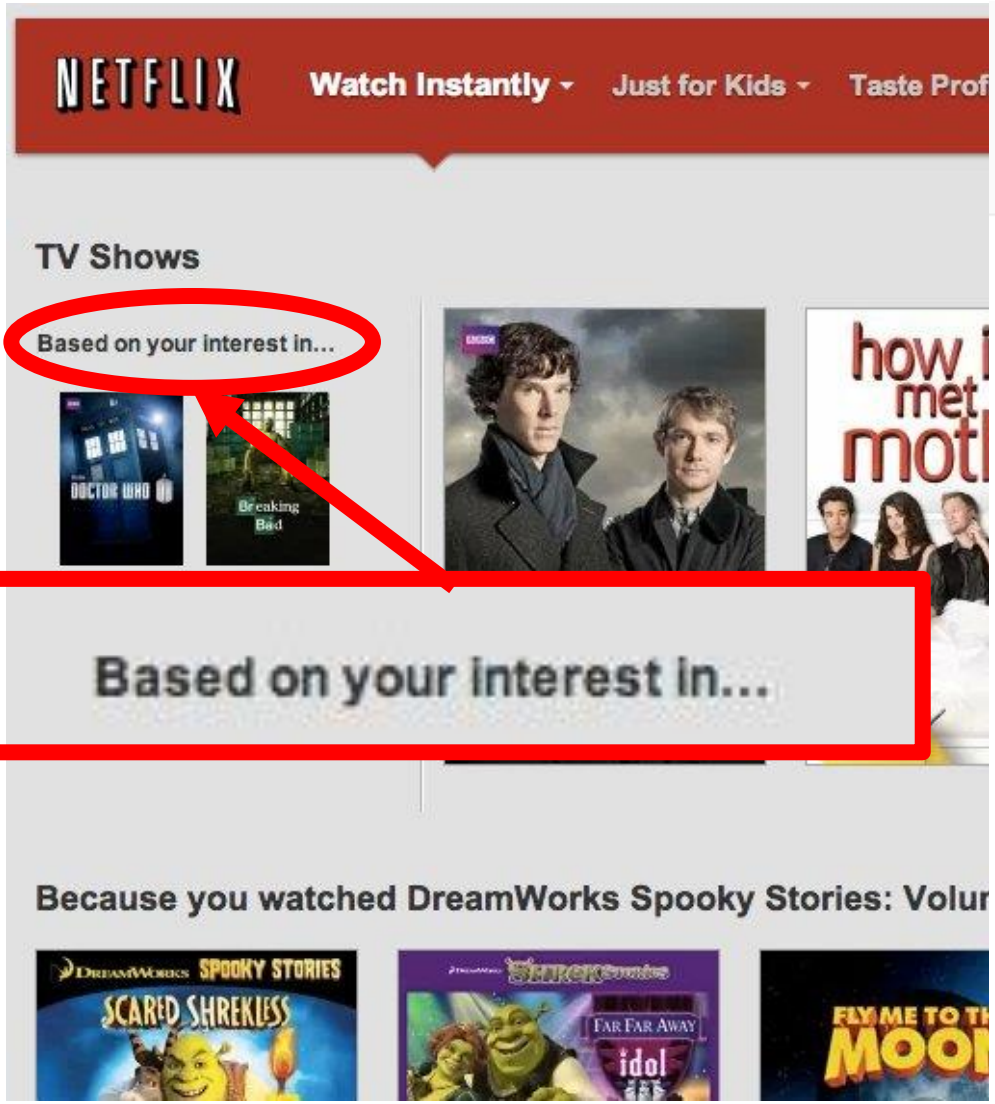
Einführung in Data Science

Unser Plan für heute:

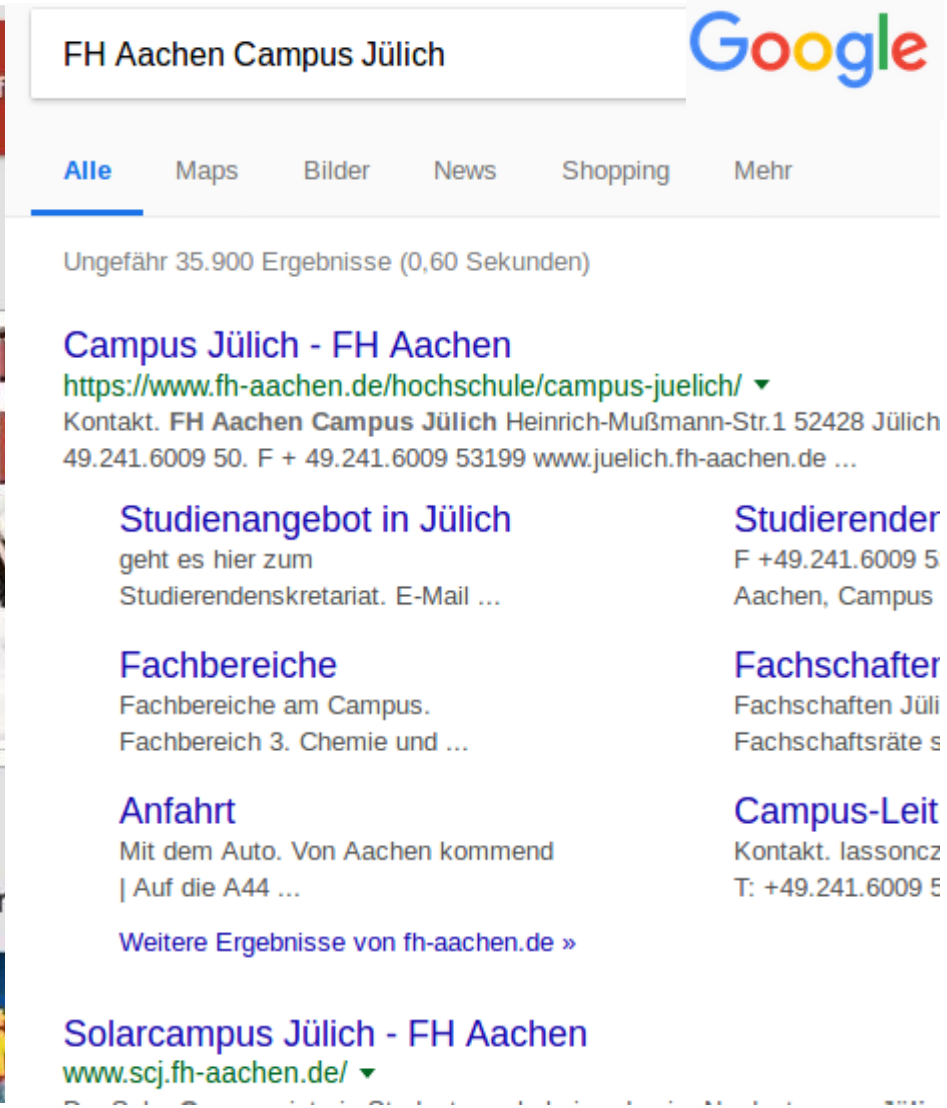
1. Was ist Data Science?
2. Fallbeispiel
3. Organisatorisches

(*) Ja, Sie finden die Folien später als PDF-Datei in ILIAS.

