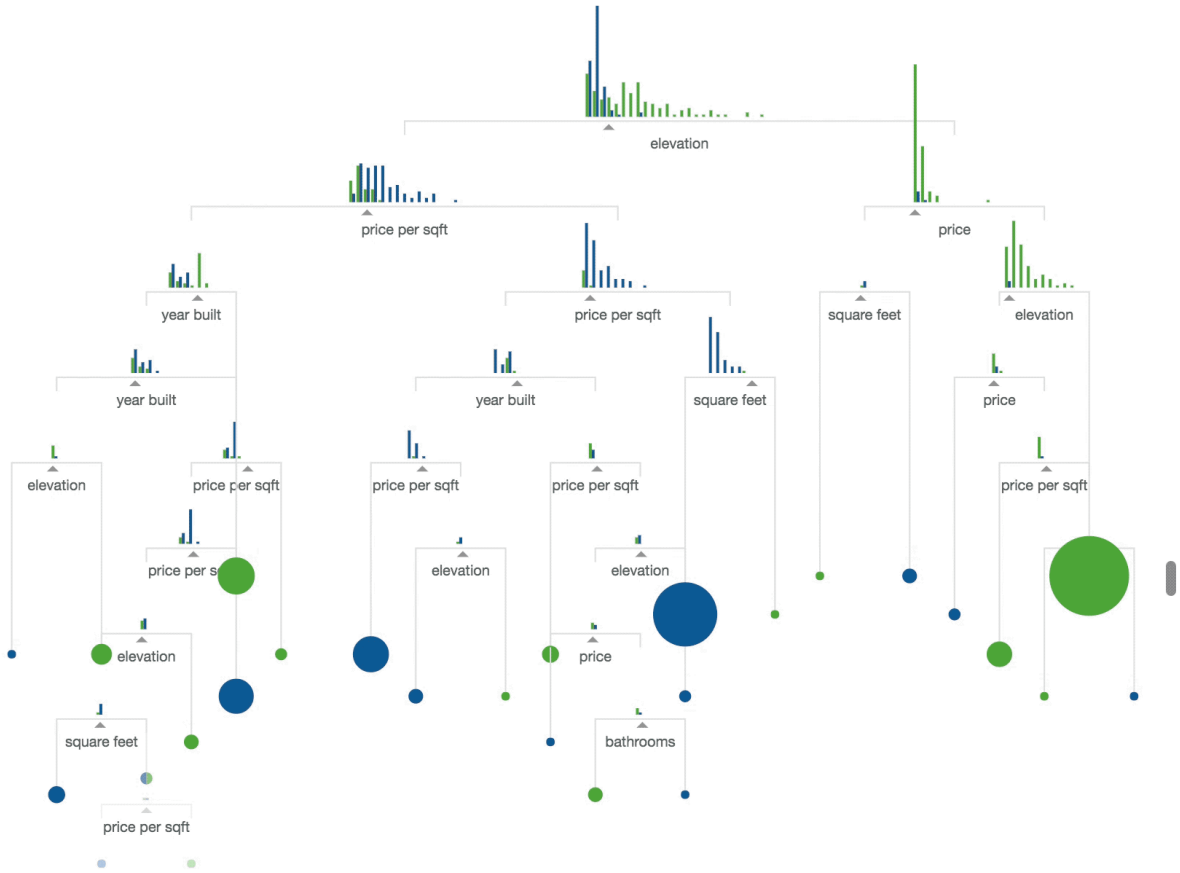
Was ist Machine Learning?



Studiengang: Systemtechnik[NTB]

Modul: Machine Learning

Teil 1 – Einführung in das Maschinelle Lernen

Dozent: Prof. Dr. Klaus Frick | ICE
Prof. Dr. Christoph Würsch | ICE

NTB
Interstaatliche Hochschule für Technik Buchs
FHO Fachhochschule Ostschweiz

Inhalt

- Ein erstes Beispiel: Spam Detekoren
- Grundlagen der Datendarstellung
- Definition von Machine Learning
- Arten des Maschinellen Lernens (Machine Learning)
  - Überwachtes Lernen (supervised learning)
  - Unüberwachtes Lernen (unsupervised learning)
  - Reinforcement Learning
- Der Machine Learning Prozess
- Umgang mit Daten mit dem Pandas-Paket in Python (Notebook)

Inhalt

- Ein erstes Beispiel: Spam Detekoren
- Grundlagen der Datendarstellung
- Definition von Machine Learning
- Arten des Maschinellen Lernens (Machine Learning)
  - Überwachtes Lernen (supervised learning)
  - Unüberwachtes Lernen (unsupervised learning)
  - Reinforcement Learning
- Der Machine Learning Prozess
- Umgang mit Daten mit dem Pandas-Paket in Python (Notebook)

Spam Detektor

- Spam Detektoren gelten als erste verbreitete Anwendung von künstlicher Intelligenz (*artificial intelligence* AI).
- Der Task besteht darin, für jede erhaltene E-Mail Nachricht bzw. für jeden neuen Kommentar in einem Online-Forum, auf Facebook, YouTube & Co folgende Entscheidung zu treffen:

Handelt es sich um eine unerwünschte, boshafte oder betrügerische Nachricht (spam) oder um eine statthafte, dem Kontext entsprechende Nachricht (ham)?

