Karlsruher Institut für Technologie

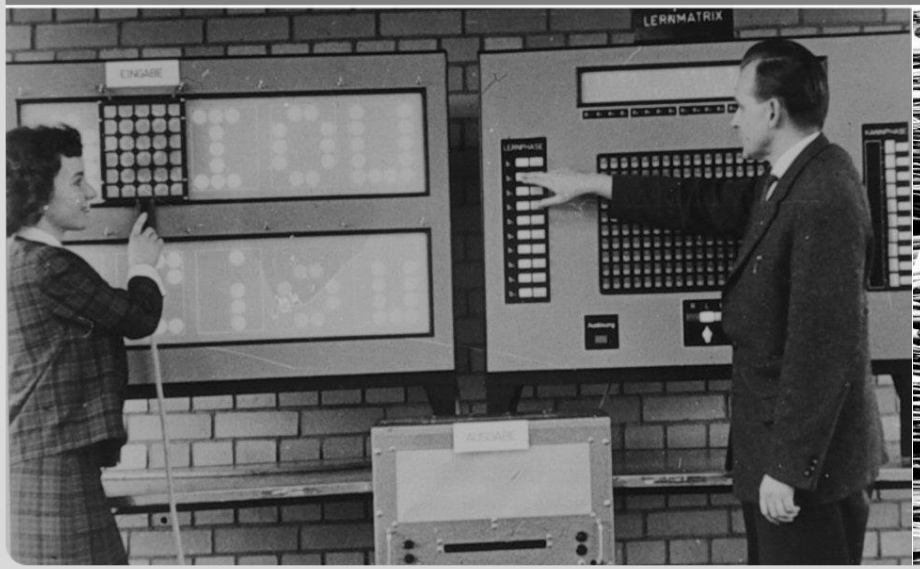
Institut für Technik der Informationsverarbeitung

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 – Nathalie Brenner

Prof. Dr.-Ing. Eric Sax



WIEDERHOLUNG ÜBUNG 2

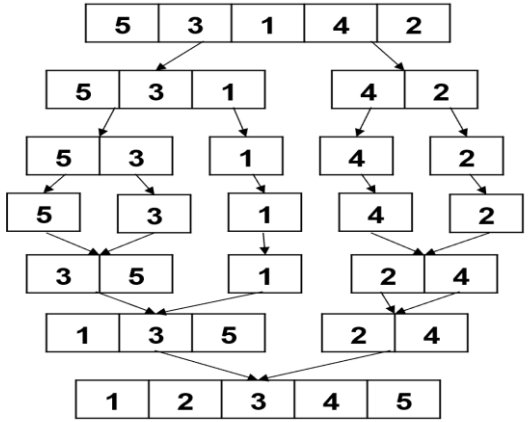
Wiederholung Übung 2

Sortieralgorithmen

Bubble Sort

j=1	j=2	j=3	j=4	j=5
5	3	1	4	2
3	5	1	4	2
…				
1	2	3	4	5

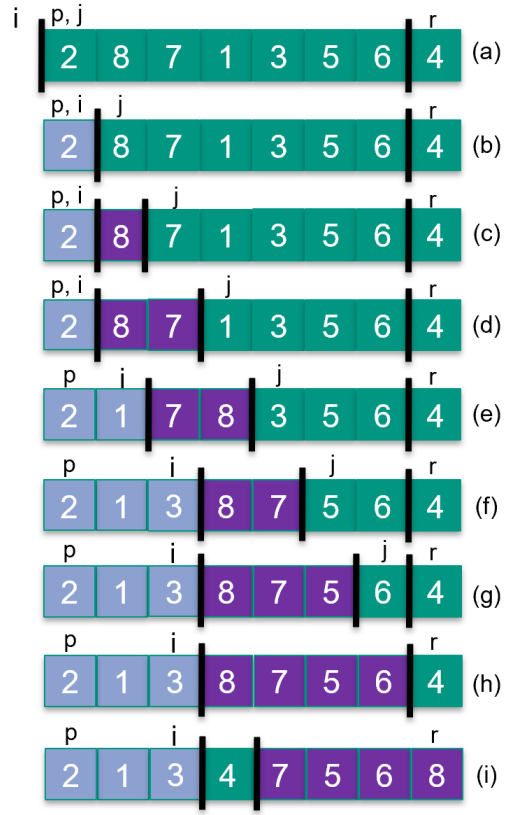
Merge Sort



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
```

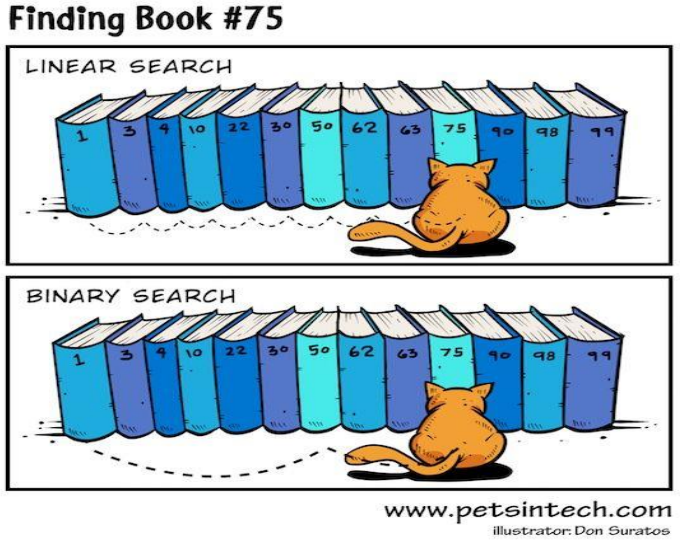
Quick Sort



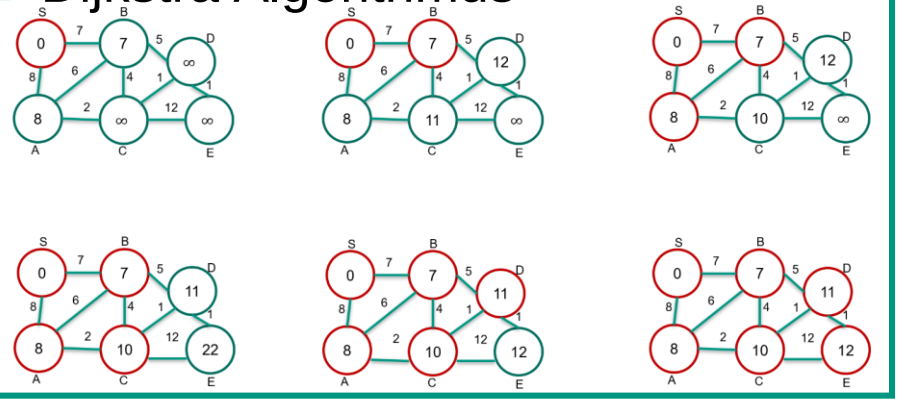
Wiederholung Übung 2

Suchalgorithmen und Optimierungsalgorithmen

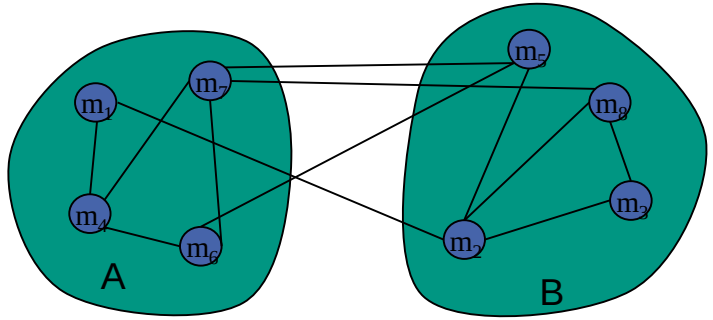
Lineare Suche und Binäre Suche



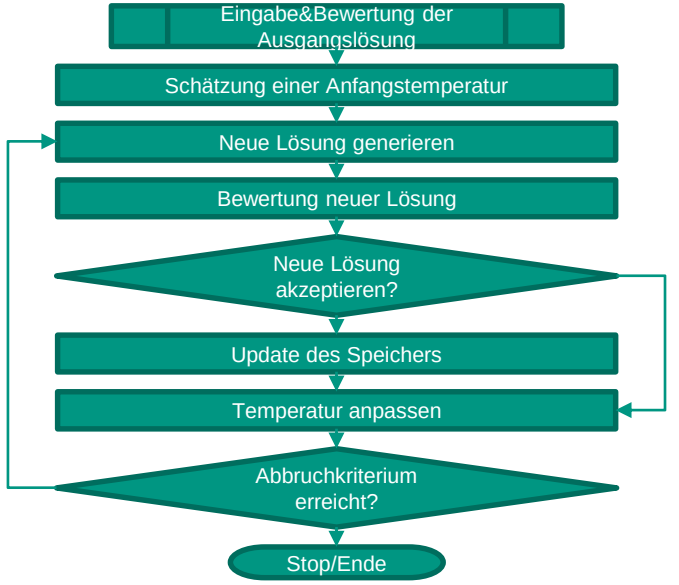
Dijkstra Algorithmus



Partitionierung und Kostenreduktion



Simulated Annealing



INHALT ÜBUNG 3

Big Data, Data Mining und Prozesse

Schlagwort, Sammelbegriff oder Synonym? Und wozu das Ganze?

