Betriebswirtschaftslehre II
Vorlesung 9: Business Intelligence – Data Mining

Wintersemester 2018/19

Prof. Dr. Martin Schultz
martin.schultz@haw-hamburg.de

Agenda

1. Überblick Data Mining
2. Klassifikation
3. Segmentierung
4. Abhängigkeitsanalyse
5. Abweichungsanalyse
6. Text Mining

Inhalte der Vorlesung und Übung

	Termin	Vorlesung	Übung
1	28.09.2018	Einführung und Grundlagen	-
2	05.10.2018	Geschäftsprozessmodellierung	Übung 1 – Gruppe 3/4
3	12.10.2018	Anwendungssysteme in Unternehmen	Übung 1 – Gruppe 1/2
4	19.10.2018	ERP-Systeme	Übung 2 – Gruppe 3/4
5	26.10.2018	ERP-Systeme: ReWe und Einführungsprojekte	Übung 2 – Gruppe 1/2
6	02.11.2018	Business Intelligence - OLAP	Übung 3 – Gruppe 3/4
7	09.11.2018	Business Intelligence - ETL	Übung 3 – Gruppe 1/2
8	16.11.2018	Business Intelligence – Dashboards	Übung 4 – Gruppe 3/4
9	**23.11.2018**	**Data Mining**	**Übung 4 – Gruppe 1/2**
10	30.11.2018	Informationsmanagement	Übung 5 – Gruppe 3/4
11	07.12.2018	IT-Service-/ Enterprise Architecture-Management	Übung 5 – Gruppe 1/2
12	14.12.2018	Klausurvorbereitung	Übung 6 – Gruppe 3/4
	21.12.2018		Übung 6 – Gruppe 1/2
	11.01.2019		Übung 7 – Gruppe 1/2/3/4

Was sollen Sie mitnehmen…

- Wesentliche Aufgabentypen und Verfahren des Data Mining beschreiben und anwenden können

Informationsbereitstellung und Nutzertypen

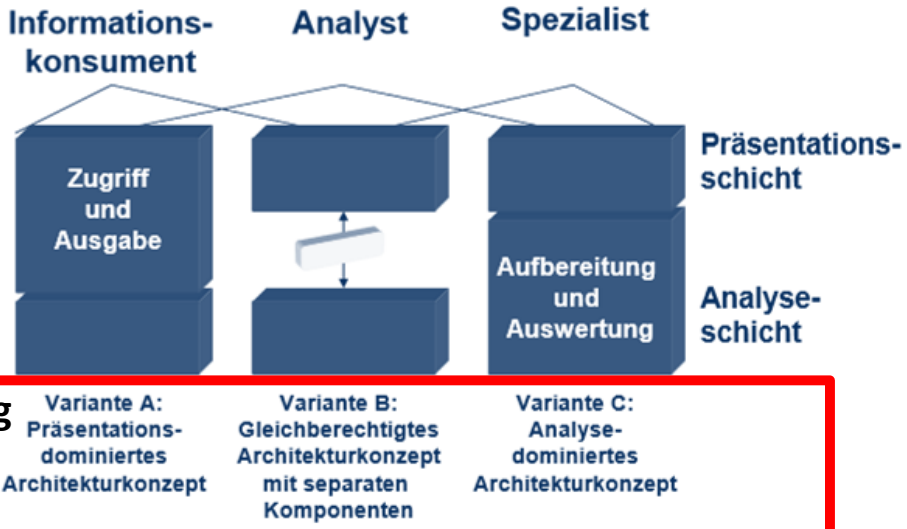
Spektrum der **Nutzer** von BI-Lösungen ist sehr heterogen (Kenntnisse, Vorlieben), viele unterschiedliche **Darstellungsformen** anwendbar mit unterschiedlichem **Interaktionsgrad** (starr, Verlinkung)

Nutzertyp **Informationskonsument → Berichtswesen, Management Support Systeme/ Dashboards**

- Nutzertyp der überwiegend Tools verwendet, die das Datenmaterial nach **festen Mustern** aufbereiteten und ausgeben

Nutzertyp **Analytiker → OLAP**

- Nutzertyp der überwiegend die Funktionalitäten der **navigationsorientierten Analyse** einsetzen und sich frei im Datenbestand bewegen will
- Einfache Methoden und Werkzeuge für Anzeige/ Ausgabe kommen zur Anwendung



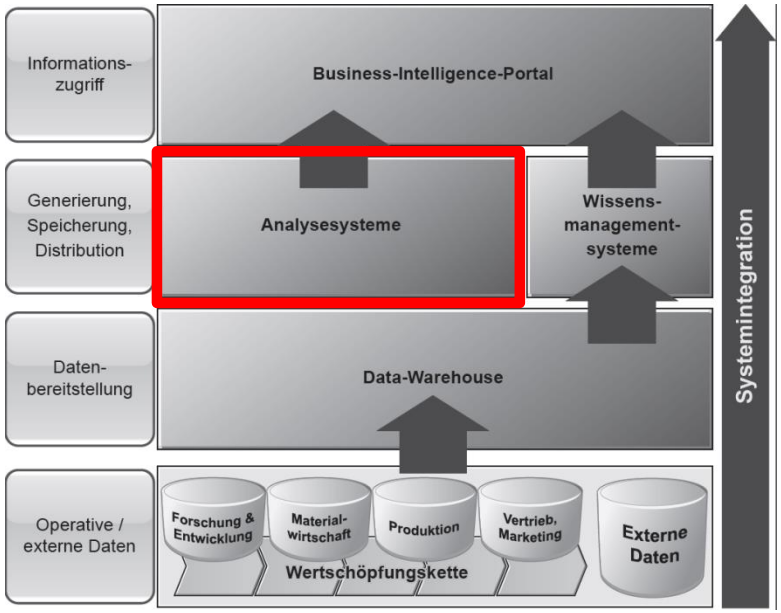
Nutzertyp **Spezialist → Decision Support, Data Mining**

- Nutzertyp der vorwiegend direkt auf die **methodenorientierten Funktionsbausteine** zurückgreift, um anspruchsvolle Datenanalysen vorzunehmen
- Nimmt dabei funktionale Komplexität und wenig benutzungsfreundliche Oberflächen in Kauf, ggf. Misstrauen gegenüber einfachen Zugriffs- und Ausgabewerkzeugen

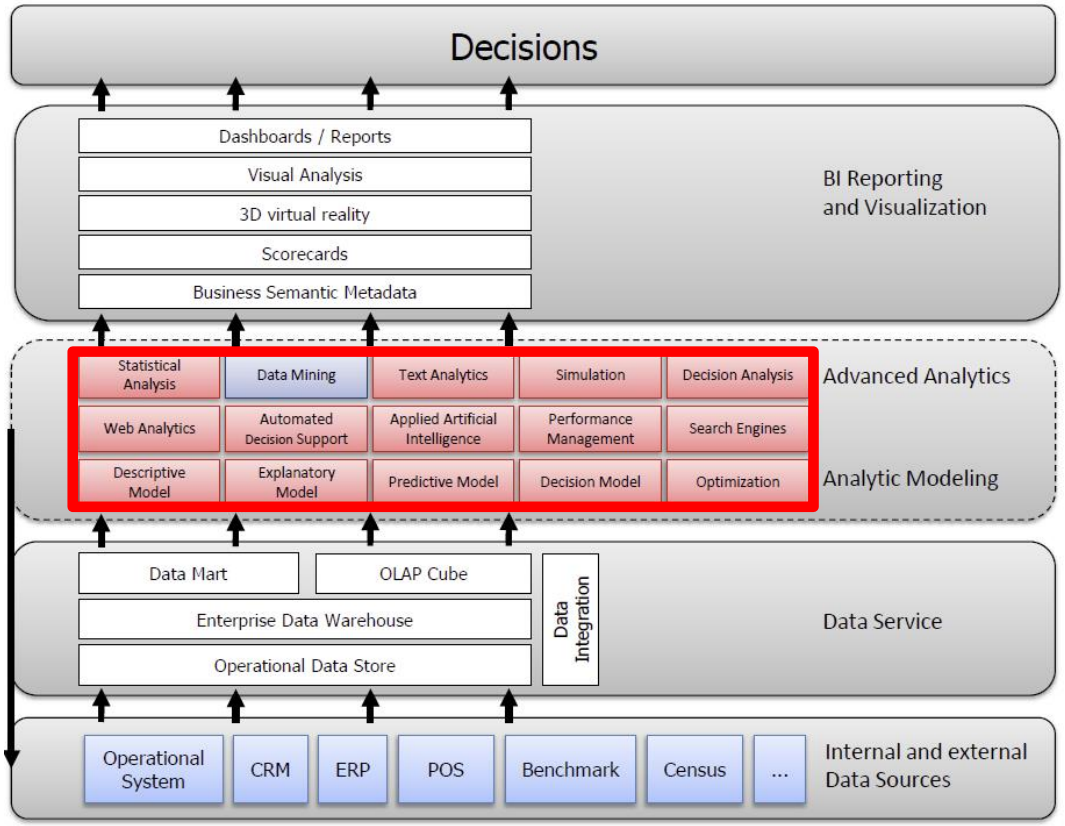
(Gluchowski 2008)

Data Mining als Bestandteil der Analyseschicht

- In Bezug auf die Schichtenarchitektur von BI-Systemen lässt sich Data Mining auf der Analyseschicht einordnen
- Ziel ist es, auf Basis der im Data Warehouse vorhandenen Daten mittels komplexer Methoden Modelle/ relevante Informationen für Entscheidungen zu generieren



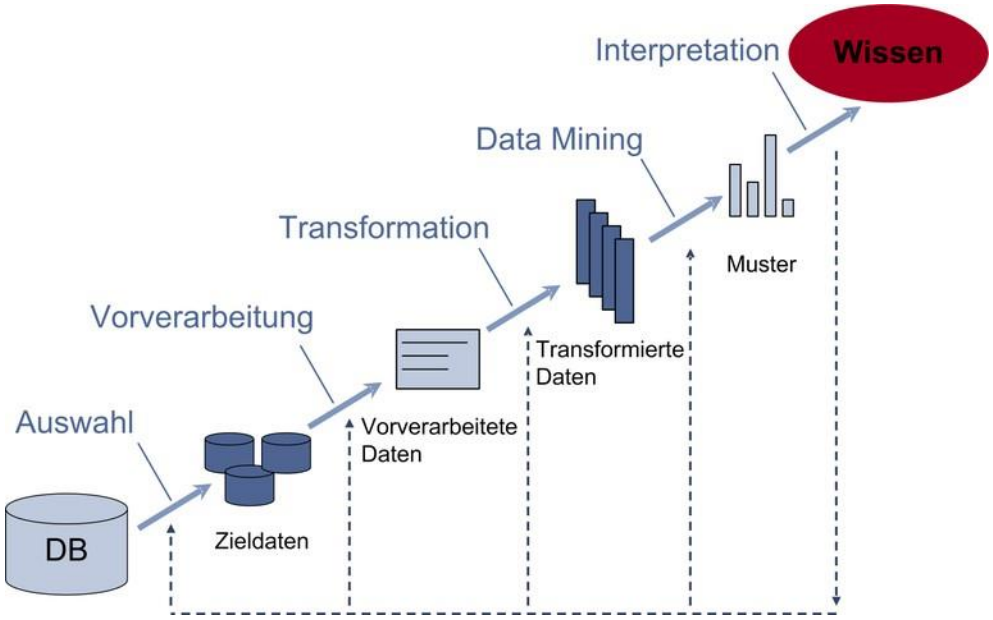
(Hansen 2009)



(Shen 2011)

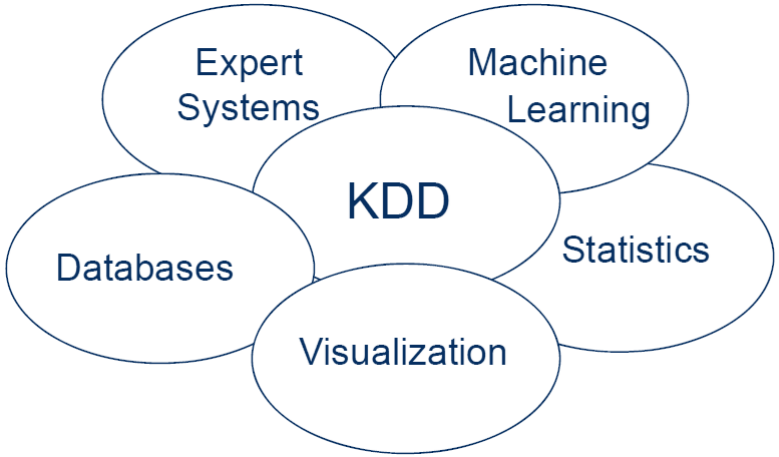
Data Mining: Begriff

- "**Data Mining** is a problem-solving methodology that finds a logical or mathematical description, eventually of a complex nature, of patterns and regularities in a set of data." (Decker 1995)
- **Knowledge Discovery in Databases (KDD):** „… non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data …" (Fayyad 1996)



