Grundlagen des Data Mining

Vorlesung

Winfried Kurth, Lehrstuhl für Praktische Informatik / Grafische Systeme der BTU Cottbus

Systematische Einordnung:

- Bereich "Datenbanken"
- "Künstliche Intelligenz"
- Techniken aus Statistik, Algebra (Verbandstheorie), Logik, Komplexitätstheorie...

Leistungsnachweis / Abschluss des Moduls:

- Prüfung (schriftlich oder mündlich)
- für Schein auch Hausarbeit möglich

Skript:

http://www-gs.informatik.tu-cottbus.de/~wwwgs/gdm2_vorles.htm

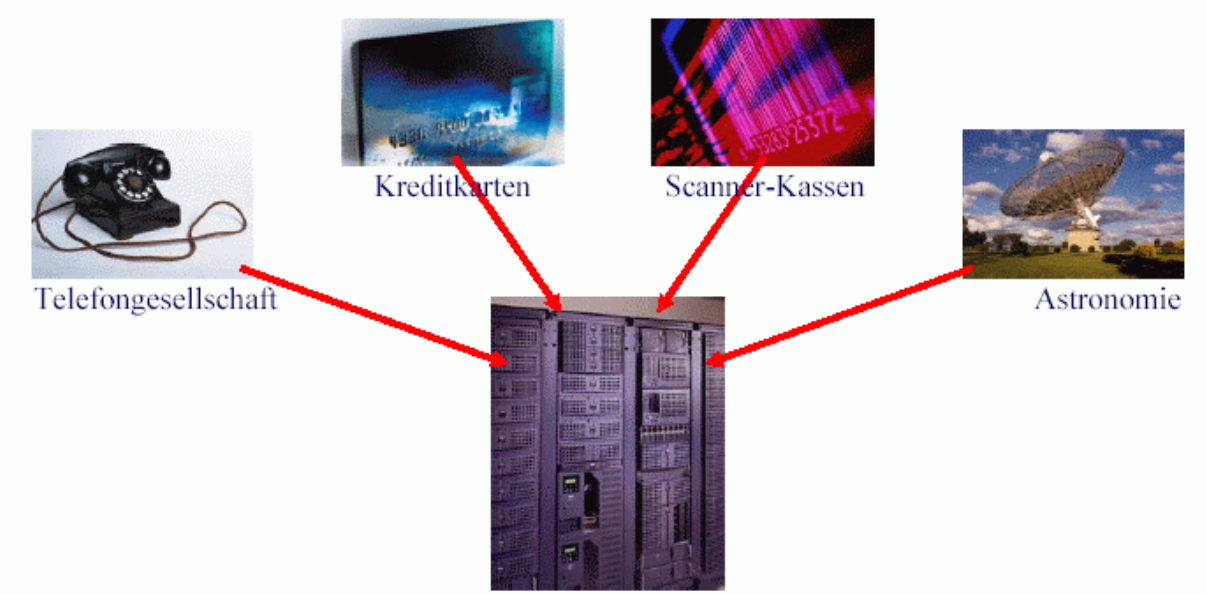
Literatur:

Ester & Sander: Knowledge Discovery in Databases
Mitchell: Machine Learning
Witten & Frank: Data Mining
Fischer: Algorithmisches Lernen
Ganter & Wille: Formale Begriffsanalyse
... und weitere

vollständige bibliografische Angaben siehe
http://www-gs.informatik.tu-cottbus.de/~wwwgs/gdm_lit.htm

1. Einleitung:
Motivation, Grundbegriffe, Methodenüberblick

Motivation



Bewältigung von Information

- Computerisierung des täglichen Lebens sorgt für die Erzeugung großer Datenmengen
  - Point-of-sale, Internet-Einkauf (& Browsing), Kreditkarten, . . .
  - Kreditkartentransaktionen, Einkaufsmuster, Produkt-Präferenzen, Zahlungshistorie, Besuchshistorien . . .
- Reise: Ein Trip einer Person generiert Information zu Zielort, Flug/Bahn-Präferenzen, Hotel, Mietwagen, Adressen, Restaurantwahl . . .
- Die Daten können sehr einfach (in Datenbanken) gespeichert werden
- Die Daten können nicht mehr manuell verarbeitet oder untersucht werden

(aus Koch 2003)

Überangebot von Daten

- Nur eine kleine Menge der Daten, die gesammelt wird, wird analysiert (Schätzungen gehen von maximal 5% aus)
- Große Mengen der Daten werden gesammelt und gespeichert, da befürchtet wird, ansonsten wichtige Informationen zu verlieren/übersehen
- Die Datenmenge wächst so schnell, dass ältere Daten nie analysiert werden
- Datenbanksysteme können keine Anfragen der folgenden Form beantworten
  - "Wer könnte wahrscheinlich Produkt X kaufen"
  - "Liste alle Berichte zu Problemen auf, die dem gegebenen ähnlich sind"
  - "Markiere alle betrügerischen Transaktionen"
- Aber dies sind wichtige Fragestellungen!

Beispiele für Datenflut:

Industrielle Prozeßdaten

Analyse der Altpapieraufbereitung bei Kübler+Niethammer
8 Deinkingzellen à 54 Sensoren à 9000 Meßwerte/Tag
➪ 3.888.000 Meßwerte/Tag

Umsatzdatenbanken

Warenkorbanalyse für die Scannerkassen bei *WalMart*
20 Millionen Transaktionen/Tag
➪ Datenbank 24 Terabytes

Molekularbiologie

Human Genome Database Project zur
Entschlüsselung des genetischen Codes des Menschen
60 000–80 000 Gene — 3 Milliarden DNA-Basen

Visuelle Daten

NASA *Earth Observing System* sammelt
Oberflächenbilder tieffliegender Satelliten
50 Gigabytes/Stunde

Textinformationen

Ca. 10 Milliarden HTML-Seiten im *World Wide Web*
Suchmaschinen, Indexierer, Extrahierer, Emailfilter

(aus Schukat 2002)

- Computer speichern in Unternehmen und Behörden Daten in großer Zahl
  - Kundendaten, Lieferantendaten, Personaldaten
  - Lagerverwaltung, Produktdaten, Daten zur Planung von Produktionsprozessen
  - Vertriebsplanung, ...
- Meist besteht eine enge Kopplung mit Datenbanksystemen. Viele Einzelinformationen sind abrufbar.
- **Regelhaftigkeiten, Strukturen, Muster** bleiben meist verborgen !

"The key in business is to know something that nobody else knows."

PHOTO: HULTON-DEUTSCH COLL

— Aristotle Onassis

"To understand is to perceive patterns."

— Sir Isaiah Berlin

PHOTO: LUCINDA DOUGLAS-MENZIES

(aus Koch 2003)

Begriffsklärungen

Daten

Beispiele für Daten:

- *Kunde X hat Bier gekauft !*
- *Der QRS-Dauer des Patienten ist 140 msec !*

Eigenschaften von Daten:

- beschreiben Einzelfälle (Personen, Zeitpunkte, Orte)
- sind vielfach in großer Zahl vorhanden
- sind oft leicht zu beschaffen und zu erfassen (Internet, Scannerkassen, etc)
- lassen meist keine Vorhersagen zu

Wissen

Beispiele von Wissen:

- Der *5-er-Bus* fährt im 10-Minuten-Takt
- Die Erdbeschleunigung beträgt 9.81 m/s^2

Kennzeichen von Wissen:

- beschreibt allgemeine Muster, Strukturen, Gesetze und Prinzipien
- soll aus möglichst wenigen und einfachen Aussagen bestehen
- ist i.a. schwer zu finden bzw. zu beschaffen
- läßt Voraussagen zu

Bewertungskriterien für Wissen

Wissen muss bewertet werden!

Nicht jede allgemeine Aussage ist wichtig oder nutzbar!

Kriterien nach denen man Wissen bewerten kann:

- Korrektheit : *Wie wahrscheinlich?*
- Allgemeinheit : *Wann und unter welchen Bedingungen anwendbar?*
- Nutzbarkeit : *Welche Vorhersagekraft ist gegeben?*
- Verständlichkeit : *Liegt es in einfachen und übersichtlichen Regeln vor?*
- Neuheit : *War es bisher unbekannt bzw. so nicht erwartet!*

Historisches Beispiel für "Daten" versus "Wissen":

Tycho Brahe (1546-1601)

- dänischer Astronom; bedeutenster Astronom vor Erfindung des Fernrohrs
- 1582 erbaute er eine Sternwarte in der Nähe von Kopenhagen; ab 1599 Hofastronom von Rudolf II in Prag
- bestimmte Positionen der Sonne, des Mondes und der Planeten mit hoher Präzision und zeichnete diese Daten über viele Jahre hinweg auf.

Brahes Problem:

- konnte die gesammelten Daten nicht in einem einheitlichen System zusammenfassen
- sein Modell des (geozentrischen) Planetensystems bewährte sich nicht

Johannes Kepler (1571-1630)

- deutscher Astronom und Mathematiker
- ab 1600 Gehilfe von Tycho Brahe; ab 1601 dessen Nachfolger
- vertrat das Modell des kopernikanischen Planetensystems
- benutzte Brahes Datensammlung $\Longrightarrow$ Mars läuft auf Ellipsenbahn um die Sonne

Die Kepler'schen Gesetze (1609 und 1619)

1. Alle Planeten bewegen sich auf Ellipsen, in deren Brennpunkt die Sonne steht.
2. Eine von der Sonne zum Planeten gezogene Linie,überstreicht in gleichen Zeiten gleiche Flächen.
3. Die Quadrate der Umlaufzeiten zweier Planeten verhalten sich wie die Kuben der großen Ellipsenachsen ihrer Umlaufbahn.

Wie findet man Wissen ?

Es gibt keine universelle Methode um Wissen zu entdecken!

Probleme:

- Riesige Datenmengen in Datenbanken sind heute verfügbar. *Wir ertrinken in einem Meer von Informationen, aber wir hungern nach Wissen.*
- Manuelle Analysen sind kaum mehr durchführbar.
- Einfache Methoden (Diagramme,etc.) stoßen schnell an ihre Grenzen.

Lösungsversuche:

- Interaktive Datenanalyse-Programme
- Knowledge Discovery in Data Bases und Data Mining Methoden

(aus Schwenker 2004)

Was ist "Data Mining"?

Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summerize the data in novel ways that are both understandable and useful to the data owner.
(Hand/Mannila/Smyth)

(aus Klawonn 2004)

Oberbegriff: "Knowledge Discovery in Databases" (KDD)

Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data
(Fayyad, Piatetsky-Shapiro & Smyth, 1996)

Iterativer Prozeß

- Datenauswahl und Stichprobenziehung
- Pre-processing und Datensäuberung
- Transformation und Dimensionsreduktion
- Data mining bzw. Datenanalyse
- Visualisierung und Bewertung

(aus Wilhelm 2001)

Definition KDD

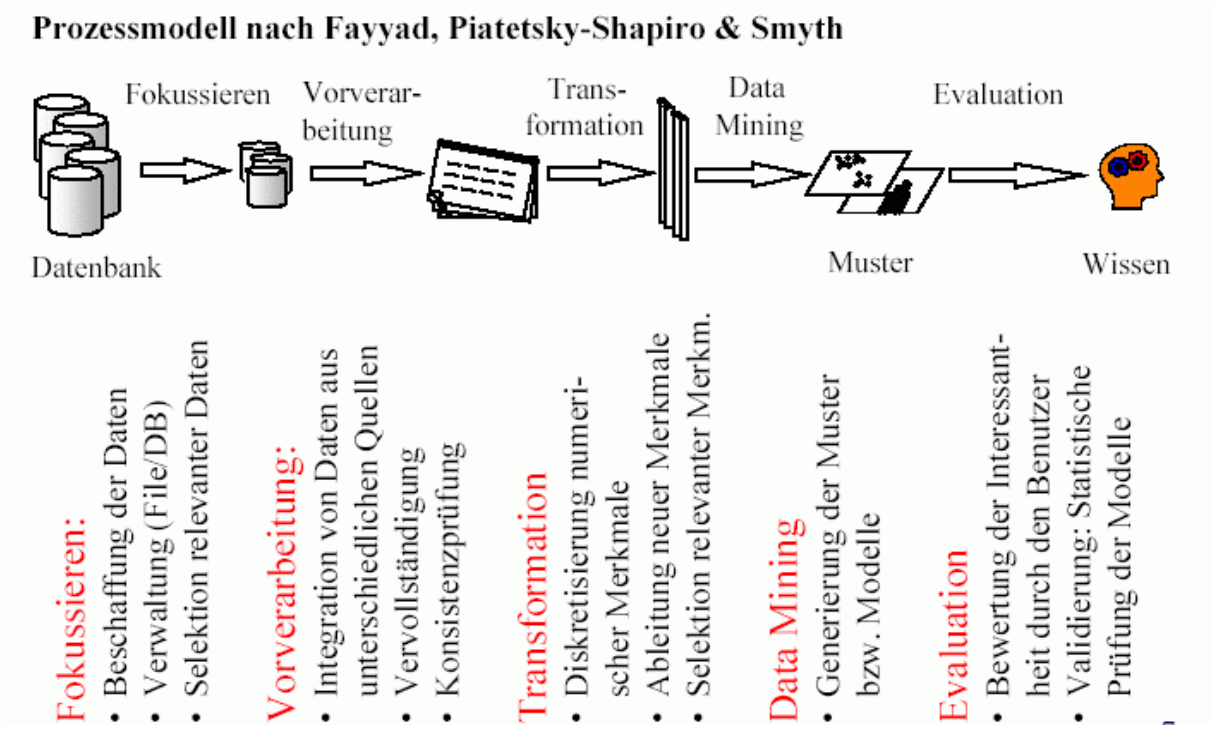
[Fayyad, Piatetsky-Shapiro & Smyth 1996]

Knowledge Discovery in Databases (*KDD*) ist der Prozess der (semi-) automatischen Extraktion von Wissen aus Datenbanken, das

- *gültig*
- *bisher unbekannt*
- und *potentiell nützlich* ist.

Bemerkungen:

- (*semi-*) *automatisch*: im Unterschied zu manueller Analyse. Häufig ist trotzdem Interaktion mit dem Benutzer nötig.
- *gültig*: im statistischen Sinn.
- *bisher unbekannt*: bisher nicht explizit, kein „Allgemeinwissen".
- *potentiell nützlich*: für eine gegebene Anwendung.



(aus Böhm 2004)

KDD-Prozess

Vorstufen

- Bestimmung des Nutzenpotenzials
- Anforderungs-/Durchführbarkeitsanalyse

Hauptstufen

- Sichtung des Datenbestandes.
- Datenvorverarbeitung!
  - Vereinheitlichung und Transformation der Daten in ein einheitliches Datenformat.
  - Datensäuberung: fehlerhafte/unvollständige Eingaben feststellen und ggf. solche Datensätze/Attribute aus dem Datensatz entfernen

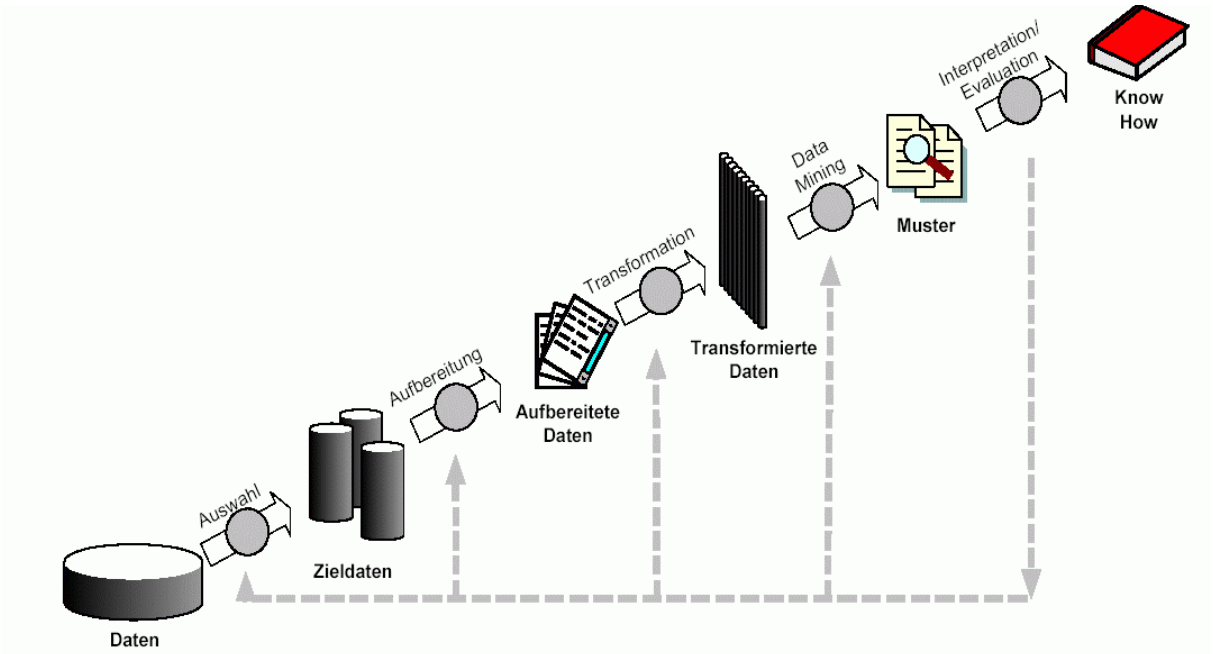
  - Datenreduktion: Stichprobe, Attributauswahl, Beschränkung auf Prototypen

- Data Mining (mit verschiedenen Verfahren)
- Visualisierung der Resultate
- Interpretation, Analyse und Bewertung der erzielten Resultate.
- Anwendung und Dokumentation.

Allgemeines

- Die einzelnen Stufen sind nicht strikt von einander getrennt.
- Der gesamte KDD-Prozess ist in seiner Gesamtheit und seinen Teilaspekten iterativ, d.h. mehrere Durchläufe sind erforderlich.

(aus Schwenker 2004)



Quelle: Fayyad, U.M. et al. (1996), "From Data Mining to Knowledge Discovery: An Overview", in: Fayyad, U.M. et al. (Hrsg.): "Advances in Knowledge Discovery and Data Mining", Menlo Park (CA), S. 1-34.

Data Mining – Definitionen

Data Mining:

a step in the KDD process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, produce a particular enumeration of patterns
(Fayyad, Piatetsky-Shapiro & Smyth, 1996)

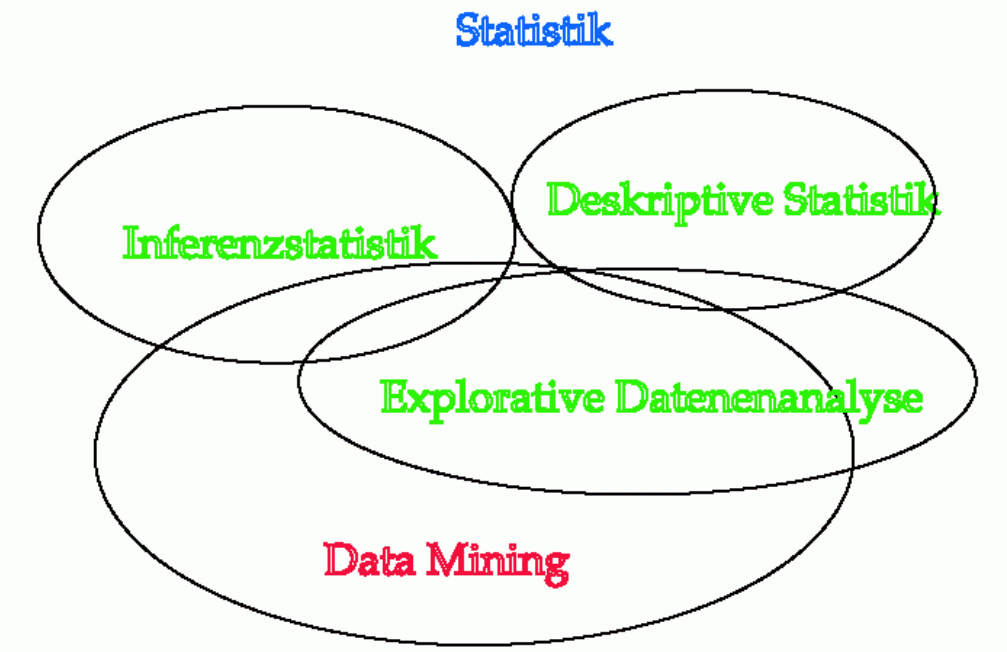
the extraction of previously unknown information from databases that may be large, noisy and have missing data
(Fayyad 1997, Chatfield 1997)

Decision Trees, Neural Networks, Rule Induction, Nearest Neighbors, Genetic Algorithms.
(Meta Group, 1997)

used to discover patterns and relationships in data, with an emphasis on large observational data bases. It sits at the common frontiers of several fields including Data Base Management, Artificial Intelligence, Machine Learning, Pattern Recognition, and Data Visualization.
(J. Friedman, 1997)

the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners
(Hand, 1998)

Statistik vs. Data Mining



(aus Wilhelm 2001)

- Beim Data Mining werden meistens bereits beobachtete/gemessene Daten untersucht, die nicht notwendigerweise speziell für den Datenanalyseprozess gesammelt wurden (im Gegensatz zu experimentellen Daten, bei denen die Datenaufnahme auf den Datenanalyseprozess zugeschnitten wird).