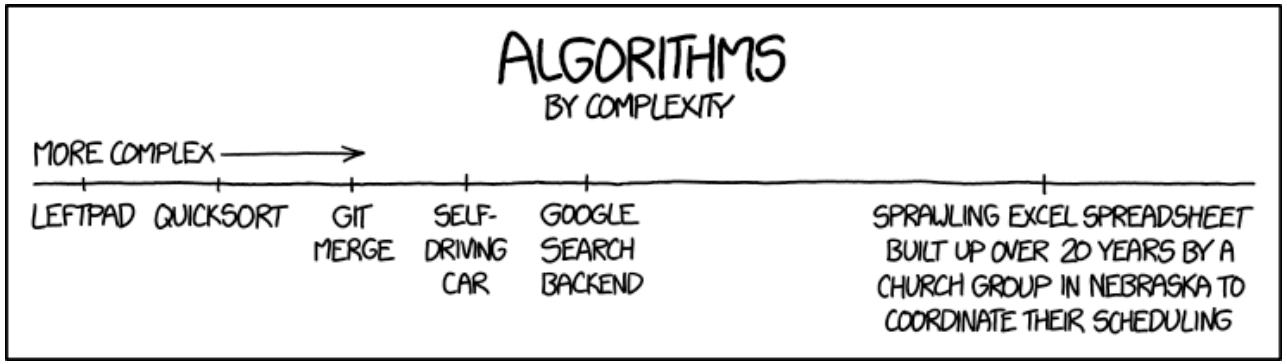
- Big Data steht für große Menge an digitalen Daten, sowie deren Erfassung, Analyse und Auswertung

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….

- Charakteristika, Notwendigkeit und Vorgehensweisen zur Analyse großer Datenbestände beschreiben

1. … Notwendigkeit zur Analyse großer Datenbestände nennen
2. … Charakteristika zur Abgrenzung von Big Data nennen
3. … Grundbegriffe im Bezug zu Big Data nennen
4. … den Begriff des „Maschinelles Lernens“ abgrenzen
5. … den Begriff und das Vorgehen beim „Trainieren“ abgrenzen

CHARAKTERISTIKA ZUR ANALYSE GROSSER DATENMENGEN
TEIL 1 – DIE 5V‘S

„Big Data“ als Überbegriff

Anwendungen

- Wo wird „Big Data“ angewendet?
- Wann wird „Big Data“ angewendet?
- Wozu wird „Big Data“ angewendet?

Denken Sie kurz über die Fragen nach und diskutieren Sie im Anschluss mit Ihren Nachbarn über Ihre „Ergebnisse/Entschlüsse“

Big Data - unerwartete Anwendungen?

Wilderern zuvor kommen

- Studie zur Schätzung der Wahrscheinlichkeit des Vorkommens von Wilderei
- Analyse von 25.000 Datenpunkten: seit 1972 in 605 Distrikten gesammelt
- Große Datenmenge, jeder Analysezyklus nimmt 20-25 Minuten Zeit in Anspruch

- Verbrechensmuster erkennen
- An „Hot Spots“ höhere Präsenz/Patrouillen aufweisen
- Dauerhafte Aktualisierung, da Wilderer Taktik ändern
  → andauernd neue, große Datenmengen zur Analyse

Big Data - unerwartete Anwendungen?

Risikomanagement - Insiderfahndung

- Bundesanstalt für Finanzdienstleistungsaufsicht (BAFin) überwacht Märkte für Wertpapiere und Derivate
- Ad hoc-Meldungen können Aktienkurs stark beeinflussen, bei früherem Erhalt können „Insider“ kursrelevante Informationen ausnutzen
  → Insidergeschäft (Aufzudecken im Sinne des Anlegerschutzes)

- 2000 ad hoc-Meldungen pro Jahr
- 5.6 Millionen Wertpapiertransaktionen täglich im Rechenzentrum der BAFin
- Alle Transaktionen in relationalen Datenbanken (riesige Datenmengen in kürzester Zeit)

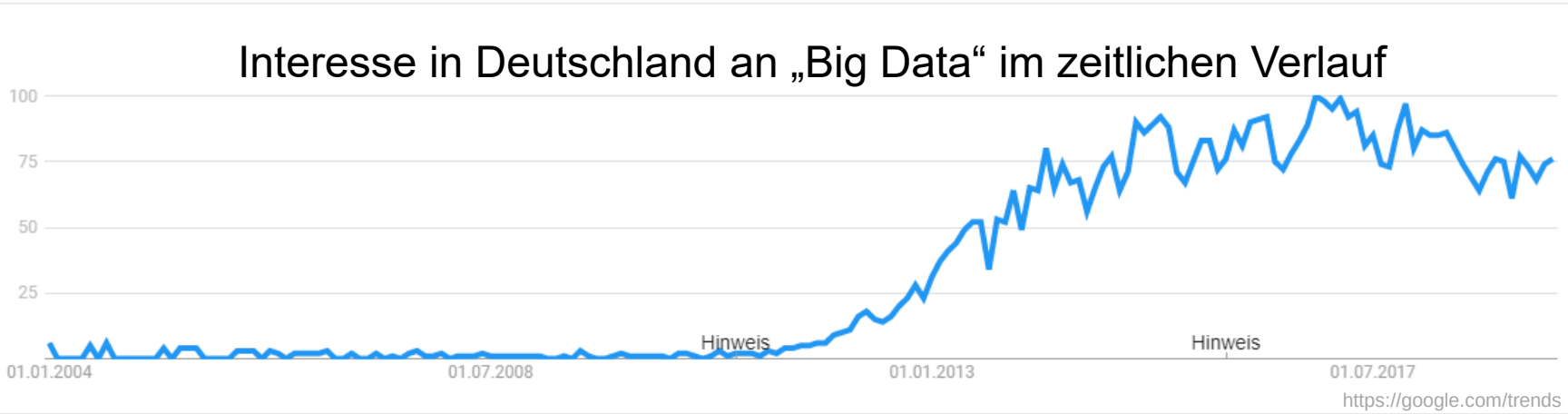
Auffällige Transaktionen möglichst schnell aufdecken

Big Data - Charakteristika

Schlagwort, Sammelbegriff, Synonym

- Erstmalige Publikation des Begriffes „Big Data“ im Jahre 1997
  „Application-Controlled Demand Paging for Out-of-Visualization“ by Michael Cox and David Ellsworth
- 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*

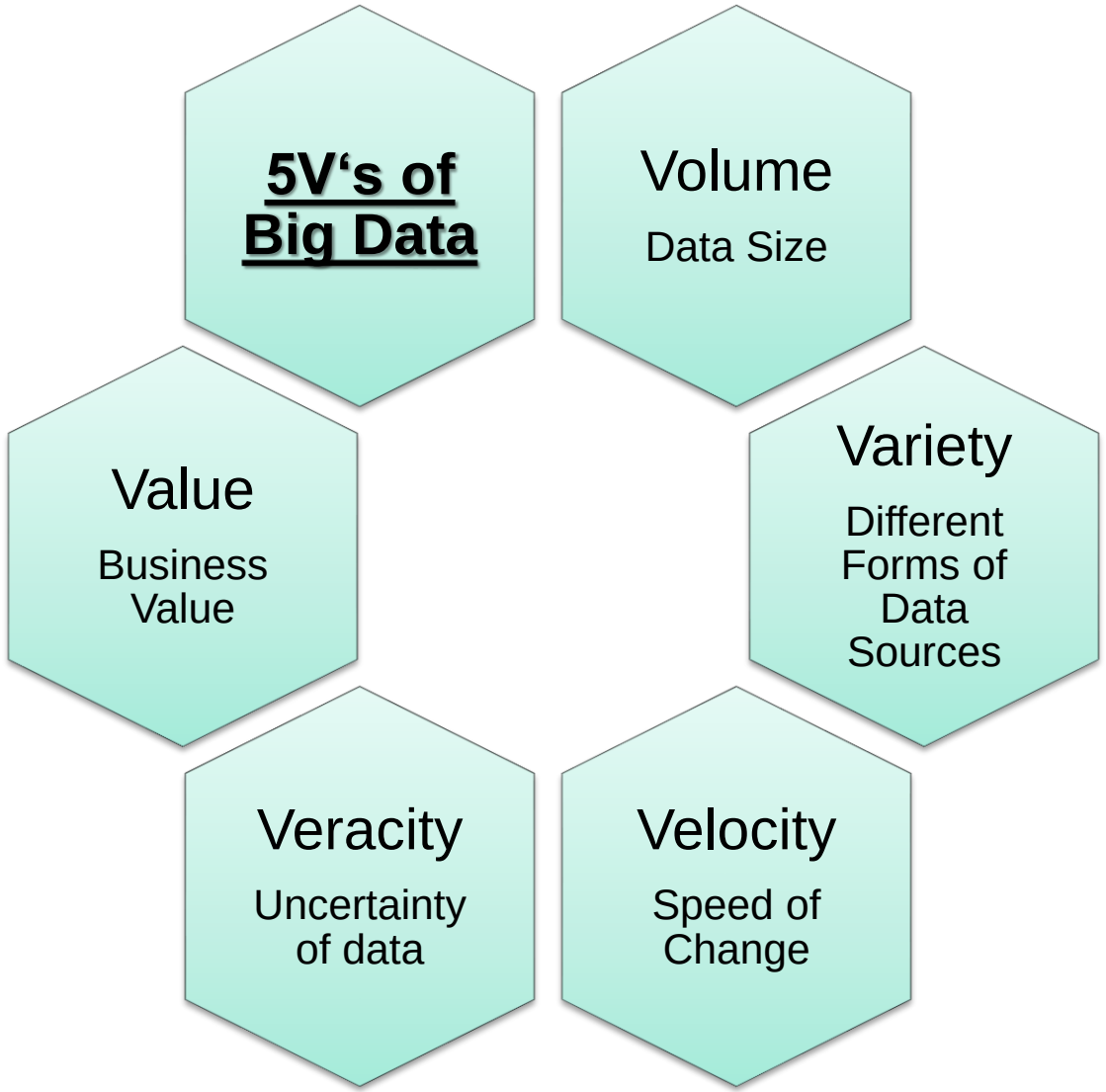
- Akzeptanz erst nach Aufnahme des Begriffes von größeren IT-Häusern wie IBM, SAP oder Oracle
- Anstieg der Begriffsnennung von 2009 bis 2012 um 1211%



Begriff muss abgrenzbar und bewertbar sein
→ Die 5V‘s von Big Data!

