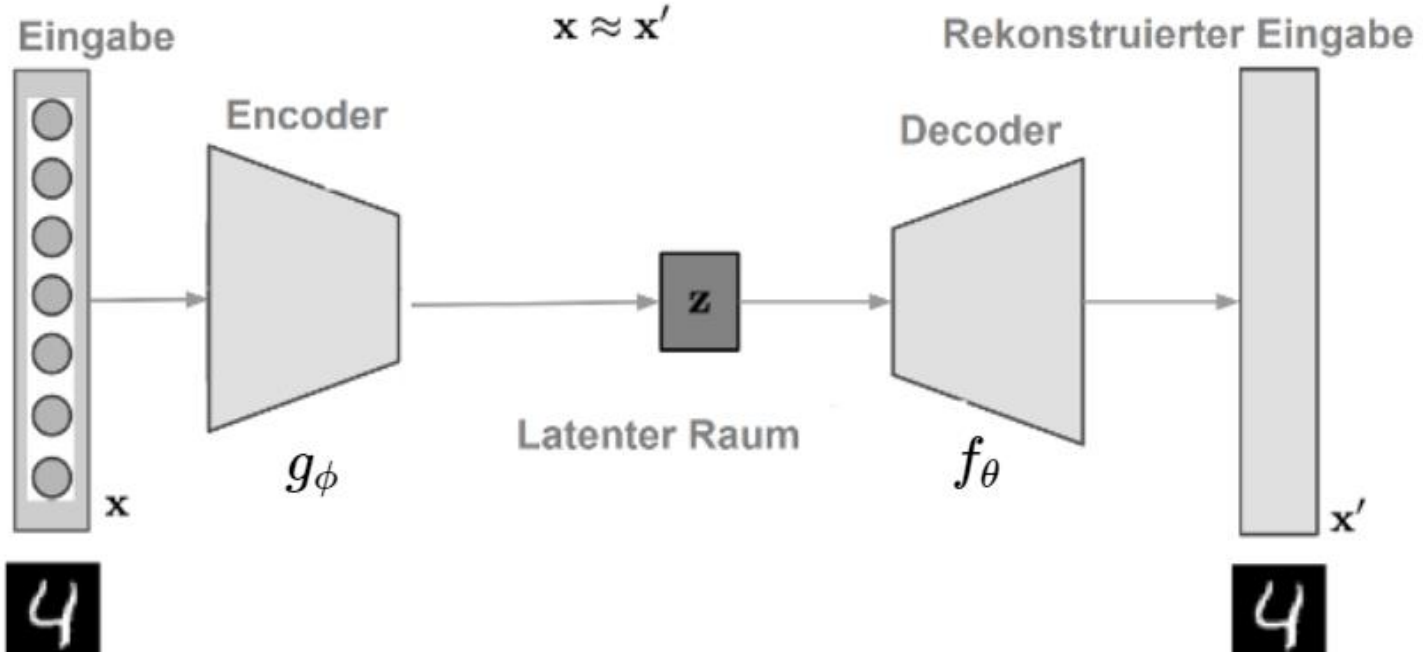
## Automatische Feature-Extraction

*Theorie, Methodik und Umsetzung von Autoencoders*

### Autoencoder

Für jeden Eingangsvektor $x$ der Dimension $d$ des kompletten Datensaztes der Länge $n$ generiert das neuronale Netz eine Rekonstruktion $x'$ durch:

