Vorbesprechung

Praktikum:
Data Warehousing und Mining

Team

- Matthias Bracht
  - bracht AT kit DOT edu
- Frank Eichinger
  - eichinger AT kit DOT edu
- Ursula Kotzur
  - ursula DOT kotzur AT gmx DOT de
- Emanuel Pongracz
  - emanuel DOT pongracz AT gmx DOT net

Agenda

- Behandelte Themen im Praktikum
- Organisatorisches
  - Termine
  - Scheinvoraussetzungen
  - Gruppeneinteilung
- 1. Vorlesungsteil:
  - Vorgehen beim Data Mining
  - Preprocessing

Motivation

- Grosse Datensammlungen in Unternehmen
  - Jede Abteilung hat eigene Datenbestände
  - Daten beschreiben alle Aspekte der Organisation
- Wissen in Daten nicht offensichtlich
  - Zu viele Attribute
  - Niemand hat Überblick über alle Daten
  - Mitarbeiter wechseln, alte Daten werden uninterpretierbar
  - Daten im Unternehmen verstreut
- „We are drowning in information, but starving for knowledge!“ (John Naisbitt)
- Thema
  - Wie in der Vorlesung:
    Wie kommt man in diesem Szenario zu Wissen?
  - … praktisch an Beispielen mit marktüblicher Software

Data Warehousing

- Ziel
  - Integration von Unternehmensdaten in zentralen Datenbestand
  - Anfragen / Analysen auf diesem Datenbestand

- Charakteristika
  - Materialisierte Sichten auf unterschiedliche andere Quellen
  - Daten aus unterschiedlichen Quellen im Unternehmen
  - Daten sind meist aggregiert
  - ⇨ OLAP (Online Analytical Processing)

OLTP vs. OLAP
(Datenbank vs. Data Warehouse)

- Anfragecharakteristika

	transaktional	analytisch
Fokus	Lesen, Schreiben, Modifizieren, Löschen	Lesen, periodisches Hinzufügen
Transaktionsdauer und -typ	Kurze Lese- / Schreibtransaktionen	Lange Lesetransaktionen
Anfragestruktur	Einfach strukturiert	komplex
Datenvolumen einer Anfrage	Wenige Datensätze	Viele Datensätze

nach Bauer, Günzel (Hrsg):

Data-Warehouse-Systeme – Architektur, Entwicklung, Anwendung

Data Warehousing in diesem Praktikum

- Benutzung der Tools
  - Oracle und Cognos ReportStudio
- Oracle
  - Anfragen auf dem relationalen Datenbestand
  - Datenwürfel modellieren
  - Datenwürfel erstellen
- Cognos
  - Anfragen auf dem Datenwürfel
  - Erstellen von Analysen

Data Mining

- Menge von Techniken
  - Klassifikation
    Ist der Kunde kreditwürdig?
  - Regression
    Wieviel verdient der Kunde?
  - Clustering
    Welche Kundengruppen gibt es?
  - Association Rules
    Welche Produkte werden zusammen gekauft?

- Ziel
  - Finden interessanter Muster und Eigenschaften in großen Datenbeständen

Data Mining in diesem Praktikum

- Benutzung der Tools
  - IBM SPSS Modeler (früher: Clementine)
  - Weka
  - Knime
  - FrIDA
- Daten aus dem Data Mining Cup

Synergieeffekte Data Warehousing und Data Mining

- Aufwändigster Schritt: Datenbereinigung
  - Fällt bei Data Warehousing und Data Mining an
  - ⇨ Daten des Data Warehouse eignen sich für Data Mining
- Data Mining als Analysekonzept im Data Warehouse
- **Problem:**
  - Data Mining benötigt operative, transaktionsorientierte Daten
    (z. B. Kassenbons)
  - Data Warehouse hält häufig aggregierte Daten vor
    ⇨ feingranulare Informationen gehen verloren

Data-Mining-Cup

- Aufgabenstellung ab Donnerstag unter
  - http://www.data-mining-cup.de
- Teilnahme als Team „Inst_KIT_1“
- Kombination der einzelnen Gruppenlösungen
- Gesamtabgabe: 31. Mai 2010

Data-Mining-Cup 2009

- Thema im letzten Jahr: Bücherverkauf
  - Fragestellung: Wo wird welches Buch wie oft verkauft?
  - Ziel: Einkauf angemessener Büchermengen
- Unsere Einreichung: 5. Platz weltweit
- Präsentation der Lösung durch vier Studenten in Leipzig

DATA MINING CUP 2009

DATA MINING CUP 2009 Description of Features

Feature	Type	Description	Attributes
ID	Integer	• Unique location id	random unique key
WGxxxxx	Integer	• Number of total items sold within 12 months within a category • Categories with 5 digits • First digit: Information about the type of product e.g. hardcover vs. paperback • Second to fifth digit: Hierarchical information about the type of content e.g. second digit Fiction and third digit subcategory Science-Fiction	independent variables
T1…T8	Integer	• Number of items sold within 12 months per title	target value

Organisatorisches

Praktikum:
Data Warehousing und
Data Mining

Tutorien

- Teams
  - Besuchen gemeinsam ein Tutorium
  - Geben DMC-Lösungen zunächst gemeinsam ab
- Tutorien
  - Je 1,5 Stunden pro Team, Woche
- Tutoren
  - Betreuen je zwei Teams
  - Führen Tutorien durch
  - Sind Ansprechpartner nach den Veranstaltungen

Weitere Veranstaltungen

- Vortrag: Prof. Thomas Ruf, GfK
  - voraussichtlich am Montag, 21.6., 9:45 Uhr
  - Data Warehousing und Mining in der Marktforschung

