Datenbanken und Informationssysteme

VL 08, Data Mining

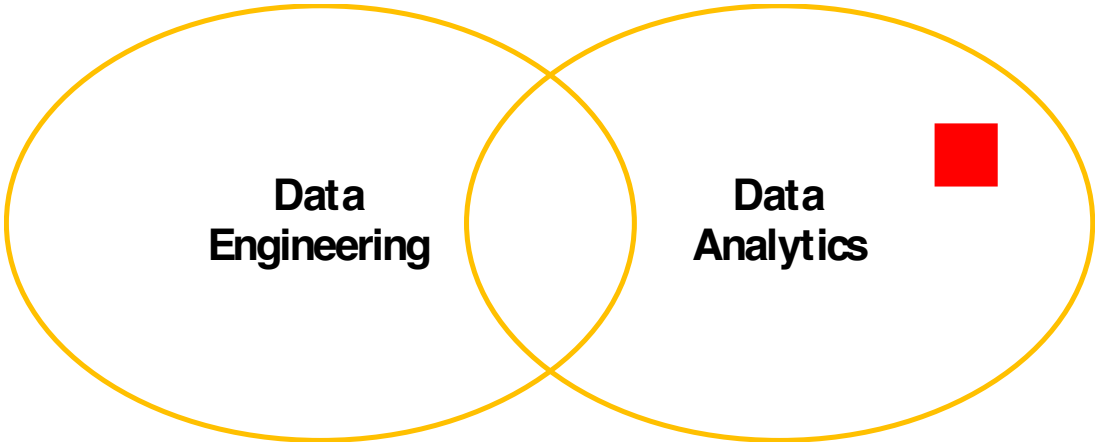
SoSe 2022

Univ.-Prof. Dr.–Ing. habil. Norbert Gronau
Lehrstuhlinhaber | Chairholder

Mail August-Bebel-Str. 89 | 14482 Potsdam | Germany
Visitors Digitalvilla am Hedy-Lamarr-Platz, 14482 Potsdam
Tel +49 331 977 3322

E-Mail ngronau@lswi.de
Web lswi.de

Kapitel 8: Data Mining



Systematische Auswertung von Daten (1854)

ON THE COMMUNICATION OF CHOLERA BY IMPURE THAMES WATER.

By JOHN SNOW, M.D.

DISTRICTS AND SUB-DISTRICTS.	Population in 1851.	Deaths from Cholera in the Seven Weeks ending August 26.	Supply of Water in the House of Attack.				
			Southwark and Vauxhall.	Lambeth.	Pump-wells & Springs.	River Thames, Ditches, &c.	Not ascertained.
ST. SAVIOUR, SOUTHWARK.							
Christchurch	16,022	23	11	13	..	..	1
*St. Saviour	10,700	125	115	..	..	10	..
ST. OLAVE, SOUTHWARK.							
*St. Olave	8,015	53	44	..	..	3	6
*St. John, Horsleydown ..	11,360	51	46	..	..	3	2
BERMONDSEY.							
*St. James	15,359	123	102	..	..	21	..
*St. Mary Magdalen ..	13,934	87	83	..	..	4	..
*Leather Market	15,295	81	81	..	..	..	..
ST. GEORGE, SOUTHWARK.							
Kent-road	18,126	57	52	5	..	..	..
Borough-road	15,862	71	61	7	..	..	3
London-road	17,836	29	21	8	..	..	..
NEWINGTON.							
Trinity	20,922	53	52	0	..	..	..
St. Peter, Walworth ..	29,861	90	84	4	..	..	2
St. Mary	14,033	21	19	1	1	..	..
LAMBETH.							
Waterloo-road, 1st. ..	14,088	10	8	2	..	..	..
Waterloo-road, 2nd. ..	18,348	36	25	8	1	2	..
Lambeth Church, 1st. ..	18,400	18	6	9	..	1	2
Lambeth Church, 2nd. ..	26,784	53	34	13	1	..	5
Kennington, 1st.	24,261	71	63	5	3	..	..
Kennington, 2nd.. ..	18,848	38	34	3	1	..	..
Brixton	14,610	9	5	2	..	..	2
Norwood	3,977	8	..	2	1	5	..
WANDSWORTH.							
*Clapham	16,290	24	19	..	5	..	..
*Battersea	10,560	54	36	..	4	8	6
*Wandsworth	9,611	11	3	..	2	6	..
Putney	5,280	1	..	..	..	..	1
Streatham	9,023	6	..	1	5	..	..
CAMBERWELL.							
Dulwich	1,632	..	..	..	..	..	..
*Camberwell	17,742	96	72	..	24	..	..
*Peckham	19,444	59	45	..	..	..	14
St. George	15,849	42	34	4	..	..	4
ROTHERHITHE.							
*Rotherhithe	17,805	103	69	..	..	34	..
Total	432,435	1,510	1,224	93	48	97	48