Chamoni 2013

KDD/ Data Mining
als interdisziplinäres Forschungsgebiet

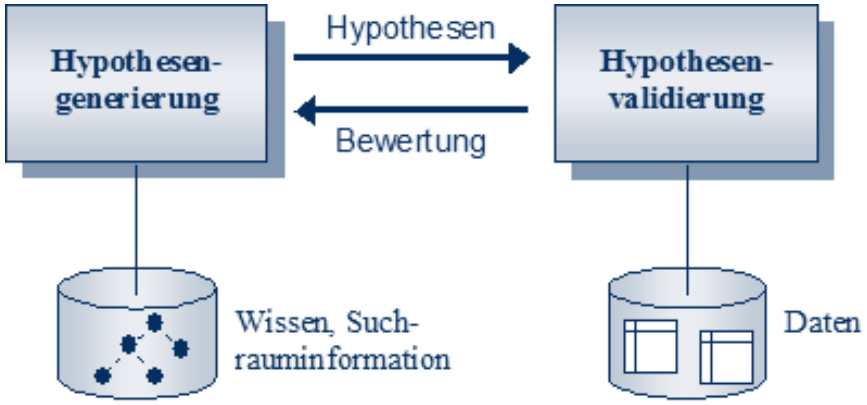


Data Mining: Abgrenzung

Wachsende Datenmenge führen dazu, dass „klassische“ **hypothesengestützte** Auswertungen von Daten (OLAP, Statistik) an Ihre Grenzen stoßen

- zu zeitaufwendig ***alle*** **möglichen Zusammenhänge** in den Daten zu überprüfen
- Auswahl der zu untersuchenden Hypothesen hochgradig subjektiv

Lösung: computergestützte, „hypothesenfreie“ Suche nach Mustern in Daten



hoch — Grad der Hypothesenfreiheit — niedrig

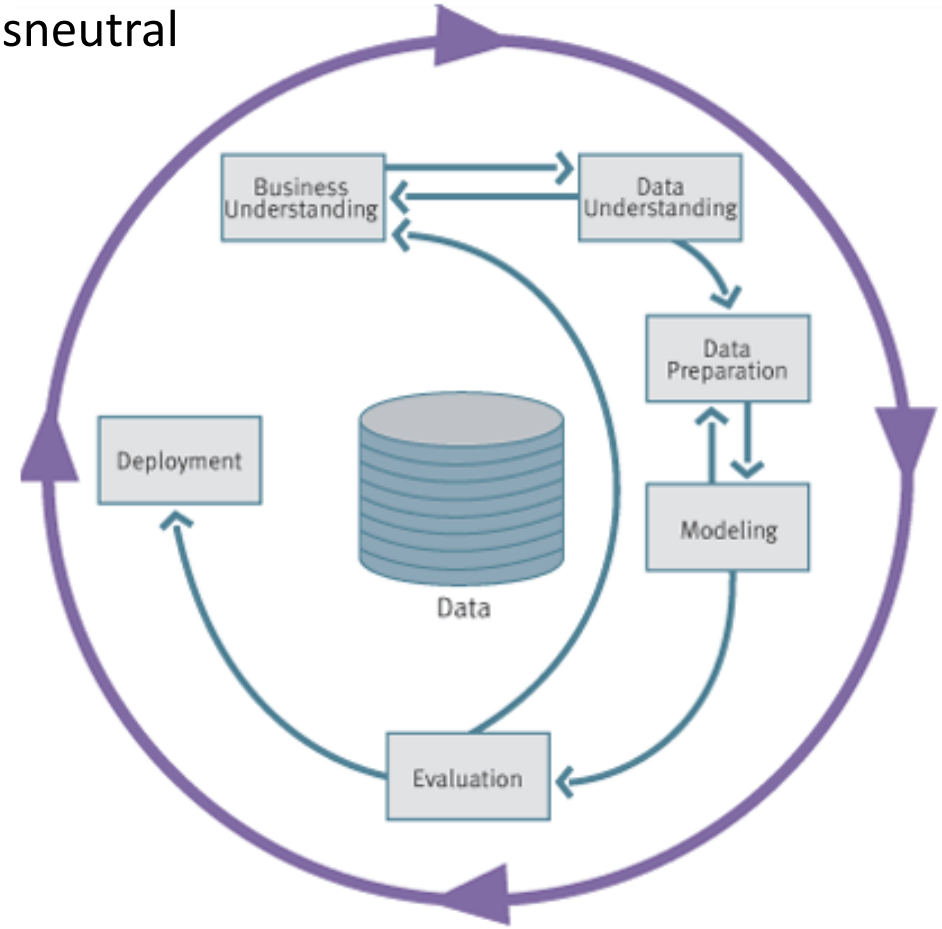
Anwendungs-gebiet	unbestimmt	vorgegeben	vorgegeben	vorgegeben
Art der Ergebnisse	unbestimmt	unbestimmt	vorgegeben	vorgegeben
Hypothesen-generierung	datengetrieben	datengetrieben	datengetrieben	Anwender
Hypothesen-validierung	System	System	System	System

Herausforderung für die Zukunft — Data Mining im engeren Sinne — Data Mining im weiteren Sinne

Vorgehensmodell für Data Mining - CRISP

Cross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining ist ein robustes, allgemeines Modell zur Systematisierung von Data Mining-Projekten

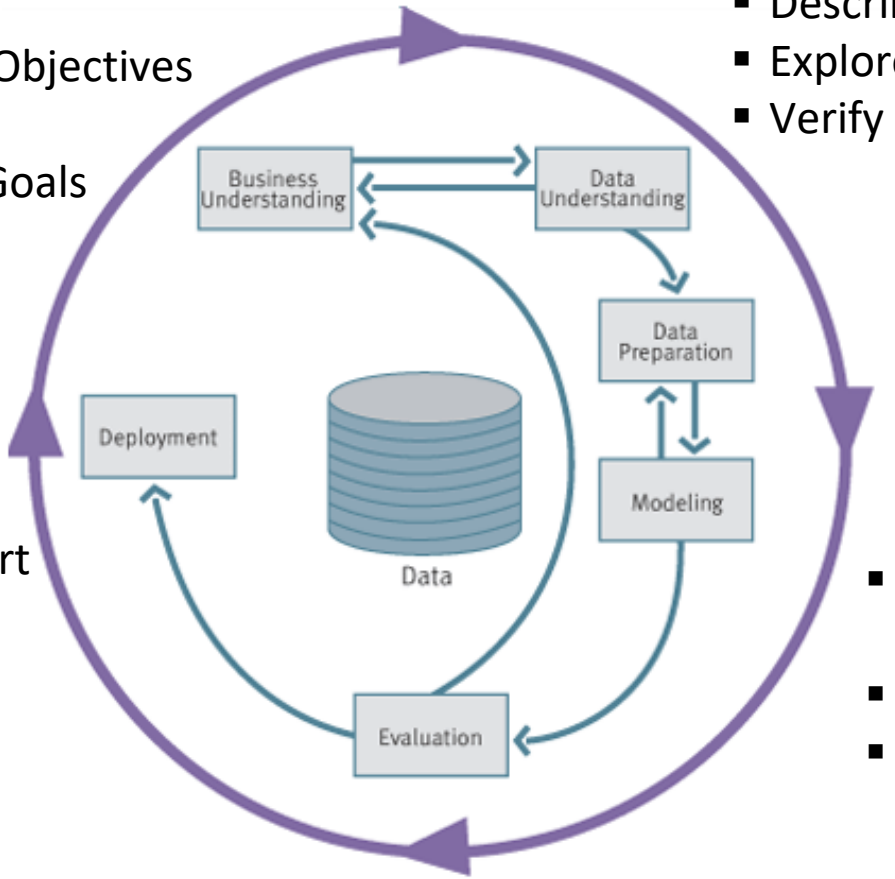
- ist branchen-, tool- und anwendungsneutral
- basiert auf praktischen Erfahrungen
- gewährleistet hohe Qualität
- reduziert Kosten und Zeitaufwand
- unterstützt die Dokumentation und Argumentation



Vorgehensmodell für Data Mining - CRISP

- Determine Business Objectives
- Assess Situation
- Determine Analysis Goals

- Collect Initial Data
- Describe Data
- Explore Data
- Verify Data Quality



- Select Data
- Clean Data
- Construct Data
- Integrate Data
- Format Data

- Plan Deployment
- Plan Maintenance
- Produce Final Report

- Select Modelling Technique
- Build Model
- Assess Model

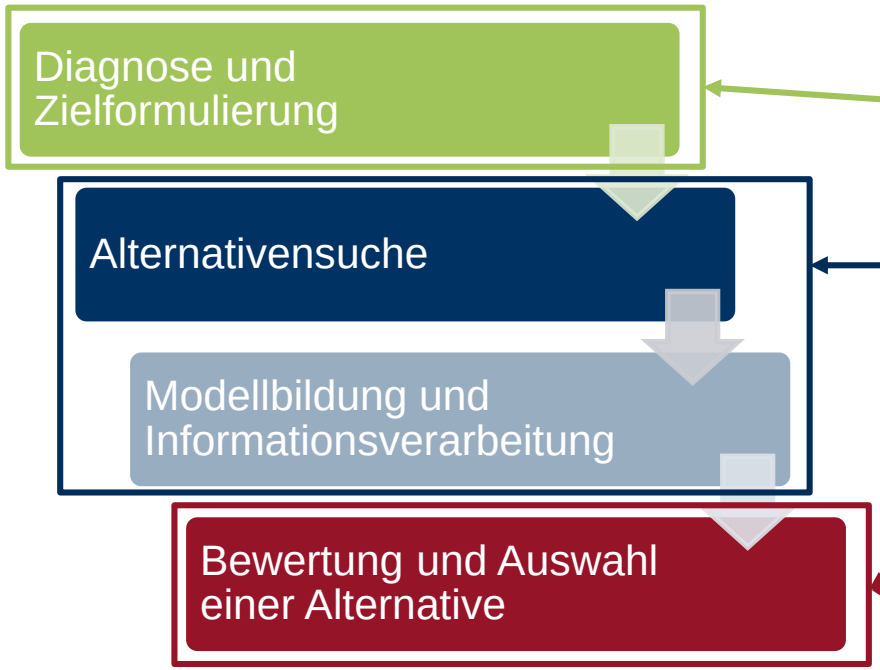
- Evaluate Results
- Determine Next Steps

(Wirth and Hipp 2000)

Data Mining und Management-/ Entscheidungsaufgaben

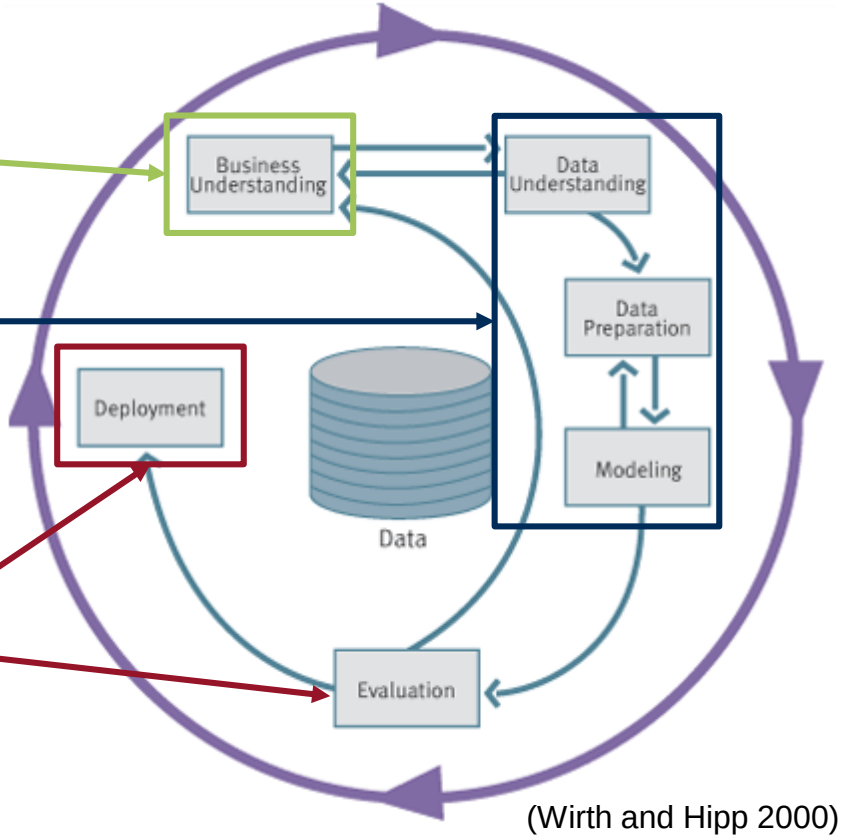
Bei der Anwendung von Data Mining für die Unterstützung von Entscheidungen sind Herausforderungen auf **fachlicher** und **technischer** Ebene zu lösen

