Maschinelles Lernen in der klinischen Bioinformatik: Einführung

Dr. Meik Kunz

Lehrstuhl für Medizinische Informatik

16.10.2019

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Machine Learning

- „A computer program is said to learn from experience **E** with respect to some class of tasks **T** and performance measure **P**, if its performance at tasks in T, as measured by P, improves with experience E.“ – Mitchel 1997

  → Ein Programm soll mit zunehmender Erfahrung bei der Lösung einer Aufgabe besser abschneiden

- ML erlaubt das Bearbeiten von sehr komplexen und rechenintensiven Aufgaben

- auf Basis vorhandener Datenbestände und Algorithmen Muster und Gesetzmäßigkeiten erkennen und Vorhersagen treffen

Definitionen/Begriffe

- Künstliche Intelligenz (Artificial Intelligence)
  - breitester Begriff für Techniken, die es dem Computer ermöglichen die menschliche Intelligenz nachzuahmen
- Machine Learning
  - Teilgebiet von KI
  - erlaubt Bearbeiten sehr komplexer und rechenintensiven Aufgaben, um Muster und Gesetzmäßigkeiten zu erkennen und vorherzusagen
  - beinhaltet statistische Techniken
  - y = *f(x)*
- Deep Learning
  - Teilbereich des Machine Learning
  - nutzt neuronale Netze

Anwendungsgebiete

Anwendungsgebiete: KI in der Medizin

- Diagnostizieren von Krankheiten
- Radiologie (Imaging)
- Personalisierte Medizin
- Natural Language Processing
- Medikamente entwickeln (Drug Discovery)
- Chirurgie-Robotik

Diagnostizieren von Krankheiten

- Pneumonie (Lungenentzündung) in den 1990er Jahren
- Neuronales Netzwerk um Pneumonie vorherzusagen
  - Behandlung und Überleben der Patienten zu verbessern
  - Kosten und Zeit zu sparen
- Training mit Daten von 250k Patienten aus 78 Krankenhäusern
- Ergebnis: Patienten mit Pneumonie + Asthma hatten besseren Outcome
  - Kausaler Zusammenhang? NEIN!

➡ Expertise ist unersetzbar

Diagnostizieren von Krankheiten

- ein Drittel aller AI SaaS Firmen im Gesundheitswesen sind spezialisiert auf medizinische Diagnostik

- ca. 10% und der Todesfälle 6-17% der Komplikationen von Patienten verursacht durch falsche Diagnose
  - Ineffektive Kollaboration
  - lückenhafte Kommunikation
  - keine ausreichenden diagnostischen Tests
  - (unzureichende Expertise)

Bedarf an medizinischer Diagnostik basierend auf KI

Emerj, 2019 https://emerj.com/

Diagnostizieren von Krankheiten

Liquid Biopsy

- Früherkennung enorm wichtig für Heilungschance
- Blutbasierter Analytik von Tumoren ohne invasiven Eingriff
  - Zirkulierende Tumorzellen
  - Zellfreie Tumor-DNA
- Einsatzgebiete
  - Krebsfrüherkennung und wiederkehrende Tumore
  - Abschätzung des Metastasierungsrisikos
  - Identifizierung therapeutischer Zielstrukturen und Resistenzmechanismen
  - Tumor-Monitoring
- ML-Modelle können trainiert werden, um Risiko abzuschätzen bzw. Vorhersagen zu treffen

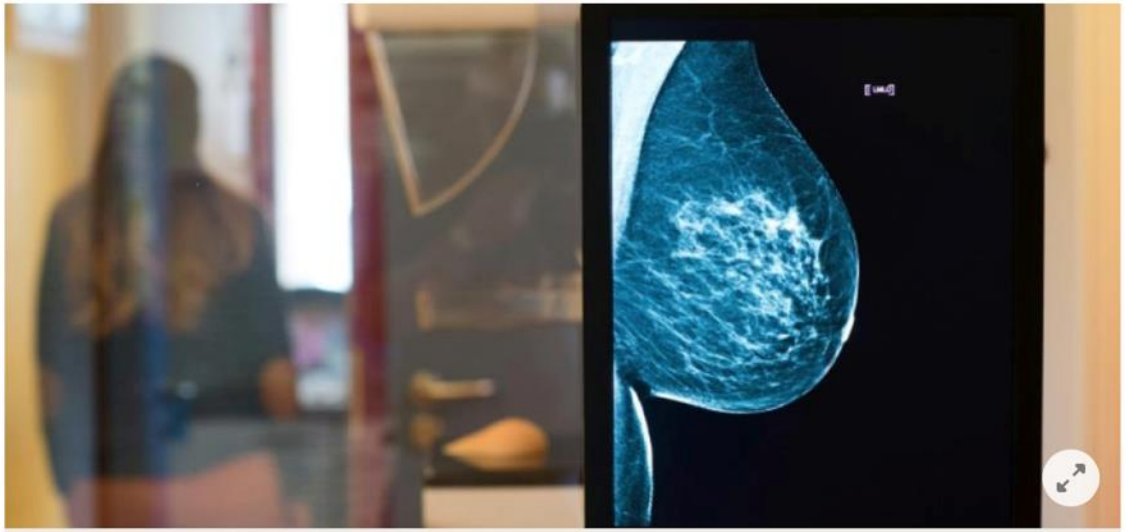
Diagnostizieren von Krankheiten

Liquid Biopsy

NEWS1 (AFP - JOURNAL) KREBS

Heidelberger Forscher entwickeln Bluttest zur Erkennung von Brustkrebs

Veröffentlicht am 21.02.2019 | Lesedauer: 2 Minuten



Röntgenbild von einer weiblichen Brust

Quelle: dpa/AFP/Archiv

In aktueller Studie zeigt Test Treffsicherheit von 75 Prozent

Diagnostizieren von Krankheiten

Liquid Biopsy

NEWS1 (AFP - JOURNAL) KREBS

Blamage mit Bluttest

29.04.2019, 19:47 Uhr

Skandal bedroht Exzellenzprädikat der Uni Heidelberg

Ein Bluttest der Uni Heidelberg wurde in Medien als „Weltsensation“ gefeiert. Nun ist er als unbrauchbar entlarvt - und der Ruf der Uni schwer beschädigt. VON SASCHA KARBERG

Diagnostizieren von Krankheiten

Chatbots

Babylon Health

- Videotelefonie oder Chat mit Arzt
- Muster in Symptomen des Patienten erkennen
- gleicht Symptome mit Datenbank ab
- schlägt geeignete Behandlung vor
- empfiehlt Arzt aufzusuchen
- speichert und wertet Gesundheitsdaten aus

Diagnostik: Biomarker-Signaturen identifizieren

Clinical Study | S Schweitzer, M Kunz and others | Plasma steroid profiling in adrenal tumors | **180**:2 | **117**–125

Plasma steroid metabolome profiling for the diagnosis of adrenocortical carcinoma

Sophie Schweitzer[1,*], Meik Kunz[2,3,4,*], Max Kurlbaum[1,5], Johannes Vey[2,3], Sabine Kendl[1], Timo Deutschbein[1], Stefanie Hahner[1,3], Martin Fassnacht[1,3,5], Thomas Dandekar[2,3] and **Matthias Kroiss[1,3,5]**

[1]Division of Endocrinology/Diabetology and Core Unit Clinical Mass Spectrometry, Department of Internal Medicine I, University Hospital Würzburg, [2]Department of Bioinformatics, Biocenter, University of Würzburg, [3]University of Würzburg, Comprehensive Cancer Center Mainfranken, [4]Chair of Medical Informatics, Friedrich-Alexander University of Erlangen-Nürnberg, Erlangen, Germany, and [5]University Hospital Würzburg, Central Laboratory, Core Unit Clinical Mass Spectrometry, Würzburg, Germany

*(S Schweitzer and M Kunz contributed equally to this work)

Correspondence should be addressed to M Kroiss or M Kunz
Email
Kroiss_M@ukw.de or meik.kunz@uni-wuerzburg.de

Statistisches Regressionsmodell identifiziert Blut-basierte 6-Steroid-Signatur mit hoher Genauigkeit

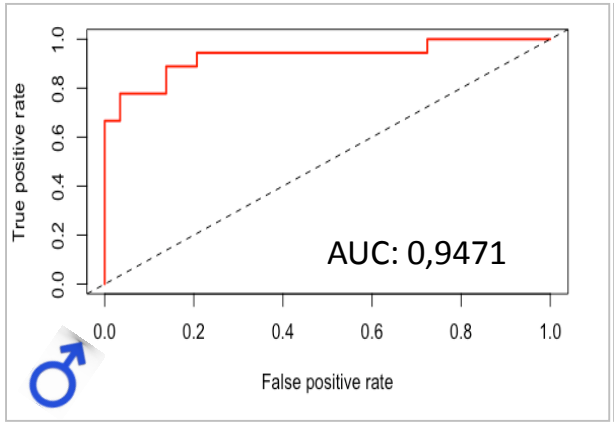
MEN		Diagnosis	
		Adenoma	Carcinoma
Prediction	Adenoma	28	3
	Carcinoma	1	12
		29	15

