Organisatorisches

Vorlesung

- Beginn: 13. April 2005
- Mittwoch, 14.15 – 15.45 Uhr in Raum 0443

Übungen

- Dienstag, 12.30 h - 14.00 h, in Raum 0443
- Beginn: **26. April 2004**
- wird als Präsenzübung abgehalten (s. nächste Folie)
- praktische Übungen mit Weka (http://www.cs.waikato.ac.nz/~ml/weka/)
  (Bonus in der Klausur oder mündlichen Prüfung bei erfolgreicher Teilnahme)

2. Grundlagen

Inhalt dieses Kapitels

2.1 Datenbanksysteme [Kemper & Eickler 1999]

Grundbegriffe, relationale Datenbanksysteme, Anfragesprache SQL, Methode der Anfragebearbeitung, physische Speicherung der Daten, Indexstrukturen zur effizienten Anfragebearbeitung

2.2 Statistik [Fahrmeier, Künstler, Pigeot & Tutz 1999]

univariate und multivariate Deskription, Wahrscheinlichkeitsrechnung, diskrete und stetige Zufallsvariablen, Approximation von Verteilungen, Parameterschätzung, Testen von Hypothesen

2.3 OLAP [S. Chaudhuri & U. Dayal, 1997]

OLTP, Kennzahlen, mutlidimensionales Datenmodell, Stern- und Schneeflockenschema, Cubes

2.4 Preprocessing [Pyle 1999]

Ziele der Vorverarbeitung, typische Vorverarbeitungsschritte, Beispiele

2. 1 Datenbanksysteme

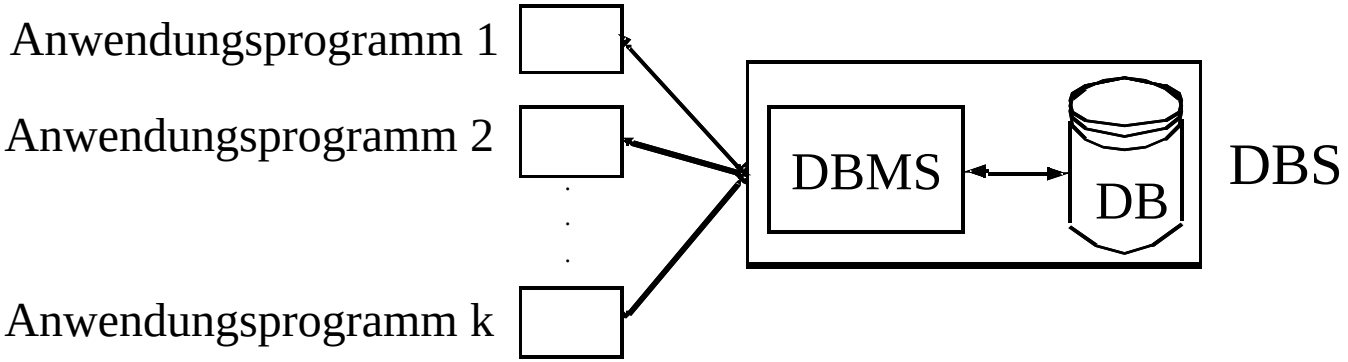
Definition

Ein *Datenbanksystem* (*DBS*) ist ein Software System zur dauerhaften Speicherung und zum effizienten Suchen in großen Datenmengen.

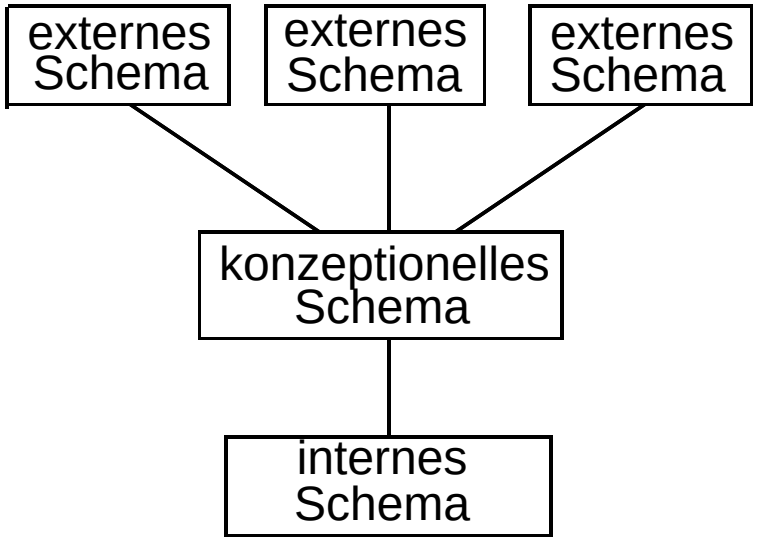
Komponenten

Datenbank (DB): Sammlung von Daten einer gegebenen Anwendung

Datenbank-Management-System (DBMS): Computer Programm zum Management von Datenbanken beliebiger Anwendungen in einem spezifizierten Format



Drei-Ebenen-Architektur



Spezielle Sichten verschiedener Benutzer / Anwendungsprogramme auf dieselbe Datenbank

Logische Sicht der ganzen Datenbank

Physische Speicherung der Datenbank

Beispiele

Kunde (KName, KAdr, Kto)
Auftrag (KName, Ware, Menge)
Lieferant (LName, LAdr, Ware, Preis)

```
select distinct Lname
from Lieferant, Auftrag
where Lieferant.Ware = Auftrag.Ware and KName =
'Huber'
```

```
select Ware, min (Preis), max (Preis), avg (Preis)
from Lieferant
group by Ware
order by Ware
```

2.1 Anfragebearbeitung

Prinzip

- eine SQL-Anfrage spezifiziert nur das „Was“
- der Anfrageoptimierer des DBMS bestimmt einen möglichst effizienten Anfrageplan, um die gegebene SQL-Anfrage zu beantworten
- Anfrageplan als *Operatorbaum*:
  - Die Blätter eines Operatorbaumes enthalten die auftretenden *Relationen*.
  - Die inneren Knoten repräsentieren die verwendeten *Operationen*.

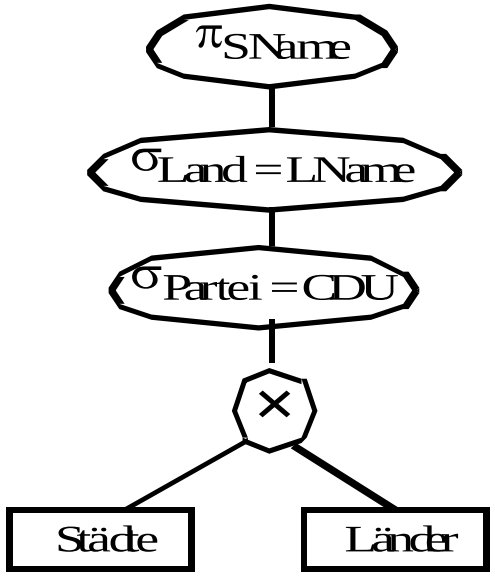
Ablauf

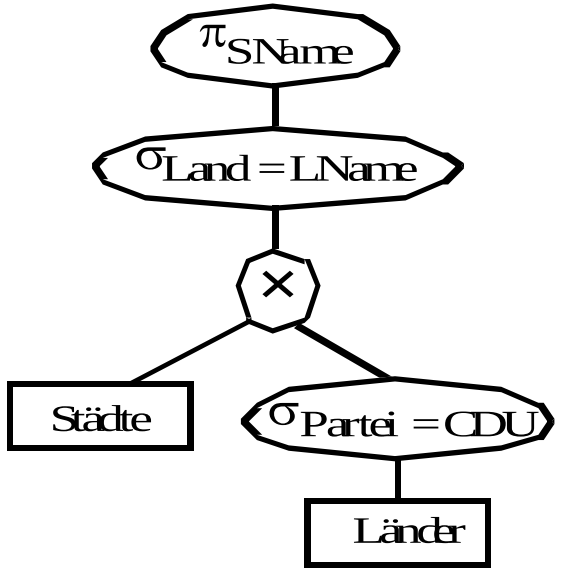
- Generierung von Anfrageplänen mit Hilfe von *heuristischen Regeln* (z.B. Selektionen vor Joins)
- Bewertung der Anfragepläne basierend auf einem *Kostenmodell* (Kostenmaß: Anzahl zu bearbeitender Tupel) und statistischen Angaben über die Ausprägung der Datenbank

Beispiel

```
Städte (SName, SEinw, Land)
Länder (LName, LEinw, Partei)

select Sname from Städte,Länder
where Land=Lname and Partei=CDU
```





2.1 Physische Speicherung der Daten

Prinzip der Magnetplatten

- *Seiten* (Blöcke) als *kleinste Transfereinheit* zwischen Haupt- und Sekundärspeicher
- *Feste Größe* zwischen 128 Byte und 16 KByte
- *Direkter Zugriff* auf eine Seite mit gegebener Seitennummer

Wahlfreier Zugriff

- Positionierung des Schreib-/Lesekopfes

  Zeit für die Kammbewegung [6 ms]
- Warten auf den Sektor / die Seite

  im Durchschnitt die halbe Rotationszeit der Platte [4 ms]
- Übertragung der Seite

  Zeit für Schreiben bzw. Lesen [0,1 ms / KByte]

sehr teuer im Vergleich zu Hauptspeicher-Operationen

2.1 Physische Speicherung der Daten

Sequentieller Zugriff

- Zugriff auf eine Menge von Seiten mit aufeinanderfolgenden Adressen
- ab der zweiten Seite entfällt der große Aufwand zur Positionierung des Schreib-/Lesekopfes und für das Warten auf die Seite
- sequentieller Zugriff ist wesentlich effizienter als wahlfreier Zugriff

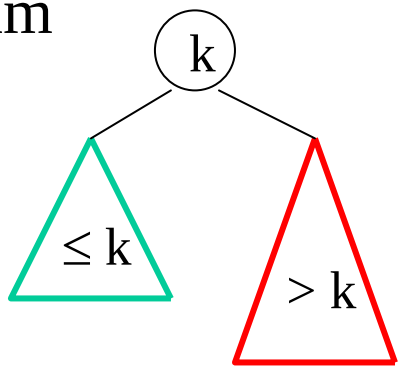
Kostenmaß für die Anfragebearbeitung

- Annahme: Zugriff auf Seiten erfolgt unabhängig voneinander
- sequentieller Zugriff ist dann nicht möglich
- Zeitaufwand für den wahlfreien Seitenzugriff ist um Größenordnungen höher als die Zeit für eine Operation im Hauptspeicher