Empfehlungsdienste



Web-Suche



„Neuronale Kunst“

Gatys et al. arXiv 1508.06576

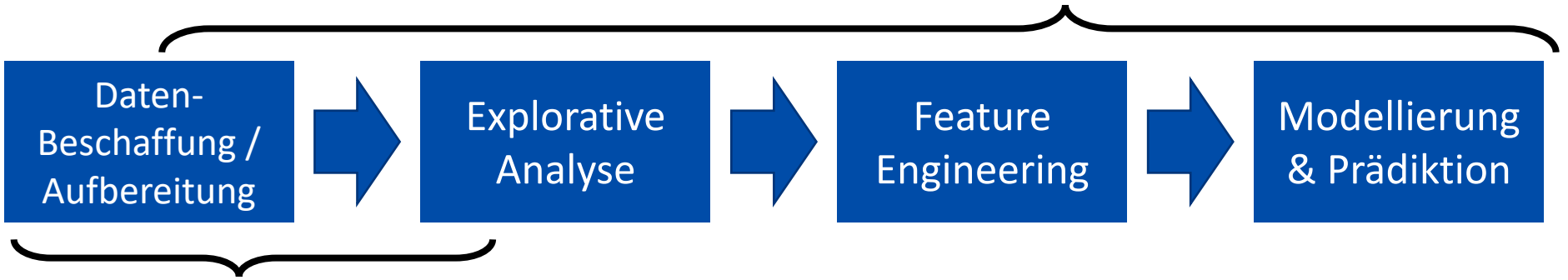
Hautkrebs-Diagnostik

1. Was ist Data Science?

Überblick

Erkenntnisse
Vorhersagen
Daten-Produkte

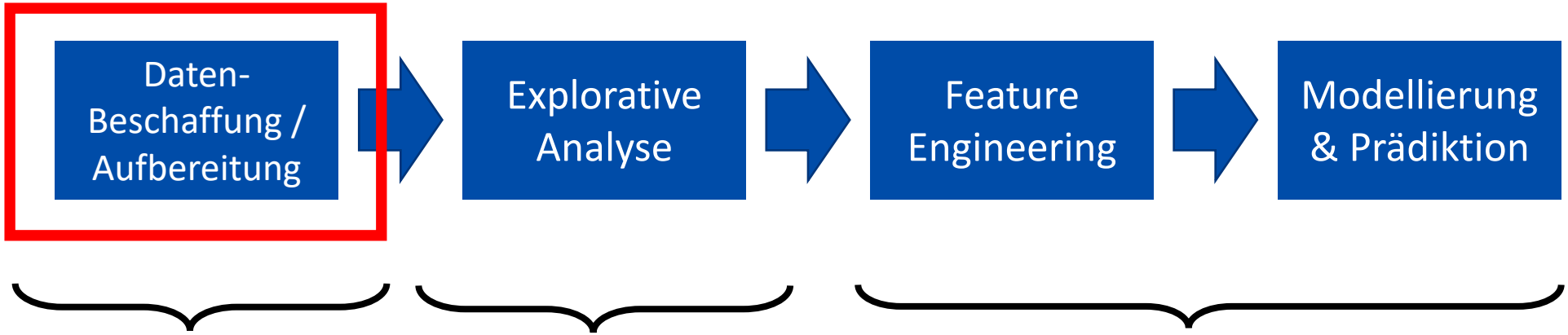
Data Scientist
Daten analysieren, visualisieren, interpretieren
Zusammenhänge erkennen



Data Engineer
Daten sammeln, speichern, bereinigen,
zusammenführen, bereitstellen, ETL

IT-Infrastruktur
Daten-Schnittstellen

Überblick



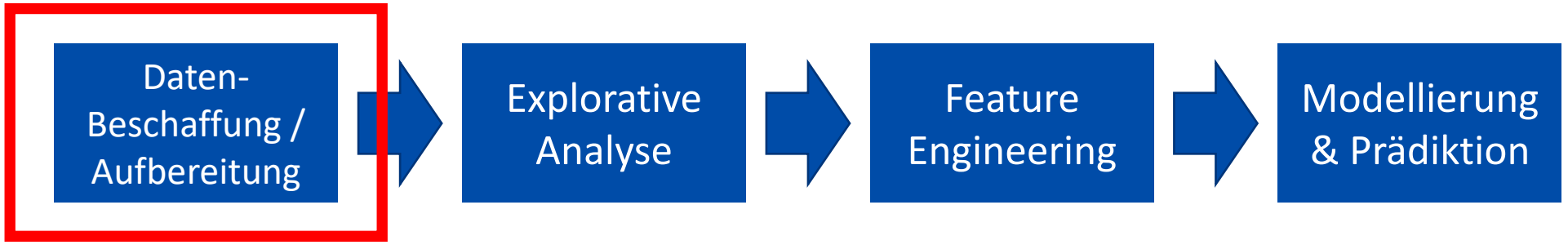
ETL Strecke
Extract
Transform
Load

EDA
Exploratory
Data
Analysis

„Machine Learning“
(Begriff wird leider oft unscharf für die Bezeichnung beider Schritte benutzt)

Typische Bezeichnungen im Business Jargon

Datenbeschaffung & -aufbereitung

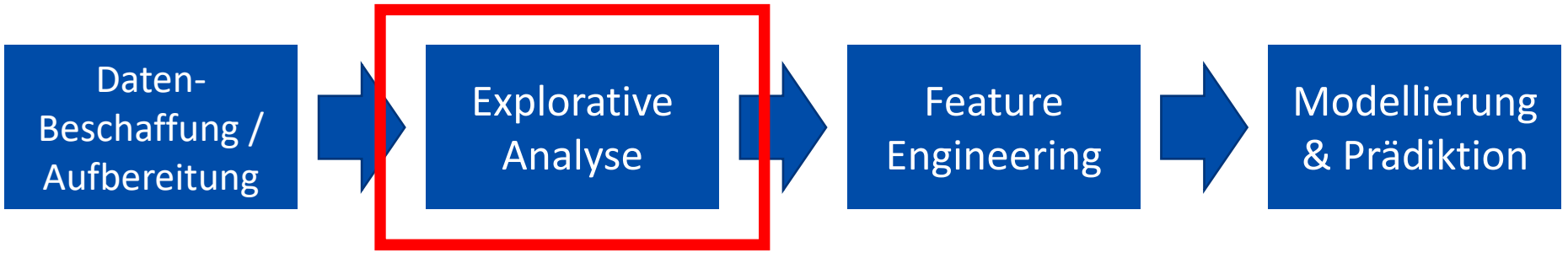


- **Datenbeschaffung**
  oft: unstrukturierte, heterogene Quellen
  (Dateien, Datenbanken, APIs, Webservices, Webscraping)
- **Datenbereinigung/-transformation**
  z.B. Anonymisierung, Artefakt-Detektion und –behandlung, Datenzusammen-führung
- **Visualisierung / Schnittstellen**

Typisch für Datenaufbereitung:

- Kette von Arbeitsschritten (Pipelines)
- **Domänenspezifisch!**
- abhängig von Datenmenge (Small/Big Data; verteiltes Speichern, verteiltes Rechnen)