Big Data - Charakteristika

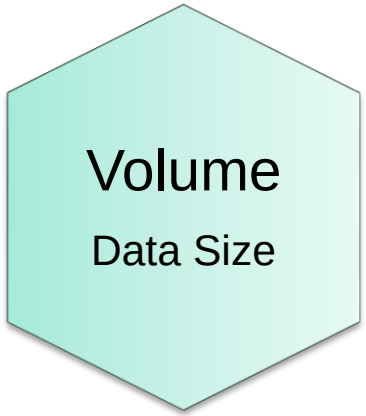
Begriffe: Die 5 V‘s



- **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

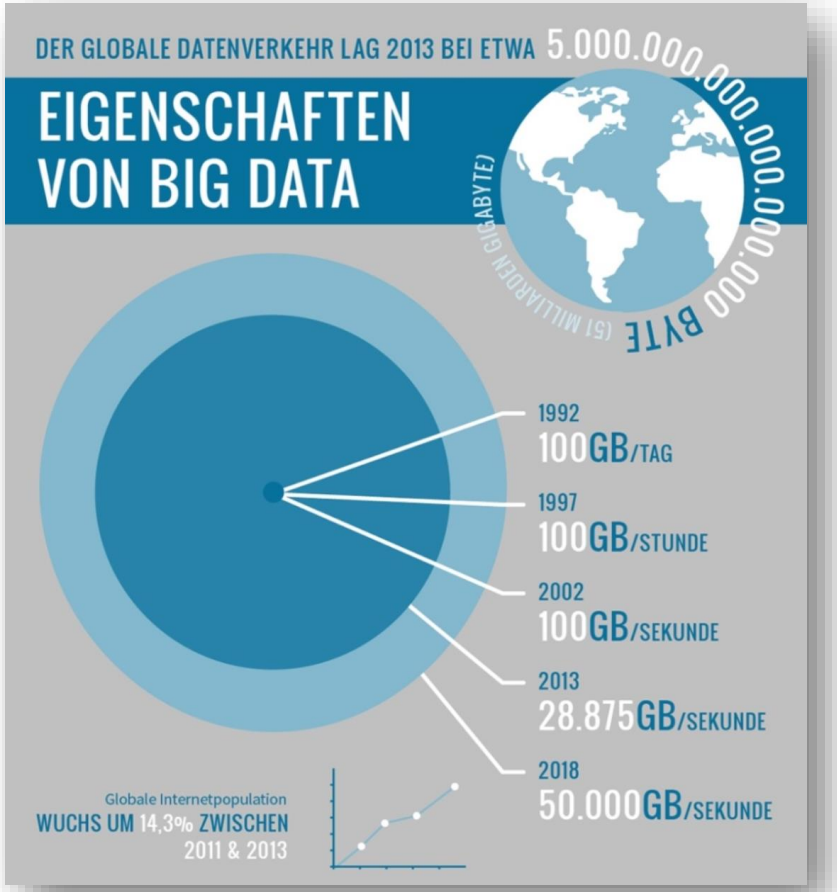
Big Data - Die drei grundlegenden V‘s

Volume



- 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
  - usw.

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



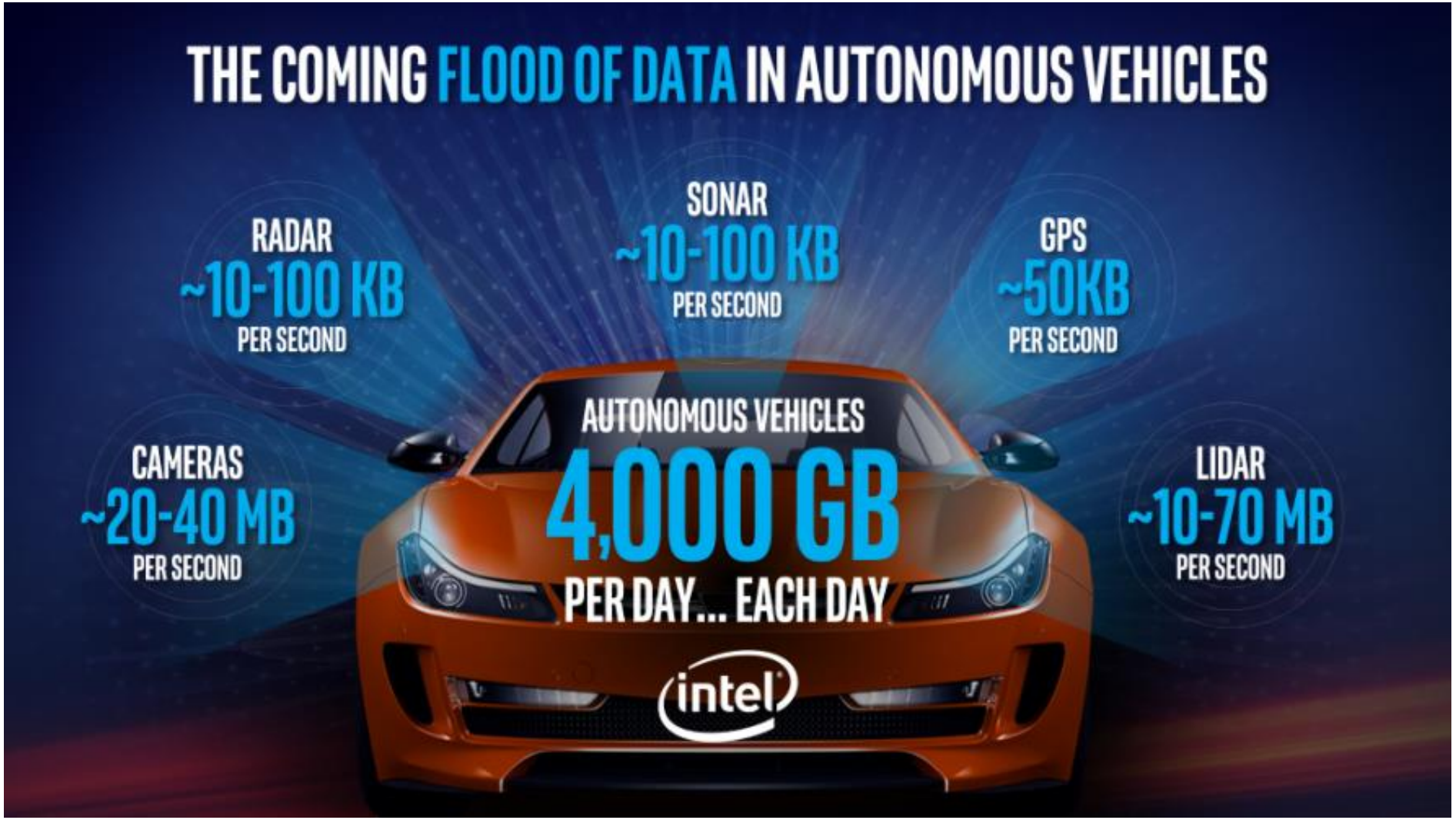
vouchercloud

Big Data - Die drei grundlegenden V‘s

Volume – Am Beispiel Fahrzeug

Volume

Data Size



Intel

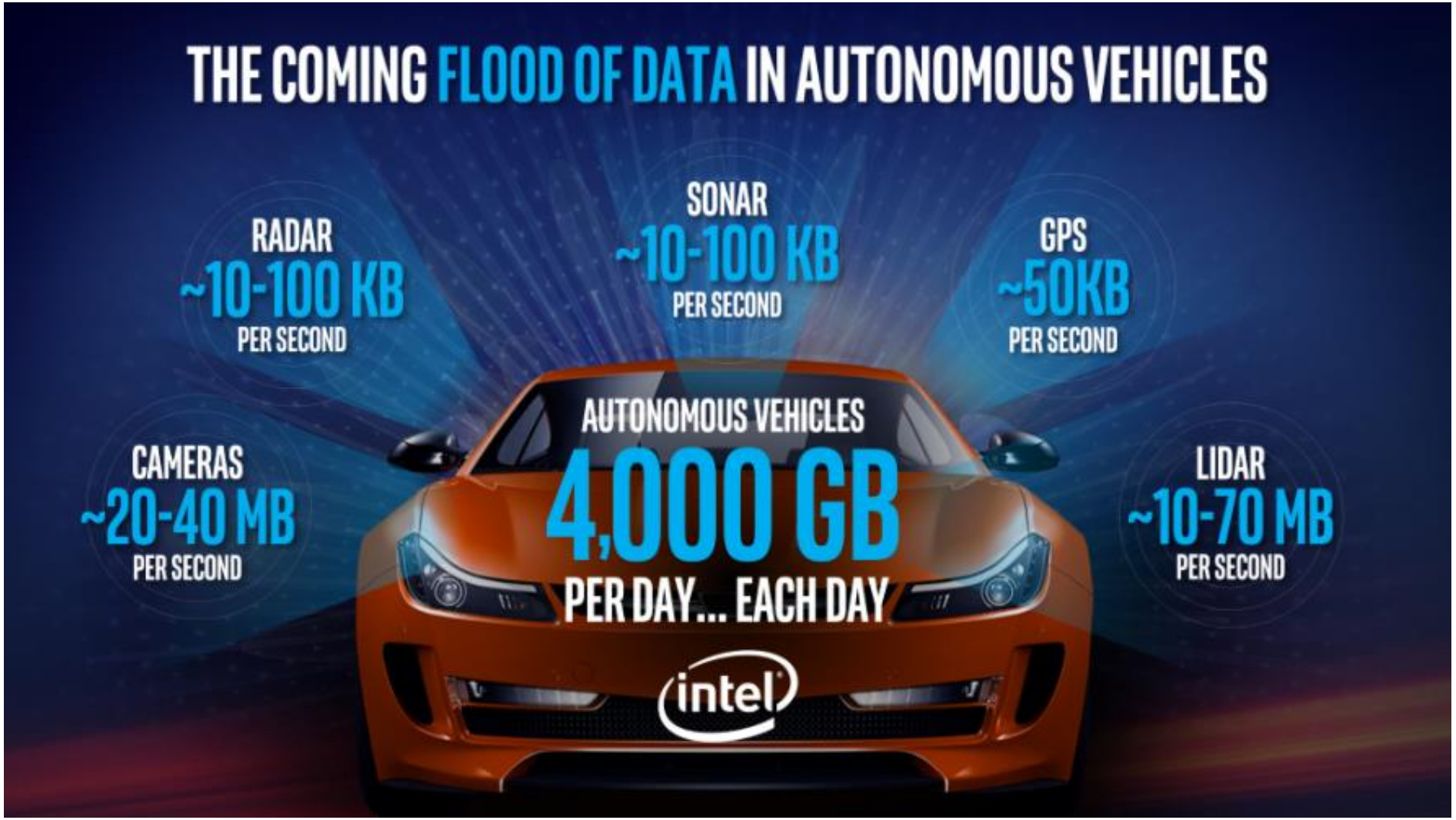
Big Data - Die drei grundlegenden V‘s

Velocity – Am Beispiel Fahrzeug

Velocity

Speed of Change

- GPS-Daten
- Kilometerstand
- Verbrauch
- Gurtstraffungen
- Reifendruck



Intel

Big Data - Die drei grundlegenden V‘s

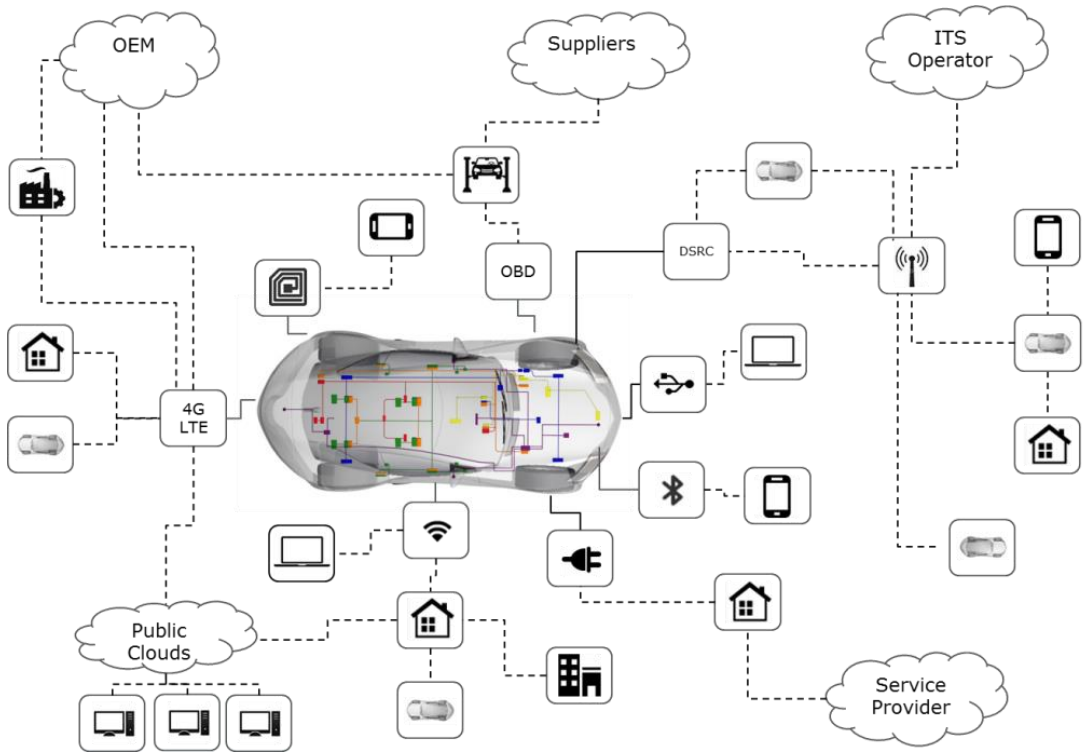
Variety – Am Beispiel Fahrzeug

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 - Die 5V‘s

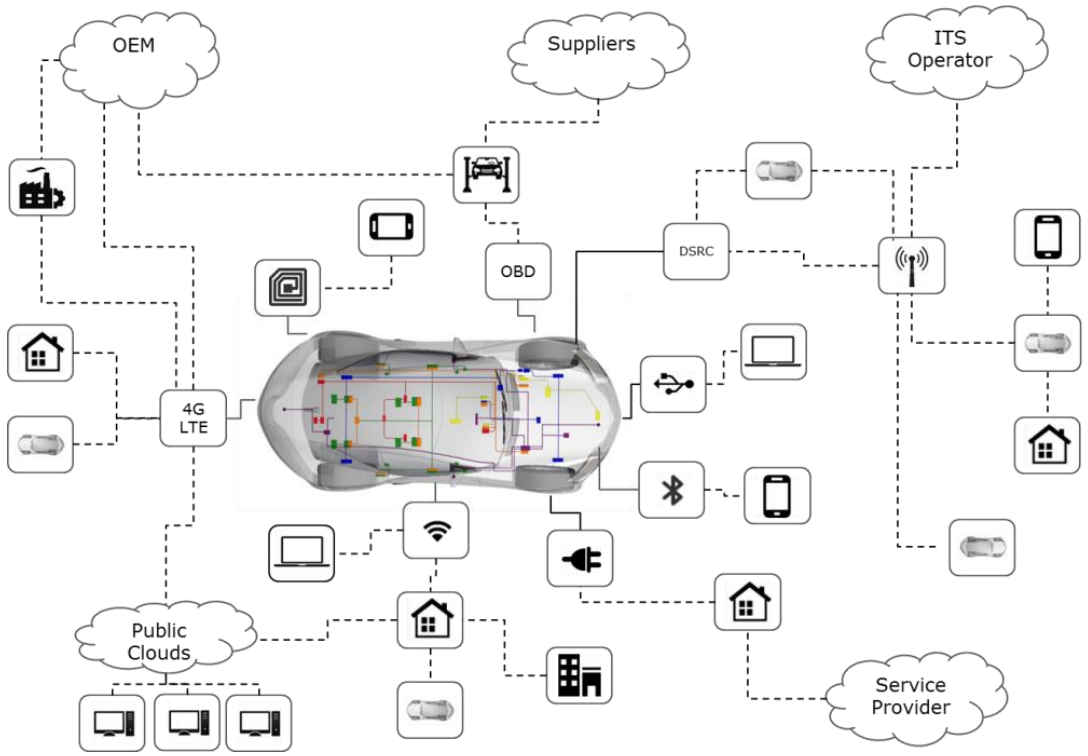
Veracity – Am Beispiel Fahrzeug

Veracity

Uncertainty of data

- Datenqualität und Zuverlässigkeit
- Daten stammen aus Quellen, deren Richtigkeit nicht immer ohne Weiteres angenommen werden kann

- Viele Maschinen und Systeme geben auch fehlerhafte Daten aus
- Daten müssen bereinigt werden