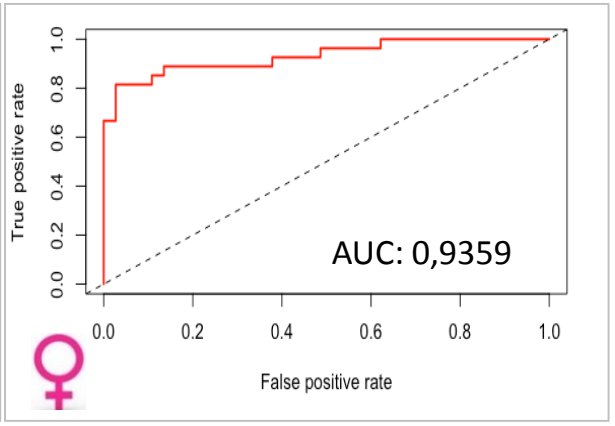
Sens: 80.0%; Spez: 97%;
PPV: 92%; NPV: 90%;
Accuracy: 91%

Männer	Frauen
Corticosteron	DHEAS
Progesteron	DHT
Estradiol	17-OHP
11-Deoxycortisol	11-Deoxycorticosteron
Androstendion	Androstendion
DHEA	DHEA

WOMEN		Diagnosis	
		Adenoma	Carcinoma
Prediction	Adenoma	36	6
	Carcinoma	1	21
		37	27

Sens: 78%; Spez: 97%;
PPV: 95%; NPV: 86%;
Accuracy: 89%



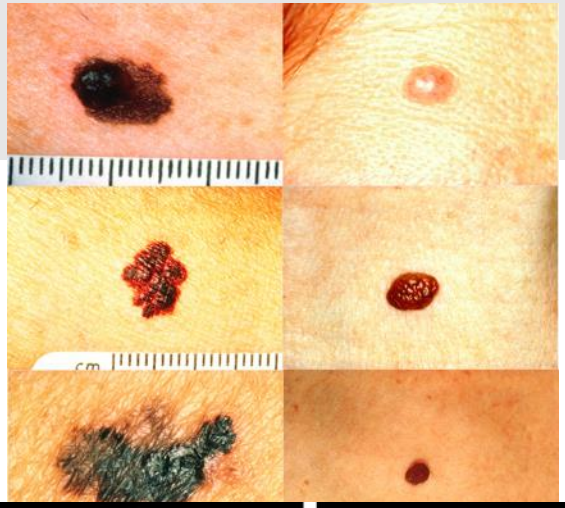


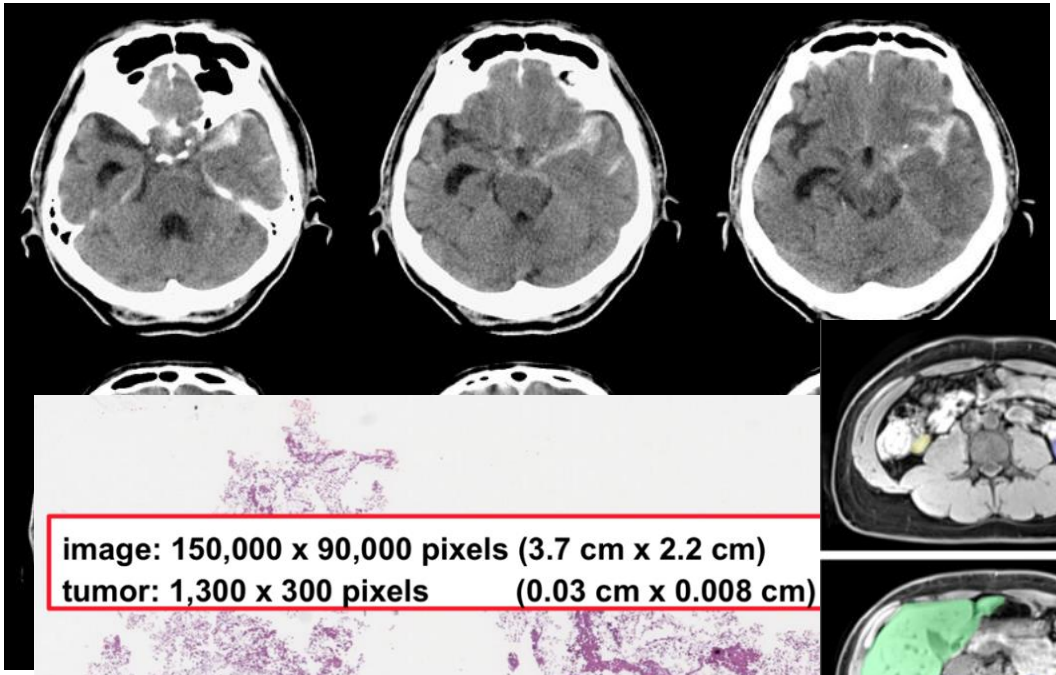
Schweitzer*, Kunz* et al., Europ. J Endoc 2019

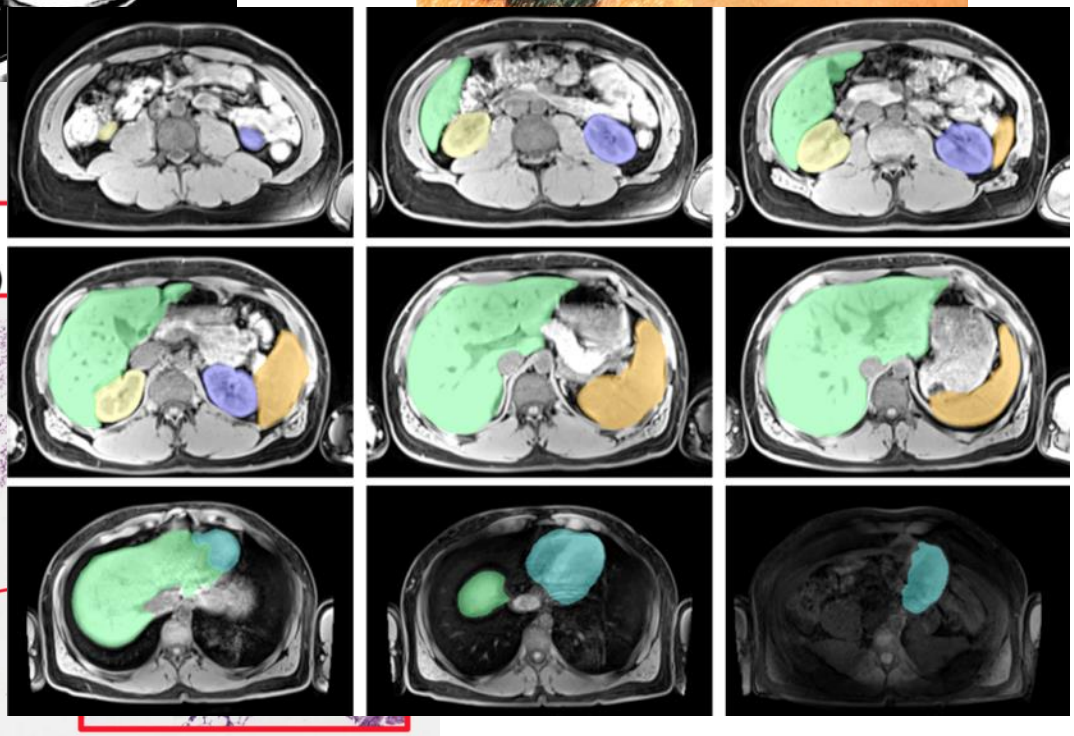
Radiologie (Imaging)

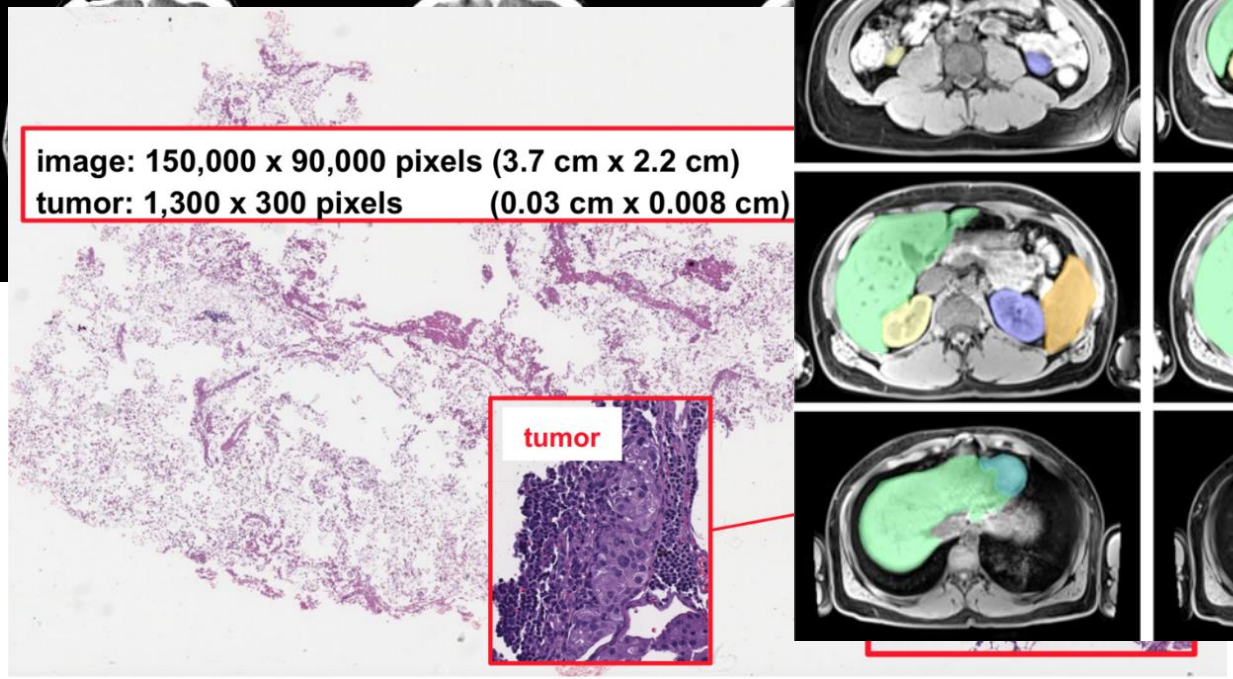
- Einsatz von KI bei Auswertung von bildgebenden Verfahren
  - Röntgenbilder
  - MRT
  - CT
  - Ultraschall
  - Hautbilder
  - histologische Präparate
- einmal trainiert, schnelle und kostengünstige Auswertung
- Unterstützung/Entlastung von Radiologen
- z.B. Vorsortierung von dringenden Fällen