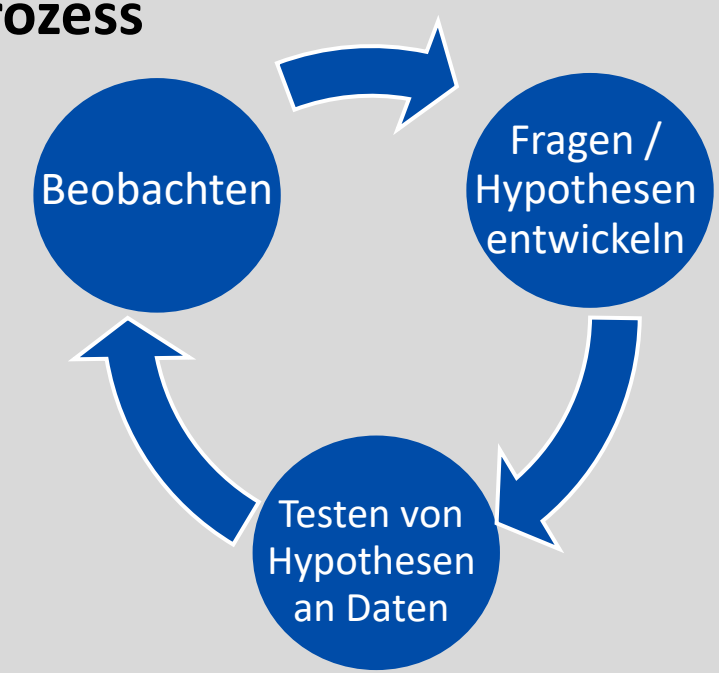
Explorative Datenanalyse



Typische Werkzeuge

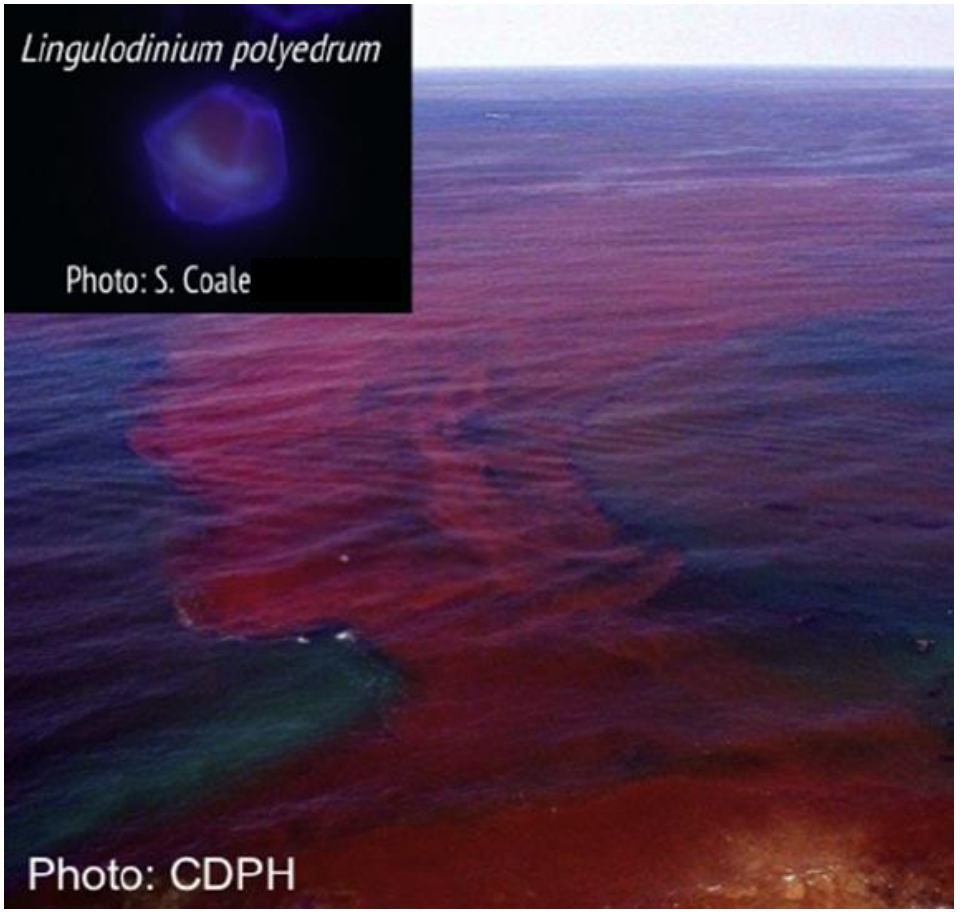
- (interpretierte) Programmiersprache (z.B. Python, R, Matlab)
- Visualisierung
- deskriptive Statistik (Beschreibung von Stichproben)
- statistisches Testen
- Schätzer für Beziehungen zwischen Variablen (bi-/multivariate Analysen)
- Cluster-Methoden
- …

Prozess



Explorative Datenanalyse | Beispiel

- Fischerei-Industrie: Millionenverluste durch gefährliche Algenblüten
- Fernziel: Vorhersage und gezielte Verhinderung von Algenblüten

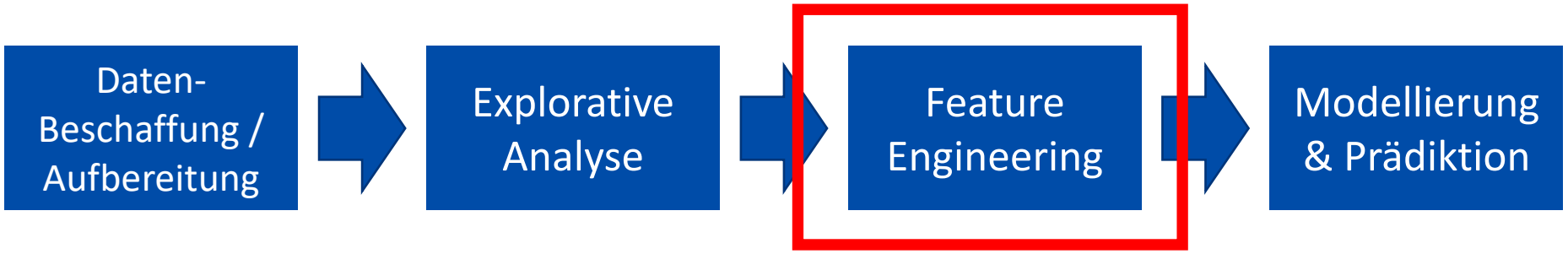


Anderson et al, Coastal and Marine Hazards, 495-561, 2015

Explorative Datenanalyse | Beispiel

Bialonski et al, J. Plankton Res. 38, 1077-1091, 2016

Feature Engineering



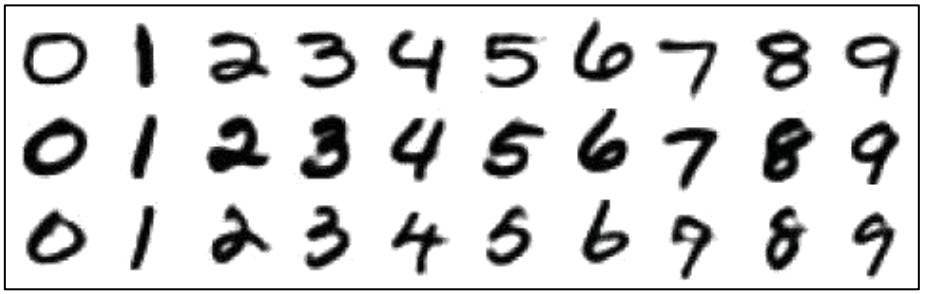
Typische Werkzeuge

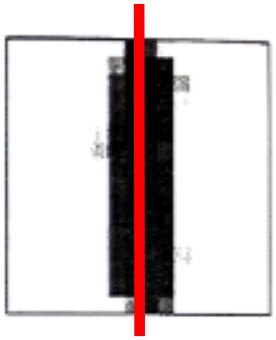
- Methoden der explorativen Analyse ([interpretierte] Programmier-sprache, Statistik, Cluster-Methoden, Visualisierung)
- Domänenexpertise
- (ad hoc) Heuristiken