  Beispiel: Geldtransaktionen in einer Bank werden zur Sicherheit aufgezeichnet. Es können aber Anforderungen, Kundenpräferenzen oder -wünsche daraus abgeleitet werden – z.B. wie viel Geld sollte in einem Bankautomaten tagesabhängig zur Verfügung stehen.

(aus Klawonn 2004)

- Beim Data Mining soll das Ergebnis meistens in einer Form dargestellt werden, die verständlich ist für die mit den Daten vertrauten Personen, nicht nur für Statistikexperten.

  Im Idealfall werden die Data Mining Methoden direkt vom Nutzer der Daten angewendet.

Statistik versus Data Mining:

- Datenbanken versus Datenmatrizen
- Beobachtungsstudien versus experimentelle Studien
- sehr große Datenbanken: manche statistische Methoden nicht skalierbar
- heterogene Datenbanken:
  - selection bias (opportunistische Datensammlung)
  - Populationsdrift
  - abhängige Beobachtungen
- geringe Datenqualität, Datenfehler
- Natur neuer Datentypen: verknüpfte Hypertextdaten, graphische Bilder, Audio- und Videodateien
  - spatial-mining
  - text-mining
  - web-mining
- wachsende und sich verändernde Datenbanken
- teilweiseTotalerhebung
- Bedarf an automatisierter Analyse (oft in Echtzeit)

(aus Wilhelm 2001)

Data Mining als eine Renaissance klassischer Verfahren?

- Prozesssicht des Data Mining folgt dem Paradigma der explorativen Datenanalyse
- Selektion und Aufbereitung der Daten erfolgt nach bekannten Mustern aus der Praxis der Datenbanken und der explorativen Datenanalyse
- Teile der Methodik wurden aus der statistischen Methodologie übernommen

- Konzentration auf die Extraktion von Informationen aus Data Warehouse, Data Marts und dem Internet
- Fokussierung auf Algorithmen, mit denen insb. umfangreiche Datensätze verarbeitet werden können
- Zielsetzung, Informationen automatisch zu generieren

(aus Skiera 2003)

Einsatzgebiete des Data Mining:

- Sport
  - IBM Advanced Scout hat die NBA Statistiken analyisiert um Vorteile für die New York Knicks and Miami Heat zu erlangen.
- Astronomie
  - Cal Tech und das Palomar Observatory entdeckten 22 Quasare mittels Data Mining
- Internet Web Surf-Hilfe
  - IBM Surf-Aid analysiert Web Zugriffshistorien um die Präferenzen und das Verhalten von Kunden abzuleiten und damit die Effektivität von Werbung auf Websites und der Organisation von Websites zu verbessern.

(aus Koch 2003)

Anwendungsbedarf nach Industriezweigen:

- Großhandel
- Finanzen
- Telekommunikation
- Verkehr
- Gesundheit

Fälschungssicherheit

- Mobilfunk — *'cloning'* der Gerätekennung
- Kreditkartenmißbrauch — physikalisch/elektronisch
- Rechnermißbrauch — Angriff, Einbruch

Kreditwesen

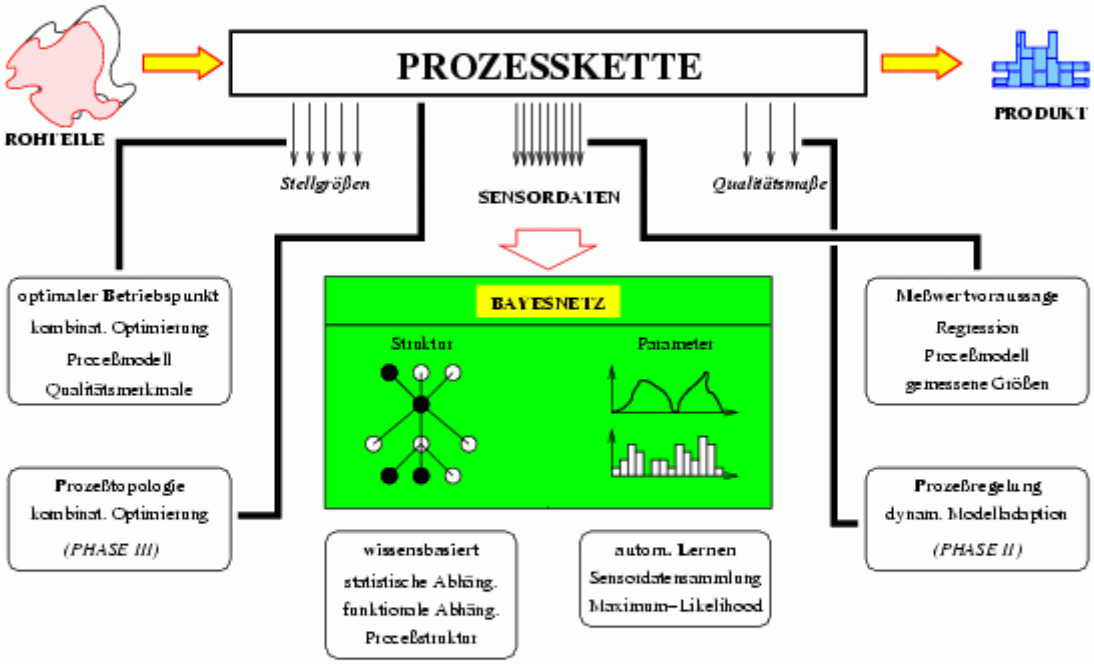
- Kreditwürdigkeit, Zahlungsfähigkeit
- Risikokapital, Unternehmenssolvenz
- Anlageberatung

Kundenbetreuung

- Kundenbindung
  (Beispiel: 5% Reduktion der Fluktuation ⇨ 200% Gewinn)
- Direktmarketing (Handel, Bank, Versicherung)
- Warenkornanalyse im Einzelhandel

Prozeßtechnik

- **Meßwerte** erfassen + auswerten ➪ *Sensoren*
- **Stellgrößen** berechnen + anlegen ➪ *Aktoren*



AUFGABENSTELLUNG: statt Erfahrung und Gefühl ...

- Prozeßvisualisierung
- Entscheidungsunterstützung
- Automatische Regelung

Marketing

Aktive Orientierung an Kundenwünschen ➪ *Wettbewerbsvorteil*

- **Segmentierung** — welche Art Kunden hat die Firma?
- **Klassifikation** — ist Person potentieller Kunde?
- **Konzeptualisierung** — welche Kundeneigenschaften?
- **Prädiktion** — welcher Umsatz im Folgejahr?
- **Deviation** — warum ist Kundenverkalten verändert?
- **Dependenz** — wie beeinflußt Marketing Kundenverhalten?

Relationale Datenbank eines Versandhauses

- *Kundentabelle* — KuNr, PLZ, GJ (Geburtsjahr), ...
- *Umsatztabelle* — BestNr, KuNr, Betrag, ...

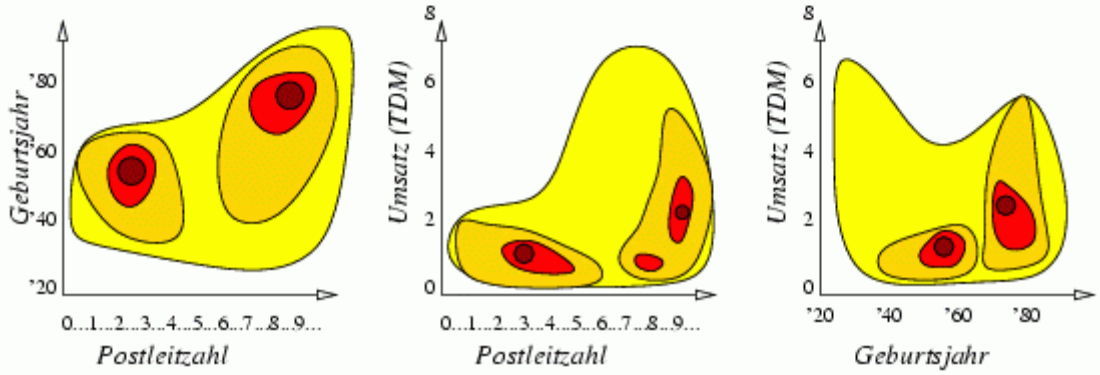
Clusteranalyse der Verbundtabelle

$$(PLZ, GJ, Umsatz) \in \mathbb{R}^3$$

Gewichteter euklidischer Abstand ($\boldsymbol{g} = (10^{-5}, 10^{-2}, 10^{-4})$) ergibt:

$$\boldsymbol{\mu}^{(1)} = \begin{pmatrix} 27\,374 \\ 1954.16 \\ 1122.44 \end{pmatrix}, \quad \boldsymbol{\mu}^{(2)} = \begin{pmatrix} 86\,356 \\ 1969.35 \\ 1618.99 \end{pmatrix}$$

Visualisierung:



(aus Schukat-Talamazzini 2002)

weitere typische Anwendungen:

❑ Palomar Observatory Sky Survey Projekt

- 3 Terabyte Bilddaten
- schätzungsweise 2 Milliarden astronomisch relevante Objekte
- Auswertung: SKICAT (http://www-aig.jpl.nasa.gov/public/mls/skicat/skicat_home.html)
  - Bildsegmentierung und Feature-Belegung (40 Attribute)
  - Klassifikation von Objekten (Sterne bzw. Galaxien)

❑ NASA Earth Observing System (EOS)

- 1,9 TeraByte Datenvolumen pro Tag
  (10 PetaByte Gesamtvolumen)
- Erfassungszeitraum: 15 Jahre
- nur 10% des Datenmaterials wird analysiert
- Echtzeit-Übernahme des Messdatenstroms
  (51 MegaBit/sec bzw. 553 GigaByte pro Tag)
- jährlich ca. 100.000 Benutzer der EOS-Datenbank
  mittlere Objektgröße als Resultat einer Anfrage: 10 MegaByte

Betriebswirtschaftlicher Bereich

❑ Woolworth

- 800 Filialen, 45000 Produkte, 16.000 Angestellte
- 300 Millionen Kundentransaktionen / Jahr
- zwischen 32-750 parallele Benutzer des DWS
- Zentrales DHW integriert 20 operative Systeme
  (Verkaufszahlen, Lagerverwaltung, Planungsinstrumente, ...)