- Ausflug zu IBM nach Böblingen
  - voraussichtlich am Freitag, 25.6., ganztägig

Scheinvoraussetzungen

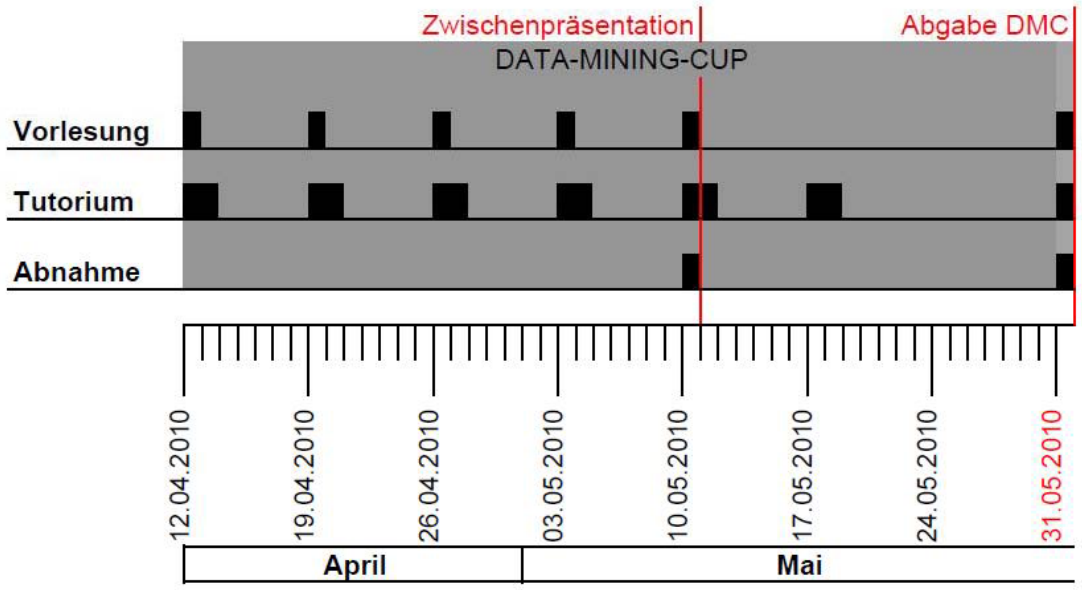
- Für jede Leistung sind Punkte erreichbar
  - Zwischenpräsentation Data Mining Cup: 2 Punkte
    - Jedes Team präsentiert Lösung in 15 Minuten
  - Ergebnis Data-Mining-Cup: 7 Punkte
    - Bis zu 7 Punkte für Lösung der Tutoriumsgruppe
  - Weitere Blöcke: 9 Punkte

  - Summe: 18 Punkte

- Scheinvoraussetzung:
  - Erlangen von 10 Punkten und mehr, Bearbeitung jeder Aufgabe und Teilnahme an der Exkursion!

- Schein ist unbenotet
- Praktikum ist prüfbar!

Veranstaltungstermine bis Ende Mai



- Danach: zwei bis drei weitere Blöcke zu den Themen Data Warehousing und Mining

Was passiert heute noch?

- Bestätigung der Teilnahme
- Vorlesungsarbeitsbereich unter https://studium.kit.edu/?
- Verteilung auf Tutorien
- Danach:
  - Data Mining: Vorgehen
  - Preprocessing

Tutorientermine

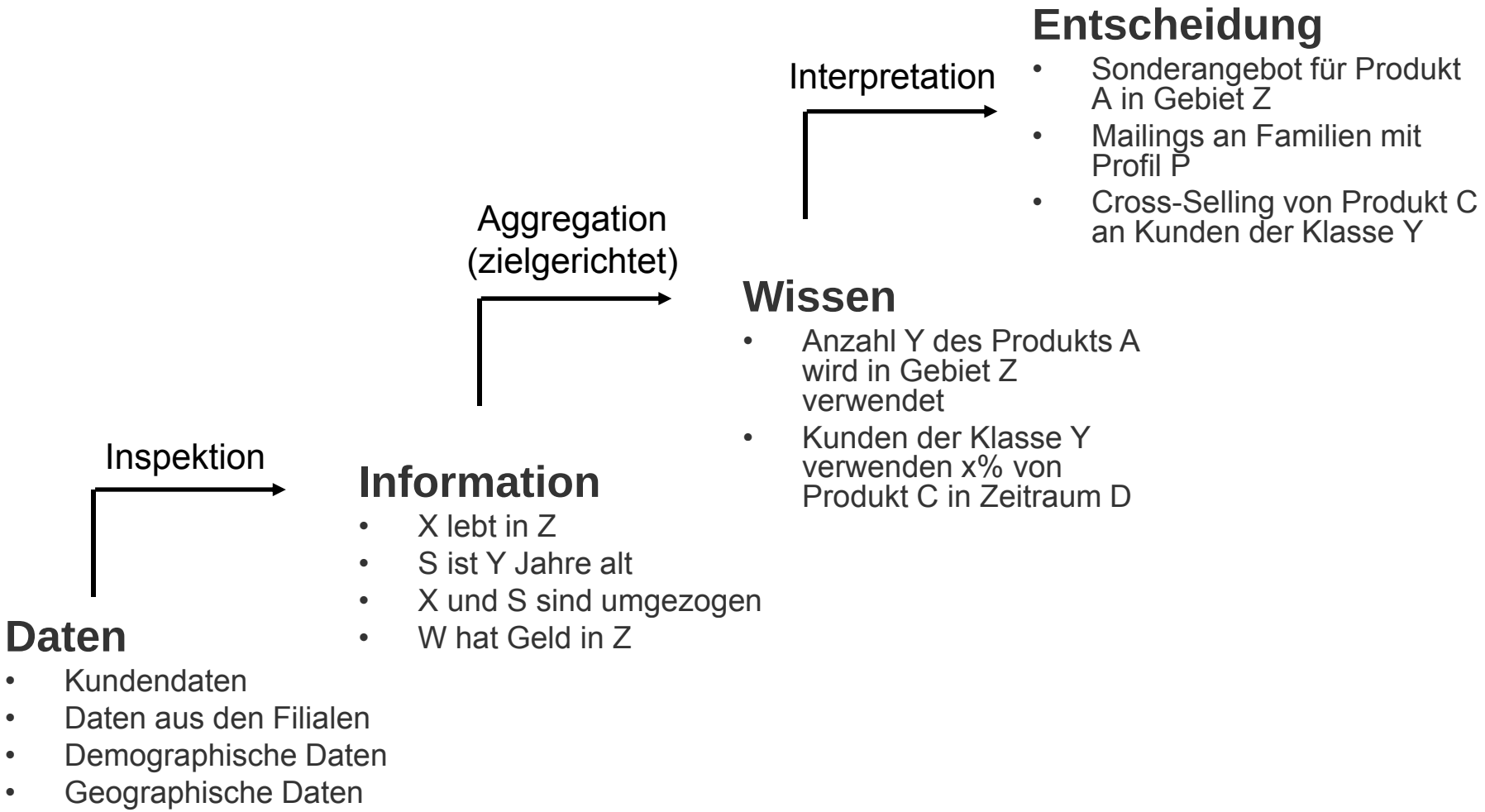
Ursula Kotzur	Montag	11:30 Uhr	David, Philippe, Michael, Alexander, Thomas M., Elvi
	Mittwoch	9:45 Uhr	Patricia, Muhannad, Hong, Marusa, Jingyu, Dominik
Emanuel Pongracz	Montag	11:30 Uhr	Fabian, Daniel, Stefan, Tihomir, Ivan, Thomas K., Nguyen
	Montag	14:00 Uhr	Andreas, Sven, Zhen, Raimund, Andriy, Patrick, Jens

Literaturempfehlungen

- J. Han und M. Kamber: "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2006.
- I. H. Witten und E. Frank: "Data Mining - Practical Machine Learning Tools and Techniques", Morgan Kaufmann, 2005.
- D. Hand, H. Mannila und P. Smyth: "Principles of Data Mining", MIT Press, 2001.
- L. I. Kuncheva: "Combining Pattern Classifiers", Wiley-Interscience, 2004.
- A. Bauer, H. Günzel: "Data Warehouse Systeme – Architektur, Entwicklung, Anwendung", dpunkt.verlag, 2004.
- T. Mitchell: “Machine Learning”, McGraw Hill, 1997.