Prozessmodell in der Entscheidungstheorie



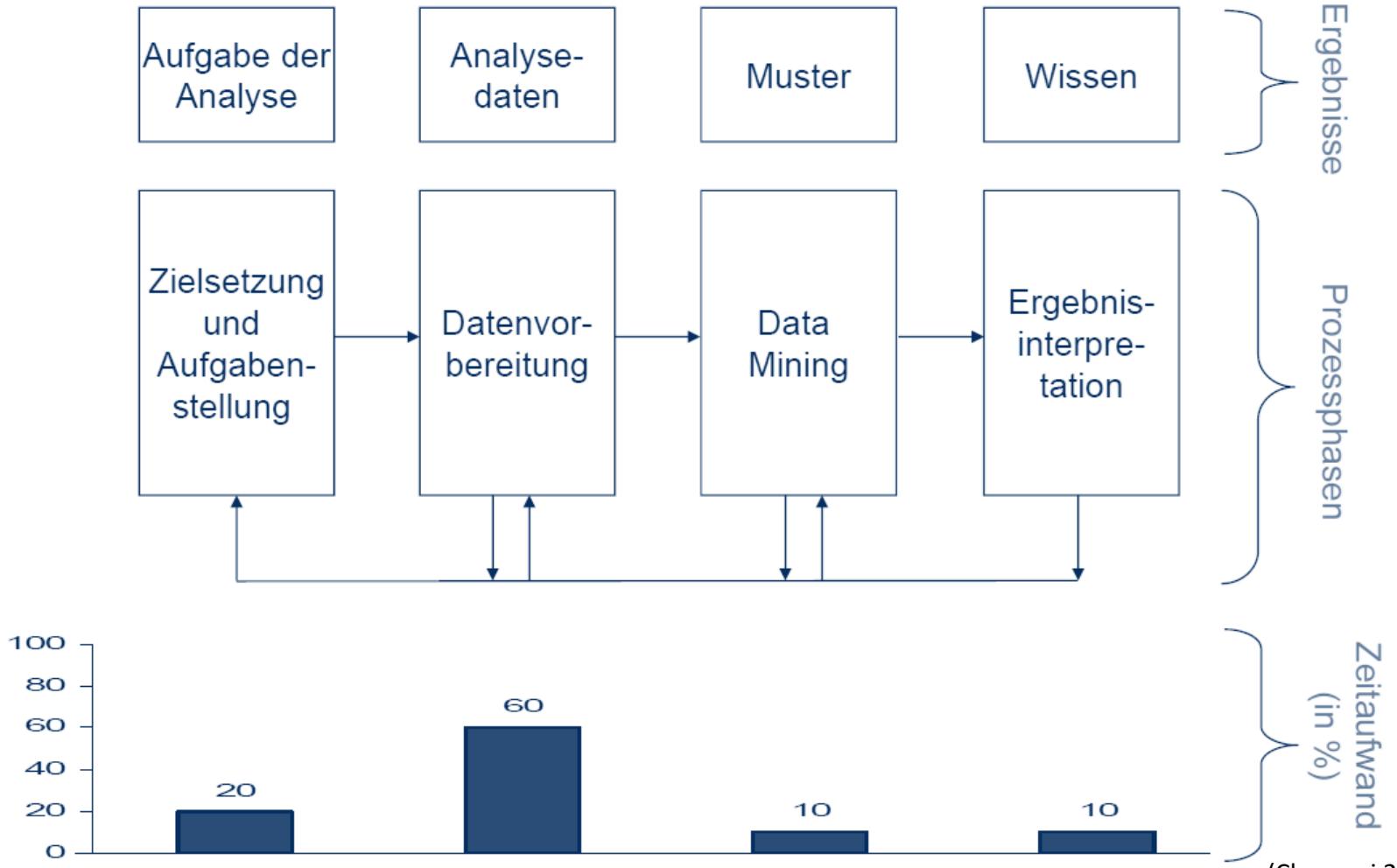
(Geldermann 2013)

Prozessmodell für Data Mining



(Wirth and Hipp 2000)

Prozessmodelle und Zeitaufwand



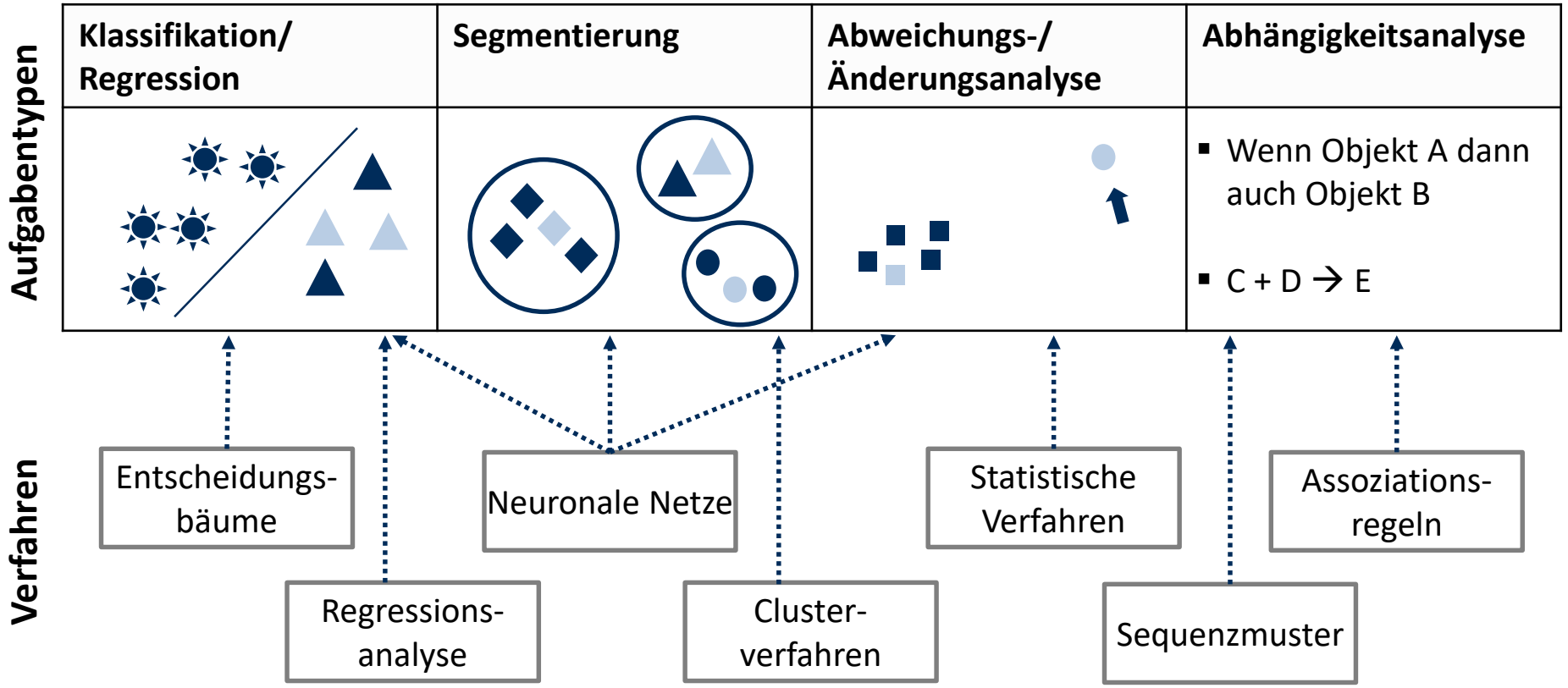
(Chamoni 2013)

Data Mining: Software Überblick

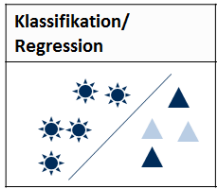
Gartner defines Advanced Analytics as, "the analysis of all kinds of data using **sophisticated quantitative methods** (e.g. statistics, descriptive and predictive data mining, simulation and optimization) to produce insights that **traditional approaches** to business intelligence (BI) - such as query and reporting – **are unlikely to discover."**

Data Mining: Aufgabentypen und Verfahren

- Im Data Mining haben sich im Wesentlichen vier übergeordnete Aufgabentypen herauskristallisiert, auf die sich viele betriebswirtschaftliche Fragestellungen abbilden lassen
- Diese Aufgabentypen lassen sich jeweils wiederum durch mehrere Verfahren/ Algorithmen bearbeiten

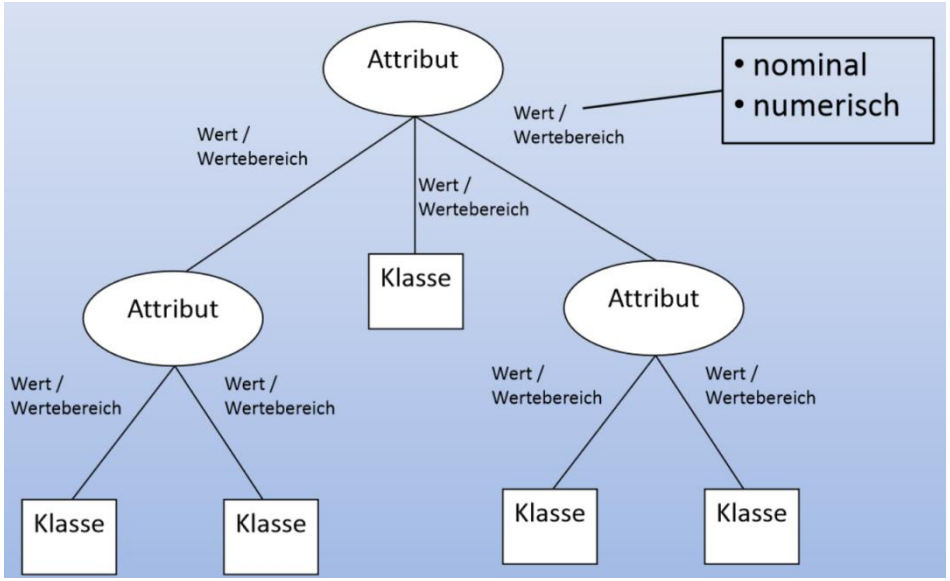


Klassifikation: Beschreibung des Aufgabentyps

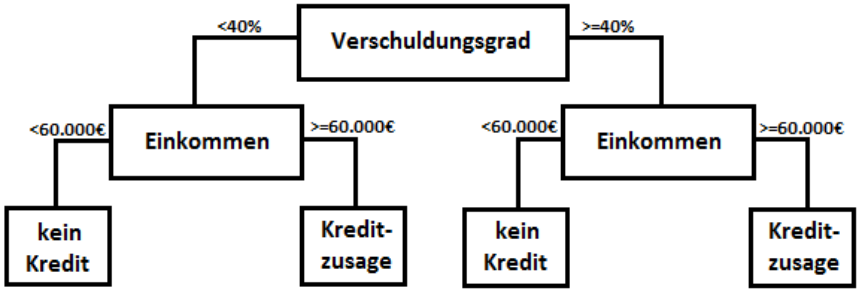


- Ein Klassifikationsmodell ist eine **Abbildung**, die die Zuordnung von Elementen zu **vorgegebenen Klassen** beschreibt.
- Die Grundlage für ein Klassifikationsmodell bildet ein Datenbestand, dessen **Datenobjekte jeweils bereits einer vorgegebenen Klasse zugeordnet sind**.
- Zum erstellen des Modells ist es erforderlich, dass ein geeigneter Datensatz mit verlässlichen Erfahrungswerten zum Entscheidungsproblem (***Trainingsdatensatz***) vorliegt
- Ein Klassifikationsmodell wird zur Prognose der Klassenzugehörigkeit von Datenobjekten eingesetzt, deren Klassenzugehörigkeit bislang noch nicht bekannt ist.

http://www.enzyklopaedie-der-wirtschaftsinformatik.de

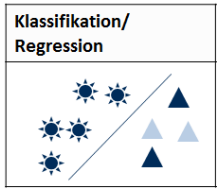


Beispiel: Einteilung von Datensätzen, die Angaben über Kunden enthalten, so dass damit die Kundengruppe erkannt werden kann, in die der Kunde voraussichtlich gehört



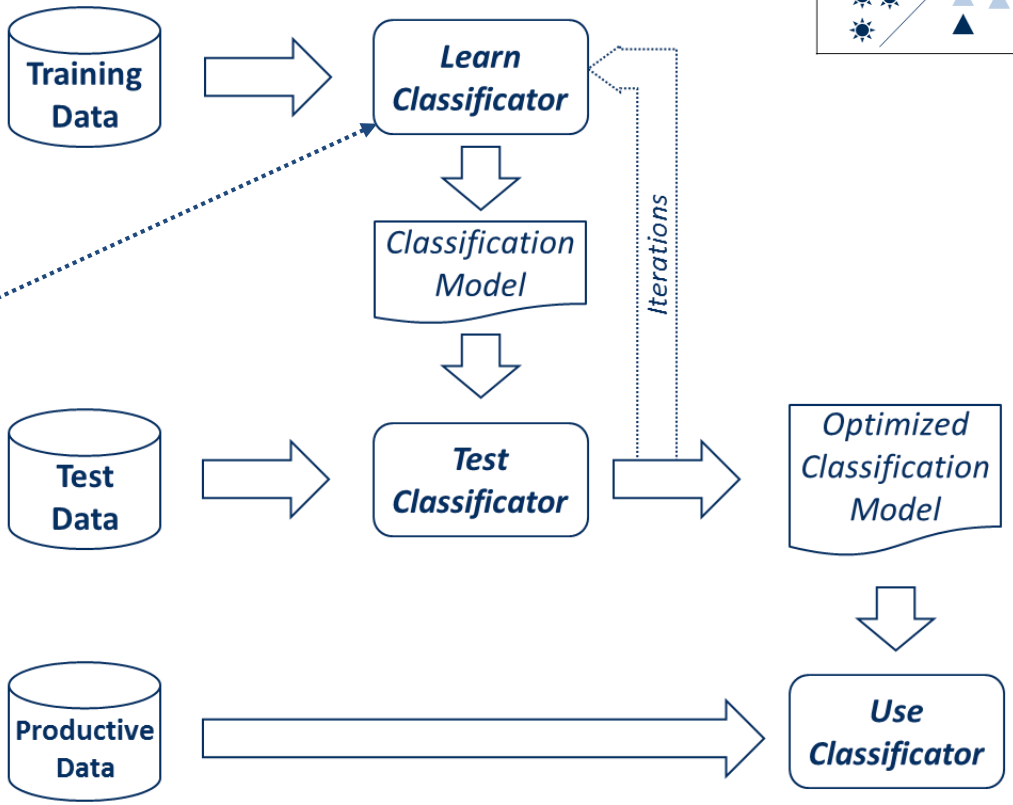
HAW HAMBURG

Klassifikation: Verfahren und Vorgehen

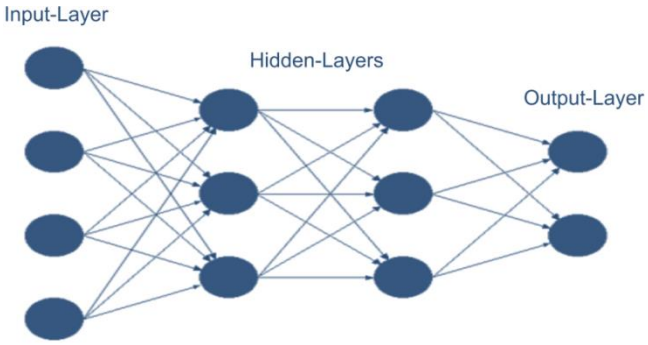


- Entscheidungsbäume
- lineare/quadratische Klassifikatoren
- Neuronale Netze/Perzeptron
- Support vector machines
- Statistisches Lernen/ Bayes Klassifikator
- k-Nearest Neighbor (k-NN)
- ...