- Im Fahrzeug: Anomaliererkennung kann durch fehlerhafte Daten stark verfälscht werden
- Entscheidung, welchen Daten vertraut werden kann und muss



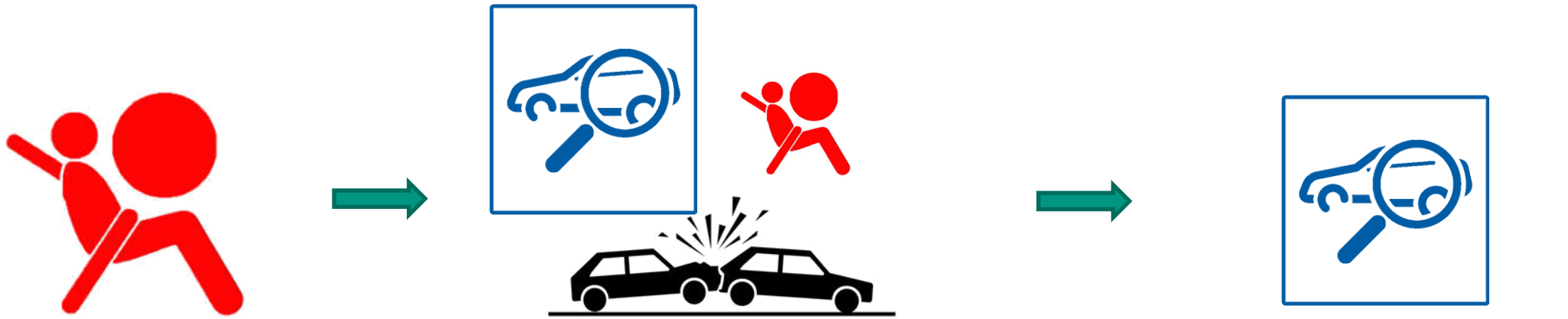
Big Data - Die 5V‘s

Value – Am Beispiel Fahrzeug

Value

Business Value

- Ziele
  - Extraktion von nützlichem Wissen und aussagekräftigen Informationen aus großen Datenmengen
  - Geschäftliche Entscheidungsfindung verbessern
- Daten können als strategisches Gut angesehen werden
- Erkenntnisse können wirtschaftlichen Vorteil bieten



Ereignis
Airbag wird ausgelöst

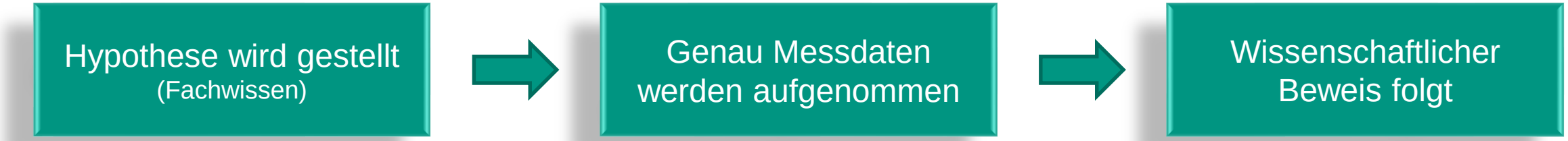
Korrelation
Bremsbeläge veraltet, Auffahrunfall, Airbag, Werkstatt

Vorhersage
Bremsbeläge veraltet kann zu Unfall führen → Werkstattfrühzeitig aufsuchen

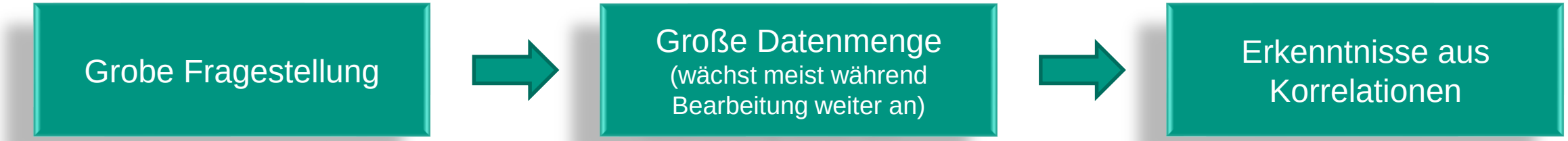
Big Data - Potential

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 Frage… …sa… …in… könnte und wie die drei grundle…

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

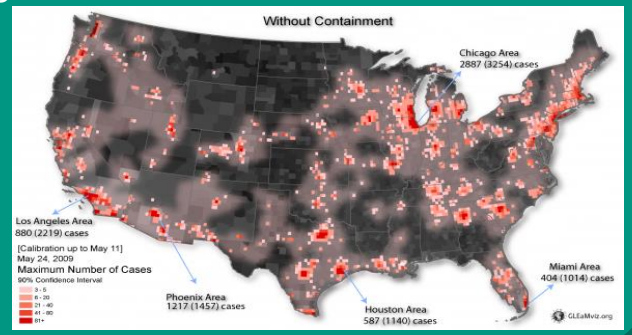
Variety
Wetter-daten

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

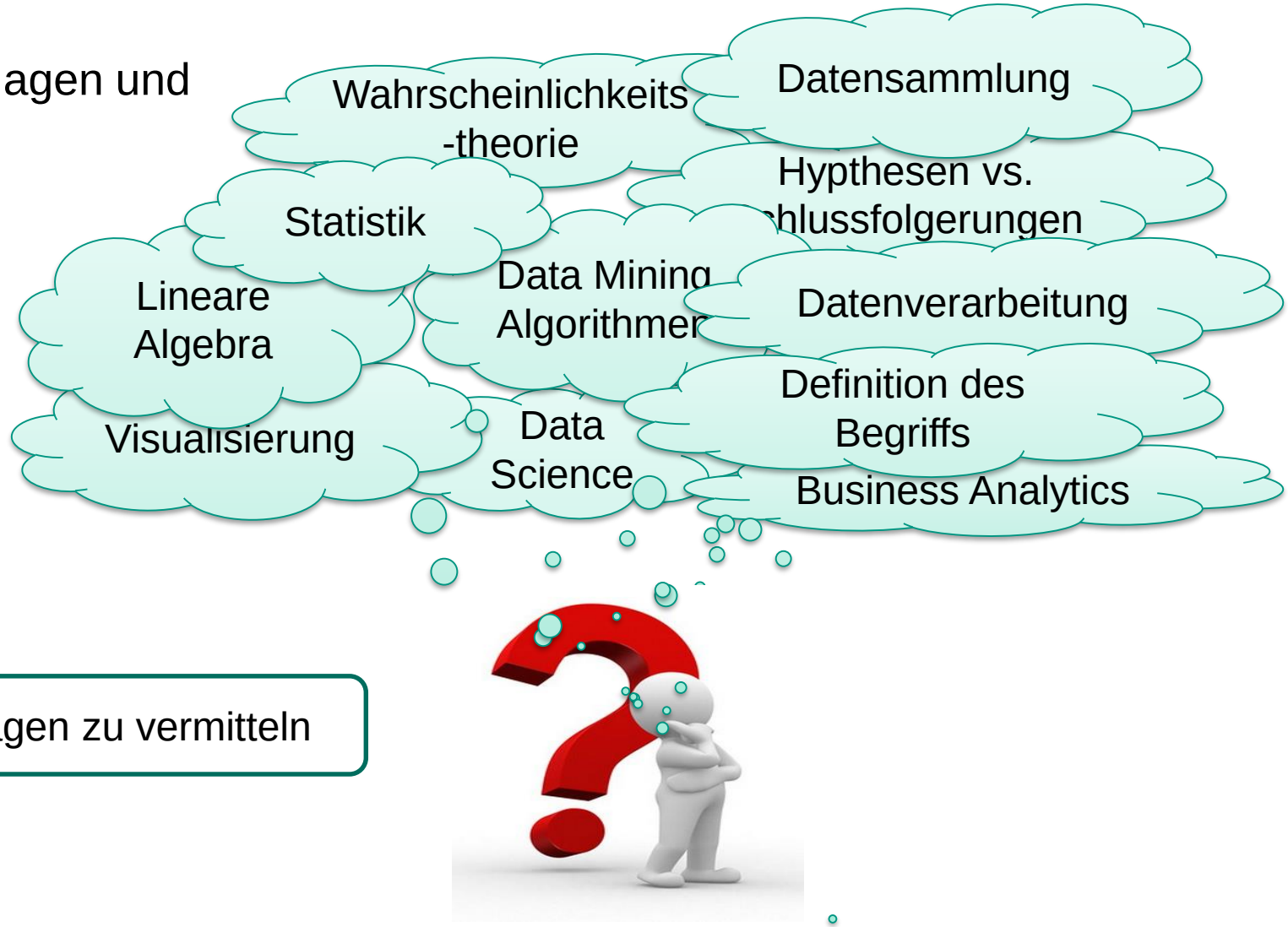


- 5V‘s
  - Volume
  - Variety
  - Velocity
  - Veracity
  - Value
- Hypothesengestützt vs. Schlussfolgerungen

CHARAKTERISTIKA ZUR ANALYSE GROSSER DATENMENGEN
TEIL 2 – GRUNDLAGEN

Big Data