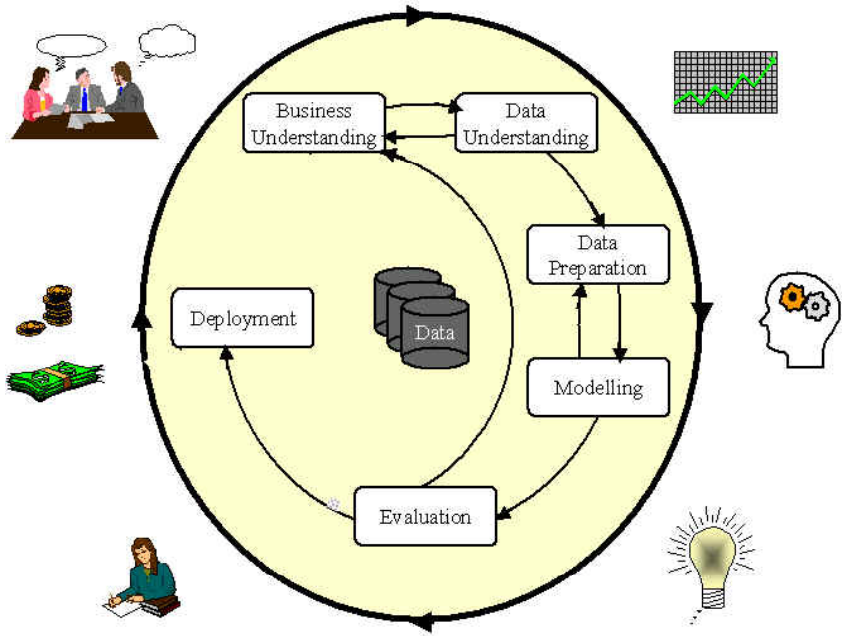
Data Mining: Vorgehen

Von Daten zur Entscheidung (Gianotti und Pedreschi)



Vorgehensmodell: CRISP-DM

- „CRoss Industry Standard Process for Data Mining“
- Zusammenschluss verschiedener Hersteller- und Anwenderfirmen
- Definiert allgemeines Prozessmodell
- “Modeling” ist eigentlicher Data-Mining-Schritt



www.crisp-dm.org

Business Understanding

- Identifiziere Geschäftsziele
- Aneignen von Domänenwissen
- Analysiere Situation und Umfeld
- Formuliere Data-Mining-Ziele (und Erfolgskriterium!)
- Erstelle Projektplan
  - Zeitaufwand:
    - Data Understanding 20-30%
    - Data Preparation 50-70% (!)
    - Modeling + Evaluation 10-20%
    - Deployment 5-10%

Data Understanding

- Initiale Daten sammeln
  - Quellen identifizieren und zusammenstellen
- Daten beschreiben
  - Metadaten, z.B. Volumen, Tabellen und Attribute
- Daten erforschen
  - Visualisierung, Anfragen, Statistik
- Datenqualität sicherstellen
  - Missing Values, ...

Data Preparation

- Selektieren
- Säubern
  - Falsche und fehlende Werte ersetzen
- Zusammenstellen
  - Abgeleitete/aggregierte Attribute berechnen
  - Numerische Attribute normieren
- Integrieren
  - Daten aus verschiedenen Quellen
  - Semantische Ungleichheiten beachten
- Formatieren

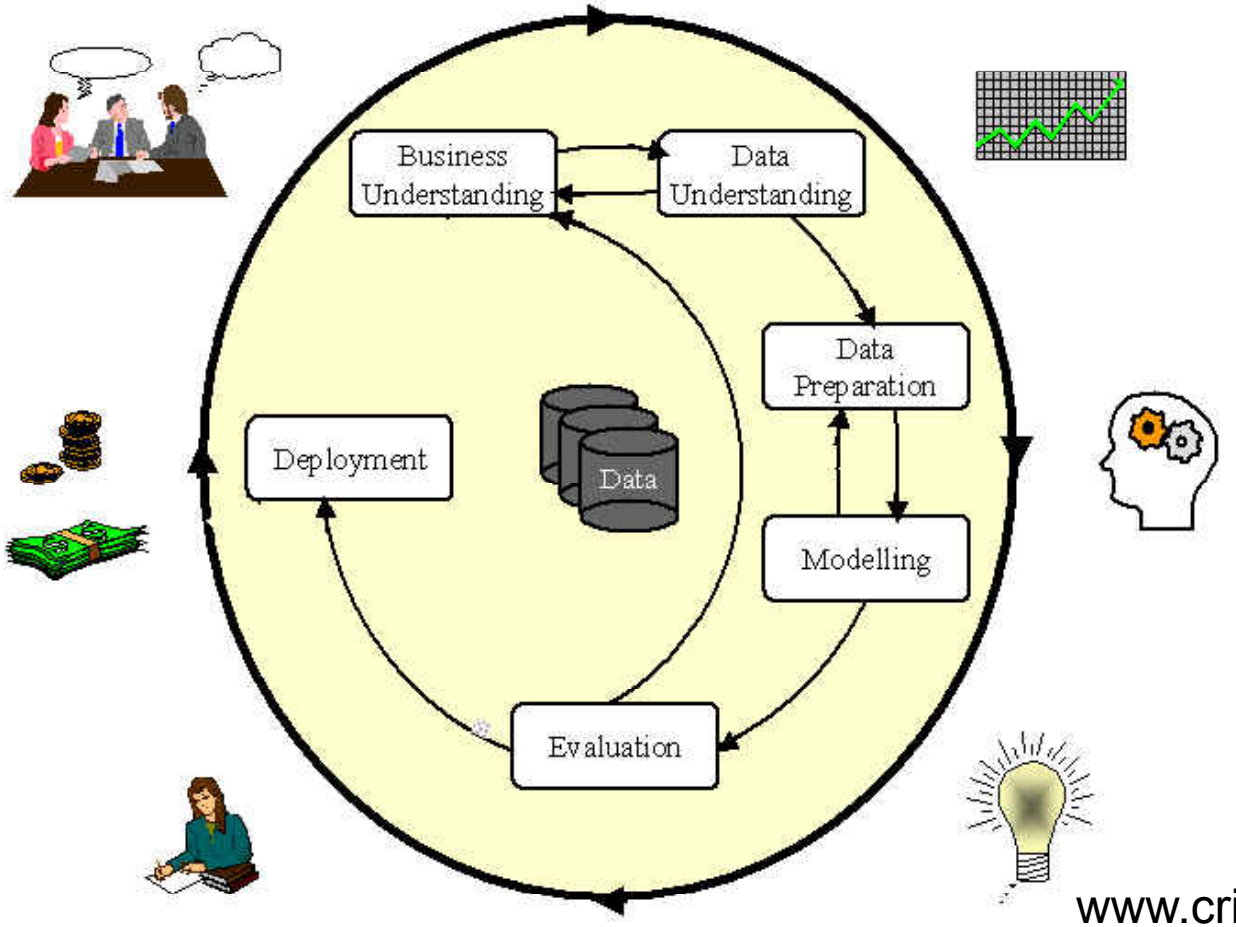
Modeling

- Verfahren auswählen
- Trainings- und Testdaten separieren
- Modell lernen
  - Parameter geeignet einstellen
    (in der Regel mehrere Iterationen erforderlich)
- Ergebnis prüfen
  - Anhand von allgemeinen Kriterien
  - Im Vergleich zu anderen Verfahren
  - Ggf. neue Parameter (oder Verfahren!) und nochmal bauen…