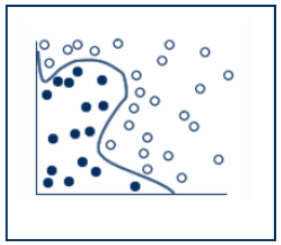
Vorgehen



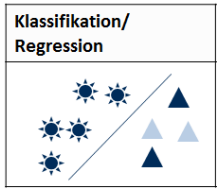
Neuronale Netze



Support vector machines



Klassifikation: Entscheidungsbaum - Beispiel



- Übersicht von Versicherungsnehmern und deren Einteilung die Klassen „Hoch“ und „Gering“ auf Basis des bisher verursachten Schadens

Alter	Geschlecht	Autotyp	Schaden
45	W	Van	Gering
35	W	Coupe	Gering
22	W	Van	Gering
19	M	Coupe	Hoch
38	W	Coupe	Gering
43	W	Coupe	Gering
37	W	Van	Gering
24	M	Van	Hoch
27	M	Coupe	Hoch
19	M	Van	Hoch
40	M	Van	Gering
40	M	Coupe	Hoch
23	W	Coupe	Gering

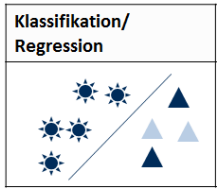
Aufgabenstellung:

Wie sieht ein Klassifikationsbaum aus, der die Variable „Schaden“ am besten vorhersagt?

Zeit: 2 min

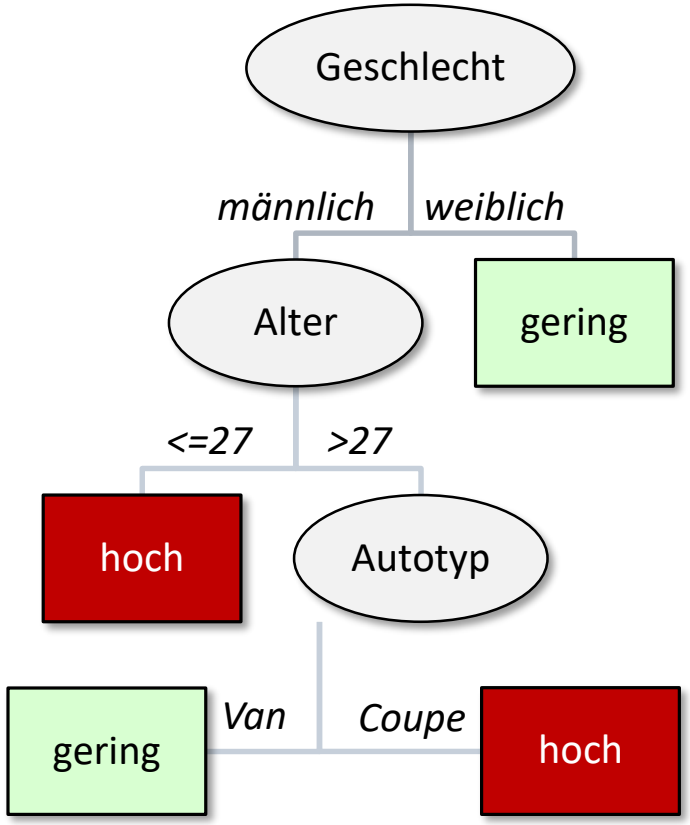
HAW HAMBURG

Klassifikation: Entscheidungsbaum - Beispiel



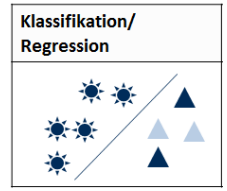
- Übersicht von Versicherungsnehmern und deren Einteilung die Klassen „Hoch“ und „Gering“ auf Basis des bisher verursachten Schadens

Alter	Geschlecht	Autotyp	Schaden
45	W	Van	Gering
35	W	Coupe	Gering
22	W	Van	Gering
19	M	Coupe	Hoch
38	W	Coupe	Gering
43	W	Coupe	Gering
37	W	Van	Gering
24	M	Van	Hoch
27	M	Coupe	Hoch
19	M	Van	Hoch
40	M	Van	Gering
40	M	Coupe	Hoch
23	W	Coupe	Gering

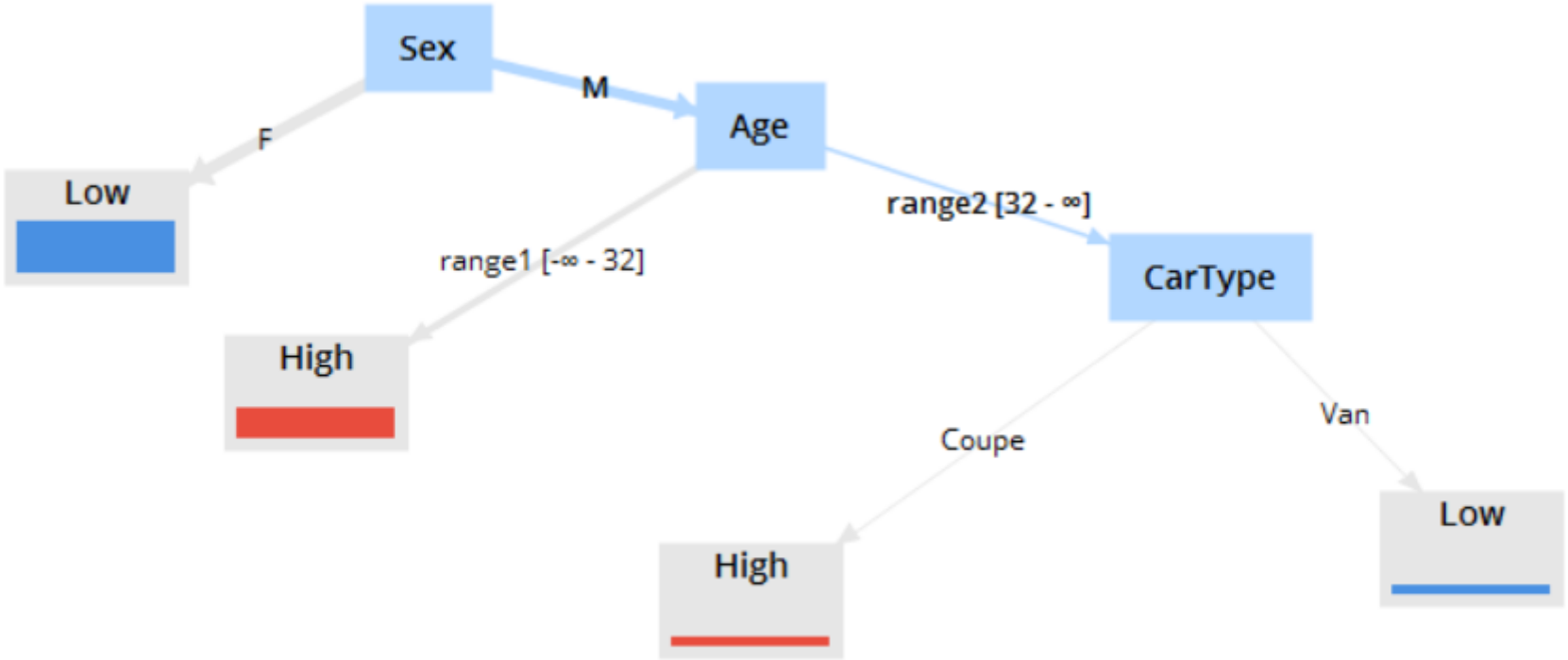


HAW HAMBURG

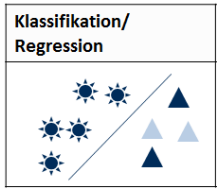
Klassifikation: Entscheidungsbaum – Beispielergebnis



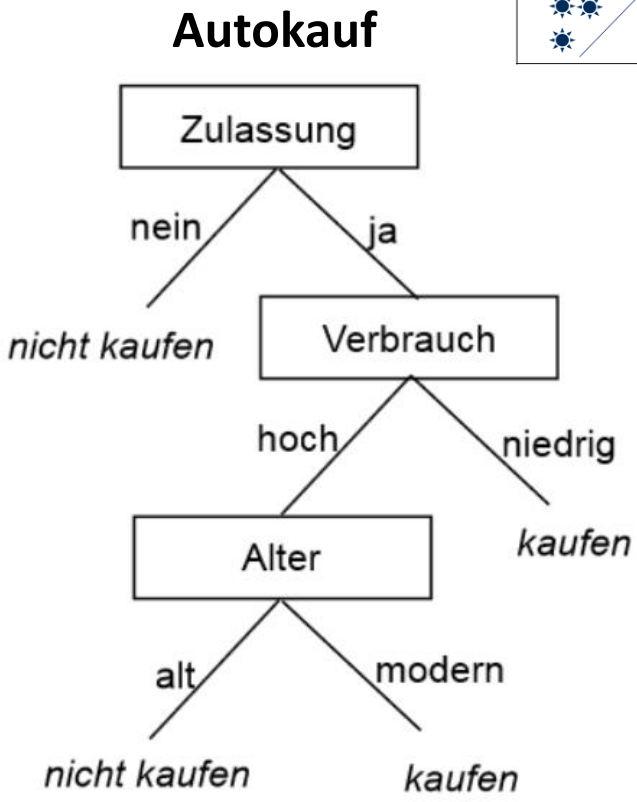
- Algorithmus: **Ch**i-square **A**utomatic **I**nteraction **D**etectors (CHAID)



Klassifikation: Entscheidungsbaum – Verfahren

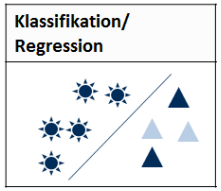


- **Vorgehen:** Aufbau eines Baums
  - Knoten entspricht Entscheidungskriterium
  - Blatt entspricht Entscheidung
- **Vorteile:**
  - Ergebnis leicht interpretierbar
  - Übersetzbar in Regelsystem
- **Diverse Implementierungen:**
  - ID3/ C4.5/ C.50/ J48
  - C&R Tree
  - Quest
  - CHAID



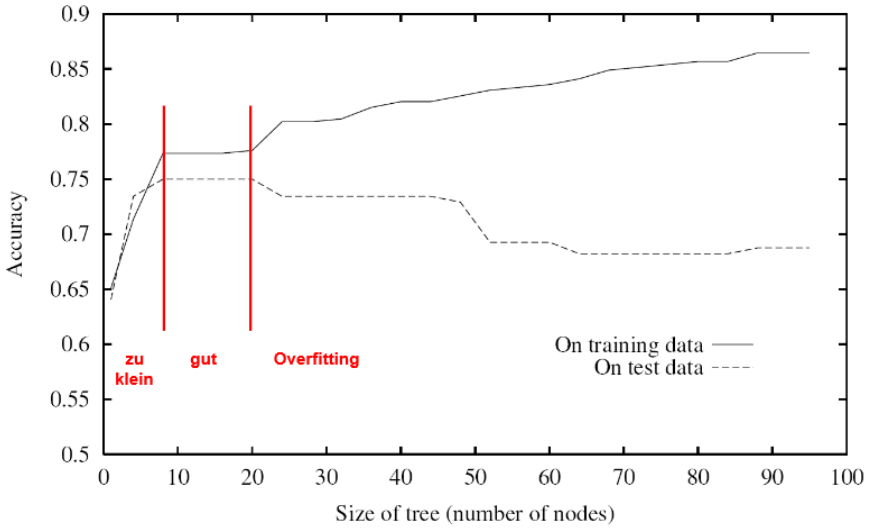
Eine Regel (von mehreren):
Wenn
Zulassung vorhanden und
Verbrauch niedrig,
dann kaufen.