Benötigte Grundlagen

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



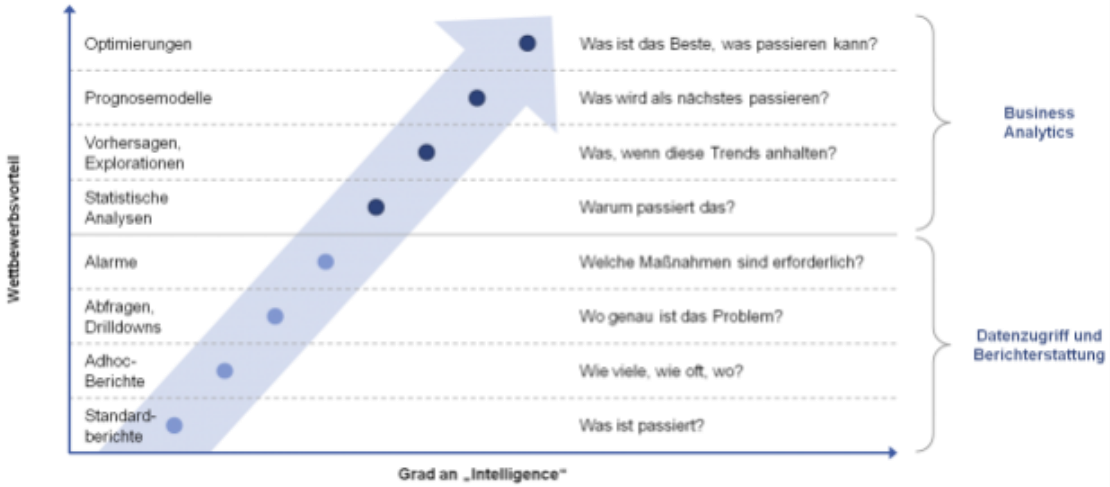
Ziel in IT2 diese und weitere Grundlagen zu vermitteln

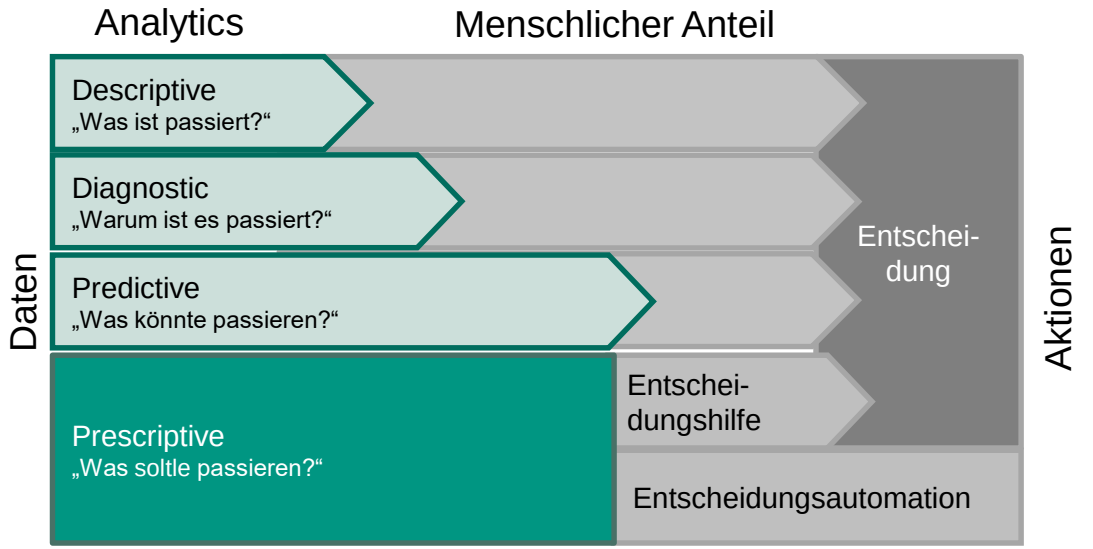
Big Data - Grundlagen

Business Analytics/Analyseformen

- Analytics:
  Umfassende Nutzung von Daten, statistische und quantitative Analysen, erklärende/voraussagende Modelle
- Business Analytics
  Analytics-Methoden/-Modelle im betrieblichen Kontext einsetzen, datengetriebene Managemententscheidungen herbeiführen
  Vorhersagen, Prognosen, Optimierungen

- Aufteilung Business 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

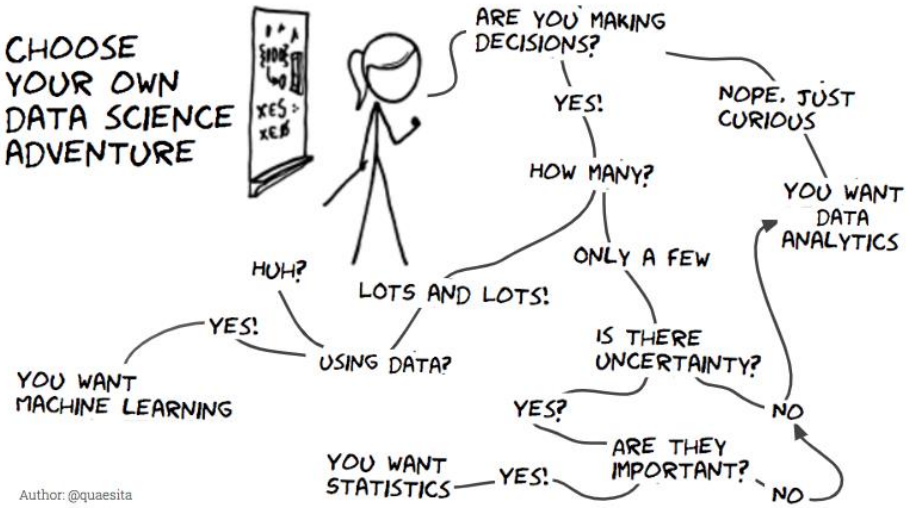


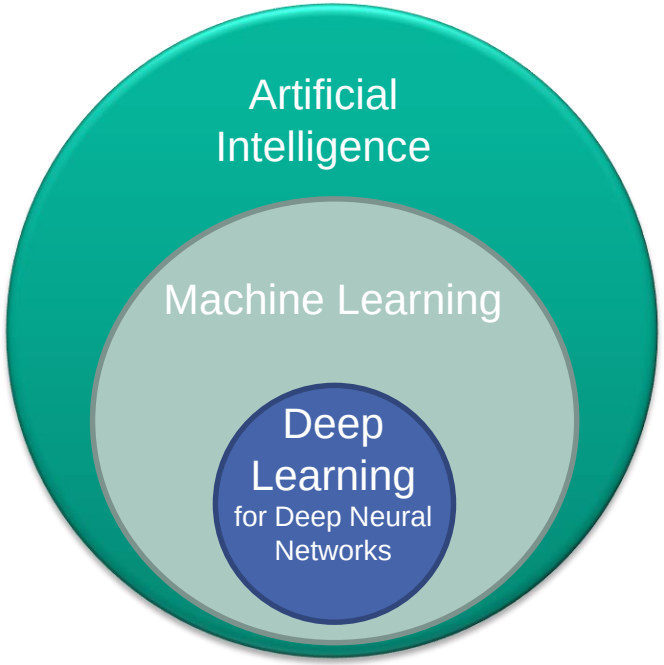


Big Data - Grundlagen

Data Science, Data Mining und Maschinelles Lernen

- **Data Science** bezeichnet die Extraktion von Wissen aus Daten. Dazu werden Methoden, Prozesse oder Algorithmen verwendet um Erkenntnisse, Muster und Schlüsse zu ziehen
  - Transformation von „Geschäftsprobleme“ in „Datenprobleme“
    - Datensammlung
    - Datenbereinigung
    - Datenformatierung
  - Extraktion von Wissen durch Data Mining

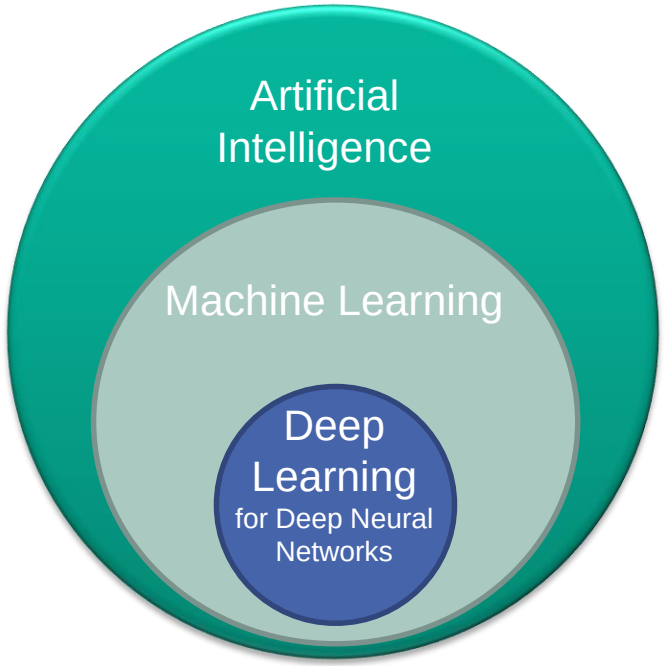




- **Data Mining** beschreibt die Eigentliche Extraktion des Wissens durch Anwendung von „Algorithmen“