Radiologie (Imaging)









Commons.wikimedia.org

Radiologie (Imaging)

- BioMind
- 225 Fälle
- AI: 87% in 15 min
- 30 Ärzte: 66% in 30 min

China Watch › Technology

AI defeats elite doctors in diagnosis competition

Advertisement feature for CHINAWATCH

share

Man vs machine: radiologist Zhang Junhai (left) of Shanghai Huashan Hospital reads a medical image display during the competition with BioMind in Beijing on June 30

thetelegraph.co.uk

Personalisierte Medizin

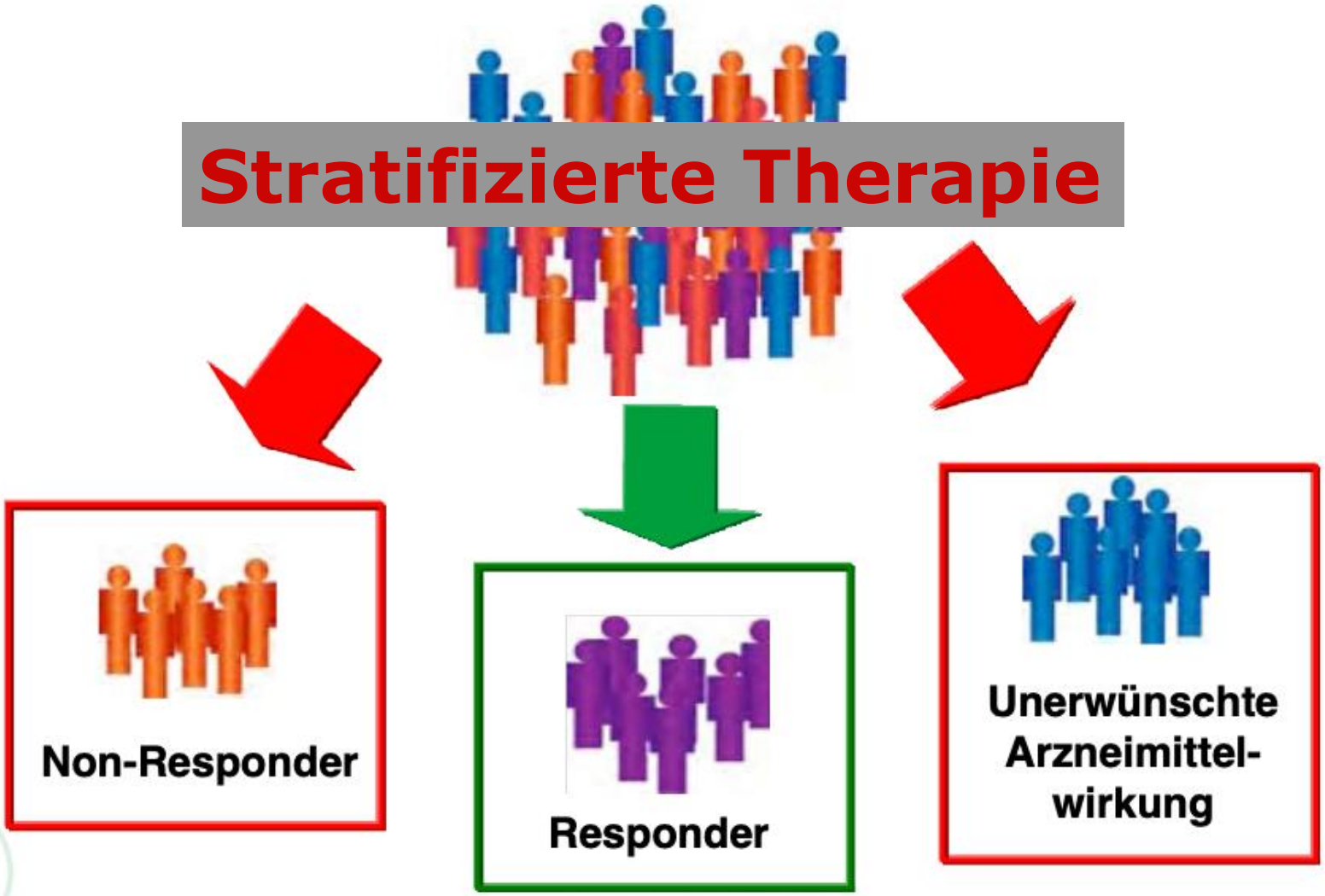
Was ist personalisierte Medizin?

- beste Therapie für den Patienten finden
  - Biomarker identifizieren
  - Patienten in klinisch relevante Subgruppen unterteilen
  - Therapieansprechen vorhersagen

Personalisierte Medizin

- Beispiel Darmkrebs
  - Cetuximab oder Panitumumab wirken bei fortgeschrittenem Darmkrebs nur, wenn KRAS noch nicht mutiert ist
  - KRAS bei 40% allerdings mutiert
  - Mutationsstatus von KRAS durch Gentest
  - Vortest mittlerweile verpflichtend

Personalisierte Medizin – Stratifizierung/optimale Therapie



AOK – Die Gesundheitskasse für Niedersachsen – Personalisierte Medizin – Chancen und Risiken Brigitte Käser 14.9.2012, vfa, Berlin

Natural Language Processing

- natürlichsprachliche Texte automatisiert in ein standardisiertes Format bringen

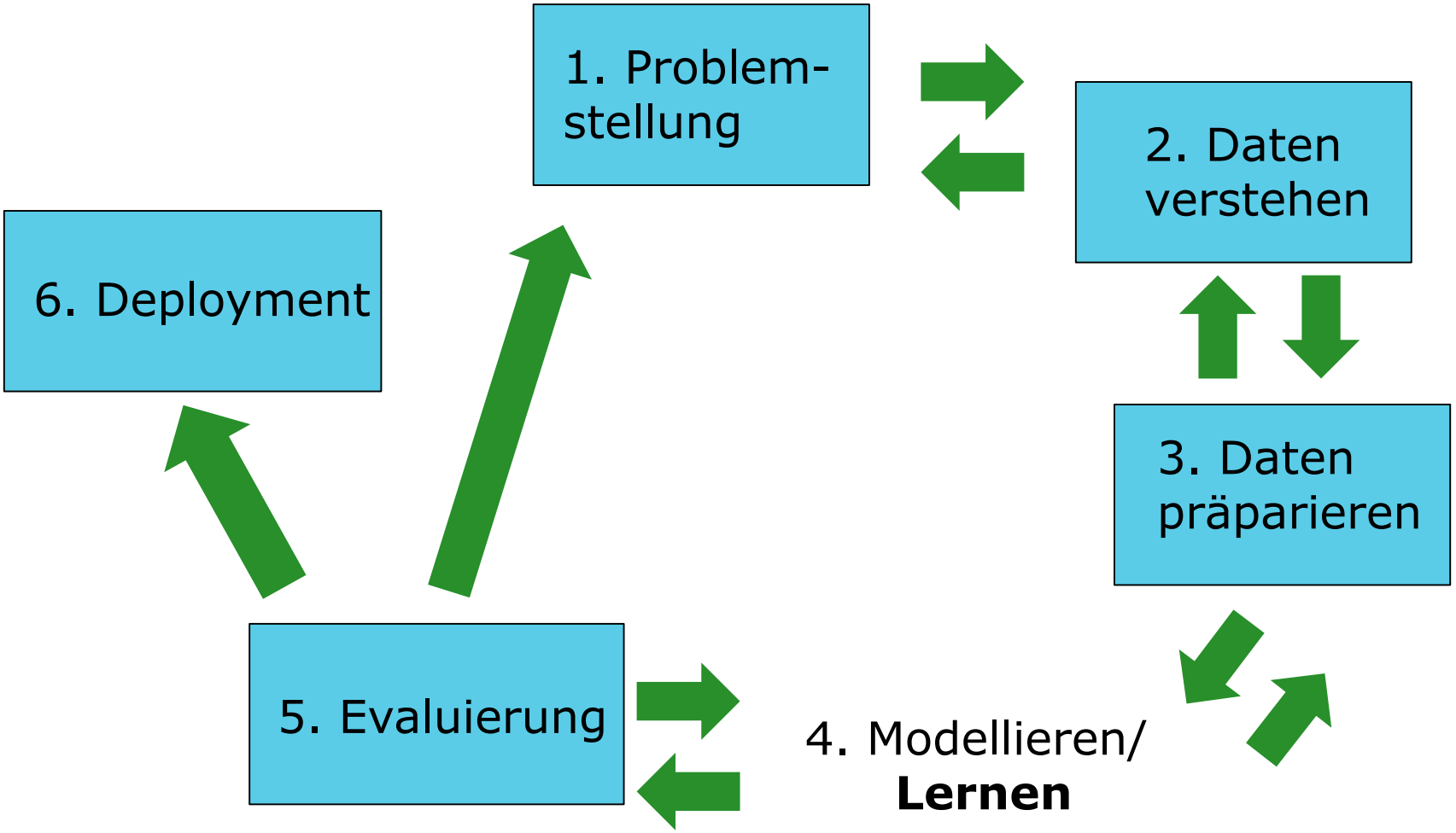
- natürlichsprachliche Texte:
  - radiologische Befunde
  - Arztbriefe
  - OP-Berichte

