Karlsruher Institut für Technologie

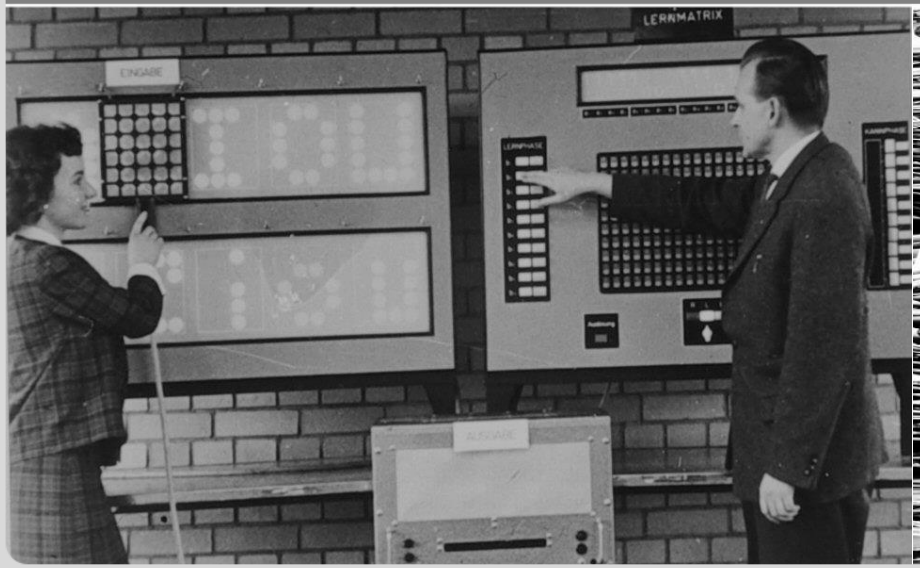
Institut für Technik der Informationsverarbeitung

**Institutsleitung**
Prof. Dr.-Ing. J. Becker
Prof. Dr.-Ing. E. Sax
Prof. Dr. rer. nat. W. Stork

# Übung 3

## Übung zu Informationstechnik II und Automatisierungstechnik – Felix Pistorius

Prof. Dr.-Ing. Eric Sax



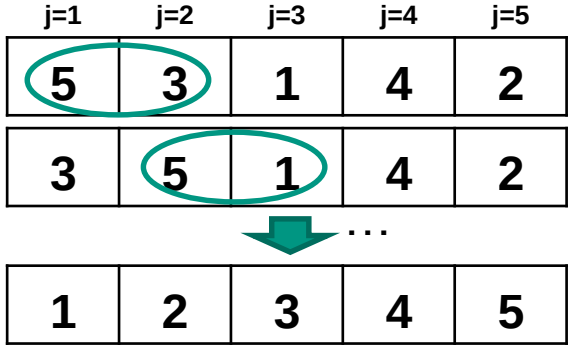


# WIEDERHOLUNG ÜBUNG 2

# Wiederholung Übung 2

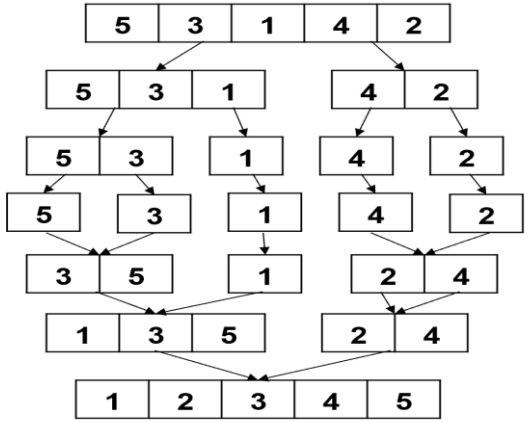
## Sortieralgorithmen

### Bubble Sort





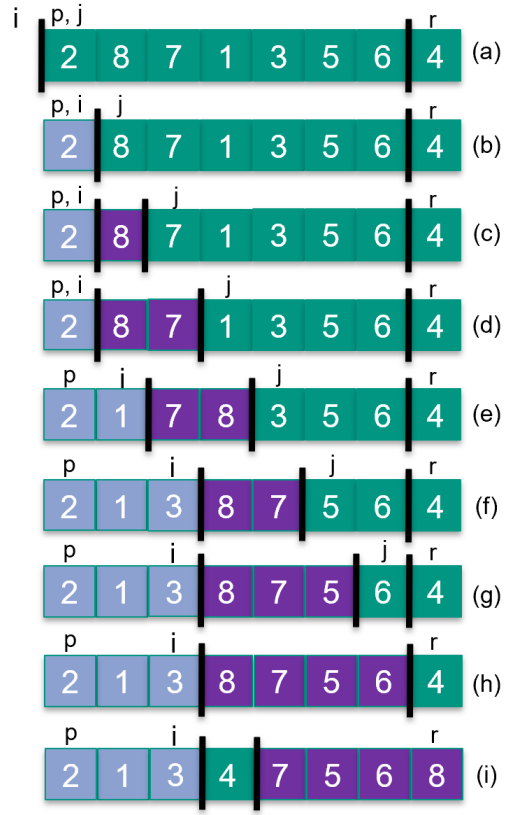
### Merge Sort




### Insertion Sort

```
InsertionSort
for ( j = 2 to length(A) ) do
    key = A[j]
    i = j - 1
    while ( i > 0 and A[i] > key ) do
        A[i+1] = A[i]
        i = i - 1
    A[i+1] = key
```

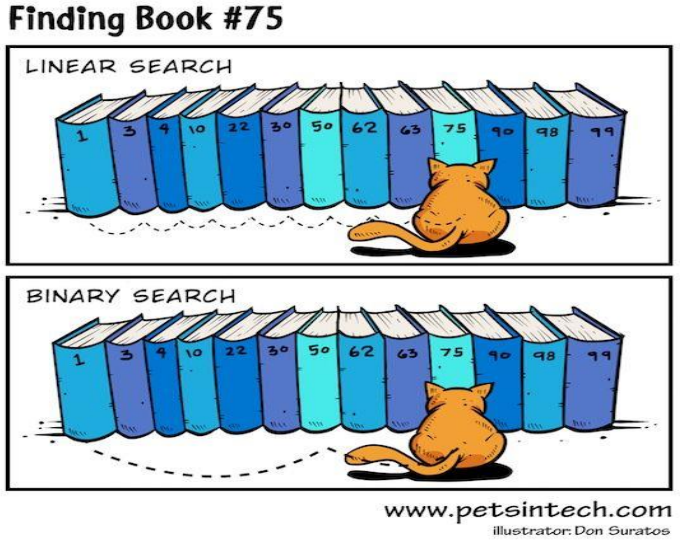
### Quick Sort



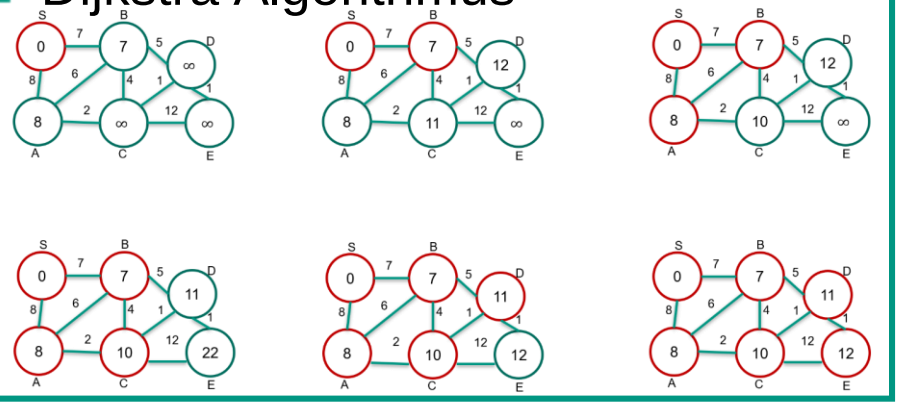
# Wiederholung Übung 2

## Suchalgorithmen

- Lineare Suche und Binäre Suche




- Dijkstra Algorithmus



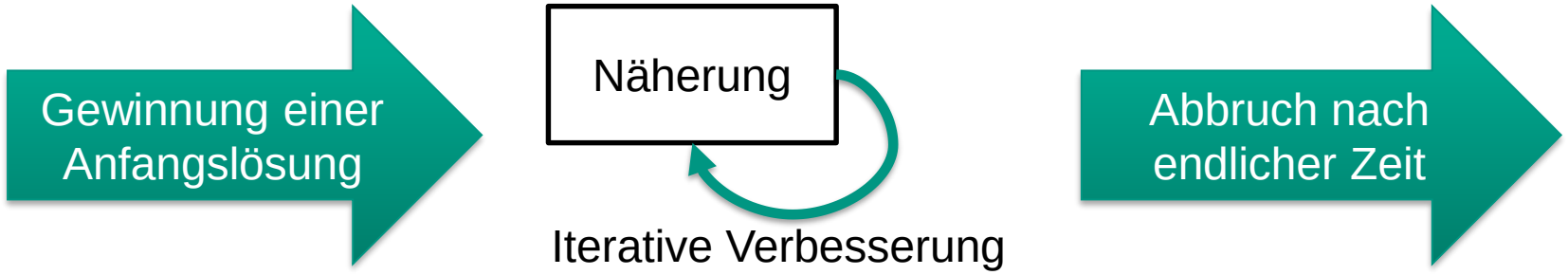
# INHALT ÜBUNG 3

# Optimierungsalgorithmen

## Motivation

- Problemstellung: große Menge möglicher Lösungen, exakte Lösung nicht möglich oder sinnvoll
  - Bewertungs- / Kostenfunktion nötig, damit jede Lösung bewertbar und somit vergleichbar ist
  - Bsp. Parameteroptimierung: alle Parameterwerte und Kombinationen durchprobieren dauert zu lange

- Lösungsidee: **Heuristische Verfahren**
  - nicht ewig nach der perfekte Lösung suchen, wenn eine schnelle, gute Lösung ausreichend ist




- In IT behandelte Verfahren:
  - Anlagerungsverfahren (für Anfangslösung)
  - Random Interchange
  - Kernighan-Lin
  - Greedy
  - Simulated Annealing

# Big Data, Data Mining und Prozesse

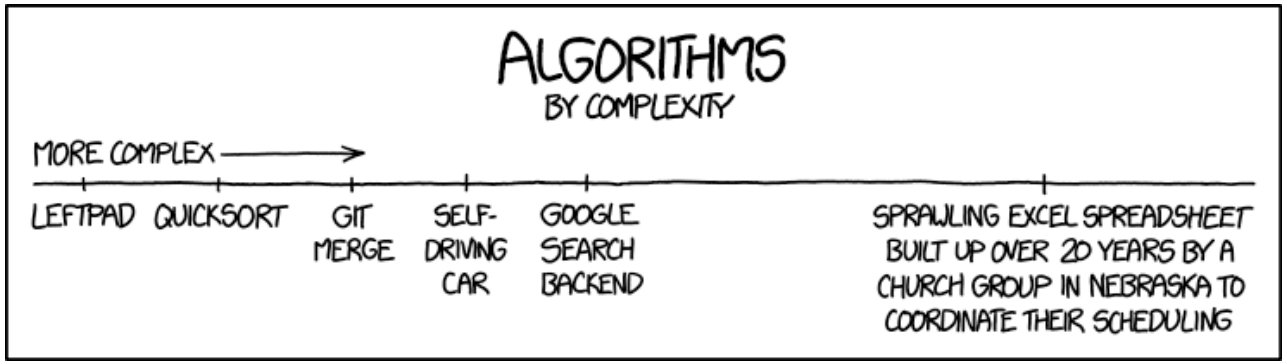
## Schlagwort, Sammelbegriff oder Synonym? Und wozu das Ganze?

- Big Data steht für große Menge an digitalen Daten,  sowie deren Erfassung und Analyse

*Big Data ist ein Synonym für die Bedeutung großer Datenvolumen in verschiedensten Anwendungsbereichen sowie der damit verbundenen Herausforderung, diese verarbeiten zu können.*
*~ Hasso Plattner*

- Datenmenge – wann ist eine Datenmenge „Big Data“?:
  Bis 2003 wurden insgesamt 5.000 Milliarden GB Daten erzeugt → 2011 gleiche Menge in 48h

- Algorithmen entwickeln sich stetig weiter
- Problemstellungen werden stetig abstrakter
- Lösung über „codieren“ nicht mehr effizient




- Big Data als Prozess
  Beinhaltet mehr als nur die Datenmenge und den Algorithmus!

# Ziele der heutigen Übung

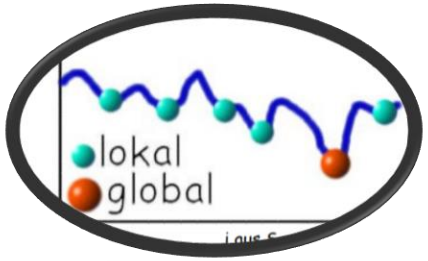
- Nach der heutigen Übung können Sie….

1. … bekannte Optimierungsalgorithmen gegenüberstellen und demonstrieren
2. … Charakteristika, Notwendigkeit und Vorgehensweisen zur Analyse großer Datenbestände beschreiben
3. … gängige Prozessabläufe zur Analyse von Big Data Problemstellungen beschreiben

# TEIL 1:
# OPTIMIERUNGSALGORITHMEN

# Optimierungsalgorithmen
## Motivation

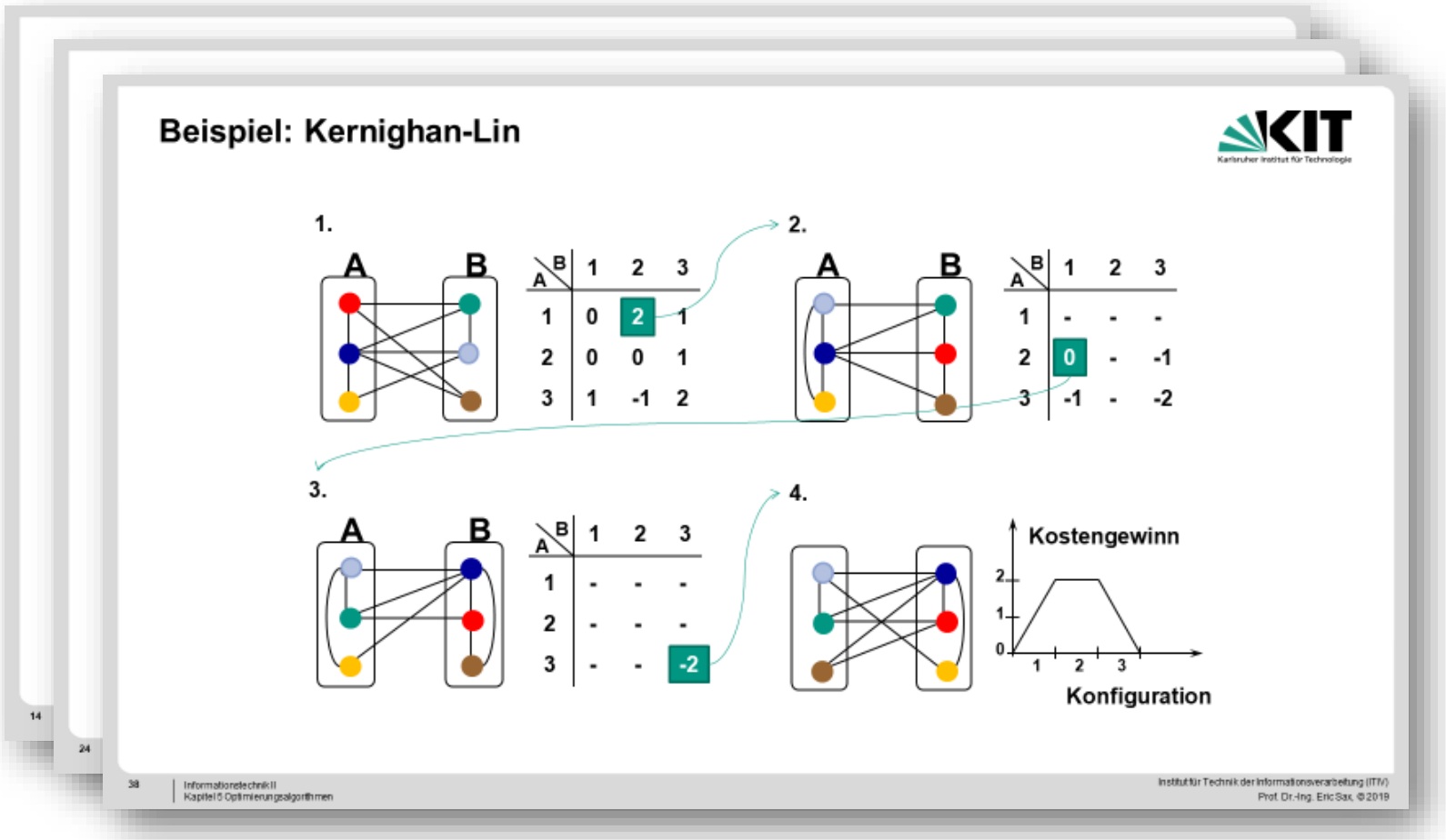
- Optimierung findet man überall – sie ist allgegenwärtig
  - Allgemein: Suchen eines Zustandes, welcher optimale Eigenschaften besitzt (Wirtschaft, Wissenschaft, Gesellschaft, ...)
- Mathematisch betrachtet ist die Optimierung ein Minimierungs- bzw. Maximierungsproblem mit den zwei großen Herausforderungen:
  - Eine geeignete Kosten- bzw. Bewertungsfunktion *f(i)* zu entwickeln
  - Das globale Minimum bzw. Maximum für den Lösungsraum *S* finden




- Numerische Lösungsverfahren (Lösen des gesamten Lösungsraumes)
  - Vorteil: Finden des globalen Optimums
  - Nachteil: Rechenaufwand meist gigantisch bzw. nicht praktikabel
- Heuristische Lösungsverfahren (Finden einer Lösung ohne Betrachtung des gesamten Lösungsraumes, mit Hilfe von Vereinfachungen bzw. dem Zufall)
  - Vorteil: Finden einer akzeptablen Lösung in annehmbarer Zeit
  - Nachteil: Nicht zwingendes Auffinden des globalen Optimums (ggf. sogar beliebig „schlecht“)

# Optimierungsverfahren
## Partitionierung

- Alg. Aufgabe: Finde eine Partition P der Modulmenge M = { $m_1$, ..., $m_i$, ...,$m_n$ } aus Blöcken B
  - Vereinfachung: Finde eine Bi-Partition P = { $B_1$, $B_2$ } mit minimaler Kostenfunktion
- Heuristische Verfahren:
  - Konstruktive Methoden:
    - Anlagerungsverfahren
  - Iteratives Verbessern:
    - Random Interchange
    - Kernighan-Lin
    - Greedy
    - Simulated Annealing