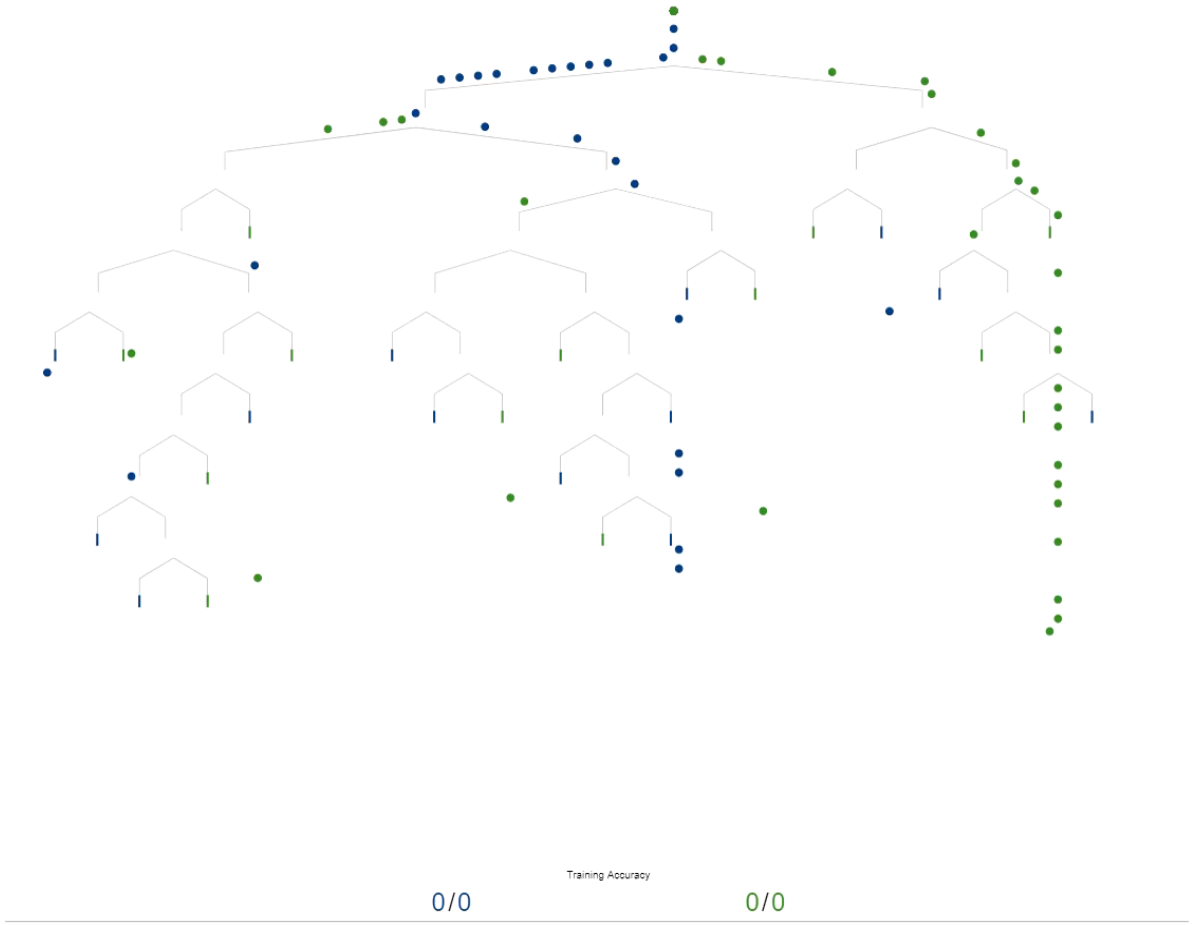
- **Kodierung der Eingangsdaten** (d.h. verwende die lineare / nicht-lineare Transformation $g_\phi(.)$)
- dies liefert eine **komprimierte Kodierung** in der dünnsten Netzwerk-Ebene, $z$
- **Dekodierung der komprimierten Eingangsdaten** durch Anwendung der linearen / nicht-linearen Transformation $f_\theta(.)$