Klassifikation: Entscheidungsbaum – Verfahren

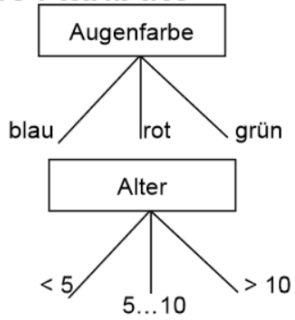


Existierende Algorithmen unterscheiden sich in

- **dem Vorgehen zur Wahl des Splitattributs** (z.B. Chi-Quadrat-Test →CHAID, Informationsgewinn → C5.0, Gini-Index → C&R Tree)
- **der Wahl des Stoppkriteriums** (z.B. Minimale Tupelzahl je Knoten, Minimaler Anteil falsch klassifizierter Tupel, Maximale Baumtiefe, Maximale Knotenanzahl)
- **der Art des Splits**
- der Wahl des **Pruning**-Verfahrens

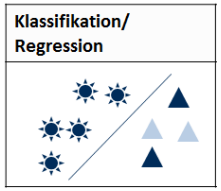


- Diskrete vs. kontinuierliche Attribute
  - Diskret
    - Ein Knoten pro Attributwert
  - Kontinuierlich:
    - Ein Knoten pro Attributintervall



- Binäre vs. n-äre Bäume
  - Zwei oder mehrer Ausgangskanten

HAW HAMBURG



Klassifikation: Entscheidungsbaum – Wahl des Splitattributs

Gini-Index: Statistisches Maß zur Darstellung von Ungleichverteilungen

- **Prinzip:** Minimierung der Heterogenität
- **Vorgehen:** Wahrscheinlichkeitsmaß, bei Stichprobe, Datentupel aus 2 unterschiedlichen Klassen (0;1) zu erhalten: ***Gini Index = 1 – p(0)² – p(1)²***
- Minimum = 0,0 → alle Objekte aus einer Klasse = Maximale Homogenität
- Maximum = 0,5 → Objekte zweier Klassen gleich häufig = Maximale Heterogenität

Beispiel Versicherung Auto

- ***Split Geschlecht:*** *1 – p(G)² – p(H)²*
  W | M: GGGGGGG | HHHHGH
  W: 1 - (7/7)² - (0/7)² = 0,0
  M: 1 - (1/6)² - (5/6)² = 0,2778
  *Mittel: 7/13 * 0 + 6/13 * 0,2778 =* ***0,1282***

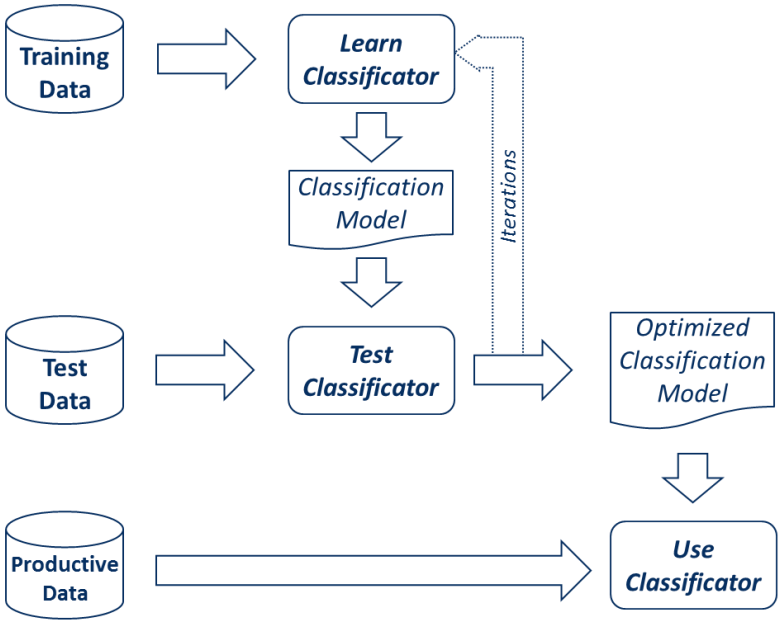
- ***Split AutoTyp:*** *1 – p(G)² – p(H)²*
  Van | Coupe: GGGHHG | GHGGHHG
  Van: 1 - (4/6)² - (2/6)² = 0,444
  Coupe: 1 - (4/7)² - (3/7)² = 0,489
  *Mittel: 6/13 * 0,444 + 7/13 * 0,489 =* ***0,468***

- ***Ergebnis: Geschlecht besseres Split-Attribut***

Alter	Geschlecht	Autotyp	Schaden
45	W	Van	Gering
35	W	Coupe	Gering
22	W	Van	Gering
19	M	Coupe	Hoch
38	W	Coupe	Gering
43	W	Coupe	Gering
37	W	Van	Gering
24	M	Van	Hoch
27	M	Coupe	Hoch
19	M	Van	Hoch
40	M	Van	Gering
40	M	Coupe	Hoch
23	W	Coupe	Gering

Klassifikation: Validierung eines Klassifikators

- Das Modell wird gelernt anhand eines Trainingsdatensatzes
- Das Modell wird dann auf einen Testdatensatz angewendet, bei dem auch die Klassenzugehörigkeit jedes Datensatzes bekannt ist
- Die Güte des Modells wird ermittelt anhand des Vergleichs der vom Modell vorhergesagten Klassenzugehörigkeit mit der tatsächlichen Klassenzugehörigkeit
  → Confusion Matrix



Confusion Matrix

	Actual -- True/False	
Predicted -- Positive/Negative	True Positive	False Positive (Type I)
	False Negative (Type II)	True Negative

Klassifikation: Validierung eines Klassifikators - Kennzahlen

- **Precision:** Anteil der korrekt als positiv klassifizierten Objekte an der Gesamtheit der als positiv klassifizierten Objekte

$$\text{Precision} = \frac{TP}{TP + FP}$$

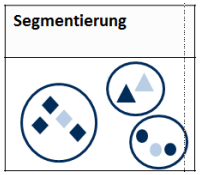
- **Recall** (Sensitivität)**:** Anteil der korrekt als positiv klassifizierten Objekte an der Gesamtheit der tatsächlich positiven Objekte

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **Accuracy:** Anteil aller Objekte, die korrekt klassifiziert wurden

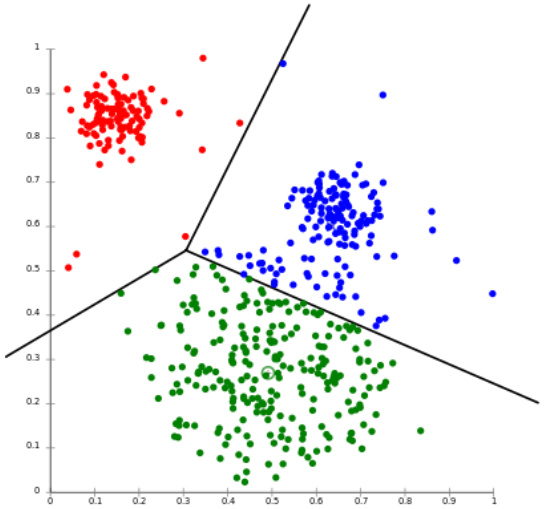
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

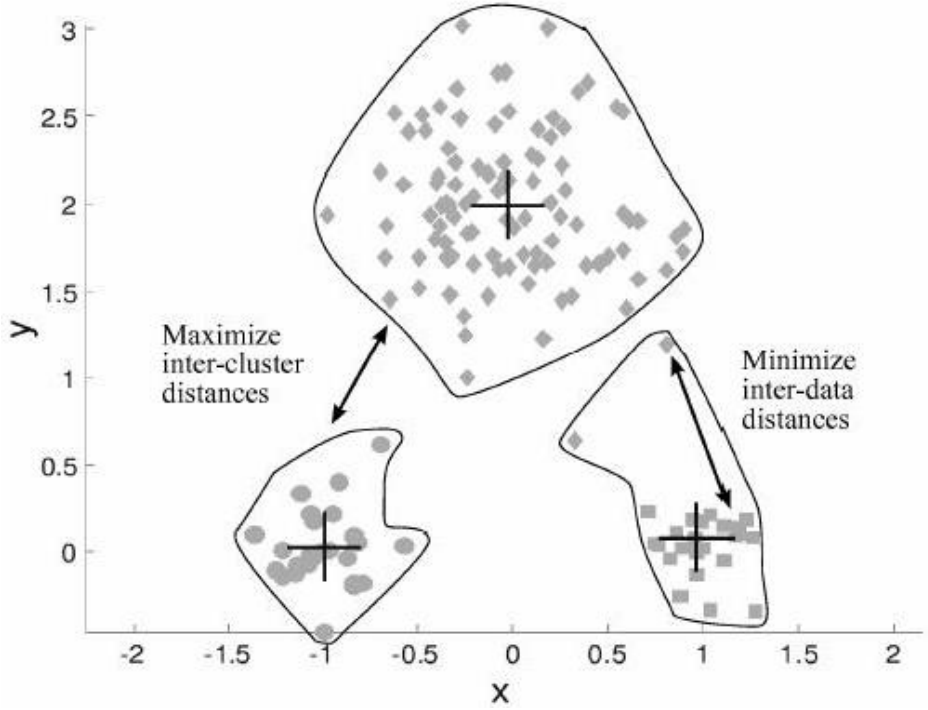
	Actual -- True/False	
Predicted -- Positive/Negative	True Positive	False Positive (Type I)
	False Negative (Type II)	True Negative

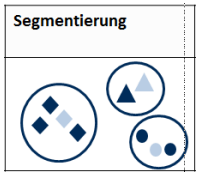


Segmentierung: Clusterverfahren

- **Ziel** der Anwendung von Clusterverfahren ist das Erkennen und Bewerten von Clustern. **Cluster** sind Mengen von Datensätzen. Dabei sollen Datensätze innerhalb eines Clusters möglichst ähnlich (homogen) und Datensätze aus unterschiedlichen Clustern dagegen möglichst unähnlich sein.
- **Voraussetzung:** Es müssen Ähnlichkeitsmaße zwischen Datensätzen sowie zwischen Clustern definiert werden
- Um die Ähnlichkeit zweier Datensätze zu bestimmen, werden oftmals **geometrische Distanzmaße** herangezogen.





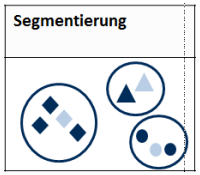


Segmentierung: Clusterverfahren - Beispiel

- **Anwender:** Produzent von Getränken (Wasser, Limonaden, Fruchtschorlen)
- **Ziel/ Frage:** Optimale Erweiterung der Produktpalette (Trendlimos oder neue Fruchtschorlen?)
- **Datenerhebung:** Befragung vor Supermärkten
- 1. Schritt: Kundensegmentierung durch Analyse von Kundenattributen:
  → Welche Kundenprofile existieren?

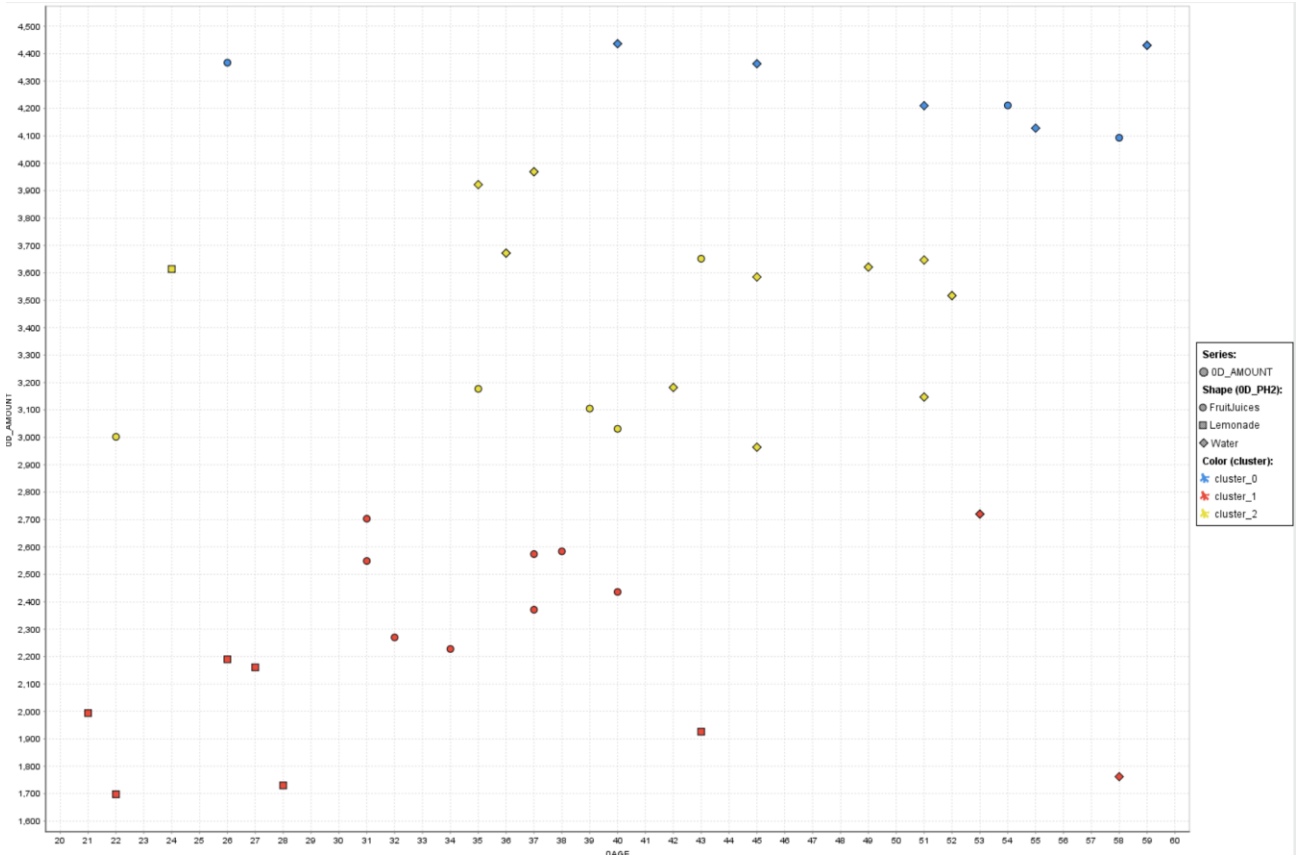
ID_NUM	Customer	Product	Age	Income	Currency
1	K1	Water	59	4430.0	EUR
2	K2	Water	55	4128.0	EUR
3	K4	Water	51	4210.0	EUR
4	K12	Water	53	2720.0	EUR
5	K13	FruitJuices	58	4093.0	EUR
6	K14	Water	58	1762.0	EUR
7	K19	Water	51	3147.0	EUR
9	K23	Water	49	3621.0	EUR
…	…	…	…	…	…
24	K20	Lemonade	43	1926.0	EUR
…	…	…	…	…	…

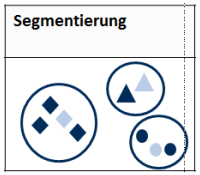
Beispieldaten aus Kießwetter 2007



Segmentierung: Clusterverfahren - Beispiel

- **Wasser (Cluster 0):** Alter: *mittel bis hoch*, Einkommen: *hoch*
- **Fruchtschorle (Cluster 2):** Alter: *mittel*, Einkommen: *mittel*
- **Limonade (Cluster 1):** Alter: *niedrig*, Einkommen: *niedrig*



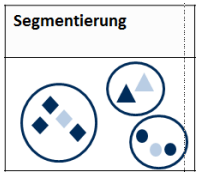


Segmentierung: Clusterverfahren - Proximitätsmaß

Proximitätsmaß: Maß welches den Ähnlichkeitsgrad zwischen zwei Datenobjekten quantifiziert

- **Ähnlichkeitsmaße**, welche die Ähnlichkeit bzw. Homogenität zweier Datenobjekte ausdrücken. Anderseits lassen sich
- **Distanzmaße**, welche die Unähnlichkeit bzw. Heterogenität zweier Datenobjekte ermitteln.