# Optimierungsalgorithmen

## Kernighan Lin

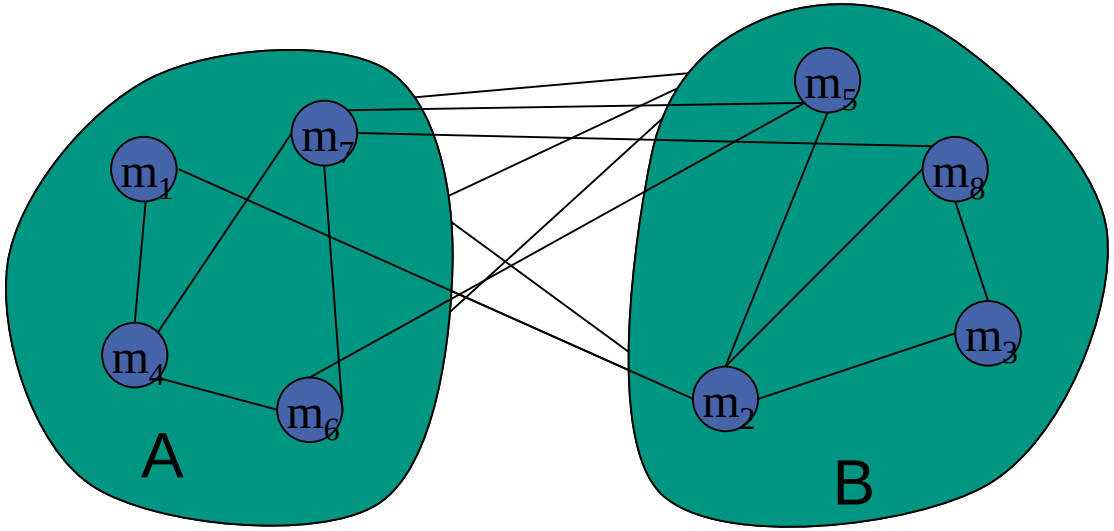
- Optimierung durch Vertauschen von immer zwei Knoten ($m_i$, $m_j$) aus unterschiedlichen Partitionen
- Kostenveränderung durch Tausch von ($m_i$, $m_j$):

$$D(m_i,m_j) = E(m_i) - I(m_i) + E(m_j) - I(m_j) - 2\,c_{ij}$$

Externe Kosten

Interne Kosten

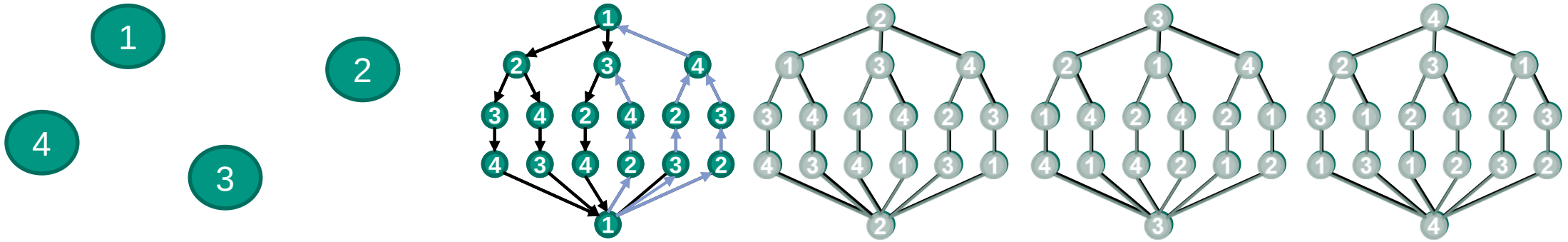
Kante zwischen $m_i$, $m_j$ bleibt extern, darf also nicht gezählt werden



# Optimierungsalgorithmen

## Travelling Salesman Problem (TSP)

- Problemstellung: Ein Handlungsreisender muss n Städte auf einer Rundreise so besuchen, dass er versucht dem kürzesten Weg dabei zu folgen.
  - TSP beschreibt sehr gut die große Klasse der Optimierungsprobleme, da es:
    - äußerst einfach formuliert ist und
    - sehr schwierig (effizient) zu lösen ist
- Frage: bei nur 4 Städte, wie viele Lösungskandidaten müssen per („klassischem“) numerischem Algorithmus berechnet werden, um die optimale Lösung zu finden?




- Zusatzfrage: wie viele bei 50 Städten?

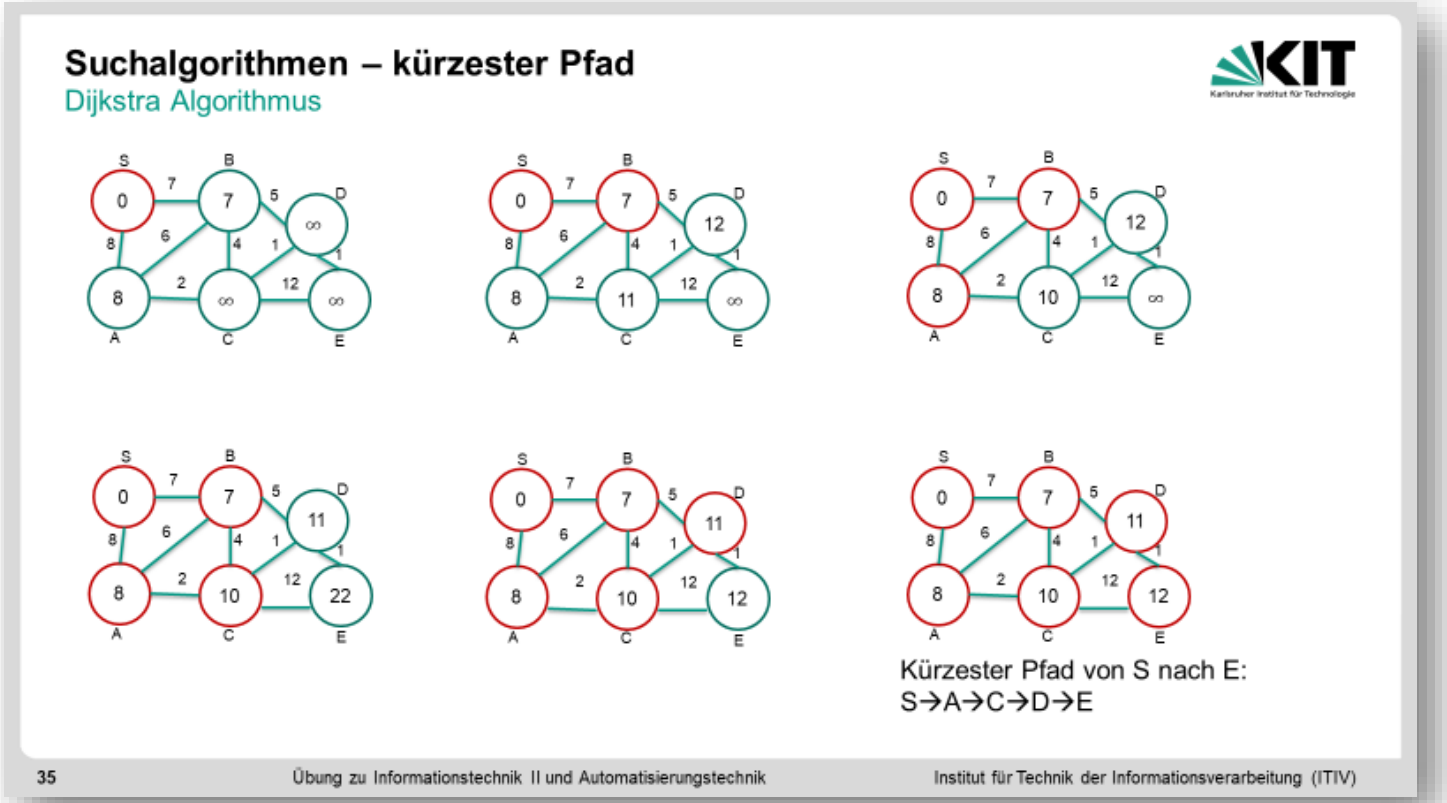
$$= \frac{(n-1)!}{2} = \frac{49!}{2} \approx 300\ Dezillionen\ Lösungen = 3 * 1062\ Lösungen$$ *(Sonne besteht aus ca. $10^{57}$ Atomen)*

(3*100 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000 000)

# Optimierungsverfahren
## Greedy

- Mit Greedy-Algorithmen ist eine Familie bzw. Klasse an Algorithmen gemeint, welche dem Grundprinzip folgen in jedem Verfahrensschritt diejenige Entscheidung zu treffen, die in diesem Moment am besten ist (ohne Berücksichtigung zukünftiger Schritte)
  - Beispiele von Greedy-Algorithmen:
    - Prim Algorithmus
    - Kruskal Algorithmus
    - **Dijkstra Algorithmus**

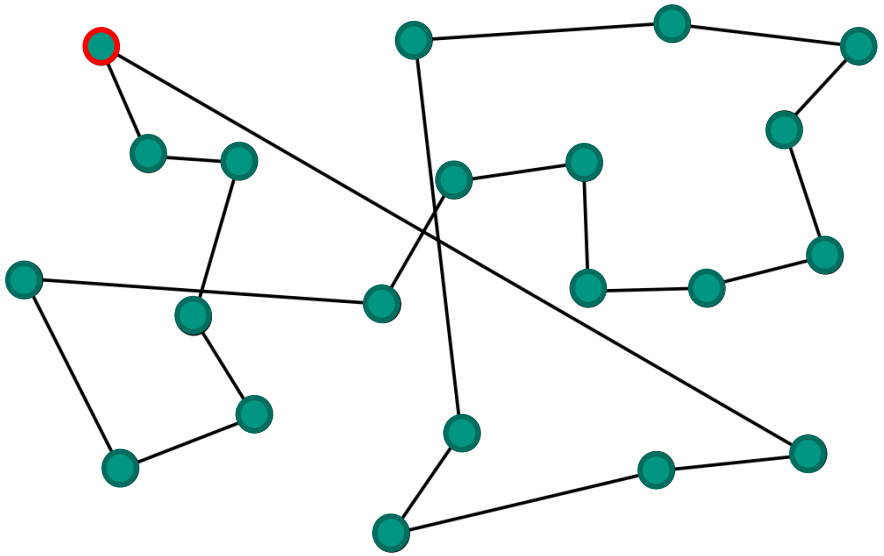
Folie zu Dijkstra aus Übung Nr. 2



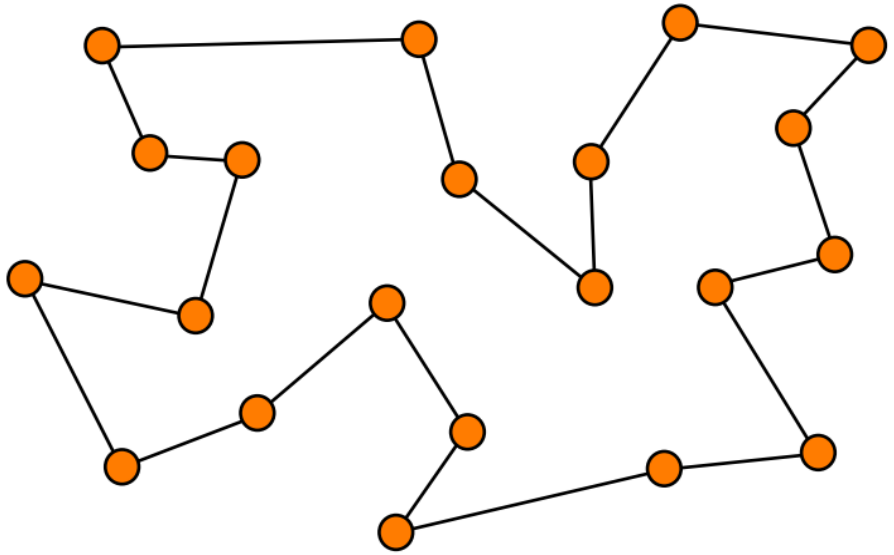
# Optimierungsverfahren

## Greedy

- Mit Greedy-Algorithmen ist eine Familie bzw. Klasse an Algorithmen gemeint, welche dem Grundprinzip folgen in jedem Verfahrensschritt diejenige Entscheidung zu treffen, die in diesem Moment am besten ist (ohne Berücksichtigung zukünftiger Schritte)
  - Problem von Greedy-Algorithmen bezogen auf TSP mit 21 Städten ($^{20!}/_{2} \approx 1{,}2 * 10^{21}$ Möglichkeiten): Führen Sie einen Greedy-Algorithmus auf das unten gezeigte TSP mit 21 Städten mittels einer Nearest-Neighbor Heuristik durch! (Beginnen Sie bei der roten „Stadt“)



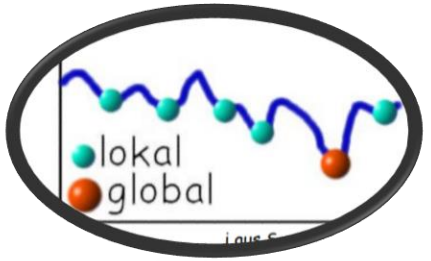
Ergebnis nach Nearest-Neighbor Heuristik

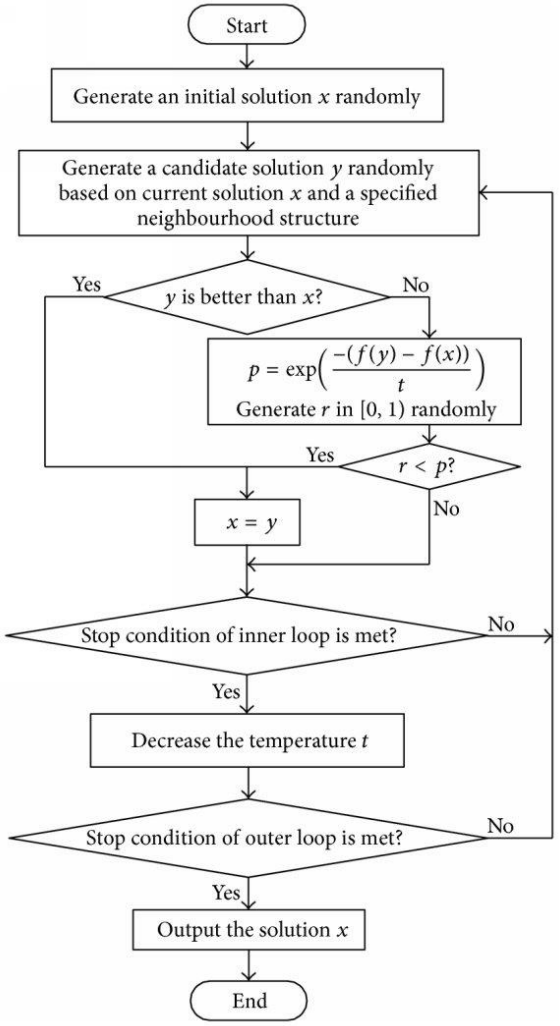


Optimale Lösung

# Optimierungsverfahren
## Simulated Annealing

- Durch das integrierte Zufallsprinzip wird dem Simulated Annealing erlaubt lokale Optima zu überwinden („Hill Climbing“) und somit das globale Optimum zu finden.
- Grundstruktur des Algorithmuses:
  1. Lege eine Initiallösung fest (erster lokaler Suchraum)
  2. Wähle im Umkreis des Zentrums (lokaler Suchraum) einen Lösungskandidaten aus
  3. Entscheidung ob der Lösungskandidat die neue Lösung werden soll
     - Besser Lösung in jedem Fall übernehmen
     - Schlechtere Lösung wird nur übernommen wenn $p(\Delta E, T) \geq random[0,1)$ ist, wobei $p(\Delta E, T) = e^{-\Delta E/T}$ ist
  4. Bestimmung des neuen Zentrums und Abkühlung
     - Verschiebung des Zentrums (oder eben nicht)
     - Abkühlung: $T = \alpha * T, wobei\ \alpha\ \in [0,1)$
  5. Weiter bei 2. bis Abbruchkriterium erfüllt



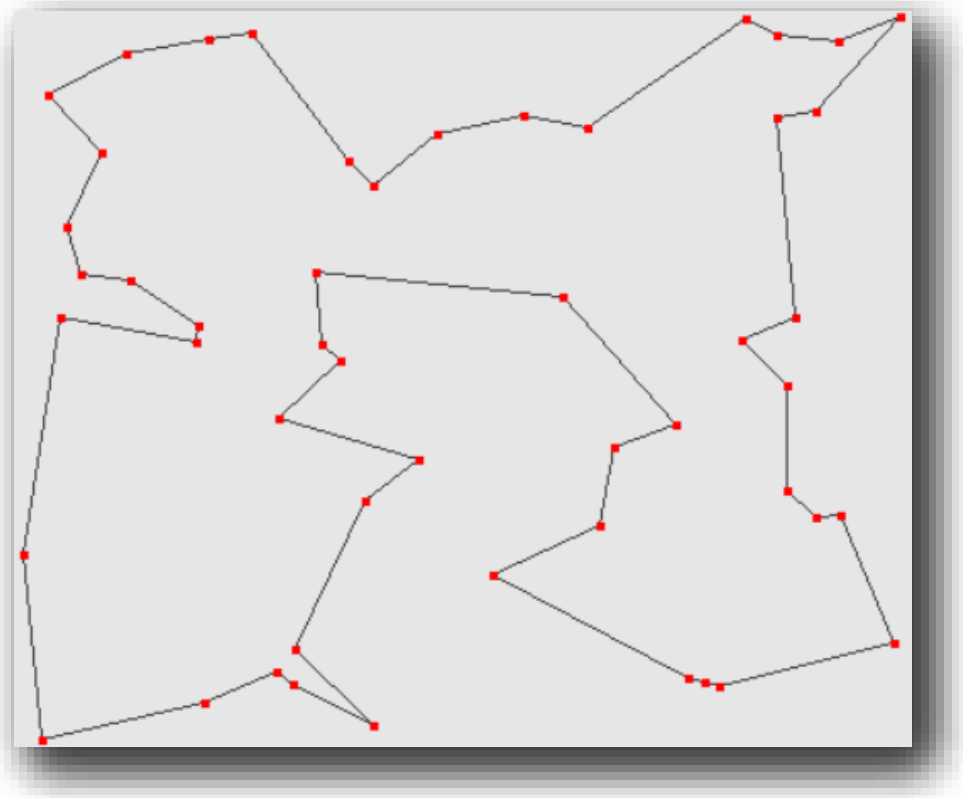




➢ Am Anfang der Optimierung → Suche eines „guten“ lokalen Suchraums

➢ Am Ende der Optimierung → Suche des Minimums im lokalen Suchraum

# Optimierungsalgorithmen
## Vergleich

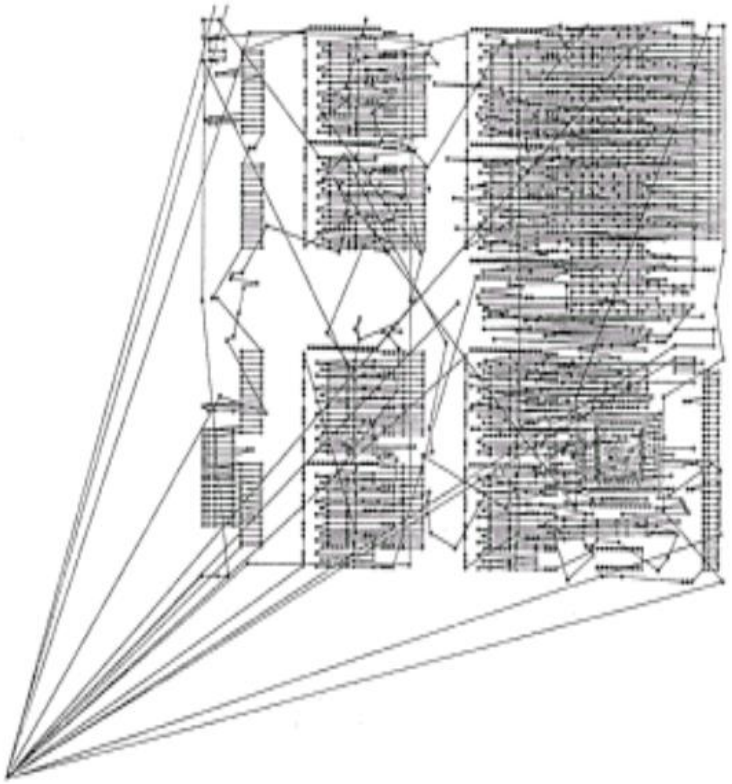
- Vergleich der Verfahren am „Travelling Salesman Problem“ (TSP)
  - Beispiel TSP mit 50 Städten (→ $3*10^{62}$ mögliche Lösungen)
  - Numerisches Verfahren:
    - Bei $10^6$ Lösungen pro Sekunde (1µs / Lösungskandidat) wären ca. $9*10^{48}$ Jahre Berechnungszeit nötig um perfekte Lösung zu finden
    - selbst bei „nur“ 21 Städten sind es ca. 38.573 Jahre
  - Nearest-Neighbor-Alg. (Greedy Algorithmus)
    - $50 * (49 + 48 + \cdots + 1) = 50 * \frac{49^2+49}{2} = 61.250$ Lösungskandidaten abgesucht
    - Gefundene Lösung nach 3 Sekunden Rechenzeit
  - Simulated Annealing
    - mit Initiallösung E = 11603
      - Parameter $T_0 = 10$; $\alpha = 0{,}999$
    - 36.188 Lösungskandidaten abgesucht
    - Gefundene Lösung nach 60 Sekunden Rechenzeit
  - Optimale Lösung: unbekannt!



