

# Data Warehousing und Data Mining

Das multidimensionale Datenmodell

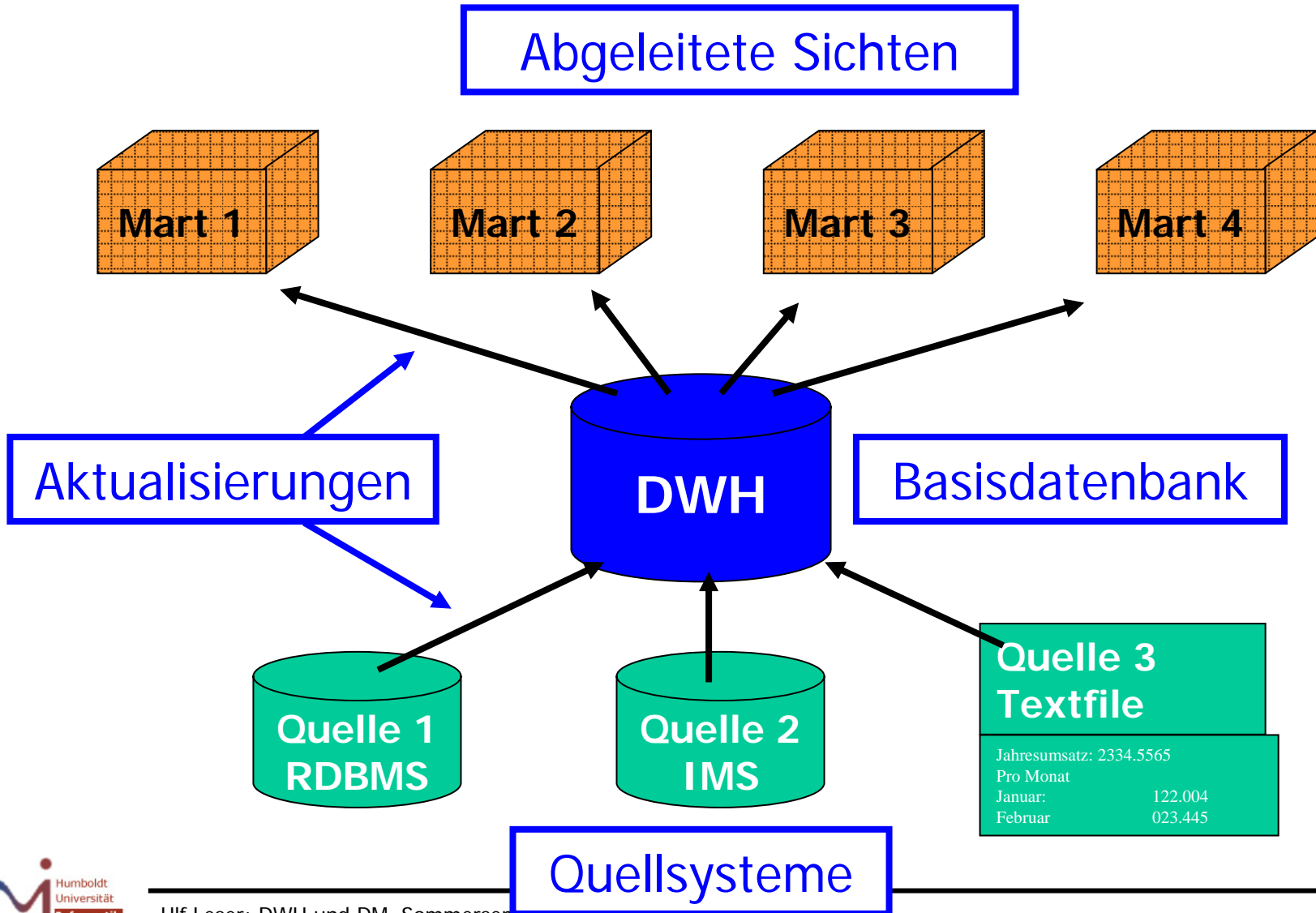


Ulf Leser

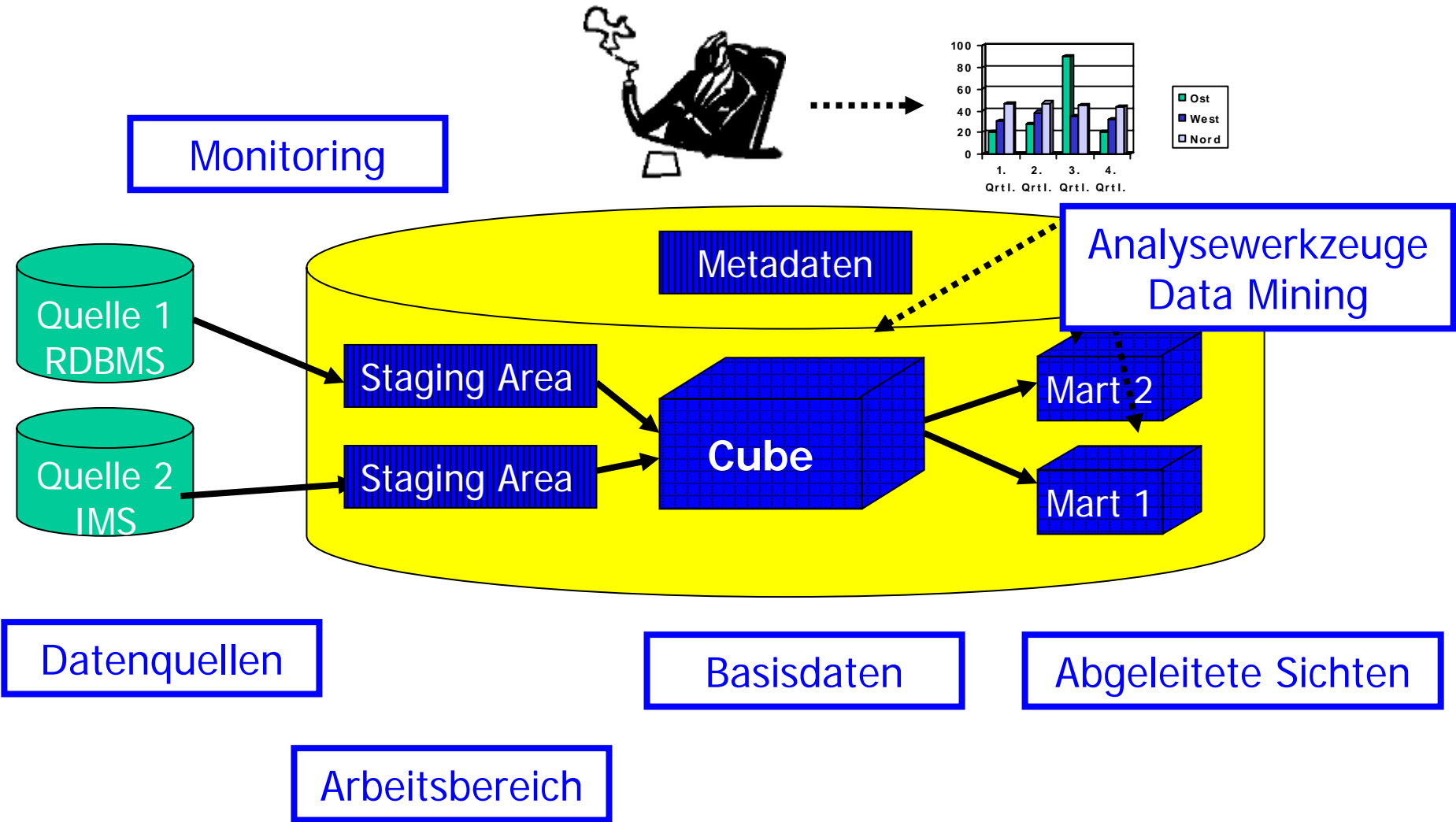
Wissensmanagement in der  
Bioinformatik



# DWH Grobarchitektur: Hubs and Spokes



# DWH Architektur & Komponenten



# Datenquellen

---

- Meist sehr heterogen
  - Technisch: RDBMS, IMS, Mainframe, Textfiles, ...
  - Logisch: Schema, Format, Repräsentation,...
  - Syntaktisch: Datum, Währung, Zahlenkodierung, ...
  - Verfügbarkeit: Kontinuierlich, Periodisch, ...
  - Qualität: Fehlende / falsche Werte, Duplikate, ...
  - Rechtlich: Datenschutz (Kunden & Mitarbeiter!)
- Zugriff
  - Push: Quelle erzeugt regelmäßig Extrakte
  - Pull: DWH stößt Zugriff an / Online-Zugriff
- Individuelle Behandlung notwendig