➪ enormer Integrationsaufwand

❑ BMW-Group

- 97.300 Angestellte, 40Mrd. Euro Umsatz
- zweistufige IT-Infrastruktur
- Zentrale IT pflegt seit 20 Jahren UWDM
- Resort-IT-Projekte müssen sich gegenüber dem UWDM abgleichen

➪ minimaler Integrationsaufwand beim Aufbau eines zentralen DWS

Klassisch statistischer Bereich

- **GfK Nürnberg**
  Gesellschaft für Absatz, Markt- und Konsumforschung
  - weltweit tätig, Marktführerschaft in Europa, Kooperationen in USA und Asien
  - TV-Einschaltquotenermittlung, Außenwerbung, Regionalforschung, Ad-Hoc-Marktforschung, Konsumentenverhalten
  - DWH im Bereich Non-food
    - 250.000 betrachtete Artikel, 8000 Geschäfte in Deutschland (15 Kanäle, 50 Regionen)
    - wöchentliche/monatliche Berichtsperiodizität
    - Aufzeichnung aller Abverkäufe von Gebrauchs- und Konsumartikel
    - Datenbestand online: 5 Jahre für Trendanalyse
- **Beispiel IMS Health**
  - erfasst seit 1969 alle in Apotheken eingelösten (Kassen-) Rezepte
    - Identifikation von Arzneimitteln und verschreibenden Arzt
    - Klassifikation nach geographischen Aspekten, Fachrichtungen, Wirkstoffen
    - Kunden: überwiegend Pharmafirmen

(aus Lehner 2003)

Data Mining - Aufgaben:

Explorative Datenanalyse besteht – wie der Name schon sagt – in der Analyse oder Sichtung der Daten zur Gewinnung allgemeiner Erkenntnisse ohne vorgegebenes Ziel wie die Vorhersage eines Attributs aus anderen.

Die bereits vorgestellten Visualisierungstechniken werden häufig im Rahmen der explorativen Datenanalyse angewendet.

Berechnung von Kovarianzen und Korrelationskoeffizienten wird ebenfalls im Rahmen der explorativen Datenanalyse vorgenommen.

Deskriptive Modellierung hat zum Ziel, die Gesamtheit der Daten oder deren Entstehungsprozess zu beschreiben:

- Betrachtung/Schätzung der Wahrscheinlichkeitsverteilungen einzelner Attribute oder gemeinsamer Verteilungen
- Gruppierung der Daten (Clusteranalyse oder Segmentierung)

  Bei der Clusteranalyse versucht man „natürliche“ Gruppen in den Daten zu finden. Bei der Segmentierung wird die Anzahl der Gruppen, in die die Daten eingeteilt werden sollen, (willkürlich) festgelegt.

vorhersagende Modellierung: Klassifikation und Regression: Hierbei geht es darum, den Wert eines Attributs aus anderen vorherzusagen. Vorhersage ist hier allgemein gemeint und bezieht sich nicht notwendigerweise auf einen zeitlichen Verlauf.

- Ist das vorherzusagende Attribut nominal, spricht man von **Klassifikation**.
- Ist das vorherzusagende Attribut eine quantitative Variable, spricht man von **Regression**.

Erkennung von Mustern und Regeln: Durch die Identifikation von typischen und insbesondere atypischen Mustern in den Daten können Besonderheiten oder spezielle Phänomene erkannt werden. Beispielsweise kostet der Betrug bei der Benutzung von Mobiltelefonen die Telekommunikationsindustrie weltweit mehrere 100 Millionen Euro. Üblicherweise weist die betrügerische Benutzung ein stark vom Durchschnitt abweichendes Verhalten auf. (Assoziations-)Regeln der Form *Wenn ein Kunde Farbe und Rauhfasertapete kauft, kauft er meistens auch Pinsel und Spachtel* können unter anderem bei der Warenkorbanalyse gefunden werden und sind interessant für Angebots-, Verkaufs- und Marketingstrategien.

Retrieval by Content: Der Benutzer hat ein Muster, das ihn interessiert und anhand dessen ähnliche oder gleichartige Muster gefunden werden sollen.

Beispiele:

- Dokumentensuche (z.B. im Web) nach Schlüsselwörtern
- Auffinden von Bildern auf der Basis von Bildinhalten
- Identifikation von Musikstücken

(aus Klawonn 2004)

Data–Mining Aufgaben

- Klassifikation : *Wird der Kunde sein Darlehen zurückzahlen?*
- Prognose : *Wie entwickelt sich der Dollar-Kurs?*
- Abhängigkeitsanalysen : *Welche Produkte werden zusammen verkauft?*
- Konzeptbeschreibung : *Welche Lesegewohnheiten haben Leser von Data-Mining Büchern?*
- Segmentierung : *Welche QRS-Dauer haben typischerweise Infarkt-Patienten?*
- Abweichungsanalyse : *Gibt es jahreszeitliche Umsatzschwankungen?*

(aus Schwenker 2004)

Data Mining - Komponenten:

Modell oder Musterstruktur: Hierdurch wird festgelegt, welche Art von Informationen oder Strukturen man in den Daten suchen möchte.

Beispiel: lineare Regressionsfunktion, nichtlineare Regressionsfunktion, Klassifikator, etc.

Ziel-/Bewertungsfunktion: Mit dieser Funktion kann die Qualität des (berechneten) Modells bestimmt werden.

Beispiel: Fehlerfunktion bei der Regression (mittlerer quadratischer Fehler, mittlerer absoluter Fehler, maximaler absoluter Fehler)

Optimierungs-/Suchmethode: Der Algorithmus oder das Verfahren, mit der die Zielfunktion optimiert werden soll.

Beispiele:

- direkte Berechnung der Parameter bei linearer Regression aus dem hergeleiteten Gleichungssystem
- Gradientenverfahren (etwa bei Multilayer-Perceptrons)
- Evolutionsstrategien bei nichtlinearer Regression

Datenmanagementstrategie: Solange alle Daten problemlos im Hauptspeicher des Rechners gehalten werden können, sind keine Datenmanagementstrategien erforderlich.

Bei großen Datenmengen etwa aus großen Datenbanken, die nicht vollständig im Hauptspeicher gehalten werden können, werden effiziente Indizierungsstrategien, geeignete Datenstrukturen und Datenbankanfragen benötigt.

Auch bei der Auswahl der Optimierungs-/ Suchstrategie zur Optimierung der Zielfunktion sollte man auf die Komplexität in Abhängigkeit der Datenzugriffe achten.

(aus Klawonn 2004)

Im Zusammenhang mit Data Mining fällt oft auch der Begriff "Data Warehouse".

Data Warehouse:

- Datenbanksystem, das alle zur Gestaltung der Geschäftsprozesse des Unternehmens und zur Unterstützung sonstiger Managemententscheidungen (MDSS) erforderlichen Daten beinhaltet
- Einheitliche, zentrale Schnittstelle
- Operative Informationssysteme des Unternehmens und unternehmensexterne Quellen werden zusammengeführt

A data warehouse is a subject oriented, integrated, non-volatile, and time-variant collection of data in support of management's decisions. (W.H. Inmon)

Data Warehousing beinhaltet

- die Verwaltung großer, sich mit der Zeit ändernder Datenmengen,
- die aus verschiedenen Quellen/Datenbanken stammen können,
- sowie Techniken zur Aufbereitung und zum komfortablen Umgang eines Benutzers mit den Daten.

(Klawonn 2004)

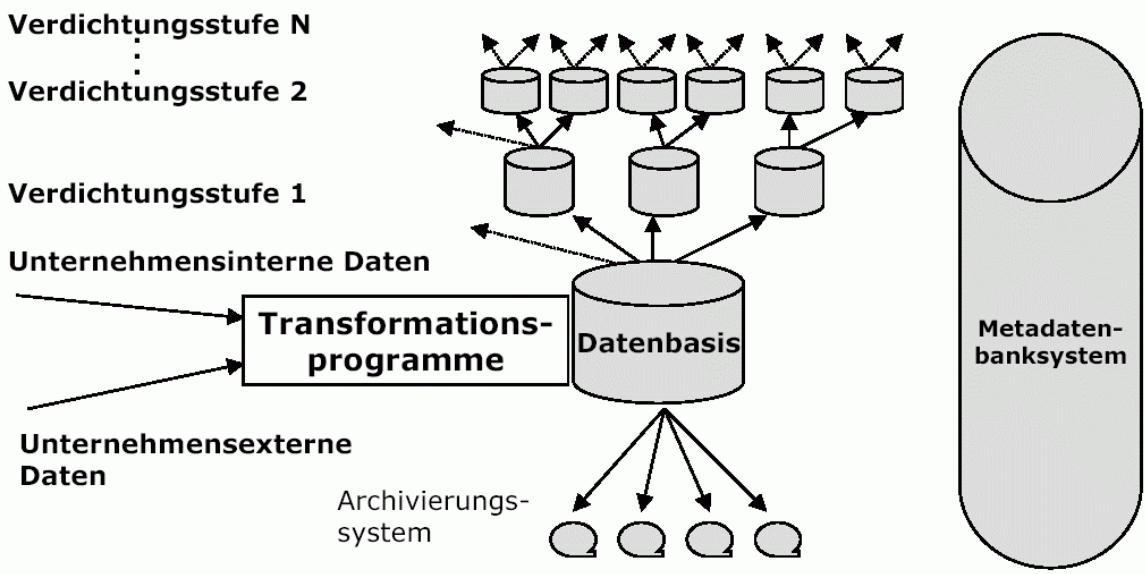
Eigenschaften eines Data Warehouse (nach Skiera 2003):

- Orientierung an betriebswirtschaftlichen Bezugsgrößen
  - Operative Informationssysteme beinhalten Daten bzgl. der operativen Abwicklung der Geschäftsprozesse
  - MDSS beinhalten mehrdimensionale Gruppierungen der Daten nach betriebswirtschaftlichen Kategorien
- Zeitraumbezug
  - Operative Informationssysteme betrachten meist zeitpunktbezogene Daten
  - MDSS betrachten meist zeitraumbezogene Daten
- Struktur- und Formatvereinheitlichung
  - Daten entstammen unterschiedlichen Quellen
  - Syntaktische Datenintegration
  - Vermeidung semantischer Inkonsistenzen