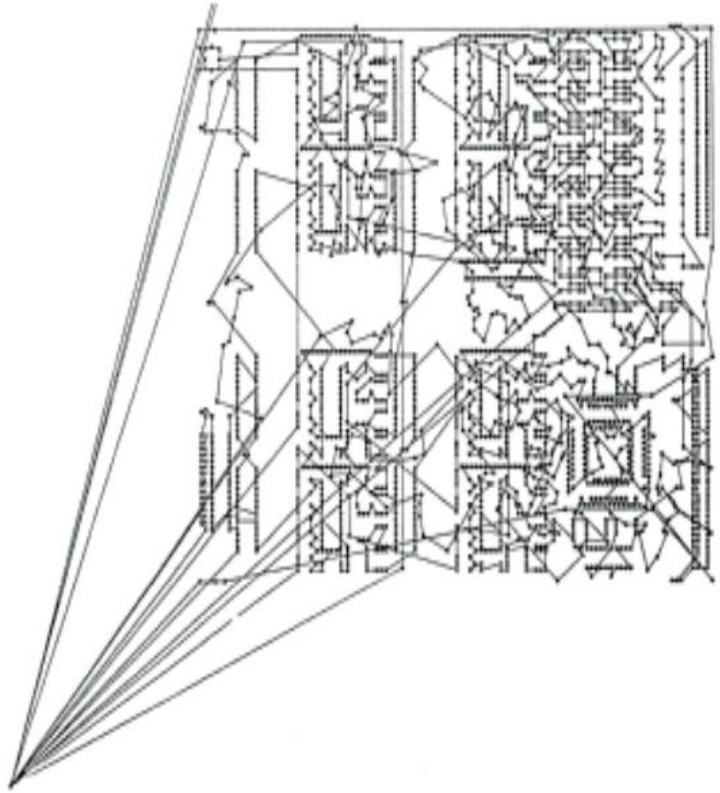
# Optimierungsalgorithmen

## Beispiel aus der Praxis

- Verfahrewege bei einer Fräsmaschine zur Platinenherstellung



Vorher



Nachher

# Ziele der heutigen Übung

- Nach der heutigen Übung können Sie….

1. … bekannte Optimierungsalgorithmen gegenüberstellen und demonstrieren
2. … Charakteristika, Notwendigkeit und Vorgehensweisen zur Analyse großer Datenbestände beschreiben

2.1 … Notwendigkeit zur Analyse großer Datenbestände nennen

2.2 … Charakteristika zur Abgrenzung von Big Data nennen

2.3 … Grundbegriffe im Bezug zu Big Data nennen

2.4 … den Begriff des „Maschinelles Lernens“ abgrenzen

2.5 … den Begriff und das Vorgehen beim „Trainieren“ abgrenzen+

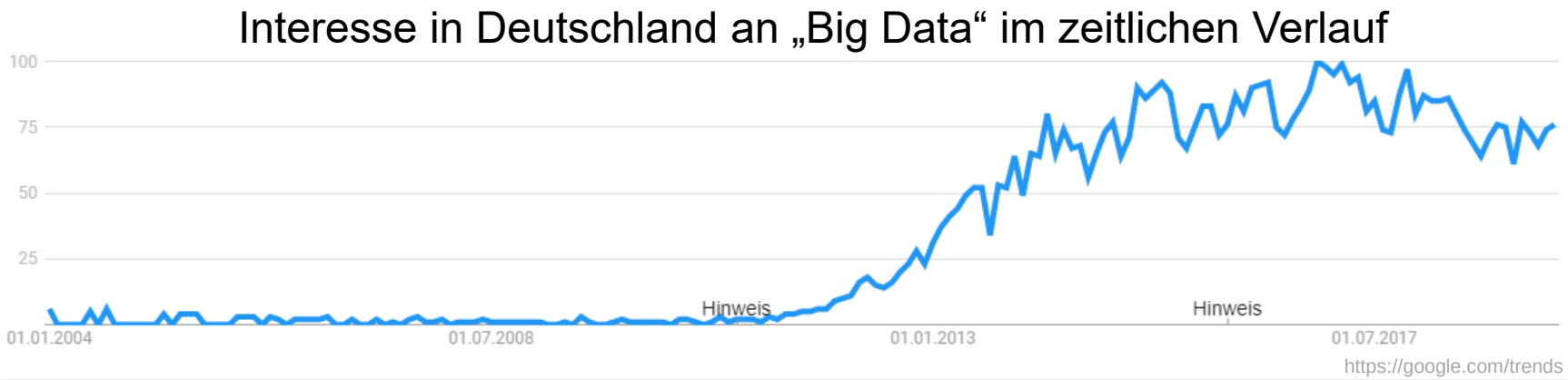
# TEIL 2:
# CHARAKTERISTIKA ZUR ANALYSE GROSSER DATENMENGEN
# TEIL A – DIE 5V‘S

# Big Data - Charakteristika

## Schlagwort, Sammelbegriff, Synonym

- Erstmalige Publikation des Begriffes „Big Data“ im Jahre 1997
  Definition damals: Daten die nicht mehr in den Hauptspeicher oder einen Massenspeicher passen
- Begriff selbst wird allerdings frühzeitig kontrovers diskutiert
  Erster Artikel auf Wikipedia (2009) wurde prompt gelöscht, mit folgender Begründung:
  *„Delete as per nome – it is simply a combination of big and data, dictionary words which have no place here. I‘m not even sure it‘s a neologogism, and even if it was it doesn‘t need an article“* ~ *John Blackburne*

- Akzeptanz erst nach Aufnahme des Begriffes von größeren IT-Häusern wie IBM, SAP oder Oracle
- Anstieg der Begriffsnennung von 2009 bis 2012 um 1211%




Begriff muss abgrenzbar und bewertbar sein
→ Die V‘s von Big Data!

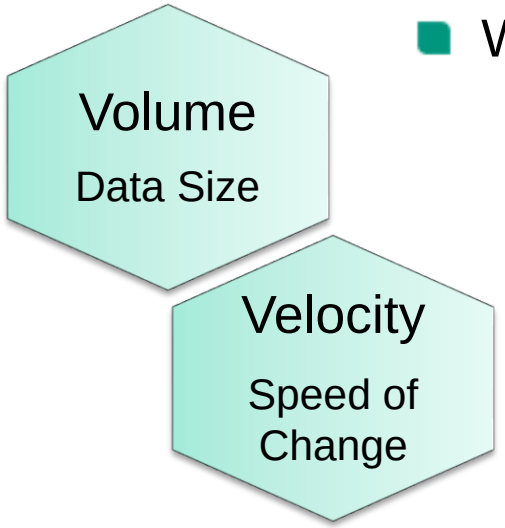
# Big Data - Charakteristika

## Begriffserklärung

- Definitionsversuche von „Big Data“:
  - Der Begriff „Big Data“ bezeichnet eine Datenmengen, welche beispielsweise zu groß, zu komplex, zu schnelllebig oder / und zu schwach strukturiert sind, um sie mit manuellen und herkömmlichen Methoden der Datenverarbeitung auszuwerten [*Wikipedia*]
  - Es bezeichnet also nicht die DatenMENGE alleine, sondern gleichbedeutend auch die Herausforderung Datenmengen zu sammeln, zu speichern und gewinnbringend auszuwerten
  - Aufgabe / Frage:
    - Bennen und beschreiben Sie DIE V‘s von „Big Data“
    - Ist das alles was „Big Data“ ausmacht? Fallen Ihnen noch weitere V‘s ein?
    - Warum ändert „Big Data“ die Analyse von Daten?

# Big Data - Die drei grundlegenden V‘s

## Volume und Velocity




- Wie groß ist groß?
  - Byte: 1 Reiskorn
  - KB: 1 Tasse Reis
  - MB 8 Reissäcke
  - GB 3 Trucks voll Reis
  - TB 2 Containerschiffe
  - PB bedeckt Karlsruhe
  - EB: bedeckt Deutschland
  - ZB: füllt den Pacific
  - YB: Erdgroßer Reisball

- Änderung der Definition von „vielen Daten“ ändert sich stetig
- 1992 : 100GB/Tag
  2018: 50.000GB/Sekunde
- Big Data passt sich der Datenproduktion an

# Big Data - Die drei grundlegenden V‘s

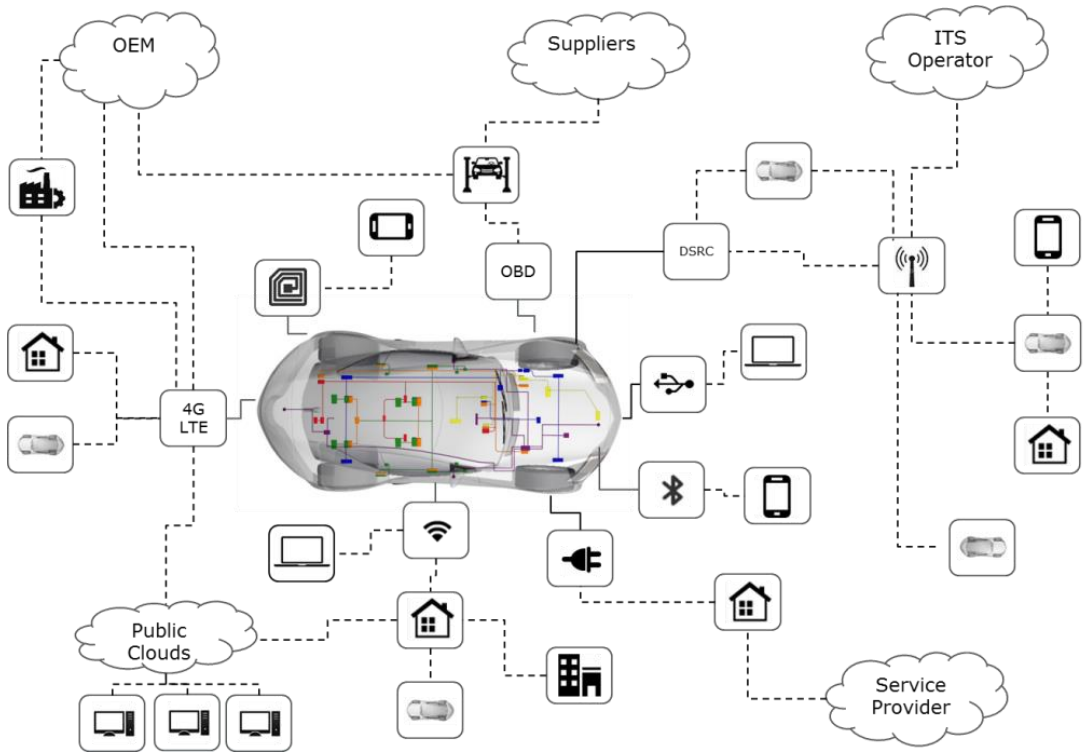
## Variety – Am Beispiel Fahrzeug

**Variety**

Different Forms of Data Sources

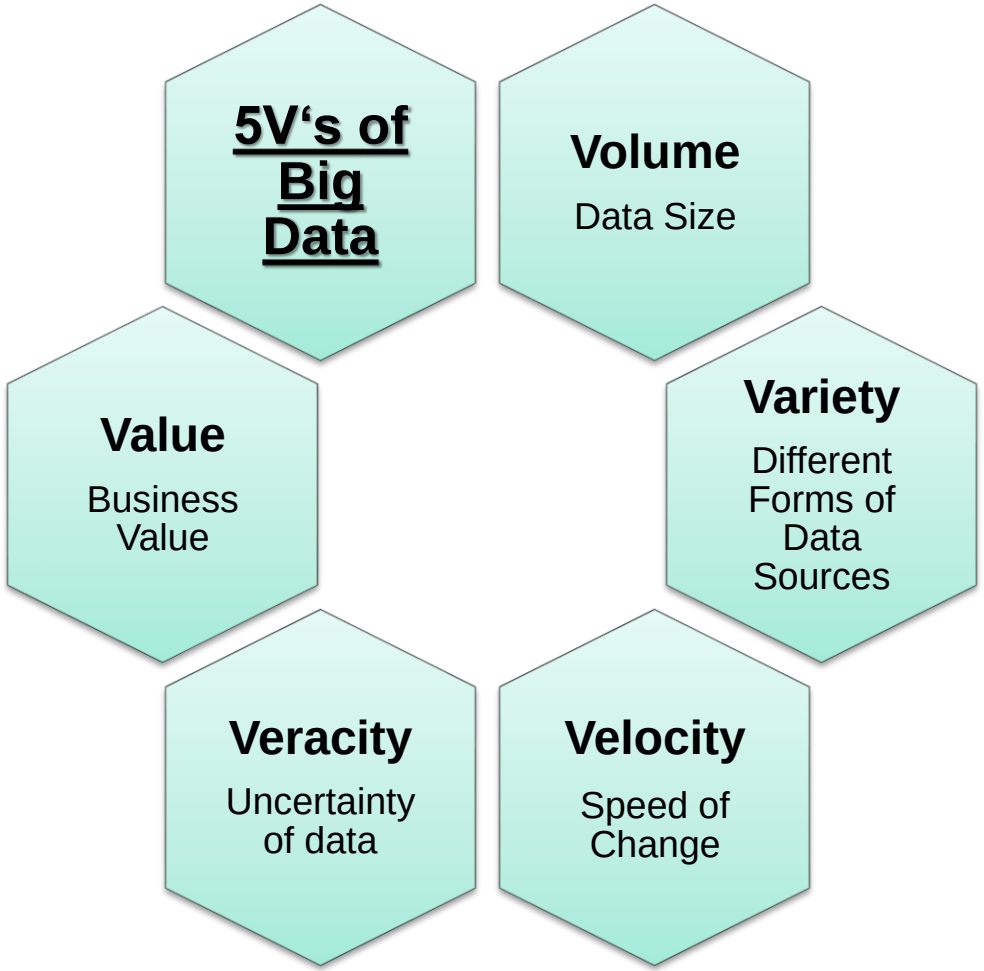
- Daten werden auf einer Vielzahl von unterschiedlichen Quellen produziert und im Anschluss oft fusioniert
- Herausforderungen
  - unterschiedliche Datenquellen
  - unterschiedliche Datentypen
  - unterschiedliche Datenformen

- Vom Fahrzeug aufgenommene und teilweise übertragene Daten
  - GPS-Daten
  - Kilometerstand
  - Verbrauch
  - Gurtstraffungen
  - Reifendruck
  - Und vieles mehr



# Big Data - Charakteristika

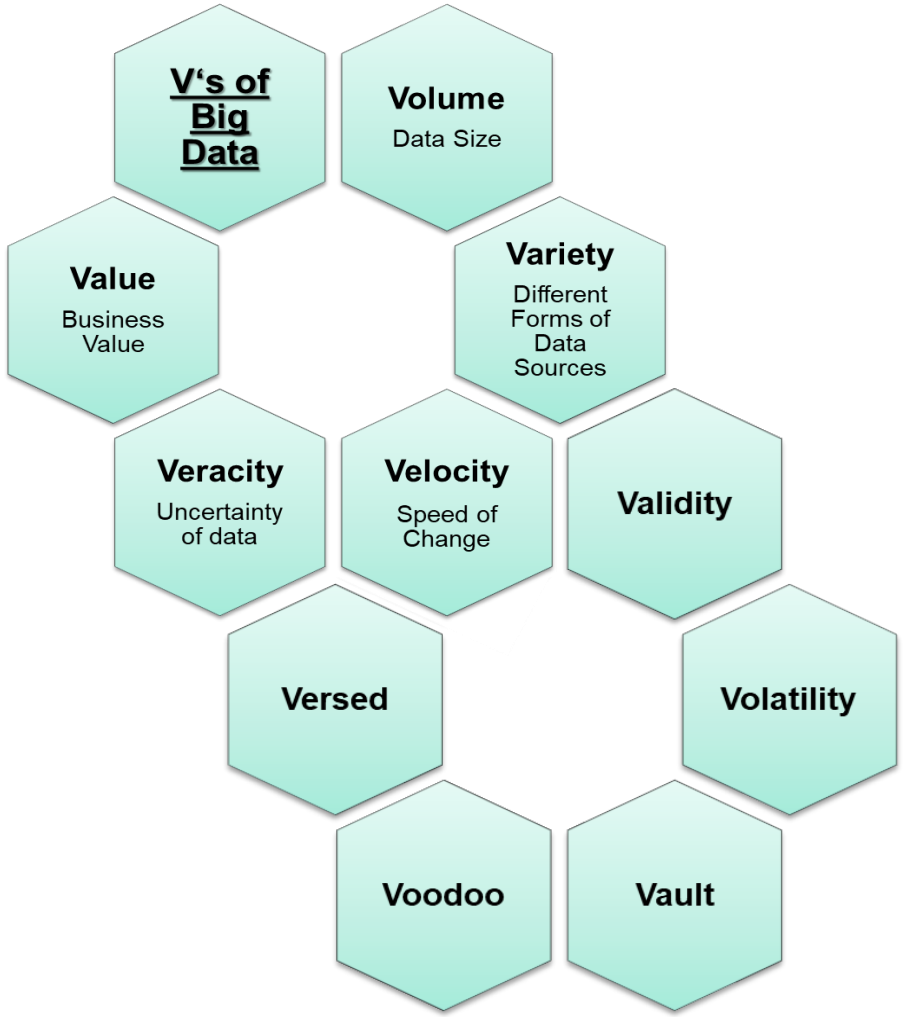
## Die 5 V‘s




- **Volume:**
  Bezeichnet schlicht die enorme Menge an Informationen. Mit der Speicherung, aber vor allem mit der Verarbeitung, sind die meisten Datenbanksysteme (aktuell im Petabyte-Bereich) überfordert
- **Variety:**
  Damit ist die Vielfalt der zur Verfügung stehenden Daten und -quellen gemeint. Hier kann man z.B. unterscheiden zwischen Unternehmens- und Fremddaten oder zwischen Bild und Textdaten.
- **Velocity:**
  Geschwindigkeit mit der Daten generiert, ausgewertet und weiterverarbeitet werden sollen.
- **Veracity:**
  Zuverlässigkeit, Sinnhaftigkeit, Vertrauenswürdigkeit: Veracity beschreibt das Problem, dass das Ergebnis einer Big-Data-Analyse stark davon abhängt, welche Qualität die Daten haben.
- **Value:**
  Für sich gesehen besitzen Daten keinen Wert. Value meint den wirtschaftlichen Wert von Big Data für ein Unternehmen oder Nutzer, der durch geeignete Analysen gewonnen werden kann.