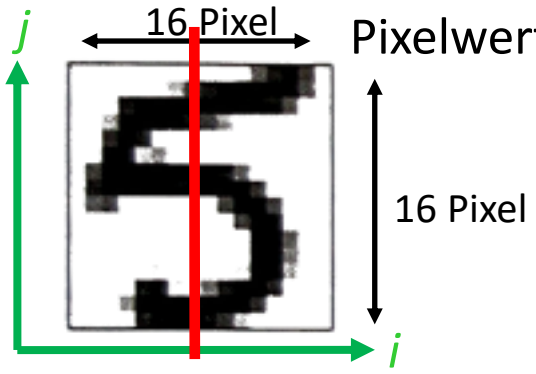
- **Features:** Merkmale der Daten, die *nützlich* für weitere Aufgaben sind (z.B. für Modellierung, Klassifikation, etc …)
- **Feature Engineering:** Prozess, Features zu finden, die die Daten in nützlicher Weise charakterisieren

Feature Engineering | Beispiel

- Automatische Erkennung handschriftlicher Postleitzahlen
- Datensatz:
  9298 digitalisierte Ziffern
  (US Postal Service)







Pixelwerte $p_{ij} \in [0, 1]$

Features

F

Schwärze

$$x_1 = \sum_{i,j} p_{ij}$$

Feature Engineering | Beispiel

Model Fitting:
Prozess, ein Model an die Daten anzupassen

Model: $f(x_1) = m x_1 + c$

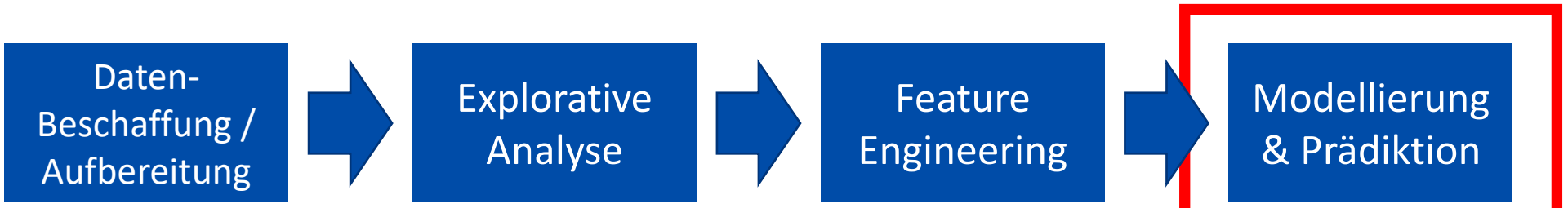
freie Parameter:
m, c

Betrachte Ziffer mit Features (x_1, x_2) :

„1“ falls $x_2 < f(x_1)$

„5“ falls $x_2 > f(x_1)$

Modellieren mit maschinellem Lernen



Unüberwachtes Lernen

Funktion $f : X \rightarrow ?$

Daten: $(x_i), \quad i = 1, \ldots, N$

Überwachtes Lernen

Funktion $f : X \rightarrow Y$

Daten: (x_i, y_i)

Verstärkendes Lernen

Strategie $\pi : S \rightarrow A$

Daten: Environment, Reward

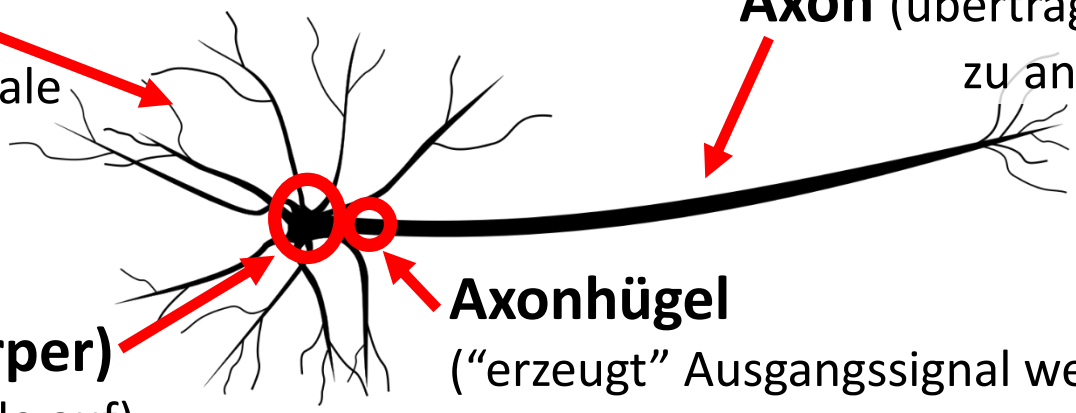
Symbolisten	Analogisten	Konnektionisten	Evolutionären	Probabilisten
▪ Entscheidungs-bäume ▪ Zufallswälder ▪ …	▪ Stütz-vektoren-Maschinen (SVM) ▪ …	▪ künstliche neuronale Netze ▪ Tiefes Lernen (Deep Learning)	▪ genetische Algorithmen ▪ symbolische Regression ▪ …	▪ graphische Modelle ▪ …

Pedro Domingos. Basic Books 2015

Das künstliche Neuron

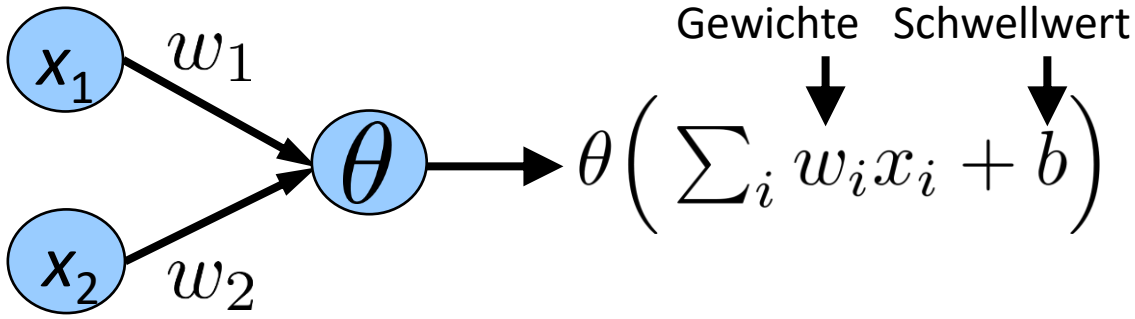
Dendriten
(übertragen Signale von anderen Neuronen)

Axon (überträgt Ausgangssignal zu anderen Neuronen)

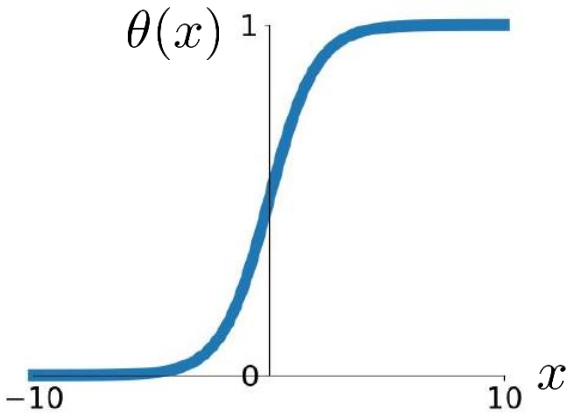


Soma (Zellkörper)
(summiert Signale auf)

Axonhügel
(“erzeugt” Ausgangssignal wenn Signalsumme einen Schwellwert übersteigt)