Evaluation & Deployment

- Evaluation
  - Messen an den Business Objectives
  - Fehler im Prozess identifizieren
- Deployment
  - Deployment-Plan
  - Wie lange soll das Modell genutzt werden?
  - Erfahrungen sammeln und dokumentieren

CRISP-DM



www.crisp-dm.org

Data Preprocessing

Beispiel: Teilnehmerliste eines Praktikums

- Ziel:
  - Alle Studenten sollen teilnehmen!
- Vorgehen
  - Liste wurde handschriftlich ausgefüllt
  - Dann in Teilnehmerdatenbank übertragen
- Probleme
  - Feld männlich/weiblich fehlt
    - Ist Conny männlich oder weiblich?
  - Feld Fachsemester ist nicht vielsagend
    - ein Masterstudent ist im 3. Semester, ein anderer im 9.
  - Beim Übertragen in Datenbank treten Fehler auf
    - E-Mail-Adressen sind undeutlich geschrieben
    - Übertragender ist im Stress und liest nur oberflächlich

Teilnehmerliste des Praktikums II

- Probleme (fortges.)
  - Einträge im Feld „Studiengang“ (Auszug): „InfoDipl.“, „InfoMa“, „InfoMaster“, „Infowirt.“, Infowirt.Ma“, „Info Erasm“
    - Wer ist in einem Diplomstudiengang?
    - Suche nach „Dipl(om)“ findet nicht alle Treffer
- Was ist zu tun?
  - Hier:
    - Alle Angemeldeten können teilnehmen.
    - „Politisch korrekt“
  - Aber:
    - Was, wenn Unternehmenserfolg von Prognose abhängt?
  - Dann:
    - Datenqualität essentiell
    - Daten müssen vorverarbeitet werden

Eigenschaften von Produktivdaten

- Daten sind meist…
  - Unvollständig
    - Enthalten NULL-Werte
    - Enthalten Aggregate
    - Interessante Informationen fehlen
  - Verunreinigt:
    - Enthalten Fehler
    - Enthalten Ausreißer
  - Inkonsistent:
    - Daten verschiedener Quellen unterscheiden sich

Data Preprocessing – Vorgehen

- Analyse der Daten
  - „Ansehen“ von einzelnen Tupel / Aggregaten von Tupeln
  - Deskriptive Statistik
  - Visualisierung der Eingangsdaten

- Durchführung des Data Preprocessing
  - Datenbereinigung
  - Datenintegration
  - Datentransformation
  - Datenreduktion

„Ansehen“ der Daten

- Nutzen:
  - Oft sind Eigenschaften am leichtesten beim direkten Betrachten der Daten zu entdecken
- Meist erster Schritt des Data Preprocessing
- Beispiele
  - Entdecken von NULL-Werten
  - Skalentypen der Werte
  - Größe der Wertebereiche
  - Diskrepanz zwischen Attributlänge und Datenlänge
  - …

Skalentypen

Skalentyp	Wertebereich	Mögliche Operationen	Beispiele
Nominale Größen	diskret, endlich	Gleichheit	Geschlecht Augenfarbe
Ordinale Größen	diskret, endlich, Ordnung	Gleichheit, größer / kleiner als	Prüfungsnoten Schulabschluss
Intervallgrößen	kontinuierlich bzw. ganzzahlig, unendlich, „gleichabständig“	Gleichheit, größer / kleiner als Differenz	Celsius-Skala Datum
Ratiogrößen	kontinuierlich bzw. ganzzahlig, unendlich, „natürlicher Nullpunkt“	Gleichheit größer / kleiner als Differenz Verhältnis	Abstand Alter Masse Kelvin-Skala

- Anwendbarkeit der Statistiken abhängig vom Skalentyp
  - Mittelwert des Geschlechts
  - Modalwert der Prüfungsnoten

Deskriptive Statistik

- Nutzen
  - Identifikation typischer Dateneigenschaften
  - Identifikation von Ausreißern und Datenfehlern

- Wichtige Statistiken
  - Maße für die Zentralität
    - Mittelwert
    - Median
    - Modalwert
  - Maße für die Verteilung
    - Interquartilsabstand
    - Varianz
    - Skewness (Schiefe)
    - …

Maße für Zentralität

- Mittelwert

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \ldots + x_N}{N}$$

  - Entspricht `average (avg( ))` in SQL
- Median
  - „Mittlerer Wert“ aller sortierten Werte
  - Durchschnitt der zwei „mittleren Werte“ bei gerader Wertanzahl
- Modalwert
  - Häufigster Wert
  - Abhängig von Anzahl der Werte: unimodal, bimodal, …

Maße für die Verteilung

- Quartil
  - Seien Daten aufsteigend sortiert
  - 1. Quartil enthält unterste 25% der sortierten Werte
  - 2. Quartil enthält untere 25% - 50% der sortierten Werte
  - usw.
- Interquartilsabstand
  - Abstand zwischen oberem und unterem Quartil
  - Einfaches Maß für die Verteilung der Daten
- Varianz

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right]$$

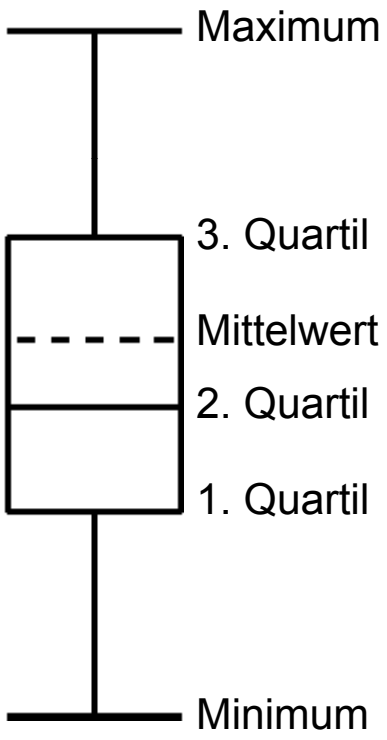
  - Nur sinnvoll, wenn Mittelwert als Zentrum der Daten
  - Maß für die Verteilung der Daten

Visualisierung der Eingangsdaten

- Nutzen
  - Menschliches Gehirn ist auf Erfassung graphischer Inhalte optimiert
  - Mehrere Aspekte können simultan untersucht werden
- Wichtige Visualisierungen
  - Boxplot
  - Histogramm
  - Scatterplot