# Big Data - Charakteristika

## weitere V‘s

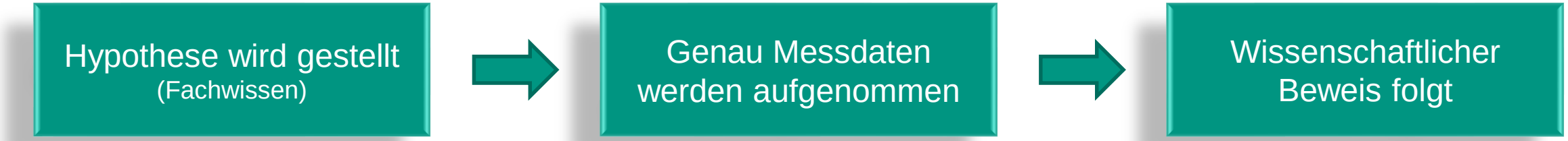



- **Validity or Vagueness:**
  Die Aussagen der Big Data Lösungen basieren nur auf den verwendeten Daten, weswegen die Aussage für weitere Daten nicht mehr valide sein muss oder vage bleibt.
- **Volatility:**
  Damit ist das Problem gemeint, wie lange die Daten noch Gültigkeit haben und somit nützlich und gespeichert bleiben müssen bzw. eben nicht mehr verwendet werden dürfen.
- **Vulnerability or Vault:**
  Viele Big Data Lösungen basieren auf sensiblen, personenbezogen Daten, welche gesondert abgesichert werden müssen.
- **Visualization or Voodoo:**
  Damit ist die Herausforderung gemeint den Lösungsweg von Big Data Anwendungen für den Menschen nachvollziehbar bzw. interpretierbar darzustellen.
- **Versed:**
  Big Data Lösungen setzten Wissen in unterschiedlichen Fachgebieten (Mathematik, Datenbanken, Statistik, Programmierung, Business- und Datenverständnis etc.) voraus. Expterten müssen als Team zusammenarbeiten.

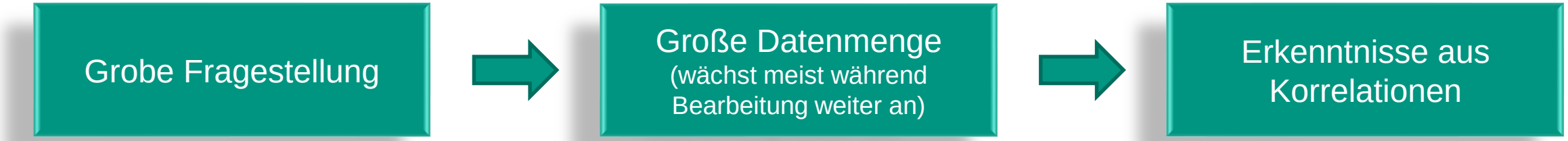
# Big Data - Potential

## Hypothese vs. Schlussfolgerung

- Hypothesengestützte Erkenntnisgewinnung
  Kausalität: auf der Annahme einer Hypothese/Grundannahme werden die dafür benötigten Daten aufgenommen und ausgewertet




- Schlussfolgerungen werden auf Korrelationen aufgebaut
  Fragestellung zu Anfang darf sehr grob gehalten werden
  Echtzeitfähigkeit möglich/zu erreichen



# Big Data - Hypothese vs. Schlussfolgerung

## Zwischenübung

- Gegeben sind folgende Hypothesengestützte Annahmen. Diskutieren Sie wie man eine jeweilige grobe Fragestellung für einen Big Data Ansatz formulieren könnte und wie die drei grundlegendes V‘s darin enthalten sein könnten:

| | Hypothese |
|---|---|
| 1. | Wer einen Golf fährt, wird wieder einen Golf kaufen |
| 2. | Werden die Oliven dieses Jahr wieder so gut wachsen, wie letztes Jahr? |
| 3. | Wird sich die Schweinegrippe die Florida ausbreiten? |
| 4. | Werden vermehrt Taschenlampen kurz vor einem Hurricane gekauft? |

# Big Data - Hypothese vs. Schlussfolgerung

## Zwischenübung – Lsg (Vorschlag)

- Gegeben sind folgende Hypoth… …kutieren Sie wie man eine jeweilige grobe Fragestellung f… …ulie… grundlegendes V‘s darin entha…

| | Hypothese / grobe Fragestellung |
|---|---|
| 1. | Wer einen Golf fährt, wird wieder… <br> Gibt es Auffälligkeiten bezüglich des … …altens von Golffahrer… |
| 2. | Werden die Oliven dieses Jahr w… …so gut wachsen, wie le… …es J… <br> Wie wird sich die Olivenernte i… …nster Zeit verhalten? |
| 3. | Wird sich die Schweinegri… …e die Florida ausbreiten? <br> Wie breitet sich die Sch… …einegrippe aus? |
| 4. | Werden vermehrt Taschenlampen kurz vor einem Hurrica… <br> Welche Produkte werden im Bezug auf Hurricanewarnung… |

2004 untersuchte Walmart die Kundendaten nach Auffälligkeiten im Verkaufsmuster während einer Hurricanewarnung. Erkenntnis war ein enormer Anstieg an „Strawberry Pop Tarts“


Volume
Nutzer-daten

Klimaveränderungen der letzten Jah… zugrunde liegende Daten ergeben Vorhersagen über Ernteverhalten verschiedener Obst-/Gemüsesorten

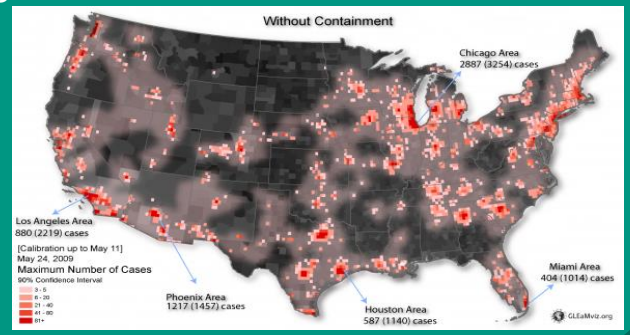
Variety
Wetter-daten

**Italien geht das Olivenöl aus**

Italien wird in Sachen Olivenöl bald von Importen abhängig sein. Italien gehe bereits im April das Olivenöl aus, so der italienische Klimaforscher Riccardo Valentini, Chef des Europa-Mittelmeer-Zentrums für Klimawandel in Italien. Der Ernteertrag ist um 57 Prozent gefallen, die Saison 2018/19 gilt in Italien bereits jetzt als schlechteste Saison seit 25 Jahren.

Google konnte im Jahre 2009 die Verbreitung der Schweinegrippe durch Untersuchung der Suchanfragen der Nutzer vorhersagen

Velocity
Such-anfragen

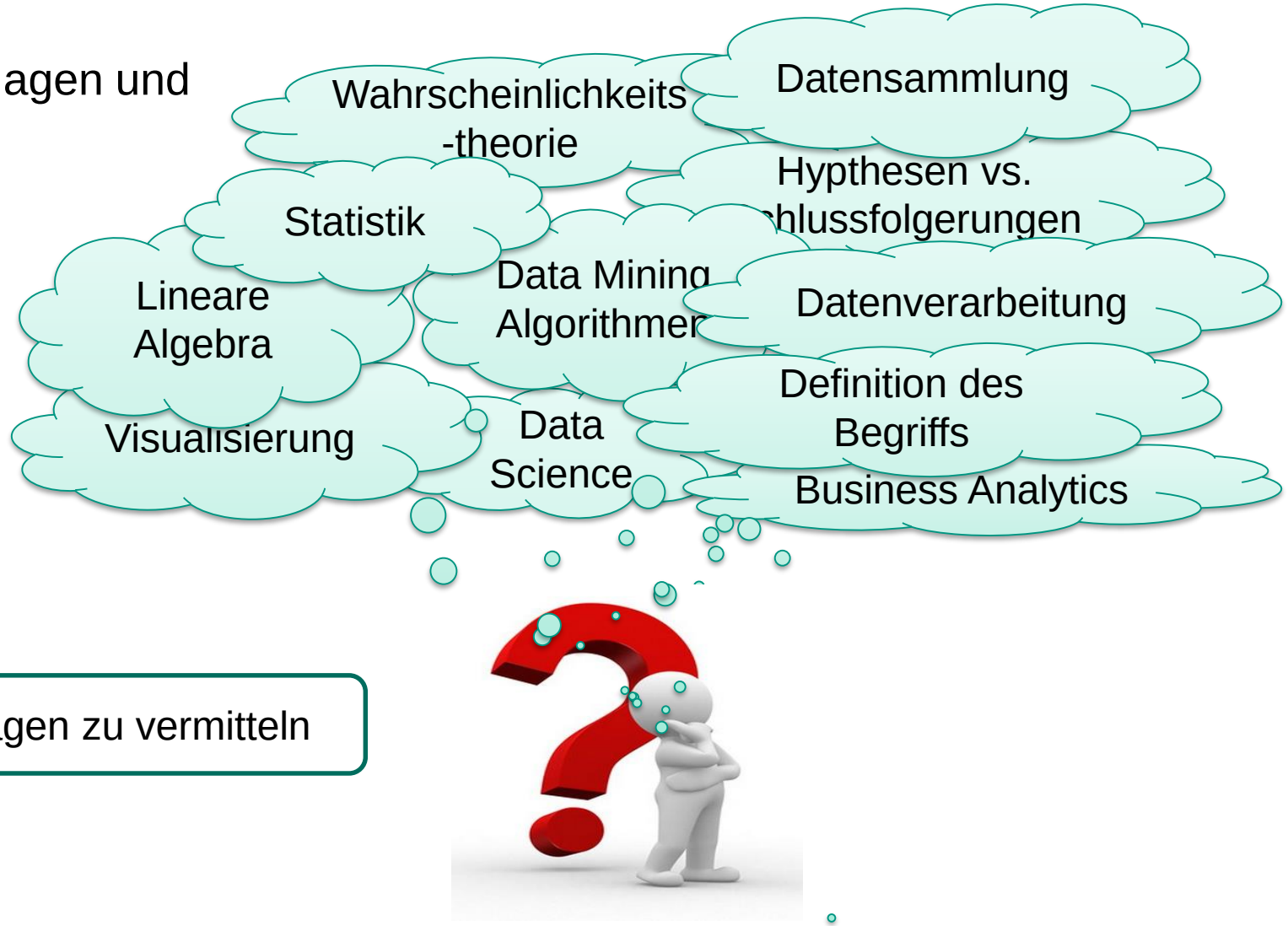


- V‘s of Big Data
  - Volume
  - Variety
  - Velocity
  - Veracity
  - Value
- Hypothesengestützt vs. Schlussfolgerungen

# TEIL 2:
# CHARAKTERISTIKA ZUR ANALYSE GROSSER DATENMENGEN
# TEIL B – GRUNDLAGEN

# Big Data
## Benötigte Grundlagen

- Big Data als umfassender Begriff
- Es werden unterschiedliche Grundlagen und Begriffe benötigt



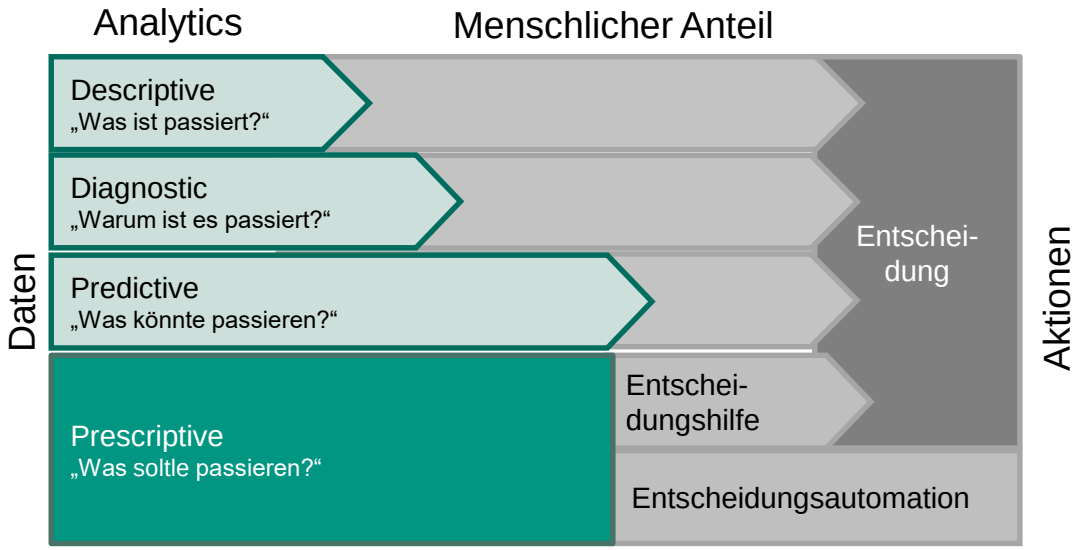

Ziel in IT2 diese und weitere Grundlagen zu vermitteln

# Big Data - Grundlagen

## Unterschiedliche Zielsetzungen von Big Data

- Es geht darum Daten zu analysieren, um daraus Erkenntnisse zu gewinnen.
- Erkenntnisse dienen als Grundlage einer Entscheidung und lösen Handlungen aus.

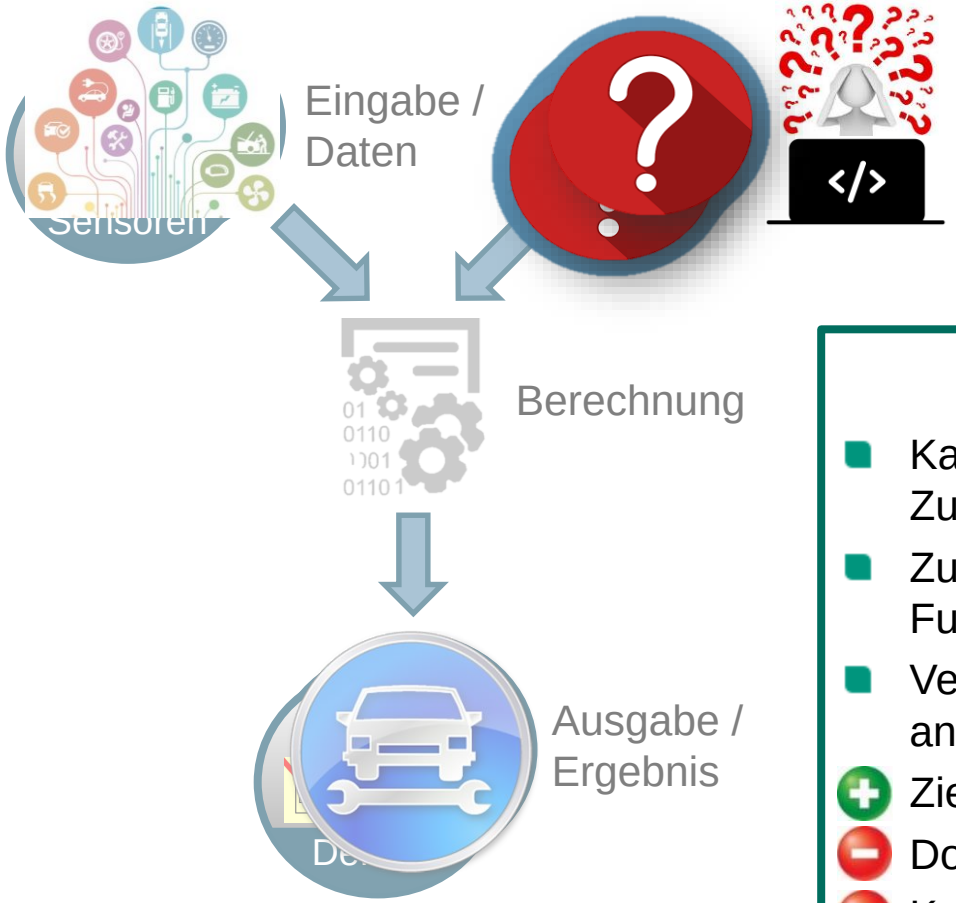
- Aufteilung Analytics in
  1. Descriptive Analytics
     Auswertung von zur Verfügung stehender Daten
  2. Diagnostic Analytics
     Analyse der Hintergründe des Ereignisses
  3. Predictive Analytics
     Vorhersage zukünftiger Ereignisse auf Basis historischer Wirkungszusammenhänge (Korrelationen)
  4. Prescriptive Analytics
     Zur Ableitung von Handlungsempfehlungen



# Motivation: Maschinelles Lernen

## Klassische Entwicklung / Analyse vs. Maschinelles Lernen

- Klassische Entwicklung / Analyse




**klassisch**

- Kausale Überlegung, welcher Zusammenhang existiert
- Zusammenhang wird als Funktion programmiert
- Verifikation dieser Annahmen anhand neuen Eingaben / Daten
- Zielgerichtet
- Domänenwissen notwendig
- Keine neuen Zusammenhänge

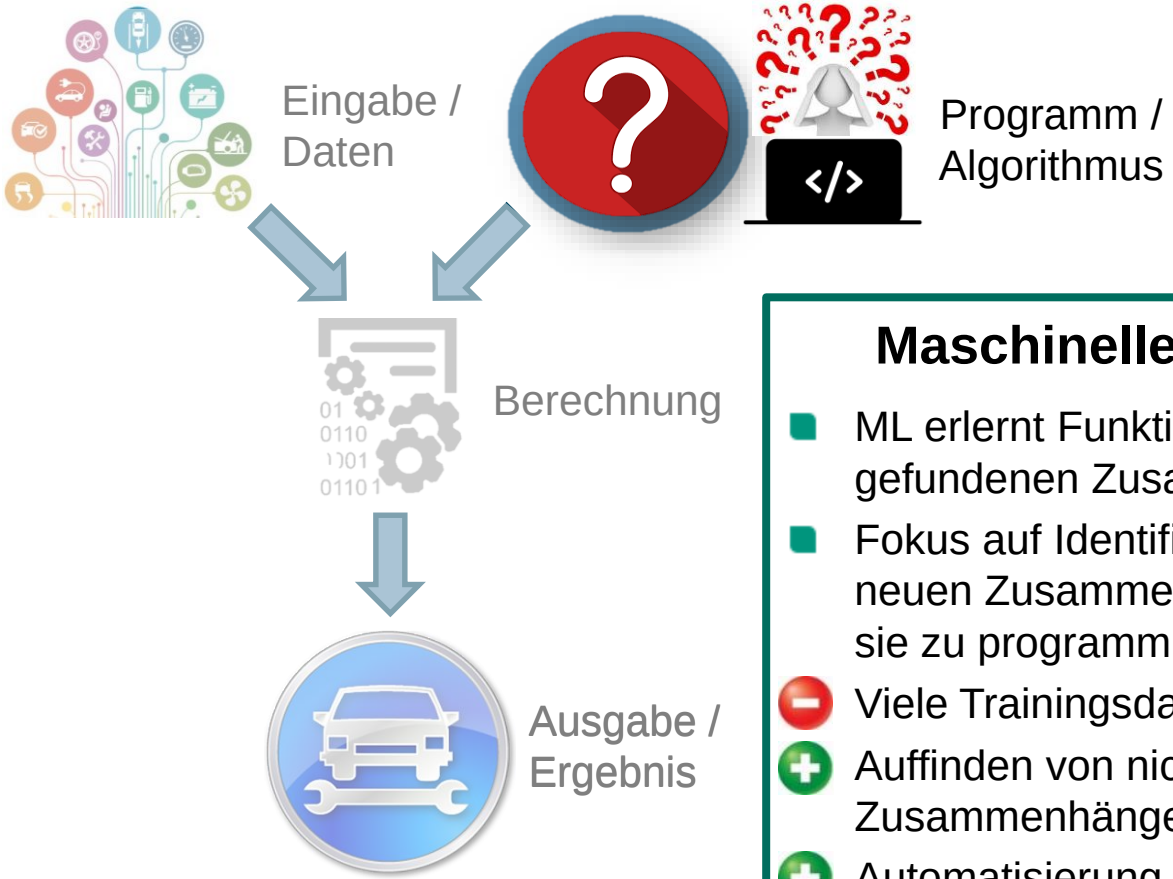
→ Die wirklichen Zusammenhänge sind nicht bekannt

→ Zu viele Zusammenhänge / Abhängigkeiten / Variationen / Ausnahmen um effizient zu programmieren