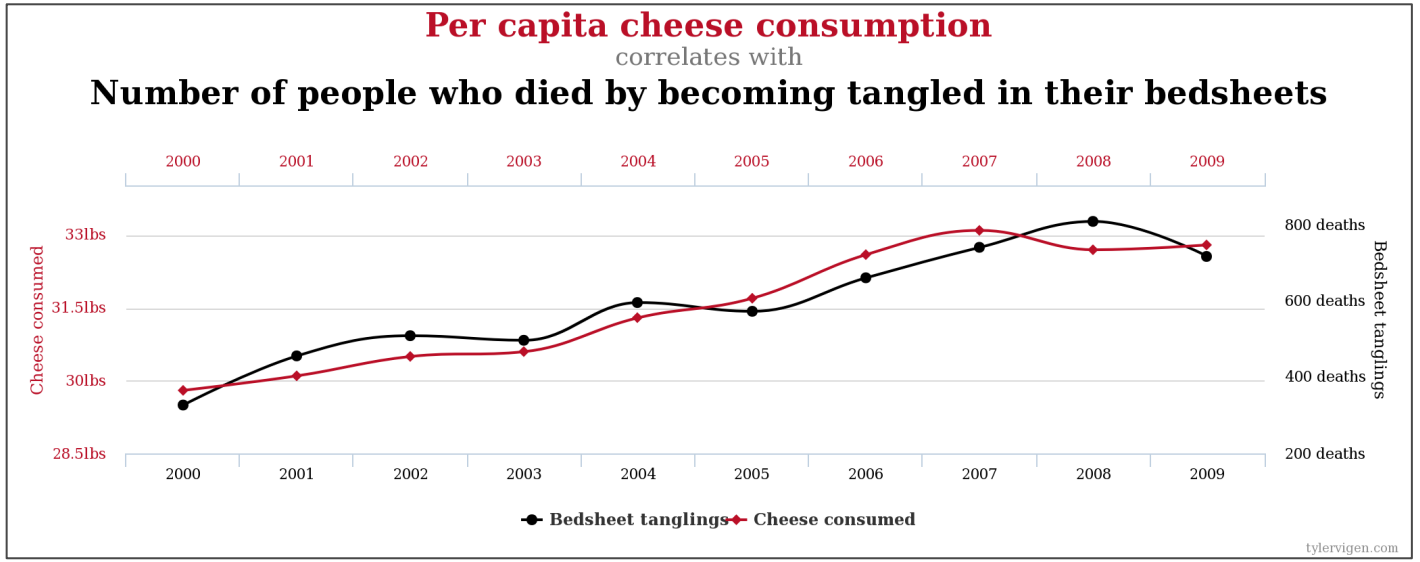
Bedeutung von Zusammenhängen

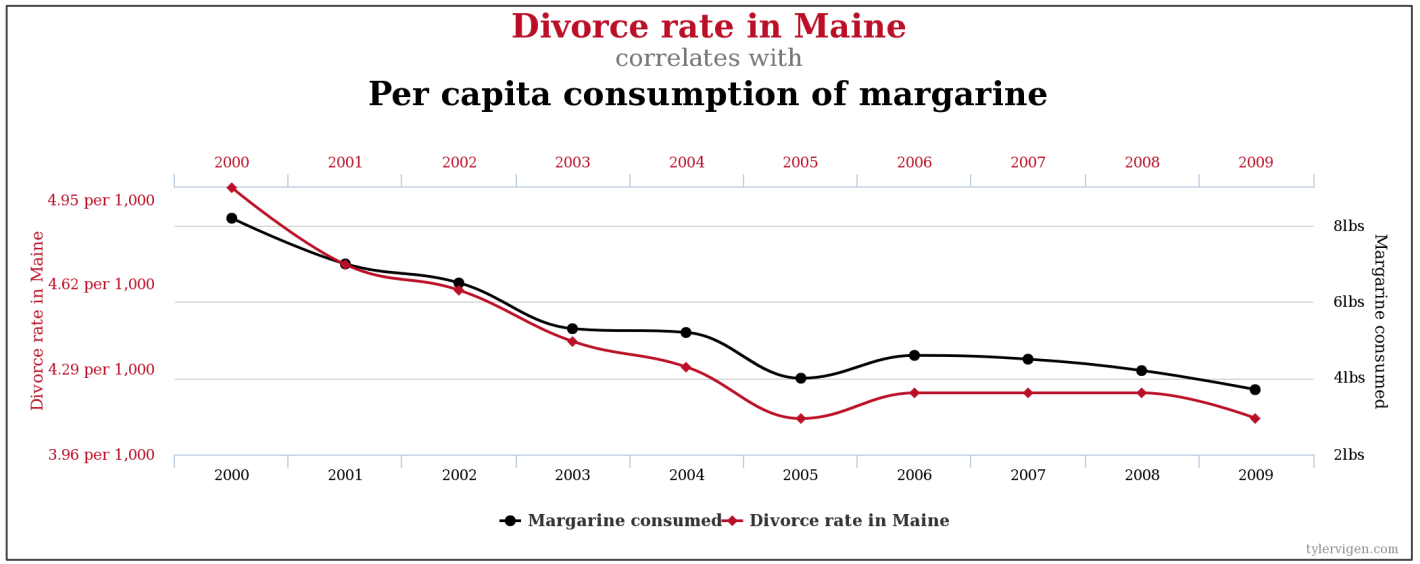
- Kausalität
  - > Ursache-Wirkungs-Beziehung zwischen zwei Variablen

- Korrelation
  - > Statistischer Zusammenhang zwischen zwei Variablen

- Korrelation kann auf Kausalität beruhen
  - > Beispiele: Ursachen für die Korrelation von X und Y
    - » $X \rightarrow Y$ (Kausalität)
    - » $Y \rightarrow X$ (Kausalität)
    - » $Z \rightarrow X$ *und* $Z \rightarrow Y$ (Scheinkorrelation)
    - » X und Y korrelieren ohne Ursache (Unsinnskorrelation)

