Data Science
8. Teil – Unüberwachtes Lernen und Big Data Analytics

Vorlesung an der DHBW Stuttgart, Prof. Dr. Monika Kochanowski

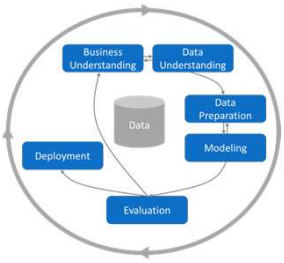
1

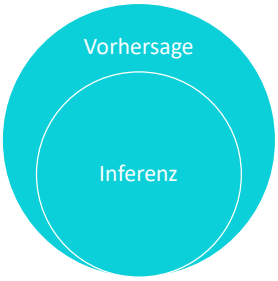
Inhalte der heutigen Vorlesung

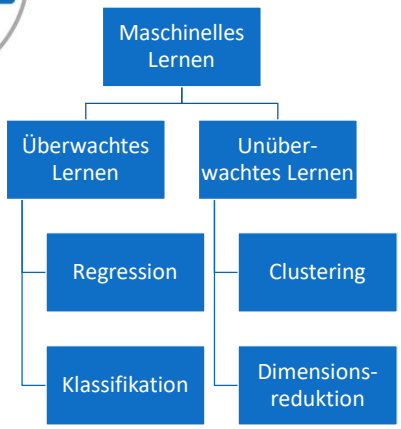
- Unüberwachte Lernverfahren
- PCA (Dimensionsreduktion)
- K-Means (Clustering)
- Hierarchische Verfahren (Clustering)
- Einsatz von Clustering für Big Data
- Association Rule Learning (Unüb. Verfahren)

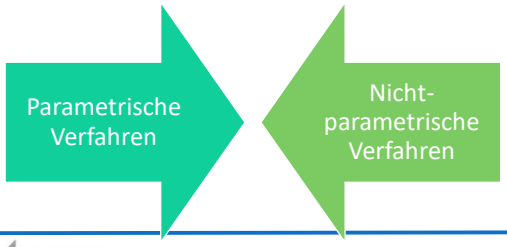
2

Wiederholung









3

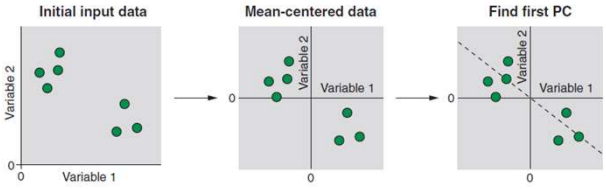
Unsupervised Learning

- **Inhärente Struktur der Daten erkennen und daraus lernen**

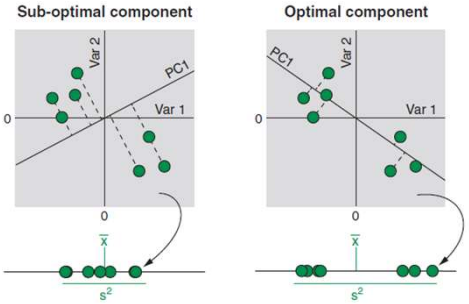
- **Verfahren zur Reduktion der Dimensionalität**
  - Principal Component Analysis (PCA)
  - Self-Organizing-Maps (SOM's)
  - Independent Component Analysis (ICA)
  - … etc.
- **Clustering**
  - k-Means
  - Gaussian Mixture Models (GMM)
  - …
- **Association Rule Learning**

4

PCA – Hauptkomponentenanalyse
Der Klassiker



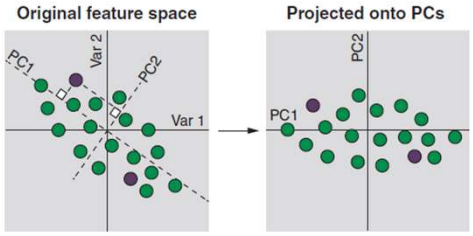
- Verwendet für
  - Visualisierung von höherdimensionalen Daten
  - Vermindert die „Curse of Dimensionality“
  - Vermindert Effekte der Kolinearität von Features/Variablen der Daten
- Ziel in Kurzform
  - Sucht die Koordinaten und das Koordinatensystem, dass die Daten am besten und aussagekräftigsten beschreibt
- Schritt 1: z-Transformation (siehe andere Folien)
- Schritt 2: Achse finden mit „Aussagekraft“
  - Muss durch den Ursprung gehen
  - Muss die Varianz der auf sie projizierten Datenpunkte **maximieren**
  - „PC1“ Erste Hauptkomponente
- Schritt 3: Nächste Hauptkomponente orthogonal suchen



Machine Learning with R, ..., Hefin I. Rhys, Manning, 2020

5

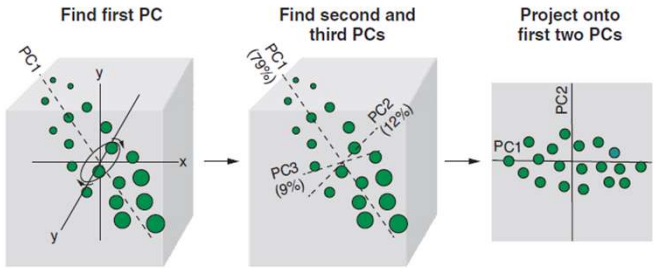
PCA – Beispiel



Im 2D-Fall ist PC2 durch PC1 definiert

Rotation der Daten in‘s Koordinatensystem PC1,PC2

Eventuell Darstellung nur für PC1 – Reduktion auf 1D

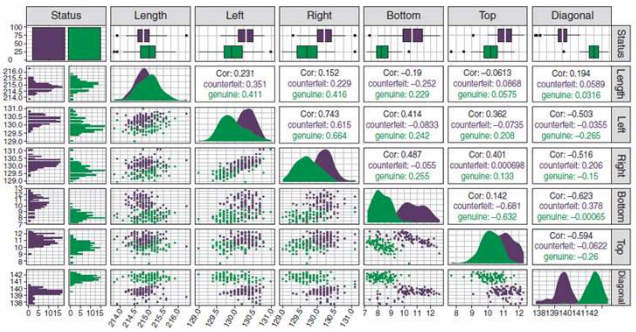


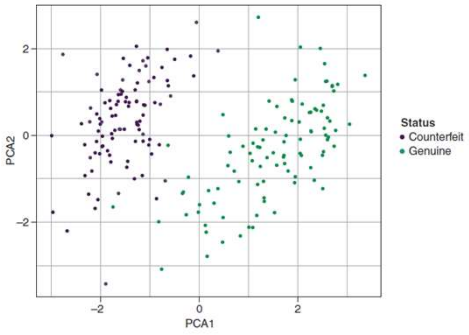
Reduktion von 3D auf 2D in‘s System von PC1 und PC2
Enthält 91%=79%+12% der Varianz der Daten

Machine Learning with R, ..., Hefin I. Rhys, Manning, 2020

8

PCA – Banknotenbeispiel





- 2D-Plot der 70.4% der Varianz der Daten enthält
- Klar zwei separate Cluster erkennbar

Machine Learning with R, ..., Hefin I. Rhys, Manning, 2020

11

PCA

Die von der PCA erzeugten Achsen sind direkt interpretierbar in Bezug auf die Originalachsen

Neue Datenpunkte können einfach auf die PC's projiziert werden

PCA ist eine rein mathematische Projektion und deshalb nicht rechenaufwendig

Die Abbildung aus den hochdimensionalen auf die niedrigdimensionalen Räume kann nicht nichtlinear sein

PCA kann nicht mit kategorialen Daten umgehen – eventuell kodieren und in numerische überführen

Die Auswahl der Dimensionen, die betrachtet werden, muss extern beschlossen werden und kann nicht aus dem Verfahren selbst gewonnen werden.