- Die Entscheidung soll nicht durch einen Menschen getroffen werden, sondern automatisch durch einen Algorithmus

https://www.youtube.com/watch?v=_rdINNHLYaQ&feature=youtu.be

YouTube Kommentare zu «Katy Perry»

- Beispiel-Kommentare «ham»:

```
1  OMG I LOVE YOU KATY PARRY YOUR SONGS
   ROCK!!!!!!!!!!!!!!!! THATS A TOTAL SUBSCRIBE
2  Katy Perry - Roar (Official):
   http://youtu.be/CevxZvSJLk8
```

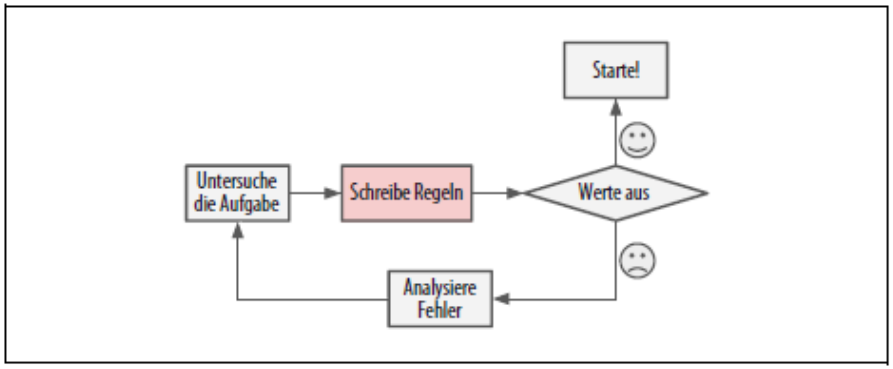
- Beispiel-Kommentare «spam»:

```
1  https://www.paidverts.com/ref/tomuciux99 esyest money
   ever. join to our team!!!!
2  Nice song .See my new track.
```

- Frage: Wie kann automatisch entschieden werden, ob ein Kommentar «ham» oder «spam» ist?

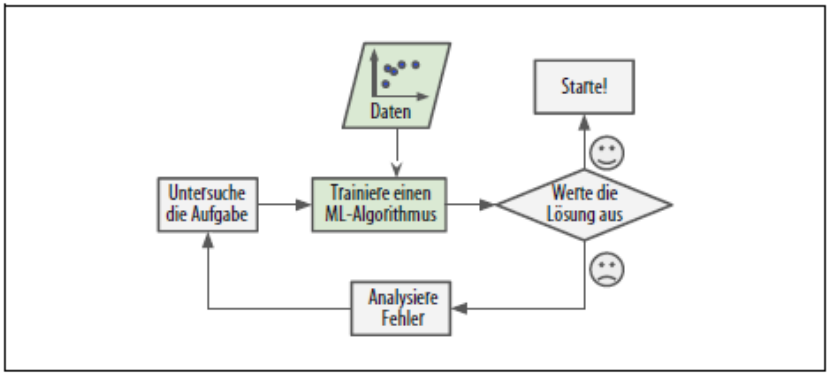
Spam Detektor 1: Expertensystem

1. Analysiere «von Hand» viele Beispiele von Spam-Kommentaren und erstelle eine Liste von Mustern, die typisch für «spam» erscheinen (ALL-CAPS, !!!!!!!, Web-Links, Schreibfehler, «money», «business», …)
2. Schreibe einen Algorithmus, der die Patterns aus der Liste in 1. erkennt und anhand fixierter Regeln entscheidet, ob es sich um «ham» oder «spam» handelt.
3. Teste das Programm und aktualisiere die Regelliste.



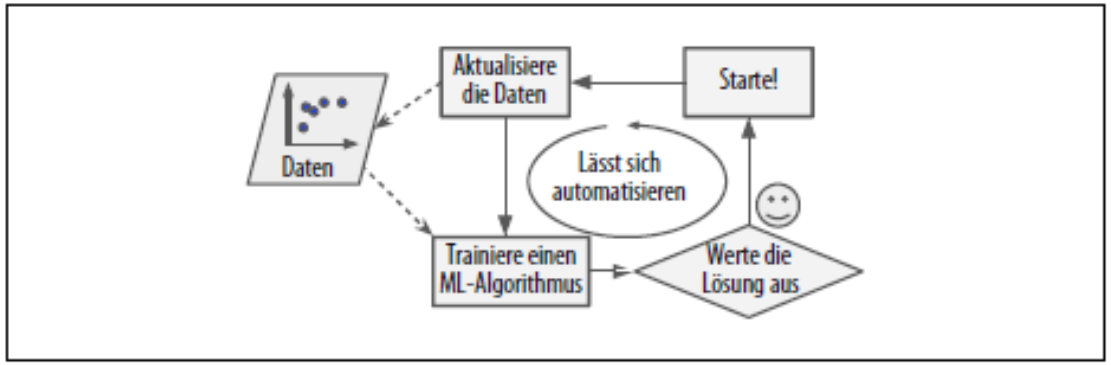
Spam Detektor 2: Machine Learning

1. Sammle viele Beispiele von Kommentaren aus der «ham» und «spam» Klasse (**Trainingsdaten**)
2. Definiere einen grossen Satz von generischen Merkmalen $x_1, \ldots, x_p$ (**features**), die aus einem gegeben Kommentar-Text abgeleitet bzw. berechnet werden können (word counts diverser Schlüsselwörter, Uhrzeit, Poster-ID, …)
3. Ein ML-Algorithmus findet in den Features der Trainingsdaten selbstständig Muster, die auf «ham» bzw «spam» hinweisen.