Beispiel: Unsinnskorrelationen



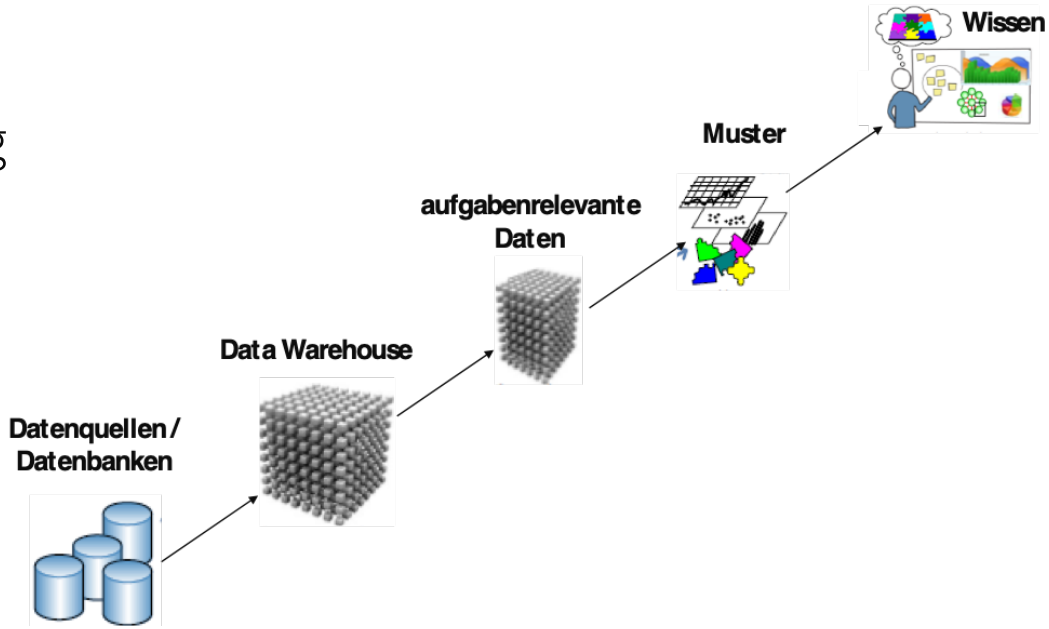


[Vige16]

Data Mining

- Definition
  - > Extraktion von interessanten (nicht-trivialen, impliziten, zuvor unbekannten und potentiell nützlichen) Informationen oder Mustern aus großen Datenmengen in großen Datenbanken
  - > Data Science ist am engsten mit Data Mining verbunden

- Ziele
  - > Wissensgenerierung
    - » Deskription
    - » Prädiktion
    - » Präskription

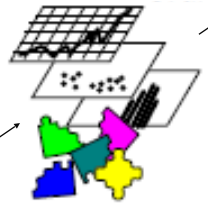


[FaPS96] [HaKa06]

Data-Mining-Prozess

Wissen

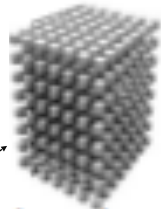
Muster



Visualisierung & Evaluation

- Analyseauswertung und Wissensrepräsentation
- Datenvisualisierung mit Diagrammen und Charts

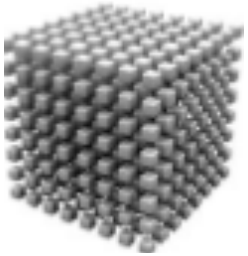
aufgabenrelevante Daten



Data Mining

- Suche nach Mustern mit Data-Mining-Paradigmen und -modellen

Data Warehouse



Transformation, Selektion & Projektion

- Dimensionalitätsreduktion

Datenquellen / Datenbanken

Datenbereinigung & -integration

- Integration der Daten aus verschiedenen Quellen
- Beseitigung von Inkonsistenzen, Rauschen
- Berechnung von fehlenden Werten

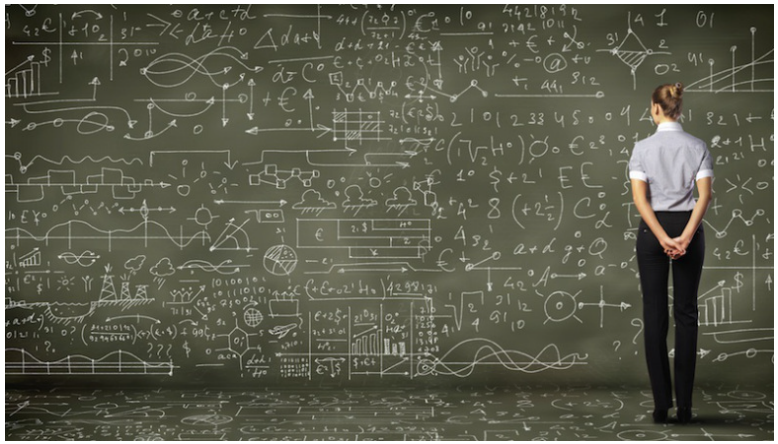
[FaPS96]
[Muel15]

Anwendungsbeispiele

- LinkedIn findet neue Ideen für Produkte und Features
- Google verfeinert die Such- und Werbealgorithmen
- GE optimiert Serviceverträge und Wartungsintervalle für industrielle Produkte
- Zynga optimiert Spielerfahrung
- Netflix verbessert das Filmempfehlungssystem
- Kaplan entdeckt effektive Lernstrategien

Data Analytics

- Hintergrund
  - > Explosion der Datenmengen und der Analysepotentiale
  - > steigende Rechen- und Speicherkapazitäten



"We are drowning in data, but starving for knowledge!"

Data Science

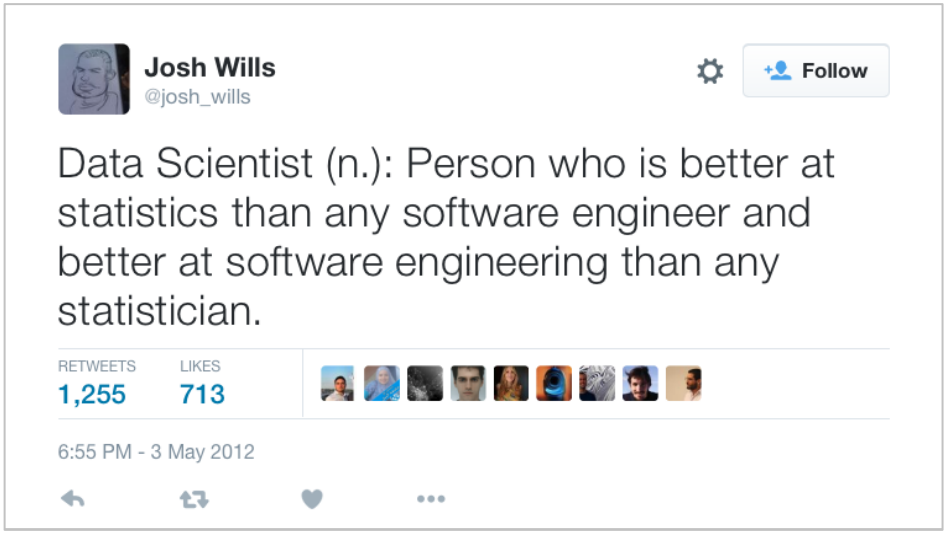
- Definition
  - > Generierung von Wissen und Erkenntnissen aus Daten durch Anwendung statistischer Methoden, Modelle und Algorithmen
- Methoden
  - > z.B. Induktive Statistik, Data Mining, Big Data Analytics
- Zunehmend Analyse heterogener, unstrukturierter Daten
  - > Text, Bilder, Videos – oft aus Netzwerken, die die Beziehungen der Beteiligten untereinander speichern und analysieren
  - > z.B. Facebook: Täglich 2,5 Mrd. "Likes", 300 Mio. Foto-Uploads