Anzahl der Seitenzugriffe als Kostenmaß

Prinzipien

Suchbaum



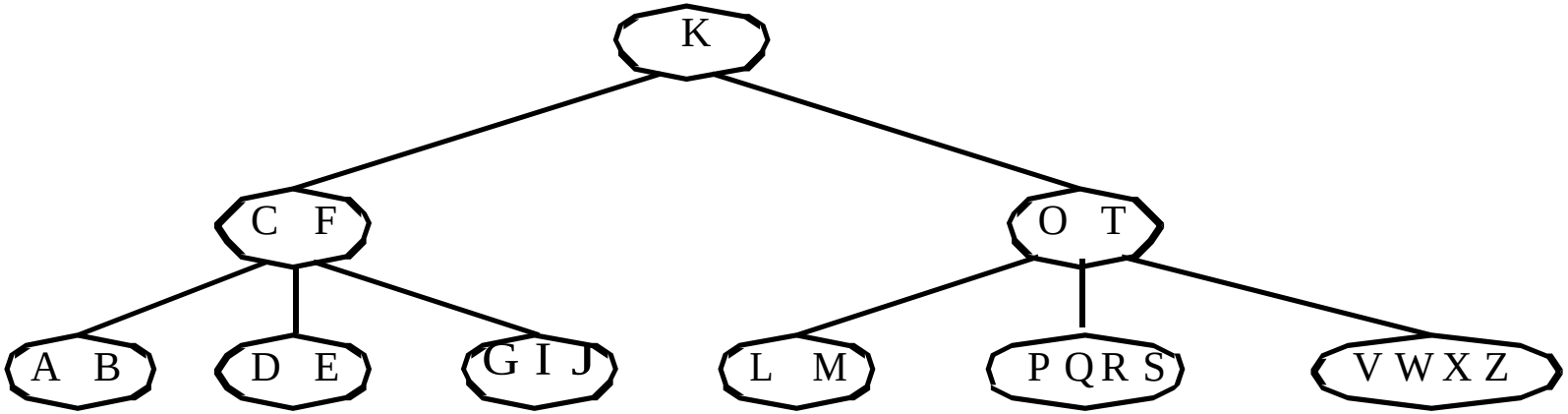
Balancierter Suchbaum

- alle Blätter des Baumes besitzen denselben Level
- die Höhe des Baumes ist O(log n) für n Datensätze
- die Operationen Einfügen, Entfernen und Suchen sind auf einen (oder wenige) Pfade beschränkt

⇒ Knoten des Baums = Seite der Magnetplatte

B-Baum

- Jeder Knoten enthält höchstens 2*m* Schlüssel.
- Jeder Knoten außer der Wurzel enthält mindestens *m* Schlüssel, die Wurzel mindestens einen Schlüssel.
- Ein Knoten mit *k* Schlüsseln hat genau *k*+1 Söhne.
- Alle Blätter befinden sich auf demselben Level.



Punktanfrage im B-Baum

```
PunktAnfrage (Seite s, Integer k);
  i:=1;
  while i < Anzahl der Einträge in s do
    if k ≤ i-ter Schlüssel in s then
      if s ist Datenseite then
        return i-ter Datensatz in s;
      else PunktAnfrage (i-ter Sohn von s, k);
    else i:= i + 1;
  if i = Anzahl der Einträge in s then
    PunktAnfrage (i-ter Sohn von s, k);
```

R-Baum

Vergleich mit B-Baum

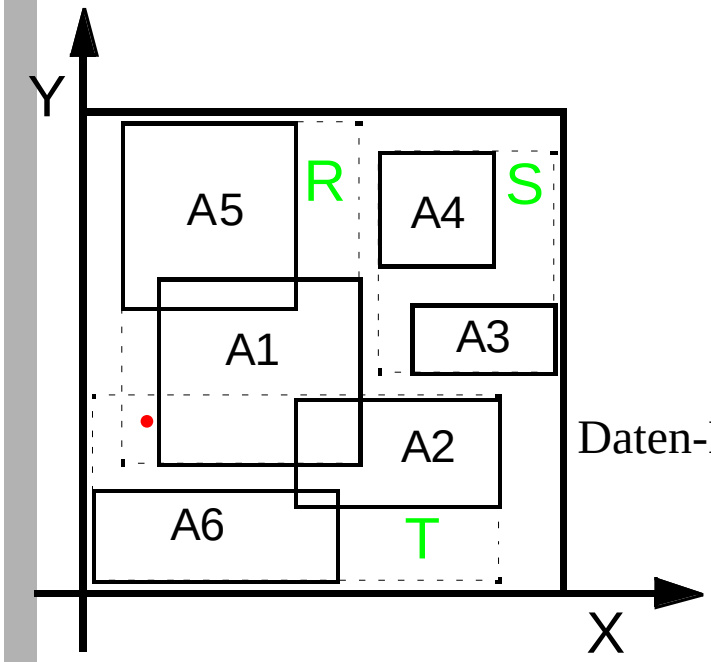
- B-Baum: eindimensionale Schlüssel (alphanumerische Werte)
- R-Baum: mehrdimensionale Schlüssel (Hyper-Rechtecke)

Definition

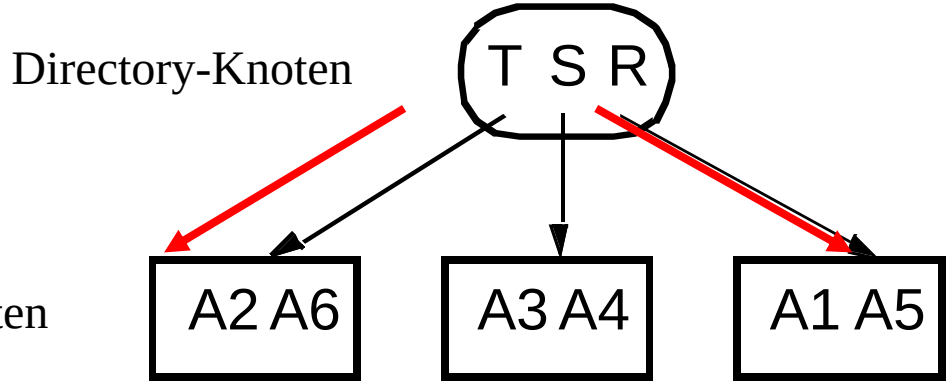
- Jeder Knoten außer der Wurzel besitzt zwischen *m* und *M* Einträge.
- Die Wurzel hat mindestens zwei Einträge, außer sie ist ein Blatt.
- Ein innerer Knoten mit *k* Einträgen hat genau *k* Söhne.
- Alle Blätter befinden sich auf demselben Level.

Punktanfrage im R-Baum

$M = 3, m = 1$



Daten-Knoten



• Anfragepunkt

Anfragebearbeitung ist *nicht* mehr auf einen Pfad beschränkt

Grundaufgaben

deskriptive Statistik

- beschreibende und graphische Aufbereitung von Daten
- auch zur Validierung der Daten

explorative Statistik

- wenn die Wahl eines geeigneten statistischen Modells unklar ist
- sucht nach Strukturen und Besonderheiten in den Daten

induktive Statistik

- basiert auf stochastischen Modellen
- zieht aus den beobachteten Daten Schlüsse auf umfassendere Grundgesamtheiten
- vorbereitende deskriptive und explorative Analysen nötig

Grundbegriffe

Stichprobenerhebung

- n Untersuchungseinheiten
- Werte $x_1,..., x_n$ eines Merkmals X beobachtet
- $h(a)$ die *absolute Häufigkeit* und die *relative Häufigkeit* des Attributwerts a in der Stichprobe

Typen von Merkmalen

- numerisch (mit totaler Ordnung < und arithmetischen Operationen)
- ordinal (mit totaler Ordnung <)
- kategorisch (keine Ordnung und keine arithmetischen Operationen)

2.2 Univariate Deskription

Lagemaße

- *arithmetisches Mittel* $\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$
- *Median*
  (seien dazu die x_i aufsteigend sortiert)

$$x_{med} = \begin{cases} x_{(n+1)/2} & \text{falls } n \text{ ungerade} \\ (x_{n/2} + x_{(n/2+1)})/2 & \text{falls } n \text{ gerade} \end{cases}$$

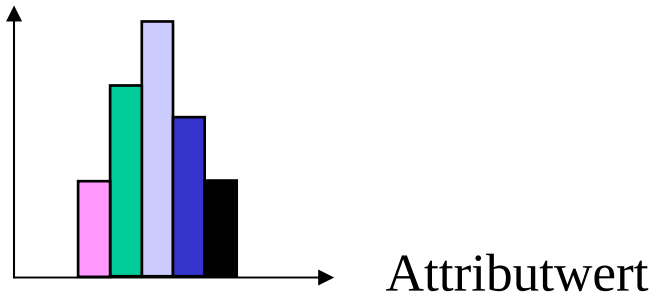
Streuungsmaße

– *Varianz* $\bar{s}^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2$

– *Standardabweichung* $\bar{s} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}$

nur für numerische Merkmale

Histogramme



Kontingenztabelle

- für kategorische Merkmale X und Y
- repräsentiert für zwei Merkmale *X* und *Y* die absolute Häufigkeit h_{ik} jeder Kombination (x_i, y_k) und alle Randhäufigkeiten $h_{.k}$ und $h_{i.}$ von *X* und *Y*

	Mittelfristige Arbeitslosigkeit	Langfristige Arbeitslosigkeit	
Keine Ausbildung	19	18	37
Lehre	43	20	63
	62	38	100

Wie sollten die relativen Häufigkeiten verteilt sein, wenn die beiden Merkmale keinerlei Abhängigkeit besitzen? $\frac{h_{ik}}{n} = \frac{h_{i.}}{n} \cdot \frac{h_{.k}}{n}$

χ^2-Koeffizient

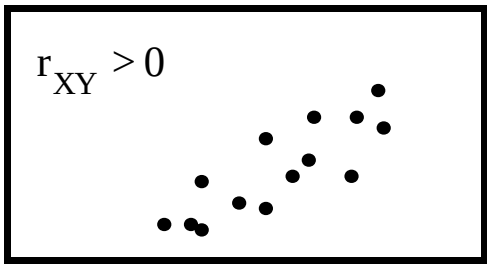
Differenz zwischen dem bei Unabhängigkeit erwarteten und dem tatsächlich beobachteten Wert von h_{ij} (Maß für die Stärke der Abhängigkeit)

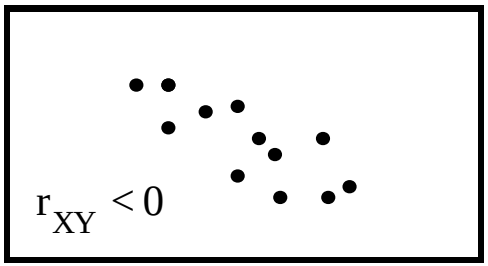
Korrelationskoeffizient

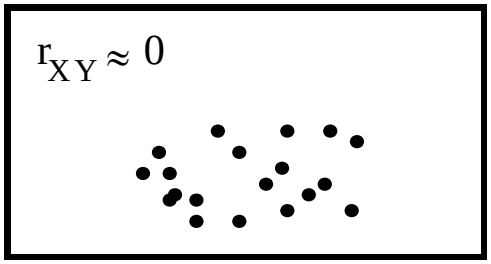
- für numerische Merkmale X und Y
- wie stark sind die Abweichungen vom jeweiligen Mittelwert korreliert?

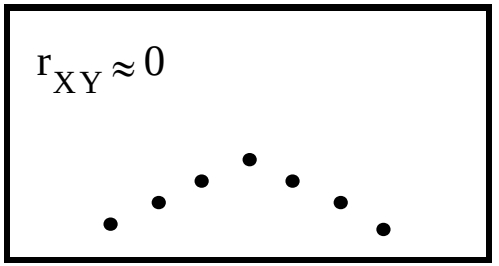
$$r_{XY} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

- Beispiele









Ereignisse und Wahrscheinlichkeitsmaße

- Ein *Zufallsvorgang* führt zu einem von mehreren sich gegenseitig ausschließenden Ergebnissen.
- Sei $\Omega = \{\omega_1, \ldots, w_n\}$ der *Ergebnisraum*, d.h. die Menge aller möglichen Ergebnisse eines Zufallsvorgangs.
- Teilmengen von Ω heißen *Ereignisse*.
- Ein *Wahrscheinlichkeitsmaß* ist eine Abbildung $P: 2^{\Omega} \rightarrow [0,1]$, die die folgenden Axiome erfüllt:

  (A1) $P(A) \geq 0$ für alle $A \subseteq \Omega$,

  (A2) $P(\Omega) = 1$,

  (A3) $P(A \cup B) = P(A) + P(B)$ für alle $A, B \subseteq \Omega$ mit $A \cap B = \varnothing$.