# Arbeitsbereich

---

- Temporärer Speicher
- Quellnahes Schema
- Sinn
  - ETL Arbeitsschritte effizienter implementierbar
    - Mengenoperationen, SQL
  - Zugriff auf Basisdatenbank möglich (Upsert)
  - Vergleich zwischen Datenquellen möglich
  - **Filterfunktion**: Nur einwandfreie Daten in Basisdatenbank übernehmen

# Basisdatenbank

---

- Zentrale Komponente des DWH
  - Begriff „DWH“ meint oft nur die Basisdatenbank
- Speichert Daten in **feinster Auflösung**
  - Einzelne Verkäufe
  - Einzelne Bons
- Historische Daten
- Große Datenmengen
  - Spezielle Modellierung
  - Spezielle Optimierungsstrategien

# DWH als ...

---

## Unterschiedliche Philosophien

- Enterprise DWH
  - Schemaintegration
- Analyseorientiertes DWH
  - Multidimensionale Modellierung

# Abgeleitete Sichten

---

- Analysten benötigt spezielle Daten
  - Aggregiert
    - Alle Verkäufe in Norddeutschland nach Lieferanten
    - Alle Verkäufe nach Niederlassung und Produkten
  - Ausgewählt
    - Alle Verkäufe in Niederlassung X
    - Alle Verkäufe von Lieferant Y
- Probleme bei Auswertung auf Cube
  - Wiederholte Durchforstung sehr großer Datenbestände notwendig
  - Hohe Detailstufe des Cubes für viele Anfragen nicht notwendig
- Vorab-Erstellung von abgeleiteten Daten
  - Data Marts
  - Prä-aggregierte, angereicherte und gefilterte Sichten

# 5. Datenanalyse

---

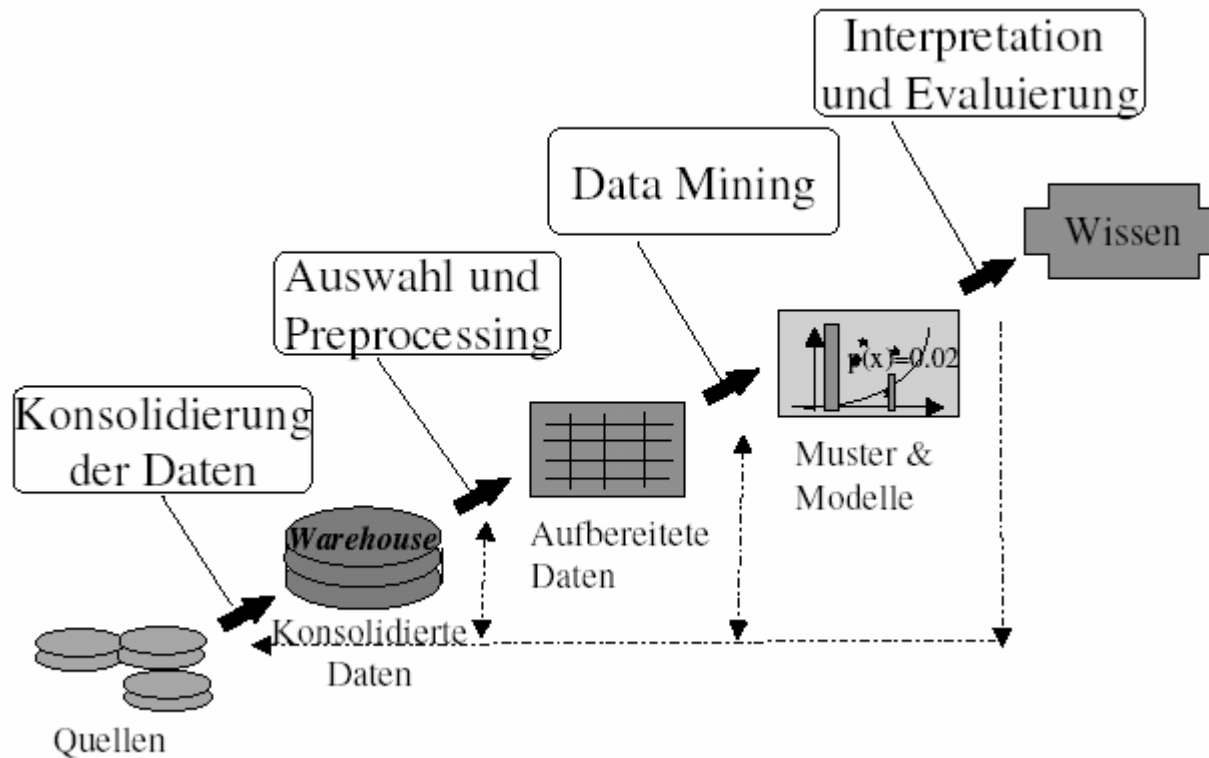
- „Einfache“ statistische Auswertung
  - OLAP Operationen
  - Reports
- OLAP Werkzeuge
  - Häufig proprietäre Systeme, eigene (geheime) Indexstrukturen
  - SAS, SPSS, Business Objects, Cognos, Excel, ...
- Funktionalität
  - Grafische Werkzeuge
  - Interaktive Datenauswahl, Filtering, Chaining, ...
  - Navigation, spez. im Cube
  - Präsentation: Grafiken, Tabellen, Reports, ...
- Die allermeisten Analysen sind **Standardreports**
  - Versus Ad-hoc Anfragen