- Nichtvolatilität
  - Daten operativer Informationssysteme ändern sich im Zeitablauf
  - Daten des Data Warehouse werden nicht verändert (Reproduzierbarkeit der Datenanalysen)
- Verzicht auf Echtzeitdaten
  - Operative Informationssysteme benötigen Echtzeitdaten
  - Daten des Data Warehouse beziehen sich auf einen Zeitraum und werden daher nicht kontinuierlich aktualisiert

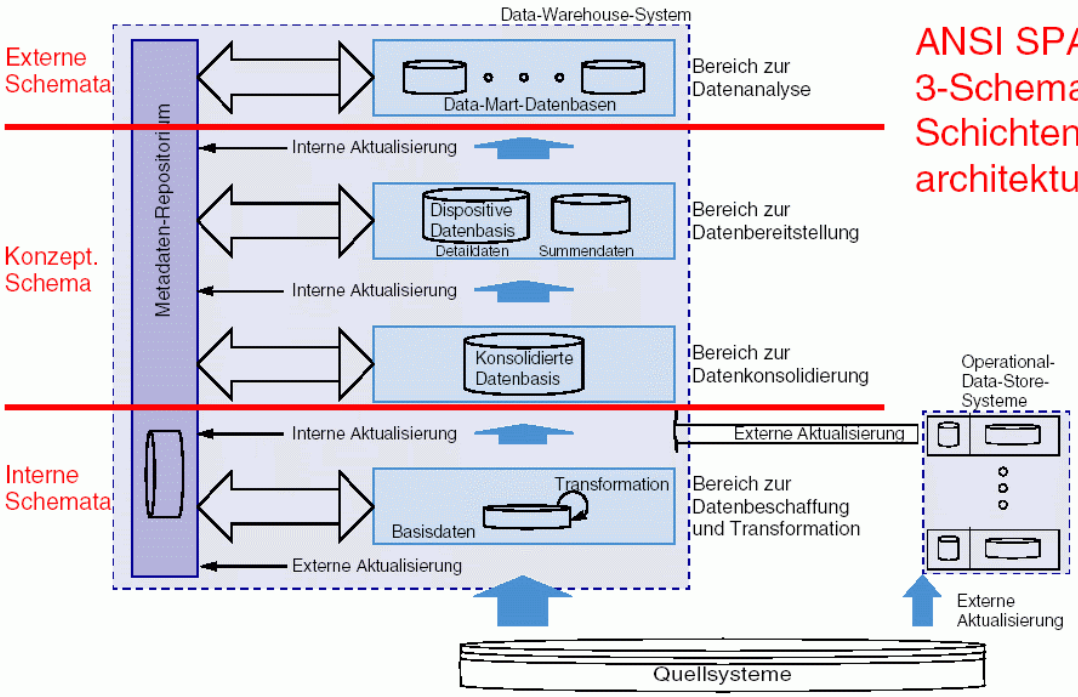
Komponenten eines Data Warehouse:

- Datenbasis
- Transformationsprogramme
  - Extraktion der Daten aus dem jeweiligen Quellsystem und ihre semantische und syntaktische Integration
- Metadatenbank
  - Dokumentation aller für die Datenverwaltung und Datenanalyse erforderlichen syntaktischen und semantischen Datenbeschreibungen (bspw. Datenquelle)
- Archivierungssystem
- Front-End Schnittstelle
  - Schnittstelle für Informationsabfragen



Quelle: Wilde, K. D. (2001): "Data Warehouse, OLAP und Data Mining im Marketing", in: Hippner, H. et al. (Hrsg.): "Handbuch Data Mining im Marketing", Wiesbaden, S. 1-19.

Komponenten eines Data-Warehouse-Systems



Eigenschaften eines Data-Warehouse-Systems

- **Auswertungsorientierte Organisation der Daten**
  - *Fachorientierung* (engl. subject orientation)
  - Modellierung eines spezifischen Anwendungsziels
- **Integration von Daten aus unterschiedlichen Quellsystem**
  - *Integrierte Datenbasis* (engl. integration)
  - Integration auf struktureller Ebene und Datenebene mehrerer Datenbanken
- **Keine Aktualisierung durch den Benutzer**
  - *Nicht flüchtige Datenbasis* (engl. non-volatile)
  - Stabile Datenbasis; einmal eingebrachte Daten werden nicht mehr entfernt oder geändert, nur lesender Zugriff
- **(Optionale Historisierung mit expliziter temporaler Modellunterstützung)**
  - *Historische Daten* (engl. time variance)
  - Daten werden über einen längeren Zeitraum gehalten

Data Warehouses sind die Basis für "multidimensionale Datenanalyse" (Analyse unter Berücksichtigung zahlreicher Attribute) und für viele Data Mining - Anwendungen.

MULTIDIMENSIONALE ANALYSE

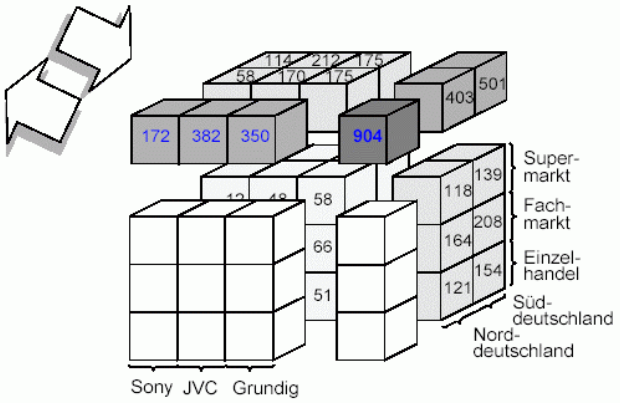
❑ Ausgangspunkt für OLAP (OnLine Analytical Processing)

- Komplex Strukturierte Statistische Tabelle
- Direkte Abbildung auf multidimensionale Datenwürfel

Verkäufe		Sony	JVC	Grundig	Σ
Nord-deutschland	Super-markt	12	48	58	118
	Fach-markt	31	67	66	164
	Einzel-handel	15	55	51	121
	Σ	58	170	175	403
Süd-deutschland	Super-markt	22	50	67	139
	Fach-markt	51	100	57	208
	Einzel-handel	41	62	51	154
	Σ	114	212	175	501
Σ		172	382	350	904

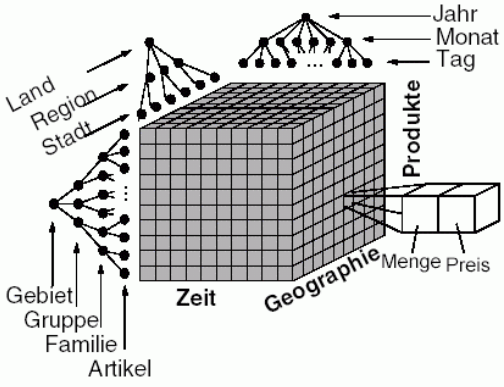
Statistische Tabelle

Multidimensionaler Datenwürfel



❑ Eigenschaften

- "Verallgemeinerung" der flachen Tabelle eines relationalen Ansatzes
- Inhärente Unterscheidung quantifizierender und qualifizierender Attribute
- Abbildung komplex strukturierter Begriffswelten in Form hierarchischer Dimensionsstrukturen
- Spezifische Operatoren zur Unterstützung des explorativen Charakters



❑ Instanz eines Datenwürfels

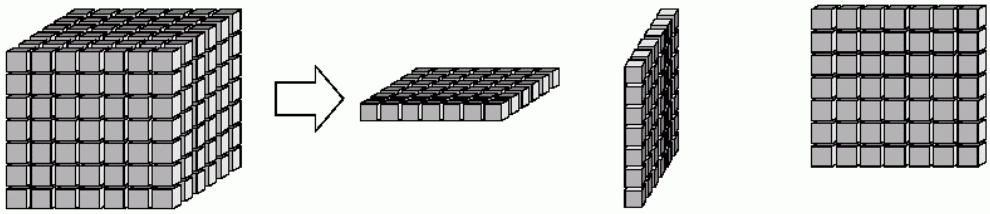
- *alle* Würfelzellen aus dem Definitionsbereich des Datenwürfels
- nicht Teilmenge wie im relationalen Modell!

❑ Achtung: Würfel ist nur eine Metapher!

Operatoren im Multidimensionalen Modell

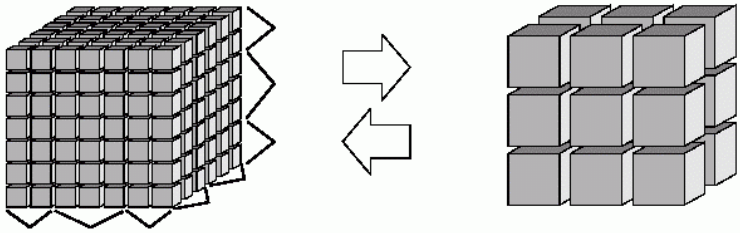
❑ **"Slice and Dice"**

- Selektion eines Teilwürfels



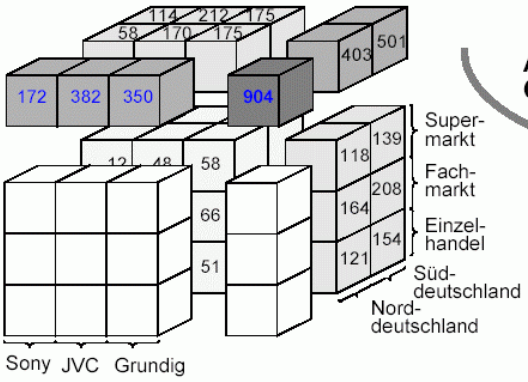
❑ **"Roll Up" / "Drill-Down"**

- Aggregation / De-Aggregation



Unterstützung vom Datenbanksystem: CUBE-Operator

Verkäufe(	Region	Geschäftstyp	Marke	Verkäufe)
	Norddeutschland	Supermarkt	Sony	12
	Norddeutschland	Supermarkt	JVC	48
	Norddeutschland	Supermarkt	Grundig	58
	Norddeutschland	Fachmarkt	Sony	31
	Norddeutschland	Fachmarkt	JVC	67
	Norddeutschland	Fachmarkt	Grundig	66
	Norddeutschland	Einzelhandel	Sony	15
	Norddeutschland	Einzelhandel	JVC	55
	Norddeutschland	Einzelhandel	Grundig	51
	Süddeutschland	Supermarkt	Sony	22
	Süddeutschland	Supermarkt	JVC	50
	Süddeutschland	Supermarkt	Grundig	67
	Süddeutschland	Fachmarkt	Sony	51
	...			