Visualisierung - Boxplot

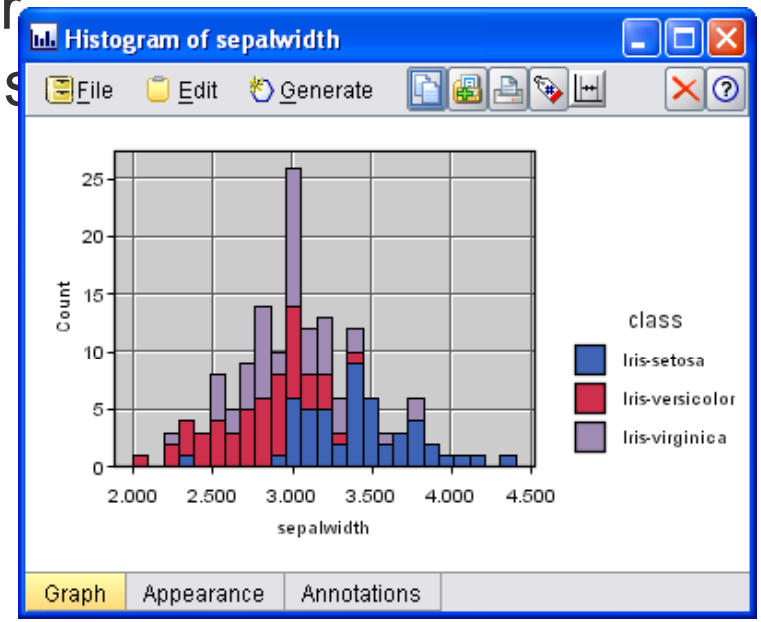
- Fasst mehrere statistische Maße zusammen
- Zeigt
  - Mittelwert, Quartile, Minimum Maximum, Interquartilsabstand
- Nutzen
  - Finden der Verteilung
  - Finden von Ausreißern



Visualisierung - Histogramm

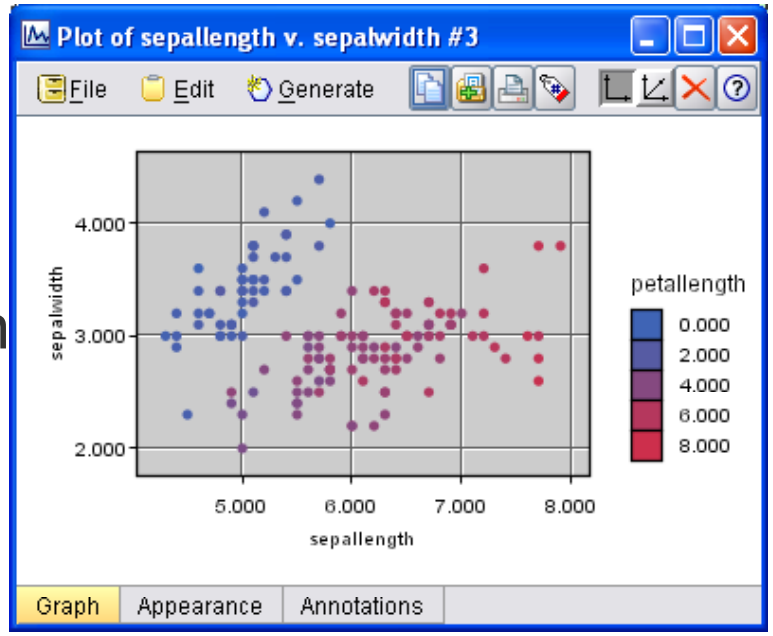
- Zeigt die Verteilung einzelner, numerischer Attribute
- Verteilung abhängig von kategorischem Attribut möglich
- Darstellung der Anzahl
- Prozentsatz interpretierbar
- Kenngröße gegebenenfalls in Buckets gruppiert

- Nutzen
  - Finden von Ausreißern
  - Finden der Verteilung
  - Erkennen von Tupel-charakteristika