- große Mengen unstrukturierter Daten automatisiert und effizient interpretieren und in strukturierte Form bringen

- Erstellen von Statistiken (z.B. zur Ursachenforschung)

Vorgehen

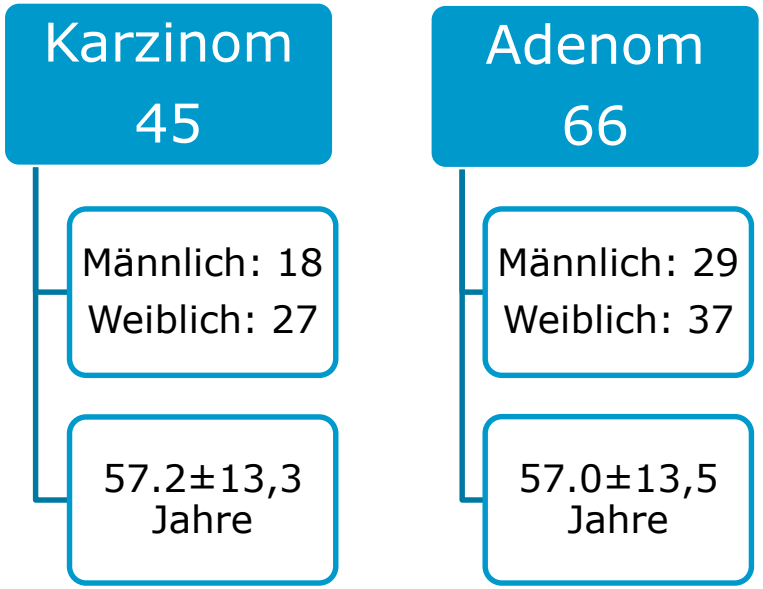
Vorgehen: Machine Learning



Nach CRISP-DM (Cross Industry Standard Process for Data Mining)

Anwendungsbeispiel: Blut-basierte Metabolomic von Nebennierenkarzinomen/-adenomen

111 Patienten: 15 Steroidhormone durch Liquid-Chromatographie-Tandem-Massenspektrometrie (LC-MS/MS)



Schweitzer*, Kunz* et al., Europ. J Endoc 2019

1. Problemstellung verstehen

Anwendungsbeispiel:
Differentielle Diagnose von Nebennierenrindentumoren

- Existieren Gruppen? Welche/wie viele Gruppen existieren
  - Zwei Gruppen: Adenom (gutartig) & Karzinom (bösartig)

- Formulierung der Aufgabenstellung
  - Binäre Klassifikation anhand numerischer Variablen

2. Daten verstehen

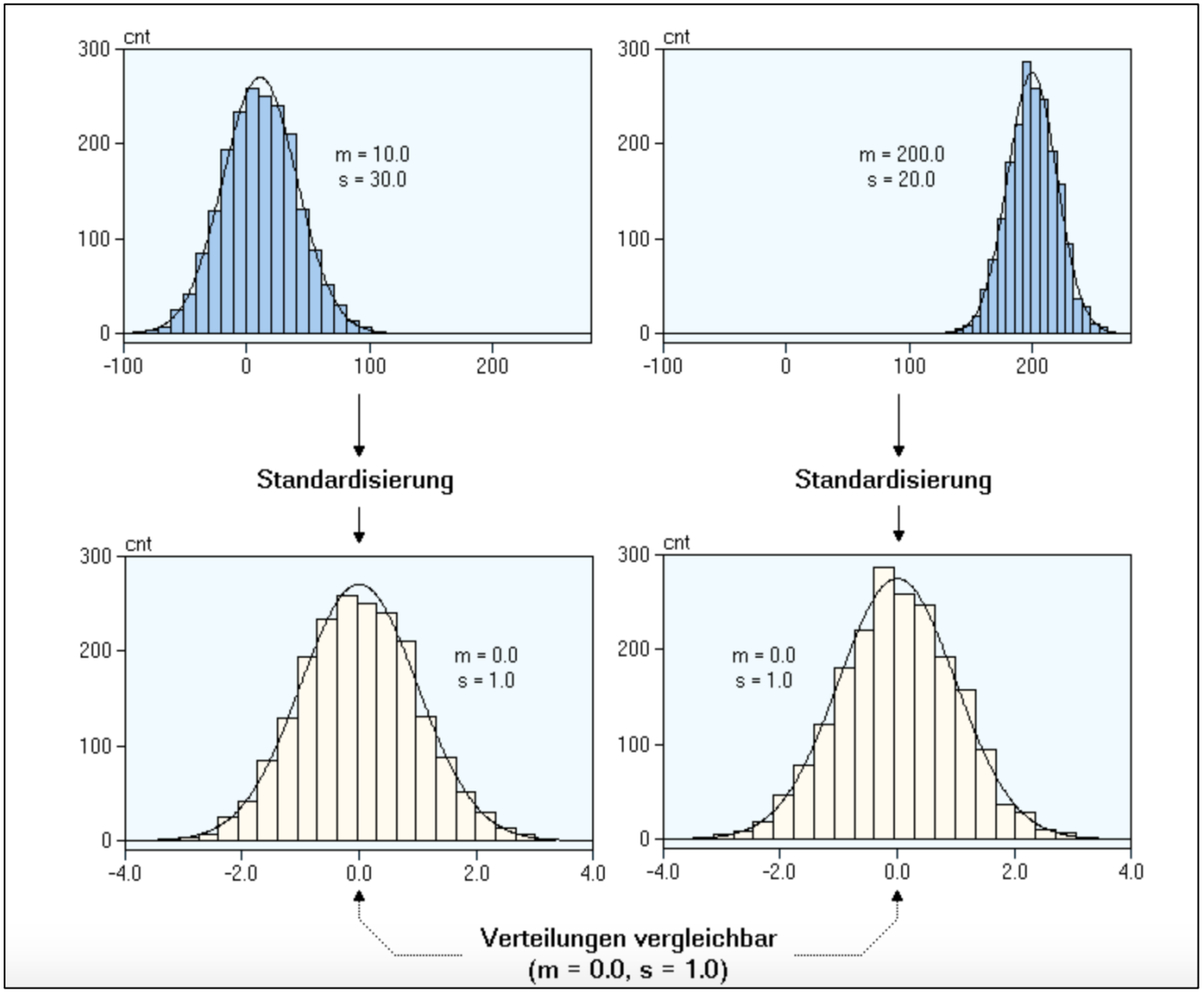
- Welche Daten stehen zur Verfügung?
  - Steroidkonzentration im Blut
- Welche Datentypen enthält der Datensatz?
  - Nur numerische Variablen
- Wie groß ist der Datensatz?
  - 111 Samples; 15 Variablen
- Wie ist das Verhältnis Samples (n) / Variablen (p)
  - $n > p$
- Sind die Klassen balanciert?
  - 66 Adenome; 45 Karzinome (59% zu 41%)

3. Daten präparieren

- Sind fehlende Werte vorhanden?
  - Imputation der fehlenden Werte (z.B. kNN, ...)
  - Löschen der entsprechenden Reihen/Spalten
- Wie ist die Datenverteilung bzw. die Varianz?
  - Transformation oder Standardisierung der Daten
- Sind Ausreißer vorhanden (technisch bedingt)?
  - Entfernen/Anpassend er Ausreißer (z.B. 1,5 x IQR)
- Sind korrelierende Variablen vorhanden?
  - Berücksichtigung von korrelierenden Variablen