"Data science... can see things as never before, but also can be a blunt instrument, missing context and nuance."

Clayton Christensen

Data Scientist

- Aufgaben
  - > Identifikation und Kombination interessanter, möglicherweise unvollständiger Datenquellen
  - > Bereinigung und Strukturierung großer Mengen formloser Daten
  - > Ermöglichung von Analysen
  - > Visuelle Darstellung der Erkenntnisse und Aufzeigen von Mustern
  - > Beratung von Führungskräften



Data-Mining-Paradigmen

- Standardtypen
  - > Clustering (Gruppenanalyse)
  - > Klassifikation
  - > Assoziationsanalyse (Frequent Pattern Mining)
  - > Regression
- Weitere Typen
  - > Ausreißeranalyse (Abweichungsanalyse)
  - > Sequentielle Musteranalyse
  - > Trendanalyse
- Methoden für spezielle Datentypen
  - > Spatial Data Mining (z.B. geographische Datenanalyse)
  - > Web Mining (z.B. Text-Analyse)
  - > Graph Mining (z.B. soziale Netzwerkanalyse)
  - > Process Mining

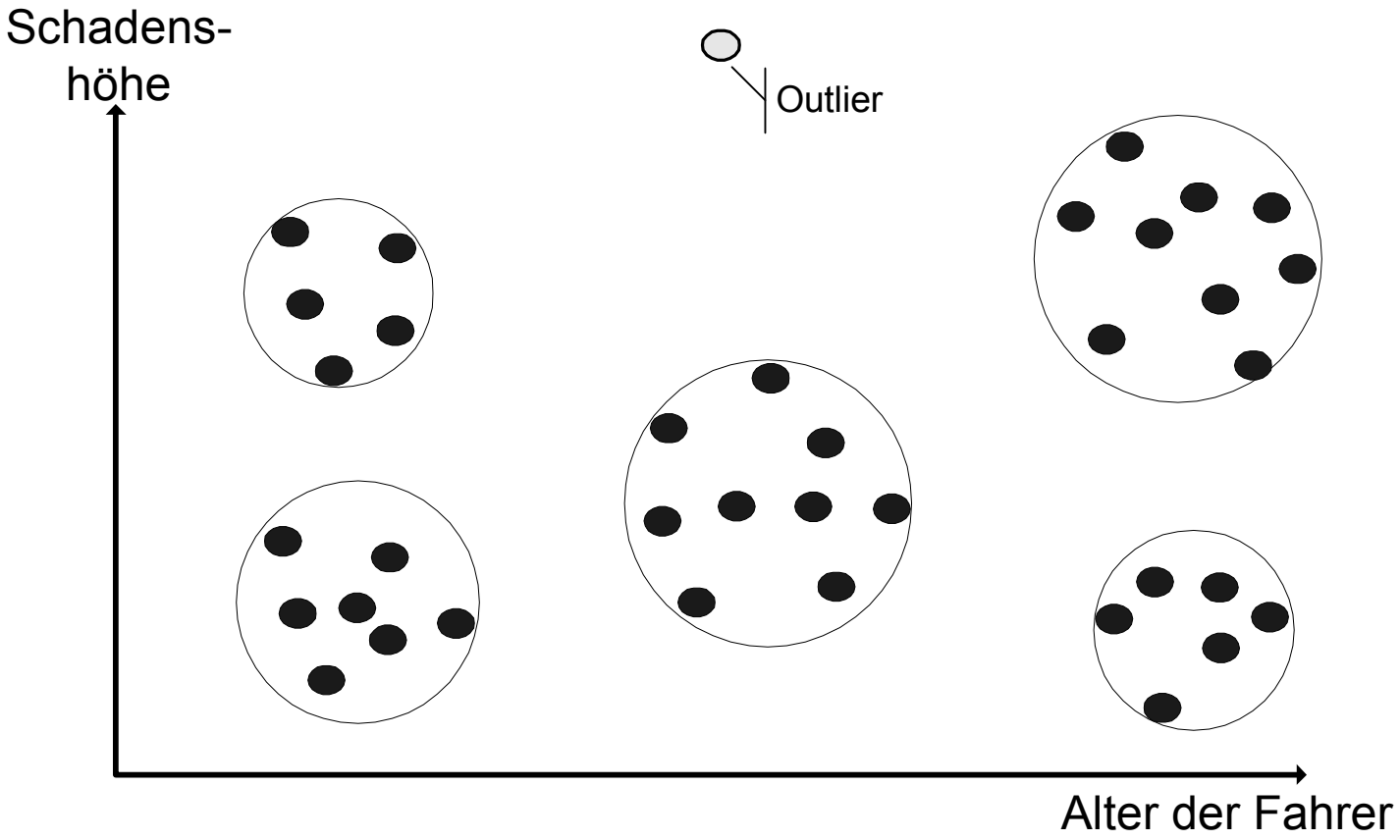
Clustering

- Voraussetzung und Vorgehen
  - > Klassenetiketten sind unbekannt (= unsupervised learning)
  - > Aggregation von Objekten zu Untergruppen durch Messung der (Un-)Ähnlichkeit zwischen Objekten
    - » Ähnlichkeitsfunktion bzw. Unähnlichkeitsfunktion (= Distanz)
- Konkurrierende Ziele
  - > Maximierung der Ähnlichkeit von Objekten innerhalb einer Klasse
  - > Minimierung der Ähnlichkeit zwischen Objekten einer Klasse
- Anwendungsbeispiele
  - > Kundenprofilierung/Segmentierung
  - > Dokument- oder Bildsammlungen
  - > Webzugriffsmuster