Bedingte Wahrscheinlichkeiten

- Seien $A, B \subseteq \Omega$. Die *bedingte Wahrscheinlichkeit* von A unter B, $P(A|B)$, ist definiert als

$$P(A|B) = \begin{cases} 0 \ \ falls \ P(B) = 0 \\ \dfrac{P(A \cap B)}{P(B)} \ \ sonst \end{cases}$$

- A und B heißen *unabhängig*, wenn gilt $P(A|B) = P(A)$ und $P(B|A) = P(B)$.

Satz von Bayes

Sei $A_1, \ldots, A_k$ eine disjunkte Zerlegung von Ω , so daß für mindestens ein i, $1 \leq i \leq k$, gilt: $P(A_i) > 0$ und $P(B|A_i) > 0$. Dann gilt für alle $1 \leq j \leq k$:

$$P(A_j|B) = \frac{P(B|A_j) \cdot P(A_j)}{P(B)}$$

a-priori-Wahrscheinlichkeit: $P(A_i)$

a-posteriori-Wahrscheinlichkeit: $P(A_i|B)$

Grundbegriffe

- *Zufallsvariable*
  Merkmal, dessen Werte die Ergebnisse eines Zufallsvorgangs sind
- *diskrete Zufallsvariable*
  endlich oder abzählbar unendlich viele verschiedene Werte $x_1, x_2, \ldots, x_k, \ldots$
- *Wahrscheinlichkeitsfunktion* $f(x) = \begin{cases} P(x_i) \ \textit{falls} \ x = x_i \\ 0 \ \textit{sonst} \end{cases}$
- *Verteilungsfunktion* $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$
- *Erwartungswert* $E(X) = P(X \leq x) = \sum_{i \geq 1} x_i \cdot f(x_i)$
- *Varianz* $Var(X) = \sum_{i \geq 1} (x_i - E(X))^2 \cdot f(x_i)$

Binomialverteilung

- *Bernoulli-Experiment*: nur zwei Ergebnisse (Treffer oder Nichttreffer), p die Wahrscheinlichkeit des Treffers
- n unabhängige Wiederholungen desselben Bernoulli-Experiments, die Gesamtanzahl der Treffer wird beobachtet
- *binomialverteilte* Zufallsvariable mit den Parametern n und p

  besitzt folgende Wahrscheinlichkeitsfunktion:

$$f(x) = \begin{cases} \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x} & \textit{falls } x \in \{0,1,\ldots,n\} \\ 0 & \textit{sonst} \end{cases}$$

- Erwartungswert und Varianz einer binomialverteilten Zufallsvariablen:

$$E(X) = n \cdot p \qquad Var(X) = n \cdot p \cdot (1-p)$$

Beispiel einer Binomialverteilung

- Anwendung: Abschätzung des (auf einer Stichprobe bestimmten) Klassifikationsfehlers auf der Grundgesamtheit
- Bernoulli-Experiment: zufälliges Ziehen eines Objekts der Grundgesamtheit und Test, ob dieses Objekt von dem Klassifikator falsch klassifiziert wird
- Treffer: Objekt wird falsch klassifiziert
- Nichttreffer: Objekt wird korrekt klassifiziert
- p: Wahrscheinlichkeit einer Fehlklassifikation in der Grundgesamtheit
- n: Größe der Trainingsmenge

⇒ gesucht ist ein Intervall $[u,o]$, so daß mit einer Wahrscheinlichkeit von z.B. mindestens 95 % gilt

$$u \leq p \leq o$$

Grundbegriffe

- überabzählbar unendlich viele verschiedene Werte $x_1, x_2, \ldots, x_k, \ldots$
- Eine Zufallsvariable X heißt *stetig*, wenn es eine Funktion (*Wahrscheinlichkeits-Dichte*) $f(x) \geq 0$ gibt, so daß für jedes Intervall [a,b] gilt:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- *Verteilungsfunktion* $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$
- *p-Quantil* x_p mit $F(x_p) = p$
- *Erwartungswert* $E(x) = \int_{-\infty}^{+\infty} x \cdot f(x)dx$

Normalverteilung

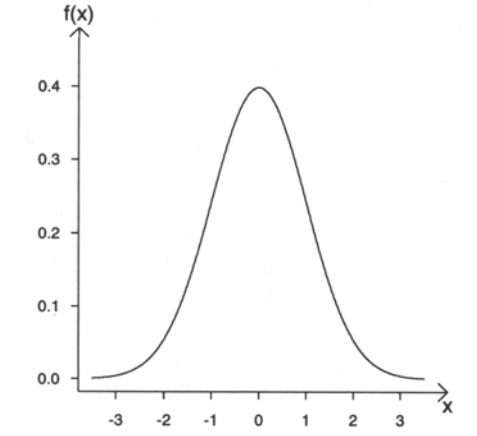
• Eine Zufallsvariable *X* heißt *normalverteilt* (bzw. *gaußverteilt*) mit den Parametern $\mu \in IR$ und $\sigma^2 > 0$, wenn sie folgende Dichte besitzt:

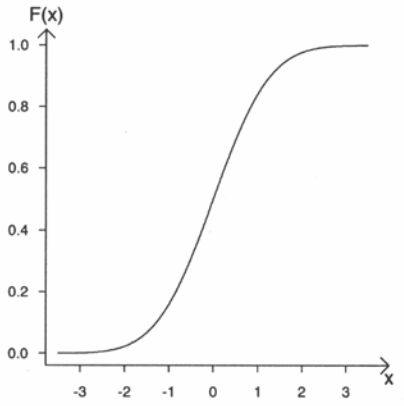
$$f(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

• Erwartungswert $E(x) = \mu$

• *standardisierte Zufallsvariable* $Z = \frac{X - \mu}{\sigma}$

standardnormalverteilt (normalverteilt mit Parametern $\mu = 0$ und $\sigma^2 = 1$)





Schwankungsintervall

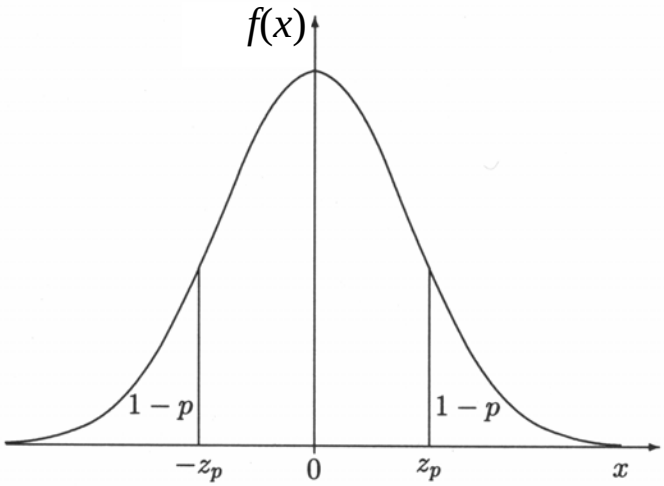
- *Schwankungsintervall* $\mu - c \le X \le \mu + c$
- es gilt $x_p = z_p \cdot \sigma + \mu$
- Wahrscheinlichkeit dafür, daß der Wert von X im Schwankungsintervall liegt:

$$P(\mu - \sigma \cdot z_{1-\alpha/2} \le X \le \mu + \sigma \cdot z_{1-\alpha/2}) = 1-$$

α *Irrtumswahrscheinlichkeit*

- es gilt z.B.

$$P(\mu - \sigma \le X \le \mu + \sigma) = 0{,}6827$$
$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0{,}9545$$
$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0{,}9973$$



Grundbegriffe

- Ziel aller Schätzverfahren: aus einer Zufallsstichprobe der Größe *n* auf die Grundgesamtheit zurückzuschließen.
- *Punktschätzung*: ein möglichst genauer Näherungswert für einen unbekannten Parameter einer Grundgesamtheit
- *Schätzfunktion* oder *Schätzstatistik* für den Grundgesamtheitsparameter θ: $T = g(X_1,\ldots,X_n)$ mit Stichprobenvariablen $X_1, \ldots, X_n$
- *Schätzwert*: durch Einsetzen der für $X_1, \ldots, X_n$ beobachteten Werte
- *Intervallschätzung* konstruiert ein Intervall, das mit vorgegebener Wahrscheinlichkeit den tatsächlichen Parameterwert enthält
- Zu gegebener *Irrtumswahrscheinlichkeit* α und Stichprobenvariablen $X_1, \ldots, X_n$ liefern die Schätzstatistiken G_u und G_o ein $(1-\alpha)$-*Konfidenzintervall* für den Grundgesamtheitsparameter θ, wenn gilt

$$P(G_u \le \theta \le G_o) = 1 - \alpha$$

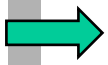
Maximum-Likelihood-Schätzer

- Seien $X_1, \ldots, X_n$ Zufallsvariablen mit gemeinsamer Wahrscheinlichkeits- bzw. Dichtefunktion $f(x_1, \ldots, x_n; \theta)$ mit unbekanntem Parameter θ
- *Likelihoodfunktion* $f(x_1, \ldots, x_n; \theta)$

  für $x_1, \ldots, x_n$ die in der Stichprobe beobachteten Werte einsetzen

  nur die Abhängigkeit von θ betrachten
- *Maximum-Likelihood-Schätzer* liefert den Wert θ, für den gilt:

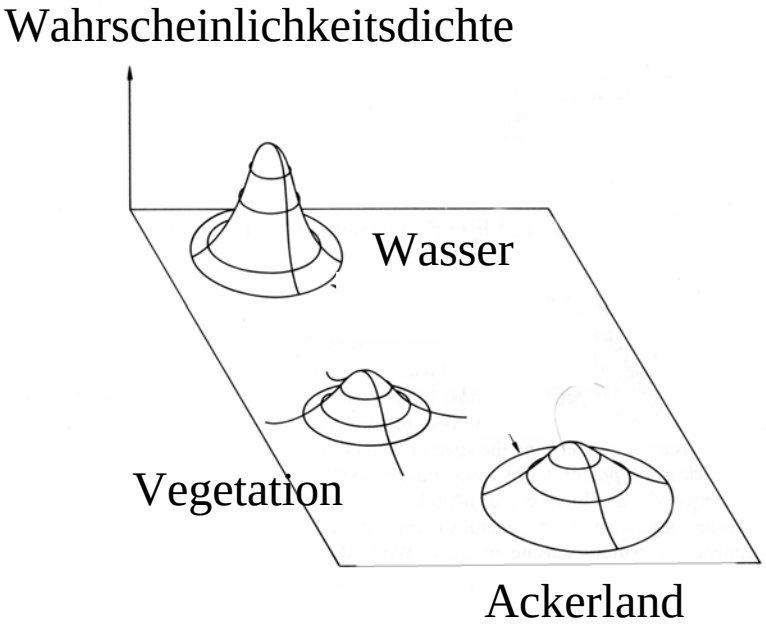
$$\forall \theta': f(x_1, \ldots, x_n; \theta) \geq f(x_1, \ldots, x_n; \theta')$$



wähle den Wert für θ, bei dem die höchste Wahrscheinlichkeit für das Auftreten von $x_1, \ldots, x_n$ besteht

Beispiel eines Maximum-Likelihood-Schätzers

- Gegeben
  Grauwertvektoren ($d_1(x,y)$, $d_2(x,y)$)
  für jedes Pixel (x,y)
- Klassen
  Wasser, Vegetation, Ackerland, etc.
- Die Grauwertvektoren jeder Klasse seien normalverteilt.
- Gesucht
  Klasse eines Pixels mit gegebenem Grauwertvektor



- *Maximum-Likelihood-Klassifikator* (vereinfacht)
  entscheidet sich für die Klasse, deren Wahrscheinlichkeitsdichte für den beobachteten Grauwertvektor maximal ist

Grundbegriffe

- *Nullhypothese* H_0 und *Alternative* H_1, die sich gegenseitig ausschließen
- Annahmen über die Verteilung oder bestimmte Parameter des interessierenden Merkmals in der Grundgesamtheit
- *Fehler 1. Art*

  H_0 wird verworfen, obwohl H_0 wahr ist
- *Fehler 2. Art*

  H_0 wird akzeptiert wird, obwohl H_1 wahr ist
- *Test zum Signifikanzniveau* α $(0 < \alpha < 1)$
  ein Hypothesen-Test, bei dem die Wahrscheinlichkeit eines Fehlers 1. Art höchstens α beträgt

2.2 Testen von Hypothesen

Tests für eine Stichprobe

• Werte für das zu untersuchende Merkmal werden in einer Stichprobe erhoben
• verschiedene Hypothesen über dieses Merkmals in der Grundgesamtheit, z.B.