  - Verwendung von statistischen Methoden auf große Datenmengen
  - Existiert bereits länger als „Artificial Intelligence“, „Machine Learning“ oder „Deep Learning“

Big Data - Grundlagen

Data Science, Data Mining und Maschinelles Lernen

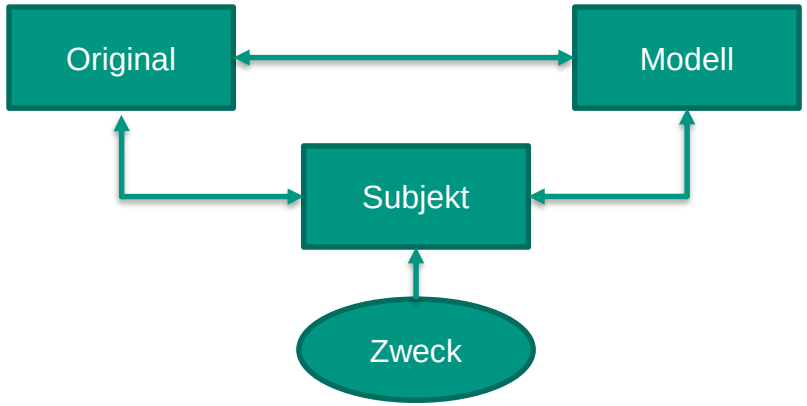


- **Artificial Intelligence**
  - Automatisierung intelligenten Verhaltens
  - Programmierung eines Computers so, dass er Probleme selbstständig löst
  - Anhand von Daten sollen Muster gefunden und Entscheidungen getroffen werden
- **Machine Learning** als Unterbegriff von AI
  - „Entwerfen eines aus Daten gelerntes Modells zur Vorhersage neuer Ergebnisse mit neuen Daten“
- **Deep Learning** als Unterbegriff von Machine Learning
  - Basieren auf sogenannten *neuronalen Netzen* mit mehreren Hidden Layers und weiteren Graphen Verfahren ⇒ Voraussichtlich Inhalt Übung 6

Maschinelles Lernen

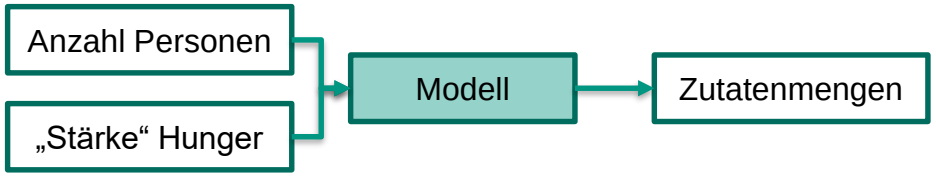
„Herkömmliche“ Modelle

- 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

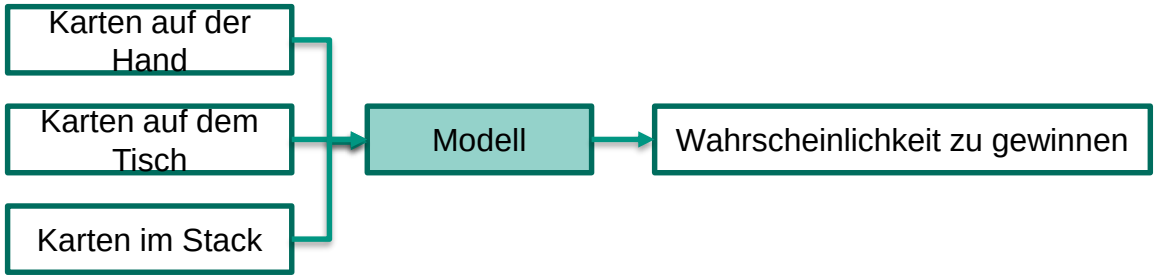


Modelle im Alltag:

Kochbuch/-rezept als Modell



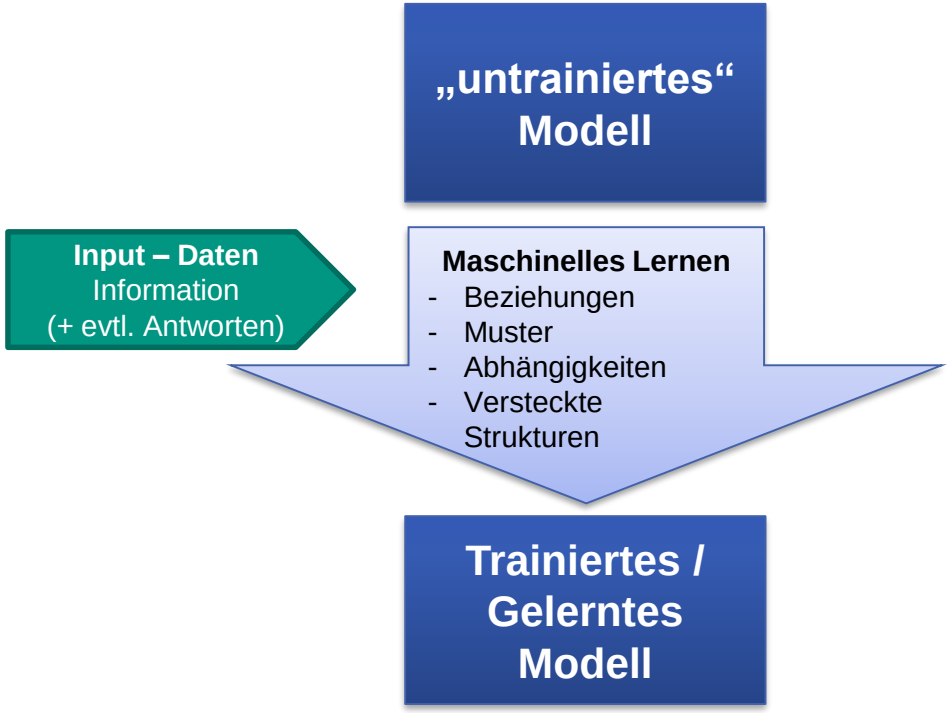
Pokern mit Modellen



Maschinelles Lernen

Maschinelles Lernen und Modelle

- Verschiedenste Definitionen von „Machine Learning“:
  - *„[Machine Learning is the] field of study that gives computers the ability to learn without explicitly programmed.“* – [Arthur Samuel, 1959]
  - *“A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.”* - [Tom Mitchel, 1997]

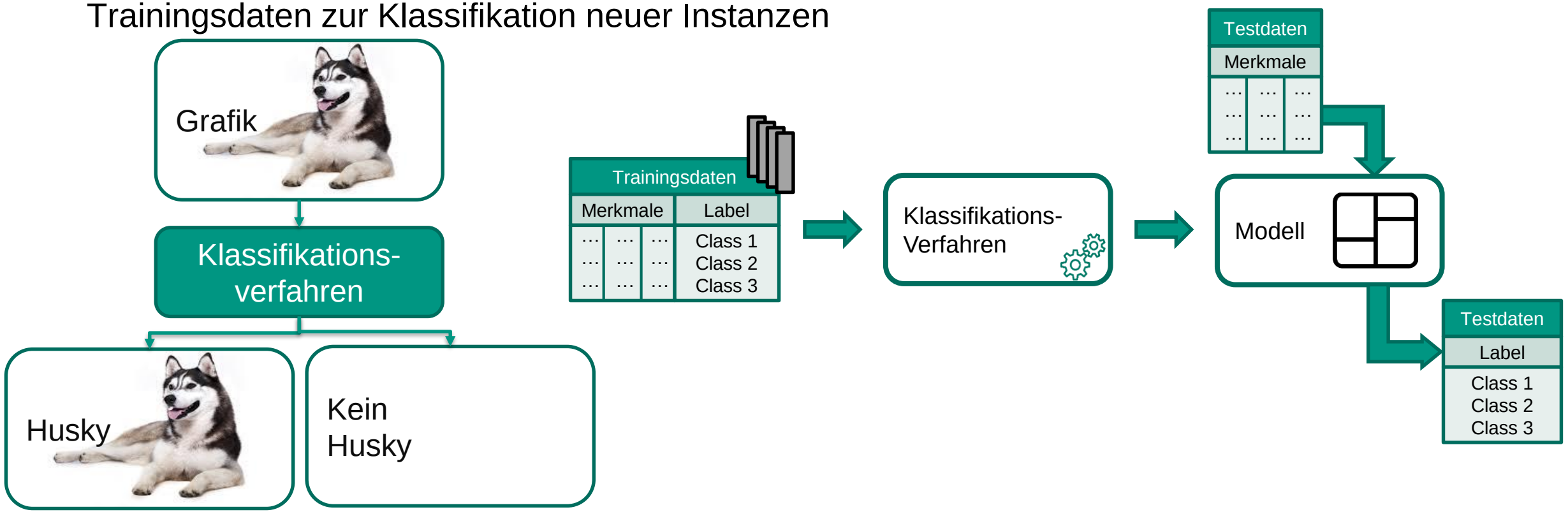


- Erstellung und Verwendung von Modellen, welche aus Daten gelernt werden
- Ziel ist aus vorhandenen Daten Modelle zu entwickeln, mit welchen verschiedene Ergebnisse für neue Daten vorhergesagt werden können
- Ergebnisse, die eventuell vorhergesagt werden können
  - Neue E-Mail als Spam erkennen
  - 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 „Klassifiktion“
  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

Lernen und Trainieren – Die Gefahr von Over- oder Underfitting

- Overfitting: Modell performiert gut auf Trainingsdaten, aber schlecht auf neuen Daten
  Rauschen kann ebenfalls mitgelernt werden
- Underfitting: Modell performiert sogar mit den Trainingsdaten schlecht
- ➢ Schlussfolgerung: Modell ist nicht ausgereift/schlecht geeignet, neues Modell erstellen

Der Versuch per Klassifikation anhand von Neuronalen Netzen mit der Aufgabe auf einer Grafik einen Husky zu erkennen, ging gründlich schief. Das Modell lernte lediglich Schnee zu erkennen, da Huskies meist vor weißem und blauem Hintergrund aufgenommen werden.