- Auswahl hängt von den Eigenschaften und dem Skalenniveau der Merkmale der betrachteten Datenobjekte ab
  - Weisen die Variablen ein **metrisches Skalenniveau** auf, können Distanzmaße auf der Basis geometrischer Abstandskonstrukte, wie die **euklidische Distanz** oder die Blockmetrik, zum Einsatz kommen.
  - Liegen die Variablen **nominalskaliert** vor, können Maße verwendet werden, die auf die Identifizierung von **Übereinstimmungen** der einzelnen Merkmalsausprägungen ausgelegt sind
    (= Anzahl übereinstimmender Attribute / Anzahl aller Attribute)

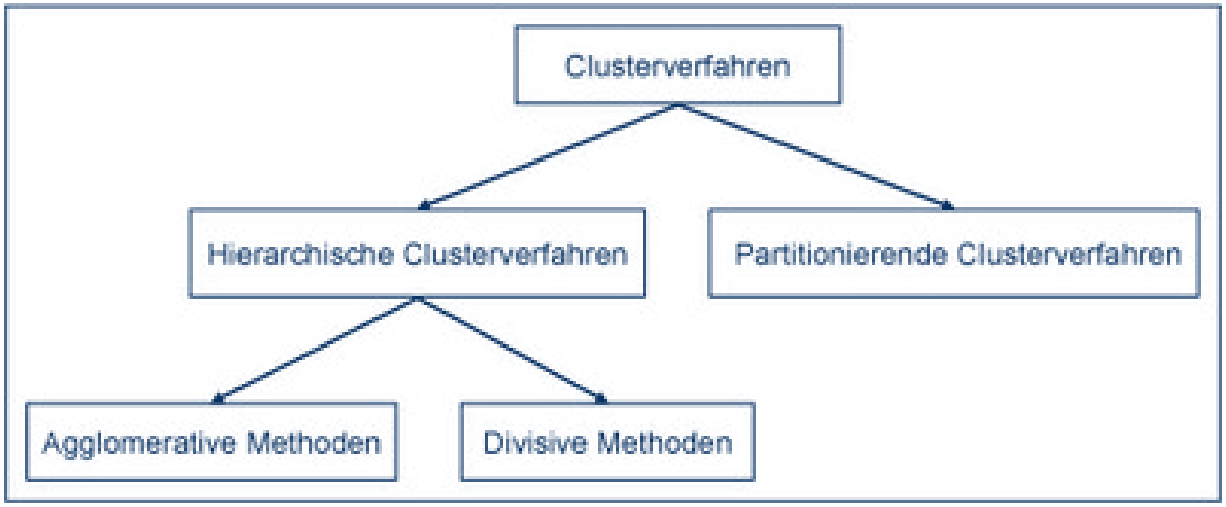


Segmentierung: Clusterverfahren - Vorgehen

partitionierenden Verfahren: versuchen, ausgehend von einer vorgegebenen Gruppeneinteilung, durch den iterativen Austausch der Datenobjekte zwischen den einzelnen Klassen die Gesamtlösung zu optimieren

agglomerativ-hierarchischen Verfahren: verfolgen einen Bottom-Up-Ansatz und gehen bei der Gruppierung von der kleinsten Partition aus. Jedes Datenobjekt repräsentiert zunächst einen Cluster und wird sukzessiv neuen, größeren Gruppen zugeteilt

divisiv-hierarchischen Verfahren: starten mit der gröbsten Partition. Diesem Top-down-Ansatz zufolge befinden sich alle Datenobjekte zunächst in einem einzigen Cluster. Anschließend erfolgt die Aufspaltung der Datenobjekte in homogenere Teilgruppen.



HAW HAMBURG

Abhängigkeitsanalyse
▪ Wenn Objekt A dann auch Objekt B ▪ C + D → E

Abhängigkeitsanalyse: Assoziationsregeln

Ziel der Assoziationsanalyse ist das Erkennen und Bewerten von gemeinsam auftretenden Datenelementen (Items).

- **Items** können Elemente von Mengen oder einzelne Attributwerte von Datensätzen sein.
  Eine Menge von Items wird als **ItemSet** oder auch Itemmenge bezeichnet.
- **Beispiel:** Items: *Artikel*, ItemSet: *Warenkorb*: ***Warenkorb {Artikel a, Artikel b}***

Voraussetzung: Vorhandensein einer Datenbasis bestehend aus einzelnen Transaktionen (z. B. Menge von Kassenbons)

Ergebnis: Regeln der Form **WENN Item *x* DANN Item y** (x→y),
wobei x und y Elemente (Items) von ItemSets sind

Beispiel-Ergebnis: **WENN *Artikel a und Artikel b gekauft werden***
DANN *wird auch Artikel c gekauft*

- **{Milch, Windeln} → {Bier}**

HAW HAMBURG

Abhängigkeitsanalyse
▪ Wenn Objekt A dann auch Objekt B ▪ C + D → E

Abhängigkeitsanalyse: Assoziationsregeln - Metriken

Metriken zur Bewertung der Regeln: **r = (X→Y)**

- **Support (s):** Anteil der Datensätze (ItemSets), die sowohl Item X als auch Item Y enthalten im Verhältnis zu allen Datensätzen (ItemSet)
- **Confidence (c):** Verhältnis der Datensätze für die Regel r = (X → Y) gilt im Verhältnis zu den Datensätzen die den linken Regelteil (X) enthalten

Bewertung:

- Niedriger Support: Spezialregel, geringe Aussagekraft
- Niedrige Confidence: „Falsche“ Regel

Beispiel: r: $\{Milk, Diaper\} \Rightarrow Beer$

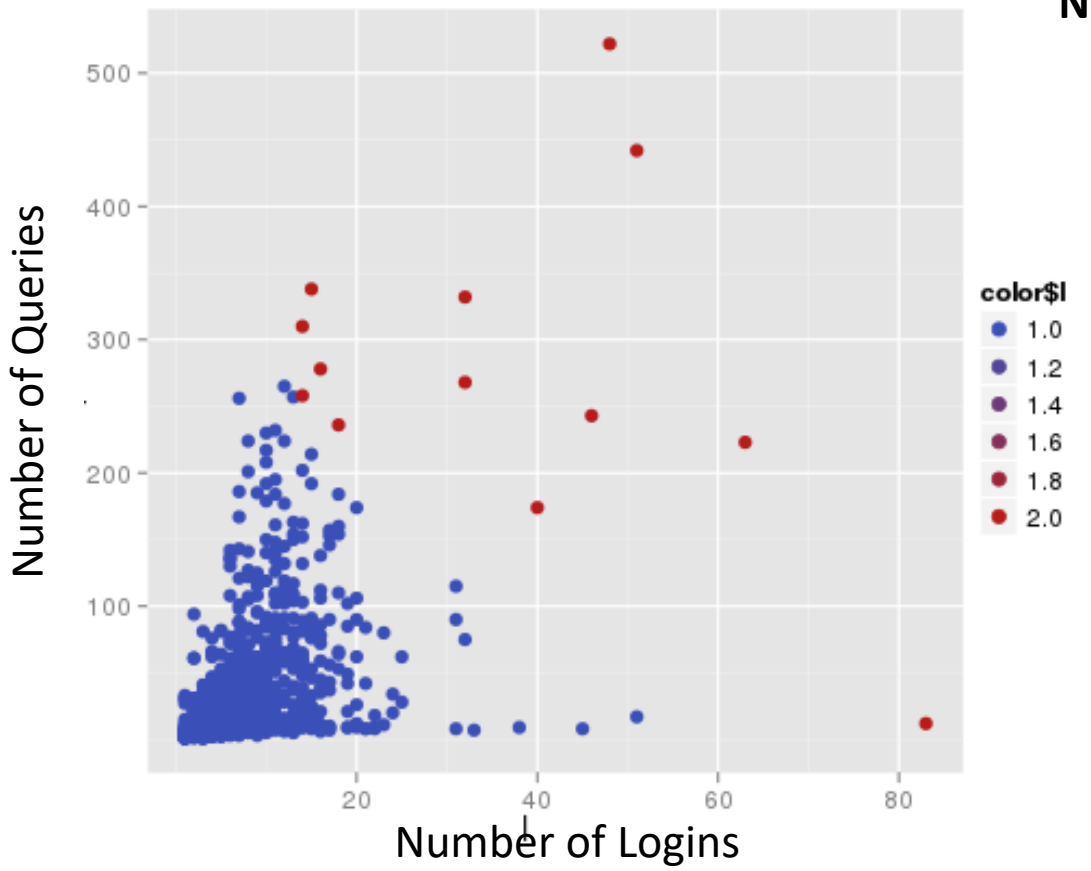
- $s = \frac{\sigma(Milk, Diaper, Beer)}{|T|} = \frac{2}{5} = 0,4$
- $c = \frac{\sigma(Milk, Diaper, Beer)}{\sigma(Milk, Diaper)} = \frac{2}{3} = 0,67$

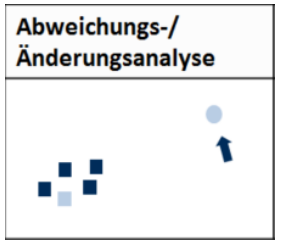
Item Set ID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Abweichungsanalyse (Outlier Detection)

- Analysen zur Erkennung von Wertabweichungen eines Merkmals von zuvor gemessenen oder normativen Werten und Erklärung der Abweichungen durch andere Attribute/ Zusammenhänge

Network Security Tool





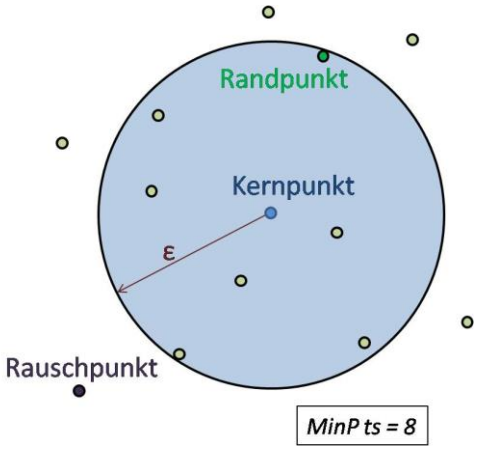
Abweichungsanalyse: Beispiel DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise

Dichteverbundene Clusterverfahren, erkennt Gebiete mit hoher Datenpunktdichte als Cluster, kann sehr gut mit Rauschen umgehen

Dichte: für den Punkt p wird durch die Anzahl der Punkte in einer ε-Umgebung für ein zu bestimmendes **ε** geschätzt.
Die ε-Umgebung von p besteht aus allen Punkten, die höchstens den Abstand ε zu p besitzen.

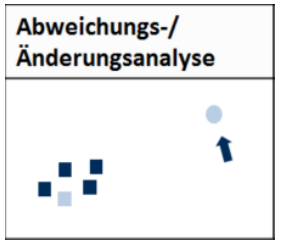
MinPts: legt fest, wie viele Punkte in der ε-Umgebung von einem Punkt p liegen müssen, damit p zu einem Cluster gehört.