H_0 : „Die zu erwartende Nettomiete in Stadtviertel *A* beträgt 15 DM/qm.“ oder

H_0 : „Die Nettomiete in Stadtviertel *A* ist normalverteilt.“.

Tests für zwei unabhängige Stichproben

• zwei unabhängigen Stichproben
• bestimmte Eigenschaften dieses Merkmals in den beiden Grundgesamtheiten vergleichen, z.B.

H_0 : „Die zu erwartende Nettomiete in den Stadtvierteln *A* und *B* ist identisch.“ oder

H_0 : „Das Einkommen weiblicher Arbeitnehmer besitzt dieselbe Verteilung wie das Einkommen männlicher Arbeitnehmer.“.

Test für einen Erwartungswert (*Gauß-Test*)

- gegeben unabhängig identisch verteilte Zufallsvariablen $X_1,..., X_n$:

  X_i normalverteilt mit den Parametern μ und σ^2, wobei σ^2 bekannt ist, oder
  X_i beliebig verteilt mit $E(X_i) = \mu$, bekanntes $Var(X_i) = \sigma^2$ und n „groß genug“

- Testprobleme

  (a) H_0 : „$\mu = \mu_0$“ gegen H_1 : „$\mu \neq \mu_0$“, (b) H_0 : „$\mu = \mu_0$“ gegen H_1 : „$\mu < \mu_0$“,
  (c) H_0 : „$\mu = \mu_0$“ gegen H_1 : „$\mu > \mu_0$“.

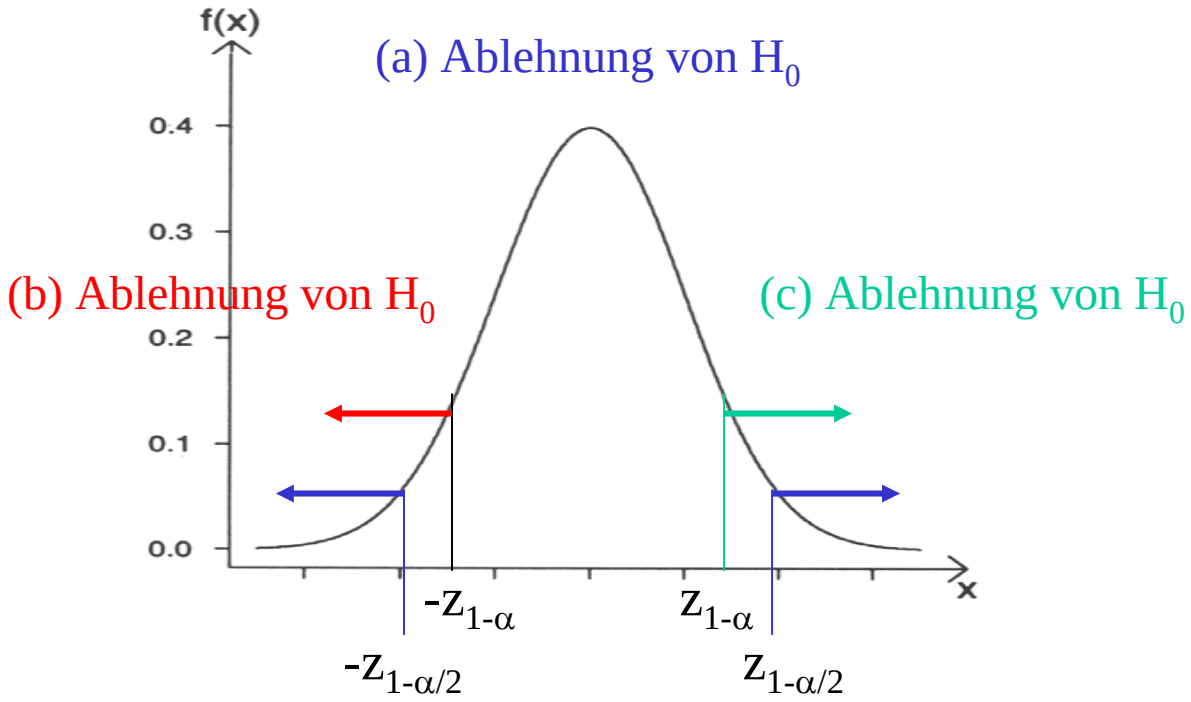
- Methode des Tests

  Falls H_0 wahr ist, ist $\frac{\overline{X} - \mu_0}{\sigma} \cdot \sqrt{n}$ standardnormalverteilt.
  H_0 wird abgelehnt und die Alternative H_1 akzeptiert, falls:

  (a) $|z| > z_{1-\alpha/2}$ (b) $z < -z_{1-\alpha}$ (c) $z > z_{1-\alpha}$

Gauß-Test

(a) H_0 : „$\mu = \mu_0$“ gegen H_1 : „$\mu \neq \mu_0$“, (b) H_0 : „$\mu = \mu_0$“ gegen H_1 : „$\mu < \mu_0$“,
(c) H_0 : „$\mu = \mu_0$“ gegen H_1 : „$\mu > \mu_0$“.



OLAP

2.3 OLAP

2.3.1 Einführung in OLAP

Wie gesehen, gibt es große Unterschiede zwischen operativen Systemen und dem DWh

Entsprechend gibt es fundamentale Unterschiede auch zwischen den jeweiligen Zugriffsarten auf diese Datenquellen:

- **OLAP = On-Line Analytical Processing** benutzt DWh
- **OLTP = On-Line Transaction Processing** benutzt operative Systeme

2.3.1 Einführung in OLAP

OLTP

- hohe Zahl **kurzer**, atomarer, isolierter, wiederkehrender Transaktionen
  - z.B. Konto-Update, Flugbuchung, Telefon-Gespräch
- Transaktionen benötigen detaillierte, aktuelle Daten
- Daten werden (oft tupelweise) gelesen und relativ **häufig aktualisiert**
- Transaktionen dienen dem **Tagesgeschäft** und haben relativ hohe Ansprüche an die Bearbeitungsgeschwindigkeit

2.3.1 Einführung in OLAP

Definition von OLAP:

- **OLAP Systeme**
  - dienen der **Entscheidungs-Unterstützung** oder
  - können in den Phasen **„Data Understanding“** bzw. **„Data Preparation“** im Rahmen des Data-Mining-Prozesses eingesetzt werden.

- **OLAP-Funktionen** erlauben
  - den schnellen, **interaktiven** Zugriff auf Unternehmensdaten
  - unter „beliebigen“ unternehmensrelevanten Blickwinkeln (**Dimensionen**)
  - auf verschiedenen **Aggregationsstufen**
  - mit verschiedenen Techniken der Visualisierung

- Hauptmerkmal ist die **multi-dimensionale** Sichtweise auf Daten mit flexiblen interaktiven Aggregations- bzw. Verfeinerungsfunktionen entlang einer oder mehrerer Dimensionen.

2.3.1 Einführung in OLAP

Multi-Dimensionalität:

- Mehrdimensionale Sichtweise auf Daten ist sehr **natürlich**.
- Sichtweise der Analysten auf Unternehmen **ist** mehrdimensional.

  $\Rightarrow$ Konzeptuelles Datenmodell sollte mehrdimensional sein, damit Analysten leicht und intuitiv Zugang finden.

- **Beispiel:** *Verkaufszahlen* können nach unterschiedlichen Kriterien / Dimensionen aggregiert und analysiert werden.
  - nach **Produkt:** *Produkt, Produktkategorie, Industriezweig*
  - nach **Region:** *Filiale, Stadt, Bundesland*
  - nach **Zeit:** *Tag, Woche, Monat, Jahr*
  - nach verschiedenen Dimensionen des Käufers: **Alter, Geschlecht, Einkommen** des Käufers
  - und nach **beliebigen Kombinationen von Dimensionen**, z.B.
    - nach *Produktkategorie, Stadt* und *Monat*

2.3.1 Einführung in OLAP

Kennzahlen:

- Die **Analyse-Gegenstände** von OLAP sind **numerische Werte,** typischerweise **Kennzahlen** genannt (oder auch Maße, Metriken oder Fakten).
  - **Beispiel:** *Verkaufszahlen, Umsatz, Gewinn, Lagerbestand,...*
- Diese numerischen Werte lassen sich auf verschiedene Weise verdichten, z.B.
  - Summenbildung
  - Mittelwertbildung
  - Minimum- oder Maximumbestimmung
  - Die zulässige Art der Verdichtung hängt vom **Skalenniveau** der Kennzahl ab.