3. Daten präparieren

- z-Transformation
- Standardisierung



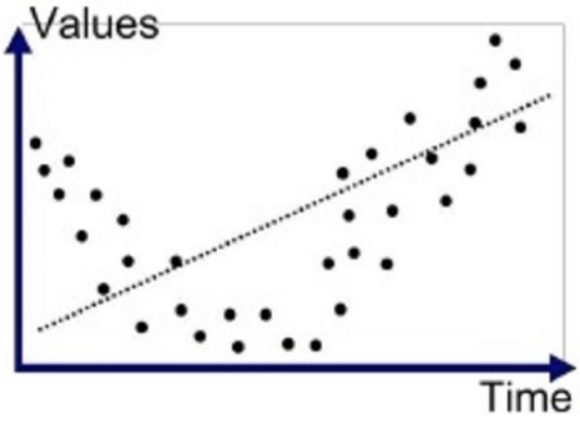
http://www.statistics4u.info/fundstat_germ/ee_ztransform.html

4. Modellieren

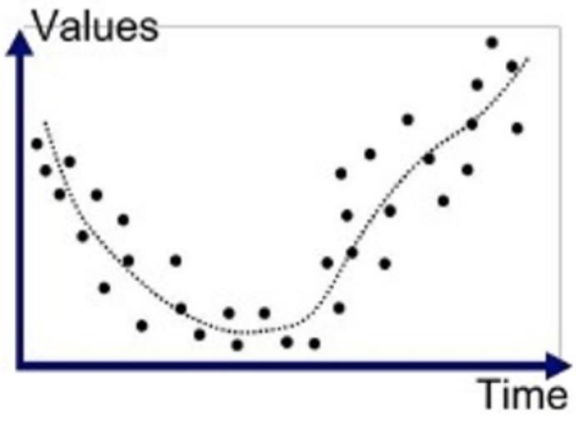
- Aufteilen der Daten in Trainings- und Testdatensatz
  - 70-80% Training; 20-30% Test
- Welche Techniken eignen sich für die vorliegenden Daten und Problemstellung?
  - Techniken für binäre Klassifikation
  - Techniken geeignet für relativ geringe Anzahl an Samples
- Feature Selection
  - Welche Variablen tragen zur Erklärung der Diagnose bei?
- Erstellen verschiedener Modelle
- Hyperparameter Tuning
  - Optimierung der Effizienz/Accuracy der Modelle durch Anpassung der Hyperparameter

4. Modellieren

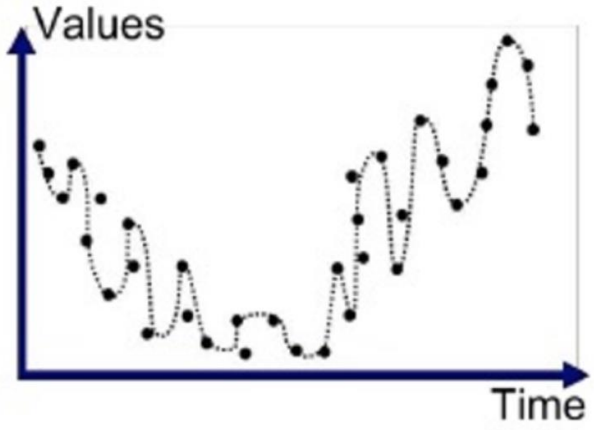
- Over-/Underfitting



Underfitting
(Modell kann zugrundeliegenden Trend nicht abbilden)



Good fit



Overfitting
(durch zu viele Variablen, Modell passt sich den Daten zu gut an, kann Rauschen nicht von reellen Signalen unterscheiden)

Modell: so einfach wie möglich, so komplex wie nötig

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

5. Evaluation

- Vorhersage des Testdatensatzes
  - Vorhersage der Diagnose von unbekannten Patienten
- Kreuzvalidierung
- Metriken zur Evaluation der Performance
  - Accuracy
  - Recall/Precision
  - AUC

6. Deployment

- ML-Modelle in die Anwendung bringen – für Kliniker bereitstellen
  - App
  - GUI
  - API

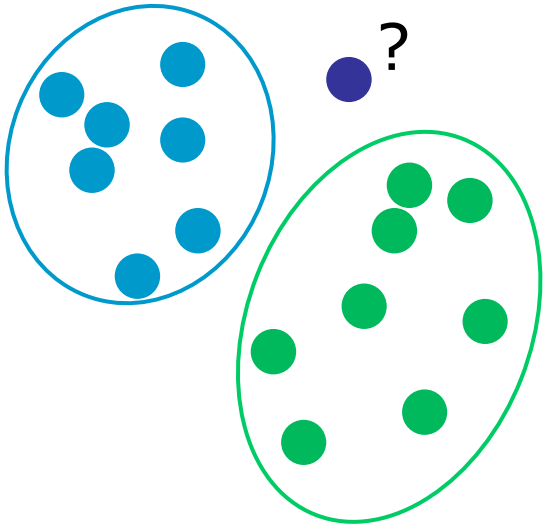
- Datenintegration
  - Einspeisen von neuen Daten
  - Anonymisierung/Randomisierung

Themen der Vorlesung

Machine Learning

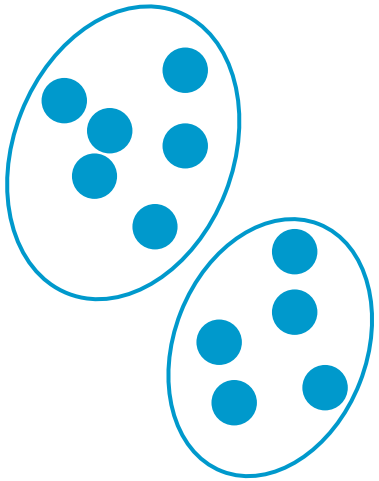
Machine Learning kann grob in drei Bereiche gegliedert werden:

Klassifikation



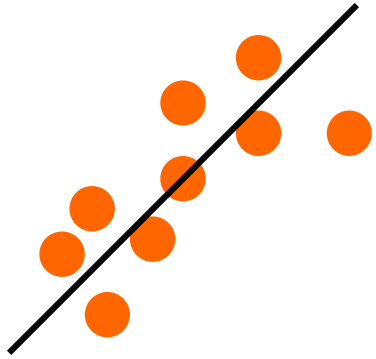
- Gruppen existieren

Clustering



- Keine Gruppen existieren

Regression



- Trends identifizieren

Biologie/Medizin: Vom Genotyp zum Phänotyp

Genotyp
(Sequenz)

Phänotyp
(Organismus)

Sequenzen
DNA, RNA, Proteine

Strukturen
Vorhersagen, Interaktionen

Netzwerke
Regulation, Interaktion

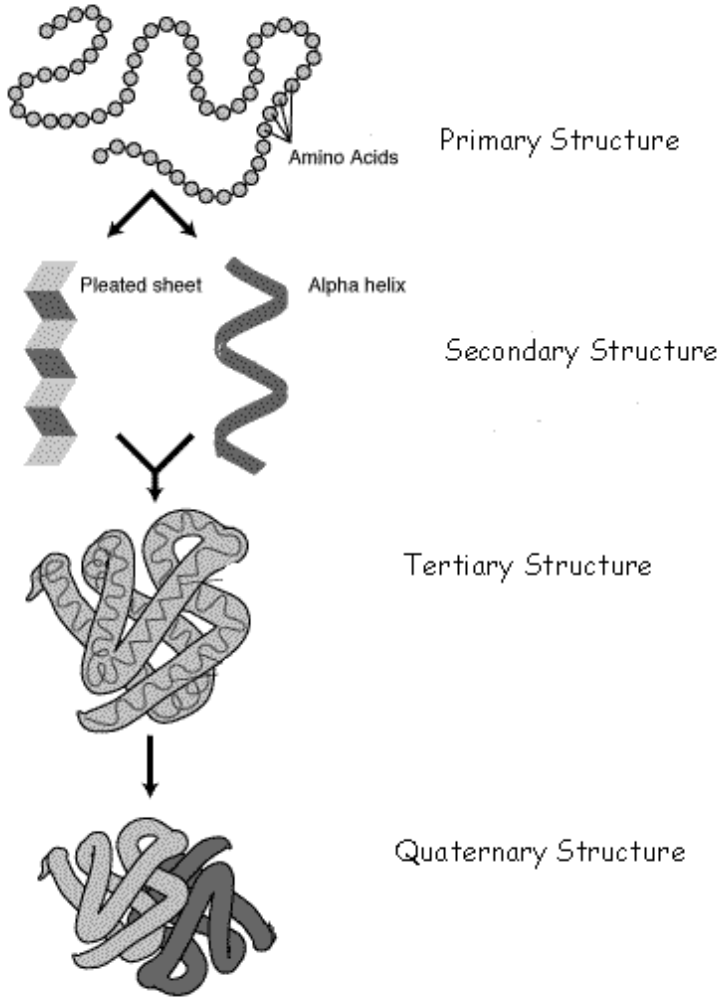
Ziel: Biologisches Verständnis zellulärer Prozesse