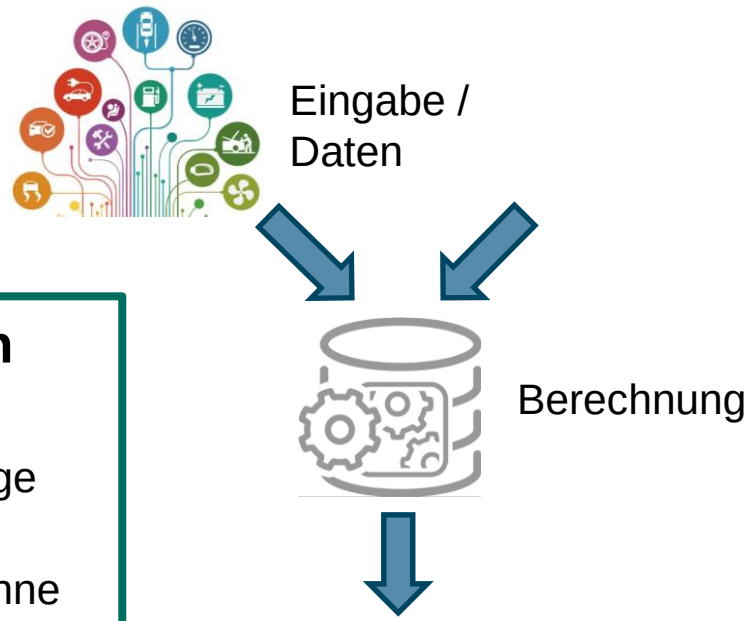
# Motivation: Maschinelles Lernen

## Klassische Entwicklung / Analyse vs. Maschinelles Lernen

- Klassische Entwicklung / Analyse




- Maschinelles Lernen




### Maschinelles Lernen

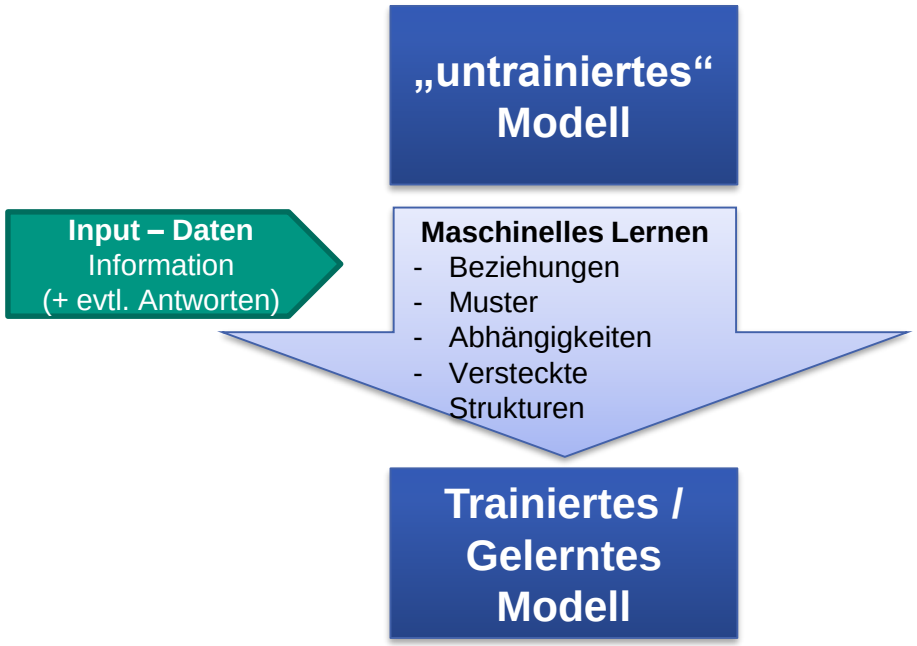
- ML erlernt Funktion der gefundenen Zusammenhänge
- Fokus auf Identifikation von neuen Zusammenhängen ohne sie zu programmieren
- ⊖ Viele Trainingsdaten erforderlich
- ⊕ Auffinden von nicht vermuteten Zusammenhängen
- ⊕ Automatisierung der Verfahren auf weitere gleichartige Daten

# Maschinelles Lernen

## Modelle bzw. ML-Modell

- Was ist ein Modell?
  - Eine Spezifikation zwischen mathematischen Beziehungen zwischen verschiedenen Variablen
  - Modellbegriff nach Stachowiak durch drei Merkmale
    - Abbildung: Ein Modell ist eine Abbildung/Repräsentation eines Originals
    - Verkürzung: Ein Modell umfasst in der Regel nicht alle Attribute des Originals
    - Pragmatismus: Modelle sind den Originalen nicht eindeutig zugeordnet, sie erfüllen eine Ersetzungsfunktion

---

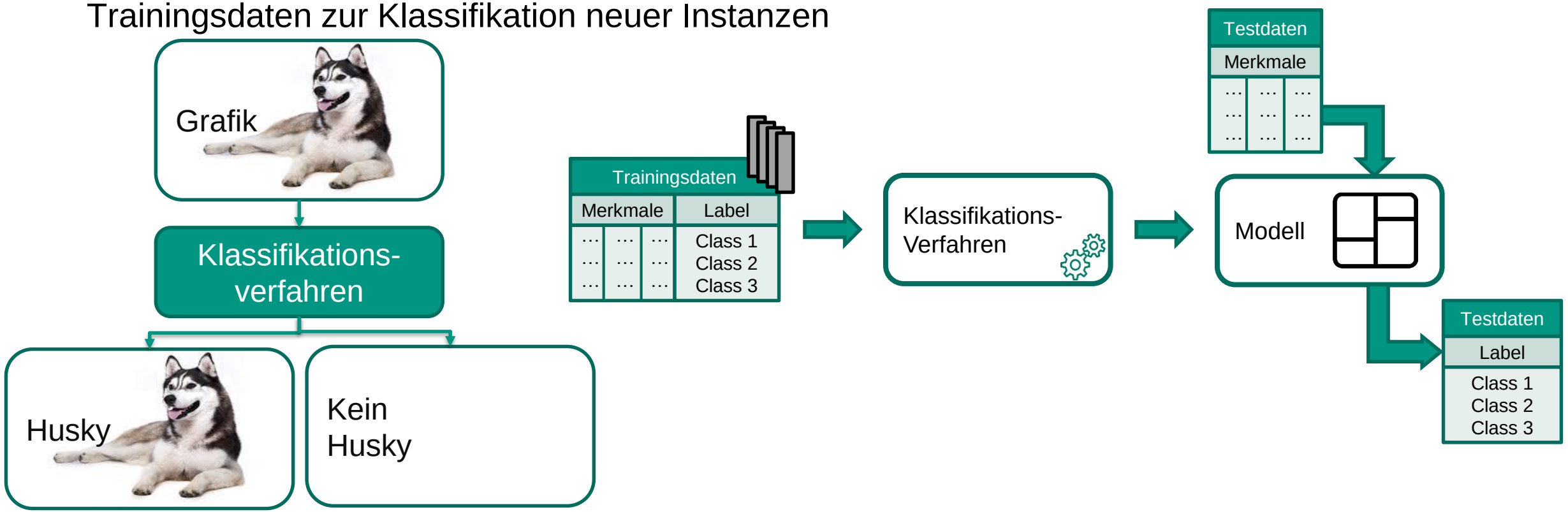



- Erstellung und Verwendung von Modellen, welche aus Daten gelernt werden
- Ziel ist aus vorhandenen Daten Modelle zu entwickeln, mit welchen verschiedene Ergebnisse für neue Daten vorhergesagt werden können
- Ergebnisse, die eventuell vorhergesagt werden können
  - Neue E-Mail als Spam erkennen
  - Werbung anzeigen, auf welche der Nutzer am ehesten klicken wird
  - Kreditkartenmissbrauch erkennen
  - Vorhersage welches Footballteam den Superbowl gewinnen wird

# Maschinelles Lernen

## Lernen und Trainieren

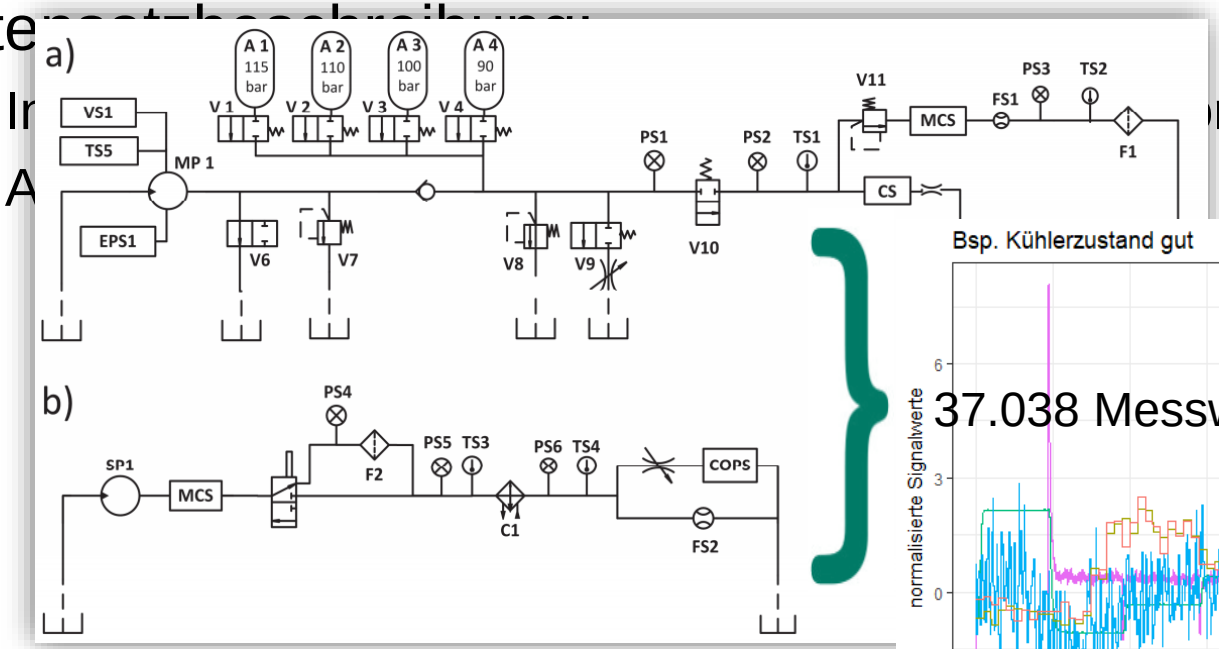
- Lernen und Trainieren am Beispiel der „Klassifikation“
  Klassifikation (Vorgriff): *Die Klassifizierung versucht für ein zur Grundgesamtheit gehörendes Individuum vorherzusagen zu welchen (einigen wenigen Klassen) dieses Individuum gehört.*
- *Def. Klassifikations-Modell:* Allgemeine Beschreibung der Regeln oder Zusammenhängen aus Trainingsdaten zur Klassifikation neuer Instanzen



# Maschinelles Lernen

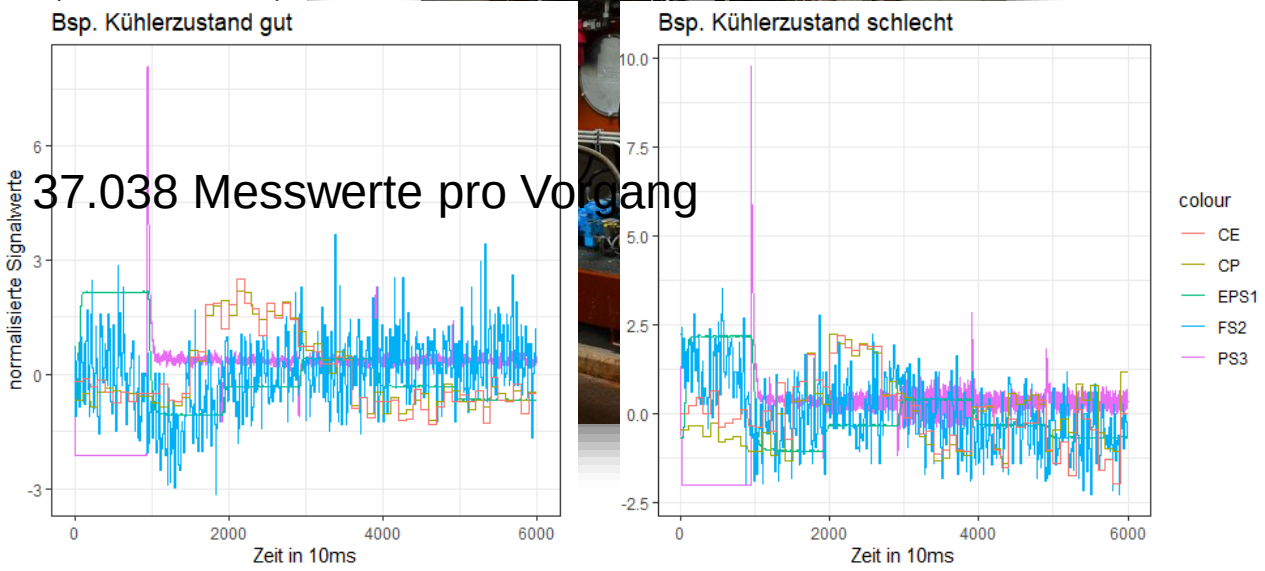
## Beispiel eines erlernten Modells

- Condition Monitoring of a Complex Hydraulic System (CMOCHS)
  - Aufgezeichnete Daten eines Hydraulikpumpen Systems (z.B. Temp., Druck, Leistung, etc.)
  - Fragestellung: Diagnose von 5 Fehlern (z.B. Kühlerzustand in gut, mittel, schlecht)
- Date[illegible]
  - In[illegible]rgän
  - A




Schematischer Aufbau des Hydraulik-Systems [1]
a) Arbeitskreislauf mit Hauptpumpe MP1
b) Kühl- und Filterkreislauf mit Kühler C1

37.038 Messwerte pro Vorgang

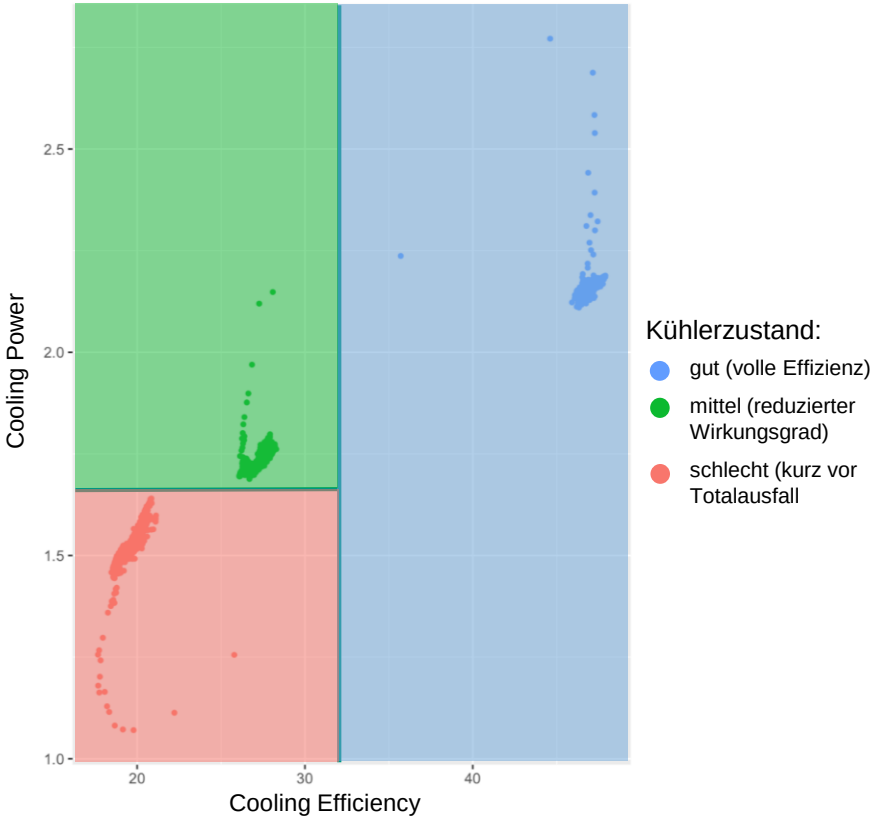



[1] Helwig, Nikolai, et al.. "Condition monitoring of a complex hydraulic system using multivariate statistics." *I2MTC*. IEEE, 2015

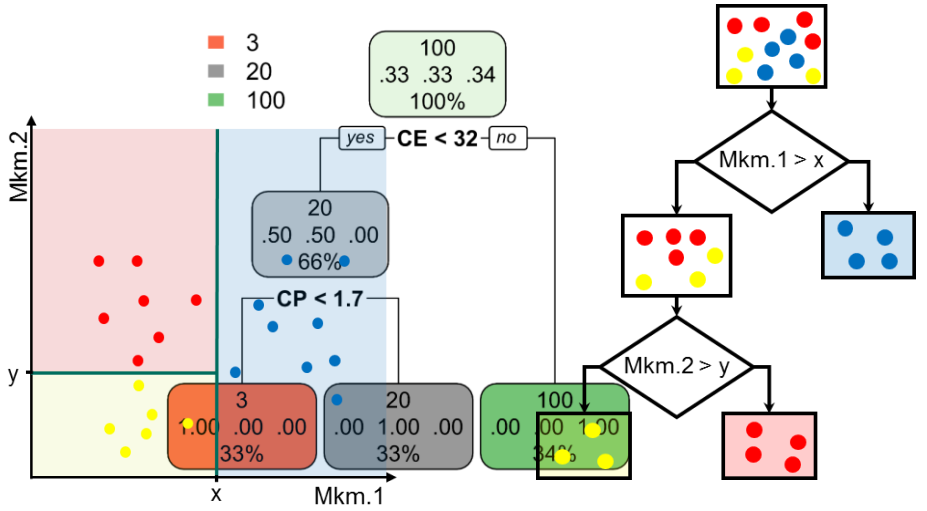
# Maschinelles Lernen

## Beispiel eines erlernten ML-Modells

- Condition Monitoring of a Complex Hydraulic System (CMOCHS)
  - Aufgezeichnete Daten eines Hydraulikpumpen Systems (z.B. Temp., Druck, Leistung, etc.)
  - Beispiele anhand Kühlerzustand eines Hydrauliksystems in 3 Klassen (gut, mittel, schlecht):




Erlerntes Modell:

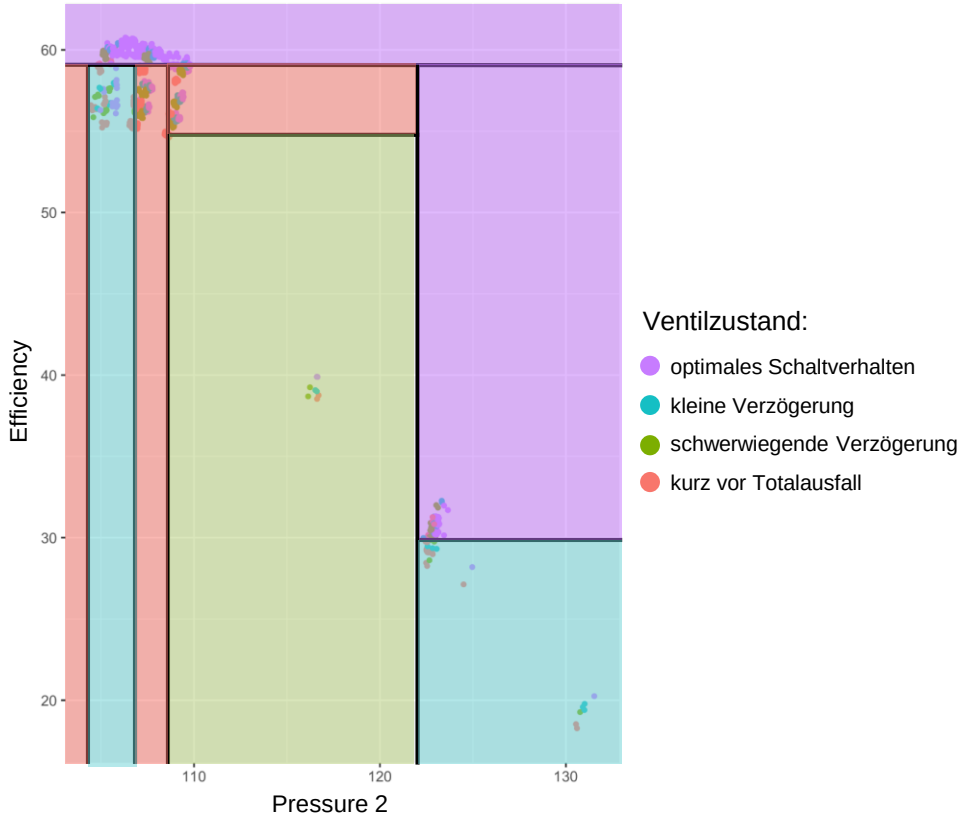



Erreichte Genauigkeit:
99,81%

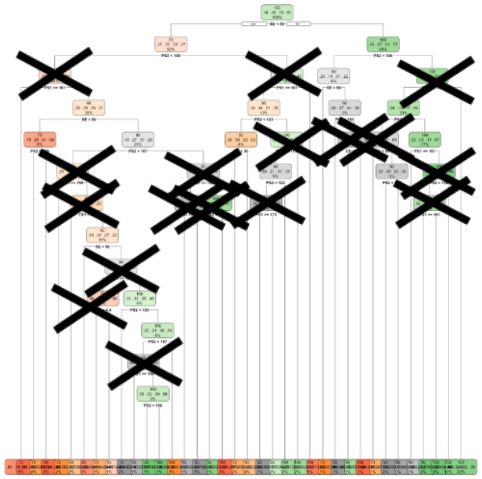
# Maschinelles Lernen

## Beispiel auf andere Fragestellung bei gleichen Daten

- Condition Monitoring of a Complex Hydraulic System (CMOCHS)
  - Aufgezeichnete Daten eines Hydraulikpumpen Systems (z.B. Temp., Druck, Leistung, etc.)
  - Beispiele anhand **Ventilzustand** eines Hydrauliksystems **in 4 Klassen**:



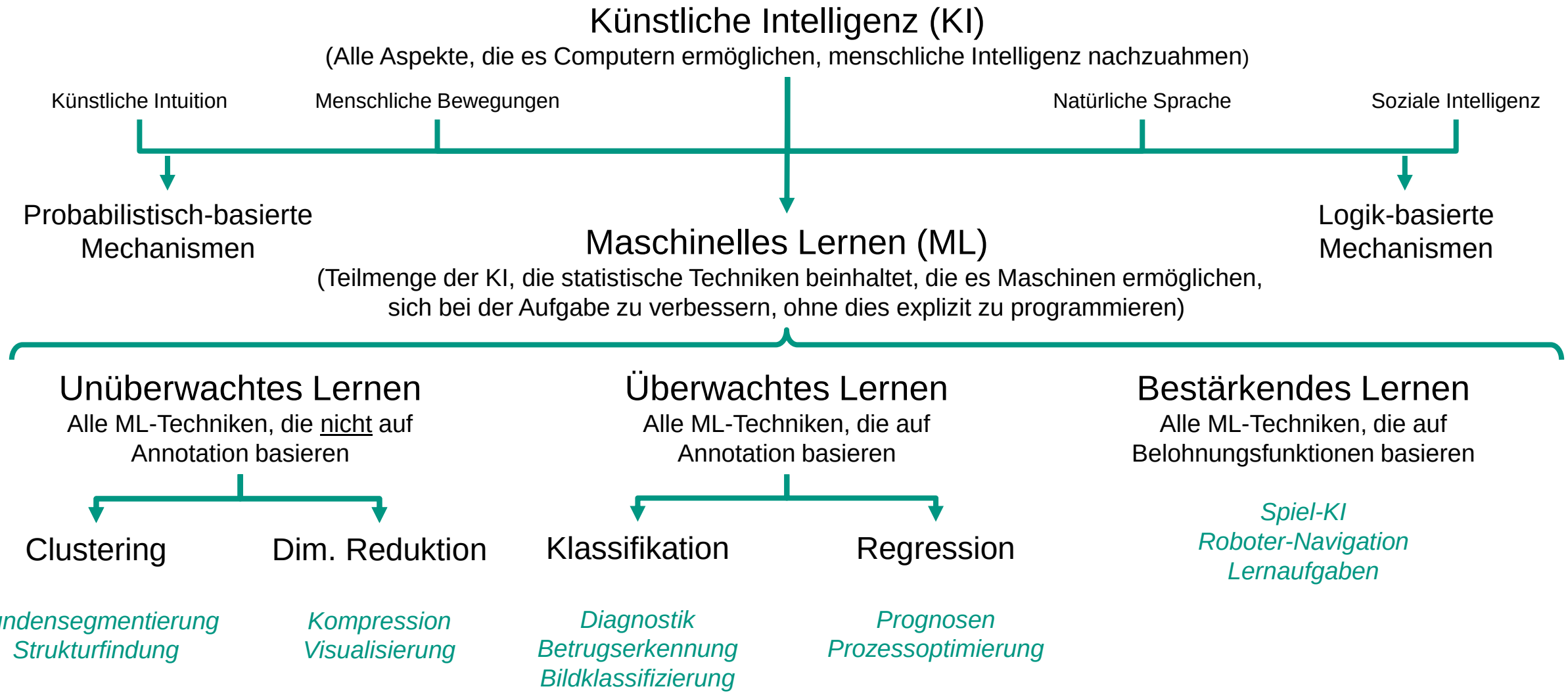

Erlerntes Modell:



Erreichte Genauigkeit:
83,19%

# Künstliche Intelligenz und Maschinelles Lernen

## Einordnung und Anwendungsbeispiele



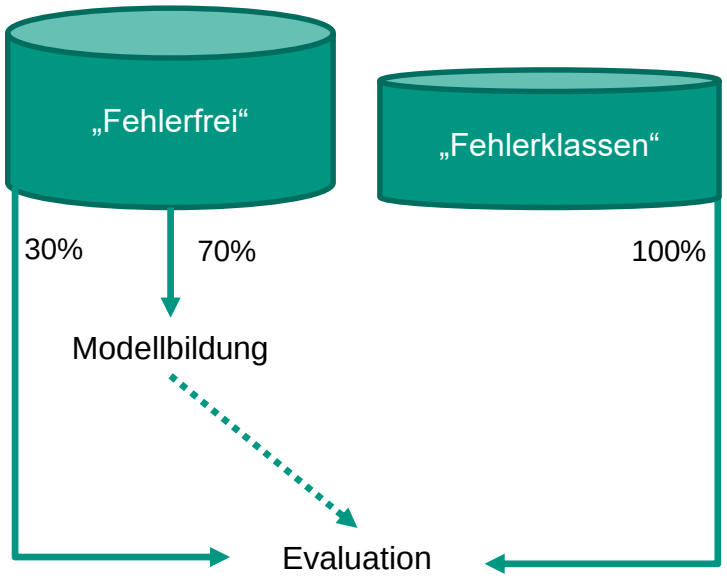
# Maschinelles Lernen

## Lernen und Trainieren anhand Klassifikation

- Daten sind vorhanden, aber wie trainiert man nun?

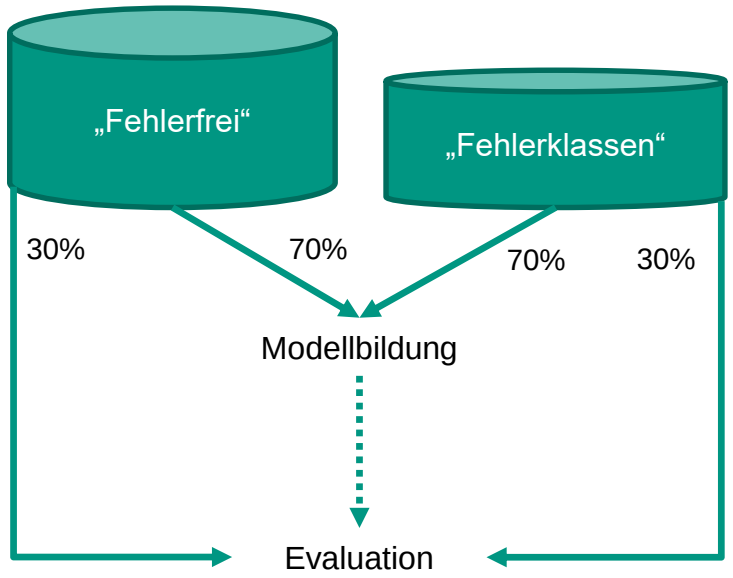
### Ein-Klassen Klassifikation

- Anomalie – Erkennung („Gut“ vs. „alles andere“)



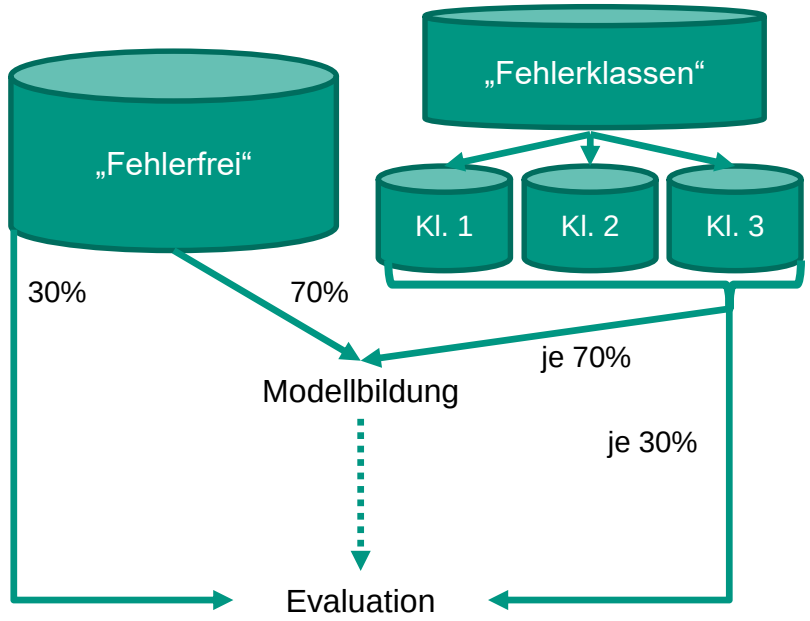

### Zwei-Klassen Klassifikation

- Erkennung „Gut“ vs. „alle bekannten Fehlerklassen“




### Multi-Klassen Klassifikation

- Erkennung „Gut“ vs. „Fehler-klasse 1“ vs. „Fehlerklasse 2“ vs. „Fehlerklasse 3“ etc.

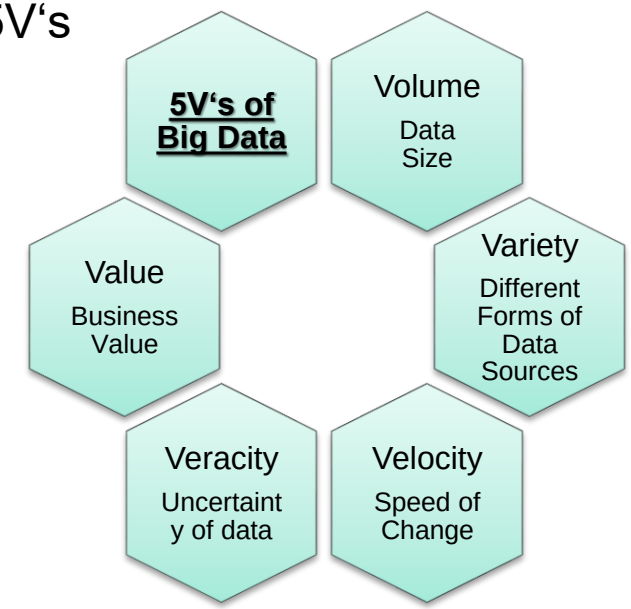


- Data Science
  - Artificial Intelligence
  - Data Mining
  - Maschinelles Lernen
- Overfitting/Underfitting

# Wiederholung

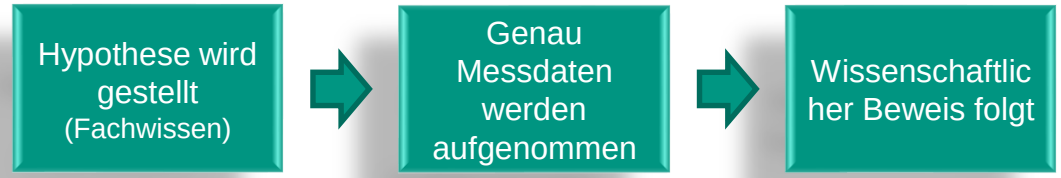
## Charakteristika zur Analyse großer Datenbestände → Big Data

- Die drei grundlegenden V‘s
  Die 5V‘s

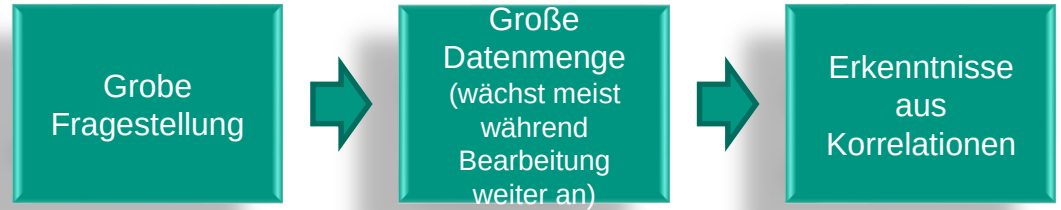



- **Volume:**
  Daten die aufgrund ihrer Menge bisher als kaum speicherbar, geschweige denn als auswertbar galten
- **Variety:**
  aus verschiedene Datenquellen strömen sortierte und unsortierte Datenmengen durch die Netze
- **Velocity:**
  Ergebnisse sollen möglichst schnell zur Verfügung stehen
- **Veracity:**
  Anspruch an hohe Datenqualität und Verlässlichkeit der Daten
- **Value:**
  Verwertbarkeit der gewonnen Erkenntnisse

- Hypothesengestützte Erkenntnisgewinnung
  Beruht auf kausalem Ansatz



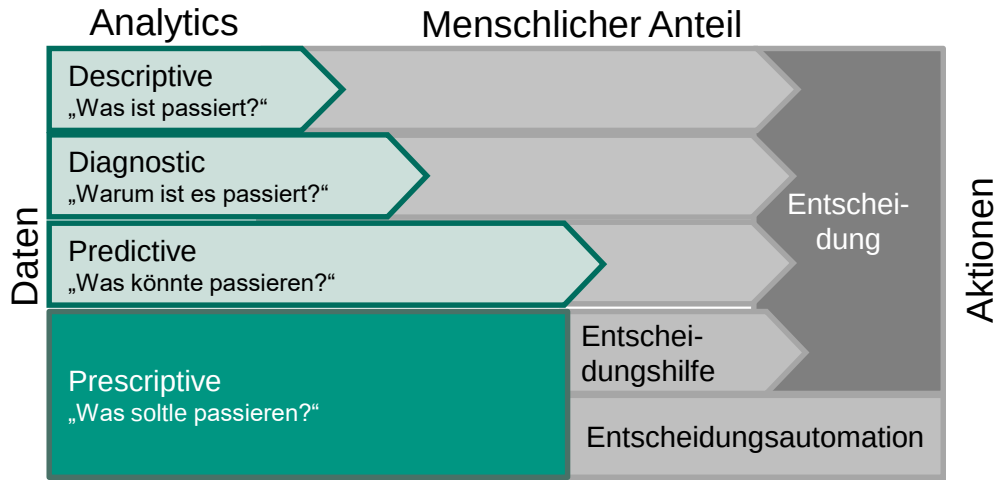

- Schlussfolgerung durch grobe Fragestellung
  Beruht auf Auffinden von Korrelationen



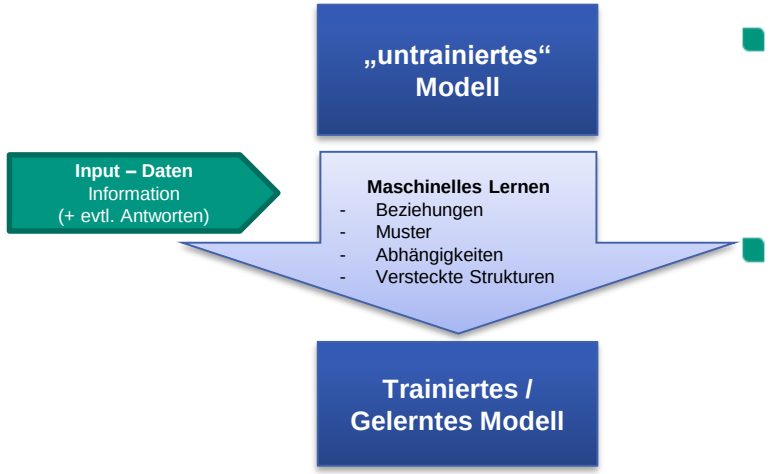
# Wiederholung

## Data Science, Data Mining, Maschinelles Lernen und Trainieren

### Zielsetzungen von Big Data




### Modelle im Bezug auf Maschinelles Lernen




- Erstellung und Verwendung von Modellen, welche aus Daten gelernt haben
- Modelle nutzen um Ergebnisse für neue Daten voraussagen zu können

### Unterschiedliche ML-Lernverfahren

**Maschinelles Lernen (ML)**

(Teilmenge der KI, die statistische Techniken beinhaltet, die es Maschinen ermöglichen, sich bei der Aufgabe zu verbessern, ohne dies explizit zu programmieren)

**Unüberwachtes Lernen**
Alle ML-Techniken, die nicht auf Annotation basieren

**Überwachtes Lernen**
Alle ML-Techniken, die auf Annotation basieren

**Bestärkendes Lernen**
Alle ML-Techniken, die auf Belohnungsfunktionen basieren

# Ziele der heutigen Übung

- Nach der heutigen Übung können Sie….