Anwendung des CUBE-Operators

Verkäufe(	Region	Geschäftstyp	Marke	Verkäufe)
	Norddeutschland	Supermarkt	Sony	12
	Norddeutschland	Supermarkt	JVC	48
	Norddeutschland	Supermarkt	Grundig	58
	Norddeutschland	Supermarkt	ALL	118
	Norddeutschland	Fachmarkt	Sony	31
	Norddeutschland	Fachmarkt	JVC	67
	Norddeutschland	Fachmarkt	Grundig	66
	Norddeutschland	Fachmarkt	ALL	164
	Norddeutschland	Einzelhandel	Sony	15
	Norddeutschland	Einzelhandel	JVC	55
	Norddeutschland	Einzelhandel	Grundig	51
	Norddeutschland	Einzelhandel	ALL	121
	Norddeutschland	ALL	ALL	403
	Süddeutschland	Supermarkt	Sony	22
	...			
	Süddeutschland	ALL	ALL	501
	ALL	Supermarkt	Sony	34
	ALL	Supermarkt	JVC	98
	ALL	Supermarkt	Grundig	155
	ALL	Supermarkt	ALL	257
	ALL	Fachmarkt	Sony	82
	...			
	ALL	ALL	Sony	172
	ALL	ALL	JVC	382
	ALL	ALL	Grundig	350
	ALL	ALL	ALL	904

Im weiteren Verlauf dieser Vorlesung werden Datenbank-architektur und Datenbankentwurf allerdings nicht mehr berücksichtigt werden.

Beispiel für Daten, wie sie im Data Mining untersucht werden:

ID	Age	Sex	Married	Education	Income
248	54	Male	Yes	High school	10000
249	?	Female	Yes	High school	12000
250	29	Male	Yes	College	23000
251	9	Male	No	Child	0
252	85	Female	No	High school	19798
253	40	Male	Yes	High school	40100
254	38	Female	No	Ph.D.	2691
255	7	Male	?	Child	0
256	49	Male	Yes	College	30000
257	76	Male	Yes	Ph.D.	30686

Begriffe:

- Eine derartige Datenmatrix wie im vorhergehenden Beispiel wird **Datensatz** genannt.
- Die einzelnen Spalten/Komponenten des Datensatzes werden **Attribute, Variablen, Felder** oder **Features** genannt.
- Eine einzelne Zeile bezeichnet man als **Individuum, Instanz, Fall, Objekt, Datum, Record** oder **Entity**.

Man unterscheidet je nach Wertebereich verschiedene Typen von Attributen (Attributarten; in der Statistik auch *Skalenniveaus* genannt):

Nominale Attribute haben einen diskreten endlichen Wertebereich.

Beispiele:

- Geschlecht (männlich/weiblich)
  (**binäres** Attribut)
- Studienrichtung
  (Medien-/Praktische/Technische Informatik)
- Nationalität (Deutsch/Englisch/Französisch/...)

Ordinale Attribute haben einen endlichen Wertebereich mit einer Ordnungs-/Präferenzstruktur.

Beispiele:

- Schulabschluss
  (Hauptschule/Realschule/Fachgymnasium/
  Gymnasium)
- Angestelltenstatus
  (Sachbearbeiter/Unterabteilungsleiter/
  Abteilungsleiter/Hauptabteilungsleiter/
  Bereichsleiter)

Intervallgrößen haben nicht nur eine feste Ordnung, sondern werden auch in gleichen Einheiten gemessen. Es gibt aber im Allgemeinen keinen ausgezeichneten Nullpunkt. Differenzen zwischen Werten ergeben Sinn, Summen oder Vielfache von Werten keinen.

Beispiele:

- Datum (Jahreszahl)
  (willkürliche) Festsetzung des Jahres 0
- Temperatur
  (in Grad Celsius oder Fahrenheit)

Ratiogrößen besitzen im Gegensatz zu Intervallgrößen außerdem einen ausgezeichneten Nullpunkt. Das Berechnen von Verhältnissen von Werten ist sinnvoll.

Beispiele:

- Abstand
- Anzahl der Kinder

Ganzzahlige Attribute können nur ganzzahlige (Integer-)Werte annehmen. Ganzzahlige Attribute können Intervall- oder Ratiogrößen sein.

Beispiele:

- Datum (Jahreszahl)
  (Intervallgröße)
- Anzahl der Kinder einer Familie
  (Ratiogröße)

Kontinuierliche Attribute können (beliebige) reelle Werte annehmen. Kontinuierliche Attribute können Intervall- oder Ratiogrößen sein.

Beispiele:

- Temperatur (Intervallgröße)
- Abstand (Ratiogröße)
- Vorsicht bei speziellen Größen wie Winkel, die sich zum Teil wie Ratiogrößen verhalten, aber nur lokal ordinal sind.

Missing Values:

Bei einigen Objekten können die Werte einzelner Attribute fehlen. Man spricht dann von **Missing Values** oder **fehlenden Werten**.

Beispielursachen für fehlende Werte:

- Ausfall eines Sensors
- Verweigerung einer Auskunft
- irrelevantes Attribut für das betreffende Objekt (schwanger (ja/nein) bei Männern)

Behandlung von Missing Values

- Bei wenigen Missing Values: Objekte mit Missing Values werden beim Data Mining Prozess nicht berücksichtigt.
- Schätzen (imputation) der Missing Values (Einsetzen des Mittelwertes, des häufigsten Wertes oder eines aus der Kenntnis der anderen Attribute geschätzten Wertes)
- Berücksichtigung/Verarbeitung/Schätzung der Missing Values während der Anwendung eines Data Mining Verfahrens (problem-/verfahrensabhängig)

Arten von Missing Values

Wir betrachten das Attribut X_{obs}. Ein Missing Value wird mit ? gekennzeichnet.

X bezeichne den tatsächlichen Wert des betrachteten Attributs, d.h., es gilt

$$X_{\text{obs}} = X, \quad \text{falls } X_{\text{obs}} \neq ?$$

Y sei die (multivariate) Zufallsvariable, die die Werte aller anderen Attribute (außer X) wiedergibt.

Missing completely at random (MCAR): Die Wahrscheinlichkeit, dass der Wert von X fehlt, hängt weder vom wahren Wert von X noch von den Werten der anderen Variablen ab:

$$P(X_{\text{obs}} = ?) \;=\; P(X_{\text{obs}} = ? \mid X, Y)$$

Beispiel: An einem Sensor wurde ab und zu vergessen, die Batterie auszutauschen. Wenn die Batterie leer war, hat der Sensor keine Daten/Werte geliefert.

MCAR wird auch als **Observed At Random (OAR)** bezeichnet.

Missing at random (MAR): Die Wahrscheinlichkeit, dass der Wert von X fehlt, hängt nicht direkt vom wahren Wert von X ab:

$$P(X_{\text{obs}} = ? \mid Y) \;=\; P(X_{\text{obs}} = ? \mid X, Y)$$

Beispiel: Bei regelmäßigen medizinischen Untersuchungen werden bestimmte Werte von schwer erkrankten Patienten nur in größeren Zeitabständen gemessen, um die Patienten (z.B. durch Blutabnahme) weniger stark zu belasten.

Nonignorable: Die Wahrscheinlichkeit, dass der Wert von X fehlt, hängt vom wahren Wert von X ab.

Beispiel: Ein Temperatursensor fällt bei Frost immer aus.

Bei MCAR und MAR können die fehlenden Werte prinzipiell (bei genügend großem Datensatz) aus den anderen Werten geschätzt werden. (Die Ursache für die Missing Values ist *ignorable*.) Im Beispiel des ausfallenden Temperatursensors können keine Aussagen über die Temperaturen unterhalb des Gefrierpunktes getroffen werden.

- Im Fall MCAR kann man davon ausgehen, dass die Missing Values der gleichen Verteilung folgen, wie die beobachteten Werte.
- Im Fall MAR folgen die Missing Values zwar nicht der gleichen Verteilung wie die beobachteten Werte. Unter Verwendung der anderen Attribute kann man aber trotzdem sinnvolle Schätzungen für die Missing Values vornehmen, sofern genügend Daten vorhanden sind.
- Bei nonignorable Missing Values ist eine realistische Schätzung für die Missing Values kaum möglich.

(aus Klawonn 2004)

Aufgabe beim Data Mining ist u.a. *Lernen* aus den Daten.

Zum Begriff "Lernen":

Lernen ist...

... jeder Vorgang, der ein System in die Lage versetzt, bei der zukünftigen Bearbeitung derselben oder einer ähnlichen Aufgabe diese besser zu erledigen. (Simon 1983)

Was heißt „besser"?
Was für den einen besser ist, ist für den anderen schlechter!
Lernen ohne Ziel!

... das Konstruieren oder Verändern von Repräsentationen von Erfahrungen. (Michalski 1986)

Lernen ist...

Wissenserwerb (Begriffe, Theorie, Sprache)

Definition eines Begriffs aus seinen Beispielen,
zusammenhängende Definitionen ergeben eine Theorie
Grammatik aus wohlgeformten Sätzen

Funktionslernen (Klassifikation, Regression)

$f(\vec{x}) = y, \quad y \in R \vee y \in [0,1]$

Suche im Hypothesenraum

Mögliche Lösungen werden geordnet aufgezählt,
bei der richtigen wird angehalten

der induktive Schluß

Uta ist ein Mensch, Uta ist sterblich, so auch Uli, Vroni, Sokrates...
also
alle Menschen sind sterblich

Wissensentdeckung ist...

... der nichttriviale Prozess der Identifikation gültiger, neuer, potenziell nützlicher und schlussendlich verständlicher Muster in (sehr großen) Datenbeständen.

Maschinelles Lernen wird in der Wissensentdeckung als data mining step verwendet.

Sogar in der Datenvorverarbeitung werden neuerdings maschinelle Lernverfahren eingesetzt.

Einige Verfahren des maschinellen Lernens (Adaptivität und Optimierung) werden nicht in der Wissensentdeckung verwendet.

Wissensentdeckung benötigt Verfahren, die sehr große Datenmengen verarbeiten können.

Wissenschaftliche Fragen

Wieviele Beispiele muß ich mindestens haben, bis ich ein ausreichend korrektes und vollständiges Lernergebnis erzielen kann? Wie sicher bin ich bei meiner Beispielmenge?