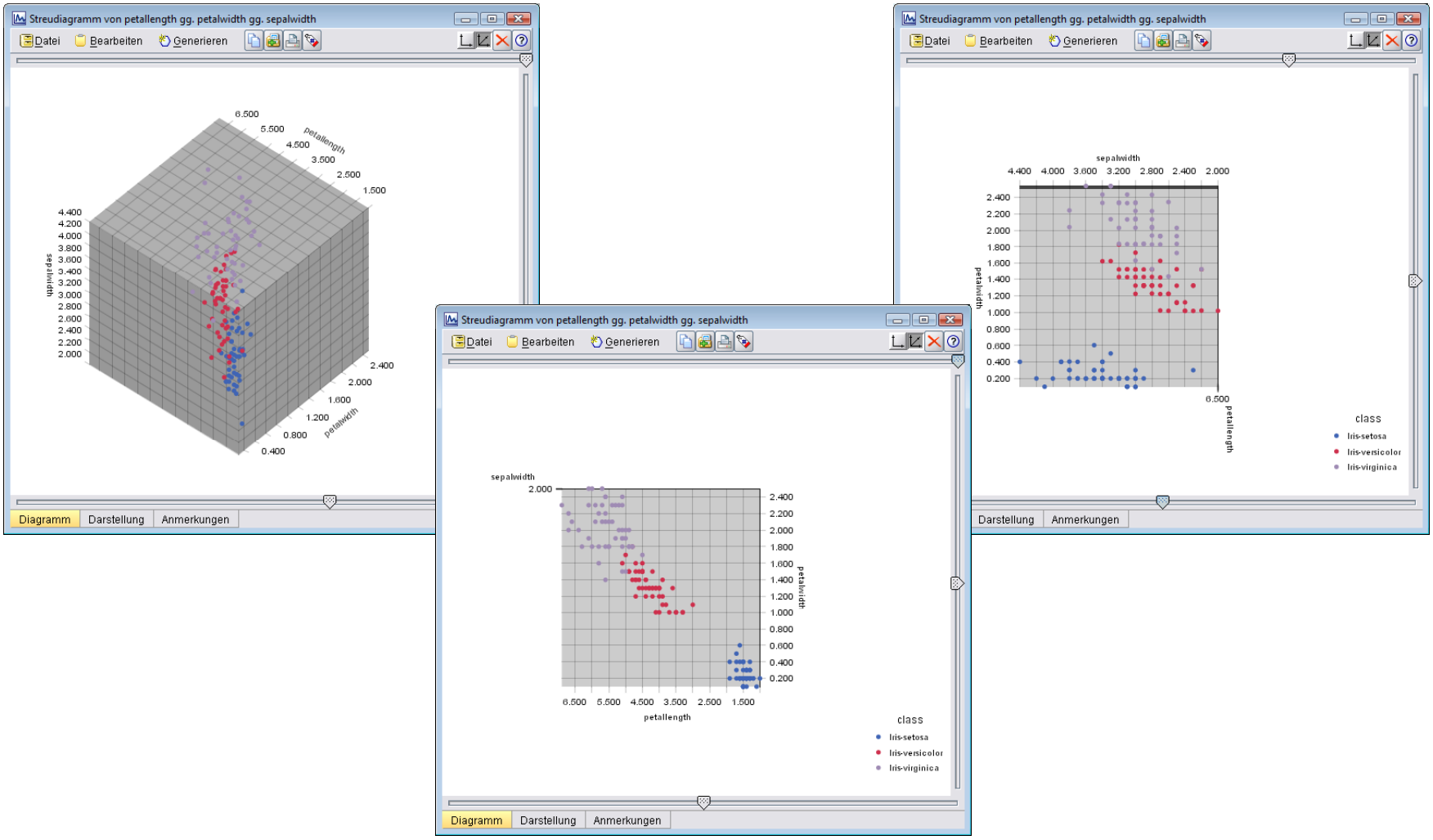
Visualisierung – Scatterplot

- Visualisiert einzelne Tupel
- Bis zu drei numerische Attribute angebbar
- Formatierung der Datenpunkte abhängig von weiteren Attributen

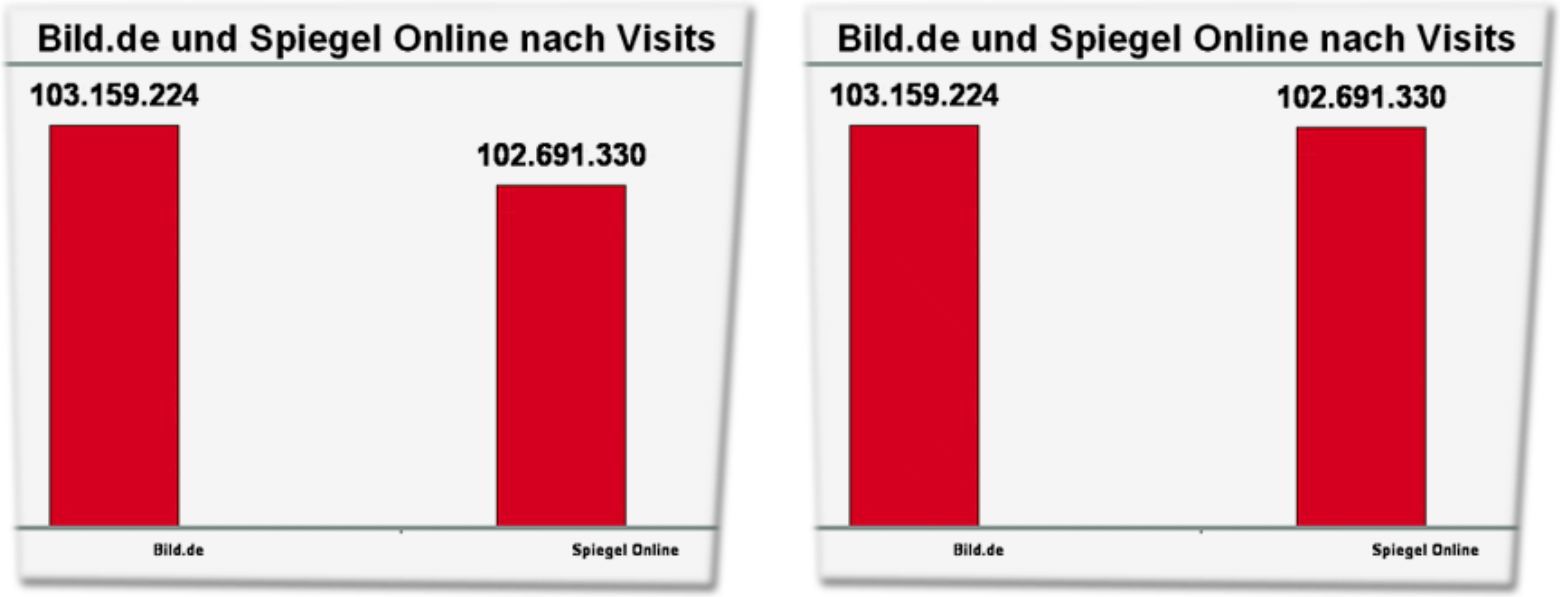


- Nutzen
  - Finden von Korrelationen
  - Finden von Clustern
  - Finden von Ausreißern