1. … bekannte Optimierungsalgorithmen gegenüberstellen und demonstrieren
2. … Charakteristika, Notwendigkeit und Vorgehensweisen zur Analyse großer Datenbestände beschreiben
3. … gängige Prozessabläufe zur Analyse von Big Data Problemstellungen beschreiben

# TEIL 3:
# BIG DATA ALS PROZESS

# Wir wollen eine Anomalie erkennen

## Zwischenübung

| Id | SepalLengthCm | SepalWidthCm | PetalLengthCm | PetalWidthCm | Species |
|---|---|---|---|---|---|
| 1 | 51 | 35 | 14 | 2 | Iris-setosa |
| 2 | 49 | 30 | 14 | 2 | Iris-setosa |
| 3 | 47 | 32 | 13 | 2 | Iris-setosa |
| 4 | 46 | 31 | 15 | 2 | Iris-setosa |
| 5 | 50 | 36 | 14 | 2 | Iris-setosa |
| 6 | 54 | 39 | 17 | 4 | Iris-setosa |
| 7 | 46 | 34 | 14 | 3 | Iris-setosa |
| 8 | 50 | 34 | 15 | 2 | Iris-setosa |
| 9 | 44 | 29 | 14 | 2 | Iris-setosa |
| 10 | 49 | 31 | 15 | 1 | Iris-setosa |
| 11 | 54 | 37 | 15 | 2 | Iris-setosa |
| 12 | 48 | 34 | 16 | 2 | Iris-setosa |
| 13 | 48 | 30 | 14 | 1 | Iris-setosa |
| 14 | 43 | 30 | 11 | 1 | Iris-setosa |
| 15 | 58 | 40 | 12 | 2 | Iris-setosa |
| 16 | 57 | 44 | 15 | 4 | Iris-setosa |
| 17 | 54 | 39 | 13 | 4 | Iris-setosa |
| 18 | 51 | 35 | 14 | 3 | Iris-setosa |
| 19 | 57 | 38 | 17 | 3 | Iris-setosa |
| 51 | 70 | 32 | 47 | 14 | Iris-versicolor |
| 52 | 64 | 32 | 45 | 15 | Iris-versicolor |
| 53 | 69 | 31 | 49 | 15 | Iris-versicolor |
| 54 | 55 | 23 | 40 | 13 | Iris-versicolor |
| 55 | 65 | 28 | 46 | 15 | Iris-versicolor |
| 56 | 57 | 28 | 45 | 13 | Iris-versicolor |
| 57 | 63 | 33 | 47 | 16 | Iris-versicolor |
| 58 | 49 | 24 | 33 | 10 | Iris-versicolor |
| 59 | 66 | 29 | 46 | 13 | Iris-versicolor |
| 60 | 52 | 27 | 39 | 14 | Iris-versicolor |
| 61 | 50 | 20 | 35 | 10 | Iris-versicolor |
| 62 | 59 | 30 | 42 | 15 | Iris-versicolor |
| 63 | 60 | 22 | 40 | 10 | Iris-versicolor |
| 64 | 61 | 29 | 47 | 14 | Iris-versicolor |
| 65 | 56 | 29 | 36 | 13 | Iris-versicolor |
| 66 | 67 | 31 | 44 | 14 | Iris-versicolor |
| 67 | 56 | 30 | 45 | 15 | Iris-versicolor |
| 68 | 58 | 27 | 41 | 10 | Iris-versicolor |
| 69 | 62 | 22 | 45 | 15 | Iris-versicolor |
| 70 | 56 | 25 | 39 | 11 | Iris-versicolor |
| 71 | 70 | 32 | 47 | 14 | Iris-versicolor |

- hierarchical cluster analysis
- DBSCAN
- One Class Support Vector Machine
- Isolation Forest
- (künstliche) Neuronale Netze
- Support Vector Machine
- Decision Tree
- Bayes-Klassifikation
- Random Forest
- Multivariate Adaptive Regressions-Splines
- Logistic Regression
- Linear Regression
- Harmonic Regression
- Diskriminanzanalyse
- Nächste-Nachbar-Klassifikation

- Regression
- Clustering
- Klassifikation

Lasst uns eine Anomalie erkennen!
Welchen Algorithmus würdet ihr verwenden?

# Wir wollen eine Anomalie erkennen

## Zwischenübung – „Lsg“ oder eben nicht…

| Id | SepalLengthCm | SepalWidthCm | PetalLengthCm | PetalWidthCm | Species |
|---|---|---|---|---|---|
| 1 | 51 | 35 | 14 | 2 | Iris-setosa |
| 2 | 49 | 30 | 14 | 2 | Iris-setosa |
| 3 | 47 | 32 | 13 | 2 | Iris-setosa |
| 4 | 46 | 31 | 15 | 2 | Iris-setosa |
| 5 | 50 | 36 | 14 | 2 | Iris-setosa |
| 6 | 54 | 39 | 17 | 4 | Iris-setosa |
| 7 | 46 | 34 | 14 | 3 | Iris-setosa |
| 8 | 50 | 34 | 15 | 2 | Iris-setosa |
| 9 | 44 | 29 | 14 | 2 | Iris-setosa |
| 10 | 49 | 31 | 15 | 1 | Iris-setosa |
| 11 | 54 | 37 | 15 | 2 | Iris-setosa |
| 12 | 48 | 34 | 16 | 2 | Iris-setosa |
| 13 | 48 | 30 | 14 | 1 | Iris-setosa |
| 14 | 43 | 30 | 11 | 1 | Iris-setosa |
| 15 | 58 | 40 | 12 | 2 | Iris-setosa |
| 16 | 57 | 44 | 15 | 4 | Iris-setosa |
| 17 | 54 | 39 | 13 | 4 | Iris-setosa |
| 18 | 51 | 35 | 14 | 3 | Iris-setosa |
| 19 | 57 | 38 | 17 | 3 | Iris-setosa |
| 51 | 70 | 32 | 47 | 14 | Iris-versicolor |
| 52 | 64 | 32 | 45 | 15 | Iris-versicolor |
| 53 | 69 | 31 | 49 | 15 | Iris-versicolor |
| 54 | 55 | 23 | 40 | 13 | Iris-versicolor |
| 55 | 65 | 28 | 46 | 15 | Iris-versicolor |
| 56 | 57 | 28 | 45 | 13 | Iris-versicolor |
| 57 | 63 | 33 | 47 | 16 | Iris-versicolor |
| 58 | 49 | 24 | 33 | 10 | Iris-versicolor |
| 59 | 66 | 29 | 46 | 13 | Iris-versicolor |
| 60 | 52 | 27 | 39 | 14 | Iris-versicolor |
| 61 | 50 | 20 | 35 | 10 | Iris-versicolor |
| 62 | 59 | 30 | 42 | 15 | Iris-versicolor |
| 63 | 60 | 22 | 40 | 10 | Iris-versicolor |
| 64 | 61 | 29 | 47 | 14 | Iris-versicolor |
| 65 | 56 | 29 | 36 | 13 | Iris-versicolor |
| 66 | 67 | 31 | 44 | 14 | Iris-versicolor |
| 67 | 56 | 30 | 45 | 15 | Iris-versicolor |
| 68 | 58 | 27 | 41 | 10 | Iris-versicolor |
| 69 | 62 | 22 | 45 | 15 | Iris-versicolor |
| 70 | 56 | 25 | 39 | 11 | Iris-versicolor |
| 71 | 70 | 32 | 47 | 14 | Iris-versicolor |

| |
|---|
| hierarchical cluster analysis |
| DBSCAN |
| One Class Support Vector Machine |
| Isolation Forest |
| |
| (künstliche) Neuronale Netze |
| Support Vector Machine |
| Decision Tree |
| Bayes-Klassifikation |
| Random Forest |
| Multivariate Adaptive Regressions-Splines |
| Logistic Regression |
| Linear Regression |
| Harmonic Regression |
| Diskriminanzanalyse |
| Nächste-Nachbar-Klassifikation |

Regression

Clustering

Klassifikation

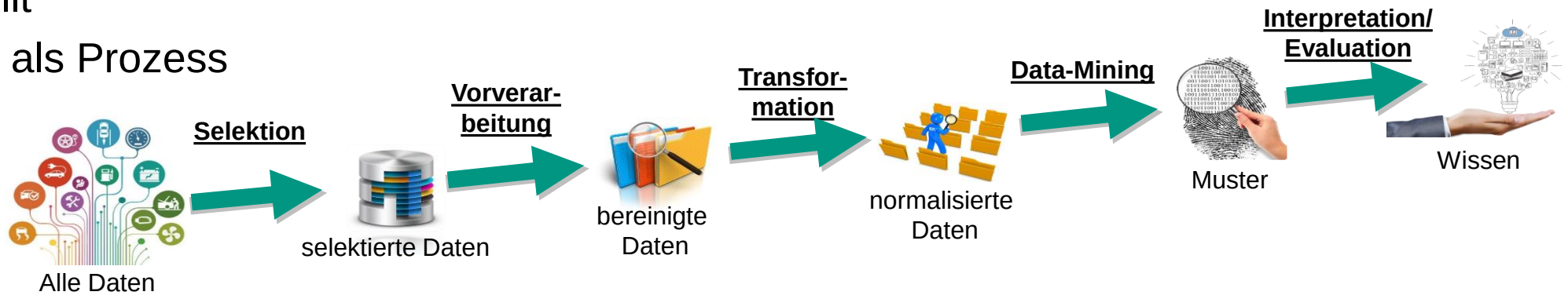
Lasst uns eine Anomalie erkennen!
Welchen Algorithmus würdet ihr verwenden?

# Big Data als Prozess

## Strukturierte Vorgehensweise

- Simpel einen beliebigen Maschinellen Lernalgorithmus auf vorliegende Daten anzuwenden ist nicht zielführend!
- ➢ Strukturelle Vorgehensweise wird benötigt!
- Data Mining erfordert die Anwendung von beträchtlichen wissenschaftlichen und technischen Kenntnissen
- Es wird ein wohlverstandenes Verfahren für die Strukturierung benötigt, welches
  - Einheitlichkeit
  - Reproduzierbarkeit
  - Objektivität

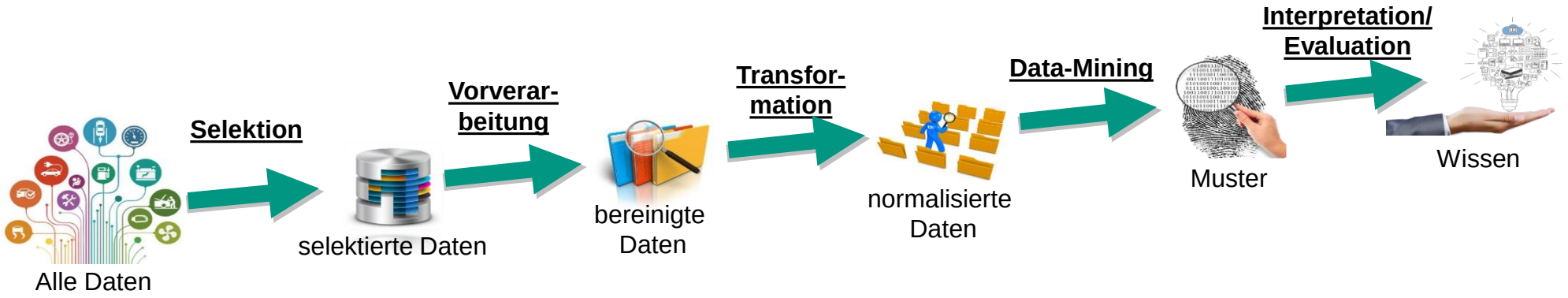
  sicherstellt
- ➢ Big Data als Prozess



# Big Data als Prozess

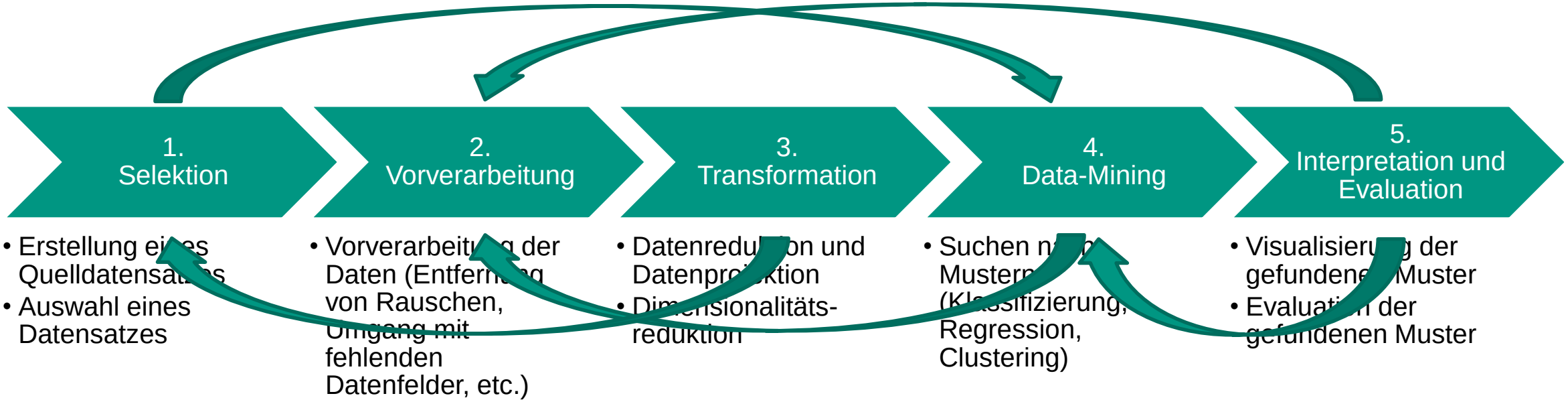
## Knowledge Discovery in Databases (KDD)

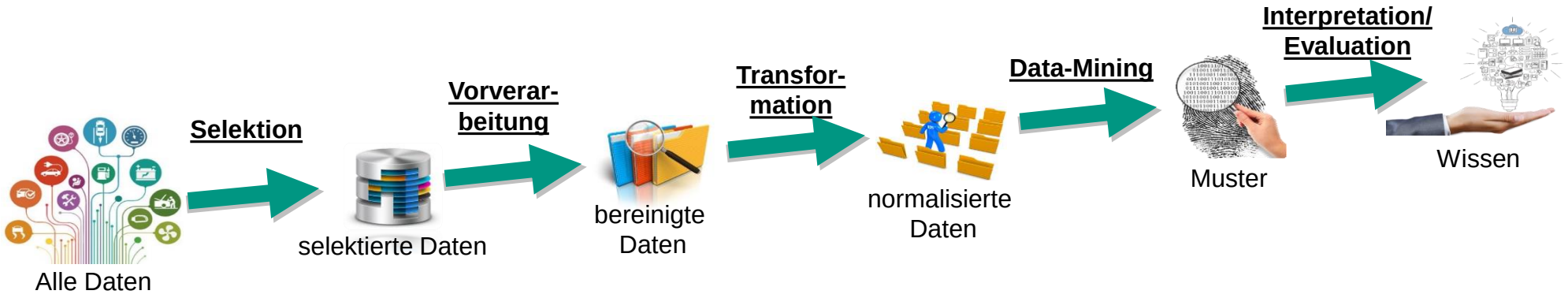
| 1. Selektion | 2. Vorverarbeitung | 3. Transformation | 4. Data-Mining | 5. Interpretation und Evaluation |
|---|---|---|---|---|
| • Erstellung eines Quelldatensatzes<br>• Auswahl eines Datensatzes | • Vorverarbeitung der Daten (Entfernung von Rauschen, Umgang mit fehlenden Datenfelder, etc.) | • Datenreduktion und Datenprojektion<br>• Dimensionalitäts-reduktion | • Suchen nach Mustern (Klassifizierung, Regression, Clustering) | • Visualisierung der gefundenen Muster<br>• Evaluation der gefundenen Muster |



# Big Data als Prozess

## Knowledge Discovery in Databases (KDD) als iterativer und zyklischer Prozess



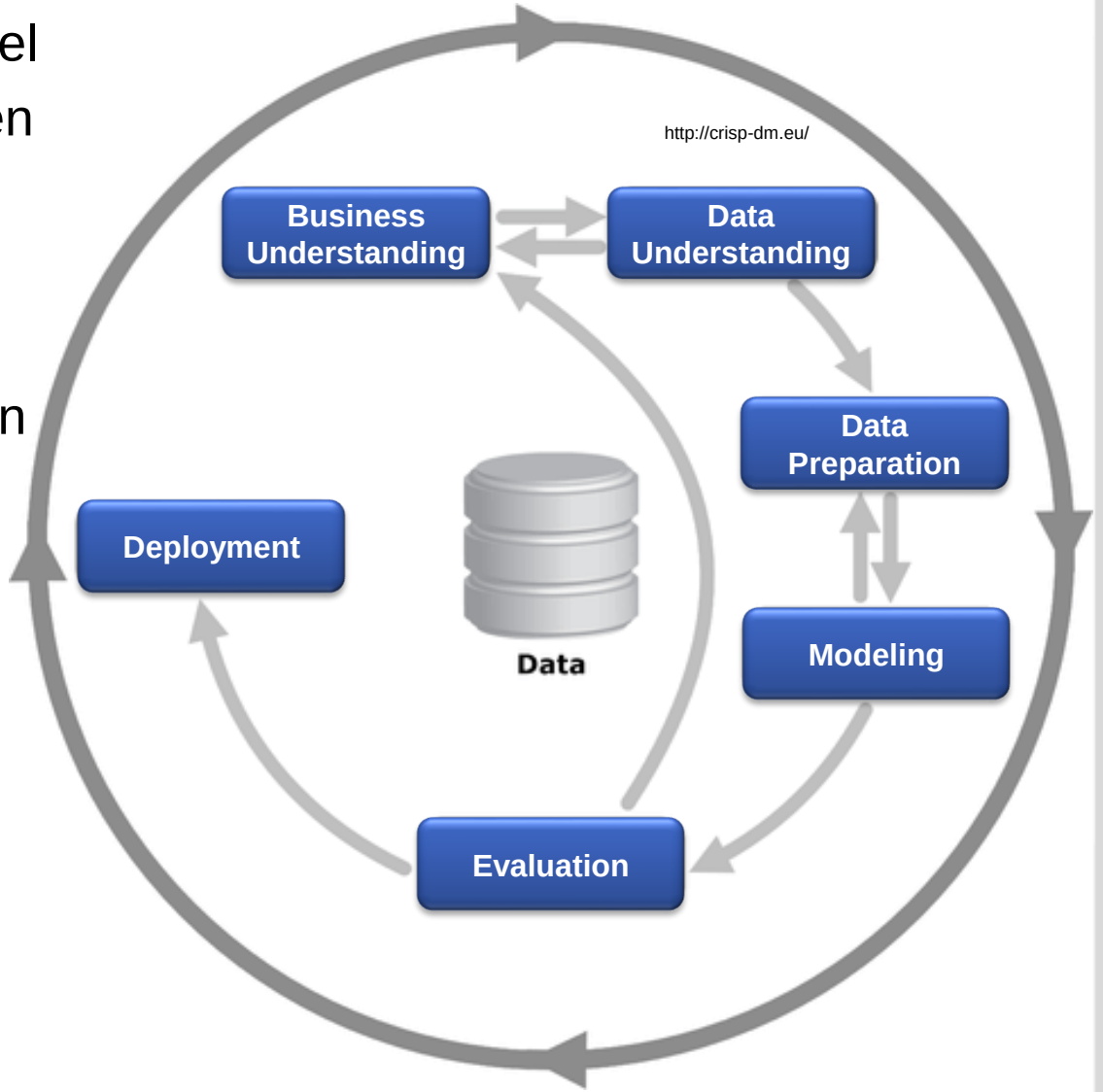
# Big Data als Prozess

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

- Iterationen sind keine Ausnahmen, sondern die Regel
- erster Durchlauf dient meist der Erkundung der Daten

- Zyklischer Charakter ist im CRISP-DM enthalten

- Phasen nehmen unterschiedlichen Arbeitsaufwand in Anspruch
  - 20-30% Data Understanding
  - 50-70% Data Preparation
  - 10-20% Modeling and Evaluation
  - 5-10% Deployment