Wie mächtig muß mein Repräsentationsformalismus sein, damit ich ein annähernd korrektes und vollständiges Lernergebnis ausdrücken kann? Wie schwach darf er sein?

Unter welchen Umständen wird der Lernalgorithmus zu einem Ergebnis kommen und anhalten? Wie schnell ist er?

Welche Zusicherungen kann der Algorithmus über sein Ergebnis garantieren? z.B.: dies ist die speziellste Generalisierung über allen Daten -- wenn sie falsch ist, sind auch die Daten falsch!
z.B.: Dies sind alle Regeln, die in den Daten verborgen sind -- wenn eine fehlt, fehlen auch die entsprechenden Daten!

(aus Morik 2003)

– Fragen, die in der algorithmischen Lerntheorie behandelt werden (gegen Ende der Vorlesung)

weitere Aufgabe des Data Mining: ***Vorhersage von Werten***

klassische statistische Methode: Regressionsrechnung
(z.B. eine Gerade durch eine Punktwolke legen)

Wenn der funktionale Zusammenhang zwischen der vorherzusagenden Größe y und den **Prädiktorvariablen** $x_1, \ldots, x_p$ bis auf die zu schätzenden Parameter bekannt ist, kann die (eventuell nichtlineare) Regressionsfunktion explizit angegeben werden.

Ist der Zusammenhang (**das Modell**) nicht bekannt, so kann man versuchen, durch eine genügend flexible Regressionsfunktion die Daten bzw. den Zusammenhang zwischen $x_1, \ldots, x_p$ und y trotzdem zu erlernen.

Beispiele:

- Vorhersage des Tageskurses des DAX (y) aus den Prädiktorvariablen Vortageskurs von Nasdaq, Dow Jones, Nemax, Dax,...
- Vorhersage der Niederschlagsmenge (y) anhand der Prädiktorvariablen Vortageswerte der Temperatur, des Niederschlags, des Luftdrucks,...
- Vorhersage des Energiebedarfs (einer Stadt) für eine bestimmte Tageszeit

Ist die Form des funktionalen Zusammenhangs nicht bekannt, so kann man bei den Prädiktorvariablen $x_1, \ldots, x_p$ z.B. einen

- linearen $y = a_0 + a_1 x_1 + \ldots + a_p x_p$
- quadratischen

$$\begin{aligned} y \;=\; & a_0 + a_1 \cdot x_1 + \ldots + a_p \cdot x_p + \\ & a_{p+1} \cdot x_1^2 + \ldots + a_{2p} \cdot x_p^2 + \\ & a_{2p+1} \cdot x_1 x_2 + \ldots + a_{2p+p(p-1)/2} \cdot x_{p-1} x_p \end{aligned}$$

- oder kubischen Ansatz verwenden.

Bei einem Ansatz mit ausschließlich linearen Termen können die Koeffizienten a_i eventuell noch als Gewichtungsfaktoren interpretiert werden. Dazu sollten die Attribute, die als Prädiktorvariablen verwendet werden, aber vorher normalisiert oder standardisiert werden.

Im Allgemeinen erhält man jedoch ein **Black Box Modell**, das zwar eventuell eine gute Approximation der Daten erreicht, aber keine Interpretation (der Koeffizienten) zulässt.

Generalisierung

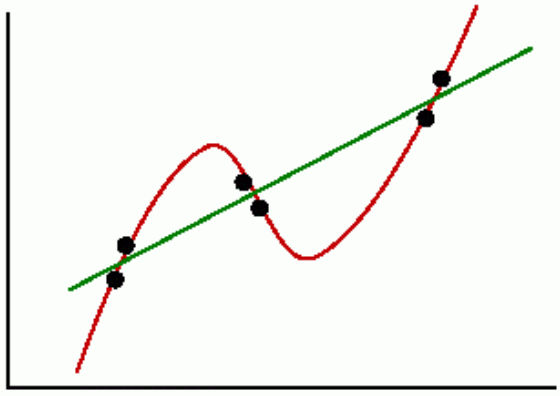
Fasst man den gegebenen Datensatz als Sammlung von Beispielen auf, anhand dessen der Zusammenhang zwischen Prädiktorvariablen und dem vorherzusagenden Attribut erlernt werden soll, muss die Regressionsfunktion „lernen“, aus den Daten zu **generalisieren**, um bei neuen Daten eine möglichst korrekte Voraussage zu treffen.

Dazu muss die Regressionsfunktion, d.h. der zur Verfügung stehende Parametersatz, genügend flexibel sein.

Das bedeutet jedoch nicht, dass eine komplexere Regressionsfunktion mit mehr Parametern zu besseren Ergebnissen führt als eine einfache.

Der Begriff "Overfitting"

Es kann zum **Overfitting** kommen:

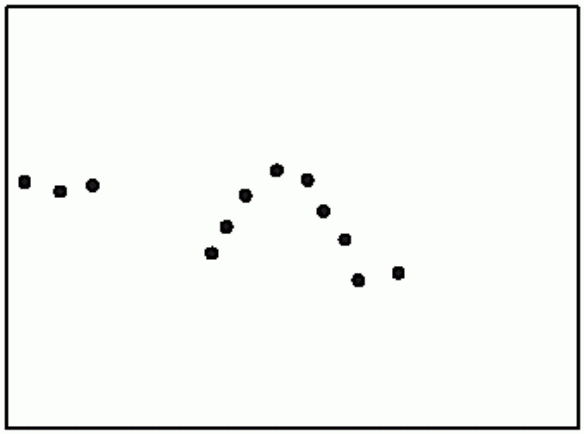


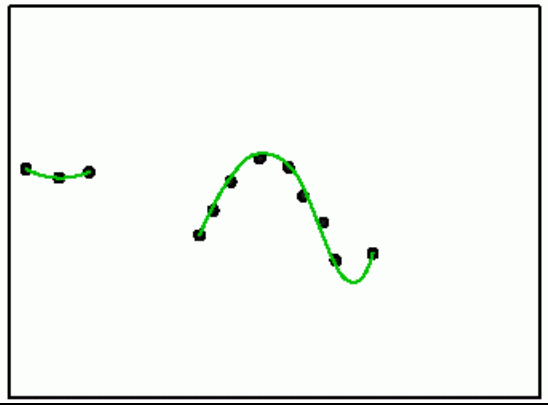
Die Funktion „lernt“ die Daten auswendig ohne zu generalisieren. Die Vorhersage ist unter Umständen schlechter als bei einer einfachen Regressionsfunktion.

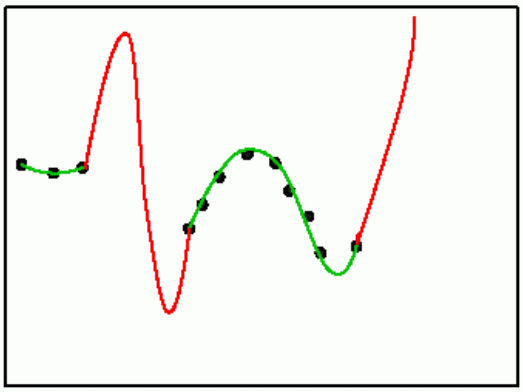
Interpolation und Extrapolation

Man unterscheidet zwischen

- **Interpolation**: Nachbildung oder Annäherung der Daten und
- **Extrapolation**: Vorhersage in Bereichen, in denen Daten fehlen.







Interpolation	Extrapolation

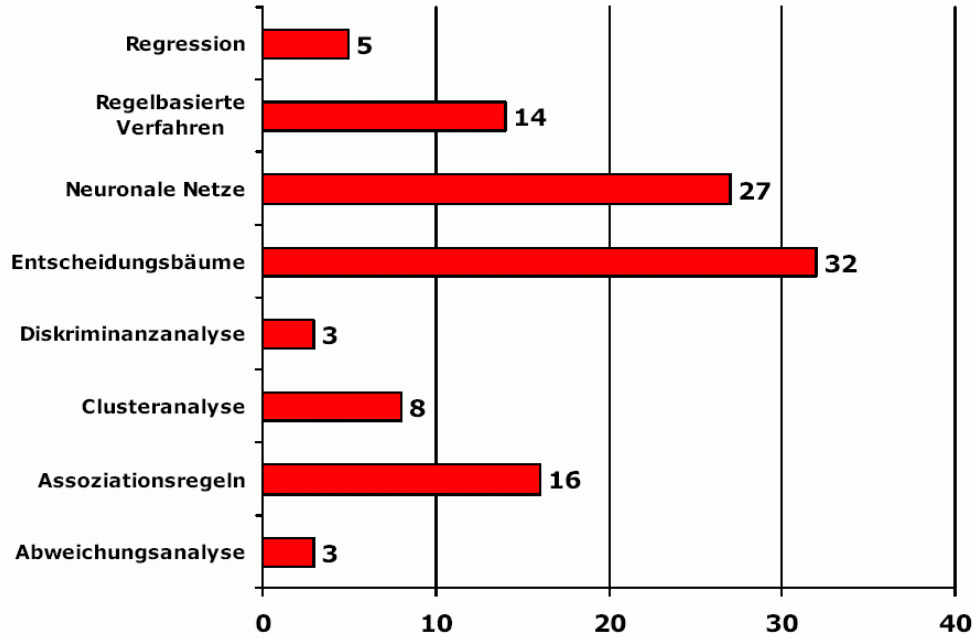
(nach Klawonn 2004)

Methodenübersicht

- Theoriegeleitete Methoden
  - Korrelationen
  - T-Test
  - ANOVA
  - Lineare Regression
  - Logistische Regression/Diskriminanzanalyse
  - Prognosemethoden

- Datengeleitete Methoden
  - Clusteranalyse
  - Faktorenanalyse
  - Entscheidungsbaumverfahren
  - Neuronale Netze
  - Assoziationsregeln

Praktische Anwendung der Methoden:



Quelle: Gaul, W., Säuberlich, F. (1999): "Classification and Positioning of Data Mining Tools", in: Gaul, W. / Locarek-Junge, H. (Hrsg.): "Classification in the Information Age", Springer, Berlin, Heidelberg, 143-152.