RNA-Strukturvorhersage

- basierend auf Faltungsenergie, optimale Basenpaare (Programme: mFold, RNAfold, ViennaPackage)
- Wichtige Algorithmen (dynamische Programmierung)
  - Nussinov-Algorithmus
    - Einfaches Modell zur Berechnung der maximalen Basenpaarungen
    - Matrix mit 1/2n(n+1)+n-1 Einträgen: Speicherbedarf=O(n^2), Laufzeit=O(n^3)
  - Zuker-Algorithmus
    - Weiterentwicklung Nussinov-Algorithmus, Nearest-Neighbour Model zum Finden der minimalen freien Energie; Speicherbedarf=O(n^2), Laufzeit=O(n^3)
  - Sankoff-Algorithmus
    - comparative sequence analysis (kombiniert Sequenzalignment, Faltung, Phylogenie)
    - multiples Alignment für Sequenz-Struktur-Konservierung in verschiedenen Sequenzen/Organismen mit ähnlicher Funktion; Speicherbedarf=O(n^4), Laufzeit=O(n^6)

Protein zeigt 1D-, 2D- und 3D-Struktur

- Lineare Abfolge von monomeren Untereinheiten
- Struktur eines Proteins durch Abfolge von Aminosäuren festgelegt
- Struktur (Aminosäuren) ermöglichen breites Spektrum an biologischen Funktionen

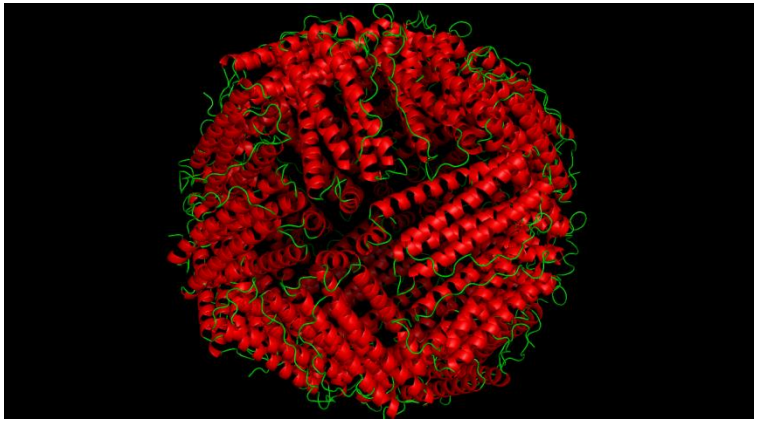


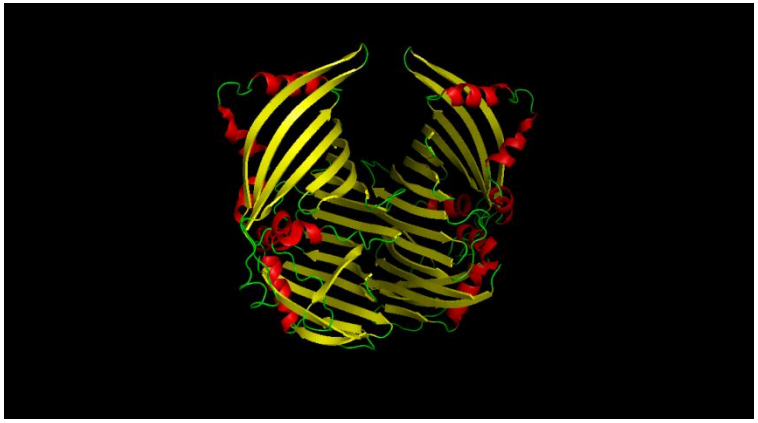
Levinthal'sches Paradoxon

- H-Brücken zwischen CO- und NH-Gruppen ermöglichen unzählige Konformationsmöglichkeiten
- Proteinfaltungsproblem: Proteine werden nicht durch Versuch und Irrtum gefaltet
- Überlegung:
  - Protein mit 100 aa, jede aa in 3 Konformationen (sterische Komplikationen, sehr konservative Annahme)
  - Anzahl aller möglichen Strukturen: $5x10^{+47}$
  - Faltung einer möglichen Struktur in 10^{-13}s
  - Alle Strukturen gefaltet: $5x10^{34}$s oder $1.6x10^{27}$ Jahre
- Es gibt also häufig genutzte Strukturen

Protein-Strukturvorhersage

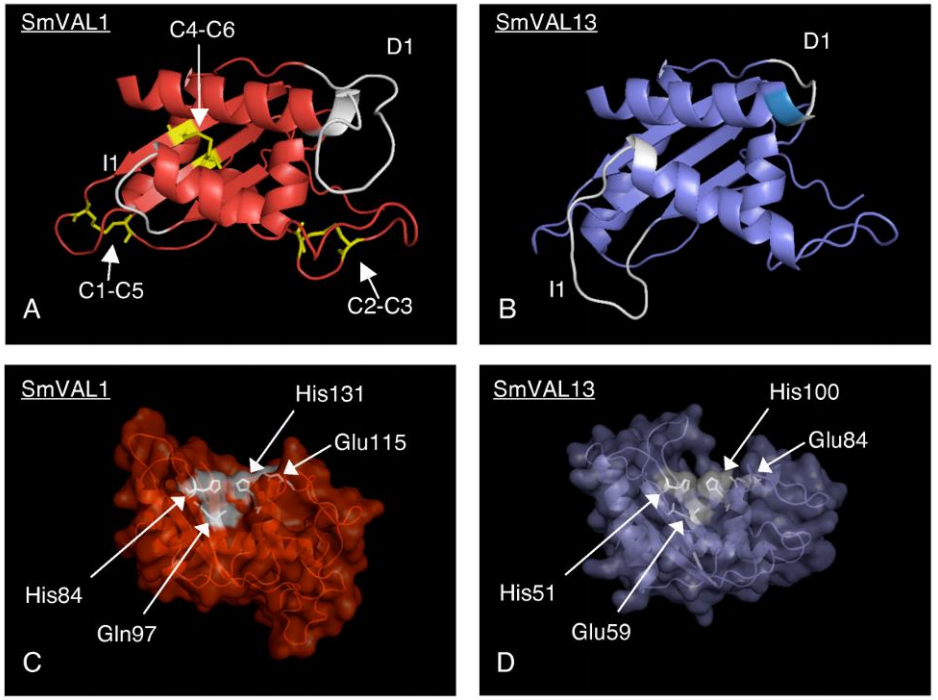
- Homologie-basierend
- *ab initio*
  - Sekundärstrukturen
  - Tertiärstruktur





Strukturvorhersage: Homologie-basierend

- Evolutionär Verwandte Proteine besitzen ähnliche Struktur
- bekannte Struktur dient als Vorlage für unbekannte Struktur
- Voraussetzung:
  - Sequenz und Struktur eines verwandten Proteins sind bekannt