Spam Detektor: Unterschiede der Ansätze

- Expertensysteme sind starr und haben einen lange Liste von Regeln, die schwer zu warten sind. Machine Learning Modelle können während dem Betrieb aktualisiert werden (s. Grafik unten).
- Bei komplizierten Zusammenhängen sind Expertensysteme nicht mehr realisierbar (Beispiel: Spracherkennung)
- Gelernte Machine Learning Modelle können dem Menschen Zusammenhänge aufzeigen, die vorher nicht offensichtlich waren.



Inhalt

- Ein erstes Beispiel: Spam Detekoren
- Grundlagen der Datendarstellung
- Definition von Machine Learning
- Arten des Maschinellen Lernens (Machine Learning)
  - Überwachtes Lernen (supervised learning)
  - Unüberwachtes Lernen (unsupervised learning)
  - Reinforcement Learning
- Der Machine Learning Prozess
- Umgang mit Daten mit dem Pandas-Paket in Python (Notebook)

Grundgesamtheit

- Wir betrachten die klassische Definition von Daten aus der Statistik.
- Wesentlich ist der Begriff der **Grundgesamtheit,** die alle **statistischen Elemente (Merkmalsträger)** beinhaltet, für die wir uns in einer speziellen Anwendung interessieren.
- Es ist wichtig, die Grundgesamtheit in zeitlicher, inhaltlicher und örtlicher Sicht genau zu definieren.

Beispiel: Spam

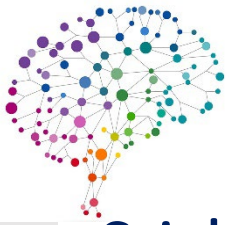
Die Elemente sind die einzelnen Kommentare, die von Menschen bzw. Bots abgegeben werden. Als Grundgesamtheit betrachten wir alle Kommentare auf der YouTube Plattform, die zu offiziellen Musikvideos im Jahr 2018 abgegeben wurden.

Merkmal (feature)

- Der Begriff des Merkmals ist zentral im Maschinellen Lernen! In der Literatur wird oft der englische Begriff **feature** verwendet.
- Ein **Merkmal ist eine messbare Eigenschaft** der statistischen Einheiten in der Grundgesamtheit.
- Die Werte, die ein Merkmal annehmen kann, nennt man **Merkmalsausprägung.** Ein sinnvolles Merkmal hat mindestens zwei verschiedene Ausprägungen.

Beispiel: Spam

Mögliche Merkmale eines YouTube Kommentars: Anzahl der ASCII Zeichen (inkl. Leerschlag), die Uhrzeit des Kommentars, die Anzahl der Rechtschreibfehler, Indikator (0/1) ob der Name des Interpreten im Kommentar enthalten ist, …



Stichprobe (sample)

- Die Grundgesamtheit aller Elemente ist in den seltensten Fällen als Ganzes zugänglich. Es wird lediglich eine Teilmenge der Grundgesamtheit erfasst: **die Stichprobe.**
- Soll basierend auf der Stichprobe eine Aussage über nicht-erfasste bzw. zukünftige Elemente gemacht werden, muss diese **repräsentativ** bzw. **unverzerrt** sein.
  - Der **Stichprobenumfang** muss der gewünschten Genauigkeit des Machine Learning Algorithmus entsprechen.
  - Bei Klassifikationsproblemen (z.B. Spam Filter): ***stratified sampling***

Beispiel: Spam

5000 Kommentare von insgesamt 25 unterschiedlichen Interpreten, wobei «spam» und «ham» Kommentare in der Stichprobe im gleichen Verhältnis wie in der Grundgesamtheit auftauchen (stratified sampling).

Datentypen

- Die Merkmalsausprägungen in einer Stichprobe sind die Daten, mit denen es der Data Scientist zu tun hat.
- **Qualitative Merkmale** lassen sich nicht in natürlicher Weise durch Zahlen sondern nur verbal beschreiben.
  - **Ordinale Merkmale** lassen sich logisch miteinander vergleichen (Schulnoten, Dienstgrade, Zeitangaben, …)
  - **Nominale Merkmale** sind nur durch Namen unterscheidbar (Herstellername, Boole'sche Variablen, Farben, Geschlecht, …)
- **Quantitative Merkmale** lassen sich in natürlicher Weise durch Zahlen beschreiben («klassische Messwerte»).
  - **Diskrete Merkmale** können nur endlich (abzählbar unendlich) viele Werte Annehmen (Seitenzahlen, Anzahl der Zeichen in einem Kommentar, …)
  - **Kontinuierliche (stetige) Merkmale** können beliebe Werte in einem Intervall annehmen (Druck, Temperatur, …)

Merkmalstypen

- Preis eines Produktes in CHF.
  - (A) Quantitativ (diskret), (B) Quantitativ (stetig), (C) Ordinal, (D) Nominal
- Herstellername.
  - (A) Quantitativ (diskret), (B) Quantitativ (stetig), (C) Ordinal, (D) Nominal
- Star Rating auf Amazon.
  - (A) Quantitativ (diskret), (B) Quantitativ (stetig), (C) Ordinal, (D) Nominal
- Datum, an dem ein Produkt verkauft wurde
  - (A) Quantitativ (diskret), (B) Quantitativ (stetig), (C) Ordinal, (D) Nominal
- Druckanstieg nach der ersten Kompressorstufe
  - (A) Quantitativ (diskret), (B) Quantitativ (stetig), (C) Ordinal, (D) Nominal
- Produzierte Stückzahl pro Minute auf einer Anlage
  - (A) Quantitativ (diskret), (B) Quantitativ (stetig), (C) Ordinal, (D) Nominal