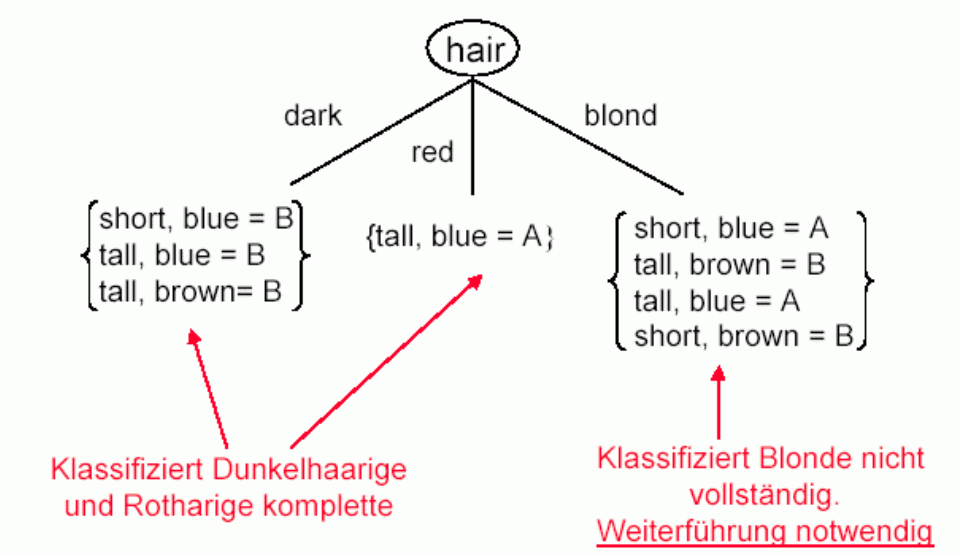
(aus Skiera 2003)

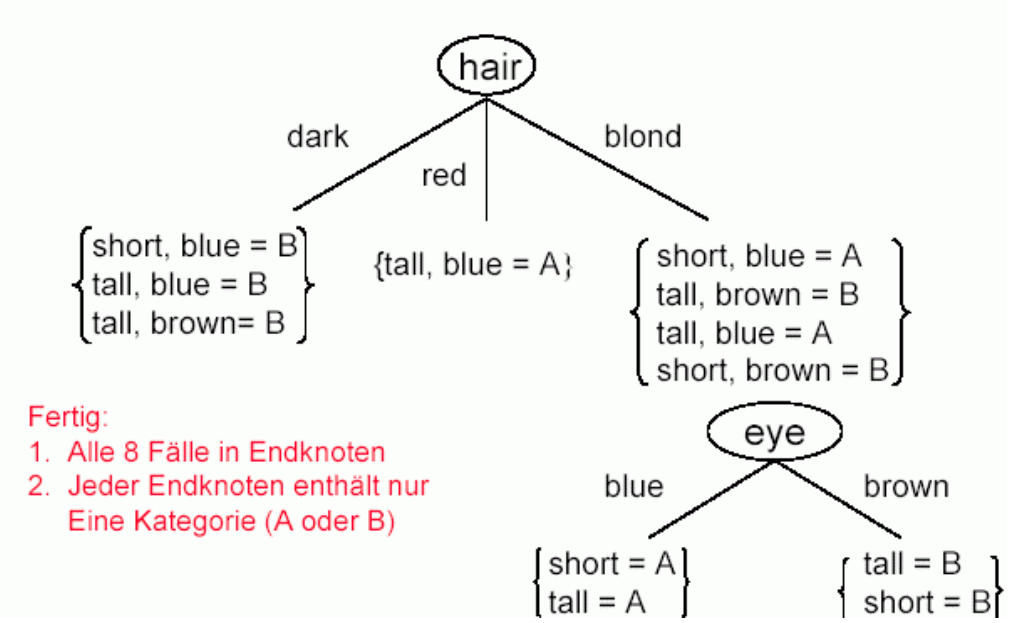
Regelbestimmung/Entscheidungsbäume - Beispiel

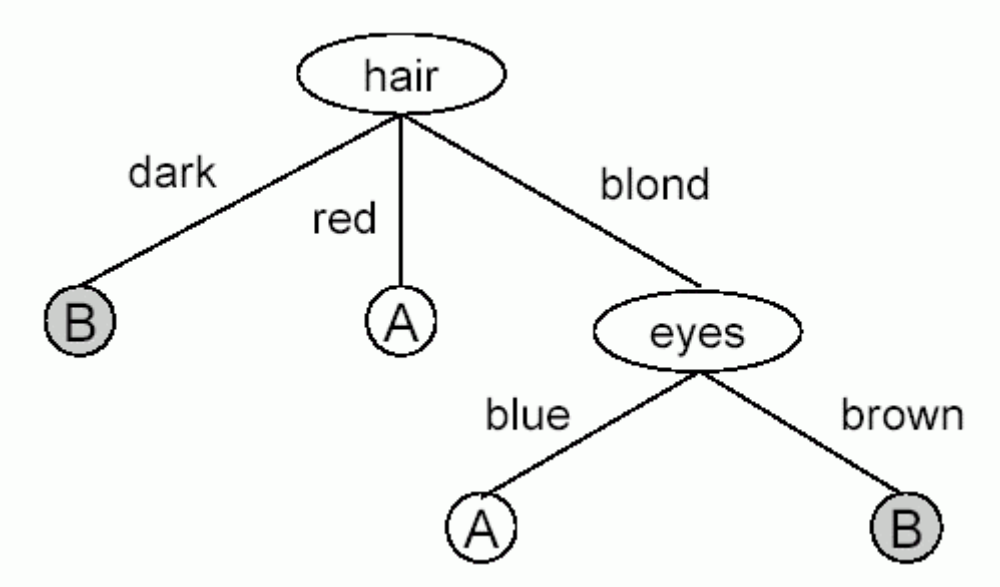
Daten:

height	hair	eyes	class
short	blond	blue	A
tall	blond	brown	B
tall	red	blue	A
short	dark	blue	B
tall	dark	blue	B
tall	blond	blue	A
tall	dark	brown	B
short	blond	brown	B

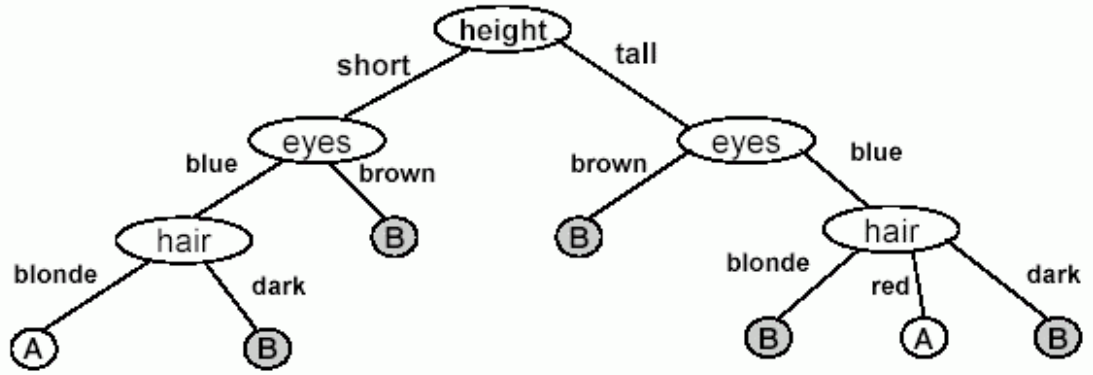
Bestimme eine Regel um eine neue Person in A oder B zu klassifizieren







- Gut: Ein Entscheidungsbaum kann immer aufgestellt werden
- Schlecht: Im schlechtesten Fall für jeden Datenpunkt ein Blattknoten



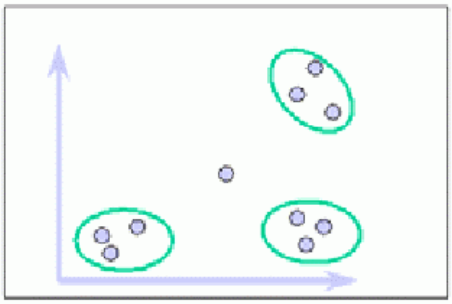
8 Fälle, 7 Knoten. Keine effiziente Zusammenfassung.

Clusterung (Datenbank-Segmentierung)

- "The art of finding groups in data" Kaufman & Rousseeuw
- Ziel: Zusammenfassung von Items aus der Datenbank in Mengen anhand (unbekannter) gemeinsamer Charakteristika
- Schwieriger als Klassifizierung, da die Klassen nicht in Voraus bekannt sind (kein Training) Examples:
- Technik: „**unsupervised learning**"

Clusterung - Beispiel

- Gibt es natürliche Cluster in den Daten (36,10), (12,8), (38,42), (13,6), (36,38), (16,9), (40,36), (35,19), (37,7), (39,8)?



Regelgenerierung

- Finden von Regeln der Form

  IF <left-hand-side> THEN <right-hand-side>
- (Gegenteil eines Regelbasierten Agenten, dem die Regeln vorgegeben werden und der darauf reagiert. Hier werden die Ergebnisse/Aktivitäten gegeben und die Regeln sollen bestimmt werden.)
- Prevalence = Wahrscheinlichkeit, dass LHS und RHS gemeinsam auftreten (“support factor,” “leverage” oder “lift”)
- Predictability = Wahrscheinlichkeit der RHS unter der Bedingung, dass LHS zutrifft (“confidence” oder “strength”)

Regelgenerierung aus einer Einkaufskorbanalyse

- <Dairy-Milk-Refrigerated> → <Soft Drinks Carbonated>
  - prevalence = 4.99%, predictability = 22.89%
- <Dry Dinners - Pasta> → <Soup-Canned>
  - prevalence = 0.94%, predictability = 28.14%
- <Paper Towels - Jumbo> → <Toilet Tissue>
  - prevalence = 2.11%, predictability = 38.22%
- <Dry Dinners - Pasta> → <Cereal - Ready to Eat>
  - prevalence = 1.36%, predictability = 41.02%
- <American Cheese Slices > → <Cereal - Ready to Eat>
  - prevalence = 1.16%, predictability = 38.01%

Nutzung von generierten Zusammenhängen

- Gutscheine, Preisnachlässe
  - Keine Preisnachlässe für zwei Produkte, die häufig zusammen gekauft werden. Preisnachlässe auf ein Produkt um das andere zu “ziehen”
- Produktplatzierung
  - Biete dem Kunden korrelierte Produkte zur selben Zeit/am selben Ort an.

Weitere Data Mining - Ansätze:

statistische Verfahren

- Regressions- und Korrelationsrechnung
- Hauptkomponentenanalyse
- Varianzanalyse
- Zeitreihenanalyse

maschinelles Lernen (KI-Verfahren)

- Begriffslernen
- instanzenbasiertes Lernen
- induktive logische Programmierung
- Bayes-Klassifikatoren
- support vector machines
- künstliche neuronale Netze
- genetische Algorithmen