2.3.1 Einführung in OLAP

Skalenniveaus

In der Statistik unterscheidet man die Attributausprägungen einer vorgegebenen Menge von Daten mittels Skalen mit unterschiedlichem Skalenniveau. Die wichtigsten Typen sind:

Nominalskalierte Merkmale:
Ausprägungen sind ‘‘Namen‘‘, keine Ordnung möglich
→ keine Aggreagation möglich

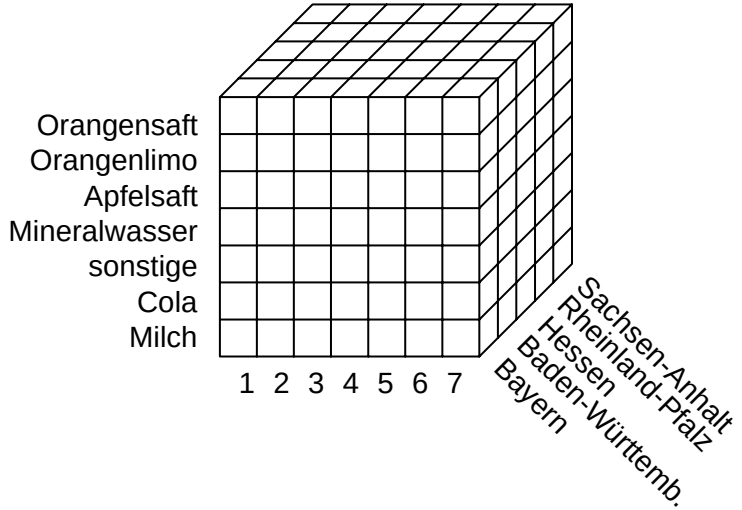
Ordinalskalierte Merkmale:
Ausprägungen können geordnet, aber Abstände nicht interpretiert werden.
→ Median macht Sinn, Mittelwert z.B. nicht

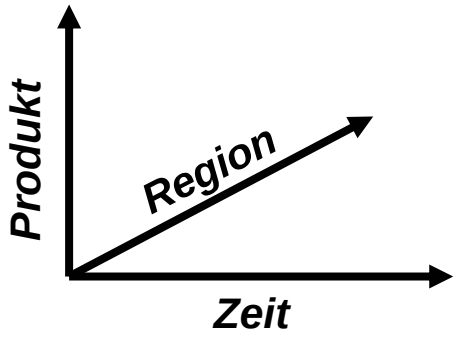
Kardinalskalierte Merkmale:
Ausprägungen sind Zahlen, Interpretation der Abstände möglich (metrisch)
→ Mittelwertbildung, Standardabweichung etc. sinnvoll

2.3.1 Einführung in OLAP

Dimensionen:

- Jede Kennzahl hängt von einer Menge von **Dimensionen** ab.
  Diese bilden den **Kontext der Kennzahlen.**
  - **Beispiel:** Die *Verkaufszahlen* (Kennzahl) hängen von den Dimensionen *Produkt, Region* und *Zeit* ab.
  - Die Dimensionen sind **orthogonal (unabhängig).**
  - Sie definieren einen sog. **Hyper-Würfel (hyper cube).**





- Es kann eine beliebige Zahl an Dimensionen geben (abhängig vom Zweck des OLAP-Systems und der enthaltenen Daten).
  In manchen Anwendungen treten bis zu 50 Dimensionen auf.

2.3.1 Einführung in OLAP

Dimension Zeit:

- **Spezielle Dimension,** die in jedem OLAP-System existiert, ist die ***Zeit.***
- Leistung eines Unternehmens wird immer anhand der Zeit bewertet:
  - aktueller Monat im Vergleich zu letztem Monat
  - aktueller Monat im Vergleich zum gleichen Monat des Vorjahres
- Dimension *Zeit* unterscheidet sich von allen anderen Dimensionen:
  - Zeit hat einen linearen Charakter:
    - Januar kommt vor Februar
  - Zeit hat Wiederholungscharakter: jeden Montag, werktags, ...
- OLAP-System muss Umgang mit der Dimension Zeit und den damit verbundenen Besonderheiten unterstützen.

Attribute und Attributelemente:

Jede Dimension ist durch eine **Menge von Attributen** charakterisiert.

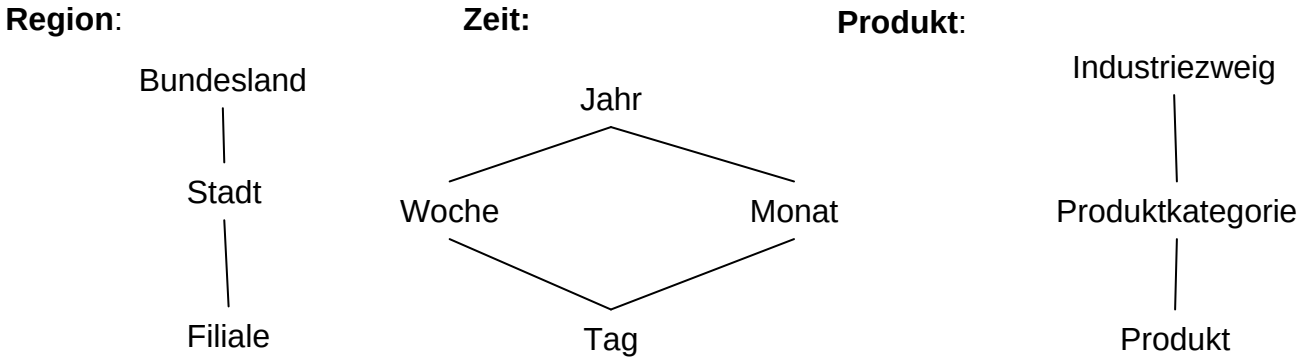
- **Beispiel:** Die Dimension *Region* ist charakterisiert durch die Attribute *Filiale, Stadt* und *Bundesland.*

2.3.1 Einführung in OLAP

Attribute und Attributelemente:

•Diese Attribute können **hierarchisch** angeordnet sein (Aggregationsstufen)

- **Beispiel:**
  - Gesamtwert ergibt sich aus den Werten mehrerer *Bundesländer.*
  - Wert für ein *Bundesland* ergibt sich aus Werten mehrerer *Städte*.
  - Wert für eine Stadt ergibt sich aus Werten mehrerer *Filialen.*



2.3.1 Einführung in OLAP

- Ein Pfad in einer solchen **Attribut-Hierarchie** (z.B. *Tag, Monat, Jahr*) wird auch **consolidation path** genannt.

- Jedes Attribut einer Dimension wird durch **Attributelemente** instantiert.
  - **Beispiel:**
    - Das Attribut ***Produkt*** der Dimension *Produkt* hat die Attributelemente: *Coca-Cola, Pepsi-Cola, Afri-Cola, ...*
    - Das Attribut ***Produktkategorie*** hat die Attributelemente: *Orangensaft, Apfelsaft, Orangenlimo, Cola,...*
    - Das Attribut ***Industriezweig*** hat die Attributelemente: *Lebensmittelindustrie, Textilindustrie, Schwerindustrie,...*

2.3.2 OLAP Funktionalität

- Bei der Analyse können beliebige Aggregationsstufen visualisiert werden:

  Drill-Down bzw. **Roll-Up-**Operationen

- Bedingungen an Dimensionen, Attribute und Attributelemente reduzieren Dimensionalität der visualisierten Daten:

  Slice & Dice - Operationen

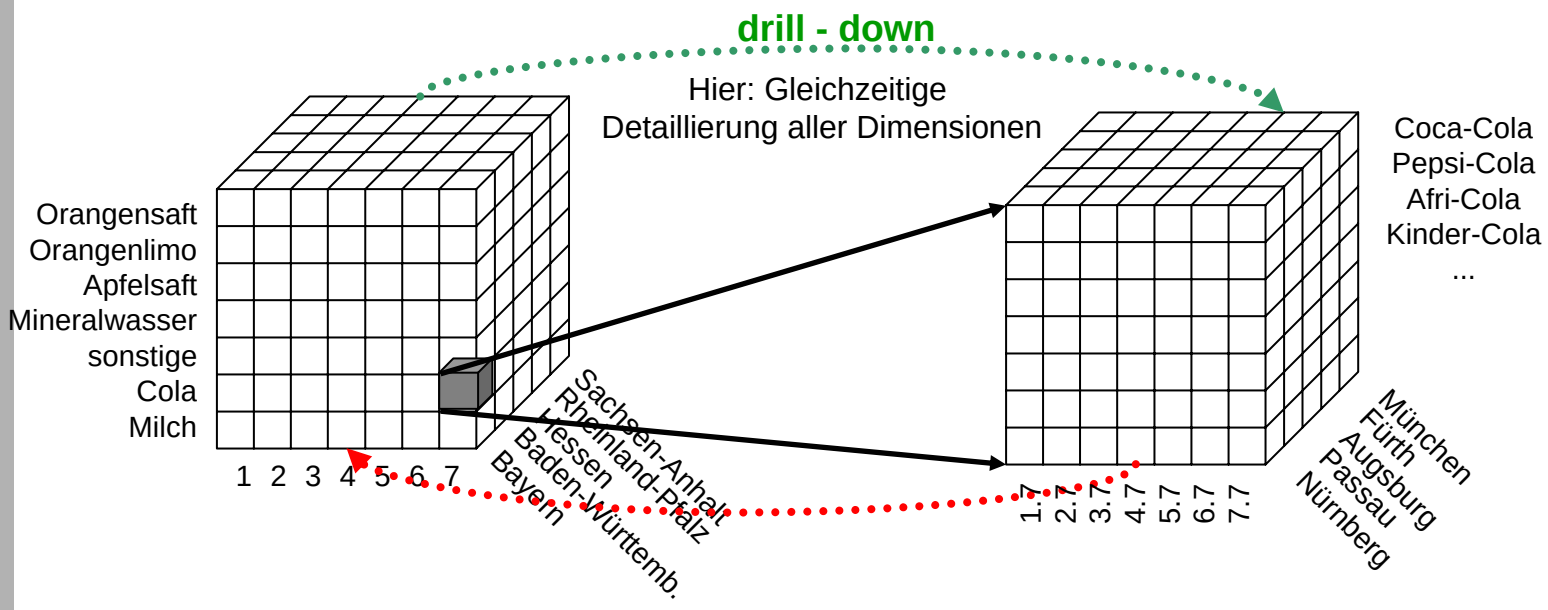
- Analyse wird durch Vielzahl von **Visualisierungstechniken** unterstützt. Bedingungen werden **interaktiv** gewählt (Buttons, Menüs, *drag & drop*), so dass Analysten und Manager keine komplizierte Anfragesprache lernen müssen.

2.3.2 OLAP-Funktionalität

Drill-Down und Roll-Up

- Entlang der Attribut-Hierarchien werden die Daten **verdichtet** bzw. wieder **detailliert** und sind so auf verschiedenen **Aggregationsstufen** für Analysen zugreifbar.

- Verdichtung/Detaillierung kann entlang einer, mehrerer oder aller Dimensionen geschehen - gleichzeitig oder in beliebiger Reihenfolge.



2.3.2 OLAP-Funktionalität

Slice & Dice:

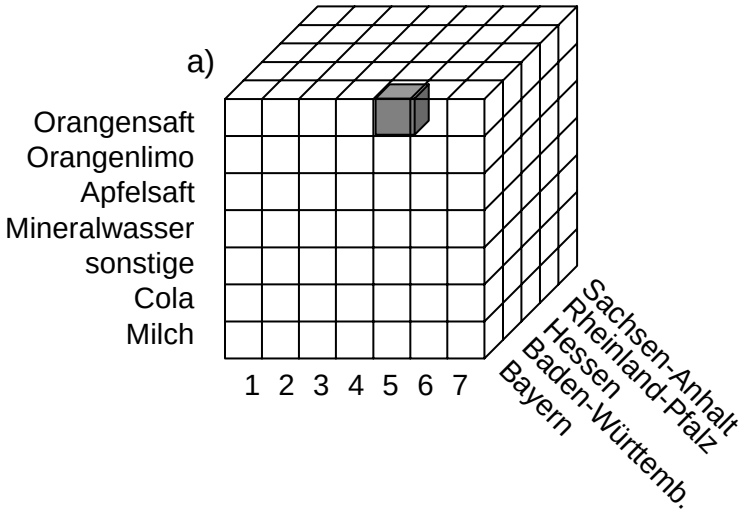
- Bei dieser Operation wird die **Dimensionalität** der visualisierten Daten **reduziert.**
- Zu einer Teilmenge der Dimensionen (sog. **page dimensions**) werden Bedingungen formuliert.
- Alle Daten in der resultierenden Tabelle genügen diesen Bedingungen.
- Die **page dimensions** tauchen in der neuen Tabelle nicht mehr explizit auf, sondern definieren implizit die Menge dargestellter Daten.

Slice & Dice entspricht dem Herausschneiden einer Scheibe (*slice*) aus dem Hyper-Würfel. Nur diese Scheibe wird weiterhin visualisiert.