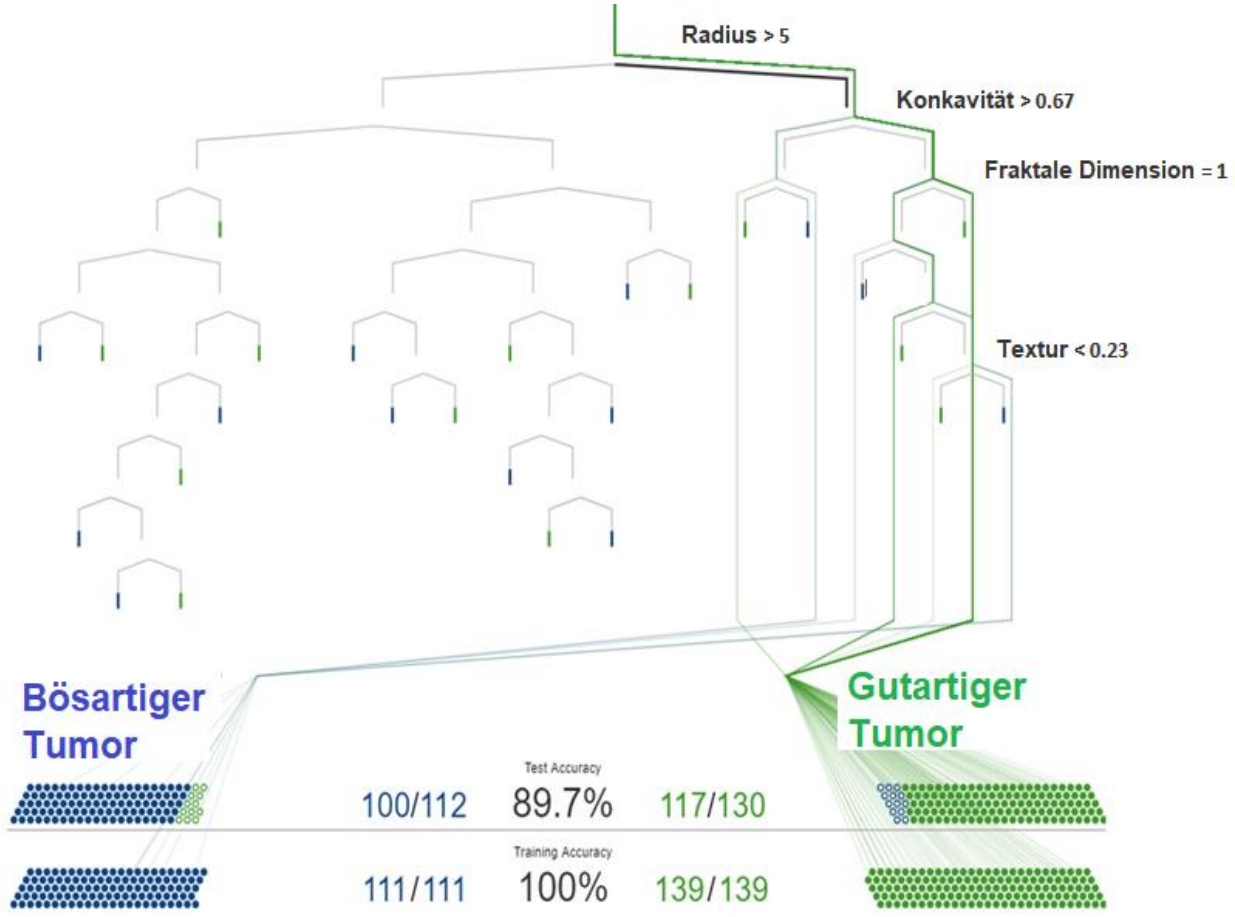


# Umsetzung in der medizinischen Diagnostik

## Klassifizierung Modell

*Anwendung von Entscheidungsbäumen*



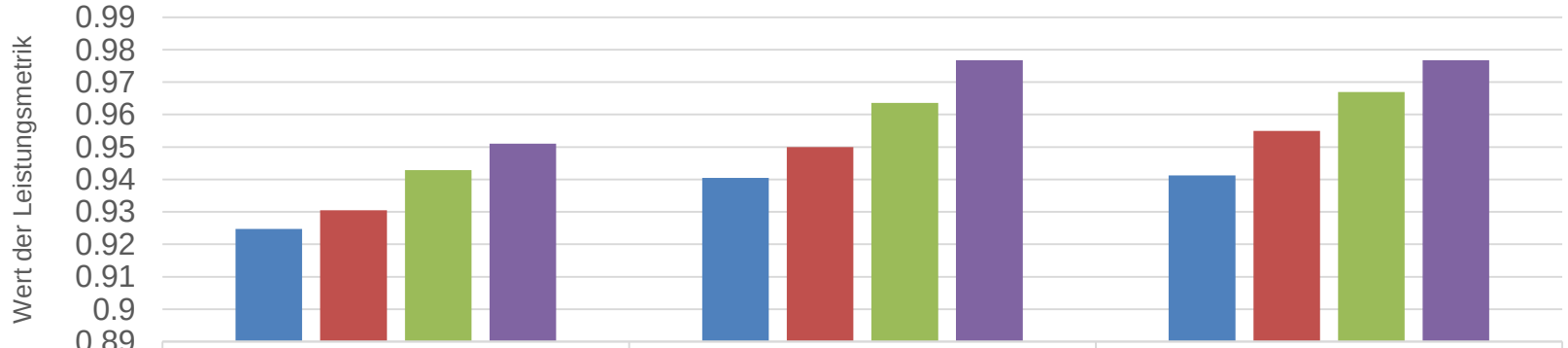



# Umsetzung in der medizinischen Diagnostik

Data-Engineering-Pipelines Auswertung

- Daten-Skalierung + Baseline-Klassifikator (Entscheidungsbäumen)
- Daten-Skalierung + Korrelation + Baseline-Klassifikator
- Daten-Skalierung + PCA  + Baseline-Klassifikator
- Daten-Skalierung + Encoder + Baseline-Klassifikator

Demo-Code Github verfügbar




| | Accuracy | Precision | Recall |
|---|---|---|---|
| Baseline (Entscheidungsbaum-Klassifikator) | 0.924774436 | 0.940450027 | 0.941190476 |
| Korr + Entscheidungsbaum-Klassifikator | 0.930459579 | 0.949913322 | 0.954957265 |
| PCA + Entscheidungsbaum-Klassifikator | 0.942901003 | 0.963545501 | 0.966920635 |
| Encoder + Entscheidungsbaum-Klassifikator | 0.950971178 | 0.976703745 | 0.976690476 |

Baseline (Entscheidungsbaum-Klassifikator) Korr + Entscheidungsbaum-Klassifikator
PCA + Entscheidungsbaum-Klassifikator Encoder + Entscheidungsbaum-Klassifikator

# Fazit

**Die Feature-Extraction** ist:

- nützlich, wenn wir die **Anzahl** der für die Verarbeitung benötigten **Ressourcen reduzieren** müssen, ohne wichtige oder **relevante Informationen** zu verlieren
- ein **wichtiger Schritt in der Data-Engineering-Pipeline**, bevor das Prädiktivmodell erstellt wird
- ist **hoch domänen-** und **datenspezifisch**
- **unterstützt** eine bessere **Analyse** und **Interpretation** der Vorhersagen

# Feature-Extraction

## Theoretische Grundlagen, Methodik und Umsetzung

*am Beispiel von medizinischen Diagnosen*