Clustering-Verfahren

- Hierarchische Clusterverfahren
  - > Agglomerative Verfahren
    - » ausgehend von einzelnen Datensätzen
    - » sukzessive Zusammenführung von jeweils zwei Clustern, bis alle Datensätze in einem einzigen Cluster enthalten sind
  - > Divisive Verfahren
    - » ausgehend von einem einzigen Cluster
    - » sukzessive Unterteilung von Clustern, bis jedes Cluster aus einem einzelnen Datensatz besteht
    - » z.B. DBSCAN
- Partionierende Verfahren
  - > Bildung einer optimalen (lokalen) Clusterung für eine festgelegte Anzahl an Clustern
  - > z.B. K-means

Beispiel: Clustering



[KeEi15]

Clustering: K-means

- Ziel: Minimierung der Summe der Abstände Q der Datenpunkte x zum Mittelpunkt des Clusters

$$Q = \sum_{i=1}^{k} \sum_{x_j \in S_i} ||x_j - \mu_i||^2$$

1. **Initialisierung:** Wähle zufällig k Mittelwerte $\mu_1^{(1)}, \ldots, \mu_k^{(1)}$ aus dem Datensatz aus. Diese bilden die initialen Mittelwerte der Cluster der ersten Iteration.
2. **Iteriere für** $t = 1$ **bis** T (oder solange sich noch Änderungen an den Clustern ergeben):
   (a) **Zuordnung zu Clustern:** Jedes Datenobjekt x_j wird jetzt dem Cluster $S_i^{(t)}$ zugeordnet, zu dessen Mittelwert $\mu_i^{(t)}$ es am nächsten liegt. Das heisst,

   nach diesen Zuordnungen enthält jedes Cluster $S_i^{(t)}$ für $1 \leq i \leq k$ nur Datenelemente, für die gilt:

   $$S_i^{(t)} = \{x_j : ||x_j - \mu_i^{(t)}||^2 \leq ||x_j - \mu_{i'}^{(t)}||^2 \text{ für alle } i' = 1, \ldots, k\}$$

   Falls ein Datenelement x_j sich für mehrere Cluster qualifiziert, wird es dennoch nur einem (bspw. zufällig gewählten) Cluster $S_i^{(t)}$ zugewiesen.
   (b) **Neuberechnung der Mittelwerte:** Jetzt werden für alle im t-ten Schritt berechneten Cluster die neuen Mittelwerte $\mu_i^{(t+1)}$ berechnet:

   $$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Klassifikation

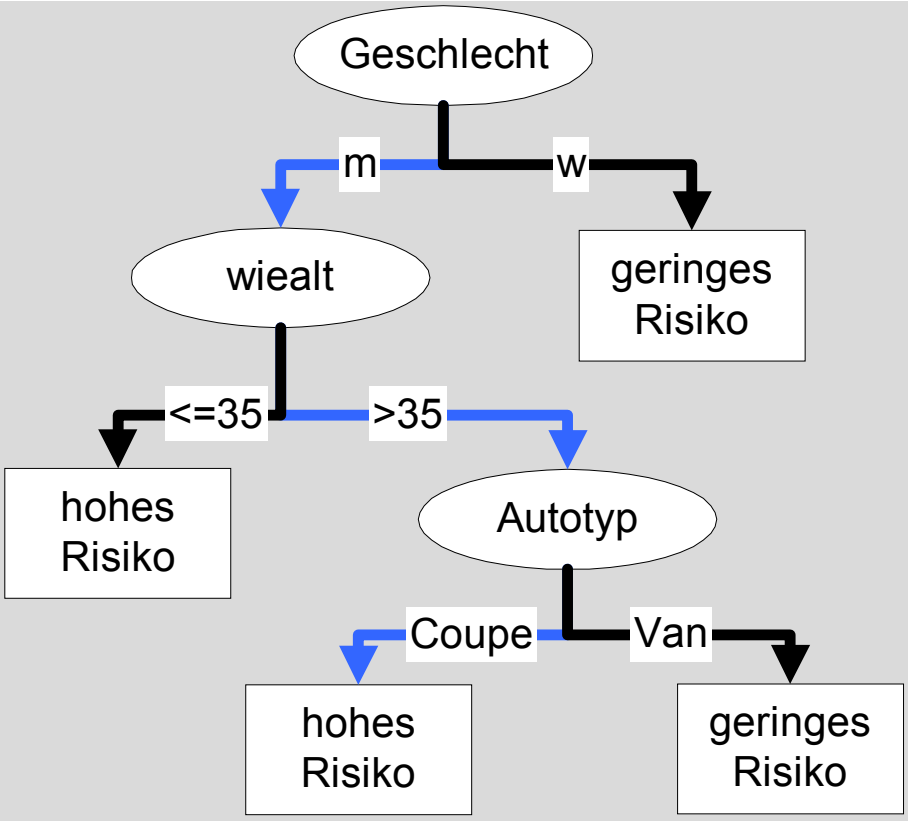
- Voraussetzung und Vorgehen
  - > bekannte Klassenetiketten für geringe Menge an Trainingsdaten
  - > Suche von Modellen, Funktionen und/oder Regeln zur Einord-nung der Klassenmitgliedschaft neuer Objekte
  - > basierend auf den Attributwerten der Trainingsbeispiele
- Verfahren
  - > z.B. Nearest Neighbor, Logit, Bagging
- Anwendungsbeispiele
  - > Klassifikation von potentiellen Kreditnehmern
  - > Klassifikation von Genexpressionswerten für Gewebeproben zur Vorhersage von Krankheitstypen und zur Empfehlung geeigneter Therapien
  - > Vorhersage unbekannter oder fehlender Werte als Data-Mining-Vorverarbeitungsschritt