2.3.2 OLAP-Funktionalität

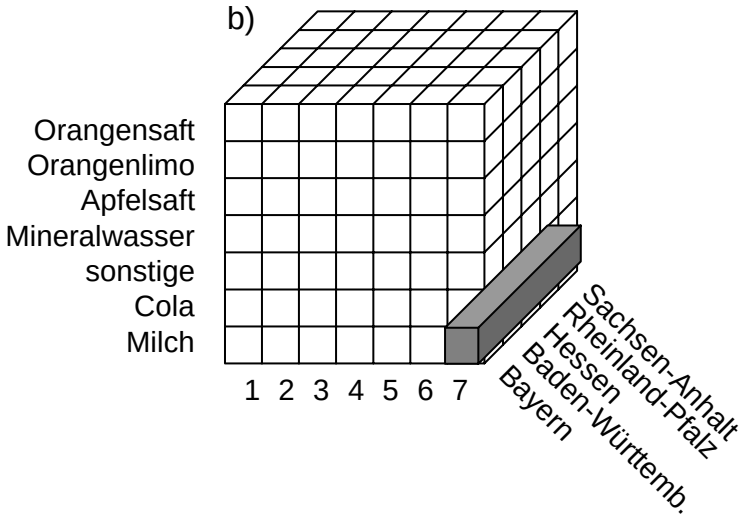
Beispiele:

Lokation bestimmter atomarer und aggregierter Werte im Hyper-Würfel.



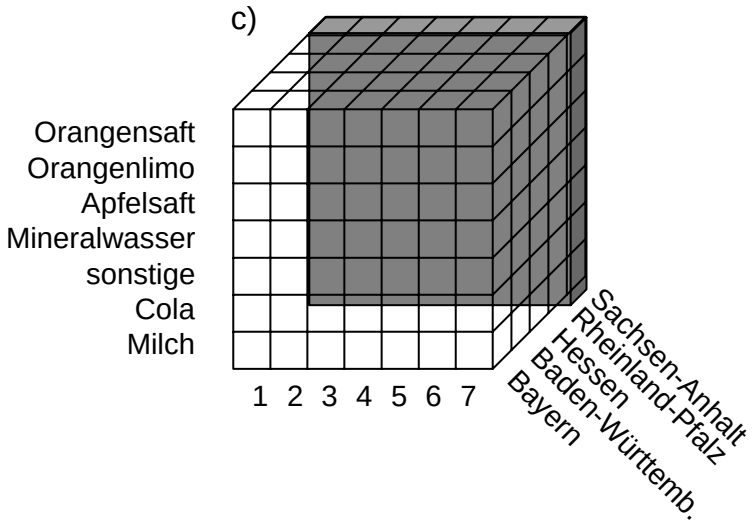
a) Verkaufszahlen für Orangensaft in Bayern im Mai

2.3.2 OLAP-Funktionalität



b) Verkaufszahlen für Milch in ganz Süddeutschland im Juli

2.3.2 OLAP-Funktionalität



c) Verkaufszahlen insgesamt für Sachsen-Anhalt

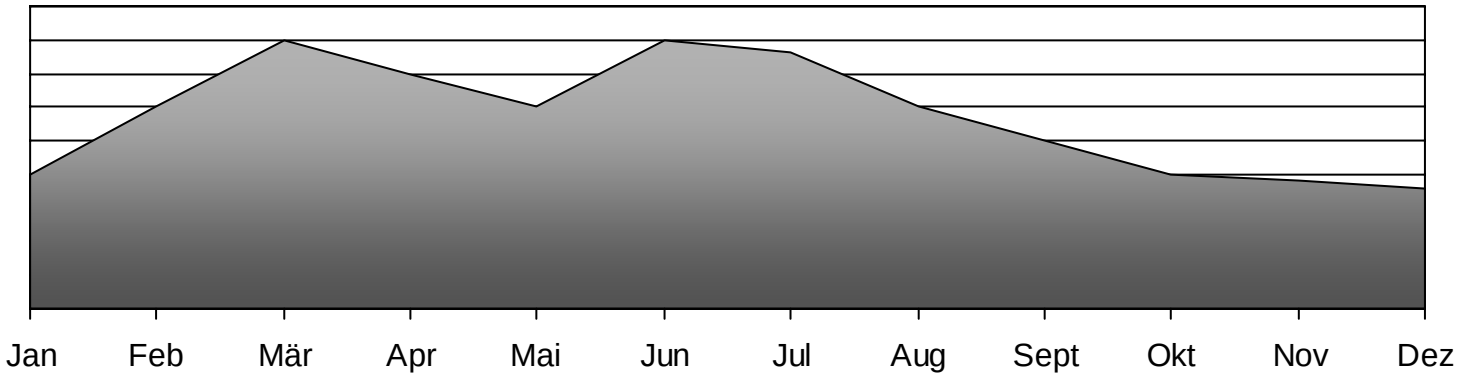
⇒ Aggregation der Verkaufszahlen über alle Monate **und** alle Produkte

2.3.2 OLAP-Funktionalität

- Analyse bezieht sich nur selten auf einen Wert:
  - sondern auf eine Folge von Werten
    $\Rightarrow$ Entwicklungen und **Trends** erkennbar (d)
  - oder auf eine Menge von Werten
    $\Rightarrow$ Vergleiche verschiedener Werte ermöglicht (e)

2.3.2 OLAP-Funktionalität

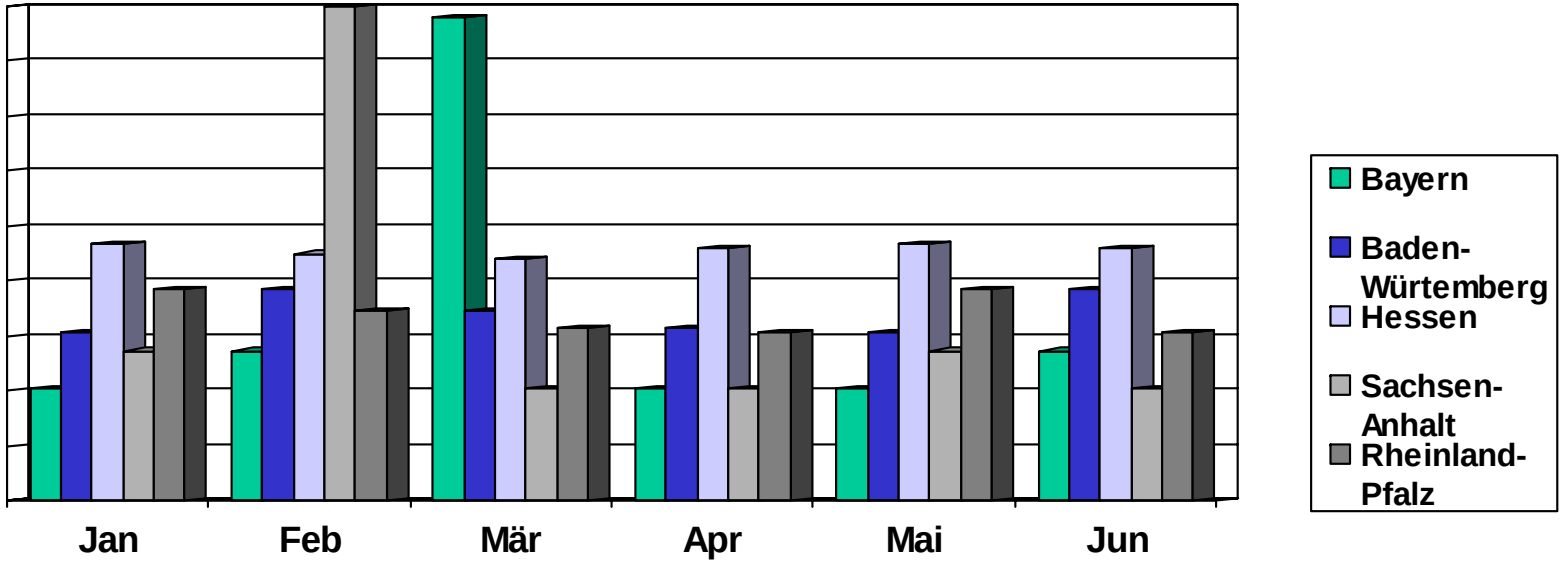
d) Entwicklung der Verkaufszahlen für Apfelsaft in Baden-Württemberg im letzten Jahr.



page dimensions: Produkt = Apfelsaft, Region = Baden-Württemberg

2.3.2 OLAP-Funktionalität

e) Vergleich der Verkaufszahlen für Apfelsaft in den Regionen Deutschlands für das erste Halbjahr



page dimensions: Produkt = Apfelsaft

2.3.3 Mehrdimensionales Datenmodell

2.3.3 Mehrdimensionales Datenmodell

Der beste Weg um zu einem OLAP-fähigen DWh zu kommen:

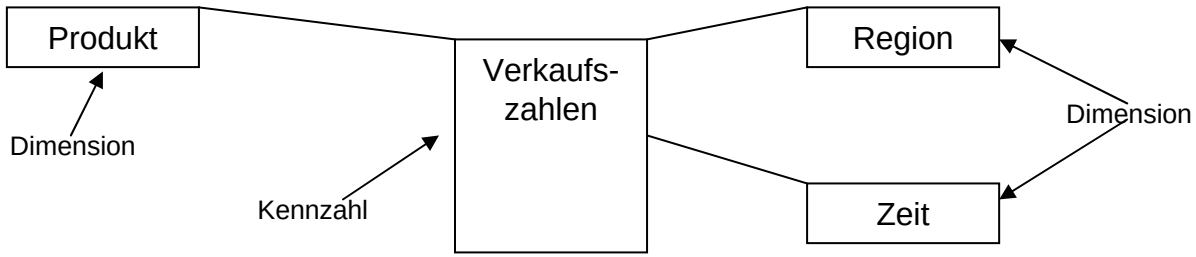
1. Erstellen eines **mehrdimensionalen** konzeptuellen Datenmodells.
2. Ableiten eines **relationalen** logischen Datenmodells.
    - Relationale DBS bilden die Implementierungsebene des DWh

Stern-Schema: (star schema)

- mehrdimensionales Datenmodell durch **Stern-Schema** realisierbar.
- Konstrukte eines Stern-Schemas:
    - **Kennzahlen:** Gegenstände der Analyse: Verkaufszahlen
    - **Dimensionen** definieren den Kontext der Kennzahlen: Produkt, Region, Zeit

2.3.3 Mehrdimensionales Datenmodell

Beispiel:



2.3.3 Mehrdimensionales Datenmodell

Vorteile des Stern-Schemas gegenüber herkömmlichen relationalen Schemata:

- Schema-Entwurf entspricht der **natürlichen Sichtweise** der Benutzer
  - Daten können in einer für Analysen adäquaten Weise zugegriffen werden.
- **Erweiterungen** und **Änderungen** am Schema sind leicht zu realisieren.
- **Beziehungen** zwischen den Tabellen sind **vordefiniert**
  - Join-Operationen können durch entsprechende Zugriffspfade unterstützt werden
  - Schnelle Antwortzeiten sind möglich
- Stern-Schema kann leicht in relationales DB-Schema umgesetzt werden.

2.3.3 Mehrdimensionales Datenmodell

- Umsetzung des Stern-Schemas in relationale Tabellen:
  - **Kennzahlentabelle (major table):** Die Gegenstände der Analyse (Kennzahlen) werden in dieser Tabelle gesichert
  - **Nebentabelle (minor tables):** Jede Dimension wird zu einer eigenen Relation / Tabelle.