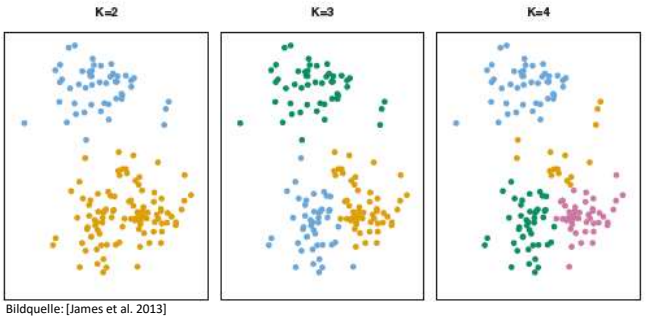
17

Inhalte der heutigen Vorlesung

- Unüberwachte Lernverfahren
- PCA (Dimensionsreduktion)
- K-Means (Clustering)
- Hierarchische Verfahren (Clustering)
- Einsatz von Clustering für Big Data
- Association Rule Learning (Unüb. Verfahren)

Clustering für unüberwachtes Lernen
Clustering Inhalte heute

- Suche eine »gute« Partition der Daten, innerhalb der Daten sehr ähnlich sind und..
  - die Daten außerhalb der Gruppe sehr unähnlich
- Definition von »ähnlich« und »unähnlich« muss vorhanden sein
- Methoden
  - K-Means
  - Hierarchisches Clustering
- Überblick über aktuelle Verfahren
  - Modellbasiert
  - Dichtebasiert
- Clustermetriken
- Beispiel in Python



Bildquelle: [James et al. 2013]

K-Means Clustering
Ähnlichkeiten mit kNN nicht rein zufällig

- »Gute« Cluster haben wenig Varianz innerhalb des Clusters $W(C_k)$
  - Verschiedene Distanzmaße möglich (gängig: Euklidische Distanz, Exkurs nächste Folie)
- Jeder Datenpunkt gehört zu genau einem Cluster
- Minimierungsproblem: $\underset{C_1,..,C_K}{\text{minimize}}\{\sum_{k=1}^{K} W(C_k)\}$
  - Setze quadrierte Euklidische Distanz ein: $\underset{C_1,..,C_K}{\text{minimize}}\left\{\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i\prime\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i\prime j})^2\right\}$

```
For i = 0 .. p-1
    assign (data_i, cluster(random(0.. K-1))

while (!stop)
      for j = 0 .. K-1
              compute_centroid(cluster_j) //Mittelwerte der Eingaben
      for i = i .. p-1
              assign (data_i, find_closest_cluster_by_centroid(data_i, all_clusters)
```

Freiwillige Übungsmöglichkeit: K_Means manuell implementieren

30

Exkurs: Distanzmetriken
Alternativen für k-NN (und auch für Cluster)

- **Euklidische Abstand**
  - »die« Distanz (wie wir sie kennen) $d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$

- **Jaccard-Metrik**
  - Distanz zwischen Mengen von Objekten (z. B. Twitter-Follower) $\mathrm{J}(\mathrm{A},\mathrm{B})=\frac{|\mathrm{A}\cup\mathrm{B}|-|\mathrm{A}\cap\mathrm{B}|}{|A\cup B|}$

- **Mahalanobis-Distanz**
  - Zwei »Vorteile« im Vergleich zu Euklidischer Distanz für Vektoren:
  - (1) Korrelation wird beachtet und (2) Skalierungsunabhängig
- **Hamming-Distanz**
  - Ähnlichkeit zwischen zwei Wörtern / Strings / gleicher Länge (z. B. bits, DNA)

GTACC
GTCCA

- **Manhattan-Distanz**
  - Dimensionen werden einzeln betrachtet $d(x,y)=\sum_{i=1}^{n}|x_i-y_i|$

[O'Neil and Schutt 2013]
Anmerkung: Je nach Metrik, insb. Euklidische Distanz, kann die z-Transformation notwendig werden.

31

K-Means Clustering
Lokales Minimum gesucht

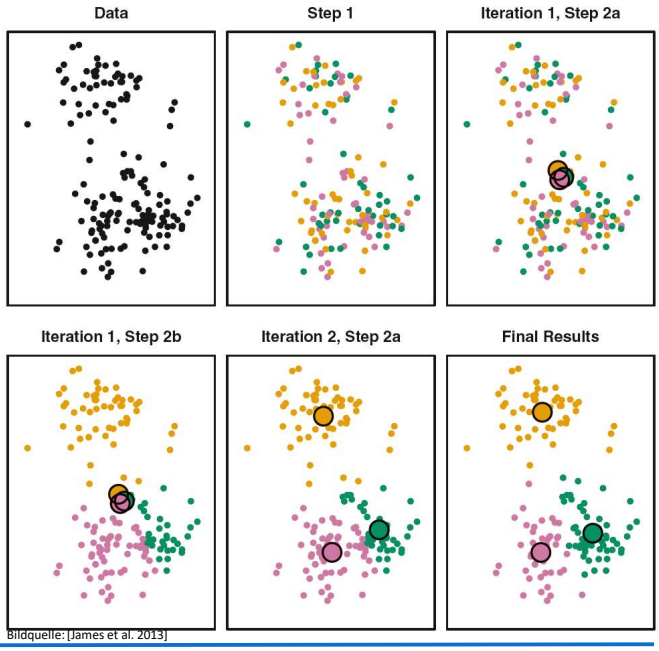
- Der Algorithmus verbessert sich in jeder Iteration und konvergiert
- Nachteil: findet lokales Optimum abhängig von Initialisierung
  - Zufällig
  - Intelligent
- Mehrfahre Ausführung
- **Wie findet man k?**

Bildquelle: [James et al. 2013]

33

K-Means Clustering
Lokales Minimum gesucht

- Der Algorithmus verbessert sich in jeder Iteration und konvergiert
- Nachteil: findet lokales Optimum abhängig von Initialisierung
  - Zufällig
  - Intelligent
- Mehrfahre Ausführung
- **Wie findet man k?**