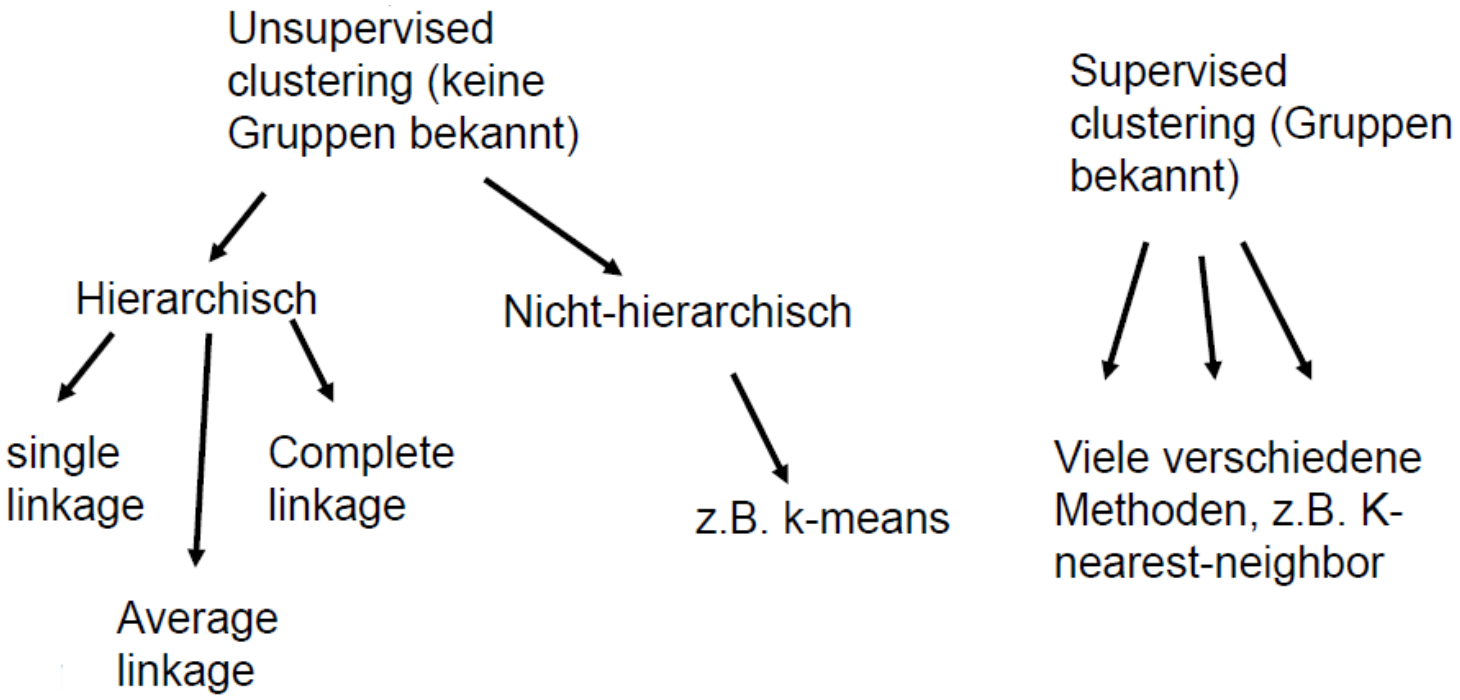
Strukturvorhersage: *ab initio*

- Vorhersage von Sekundärstrukturmerkmalen nur aus der Sequenz
- Chou-Fasman-Methode
  - Score für jede Aminosäuren, die Neigung zum Auftritt in Sekundärstruktur beschreibt (aus bekannten Strukturen und Sequenzen abgeleitet), manche Aminosäuren kommen nicht oder selten in definierter Struktur vor
  - Suche nach Anfangspunkten für Struktur (4 von 6 aa besitzen Score > Schwellenwert)
  - Erweitern der Struktur in beide Richtung bis Score (im Fenster 4) unter Schwellenwert für Struktur (Sliding Window-Ansatz)
- Garnier-Osguthorpe-Robson-Methode (GOR-Methode)
  - Ähnlich zu Chou-Fasman
  - Aminosäuren, die die aktuelle flankieren, beeinflussen die Fähigkeit diese eine definierte Struktur zu bilden

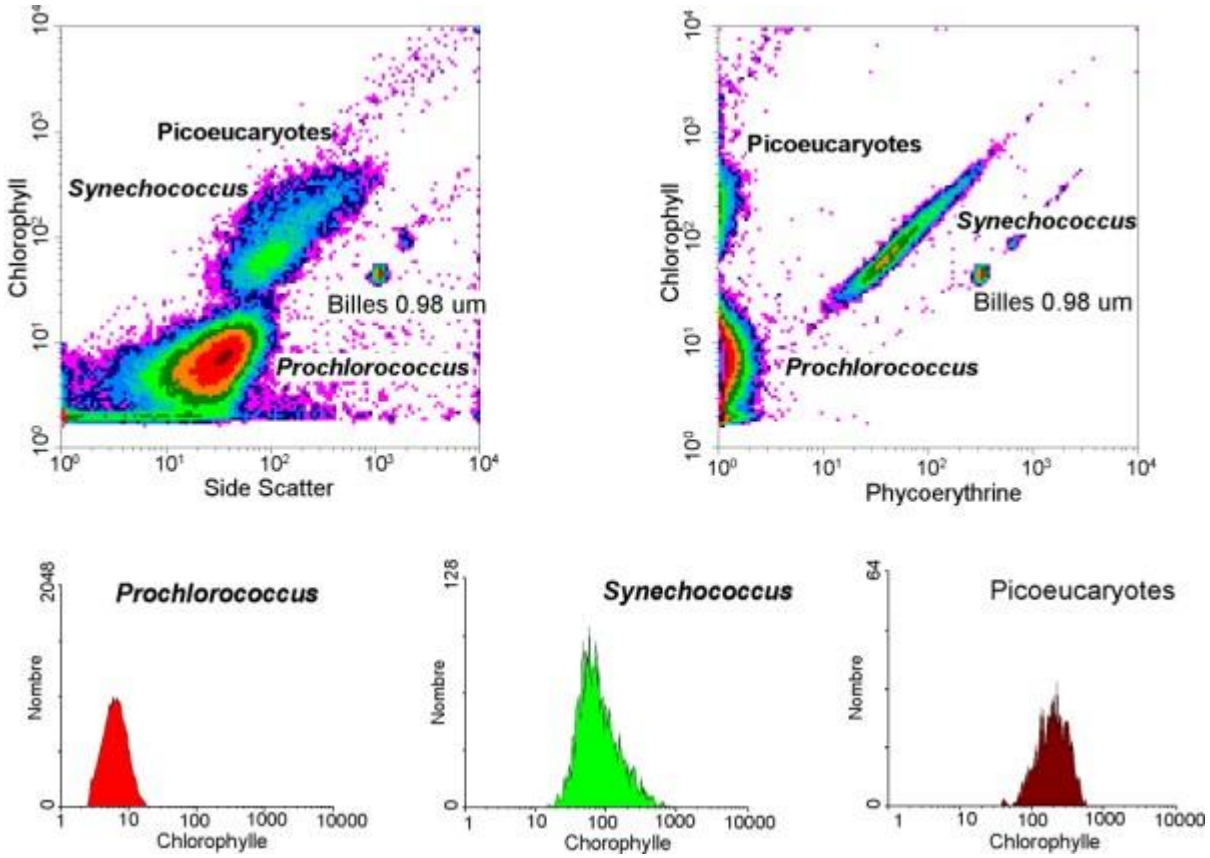
Tertiärstrukturvorhersage

- Problem: Primärstruktur gibt keine Auskunft über Anordnung im Raum, ABER: Funktion abhängig von Tertiärstruktur
- Tertiärstruktur basiert auf physikalisch-chemische Eigenschaften der Aminosäuren in Primärstruktur
  - sterische Komplikationen (Torsionswinkel)
  - hydrophober Effekt, interne Bindung
  - generelle Minimierung der freien Enthalpie
  - Optimierung der Energiefunktion
- Sehr rechenaufwendig und ungenau (nur bei kleinen Proteinen angewendet)
  - Optimierungsalgorithmen für Näherungslösungen basierend auf MFE-Minimierung helfen (Genetischer Algorithmus, Monte-Carlo, Simulated Annealing)