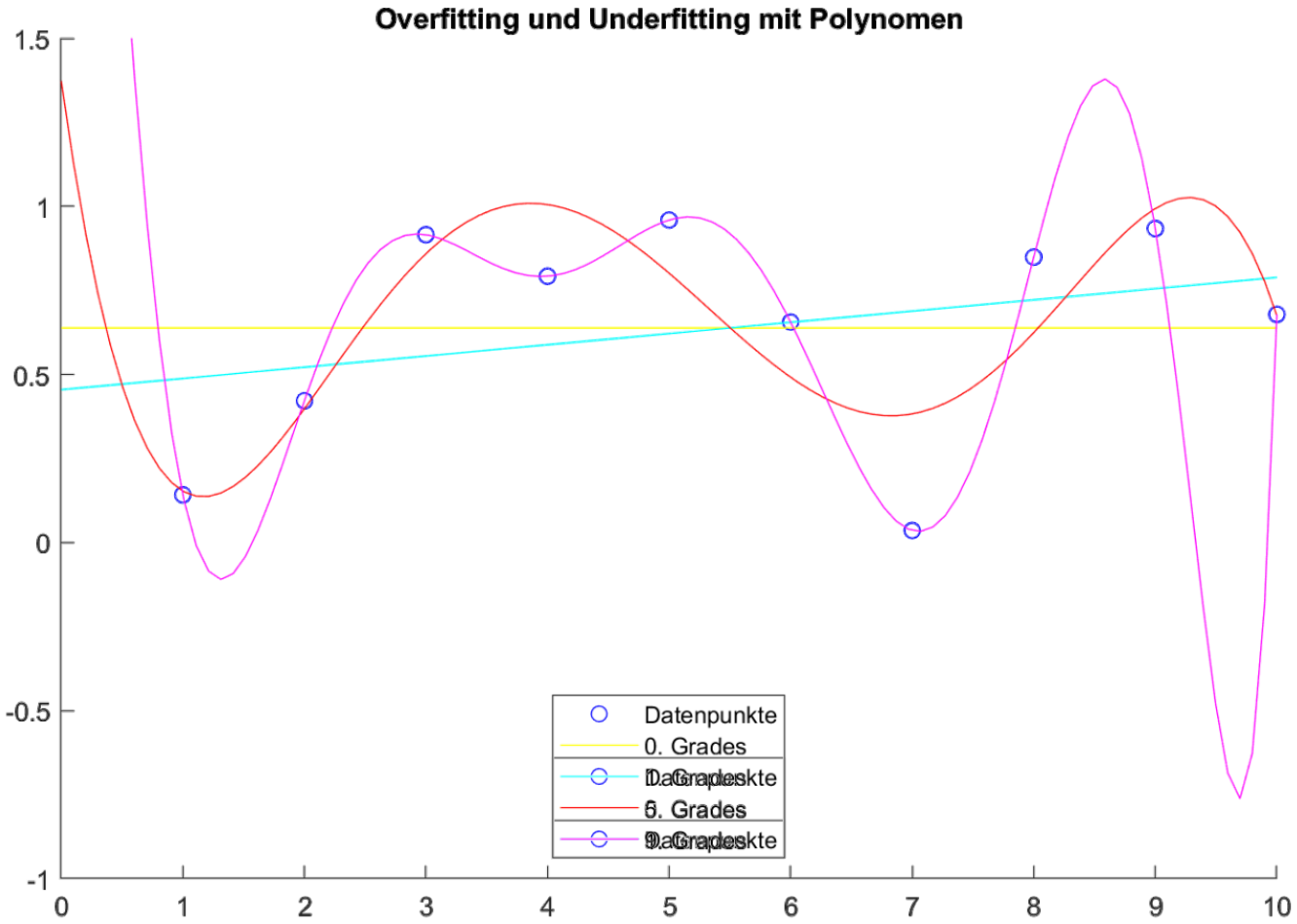
Das Modell performierte sehr gut auf den Trainingsdaten, versagte jedoch bei neuen Daten. Es erkannte Schneelandschaften eindeutig als Bilder von Huskies

Overfitting und Underfitting als Gefahrenquelle bei Machinellen Lernverfahren

Maschinelles Lernen

Over- oder Underfitting - Beispiel

- Ziel ist es mit einem Polynom möglichst alle Datenpunkte so gut wie möglich abzudecken
- Polynome verschiedenen Grades
  - 0. Grades: keine Abdeckung
  - 1. Grades: keine Abdeckung
  - 5. Grades: akzeptable Abdeckung
  - 9. Grades: perfekte Abdeckung



- 0. und 1. Grades sind nicht aussagekräftig bezüglich der Datenpunkte (Underfitting)
- 9. Grades wird beim Hinzukommen weiterer Datenpunkte diese kaum einschließen (Overfitting)

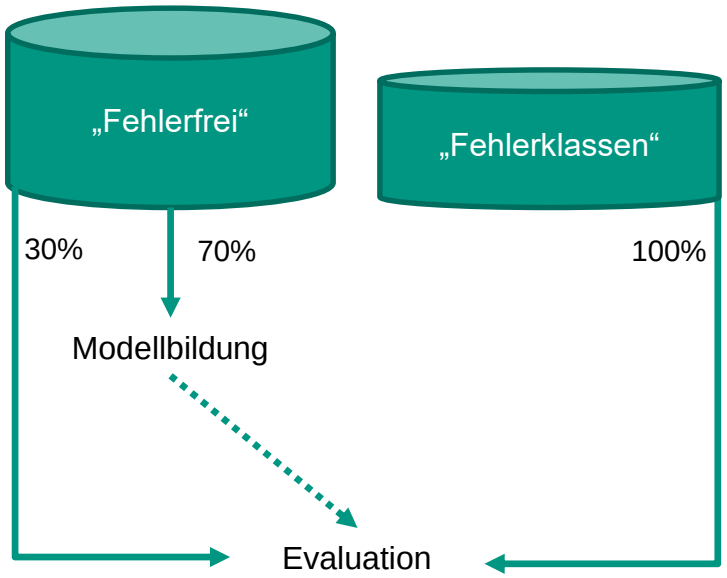
Maschinelles Lernen

Lernen und Trainieren anhand Klassifikation

- Daten sind vorhanden, aber wie trainiert man nun?

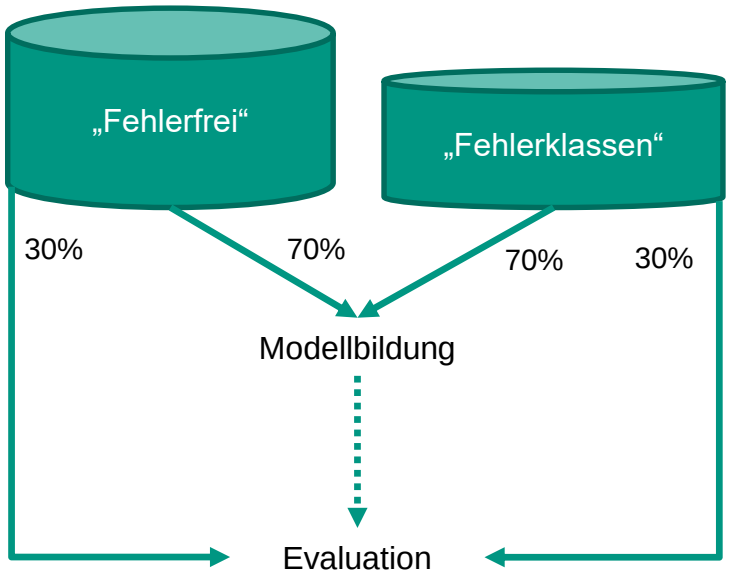
Ein-Klassen Klassifikation

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



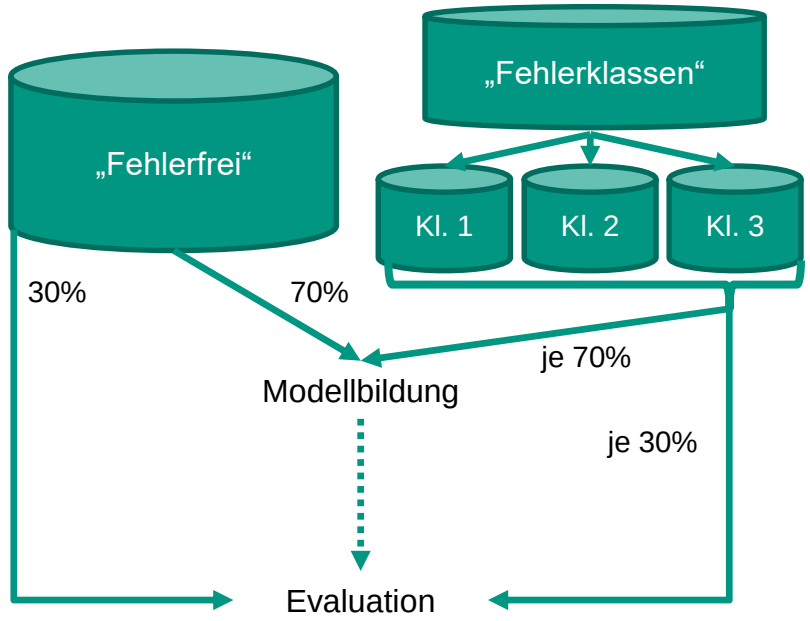
Zwei-Klassen Klassifikation

- 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

Ziele der heutigen Übung

- Nach der heutigen Übung können Sie….

- Charakteristika, Notwendigkeit und Vorgehensweisen zur Analyse großer Datenbestände beschreiben

1. … Notwendigkeit zur Analyse großer Datenbestände nennen
2. … Charakteristika zur Abgrenzung von Big Data nennen
3. … Grundbegriffe im Bezug zu Big Data nennen
4. … den Begriff des „Maschinelles Lernens“ abgrenzen
5. … den Begriff und das Vorgehen beim „Trainieren“ abgrenzen