Bildquelle: [James et al. 2013]

34

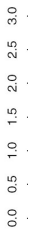
Inhalte der heutigen Vorlesung

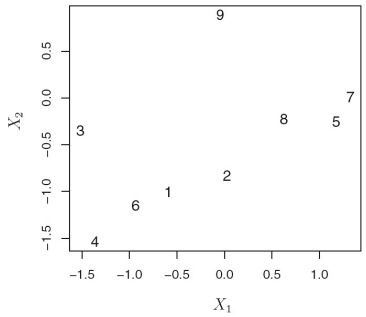
- Unüberwachte Lernverfahren
- PCA (Dimensionsreduktion)
- K-Means (Clustering)
- Hierarchische Verfahren (Clustering)
- Einsatz von Clustering für Big Data
- Association Rule Learning (Unüb. Verfahren)

Hierarchisches Clustering
Agglomeratives hierarchisches Clustering – ein Bottom-up Verfahren

- **Dendrogramm** (eine Art umgekehrter Baum)
  - Ähnlichste Zweige werden zusammengefasst
- **Wie viele Cluster würden Sie aus dem rechts gezeigten Beispiel erstellen?**

```
Beginne mit p Clustern
Finde ähnlichste Cluster und
verbinde diese zu einem Cluster
Berechne Ähnlichkeit für alle
Cluster neu
```





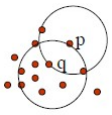
Bildquelle: [James et al. 2013]

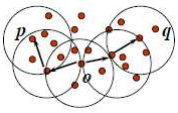
Hierarchisches Clustering
Metriken

- **Complete**
  - Complete-Link-Clusterverfahren – berechne maximale Unähnlichkeit aller Datenpunktpaare
  - $d(G_i, G_j) = \max_{x^r \in G_i, x^s \in G_j} d(x^r, x^s)$
- **Single**
  - Single-Link-Clusterverfahren – berechne minimale Unähnlichkeit aller Paare
  - $d(G_i, G_j) = \min_{x^r \in G_i, x^s \in G_j} d(x^r, x^s)$
- **Average**
  - Average-Link-Methode
  - Berechne den Durchschnitt der Entfernungen zwischen allen Paaren
- **Centroid**
  - Zentroid-Abstand
  - Berechne Entfernung zwischen den Zentroiden (Mittelwerten) von zwei Clustern

Clustering
Fortgeschrittene Algorithmen

- **Modellbasierte Algorithmen**
  - Mathematisches Modell einer Wahrscheinlichkeitsverteilung (dichte) wird angenommen
  - EM Algorithmus (Expectation Maximization)
  - Erweiterung von k-Means
  - (1) Jeden Datenpunkt einem Cluster mit einer **Wahrscheinlichkeit** zuordnen (Satz von Baves)
  - (2) **Modellparameter** neu berechnen
- **Dichte-basiertes Clustering**
  - Bekannt: **DBSCAN**, Weiterentwicklung: OPTICS
  - Ähnlich hierarchisches Clustering, aber Dichte eines Clusters wird berücksichtigt
  - Erreichbare Punkte (density-reachable) – wenn in Radius *und* Anzahl der Punkte ausreichend hoch
  - OPTICS als Weiterentwicklung (**Links**)
- Viele weitere fortgeschrittene Methoden
  - http://scikit-learn.org/stable/modules/clustering.html
  - https://en.wikipedia.org/wiki/Cluster_analysis





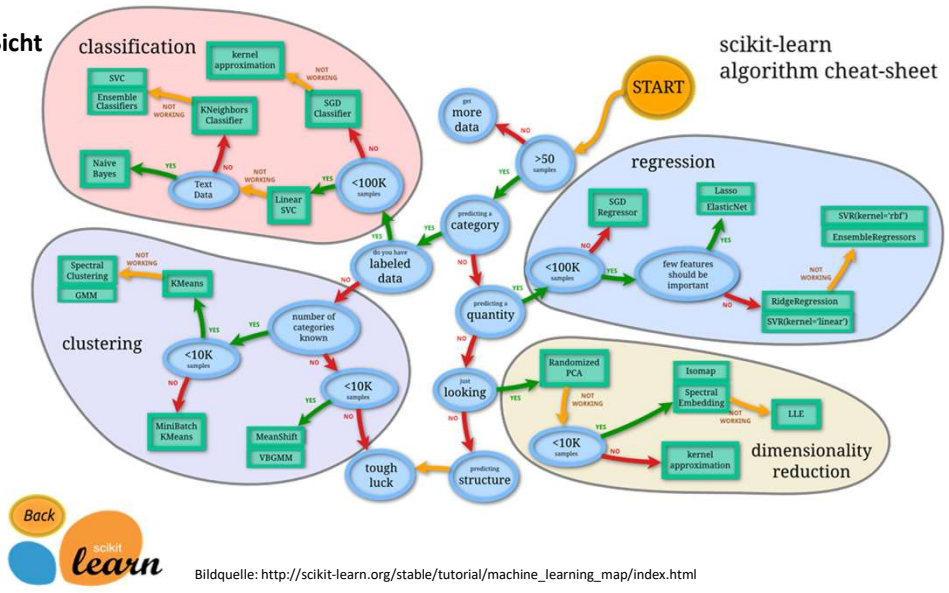
Bildquelle: "Data Mining Session 9", Dr. Jean-Claude Franchitti, *New York University*

Clustering
Metriken für Cluster-Qualität

- **Davies-Bouldin Index**
  - Hohe Ähnlichkeit im Cluster UND
  - Niedrige Ähnlichkeit zwischen Clustern (**Kompaktheit**)
  - »Gute« Lösungen haben einen niedrigen Wert $DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$
- **Dunn Index**
  - Verhältnis der *minimalen* Distanz zwischen Clustern und der *maximalen* Distanz im Cluster
  - Identifiziert dichte und gut separierte Cluster (**Separation**) $D = \frac{\min_{1 \leq i < j \leq n} d(i,j)}{\max_{1 \leq k \leq n} d'(k)}$
- **Silhouettenkoeffizienz**
  - Durchschnittliche Distanz zu Elementen *desselben* Clusters zu durchschnittlicher Distanz zu Elementen *anderer* Cluster $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$
  - Hoher Wert (für einen Datenpunkt): gut, niedrig: mögliche Ausreißer