Datenform: *tidy data*

- Angenommen, eine Stichprobe vom Umfang n liege vor. Für die Elemente in der Stichprobe seien die Ausprägungen von p Merkmalen $X_1, \ldots, X_p$ erfasst.
- Die Daten werden stets als **Datenmatrix** (Codd's dritte Normalform) dargestellt

	X_1	X_2	$\ldots$	X_j	$\ldots$	X_p
Element 1	x_{11}	x_{12}	$\ldots$	x_{1j}	$\ldots$	x_{1p}
Element 2	x_{21}	x_{22}	$\ldots$	x_{2j}	$\ldots$	x_{2p}
$\vdots$						
Element i	x_{i1}	x_{i2}	$\ldots$	x_{ij}	$\ldots$	x_{ip}
$\vdots$						
Element n	x_{n1}	x_{n2}	$\ldots$	x_{nj}	$\ldots$	x_{np}

Indizierung von Daten

- Die n Elemente einer Stichprobe sind **eindeutig indiziert**.
- Beispiele:
  - Kanonische Indizierung $1, \ldots, n$
  - Eindeutige Namen (Personen, Staaten, …)
  - Ids
  - Timestamps
- Hat der Index eine **chronologische** Interpretation (z.B. timestamp, counter, Jahreszahl, …), so nennt man die Datenmatrix eine **Zeitreihe.**
- Zeitreihen benötigen in der Regel spezielle Machine Learning Verfahren (ARIMA Modelle, Rekurrente Neuronale Netze, …).
- Für die Modelle in diesem Kurs spielt die **Reihenfolge** der beobachteten Elemente **keine Rolle**

Inhalt

- Ein erstes Beispiel: Spam Detekoren
- Grundlagen der Datendarstellung
- Definition von Machine Learning
- Arten des Maschinellen Lernens (Machine Learning)
  - Überwachtes Lernen (supervised learning)
  - Unüberwachtes Lernen (unsupervised learning)
  - Reinforcement Learning
- Der Machine Learning Prozess
- Umgang mit Daten mit dem Pandas-Paket in Python (Notebook)

Klassische Definition für Maschinelles Lernen

*«[Machine Learning is the] field of study that gives computers the ability to **learn without being explicitely programmed**.»*

(Arthur Samuel)

*«Learning is any process by which a system improves performance from experience. Machine Learning is concerned with computer programs that **automatically improve their performance through experience**»*

(Herbert Alexander Simon)

«A computer program ist said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.»

(Tom Mitchell)

Machine Learning ist ein Teilgebiet von AI

ARTIFICIAL INTELLIGENCE (AI)

Jede Technik (in Form von Software), die seine Umgebung wahrnimmt und anhand dieser Inputs menschliche Intelligenz imitiert, um die Wahrscheinlichkeit ein gewisses Ziel zu erreichen zu maximieren.

Teilgebiete: Computer Vision, Robotik, Natural Language Processing, Route Planing, Machine Learning, …

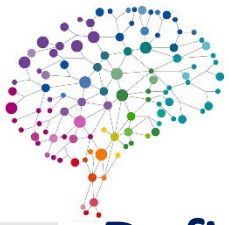
MACHINE LEARNING

Eine spezielle Sammlung an Algorithmen, realisiert in Software, die basierend auf mathematischen und/oder statistischen Verfahren es Computern erlaubt, die Wahrscheinlichkeit das Ziel zu erreichen anhand von Erfahrungen (in Form von Daten) zu erhöhen.

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Inhalt

- Ein erstes Beispiel: Spam Detekoren
- Grundlagen der Datendarstellung
- Definition von Machine Learning
- Arten des Maschinellen Lernens (Machine Learning)
  - Überwachtes Lernen (supervised learning)
  - Unüberwachtes Lernen (unsupervised learning)
  - Reinforcement Learning
- Der Machine Learning Prozess
- Umgang mit Daten mit dem Pandas-Paket in Python (Notebook)



Definition des überwachten Lernens

- Derzeit die am weitesten verbreitete Art von Machine Learning.
- Supervised Learning hat eine **klare Definition** der Aufgabe (*task*), die von einem Algorithmus erfüllt werden soll.
- Annahme: Es gibt eine wohldefinierte **Response Variable** Y und ***p* Merkmale** (Prädiktoren, features) $X_1, \ldots, X_p$, sodass

$$Y = f(X_1, \ldots, X_p) + \varepsilon$$

- Hierbei ist f eine unbekannte mehrdimensionale Funktion der Prädiktoren und ε ein Fehlerterm, der unabhängig von den Prädiktoren ist (Messfehler, …).
- Supervised Learning ist eine Sammlung von Algorithmen, um die **unbekannte Funktion f zu schätzen**. Die Schätzung erfolgt basierend auf Beispieldaten (Prädiktoren + Response).

Trainingsdaten

- Zentrale Element im (überwachten) maschinellen Lernen.
- Der Trainingsdatensatz besteht aus vielen Beobachtungen von Prädiktor-Werten **inklusive** der jeweiligen Response-Werte.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

- Finde nun eine Funktion $\hat{f}$, welche die Response-Werte in den Trainingsdaten **möglichst gut** beschreibt:

$$\hat{f}(x_{i1}, \dots, x_{ip}) = \hat{y}_i \approx y_i$$

- Die Funktion $\hat{f}$ wird meist aus eine Klasse von Ansatzfunktionen mittels Parameter ausgewählt (s. später).