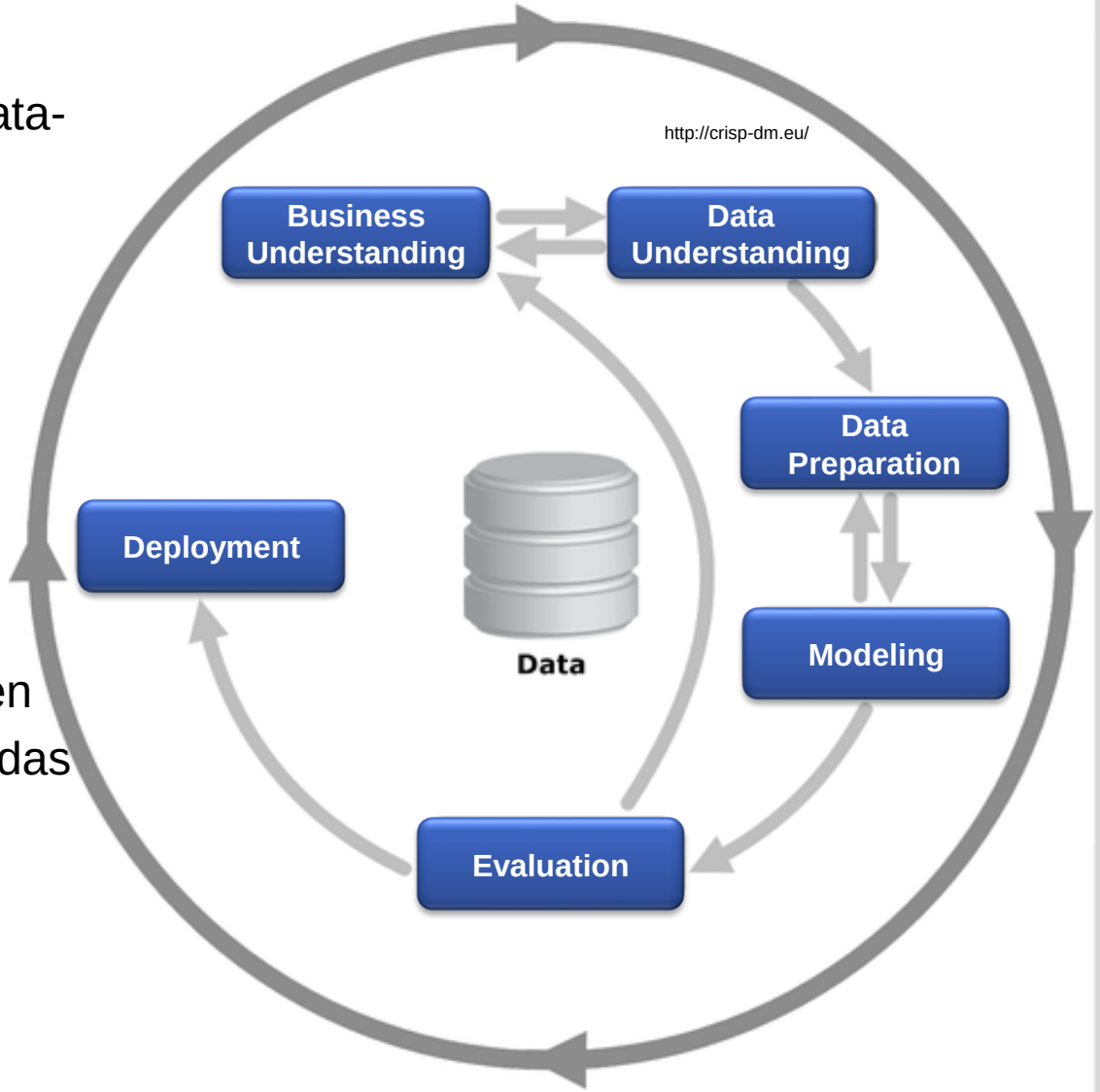


# Cross Industry Standard Process for Data Mining (CRISP-DM)

## Business- and Data Understanding

- Aufgabenverständnis: „Formulierung der Aufgabe“
  - Aufteilung des Aufgabenstellung auf verschiedene Data-Science-Aufgaben
  - Kenntnisse von Data Mining hilfreich
  - Erkenntnisse von denkbaren Anwendungsfeldern

- Datenverständnis
  - Daten sind selten genau auf die Aufgabenstellung zugeschnitten (Datensammlung meist generell oder ursprünglich zu anderem Zweck)
  - Kosten-Nutzen bei Beschaffung neuer Daten beachten
  - Durch Zunahme des Datenverständnisses kann sich das Aufgabenverständnis wiederum ändern

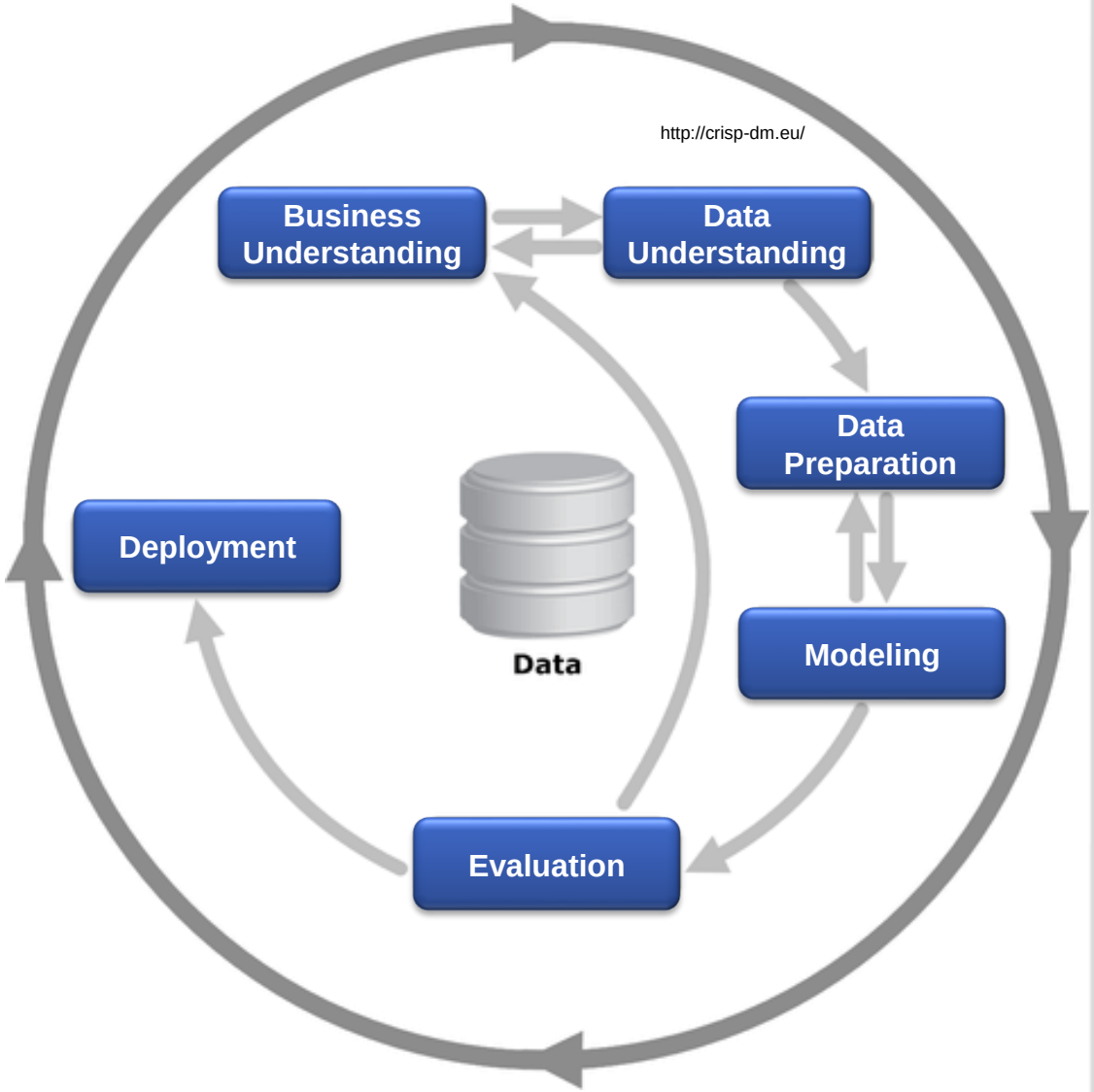


# Cross Industry Standard Process for Data Mining (CRISP-DM)

## Data Preparation

- Datenaufbereitung
  - Daten müssen bestimmte Voraussetzungen erfüllen
    - Konvertierung der Daten
    - „Glätten“ der Daten
    - Außreißerdetektion
    - Normalisierung, Skalierung, etc.

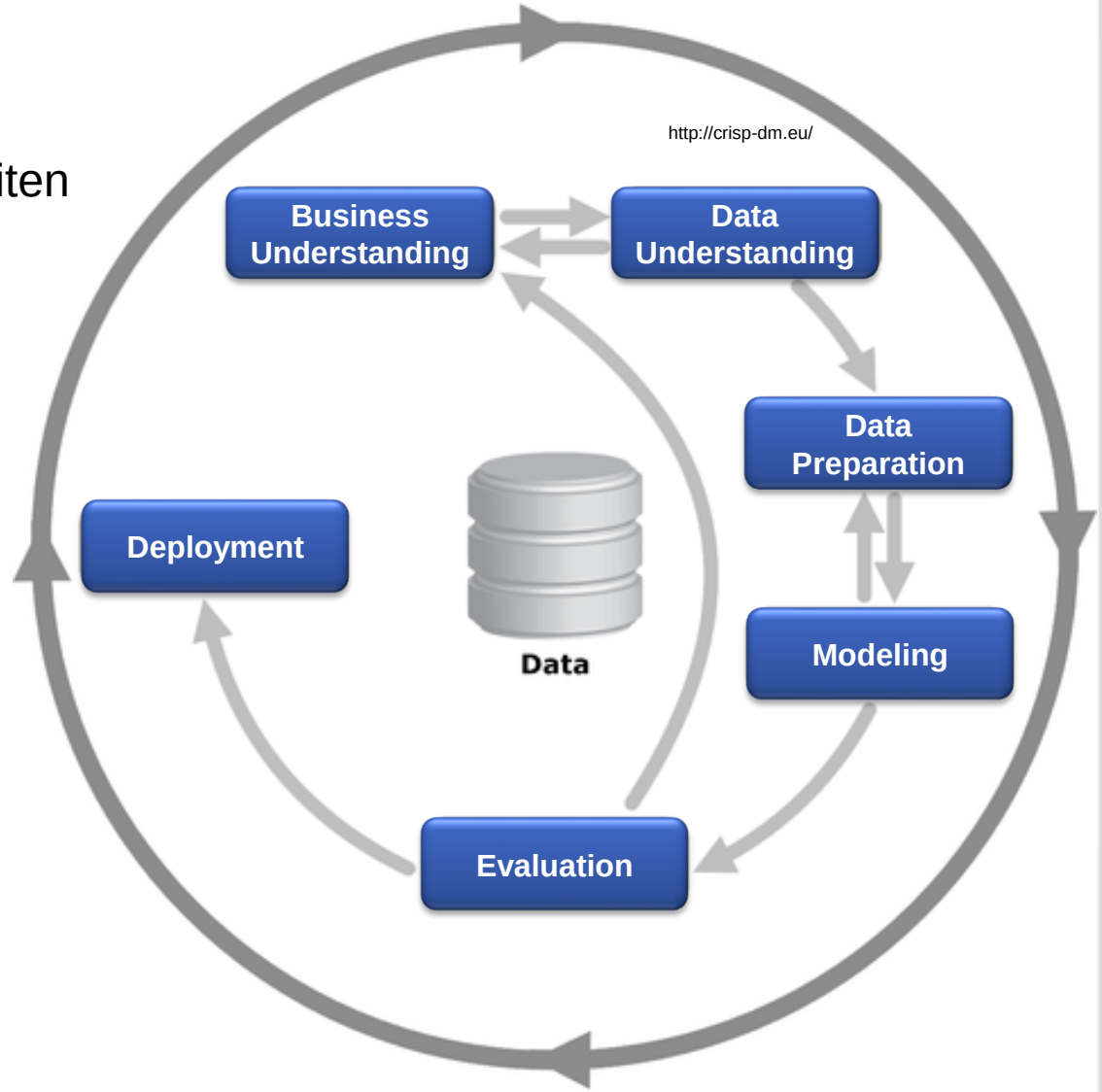
*„Im Allgemeinen müssen Data Scientists anfangs beträchtliche Zeit dafür aufwenden, die Variablen zu definieren, die später verwendet werden. Gerade hier kommen die menschliche Kreativität, der gesunde Menschenverstand und das Fachwissen ins Spiel. Die Qualität einer Data-Mining-Lösung beruht oft darauf, wie gut die Analysten die Aufgabenstellung strukturieren und die Variablen gestalten […]“* ~ *Tom Fawcett*



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

## Modeling

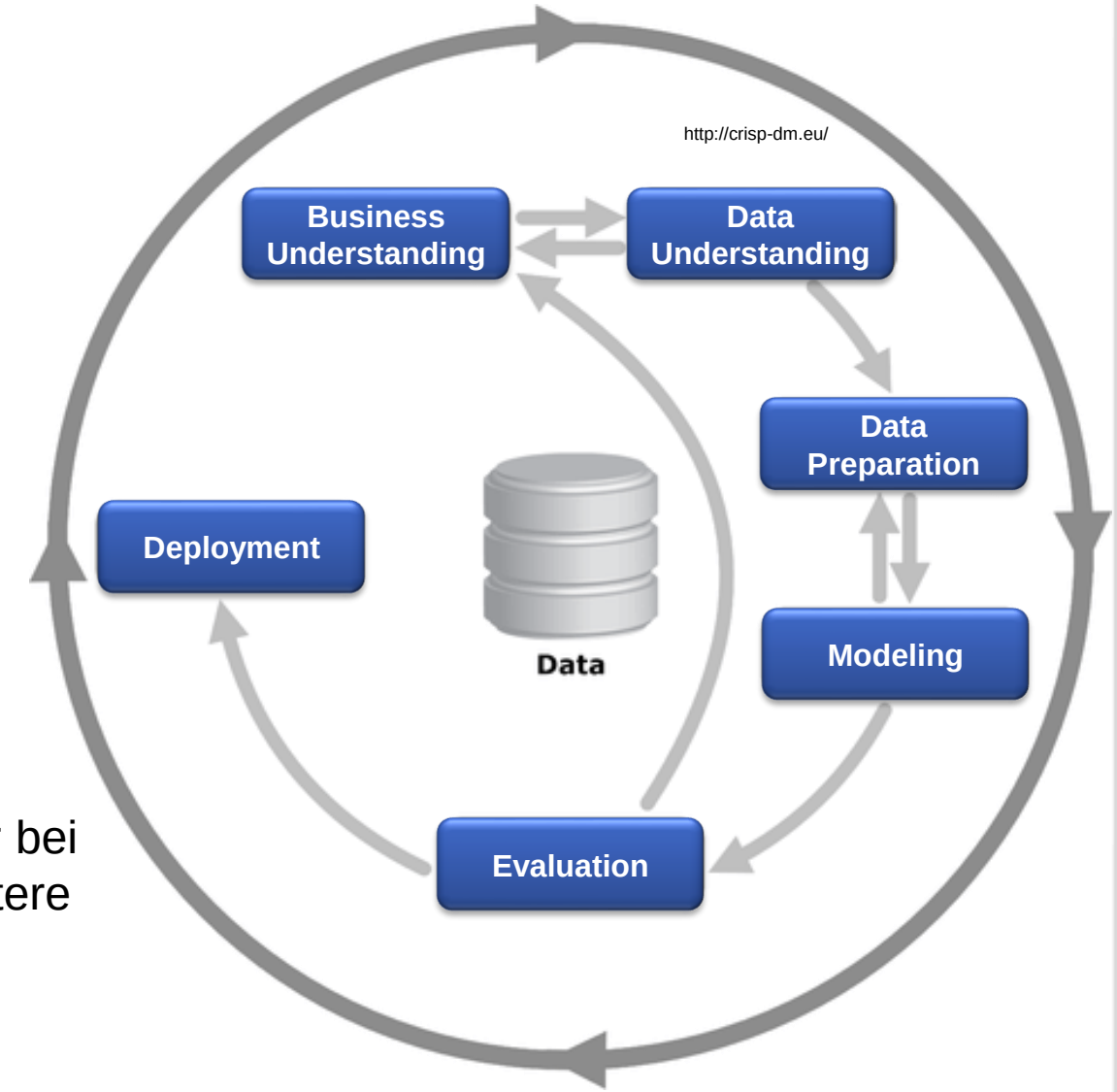
- Modellbildung
  - Anwendung von ML-Algorithmen
  - Suche nach Modellen, Mustern oder Gesetzmäßigkeiten in den vorliegenden Daten



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

## Evaluation und Deployment

- Beurteilung
  - Bewertung der Ergebnisse der Modellbildung (Zuverlässigkeit, Gültigkeit, etc.)
  - Erfüllung der Aufgabenstellung prüfen
  - Testen der Praxistauglichkeit
- Einsatz
  - Anwenden der Ergebnisse der Modellbildung
  - Implementierung eines Vorhersagemodells für Informationssysteme/Geschäftsvorgänge




- Häufig wird nach Durlaufens des Deployment nun wieder bei Phase des Aufgabenverständnisses begonnen. Eine weitere Iteration kann eine verbesserte Lösung hervorbringen (Zykluseigenschaft des CRISP-DM)

# CRISP-DM und kdd

## Was sagt „die Industrie“ dazu?

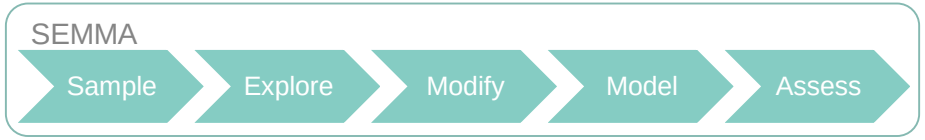
- **CRISP-DM** ist eine Beschreibung des „Workflows“ in Data-Mining-Projekten
  *„the first step towards defining a data science methodology“*
  *~ Saltz J.*
- **Kdd**-Ansätze konzentrieren sich auf die Schritte der Durchführung von Data Mining als auf die Beschreibung eines umfassenden Projektmanagement-Konzepts

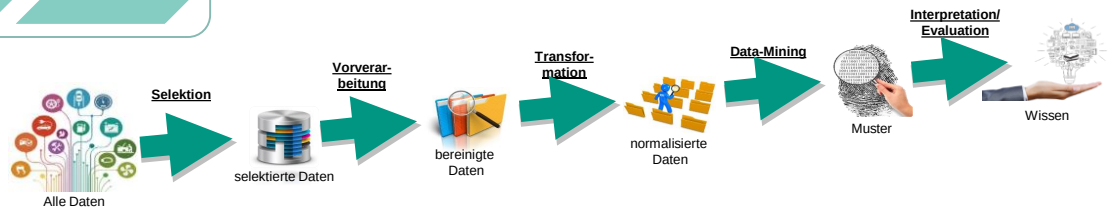
http://www.datascience-pm.com

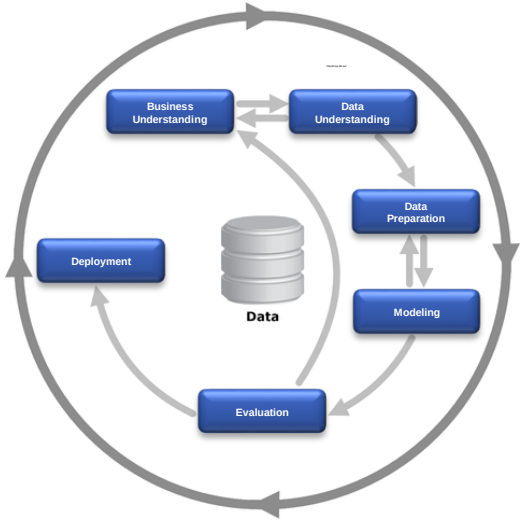
What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]

| | 2014 poll | 2007 poll |
|---|---|---|
| CRISP-DM (86) | 43% | 42% |
| My own (55) | 27.5% | 19% |
| SEMMA (17) | 8.5% | 13% |
| Other, not domain-specific (16) | 8% | 4% |
| KDD Process (15) | 7.5% | 7.3% |
| My organizations' (7) | 3.5% | 5.3% |
| A domain-specific methodology (4) | 2% | 4.7% |
| None (0) | 0% | 4.7% |

- CRISP-DM ist der am häufigsten in der Industrie genutzte Standard



- Big Data als Prozess
  - KDD und CRISP-DM
    - Business Understanding
    - Data Understanding
    - Data Preparation
    - Modeling
    - Evaluation
    - Deployment

# Ziele der heutigen Übung

- Nach der heutigen Übung können Sie….

1. … bekannte Optimierungsalgorithmen gegenüberstellen und demonstrieren
2. … Charakteristika, Notwendigkeit und Vorgehensweisen zur Analyse großer Datenbestände beschreiben
3. … gängige Prozessabläufe zur Analyse von Big Data Problemstellungen beschreiben