39

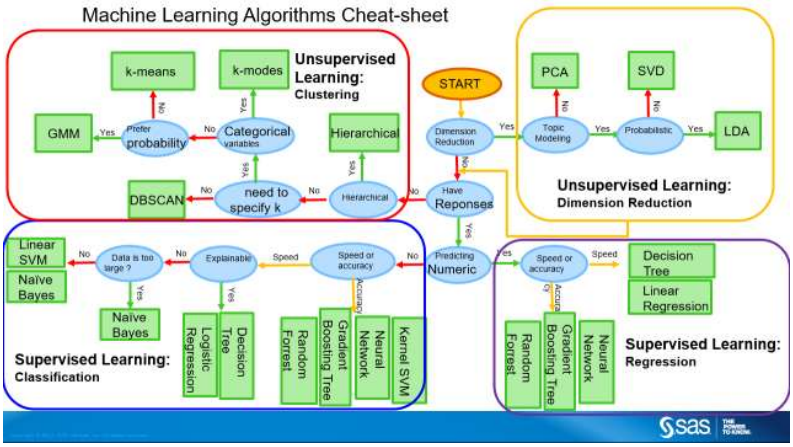
Stand
scikit-Sicht



Bildquelle: http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

40

Stand der Vorlesung
SAS Cheat Sheet Sicht



Bildquelle: https://whatsthebigdata.com/2017/05/02/types-of-machine-learning-algorithms-and-when-to-use-them/, am 10.02.2018

41

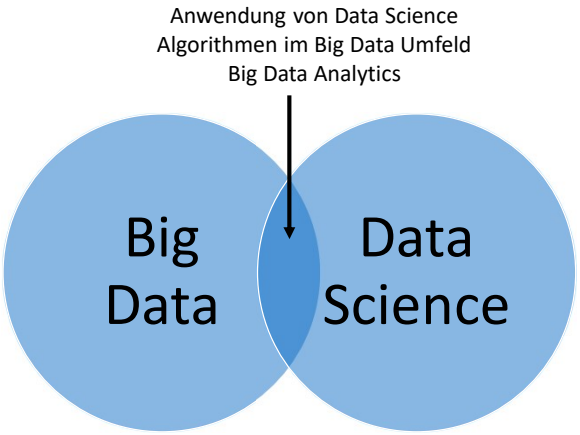
Inhalte der heutigen Vorlesung

- Unüberwachte Lernverfahren
- PCA (Dimensionsreduktion)
- K-Means (Clustering)
- Hierarchische Verfahren (Clustering)
- Einsatz von Clustering für Big Data
- Association Rule Learning (Unüb. Verfahren)

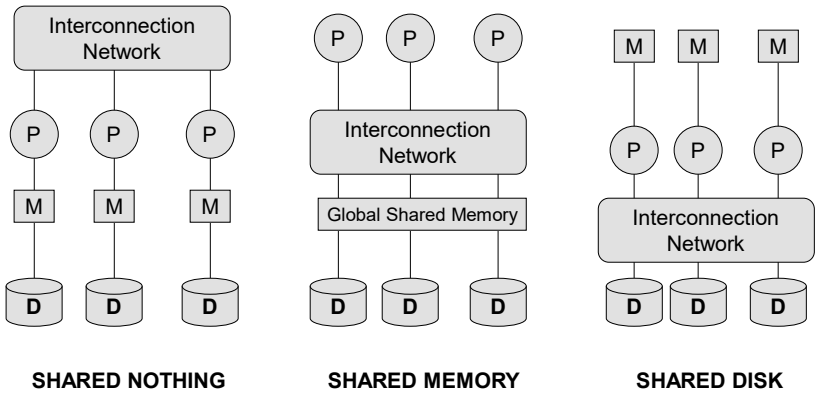
42

Big Data und Data Science
Einführung

- Big Data als »Trendthema«
  - Große Datensätze
  - Verteilung
- Hohe Verfügbarkeit erforderlich
- Hohe Performanz erforderlich
- ACID-Kriterien als »Bremsklotz«
- Aktuelle Entwicklungen (CPU, Speicher)
  - Verteilung
  - Parallelisierung der Anwendungen
- **Shared-nothing-architecture**



Exkurs: Architekturen für verteilte Systeme



Quelle: Ramakrishnan, Raghu; Gehrke, Johannes (2003): Database management systems. Third edition, international edition. New York: McGraw-Hill.

Exkurs: ACID Kriterien

Atomicity	Eine Transaktion ist die kleinste, nicht mehr weiter zerlegbare Einheit.
Consistency	Nach Durchführung der Transaktion ist die Datenbank wieder in einem konsistenten Zustand, ansonsten wird sie zurückgesetzt.
Isolation	Nebenläufige Transaktionen beeinflussen sich nicht gegenseitig. Jede wird „logisch“ ausgeführt als gäbe es keine andere.
Durability	Dauerhaftigkeit (Persistenz) durchgeführter Aktionen muss gewährleistet sein. Alle späteren Fehler betreffen die hier durchgeführten Änderungen nicht.

- Kunde überweist Firma 100 EUR
- Firma wird 100 EUR gutgeschrieben
- System stürzt ab

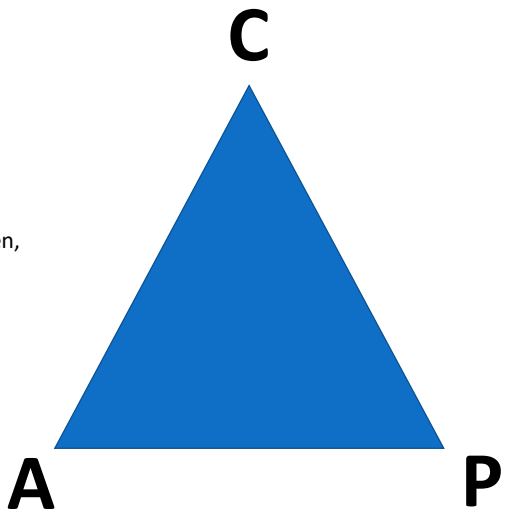
45

Das CAP Theorem
Verteilte Systeme

- **C Konsistenz**
  - Alle Nodes sehen immer denselben Zustand
- **A Verfügbarkeit**
  - Jede Anfrage wird bearbeitet
- **P Partitionstoleranz**
  - System ist einsatzbereit, selbst wenn Nachrichten, Nodes oder Partitionen ausfallen

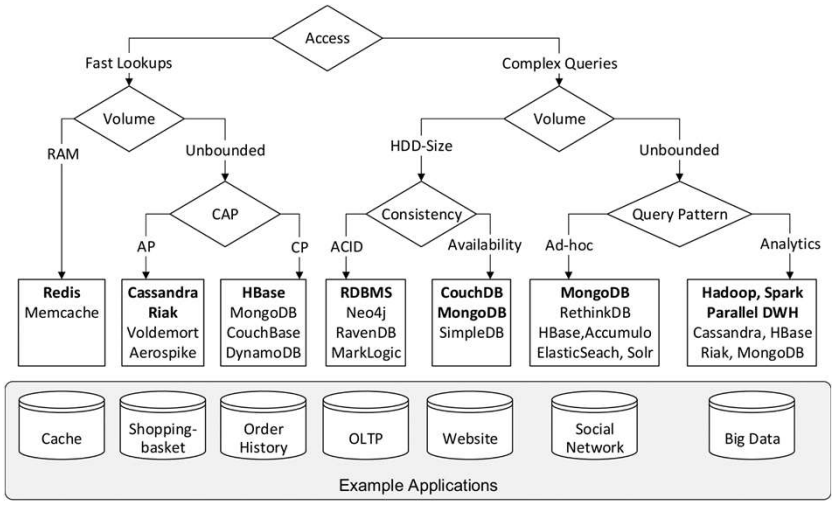
- Maximal zwei sind erfüllbar
- **Diskussion Beispiele**