Entscheidungs-/Klassifikationsbäume

- Trainingsdatenmenge
  - > große Zahl von Vergangenheitsdaten
  - > Grundlage für Prognose bzw. Einordnung neuer Objekte
  - > Beispiel: neuer Versicherungskunde wird gemäß dem Verhalten ähnlicher Kunden eingestuft
- Rekursives Partitionieren
  - > Fange mit einem Attribut an und spalte die Tupelmenge
  - > Jede dieser Teilmengen wird rekursiv weiter partitioniert, bis nur noch gleichartige Objekte in der jeweiligen Partition sind

Beispiel: Klassifikations-/Entscheidungsbaum

Schadenshöhe			
wiealt	Geschlecht	Autotyp	Schäden
45	w	Van	gering
18	w	Coupé	gering
22	w	Van	gering
38	w	Coupé	gering
19	m	Coupé	hoch
24	m	Van	hoch
40	m	Coupé	hoch
40	m	Van	gering
⋮	⋮	⋮	⋮



> (wieAlt > 35) ∧ (Geschlecht ="m") ∧ (Autotyp ="Coupé") ➔ (Risiko = "hoch")

[KeEi15]

Assoziationsanalyse

- Voraussetzung und Vorgehen
  - > Entdeckung von Abhängigkeiten zwischen Attributen oder einzelnen Attributsausprägungen innerhalb eines Datenbestandes mittels Assoziationsanalyse
  - > Beschreibung durch Regeln
    - » z.B. "Wenn X vorkommt, tritt auch Y auf"
    - » zur Bewertung einer Regel können unterschiedliche Interessantheitsmaße herangezogen werden, die auch zur Regelgenerierung genutzt werden können
- Verfahren
  - > z.B. Apriori-Algorithmus

Beispiel: Warenkorbanalyse

- Wo sollen Hygiene-artikel im Geschäft platziert werden, um den Umsatz zu verbessern?
- Werden Fenster-putzmittel gekauft, wenn Orangensaft und Hygieneartikel gekauft werden?
- Wird Saft typischer-weise zusammen mit Bananen gekauft?
- Wie beeinflusst die Umgebung, was im Laden gekauft wird?

Assoziationsregeln

- Beschreibung von Zusammenhängen/Regelmäßigkeiten
  - > Festlegung von Voraussetzungen und Regeln
  - > z.B. kauft jemand einen PC, dann kauft er auch einen Drucker
- Confidence-Wert
  - > prozentualer Anteil an Übereinstimmung von Voraussetzungs- und Regelerfüllung in der verfügbaren Datenmenge
  - > ein Wert von 80 % für die Beispielregel besagt, dass 4/5 der Leute, die einen PC kauften, auch einen Drucker kauften
- Support-Wert
  - > prozentualer Anteil an Datensätzen, überhaupt gefunden wurden, um die Gültigkeit der Regel zu verifizieren
  - > Bei einem Wert von 1 % wäre also jeder hundertste Verkauf ein "PC+Drucker"-Verkauf

Verkaufstransaktionen Warenkörbe

- Finde alle Assoziationsregeln L → R
  - > mit Support-Wert > minsupp und
  - > Confidence-Wert ≥ minconf
- Suche der sog. *Frequent Itemsets (FI)*
  - > Produktmengen, die mindestens den minsupp-Wert in den Einkaufswägen/ Transaktionen erreichen
- Der Apriori-Algorithmus basiert auf der Grundlage, dass alle Teilmengen eines FI ebenfalls FIs sein müssen

VerkaufsTransaktionen	
TransID	**Produkt**
111	Drucker
111	Papier
111	PC
111	Toner
222	PC
222	Scanner
333	Drucker
333	Papier
333	Toner
444	Drucker
444	PC
555	Drucker
555	Papier
555	PC
555	Scanner
555	Toner

[KeEi15]

Vorgehen beim Apriori-Algorithmus

- Überprüfe für jedes Produkt, ob es ein Frequent Itemset ist
  - > d.h. das Produkt ist mindestens minsupp Mal in Einkaufswägen enthalten
  - > k:=1
- Iteriere so lange für jedes FI I_k mit k Produkten, bis keine neuen FIs mehr gefunden werden
  - > generiere alle FIs I_{k+1} (mit k+1 Produkten und $I_k \subset I_{k+1}$
  - > lies alle Einkäufe ein Mal (sequentieller Scan auf der Datenbank)
  - > überprüfe, welche der (k+1)-elementigen FI-Kandidaten mindestens minsupp Mal vorkommen
  - > k:=k+1

Apriori-Algorithmus

VerkaufsTransaktionen	
TransID	**Produkt**
111	Drucker
111	Papier
111	PC
111	Toner
222	PC
222	Scanner
333	Drucker
333	Papier
333	Toner
444	Drucker
444	PC
555	Drucker
555	Papier
555	PC
555	Scanner
555	Toner

Minsupp=3

Disqua-
lifiziert

Zwischenergebnisse	
FI-Kandidat	Anzahl
{Drucker}	4
{Papier}	3
{PC}	4
{Scanner}	2
~~{Toner}~~	~~3~~
{Drucker, Papier}	3
{Drucker, PC}	3
{Drucker, Scanner}	
{Drucker, Toner}	3
{Papier, PC}	2
{Papier, Scanner}	
{Papier, Toner}	3
{PC, Scanner}	
{PC, Toner}	2
{Scanner, Toner}	[KeEi15]

Apriori-Algorithmus

VerkaufsTransaktionen	
TransID	**Produkt**
111	Drucker
111	Papier
111	PC
111	Toner
222	PC
222	Scanner
333	Drucker
333	Papier
333	Toner
444	Drucker
444	PC
555	Drucker
555	Papier
555	PC
555	Scanner
555	Toner