Ziele des überwachten Lernens

- Angenommen $\hat{f}$ ist ein Schätzer für die unbekannte Funktion. Damit können zwei (konkurrierende) Ziele verfolgt werden:
- **Vorhersage**. Für einen neuen Satz von Prädiktoren kann die Response vorhergesagt werden

$$\hat{Y} = \hat{f}(X_1, \ldots, X_p)$$

  In diesem Zusammenhang wird $\hat{f}$ oft als black-box betrachtet und es zählt nur die Vorhersagekraft (**predictive power**).
- **Interpretation.** Hier verwenden wir $\hat{f}$ dazu, den Zusammenhang von den Prädiktoren und der Response zu verstehen. Speziell:
  - Welche Prädiktoren sind wichtig?
  - Wie hängt Y von den einzelnen Prädiktoren ab?
  - Ist der Zusammenhang linear oder nichtlinear?

Beispiele: Spam Detektor

- Für einen gegebenen Kommentar berechnen wir Prädiktoren z.B. gemäss dem *bag-of-words* Modell.

```
The song is very good ...but the video makes no sense...just a nonsense video...I mean she is telling her story of
being stuck on an island, but the song doesn't fit in the situation...but nvm...The song is good
```

↓

```
{"song": 3, "video": 2, "Katy": 0, "Perry": 0, "…", 5, u.s.w. ,  "Money": 0}
```

↓

$$X_1, \ldots, X_p = (3, 2, 0, 0, 5, \ldots, 0)$$

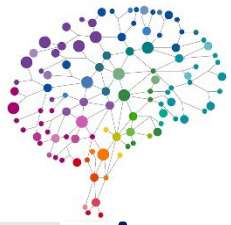
- Wir verwenden für die Response Variable ein *encoding*

$$Y = \begin{cases} 1 & \text{falls "spam"} \\ 0 & \text{falls "ham"} \end{cases}$$

- **Vorhersage**:

$$\hat{f}(3, 2, 0, 0, 5, ..., 0) = 0 \Rightarrow \text{ ham}$$

- **Interpretation.** Welche Wörter bzw. *Wortkombinationen* sind wichtig, um «ham» und «spam» zu unterscheiden?



Arten des überwachten Lernens

- Man unterscheidet zwei Arten des überwachten Lernens, je nach der Eigenschaft der Response-Variable:

Regression

Die Response Variable Y ist **numerisch** (quantitativ).

Beispiel: Vorhersage des Preises von Gebrauchtwagen basierend auf Kilometerstand, Alter, Marke, etc.

Klassifikation

Die Response Variable Y ist **nominal/ordinal** (qualitativ).

Beispiel: Vorhersage ob ein YouTube Kommentar «ham» oder «spam» ist basierend auf Worthäufigkeiten.

- Klassifikationsprobleme (Entscheidungsprobleme) kommen in klassischen Anwendungsgebieten öfter vor (Objekterkennung). Hier bezeichnet man die Response Variable Y oft als **Label.**

Erfolgsstories beim überwachten Lernen

- Spracherkennung und –synthese (Natural Language Processing): Sprecheridentifikation, Chatbots, …

  Google Duplex: https://www.youtube.com/watch?v=D5VN56jQMWM

- Textanalysis: Automatische Verarbeitung von Support-Tickets, automatische Übersetzung, …

  DeepL: https://www.deepl.com

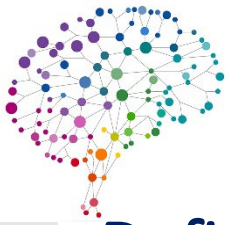
- Bildverarbeitung (computer vision): Objekterkennung, autonomes Fahren, Gesichtserkennung, Authentifizierung, …

  https://www.youtube.com/watch?v=PgnsapPGaaw

  https://www.youtube.com/watch?v=lL16AQItG1g&feature=youtu.be

- Automatische Diagnostik: Medizinische und industrielle Anwendungen

  https://emerj.com/ai-sector-overviews



Definition des unüberwachten Lernens

- Derzeit ein sehr aktives Forschungsgebiet mit vielen offenen Fragen.
- Die **Aufgabenstellung** beim unüberwachten Lernen ist **nicht klar definiert** wie beim überwachten Lernen, da keine Response Variable existiert.
- Die Trainingsdaten bestehen also nur aus den Merkmalen

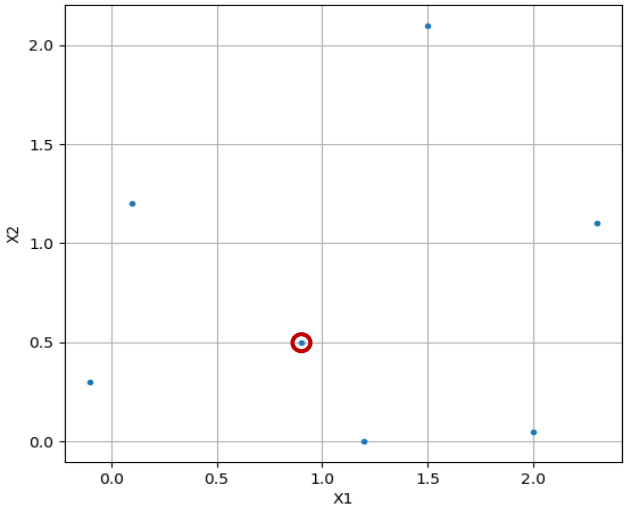
$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

- Beim unüberwachten Lernen werden also die **Merkmale modelliert**, ohne dass z.B. irgendwelche Klassenbezeichnungen (Labels) vorgegeben werden.