Aktivierungsfunktion

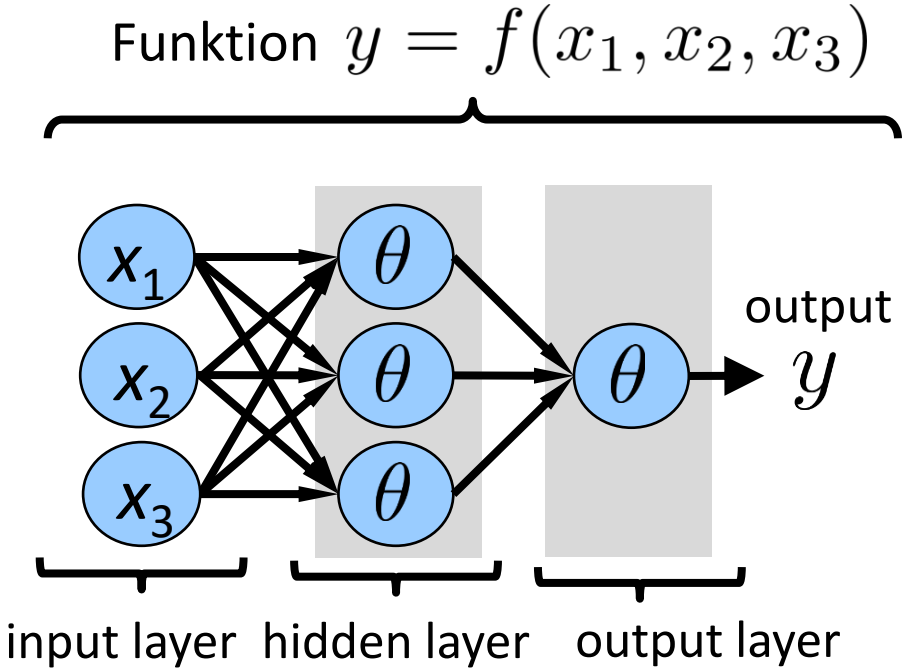


Freie (zu lernende) Parameter:

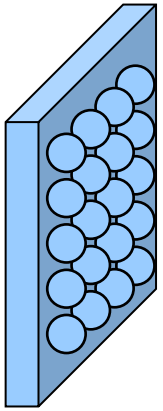
$$w_1, w_2, b$$

Künstliche neuronale Netze

- bestehen aus verknüpften künstlichen Neuronen
- Neuronen sind in Schichten (layers) angeordnet



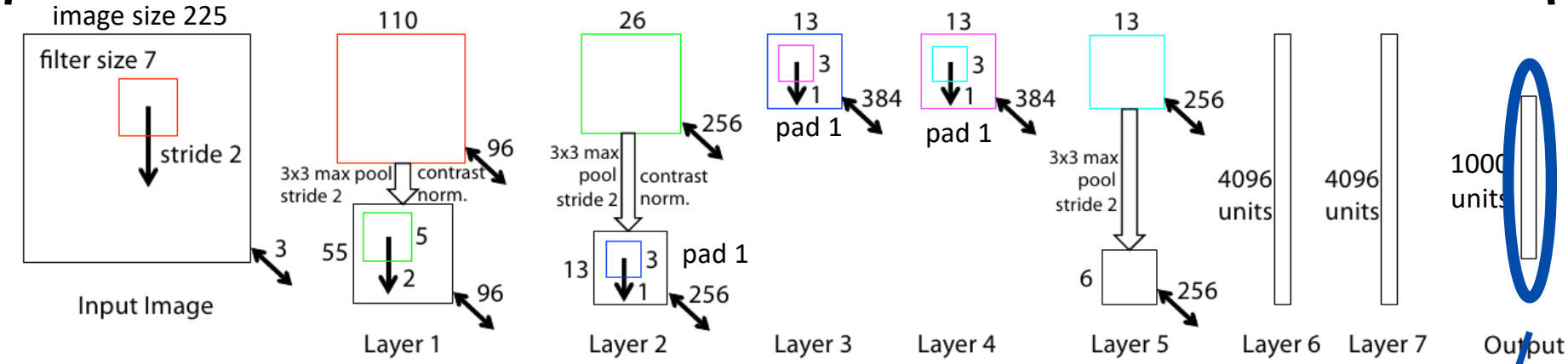
- auch zwei-dimensionale (oder mehr-dimensional) Schichten möglich



Bildklassifikation mit neuronalen Netzen

Funktion $\vec{y} = f(\text{Bild})$

ca. 57m Parameter



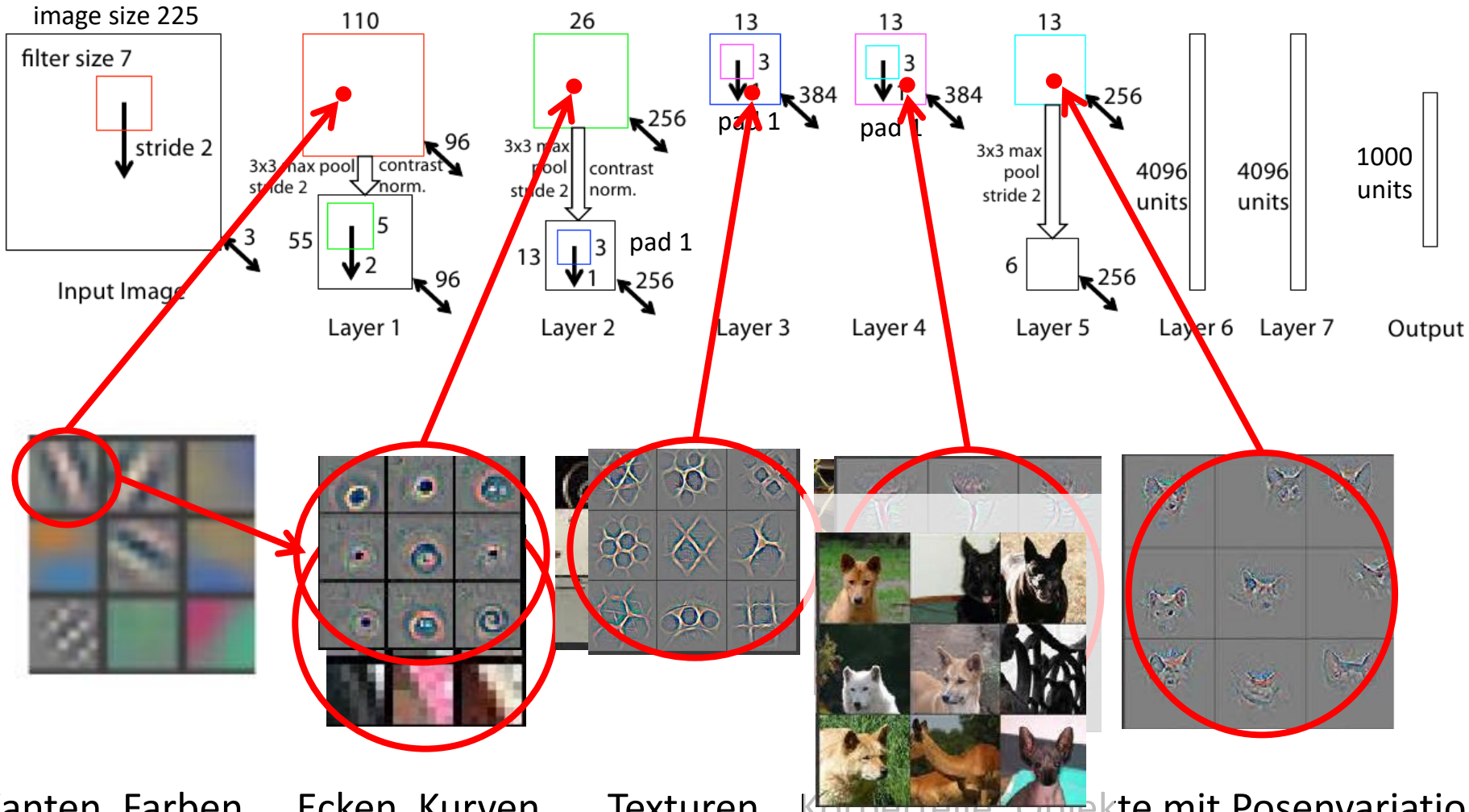
Daten (ImageNet):
13m Paare
(Bild, Bildklasse)
1000 Klassen