Zwischenergebnisse	
Fl-Kandidat	Anzahl
{Drucker, Papier}	3
{Drucker, PC}	3
{Drucker, Scanner}	
{Drucker, Toner}	3
{Papier, PC}	2
{Papier, Scanner}	
{Papier, Toner}	3
{PC, Scanner}	
{PC, Toner}	2
{Scanner, Toner}	
{Drucker, Papier, PC}	2
{Drucker, Papier, Toner}	3
{Drucker, PC, Toner}	2
{Papier, PC, Toner}	2

[KeEi15]

Ableitung von Assoziationsregeln aus den FIs

- Betrachte jeden FI mit hinreichendem Support-Wert
- Bilde alle nicht-leeren Teilmengen L ⊂ FI und untersuche die Regel
  - > L → FI – L
  - > Confidence-Wert dieser Regel berechnet sich als
    - » Confidence(L → FI – L) = support(FI) / support(L)
    - » Wenn Confidence ausreicht (d.h. > minconf), behalte diese Regel
- Betrachte FI = {Drucker, Papier, Toner}
  - > Support = 3
- Regel: {Drucker} → {Papier, Toner}
  - > Confidence = S({Drucker, Papier, Toner}) / S({Drucker})
    = (3/5) / (4/5) = 3/4 = 75 %

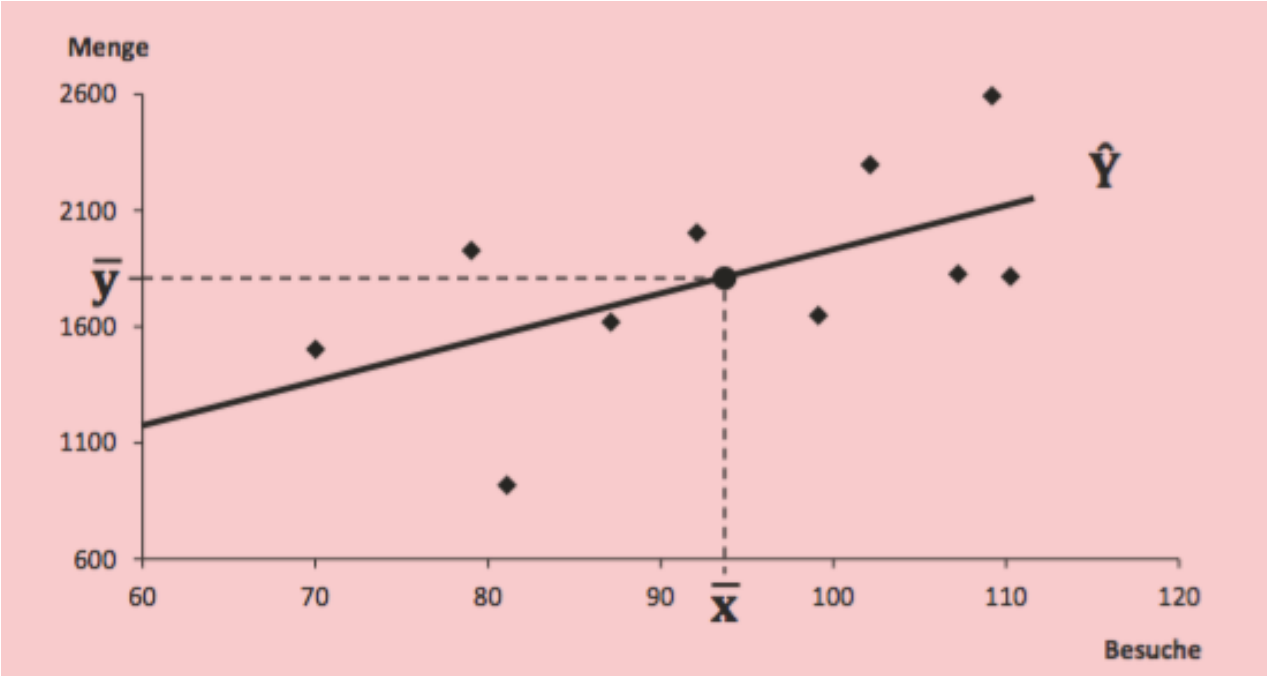
Zusammenhangsmaße

- Varianz
  - > Durchschnittliche quadrierte Abweichung vom Mittelwert
    - » Varianz: $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$
    - » Standardabweichung = $\sqrt{s^2}$
- Kovarianz
  - > Maß des linearen Zusammenhangs zweier Zufallsvariablen (gemeinsame Wahrscheinlichkeitsverteilung)
  - > Korrelationskoeffizient
    - » normiertes Maß zur Angabe der Beziehungsstärke und -richtung
    - » Wertebereich im Intervall [-1,+1]
- Multikollinearität
  - > Starke Korrelation zwischen mehreren *exogenen* Variablen

[BEPW16]

Einfache Regression

- Ermittlung einer Regressionsgeraden, die sich einer empirischen Punkteverteilung (X,Y) möglichst gut anpasst und die Abweichungen minimiert
  - > $\hat{Y} = b_0 + b_1 X$
  - > z.B.: Verkaufte Güter (Y) und Anzahl der Kundenbesuche (X)



[BEPW16]

Wie funktioniert eine Regression rechnerisch?

- Methode der kleinsten Fehlerquadrate
  - > engl.: *Least Squares Method*
  - > Optimierung der Regressionsfunktion erfolgt durch Minimierung der Summe der quadrierten Residuen

$$\sum_{n=1}^{N} e_n^2 = \sum_{n=1}^{N} [y_n - (b_0 + b_1 x_n)]^2 \rightarrow min!$$

- Regressionskoeffizient
  - > Maß für die Stärke der Wirkung von X auf Y
  - > Koeffizient der unabhängigen Variable

$$b_1 = \frac{N(\sum x_n y_n) - (\sum x_n)(\sum y_n)}{N(\sum x_n^2) - (\sum x_n)^2}$$

[BEPW16]

Literaturverzeichnis

Literaturverzeichnis

[Bapp14] Bappalige, O. (2014). Introduction to Apache Hadoop. URL: https://opensource.com/life/14/8/intro-apache-hadoop-big-data

[BEPW16] Backhaus, K., Erichson, B., Plinke, W., Weiber, R. (2015). Multivariate Analysemethoden: Eine anwendungsorientierte Einführung. Springer-Verlag.