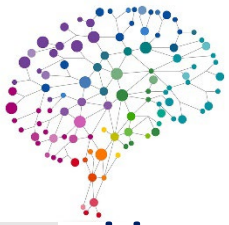
Grundüberlegung des unüberwachten Lernens

- Jeder Zeile in der Trainingsdatenmatrix kann als **Punkt im p-dimensionalen Raum** interpretiert werden. Die gesamte Matrix entspricht also einer **Punktewolke**.

$$\mathbf{X} = \begin{pmatrix} 1.2 & 0 \\ 2.3 & 1.1 \\ 0.9 & 0.5 \\ -0.1 & 0.3 \\ 2.0 & 0.05 \\ 0.1 & 1.2 \\ 1.5 & 2.1 \end{pmatrix}$$



- Für **grosse p** (viele Merkmale) und/oder n (viele Fälle) ist die geometrische Interpretation schwierig.
- Gapminder: www.gapminder.org



Hauptrichtungen des unüberwachten Lernens

- Wichtig ist beim unüberwachten Lernen der Begriff des **Abstands** bzw. die **gegenseitige Lage** von Datenpunkten.

Clustering

Bilden die Datenpunkte im p-dimensionalen Merkmalsraum Gruppen?

Beispiel: Bestimmung von «typischen» Kundengruppen im (Online) Handel.

Dimensionsreduktion

Bilden die Daten eine Punktewolke mit einer Dimension kleiner als p (z.B. eine Gerade im Raum)?

Beispiel: Reduktion der Pixeldaten handgeschriebener Ziffern auf 2 relevante Merkmale.

Visualisierung

Wie können hoch-dimensionale Daten unverzerrt und interpretierbar visualisiert werden?

Erfolgsstories beim unüberwachten Lernen

- Anomalitätsdetektion (anomality detection): Oft die erste Ansatz, industrielle Daten von Produktionsmaschinen mittels Machine Learning zu nützen, da keine Labels notwendig sind.

  https://www.youtube.com/watch?v=u7jaGXYKjVM

- Betrugserkennung (fraud detection): Anwendung von Anomalitätsdetektion auf Kreditkartenbehebungen bzw. Energiedaten (energy theft detection)

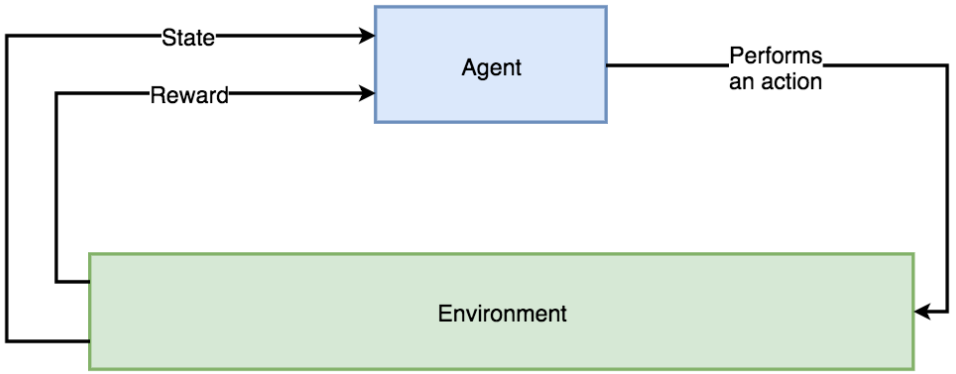
- (Deep) Autoencoder (Verallgemeinerung der PCA): Neben Anomalitätsdetektion auch Anwendungen im Bereich *text-to-speech* etc.

  https://www.youtube.com/watch?v=PgnsapPGaaw

- Kundensegmentierung: Identifiziere typische Kundengruppen anhand der Daten am *point-of-sale.*

Ausblick: Bestärkendes Lernen (nicht Teil des Kurses)

- Aktives Feld im Bereich Robotik, Chatbots, autonomes Fahren,…
- Ein **Agent** (ML Algorithms) versucht eine Aufgabe durch Ausführung von verschiedenen Aktionen (**actions**) zu erreichen.
- Er nimmt dabei seine Umwelt (environment) mittels Sensoren (Kameras, …) wahr und befindet sich nach jeder Aktion in einem neuen Zustant (**state**).
- Jede Aktion wird durch die Umgebung in Form eines **Rewards** bewertet. Ziel des Agenten ist es, den **kummulierten Reward** in der Zukunft zu maximieren.