Kennzahlentabelle:

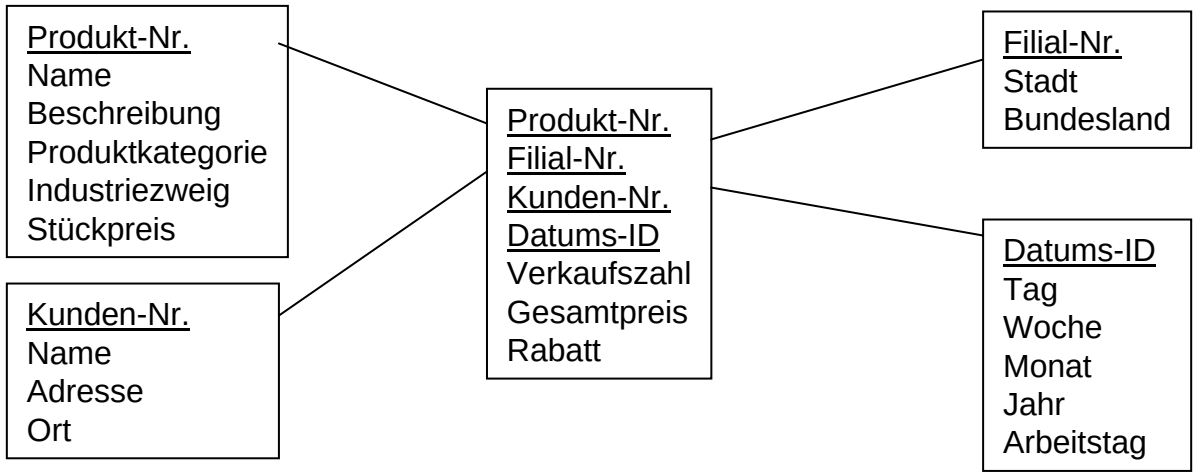
- Jedes **Tupel der Kennzahlentabelle** besteht aus
  - einem Zeiger für jede Dimensionstabelle (Fremdschlüssel), die den Kontext eindeutig definieren und
  - den numerischen Werten (**Daten**) für den jeweiligen Kontext.
- Sie enthält die eigentlichen Geschäftsdaten, die analysiert werden sollen.
- Die Kennzahlentabelle kann sehr viele Zeilen enthalten (Millionen).
- Der Schlüssel der Kennzahlentabelle wird durch die Gesamtheit der Dimensionszeiger gebildet

2.3.3 Mehrdimensionales Datenmodell

Dimensionstabelle:

- Jede **Dimensionstabelle** enthält
  - einen eindeutigen Schlüssel (z.B. Produktnummer) und
  - beschreibende Daten der Dimension (**Attribute**).
- Dimensionstabellen sind deutlich kleiner als die Kennzahlentabelle.
- Zusammenhang zur Kennzahlentabelle über Schlüssel/Fremdschlüssel-Relation

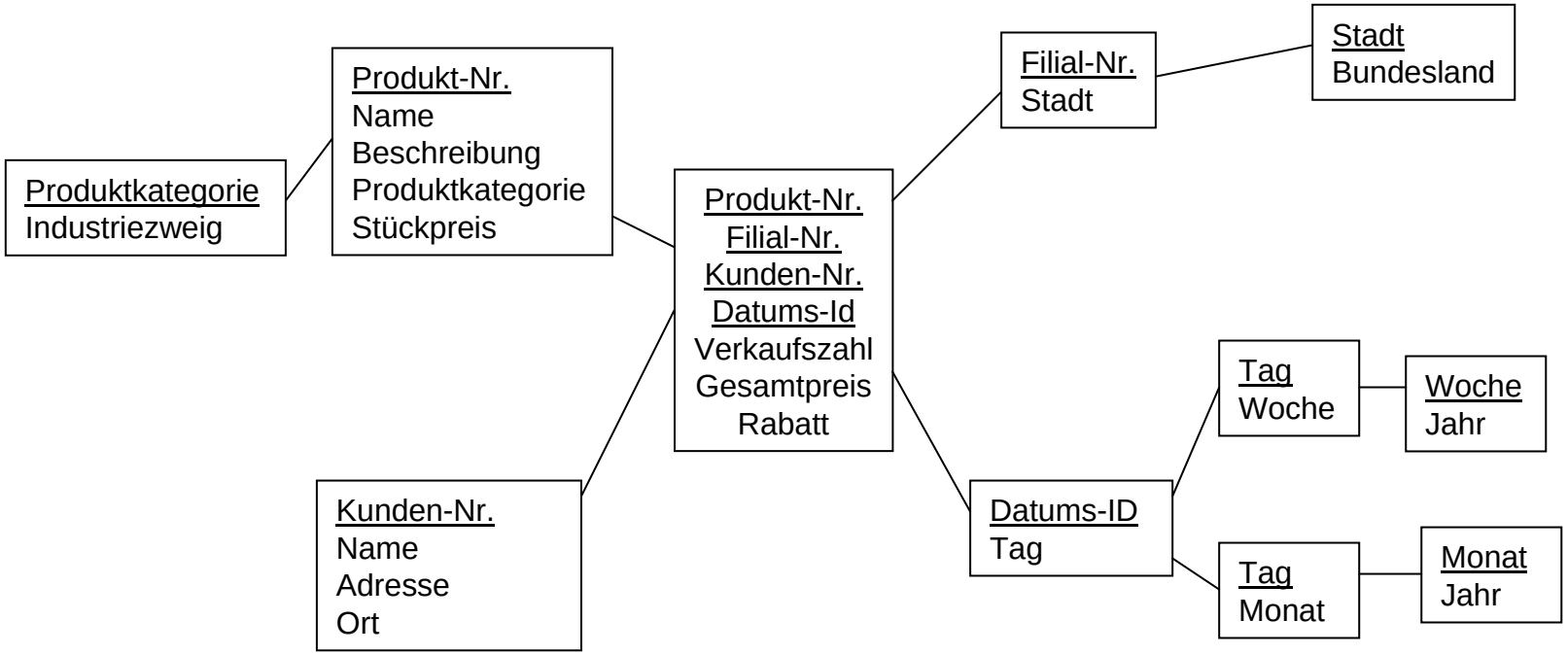
Beispiel: Tabellen abgeleitet aus einem Stern-Schema:



2.3.3 Mehrdimensionales Datenmodell

Schneeflocken-Schema:

- Stern-Schema repräsentiert die Attribut-Hierarchien in den Dimensionen nicht explizit.
- Explizite Hierarchie kann durch sog. **Schneeflocken-Schemata (Snowflake Schema)** erreicht werden.
- **Beispiel:** Schneeflocken-Schema



2.3.3 Mehrdimensionales Datenmodell

MOLAP: Multidimensional On-Line Analytical Processing

Spezifische Produkte für OLAP, die auf einer eigenen, proprietären mehrdimensionalen Datenbank beruhen.

Intern beruht die Datenbank auf einer Zell-Struktur, bei der jede Zelle entlang jeder Dimension identifiziert werden kann.

ROLAP: Relational On-Line Analytical Processing

Produkte, die eine multidimensionale Analyse auf einer relationalen Datenbank ermöglichen.

Sie speichern eine Menge von Beziehungen, die logisch einen mehrdimensionalen Würfel darstellen, aber physikalisch als relationale Daten abgelegt werden.

2.4 Preprocessing

2.4 Preprocessing

2.4.1 Introduction in preprocessing

Purpose of preprocessing

- transform datasets so that their information context is best exposed to the mining tool

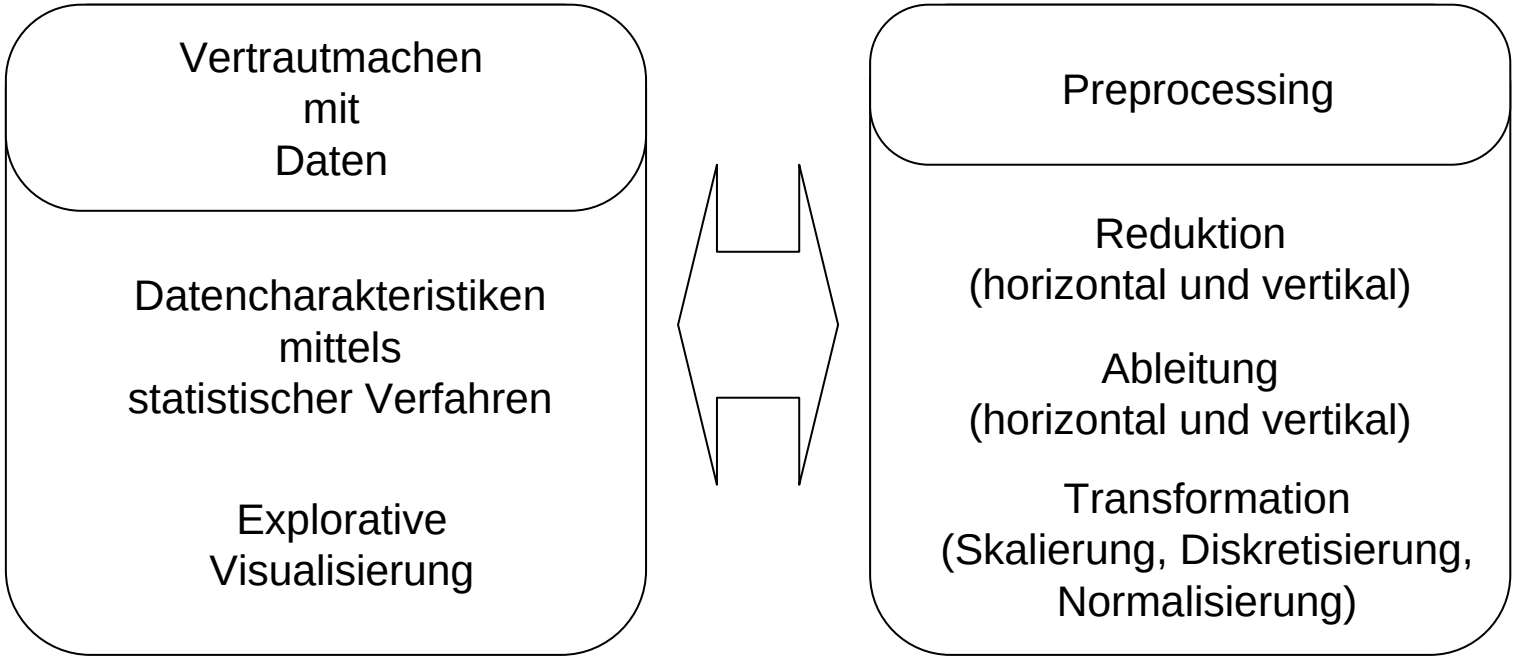
Problem: - learn the „true“**relationship**
- do not learn **noise**

2 Types of Preprocessing

prepare the data prepare the miner

2.4.1 Introduction in Preprocessing

Verknüpfung von Preprocessing und Datenverständnis



2.4.1 Introduction in Preprocessing

Weitere Aspekte des Preprocessing:

Gute Vorverarbeitung benötigt das Wissen eines Domänenexperten

Daten-Kontext

- Falsche Verteilung
- nominal vs. ordinal
- Korrelation

Domänen-Kontext

- richtige Daten im Kontext?
- Repräsentieren die Daten gesuchte Zusammenhänge?
- weitere Daten notwendig?

2.4.1 Introduction in Preprocessing

Beispiel:

Repräsentieren meine Daten die gesuchten Zusammenhänge aus Sicht der Domäne?