1000x1 Vektor mit
Klassenwahrscheinlichkeiten

Zeiler MD, Fergus F. ECCV, *Lecture Notes in Computer Science* 8689, 818-833, 2014

Tiefes Lernen von Bildklassen

ca. 57m Parameter



Zeiler MD, Fergus F. ECCV, *Lecture Notes in Computer Science* 8689, 818-833, 2014

Tiefes Lernen

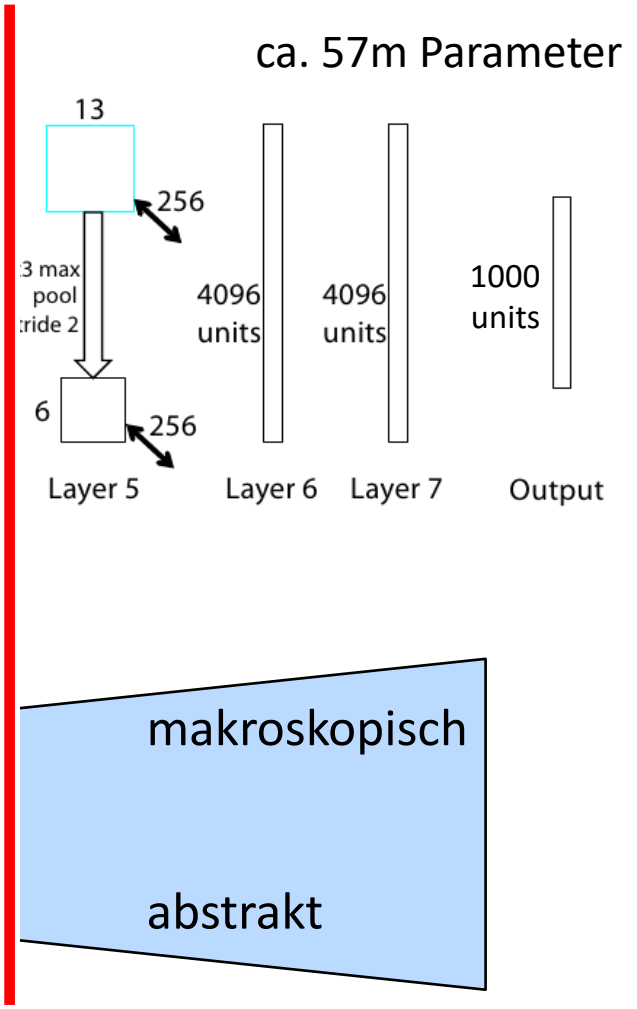
Feature Engineering

Tiefe Lernmodelle erlernen Features:

automatisches „Feature Learning“

statt

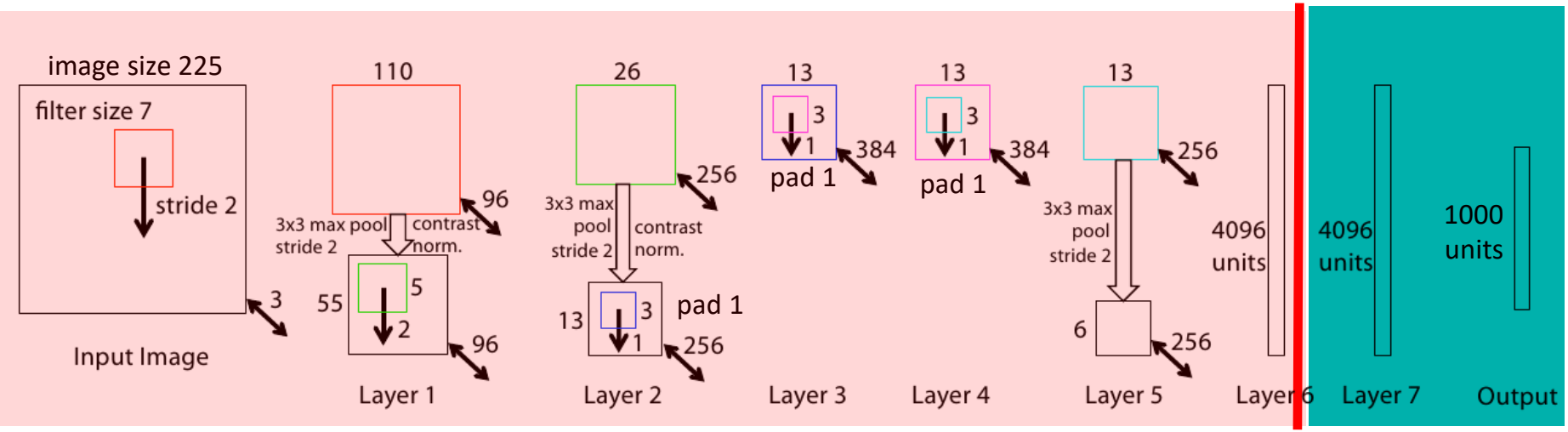
manuelles „Feature Engineering“



Kanten, Farben Ecken, Kurven Texturen Körperteile Objekte mit Posenvariation

Zeiler MD, Fergus F. ECCV, *Lecture Notes in Computer Science* 8689, 818-833, 2014

Tiefes Lernen



statischer Teil des Netzes (wurde auf einem anderen Datensatz trainiert [z.B. ImageNet])

neu-trainierter Teil des Netzwerks (wird auf Datensatz von Interesse trainiert)

"transfer learning"

Kanten, Farben Ecken, Kurven Texturen Körperteile Objekte mit Posenvariation

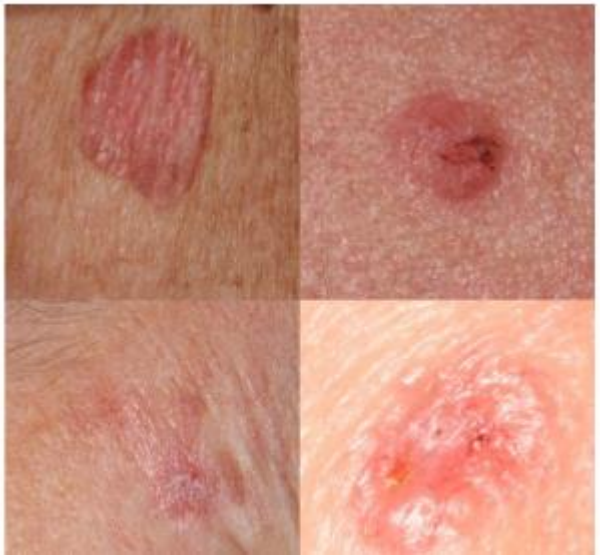
Hautkrebs-Diagnostik

Fotographien epidermaler Läsionen

Trainingsdaten:
127.000 Fotos

gutartige Strukturen

bösartige Strukturen



transfer learning:

Umtrainieren eines auf ImageNet vortrainierten Netzes

Esteva A. et al, *Nature* 542, 115-118, 2017

Hautkrebs-Diagnostik

Dermatoskop

Zusammenfassung

- Data Science & Data Engineering

- Data Science Arbeitsschritte:
  - Datenbeschaffung und –bereinigung
  - Explorative Analyse
  - Feature Engineering
  - daten-getriebene Modellierung

- Tiefes Lernen

2. Organisatorisches

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

Organisatorisches

Lernziele

- Sie können Data Science Projekte selbstständig durchführen.
- Sie können Methoden zur Erschließung und Aufbereitung von Daten sowie zu ihrer explorativen Analyse und Modellierung anwenden.
- Sie können die Ergebnisse Ihrer Untersuchungen diskutieren und beurteilen.
- Bei der Durchführung von Data Science Projekten treffen Sie Entscheidungen, die auf den Ergebnissen ihrer hypothesengetriebenen Untersuchungen basieren.