Inhalt

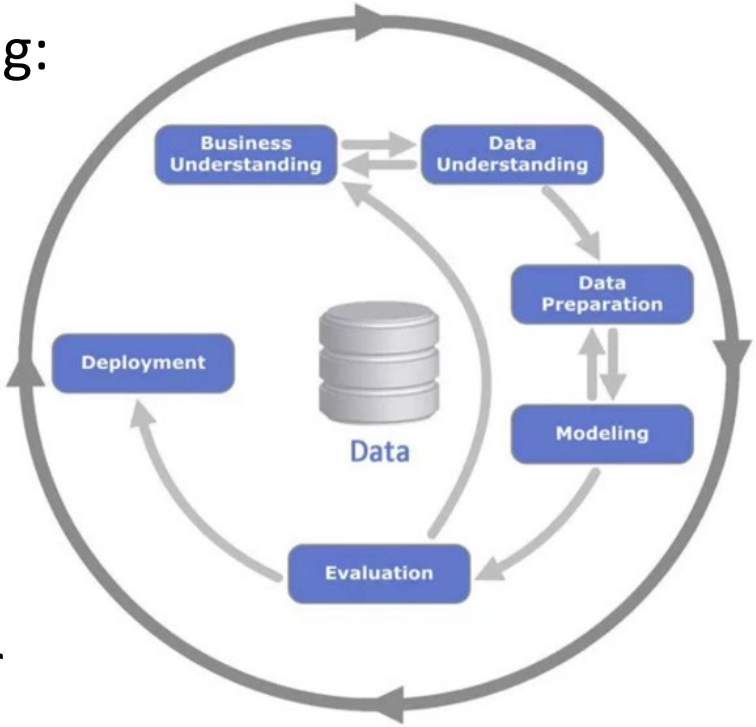
- Ein erstes Beispiel: Spam Detekoren
- Grundlagen der Datendarstellung
- Definition von Machine Learning
- Arten des Maschinellen Lernens (Machine Learning)
  - Überwachtes Lernen (supervised learning)
  - Unüberwachtes Lernen (unsupervised learning)
  - Semi-supervised Learning
  - Reinforcement Learning
- Der Machine Learning Prozess
- Umgang mit Daten mit dem Pandas-Paket in Python (Notebook)

Herzliche Gratulation!

- Sie haben eine Menge Daten gesammelt oder erhalten?
- Was tun sie als nächstes? Beginnen Sie direkt mit der Modellierung? Werfen Sie die Daten in Ihre «Deep-Learning Black Box»?
- **Nein**: Machine Learning Projekte sollten einer **Methodologie** folgen.
- Im Projekt werden dann einzelne Schritte (**Workpackages**) abgearbeitet und nach Abschluss jeweils **dokumentiert** und **evaluiert.**

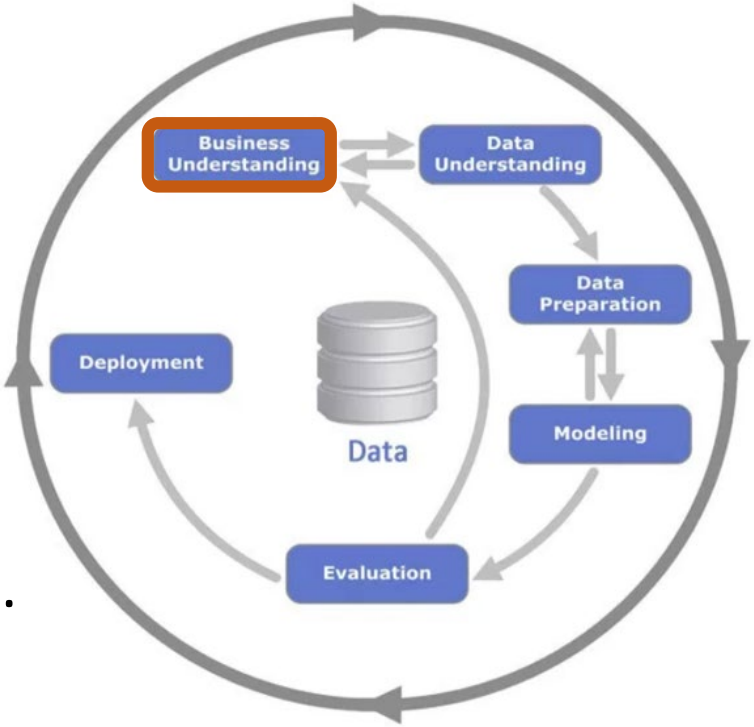
Workflow eines Machine Learning Prozesses

- Es gibt zahlreiche **Prozessmodelle** for Data Mining bzw. Machine Learning.
- Seit dem Jahr 1996 (!) in Verwendung: **Cr**oss **I**ndustrie **S**tandard **P**rocess for **D**ata **M**ining (**CRISP-DM**)
- Seither zahlreiche Überarbeitungen durch IBM, Microsoft, etc.
- CRISP-DM ist in der Praxis eine **praktische Handreichung** zur Durchführung eines ML-Projektes.
- **Wichtig:** Der Prozess beginnt bei der **Geschäftsidee**, beim praktischen Use-Case.



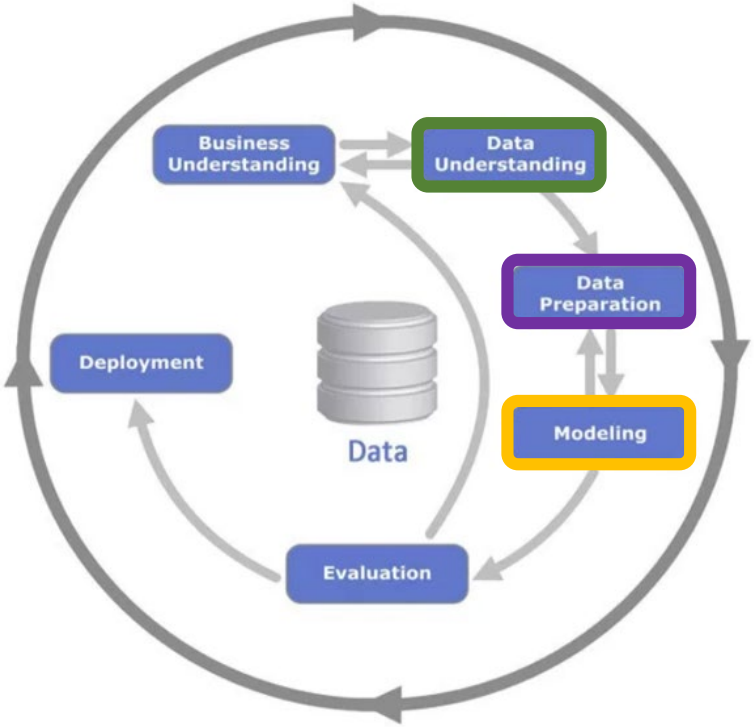
Business Understanding (keine Details in diesem Kurs)

- **Geschäftsmodell** bzw. -ziele definieren. Ist ein datenbasierter Ansatz überhaupt sinnvoll?
- **Ressourcen** analysieren:
  - Personal
  - Daten
  - Hardware
  - Software
- **Quantifizierbare** Machine Learning **Ziele** und **Erfolgskriterien** definieren.
- **Projektplan** mit **Meilensteinen** erstellen.