Verschiedene Clustermethoden

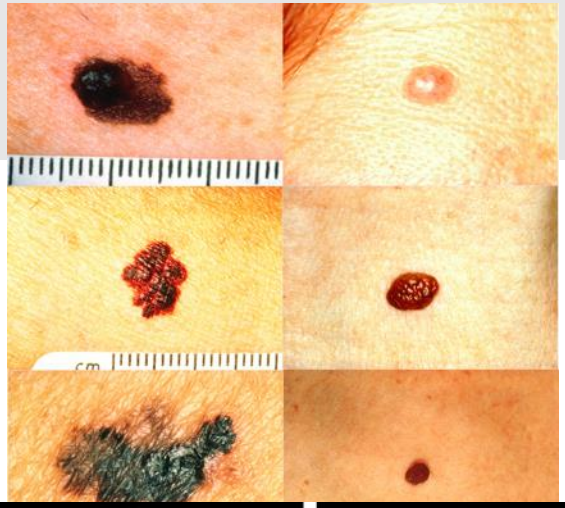


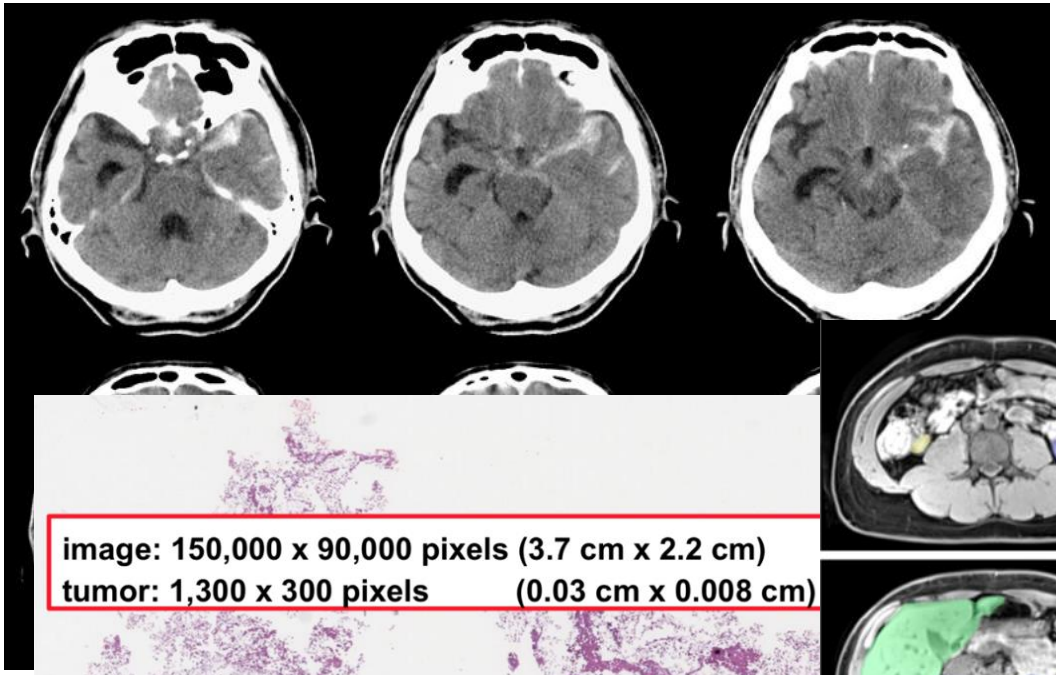
Durchflusszytometrie: SOM, t-SNE

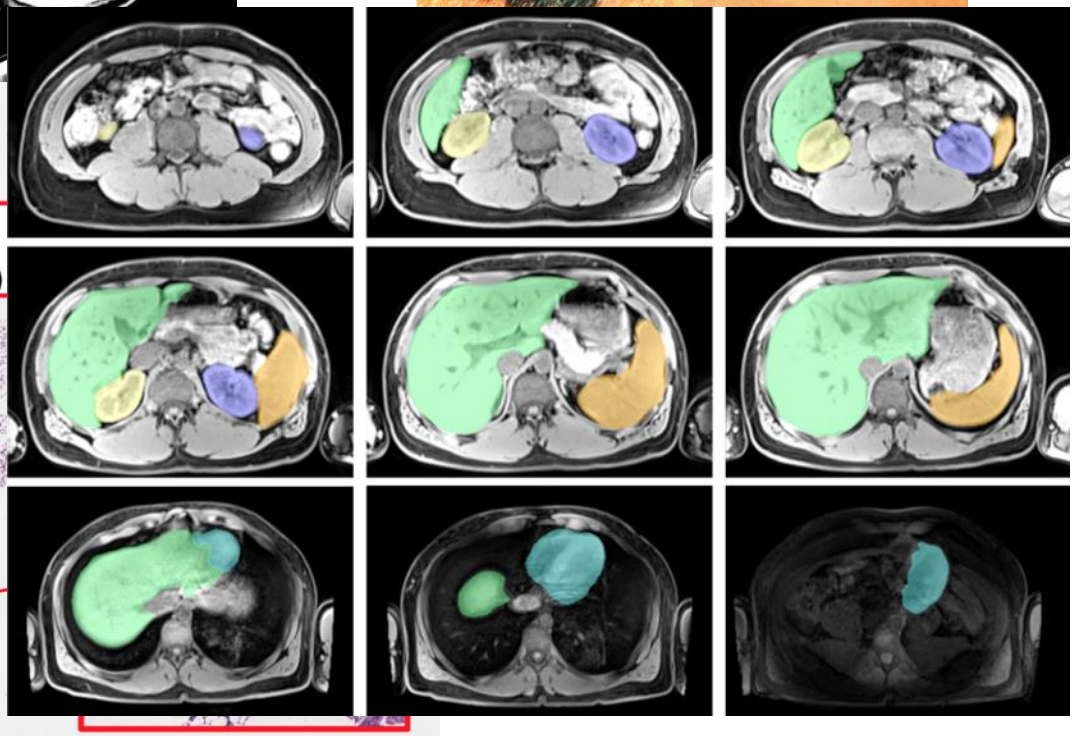


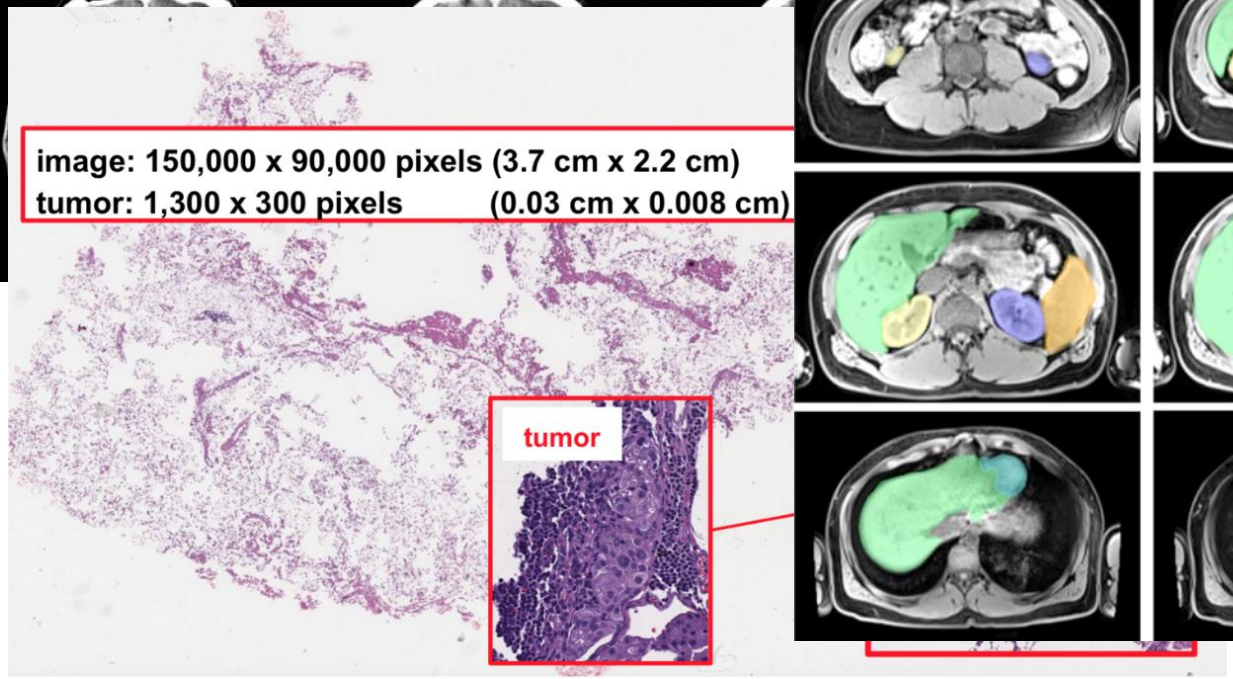
https://upload.wikimedia.org/wikipedia/commons/c/c7/Picoplancton_cytometrie.jpg

Radiologie (Imaging)



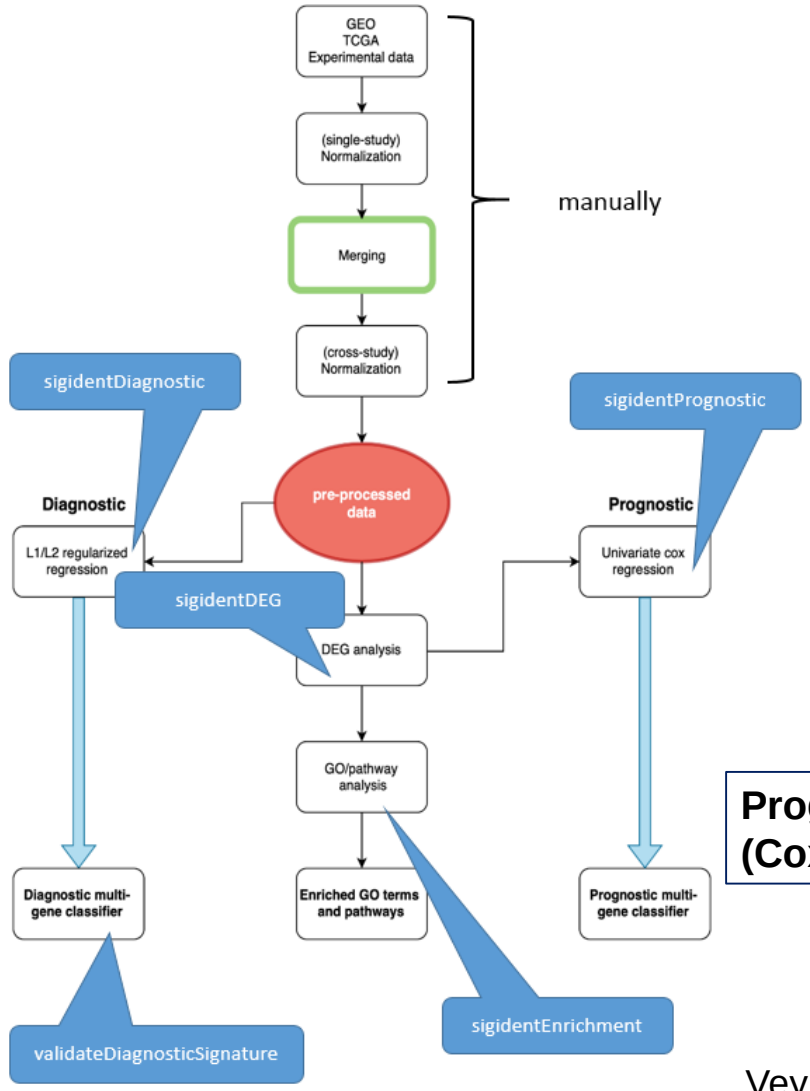






Commons.wikimedia.org

Integrierte R-Pipeline für die systematische Berechnung von Signaturen



Diagnostisch (Regression: GRM Klassifikation/Regression: SVM, Decision tree (RF, GBM))

Prognostisch (Cox Reg)

Vey et al., Cancers 2019

Übung

1. Beschreiben Sie verschiedene Anwendungsgebiete des Maschinellen Lernens in der Medizin (an je einem Beispiel erklären).
2. Nennen Sie 3 Bereiche des Maschinellen Lernens .