Kernpunkt: Die Anzahl der Datenpunkte in der ε-Umgebung des Kernpunkts beträgt mindestens MinPts

Randpunkt: Ein Randpunkt ist kein Kernpunkt, liegt aber in der ε-Umgebung eines Kernpunktes (ist als dichte-erreichbar)

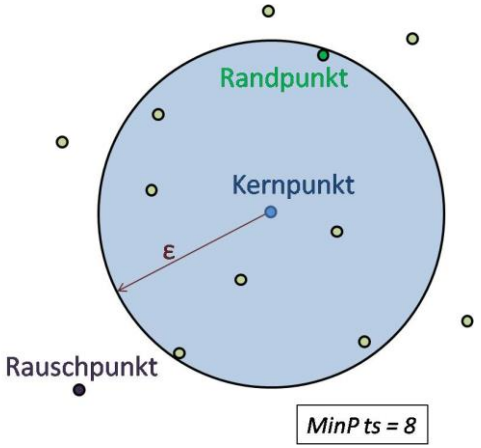
Rauschpunkt: Ein Rauschpunkt ist weder Kern- noch Randpunkt



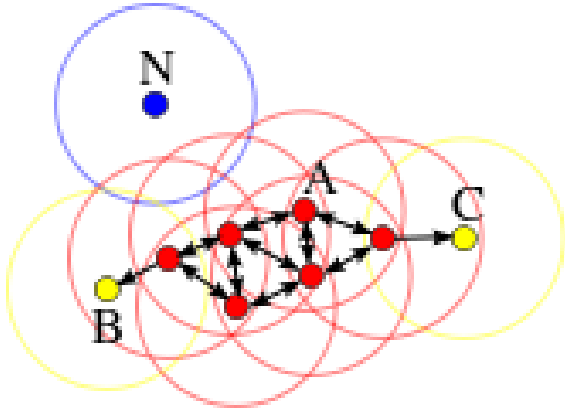
Abweichungsanalyse: Beispiel DBSCAN

Algorithmus

- Benenne Datenpunkte als Kern-, Rand- oder Rauschpunkte.
- *Lösche alle Rauschpunkte.*
- Verbinde Kernpunkte, die innerhalb einer ε-Kugel liegen, durch eine Kante.
- Eine Menge verbundener Kernpunkte bilden ein separates Cluster.
- Weise jeden Randpunkt dem Cluster eines benachbarten Kernpunkts zu.



Kernpunkte: rot
Randpunkte: gelb (dichte-erreichbar)
Rauschpunkt: blau (N)



- Für Outlier Detection: Liste der Rauschpunkte

Text Mining: Definition

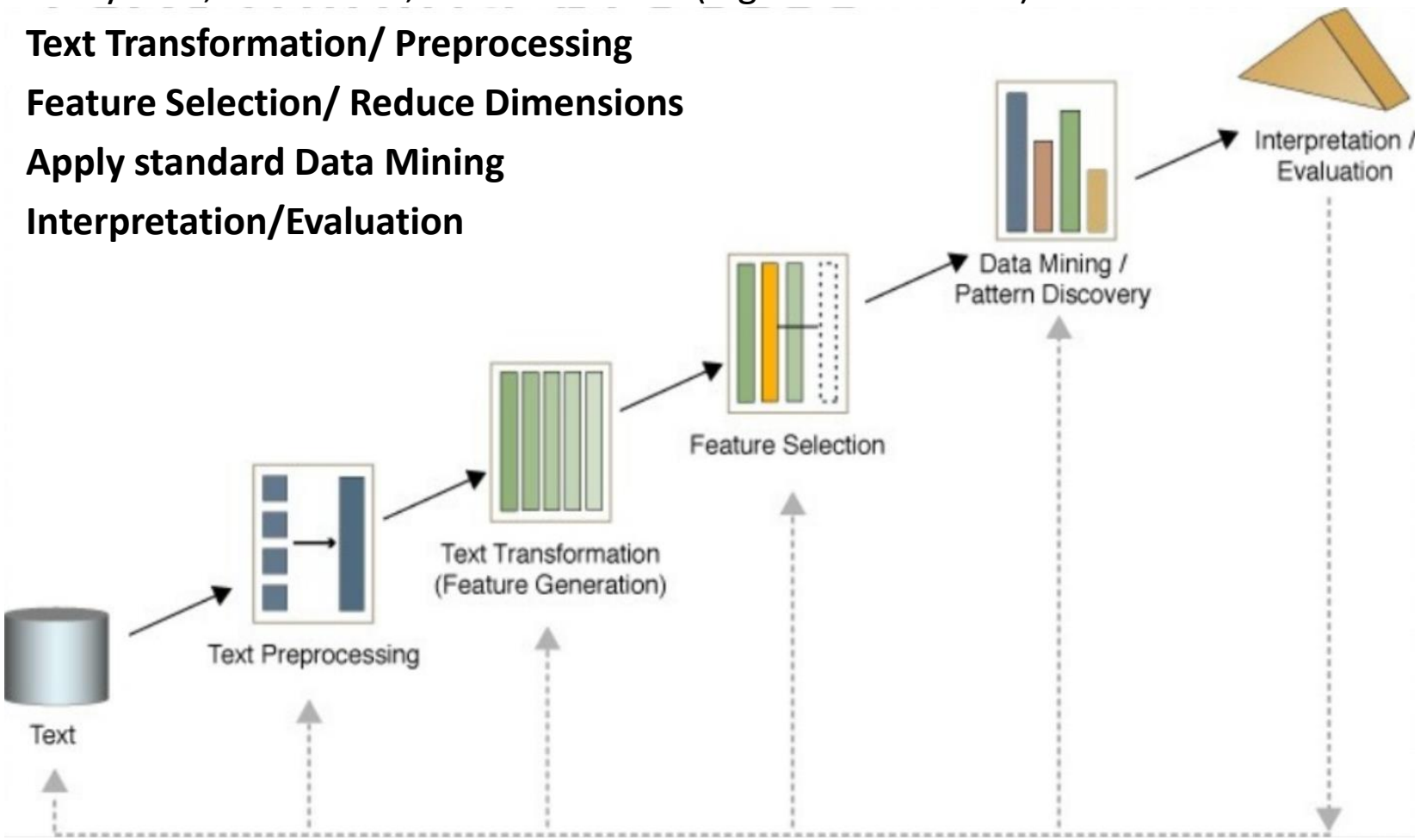
Text mining can be defined — similar to data mining — as the application of algorithms and methods from the fields machine learning and statistics to texts with the goal of finding useful patterns.

For this purpose it is necessary to pre-process the texts accordingly. Many authors use **information extraction** methods, **natural language processing** or some simple pre-processing steps in order to extract data from texts. To the extracted data then data mining algorithms can be applied

- **Natural Language Processing (NLP):** The goal of NLP is to achieve a better understanding of natural language by use of computers. The range of the assigned techniques reaches from the simple manipulation of strings to the **automatic processing of natural language inquiries**. **Linguistic analysis techniques** are used among other things for the processing of text.
- **Information Extraction (IE):** The goal of information extraction methods is the extraction of specific information from text documents. These are stored in data base-like patterns and are then available for further use.

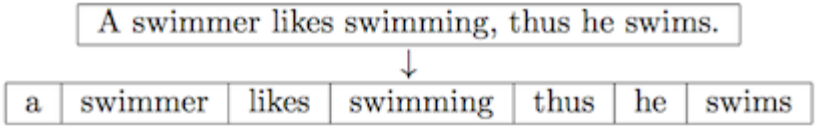
Text Mining Process

1. **Extract Documents:** Document sources can be e.g. a public web site, an internal file system, mail server, social networks (e.g. via Twitter API)
2. **Text Transformation/ Preprocessing**
3. **Feature Selection/ Reduce Dimensions**
4. **Apply standard Data Mining**
5. **Interpretation/Evaluation**

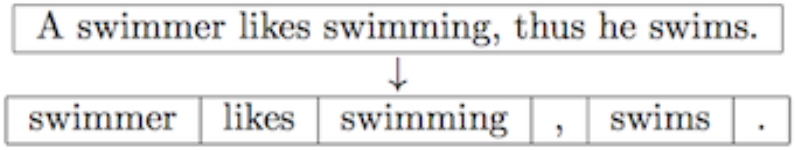


Text Mining Process: Text Preprocessing

Tokenization: is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens

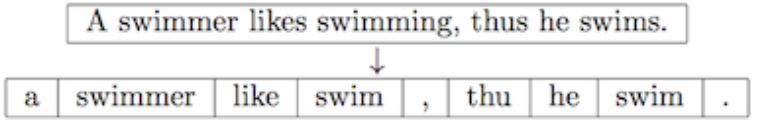


Stop word Removal: stop words usually refer to the most common words in a language. The filtered out before processing of natural language data. This step involves also removing of HTML, XML tags from web pages



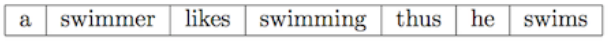
Stemming and Lemmatization: Stemming describes the process of transforming a word into its **root form**. In contrast to stemming, lemmatization aims to obtain **the canonical (grammatically correct) forms** of the words, the so-called lemmas

Stemming

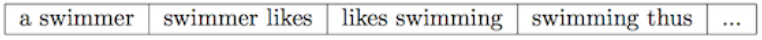


N-grams: In the n-gram model, a token can be defined as a sequence of n items

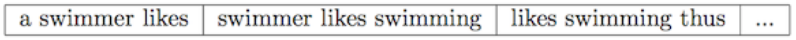
- unigram (1-gram):



- bigram (2-gram):



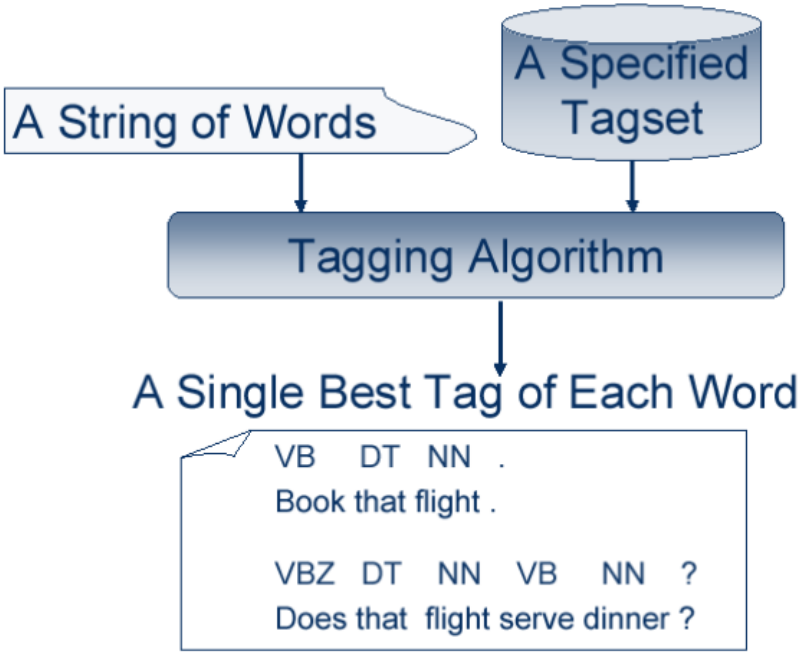
- trigram (3-gram):



Text Mining Process: Text Preprocessing - Part of Speech Tagging

Part of Speech Tagging is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context

- Rule-based taggers: large numbers of hand-crafted rules
- Probabilistic tagger: used a tagged corpus to train some sort of Model



WORD	LEMMA	TAG
the	the	+DET
girl	girl	+NOUN
kissed	kiss	+VPAST
the	the	+DET
boy	boy	+NOUN
on	on	+PREP
the	the	+DET
cheek	cheek	+NOUN

Text Mining Process: Feature Selection

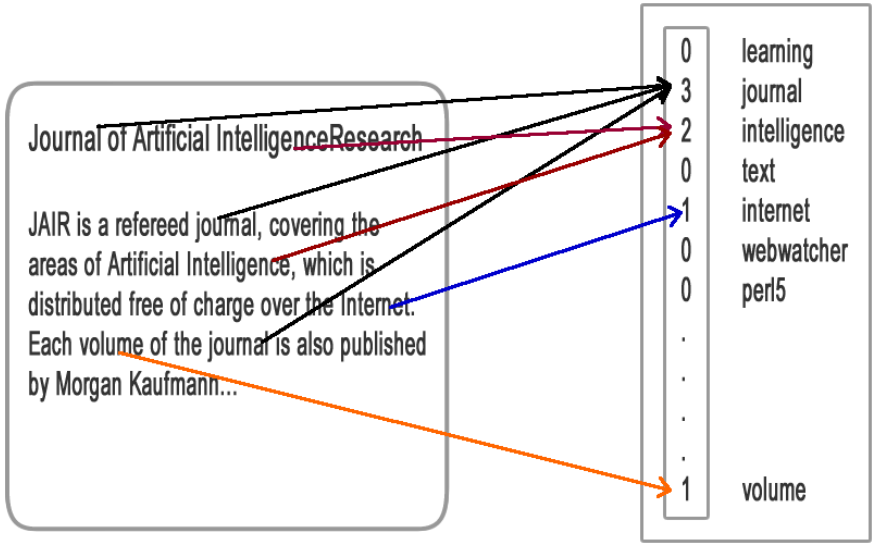
Following types of **potential features** are used to represent a document:

Characters, Words, Terms, Concepts

Bag of Words: The idea is to treat documents as unordered collections of words/ tokens

A **bag of words** can represent a document as vectors where:

- ***Dimension :*** each unique token
- ***Magnitudes:*** token weights



Token weight → term frequency: is defined as the number of times a given term t (i.e., word or token) appears in a document *d*. In practice, the term frequency is often normalized:

$$normalizedtermfrequency = tf(t,d)nd \quad (36)$$

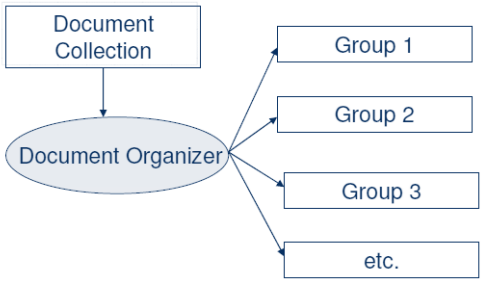
where

- tf(t,d): Raw term frequency (the count of term t in document d).
- nd: The total number of terms in document d.