Organisatorisches

Vorlesung: 2 SWS
Übung: 1 SWS
Praktikum: 1 SWS

} 45 Zeitstunden (= 3h pro Vorlesungswoche)

Vor- und Nachbereitung: 50 Zeitstunden
Hausarbeiten u.a.: 55 Zeitstunden

} ≈ 7h pro Vorlesungswoche

Gesamt: 5 ECTS Punkte; 150 Zeitstunden

Organisatorisches

Zeitstruktur

9:00 – 10:30 Uhr: Vorlesung

10:30 – 12:00 Uhr: Praktikum

Termine:

05.04., 12.04., 19.04., 26.04., 03.05., 17.05., 24.05., 07.06., 14.06., 21.06., 28.06., 05.07., 12.07.*

ILIAS:

Kursname: „Einführung Data Science FZJ (SS24)“ bzw.
„Einführung Data Science AC (SS24)“

*) Bedarfstermin. Im Semester wird bekannt gegeben, ob dieser stattfinden wird.

Organisatorisches

Übungen:

- Dienen zur Vertiefung des Lernstoffes
- Je nach Vorgabe: Zusammenarbeit in Gruppen von 4 Personen

Art der Prüfung

- Klausurart und Prüfungstermin werden noch bekannt gegeben

Kontakt

- Bevorzugt: Ansprechen in der Präsenzzeit (Vorlesung / Übung).
- E-Mail an: bialonski@fh-aachen.de, grieger@fh-aachen.de
- Beginnen Sie den Betreff der E-Mail mit: „[DS FZJ/AC]“.
- Beachten Sie bitte in ILIAS: „E-Learning -> Kommunikation“

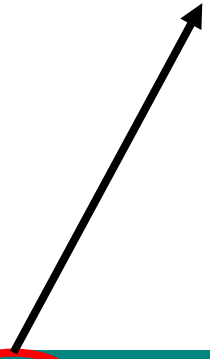
F

Organisatorisches

Fragen?

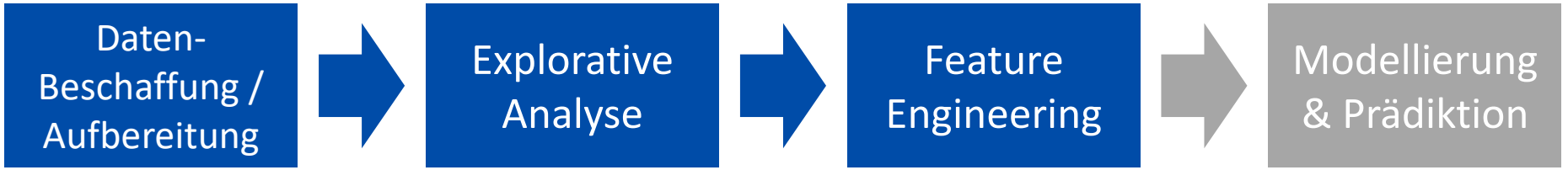
Diese Vorlesung wird weiterentwickelt. Wenn Sie Fehler auf den Folien finden, melden Sie sie gerne:

Beispiel:
E-Mail an: bialonski@fh-aachen.de
Betreff: [DS FZJ/AC] Folie 1-35
Inhalt: Beschreibung des Fehlers



3. Fallbeispiel

Fallbeispiel: Untersuchung epileptischer Anfälle



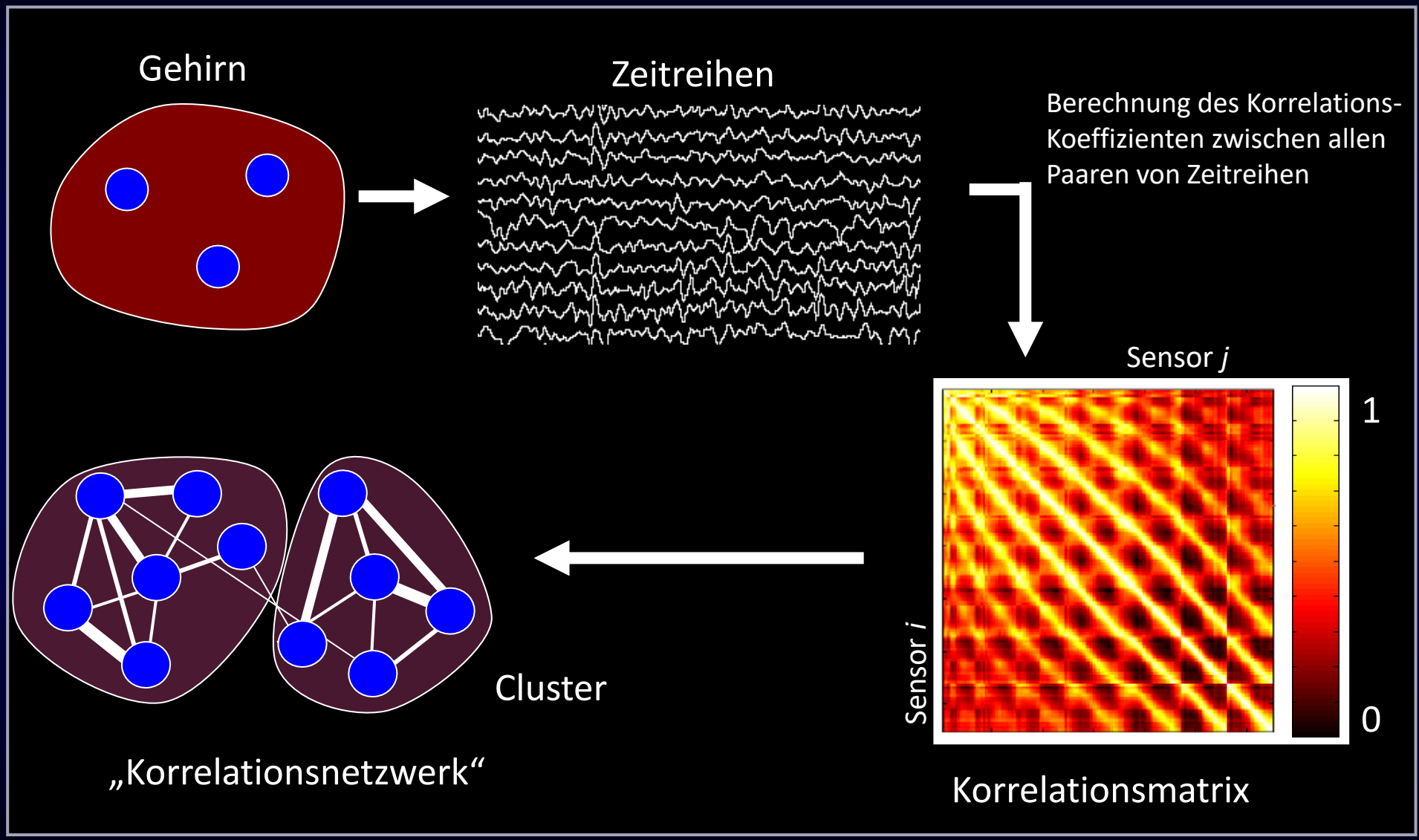
- 1% der Weltbevölkerung leidet unter epileptische Anfälle
- 25% der Epilepsiepatienten: keine effektive Behandlungsmöglichkeit