OLAP

- Visualisiert schnell die Zusammenhänge in den Daten
- Daten können schnell manipuliert werden

- Domänenexperte bekommt schnell einen Überblick über den aktuellen Stand der Daten und kann die gestellte Frage beantworten.

2.4.1 Introduction in Preprocessing

Telekom:

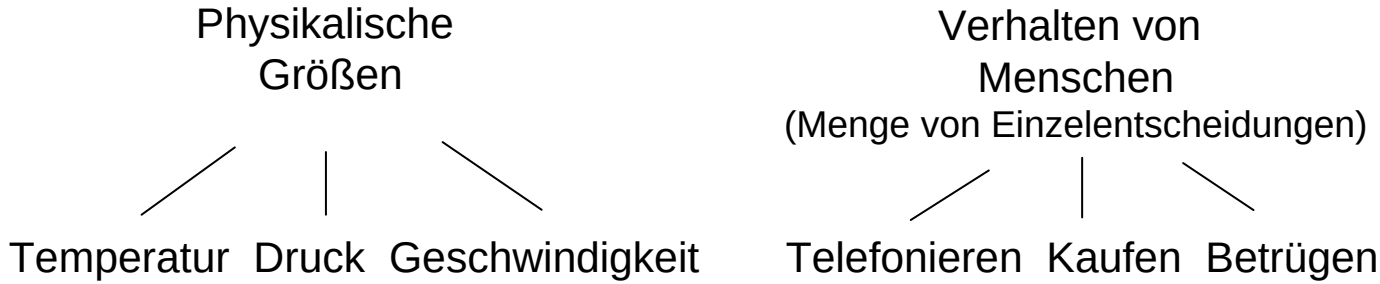
Bei der Deutschen Telekom AG sammelt man „nur“ über die eigenen Kunden schon länger Informationen über das Gesprächsverhalten. Für eine wieder einmal durchzuführende Analyse stellt sich die Frage:

Kann man mit den Daten aus dem Jahr 1997 eine allgemeingültige Aussage z.B. über das Telefonverhalten der Deutschen machen? Geht dies auch noch 1999?

- 1997 ja
- aber 1999 nicht mehr, da nicht mehr alle Telefonbesitzer auch Kunden der Deutschen Telekom AG sind.
- Außerdem ist zu beachten, daß auch 1997 mit den gesammelten Daten nur Aussagen über die Telefonbesitzer gemacht werden können, nicht aber über alle Deutschen.

2.4.1 Introduction in Preprocessing

Prinzipelle Unterschiede beim Preprocessing



Es ergeben sich unterschiedliche Probleme bei den Daten:

- komplexe Zusammenhänge
- meist nicht-linear
- konsistent
- häufig fehlende Werte (missing values)
- sehr große Datenmenge
- häufig inkonsistent

Beispiel:

Prozessoptimierung in einem Chemieunternehmen

Analyse des Verhaltens der Kunden einer Telefongesellschaft

2.4.2 Preprocessing Steps

2.4.2 Preprocessing Steps

Data Cleansing (Datenbereinigung)

- **consistency (Konsistenz)**
- **detail / aggregation level (Aggregationsniveau)**
- **pollution (Verunreinigung)**
- **relationship (Beziehungen)**
- **range (Definitionsbereich)**
- **defaults**
- **duplicate or redundant variables**
- **missing and empty values (fehlende Werte)**

Data Manipulation (Datenmanipulation)

- **reverse pivoting**
- **reducing dimensionality**
- **increasing dimensionality**
- **sparsity (schwach besetzte Werte)**
- **monotonicity (Monotonie der Daten)**
- **outliers (Ausreisser)**
- **numerating categorical values**
- **anachronisms**
- **relation between variable via pattern in the variable**
- **combinatorial explosion**

2.4.2 Preprocessing Steps

a) Data Cleansing

Consistency

- different things are represented by the same name in different systems
- same things are represented by different names in different systems

Detail / Aggregation level

- transaction record (detailed) vs. summarized transaction record (aggregated)
- general rule for data mining: detailed data is preferred to aggregated data
- level of detail in the input stream is one level of aggregation more detailed than the required level of detail in the output

2.4.2 Preprocessing Steps

Pollution

- garbage in the data, e.g. comma delimited data with comma in the data
- Human resistance, e.g. data fields are blank, incomplete, inaccurate

Relationship

- merging multiple input streams (use for example keys)
- find the right keys, eliminate double keys

Range

- variable has a particular domain, a range of permissible values
- detect outliers

Defaults

- the miner must know the default values of data capturing programs
- conditional default values can create seemingly significant patterns

2.4.2 Preprocessing Steps

Duplicate or redundant variables

- identical information in multiple variables, e.g. „date of birth“ and „age“
- problem for neural network with colinearity of variables

Missing and empty values

- empty values may not have a corresponding real-world value or have a real-world value but it was not captured
- miner should differentiate between both types of values
- data mining tools have different strategies to handle these values

2.4.2 Preprocessing Steps

b) Data Manipulation

Reverse pivoting

- modelling important things under the right point of view
- Example: Database with detailed call records
  Task is to analyse customers
  Problem: the focus of the database is not the customer

Reducing dimensionality

- eliminate features, which are not important for your task

Increasing dimensionality

- expand one dimension to represent the information in a better way
- example: Zip code can be transformed in “Lat” and “Lon”

2.4.2 Preprocessing Steps

Sparsity

- individual variables are only sparsely populated with instance values
- miner must decide e.g. to remove or to collapse the variable

Sample

- take only a part of the collected data (population)
- do not lose any information

Monotonicity

- a monotonic variable is one that increases without bound
- example variable: date, time, social number
- must be transformed in a non-monotonic form, since prediction can only be performed within the range of the learning data set

Outliers

- single or very low frequency occurrence of the value of a variable
- far away from the bulk of the values of the variable
- Is the outlier a mistake or very important information?

2.4.2 Preprocessing Steps

- **Anachronisms**

- something out of place in time
- information not actually available in the data when a prediction is needed

Relation between variable - pattern in the variable

- enough instance values to represent the variable's features are needed to detect patterns
- interactions between variables are interesting too

Combinatorial Explosion

- if you are interested in the interactions between variables you must check every combination of your variables

Number of variables	Number of combination
5	26
9	502
25	33.554.406

2.4.3 Preprocessing example for categorical data

How categorical values are best represented depends very much on the needs of the modelling tool.

Enumeration of categorical values

Time period	Rate of pay ($)
Half-day	100
Day	200
Half-week	500
Week	1000
Half-month	2000
Month	4000

Time period	...	Rate of pay ($)
Day	1	200
Half-day	2	100
Half-month	3	2000
Half-week	4	500
Month	5	4000
Week	6	1000

- both tables show the rate of pay for a period in dollars
- you don‘t see a structure if you order the time periods alphabetically (right table)

2.4.3 Preprocessing example for categorical data

Problem:

At worst (if the scale niveau is nominal) enumeration of categorical values introduces and creates patterns in the data that are not natural and that reflect throughout the data set, wreaking havoc with the modelling process.

- **Domain Knowledge can help**
- **do as little damage as possible to the natural structure**

Don‘t torture your data until they confess!

2.4.3 Preprocessing example for categorical data

Measure of distance for categorical data

Example: call detail record

customerID	distance	type of day	date/time	comm. minutes
1	Ort	Mo-Fr	19.11.98/9:55	20 min
1	Ort	Mo-Fr	20.11.98/10:10	18 min
2	Regional	Mo-Fr	19.11.98/21:00	120 min
2	Regional	Mo-Fr	20.11.98/17:00	2 min

Problem: Which records are similar to each other?

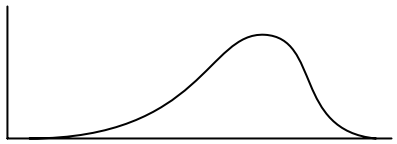
- the simple answer is: all are different
- a person would say: the first two records are similar, the last two not

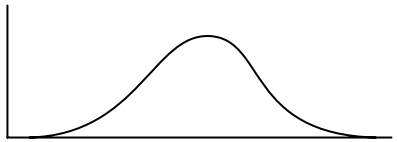
Most tools need a similarity measure to process the data automatically.
But always check that your measure is meaningful!

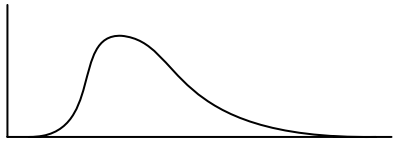
2.4.4 Preprocessing example for numerical data

Normalising a variable‘s Range with ladder of power (Tukey 1977)

p	Transformation T(x_i)	Name
...	...	...
10	x_i^{10}	dezimal
...	...	...
3	x_i^3	kubisch
2	x_i^2	quadratisch
1	x_i	**Rohdaten**
$\frac{1}{2}$	$\sqrt{x_i}$	Wurzel
0	$\log(x_i)$	logarithmisch
$-\frac{1}{2}$	$-\frac{1}{\sqrt{x_i}}$	reziproke Wurzel
1	$-\frac{1}{x_i}$	reziprok
...	...	...





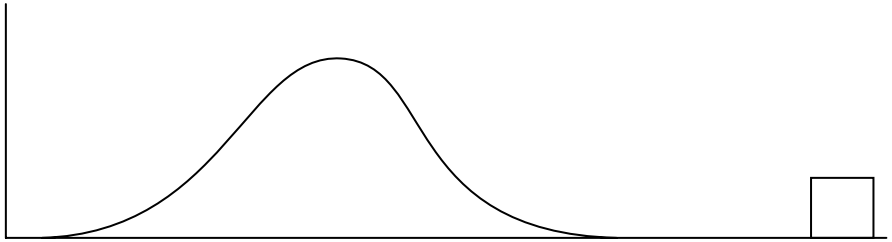


2.4.4 Preprocessing example for numerical data

Discretize numerical variable

take a range of values and map it to a new value
understand the underlying distribution

Select the right range



Why?

all variables have a particular resolution limit in practice

- accuracy of measurement
- precision of representation

if value is out of range - two different input values become the same

2.4.5 Preprocessing example for missing and empty values

Difference between missing and empty values

empty values have no corresponding real-world value
missing values have a real-world value but it was not captured

Tools can have difficulties in handling such values

ignore missing and empty values
use some metric to determine „suitable“ replacement
automated replacement techniques are critical

- Does the miner know the problems of the technique?
- Does the miner know the replacement method being used?
- What are its limitations?

Task for miner

replacement must be as **neutral** as possible
use a method understood and controlled by the miner

2.4.5 Preprocessing example for missing and empty values

Replacement: Problems and Aspects

some modelling techniques cannot deal with missing values

default replacement methods may introduce distortion

know and control the characteristics of any replacement method

important information is sometimes in the missing-value patterns

Example

the data had carefully been prepared for warehousing, including the replacement of the missing values

data warehouse data resulted in a remarkable poor quality of the learned model

quality was improved when the original source data was used

- most predictive variable was the missing-value pattern

1.3 Inhalt und Aufbau der Vorlesung

Aufbau der Vorlesung (1)

1. Einleitung
2. Grundlagen des KDD

   Statistik, Datenbanksysteme, OLAP, Preprocessing

Unüberwachte Verfahren

3. Clustering

   partitionierende und hierarchische Verfahren, Verfahren aus DBS-Sicht, neue Anforderungen und Techniken des Clustering

4. Assoziationsregeln

   einfache Assoziationsregeln, Algorithmus Apriori, Einbeziehung von Taxonomien, numerische Attribute