Visualisierung – dreidimensionaler Scatterplot

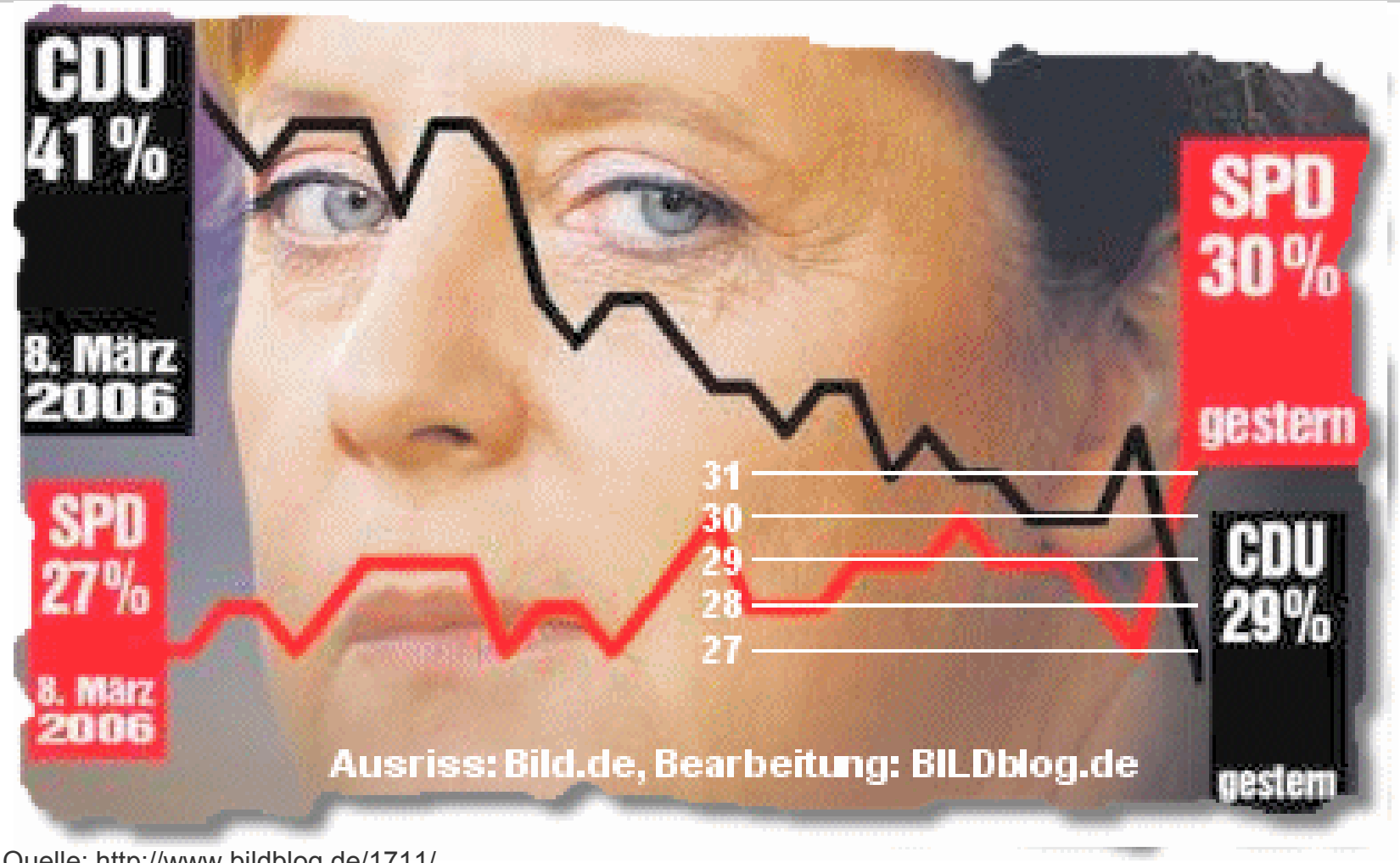


Exkurs: Risiken (I)



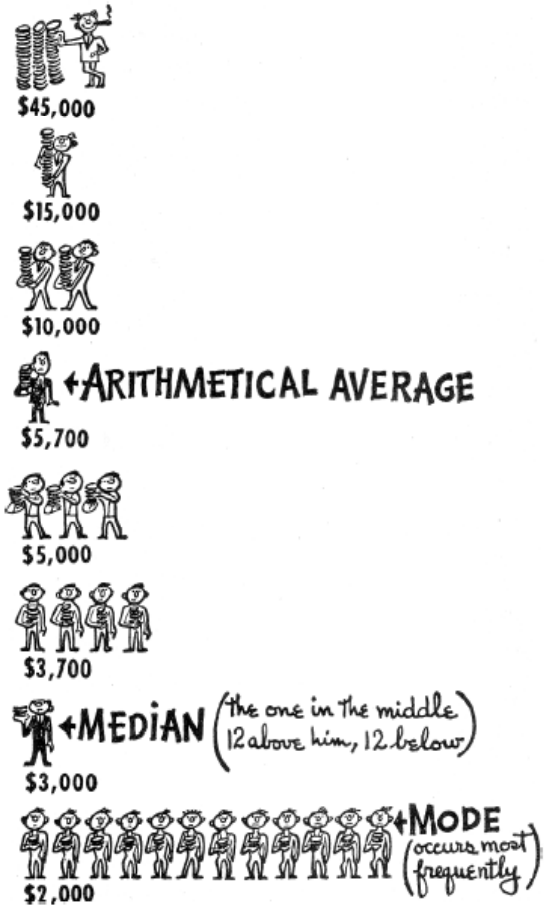
Quelle: http://www.bildblog.de/11395/

Exkurs: Risiken (II)



Quelle: http://www.bildblog.de/1711/

Exkurs: Risiken (III)



»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«

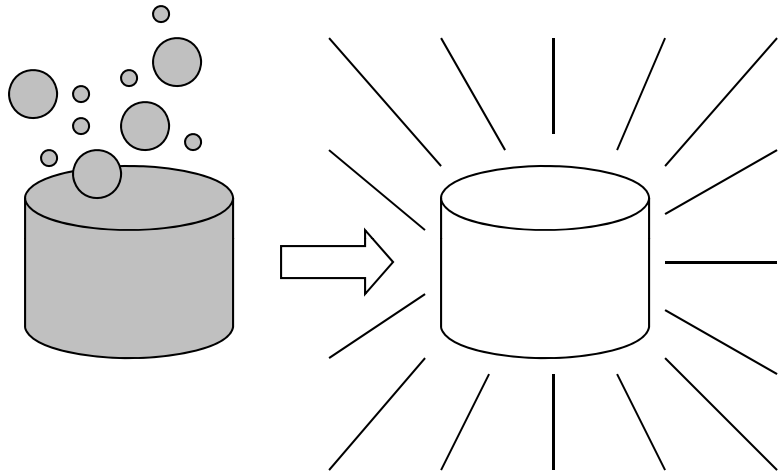
Quelle: D. Huff: How to Lie with Statistics bzw. W. Krämer: So lügt man mit Statistik.
Nach einer Auswahl von C. Borgelt: Intelligent Data Analysis

Data Preprocessing – Vorgehen

- Analyse der Daten
  - „Ansehen“ von einzelnen Tupeln / Aggregaten von Tupeln
  - Deskriptive Statistik
  - Visualisierung der Eingangsdaten

- Durchführung des Data Preprocessing
  - Datenbereinigung
  - Datenintegration
  - Datentransformation
  - Datenreduktion

Datenbereinigung



- Beseitigung von…
  - fehlenden Werten
  - verunreinigten Daten

Beseitigung von fehlenden Werten I

- Ignorieren von Tupeln
  - Notgedrungen bei Klassifikation: Klasse fehlt
  - Sinnvoll, wenn in Tupel viele Werte fehlen
  - Sonst vorsichtig:
    - Fehlender Wert kann Logik sein
    - Kritisch, wenn Häufigkeit der fehlenden Werte unter Attributen unterschiedlich
    - Beispiele:
      - Beruf: Hausfrau
      - Sensor fällt bei großer Kälte aus
- Manuelles Auffüllen
  - Nur bei geringer Zahl fehlender Werte sinnvoll
  - Auffüllender muss über Expertenwissen verfügen
- Ersetzen durch globale Konstante
  - Beispiel: Alles durch „unbekannt“ oder „-∞“
  - Aber vorsichtig:
    - Kann als besonderer Wert interpretiert werden

Beseitigung von fehlenden Werten II

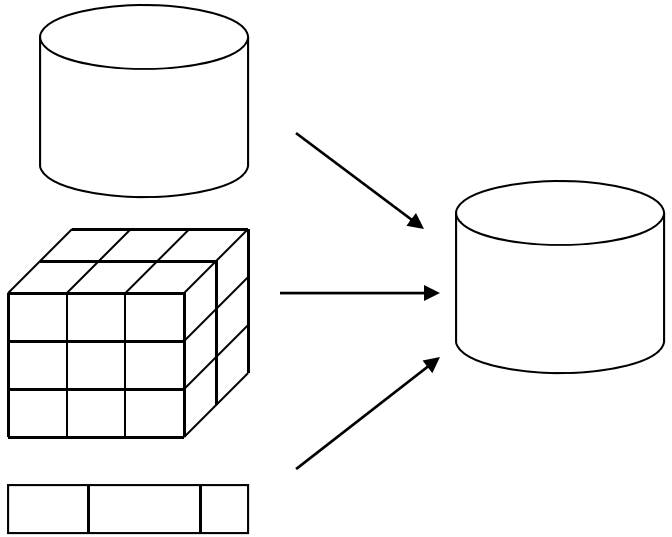
- Einsetzen des Mittelwertes
  - Beispiel: Mittelwert des Einkommens
  - Aber: nur bei metrischen Attributen sinnvoll
  - Vorsicht: Daten werden gebiast
- Einsetzen des Mittelwertes der Klasse
  - Beispiel: Mittelwert des Einkommens über alle in derselben Kreditrisiko-Klasse
  - Aber: nur bei metrischen Attributen sinnvoll
  - Vorsicht: Daten werden gebiast
- Einsetzen des wahrscheinlichsten Wertes
  - Finden des Wertes über Modalwert
  - Finden mit Klassifikationsalgorithmen
  - Vorsicht: Daten werden gebiast
- Wichtig:
  - Einige Algorithmen können mit fehlenden Daten umgehen