[Dave14] Davenport, T. (2014). Big data at work: dispelling the myths, uncovering the opportunities. Harvard Business Review Press.

[EFH+11] Edlich, S., Friedland, A., Hampe, J., Brauer, B., Brückner, M. (2011). NoSQL: Einstieg in die Welt nichtrelationaler Web 2.0 Datenbanken. Hanser.

[FaEg05] Fassott, G., Eggert, A. (2005). Zur Verwendung formative und reflektiver Indikatoren in Strukturgleichungsmodellen, in: Handbuch PLS- Pfadmodellierung. Schäffer Poeschel.

[FaPS96] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. (1996, August). Knowledge Discovery and Data Mining: Towards a Unifying Framework. In KDD (Vol. 96, pp. 82-88).

[FHKo17] Fachhochschule Köln (2011). Datenbanken Online Lexikon. URL: http://wikis.gm.fh-koeln.de/wiki_db/Datenbanken

Literaturverzeichnis

[GaUW08] Garcia-Molina, H., Ullman, J. D., Widom, J. (2008). Database Systems: The Complete Book. Pearson International.

[GMPB+09] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. Nature, 457 (7232), 1012-1014.

[HaKa06] Han, J., Kamber, M. (2006). Data mining: concepts and techniques. Morgan Kaufmann.

[Harr15] Harrison, G. P. (2015). Next Generation Databases - NoSQL and Big Data. Apress.

[JZFF11] Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). Recommender Systems: An Introduction. Cambridge University Press.

[KeBM10] Kemper, H. G., Baars, H., Mehanna, W. (2010). Business Intelligence: Grundlagen und praktische Anwendungen. Vieweg.

Literaturverzeichnis

[KeEi15] Kemper, A., Eickler, A. (2015). Datenbanksysteme. Oldenbourg.

[KrGL05] Kraft, M., Götz, O., Liehr-Gobbers, K. (2005). Die Validierung von Strukturgleichungsmodellen mit Hilfe des Partial-Least-Squares-Ansatzes, in: Handbuch PLS-Pfadmodellierung. Schäffer Poeschel.

[KuJo13] Kuhn, M., Johnson, K. (2013): Applied Predictice Modeling. Springer.

[LKKV14] Lazer, D., Kennedy, R., King, G., Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. Science, 343 (6176), 1203-1205.

[MBK+12] Mertens, P., Bodendorf, F., König, W., Schumann, M., Hess, T. (2012). Grundzüge der Wirtschaftsinformatik. Springer.

[Muel15] Müller, E. (2015). Vorlesungsskript: Big Data Analytics. Hasso-Plattner- Institut Potsdam.

[Naum15] Naumann, F. (2015). Vorlesungsskript: Datenbanksysteme II. Hasso-Plattner-Institut Potsdam.

Literaturverzeichnis

[O'Nei16] O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown Books.

[PäPa15] Pääkkönen, P., Pakkala, D. (2015). Reference architecture and classification of technologies, products and services for big data systems. Big Data Research, 2 (4), 166-186.

[Plat13] Plattner, H. (2013). A Course in In-Memory Data Management. Springer.

[PoBe13] Becker, L., Pousttchi, K. (2013). Requirements for personalized m-commerce - what drives consumers' use of social networks?. In: Journal of Electronic Commerce in Organizations 11 (2013) 4, S. 19-36.

[RaSS15] Rahm, E., Saake, G., & Sattler, K. U. (2015). Verteiltes und paralleles Datenmanagement: von verteilten Datenbanken zu Big Data und Cloud. Springer.

[RePi91] Reichwald, R., Picot, A. (1991). Informationswirtschaft, in: Heinen, E. (Hrsg.). Industriebetriebslehre: Entscheidungen im Industriebetrieb. Springer.

[Rupa16] Ruparelia, N. B. (2016). Cloud Computing. MIT Press.

Literaturverzeichnis

[RuQZ07] Rupp, C., Queins, S., Zengler, B. (2007). UML 2 glasklar: Praxiswissen für die UML-Modellierung. Carl Hanser.

[SaSH11] Saake, G., Sattler, K. U., Heuer, A. (2011). Datenbanken: Implementierungstechniken. mitp.

[SaSH13] Saake, G., Sattler, K. U., Heuer, A. (2013). Datenbanken: Konzepte und Sprachen. mitp.

[Sche97] Scheer, A. W. (1997). ARIS—vom Geschäftsprozess zum Anwendungssystem. Springer.

[StHa04] Stahlknecht, P., Hasenkamp, U. (2004). Einführung in die Wirtschaftsinformatik. Springer.

[Sull15] Sullivan, D. (2015). NoSQL for mere mortals. Addison-Wesley Professional.

[Thor13] Thor, A. (2013). Vorlesungsskript: Datenmanagement. Universität Leipzig.

Literaturverzeichnis

[Vige16] Vigen, T. (2016). Spurious Correlations. Hachette Books.

[Voss08] Vossen, G. (2008). Datenmodelle, Datenbanksprachen und Datenbank-Management-Systeme. Oldenbourg.

[WeKa16] Wedel, M., Kannan, P. K. (2016). Marketing analytics for data-rich environments. Journal of Marketing, 80(6), 97-121.

[Wies15] Wiese, L. (2015). Advanced Data Management: For SQL, NoSQL, Cloud and Distributed Databases. de Gruyter.

[Zehn13] Zehnder, C. A. (2013). Informationssysteme und Datenbanken. Springer.

[Zoss15] Zoss, A. (2015). Practical Data Visualization. Lecture at Duke University.