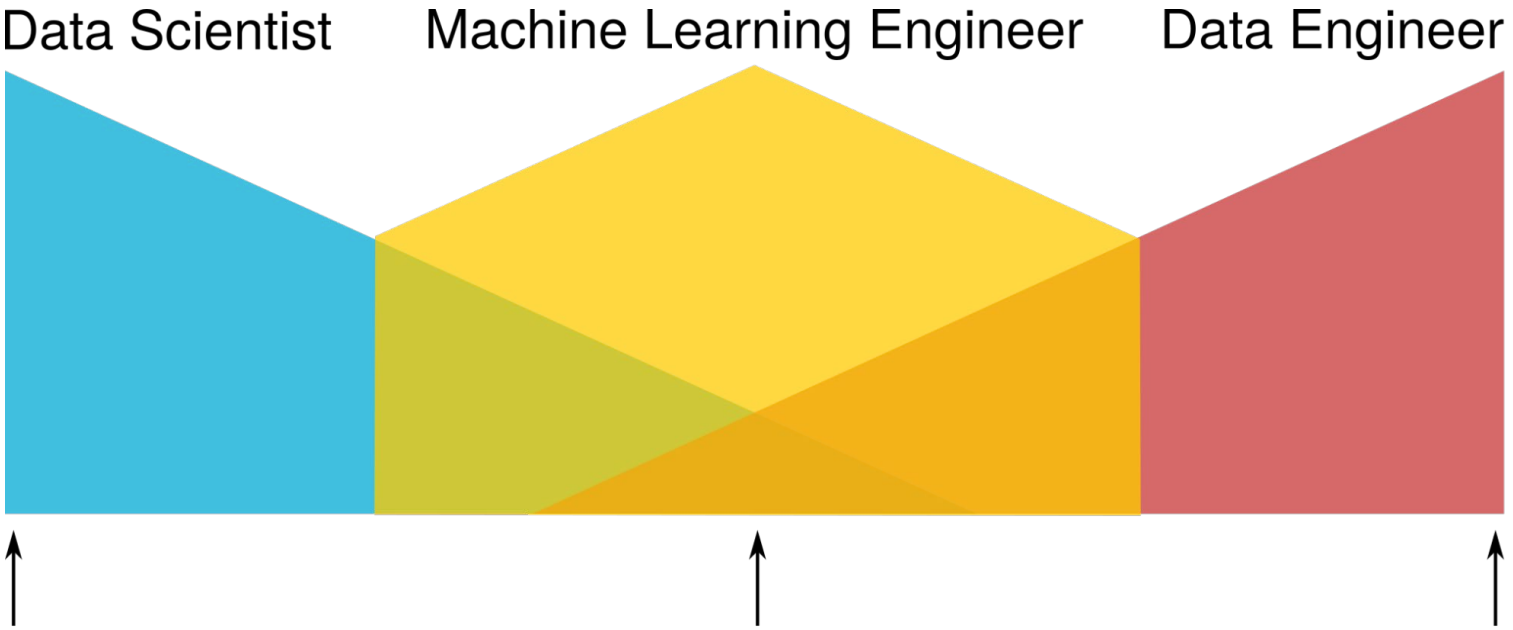


Inhalte dieses Kurses

- **Data Understanding (Kap. 2)**
  - Explorative Analyse der Daten
  - Daten kennenlernen
  - Sind die Daten zur Erreichung der Businesszielen geeignet?
- **Data Preprocessing (Kap. 3, 11, 12)**
  - Transformation, Bereinigung der Daten
  - Erzeugung und Auswahl der Prädiktoren
- **Modelling (Kap. 4 – 9)**
  - Training und Validierung von Machine Learning Modellen



Machine Learning in der Praxis



Inhalt

- Ein erstes Beispiel: Spam Detekoren
- Grundlagen der Datendarstellung
- Definition von Machine Learning
- Arten des Maschinellen Lernens (Machine Learning)
  - Überwachtes Lernen (supervised learning)
  - Unüberwachtes Lernen (unsupervised learning)
  - Reinforcement Learning
- Der Machine Learning Prozess
- Umgang mit Daten mit dem Pandas-Paket in Python (Notebook)

Einführung ins Machinelle Lernen (ML)

- Machine Learning ist eine Möglichkeit, **künstliche Intelligenz** zu realisieren.
- Grundidee ist die Modellierung ausgehend von **Daten**
- Die eigentliche Modellierung ist nur ein Teil des **Machine Learning Prozesse**s, neben explorativer Datenanalyse, Visualisierung und Datenvorverarbeitung, etc.
- Man unterscheidet verschiedene Arten von Maschine Learning. Die wichtigsten sind **überwachtes** und **unüberwachtes Lernen** sowie **Reinforcement Learning.**