=> Eventual Consistency
z. B. read your writes consistency



46

NoSQL

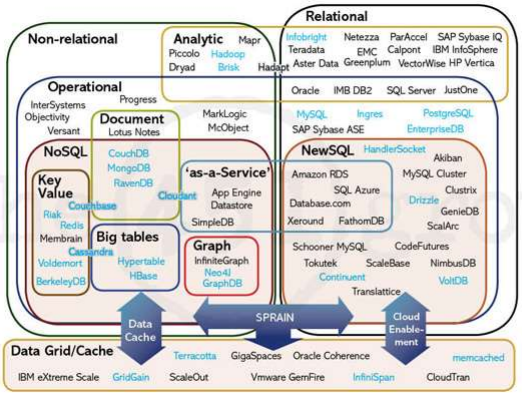


*Bildquelle: https://images.app.goo.gl/2497nX6TMEaCjXk4A, abgerufen am: 4.3.20

48

Big Data Processing Principles
Wichtige Entwicklungen

- **Eventual consistency**
- **Horizontal scalability**
- Häufig: Symmetrie
  - Alle Nodes sind gleich (**Diskussion**)
- **NoSQL – Not-only-SQL (-> NewSQL)**
  - Key Value
  - Wide Column
  - Document Oriented
  - Graph Oriented
- **Complex Event Processing**
- Mehr Funktionalität auf Applikationsebene (siehe auch ACID-Aufweichung)



Bildquelle: https://blogs.the451group.com/information_management/2011/04/15/nosql-newsql-and-beyond/

49

Anwendung der Big Data Systeme
Anforderungen und Folgen

Anforderungen an Big Data Systeme

- Fehlertoleranz
- Horizontale Skalierbarkeit
- Lesen / Schreiben mit niedriger Latenz
- Generalisierbarkeit
- Minimale Komplexität bei Umsetzung
- Ad-hoc Analyse
- …

- **CRUD vs. CR**
- **C**reate record
- **R**ead / retrieve record
- **U**pdate record
- **D**elete / destroy record

- -> **Event Sourcing**
- Zeitstempel (eine Art Event Log) für Einträge führen zu nichtveränderlichen Einträgen
- **Vorteile für verteilte Systeme?**
- **Beispiel und Nachteile**

Anfragen an Big Data Systeme
Abfragearten

- Beantworte »**einfache**« Abfrage
  - Was steht in der Datenbank aktuell?
- Beantworte »**komplexe**« Anfrage
  - Welche Anzahl / Durchschnitt / Tendenz / Veränderungen kann beobachtet werden?
  - Transformationen / Aggregationen von Daten
  - Aber: Riesige Datenmengen
  - **Ansichten vorberechnen (sog. Batch View)**

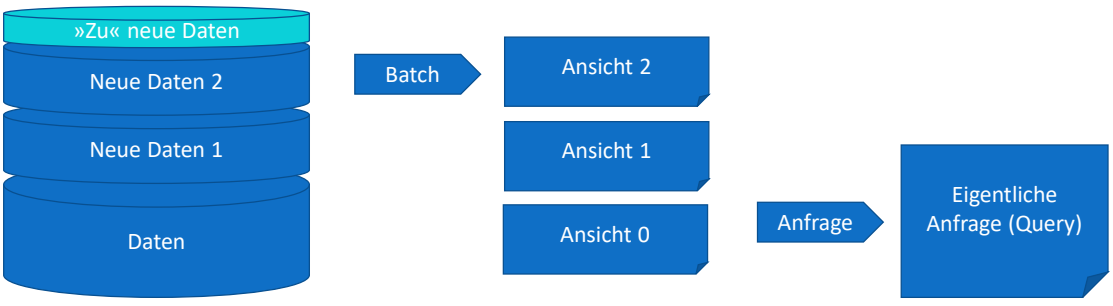
Anfragen an Big Data Systeme
Herausforderungen

1. Finde die passenden **Batch-Ansichten** für die zu Anfragen.
2. Wie berechnet man die **Ansichten**?
   MapReduce
3. Wie berechnet man die **Anfragen** auf den Ansichten?

52

Anfragen an Big Data Systeme
Herausforderungen

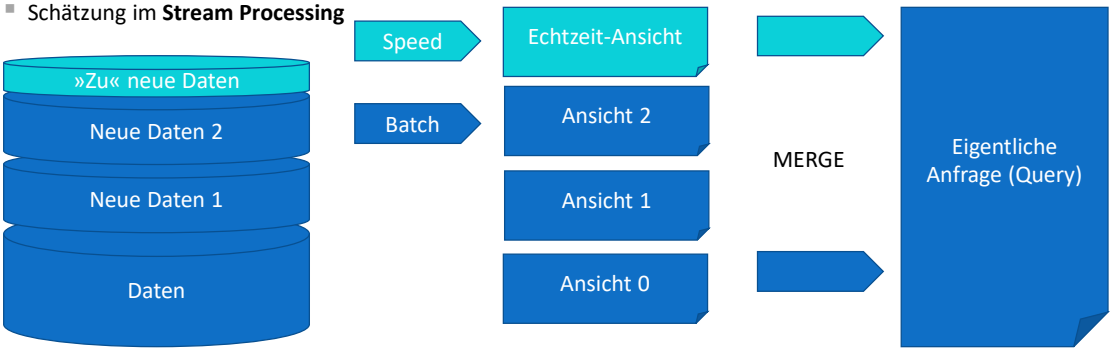
- Wechsel der Versionen leicht möglich
- Schnell verfügbar
- Skalierbar



53

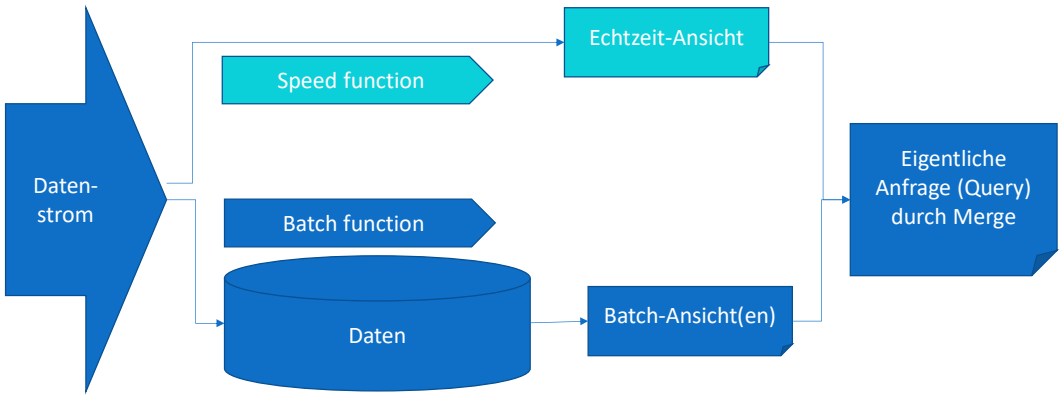
Echtzeit-Ansicht
Veraltete Ansichten durch parallelen Verarbeitungsweg vermeiden