Text Mining Process: Techniques

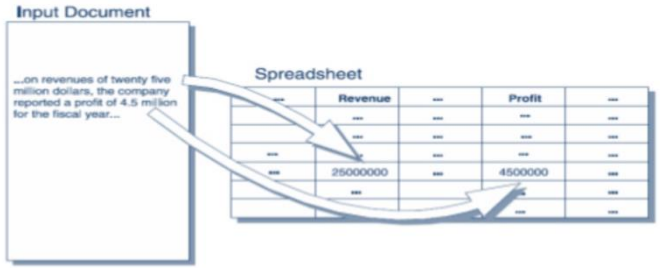
- **Document/ Text Categorization/ Classification**

- **Document/ Text Clustering**



- **Representative Sentences**

- **Information Extraction**



- **Sentiment Analysis**

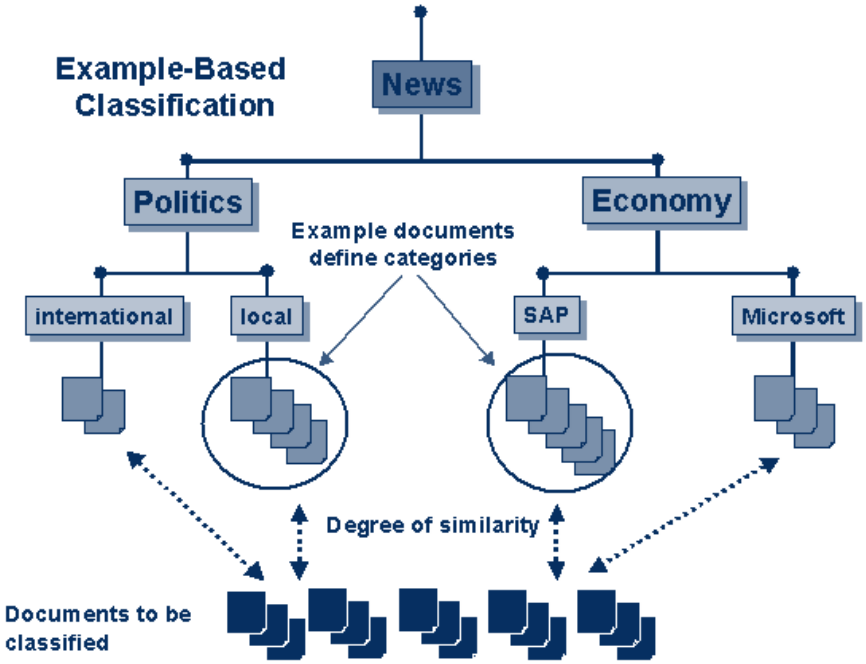
Document/ Text Categorization: Categorization by Taxonomy

Text categorization (a.k.a. **text classification**) is the task of assigning **predefined categories** to free-text documents. It can provide conceptual views of document collections and has important applications in the real world. For example, news stories are typically organized by subject categories (*topics*) or geographical codes

Statistical text categorization uses machine learning methods to learn automatic classification rules based on human-labeled training documents

Sample Taxonomies:

- **IPTC News Codes** - International standard for categorizing news content
- **IAB QAG** - The Interactive Advertising Bureau's quality guidelines for classifying ads

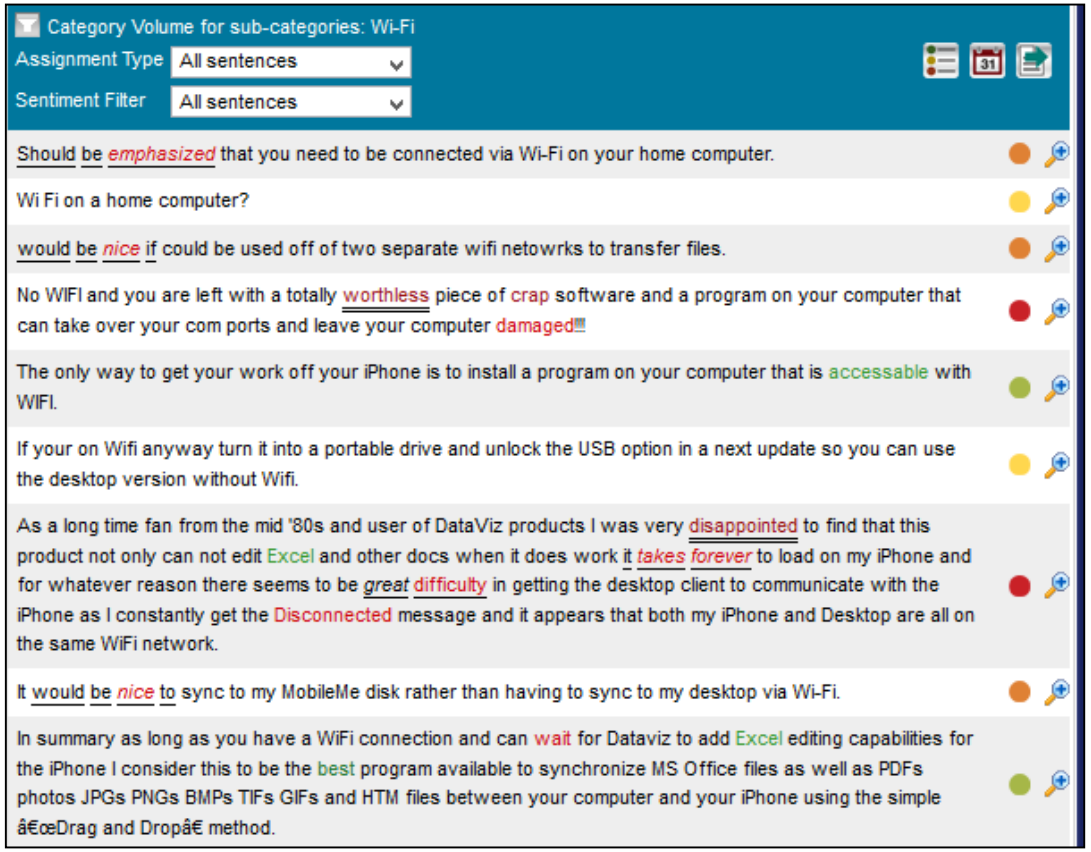


Sentiment Analysis

Sentiment = feelings, subjective impressions, not facts

- Attitudes, Emotions, Opinions, …
- Generally, a binary opposition in opinions is assumed (For/against, like/dislike, good/bad, etc.) → **Polarity**

Sentiment Analysis: Using Natural Language Processing (NLP), statistics, or machine learning methods to extract, identify, or otherwise characterize the sentiment content of a text unit.



Sentiment Analysis: Potential Questions and Challenges

Questions

- Is this product review positive or negative?
- Is this customer email satisfied or dissatisfied?
- Based on a sample of tweets, how are people responding to this ad campaign/product release/news item?
- How have bloggers' attitudes about the president changed since the election?

Challenges

- People express opinions in complex ways
- In opinion texts, lexical content alone can be misleading
- Intra-textual and sub-sentential reversals, negation, topic change common
- Rhetorical devices/modes such as sarcasm, irony, implication, etc
- Short phrases may be just as important as words: “lowest prices”, “high quality”

“Dear <hardware store>

Yesterday I had occasion to visit <your competitor>. The had an excellent selection, friendly and helpful salespeople, and the lowest prices in town.

You guys suck.

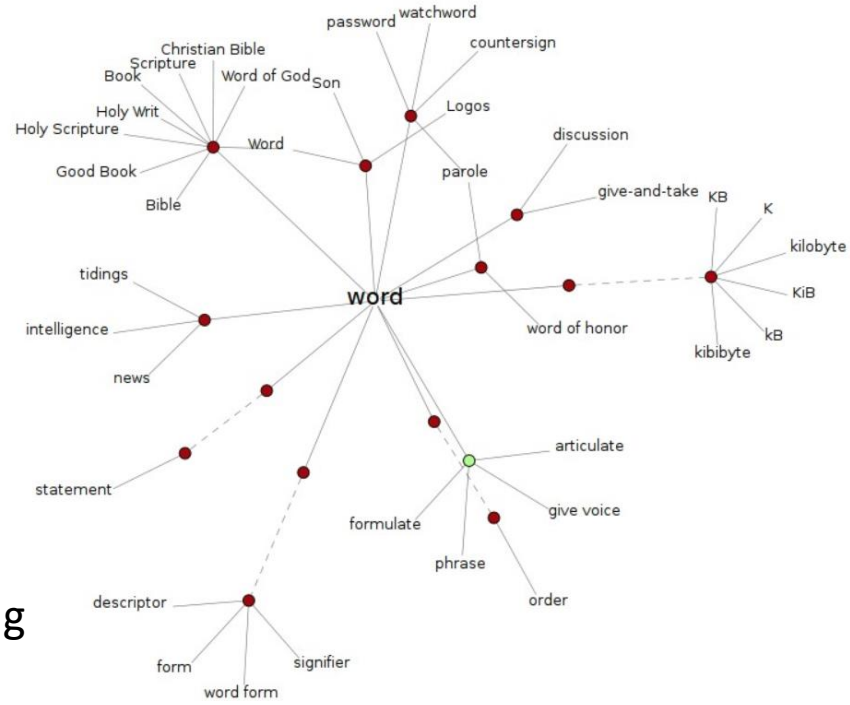
Sincerely,”

Sentiment Analysis: Polarity Keywords

Heuristic/hand made references (lexicon-based)

Wordnet: A lexical database for English with emphasis on synonymy

- Nouns, verbs, adjectives and adjectives are grouped into synonym sets (**synset**)
- Words are linked according to lexical and conceptual relations (creating a “net”)

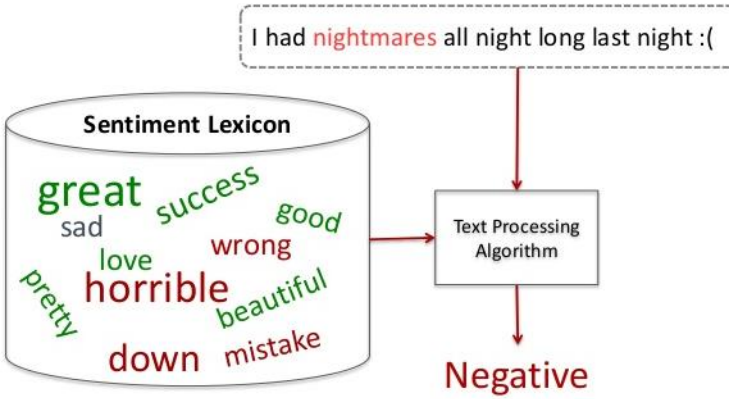


SentiWordNet: A lexical resource for opinion mining

- Based on Wordnet synsets
- each synset is assigned three sentiment scores: positivity, negativity, and objectivity

- PosScore [0,1] : positivity measure
- NegScore [0,1]: negativity measure
- ObjScore [0,1]: objective measure

$$\text{ObjScore} = 1 - (\text{PosScore} + \text{NegScore})$$



Text Mining - Example Sentiment Analysis

Sentiment Analysis for Comments in an App Store

Category Volume Report
Total: 99

Table

Category	Distinct Document	% of Document	Sentiment Score	Preview
Type: Category (14 Items)				
Start	8	8.08	1.72	
Application	17	17.17	0.71	
Excel	23	23.23	0.53	
File	25	25.25	0.35	
Document	30	30.30	0.09	
App	52	52.53	0.04	
Doc	24	24.24	0.04	
Version	25	25.25	0.02	
Exchange	37	37.37	-0.01	
Palm	19	19.19	-0.03	
Prices	36	36.36	-0.33	
Wi-Fi	12	12.12	-0.47	
Paste	8	8.08	-0.61	
Dataviz	24	24.24	-0.64	
Type: Others (1 Item)				
Global Other	3	3.03	-0.09	

ASU 2013

Category Volume for sub-categories: Wi-Fi

Assignment Type: All sentences
Sentiment Filter: All sentences

Should be *emphasized* that you need to be connected via Wi-Fi on your home computer.

Wi Fi on a home computer?

would be *nice* if could be used off of two separate wifi netowrks to transfer files.

No WIFI and you are left with a totally worthless piece of crap software and a program on your computer that can take over your com ports and leave your computer damaged!!!

The only way to get your work off your iPhone is to install a program on your computer that is accessable with WIFI.

If your on Wifi anyway turn it into a portable drive and unlock the USB option in a next update so you can use the desktop version without Wifi.

As a long time fan from the mid '80s and user of DataViz products I was very disappointed to find that this product not only can not edit Excel and other docs when it does work it *takes forever* to load on my iPhone and for whatever reason there seems to be *great* difficulty in getting the desktop client to communicate with the iPhone as I constantly get the Disconnected message and it appears that both my iPhone and Desktop are all on the same WiFi network.

It would be *nice* to sync to my MobileMe disk rather than having to sync to my desktop via Wi-Fi.

In summary as long as you have a WiFi connection and can wait for Dataviz to add Excel editing capabilities for the iPhone I consider this to be the best program available to synchronize MS Office files as well as PDFs photos JPGs PNGs BMPs TIFs GIFs and HTM files between your computer and your iPhone using the simple â€œDrag and Dropâ€ method.

http://clarabridge.com