Beseitigung von verunreinigten Daten

- Binning
  - …mit gemeinsamer Häufigkeit
    - Ersetzen durch Mittelwert
    - Ersetzen durch Median
    - Ersetzen durch nächste Bucketgrenze
  - …mit gemeinsamer Breite der Buckets
  - Hilft bei Glätten der Eingangsdaten
- Regression
  - Daten werden durch Regressionsfunktion beschrieben
- Clustering
  - Daten werden geclustert
  - Dabei können Ausreißer identifiziert werden

- Hinweis:
  - Verfahren können auch zur Datenreduktion genutzt werden

Datenintegration



- Ziel…
  - Integration von Daten aus verschiedenen Quellen

Datenintegration

- Daten aus Unternehmensquellen
  - … ähnlich Data Warehousing
  - Jetzt nicht Fokus

- Daten aus zusätzlichen Quellen
  - Frei verfügbar
    - Postleitzahlen zu Adressen
    - Umrechnungskurse zwischen Währungen
  - Extern zukaufbar
    - Schufa-Daten
    - Daten von der Post
    - Diverse andere Datenquellen

Datenintegration - Schwierigkeiten

- Entitätsidentifikationsproblem
  - Attributnamen:
    - Stimmt „Kunden-ID“ mit „Kundennummer“ überein?
  - Attributwerte:
    - Ist „m“ in Geschlecht gleich „männlich“?
- Korrelationsanalyse
  - Finden von Redundanzen:
    - Mehrinformation Jahres- gegenüber Monatseinkommen
- Skalierungsprobleme
  - Beispiele:
    - Temperaturen in Celsius bzw. Fahrenheit
    - Einkommen in Dollar bzw. Euro

Datentransformation

-3; 45; 12,0; 17

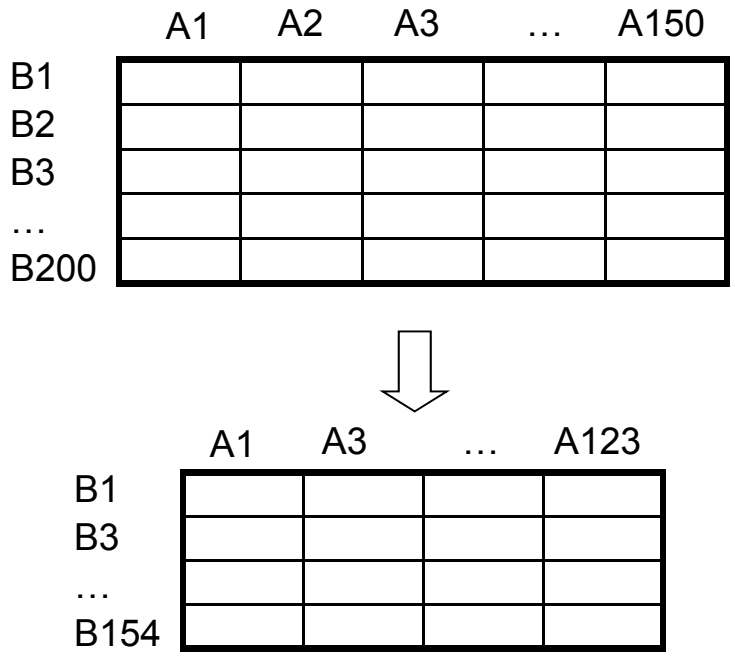
-0.03, 0.45, 0.12, 0.17

- Ziel
  - Vorbereitung der Daten für das Data Mining

Datentransformation

- Bereinigung von Daten
  - Wie eben
- Aggregation
  - Aggregat über Tageseinnahmen zu Monatseinnahmen
  - Besonders interessant, wenn auch Data Warehouse erstellt wird
- Generalisierung
  - Daten werden auf sinnvolles Niveau aggregiert
  - Beispiel: Von Adresse auf Stadt
- Normalisierung
  - Skalierung auf überschaubaren Wertebereich
  - Beispiel: auf 0,0 bis 1,0
- Attributgenerierung
  - Zusammenfassen mehrerer Attribute zu einem
  - Beispiel: Umrechnung in Vergleichswährung

Datenreduktion



- Ziel:
  - Eingrenzen des Curse of Dimensionality

Feature Selection

- Vorteile
  - Gewonnene Regeln sind leichter interpretierbar
  - Skalierbarkeit ermöglicht
- Vorgehen (allgemein)
  - Bestimmen des Attributwertes
    - … über statistische Signifikanz
    - … über Information Gain
- Vorgehen (Alternativen)
  - Schrittweise Vorwärtsselektion
    - Ausgangssituation: Leere Attributmenge
    - Rekursive Erweiterung um je ein Attribut
  - Schrittweise Rückwärtsselektion
    - Ausgangssituation: Vollständige Attributmenge
    - Rekursive Entfernung um je ein Attribut
  - Entscheidungsbauminduktion
    - Entscheidungsbaum wird generiert
    - Alle Attribute im Entscheidungsbaum werden genutzt
- Optional:
  - Expertenwissen nutzen

Sampling

- Motivation
  - Zu viele Lerndatensätze
  - Balancieren der Klassengröße
- Vorgehen
  - Auswahl einzelner Tupel
- Einfaches zufälliges Sampling
  - Zufälliges Ziehen von Tupeln
- Stratified Sampling
  - Attribut wird gewählt
  - Anteil der einzelnen Attributwerte in Ausgangsdaten gleich dem Anteil im Sample

Was fehlt noch?

- Ausblick auf nächste Woche
  - DMC-Aufgabe
  - Klassifikation, ggf. Regression
- Accounts beantragen
- Termin für die folgenden Treffen
  - Nächste Woche Montag 9:45 Uhr
- Hinweise zur verwendeten Software: in den Tutorien.
- http://dbis.ipd.uni-karlsruhe.de/1523.php
- Wiki: http://www.ipd.uni-karlsruhe.de/~ipd/wiki/mediawiki-1.5.6/index.php/DWM-Praktikum (User: Dbisstud)
- Ab Donnerstag: DMC-Aufgabe ansehen!