- Eventual Accuracy
- Echtzeit-Ansicht versioniert zu Batch-Ansicht
- Schätzung im **Stream Processing**



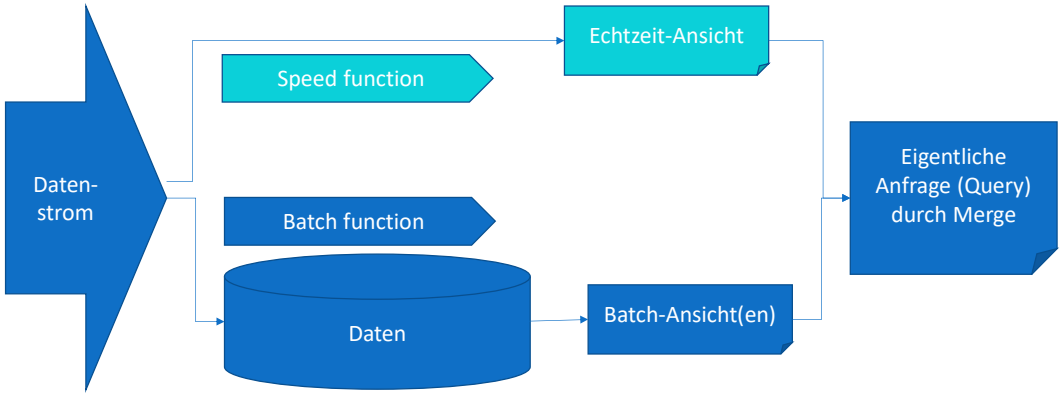
54

Lamda Architektur



55

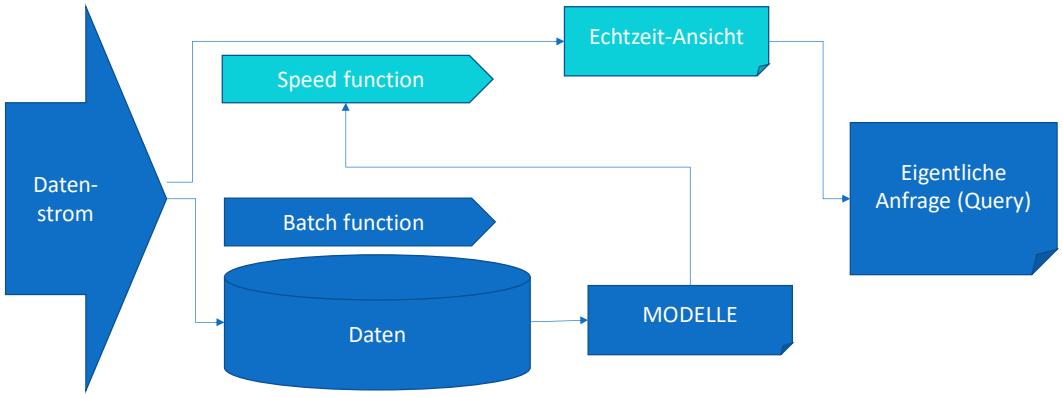
Lamda Architektur
Welche Auswirkungen auf ..?



Lesetipp: Skalierbarkeit und Architektur von Big-Data Anwendungen im Online Themenspecial IT-Trends 2014
http://www.ferari-project.eu/wp-content/uploads/2014/12/Mock_Hecker_Sylla_BigData_141.pdf

56

Lamda Architektur
Anwendung auf Data Science?



57

Exkurs: Empfehlungssysteme
Naive Ansätze

- **Welche einfachen Möglichkeiten der Empfehlung gibt es?**
- **Vor und Nachteile verschiedener Methoden**
- Beliebte Themen empfehlen
- Themen, die »ähnliche Nutzer« gewählt haben, empfehlen
- »Ähnlichkeit« zwischen Themen berechnen über die Nutzer
  - **Kollaboratives Filtern**
  - **Beispiel (siehe Bilder)**

Xavax Flexibler Röhrensiphon für Küchenspüle (Waschmaschine und… 108 EUR 14,04 prime

15 Teile GEL PINSEL SET PROFI NAILARTPINSEL MANIKÜRE UV-GEL NAGEL PINSELSET 61 EUR 1,25

Samsung Galaxy J5 2017 Panzerglasfolie, [3 Stück] FIGEMN Samsung J5 2017 Schutzfolie 5.2 Zoll… 26 EUR 8,99 prime

Soehnle 67080 Digitale Küchenwaage Page Profi 972 EUR 32,99 prime

Theo Klein 9574 - Bosch Handmixer, Spielzeug 206 EUR 8,49 prime

Bildquelle: Amazon, abgerufen am 3.4.2018 https://www.amazon.de/Bosch-MFQ40304-Handr%C3%BChrer-Styline-Colour/dp/B005J49SNW/ref=sr_1_7?ie=UTF8&qid=1522754541&sr=8-7&keywords=mixer+bosch

Empfehlungssysteme auf Basis getätigter Käufe
MapReduce für k-Means

- Beispiel: Künstler, Autoren, Lieder
- **Interaktionsmatrix**
  - Zeile: Künstler 1, 2, 3, ..
  - Spalte: Nutzer A, B, C, ..
  - Nutzer A hat Künstler 3 nicht gehört
- **Sparse** (Millionen Nutzer)
- **Jaccuard-Distanz** (Vorschau)
  - Alternativen vorhanden
- Millionen von Dimensionen
- **Ziel**: Cluster ähnlicher Künstler berechnen
- **»Trick«: Sampling**
- Weiterführende Hinweise
  - Mining of Massive Datasets, Stanford University 2013
  - Lastfm dataset