- **Fernziel**: Vorhersage epileptischer Anfälle (mittels implantierten Elektroden und Datenanalyse)
- Zusammenarbeit mit Klinik für Epileptologie (Bonn)

Datenerfassung: Ableitung von EEG Aktivität

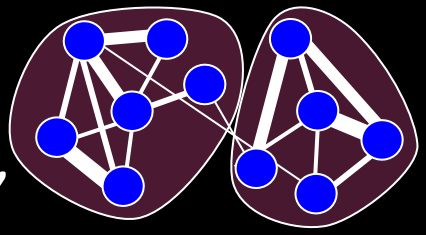
Etwa 10-100 GB / Patient; Aufzeichnungsdauer etwa 2-20 Tage
im Rahmen der prächirurgischen Diagnostik.

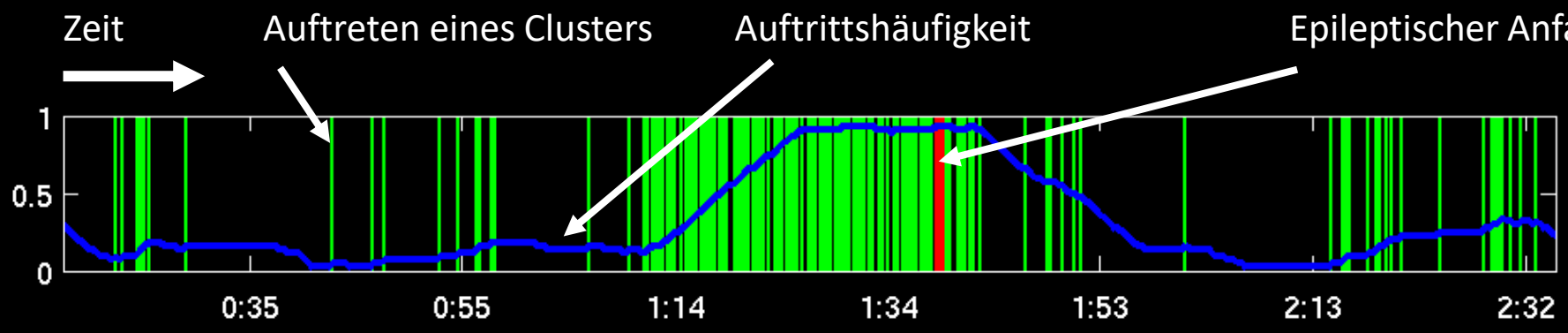
Datenaufbereitung & Analyse-Pipeline



Feature Engineering für die Anfallsvorhersage

Identifikation von Clustern in Korrelationsnetzwerken, die vermehrt vor epileptischen Anfällen auftreten.





Bisherige gefundene Features:

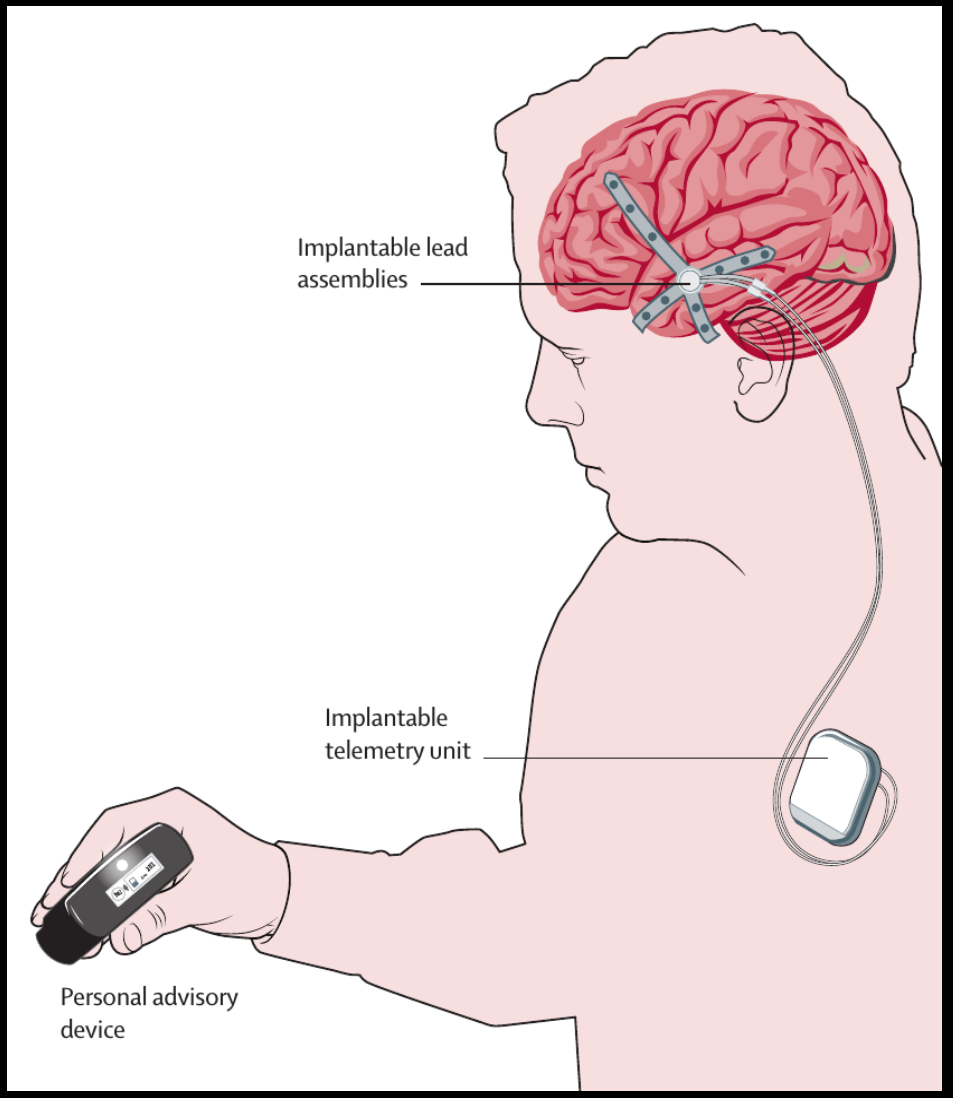
- noch nicht spezifisch genug für Anfallsvorhersage
- Hinweise auf prinzipielle Machbarkeit einer Vorhersage

Vom Feature Engineering zum Feature Learning

Ausblick:

- Längere Aufzeichnungsdauern mithilfe von Implantaten
- Feature Learning

Personalisierte Medizin
Datenwissenschaftler justieren, debuggen und aktualisieren Prädiktionsalgorithmen (*)



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