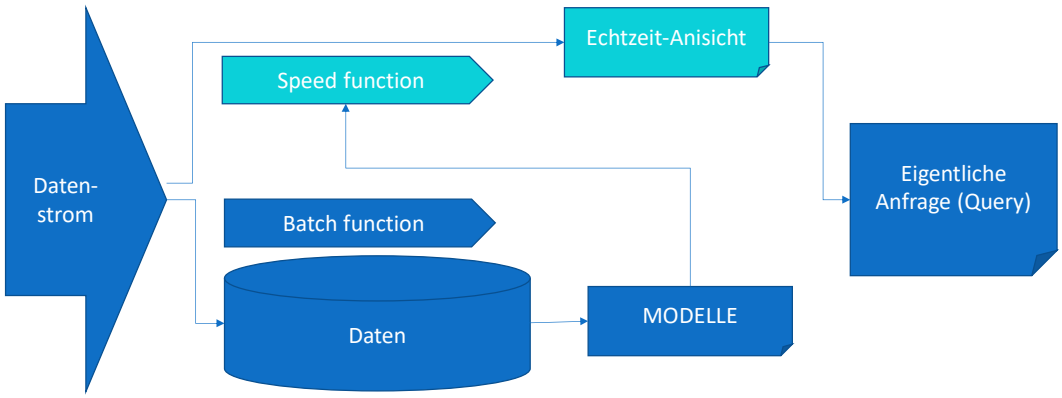
	A	B	C	D	E	F
1	1	0	0	1	1	0
2	1	0	0	1	1	1
3	0	1	1	0	1	1
4	0	0	1	1	0	0
5	1	1	0	0	0	1
6	0	1	0	1	0	1
7	0	1	1	0	0	1

Empfehlungssysteme auf Basis getätigter Käufe
MapReduce für k-Means

1. **Interaktionsmatrix** erstellen aus Rohdaten
2. **Sampling**
   - Dichte-basiertes Verfahren
   - K_Means anwenden (Cluster und Mittelpunkte = Nutzer)
3. **Validierung** der Cluster
4. Cluster auf alle Daten anwenden mit **MapReduce**
   1. Map: cluster_id als Schlüssel definieren
   2. Reduce: Gruppen von Künstlern gruppieren
   3. Clustertable: enthält Cluster und Künstlerliste
   4. Batchtable: enthält Künstler und Clusterzuordnung
5. **Validierung**
   1. Auf Basis der Verteilung der Stichprobe

60

Lamda Architektur
WIEDERHOLUNG



61

Inhalte der heutigen Vorlesung

- Unüberwachte Lernverfahren
- PCA (Dimensionsreduktion)
- K-Means (Clustering)
- Hierarchische Verfahren (Clustering)
- Einsatz von Clustering für Big Data
- Association Rule Learning (Unüb. Verfahren)

62

Hinführung

63

User ID	Spiderman	Ironman	Captain Marvel	Dr. Strange	Captain America	Hulk	Black Widow	Avengers	Winter Soldier
1	1	1	1						
2	1	1	1	1					
3		1			1	1			
4	1	1					1	1	
5			1		1			1	
6		1			1	1			
7				1	1			1	
8		1	1		1			1	
9		1			1		1	1	
10	1	1							
11			1	1					1
12			1	1					1
13						1			1
14		1							1
15		1	1						
16				1					1
17		1				1			
18		1							1
19		1							1
20	1			1	1				1

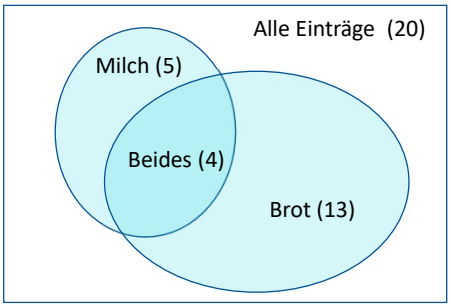
64

CART ID	MILK	BREAD	BISCUIT	CORN-FLAKES	TEA	BOURN-VITA	JAM	MAGGI	COFFEE
1	1	1	1						
2	1	1	1	1					
3		1			1	1			
4	1	1					1	1	
5			1		1			1	
6		1			1	1			
7				1	1			1	
8		1	1		1			1	
9		1			1		1	1	
10	1	1							
11			1	1					1
12			1	1					1
13						1			1
14		1							1
15		1	1						
16				1					1
17		1				1			
18		1							1
19		1							1
20	1			1	1				1

65

Support

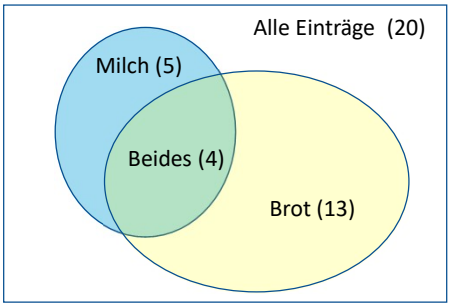
$$supp(X \cap Y) = \frac{|X \cap Y|}{N}$$



- Hier geht es um die **Häufigkeit** einer Regel
- N ist hierbei die Anzahl in der Gesamtdatenmenge (hier 20)
- Beispiel: Brot -> Milch
  - Brot kommt mit Milch 4 mal vor, daher ist der Support 4/20 = 0,2
- Diskussion

Confidence

$$confidence(X \Rightarrow Y) = \frac{supp(X \cap Y)}{supp(X)}$$



- Support siehe vorhergehende Folie
- Hier geht es darum, zu beschreiben, wie oft die Regel zutrifft, basierend auf der Anzahl wie oft **ein Teil** zutrifft (also die bedingte Wahrscheinlichkeit ausgehend von der linken Regelseite)
- Beispiele: Brot ⇒ Milch
  - Brot kommt 13 mal vor, supp(Brot) ist daher 13/20 = 0,65
  - conf(Brot ⇒ Milch) ist damit 0,2 / 0,65 = 0,31
- Beispiel 2: Milch ⇒ Brot
  - Milch kommt 5 mal vor, support ist daher 5/20 = 0,25
  - Confidence für Milch ⇒ Brot ist damit 0,2/0,25 = 0,8
- Diskussion

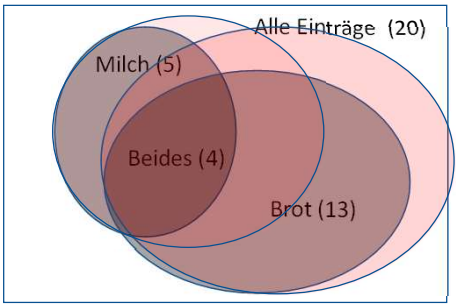
Lift

$$lift(X \Rightarrow Y) = \frac{supp(X \Rightarrow Y)}{supp(X) * supp(Y)}$$

Fall 1: Positive Korrelation wenn $lift(X \Rightarrow Y) > \mathbf{1}$

Fall 2: Negative Korrleation wenn $lift(X \Rightarrow Y) < \mathbf{1}$

Fall 3: sonst: unabhängig



- Support siehe vorhergehende Folie
- Hier geht es darum, festzustellen, ob eine Regel den Erwartungswert übertrifft basierend auf der Häufigkeit der Vorkommenden Elemente
- Beispiel: Brot ⇒ Milch
  - supp(Milch)*supp(Brot) = 0,25*0,65 = 0,16
  - lift für Brot ⇒ Milch ist damit 0,2 / 0,16 = 1,25
- Diskussion

Ein paar Beispiele zum Selbstrechnen

	Support	Confidence	Lift
Bread->Milk			
Milk->Bread			
Coffee->Bread			
Bread->Coffee			
Bread, Jam -> Maggi			

ID	MILK	BREAD	BISCUIT	CORNFLAKES	TEA	BOURNVITA	JAM	MAGGI	COFFEE
1	1	1	1						
2	1	1	1	1					
3		1			1	1			
4	1	1					1	1	
5			1		1			1	
6		1			1	1			
7				1	1			1	
8		1	1		1			1	
9		1			1		1	1	
10	1	1							
11			1	1					1
12			1	1					1
13						1			1
14		1							1
15		1	1						
16				1					1
17		1				1			
18		1							1
19		1							1
20	1			1	1				1

71

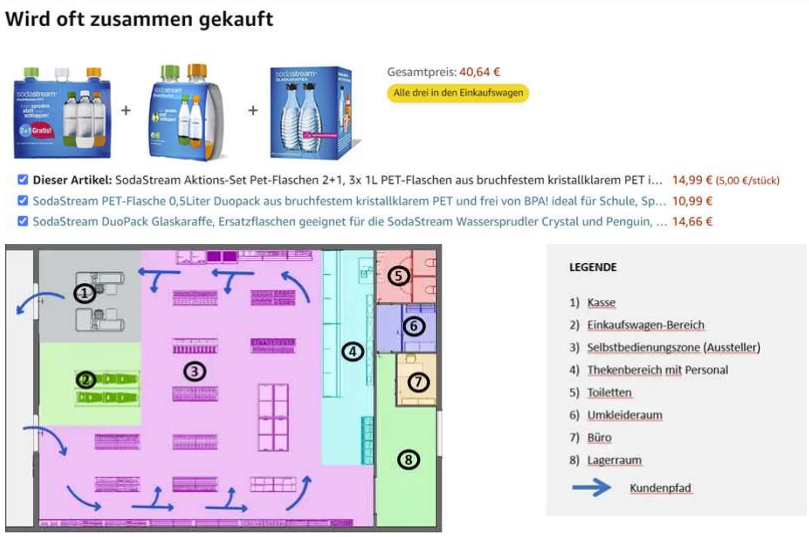
A-Priori-Algorithmus

- **Naiver Ansatz**
  - Wähle alle möglichen linken Seiten
    - Wähle alle möglichen rechten Seiten
      - Berechne die Werte
- **A-Priori-Algorithmus**
  - Alle Werte zu berechnen ist zu teuer!
  - Nutze den Support, um vorzuselektieren
- Alternativen
  - **FP-growth algorithm**
  - Nutzt eine Baum-ähnliche Struktur für effizientere Verwaltung
  - Weitere (out of scope)

```
antecedents                consequents  \
    (BREAD)                    (MILK )
    (BREAD)                    (SUGAR)
    (BREAD)                  (BISCUIT)
    (BREAD)                      (TEA)
    (BREAD)                (BOURNVITA)
    (BREAD)                    (MAGGI)
    (BREAD)                   (COFFEE)
    (BREAD)                      (JAM)
    (BREAD)           (MILK , BISCUIT)
    (BREAD)           (BOURNVITA, TEA)
    (BREAD)               (JAM, MAGGI)
    (BREAD)               (MAGGI, TEA)
```


72

Gedanken zur Anwendung



Bildquellen: amazon.de, https://biblus.accasoftware.com/de/wie-man-einen-supermarkt-plant-der-technische-leitfaden/

74

Literatur zum Thema

- *Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms" (PDF). Introduction to Data Mining. Addison-Wesley. ISBN 978-0-321-32136-7.*
- R. Agrawal, T. Imieliński, A. swami: *Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93*. In: *Mining association rules between sets of items in large databases*. 1993, S. 207, doi:10.1145/170035.170072.
- Data Mining: Concepts and Techniques Jiawei Han (Author), Jian Pei (Author), Hanghang Tong (Author) Morgan Kaufmann; 4. Edition (17. Oktober 2022)
- Data Science for Business: What you need to know about Data Mining and Data Analytic Thinking; von Foster Provost (Author), Tom Fawcett (Author) O'Reilly Media; 1. Edition (17. September 2013)
- Data Mining: The Textbook Charu C. Aggarwal (Author) Springer; 2015. Edition (27. April 2015)

75

Und die nächste Stunden sehen Sie..

- Zeitreihenanalysen
- Explainable AI

Literaturliste

- [James et al. 2013] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani: An introduction to statistical learning
  - Favorit: Sehr gut gemachte Einführung, jedoch Beispiele in R, verständlich mit Mathematik, als pdf frei erhältlich
- [Hastie et al. 2008] Trevor Hastie, Robert Tibshirani, Jerome Friedman: The elements of statistical learning
  - DIE Referenz, für Mathematiker geschrieben, als pdf frei erhältlich
- [O'Neil and Schutt 2013] Cathy O'Neil and Rachel Schutt: Doing Data Science
  - Spannend zu lesen, teilweise Erfahrungsberichte (durch Drittautoren)
- [Mueller and Guido 2017] Andreas C. Müller & Sasha Guido: An Introduction to Machine Learning with Python
  - Interessant da Python 3 tatsächlich genutzt wird für die Einführung inklusive der üblichen Bibliotheken
- [Grues 2016] Joel Grues (übersetzt von Kristian Rother): Einführung in Data Science
  - Auf deutsch gut übersetzt, nutzt Python für grundlegendes Verständnis ohne die üblichen Bibliotheken, extrem leicht lesbar
- [Alpaydin 2008]: Ethem Alpaydin (übersetzt von Simone linke): Maschinelles Lernen
  - Auf deutsch gut übersetzt, relativ viel Mathematik, in Deutschland scheint das weit verbreitet zu sein
- [Bruce et al. 2020]: Peter Bruce, Andrew Bruce, Peter Gedeck: Practical Statistics for Data Scientists
  - Das einzig wahre Statistikbuch was keines ist
- [Reinhart 2016]: Alex Reinhart (übersetzt von Knut Lorenzen): Statistics done wrong
  - Bevor man wirklich Konfidenzintervalle oder p-Werte angibt und über „Signifikanz" spricht, sollte man das gelesen haben

Literaturliste contd.

- Online-Ressource zu Visualisierung
  - https://www.visualisingdata.com/
- Storytelling with Data [Buch]: Klassiker für Überzeugungsarbeit in Präsentationen von Ergebnissen
  - http://www.bdbanalytics.ir/media/1123/storytelling-with-data-cole-nussbaumer-knaflic.pdf
- Show Me the Numbers [Buch]: Ganz konkrete Tipps für die Praxis
  - https://courses.washington.edu/info424/2007/readings/Show_Me_the_Numbers_v2.pdf
- Now you see it [Buch]: Ebenfalls ganz konkrete Inhalte