# Data Mining

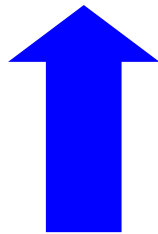
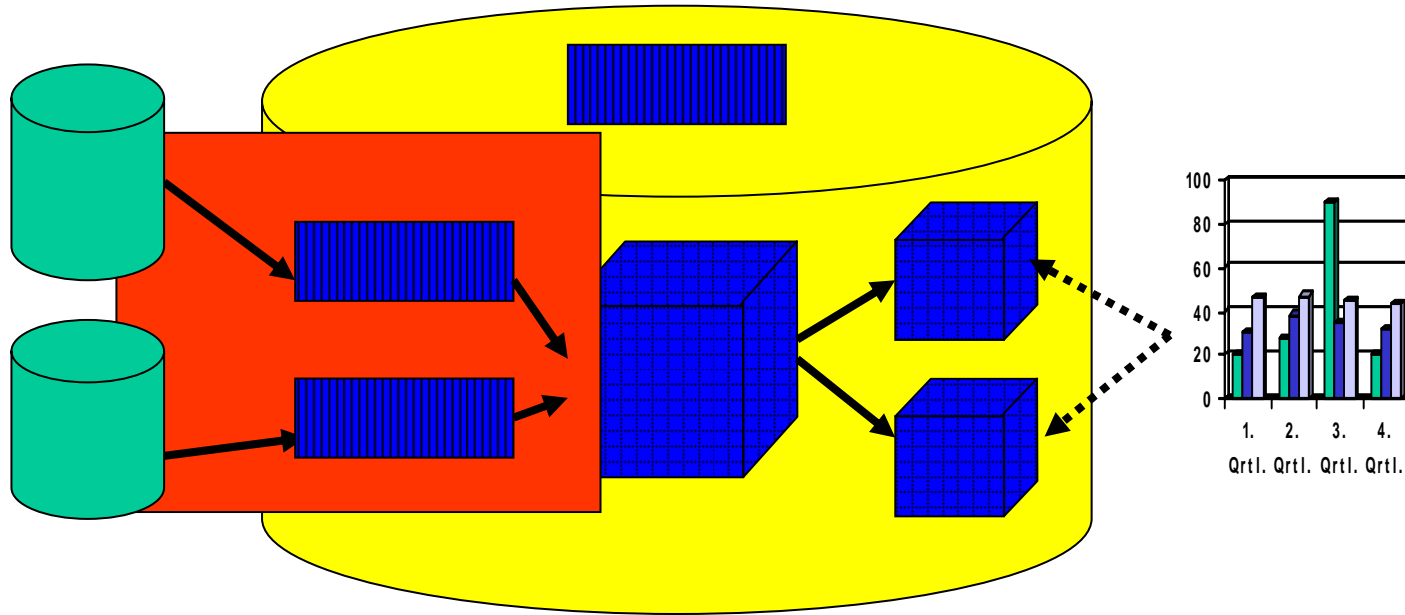
---

- „Finden verborgender, nicht-trivialer Informationen“
- Bereiche
  - Statistische Analyse
  - Maschinelle Lernverfahren
  - Knowledge Discovery in Databases (KDD)
- Suche nach Auffälligkeiten, Mustern, Regeln
  - Viele Kunden, die Windeln kaufen, kaufen auch Bier
- Suche nach Erklärungsmodellen
  - Modell: Abstraktion der Wirklichkeit
  - Korrelation versus Kausalität

# Data Mining Prozess



# 8. ETL



- Extraction
- Transformation
- Load

# ETL - Extraktion

---

- Aufgabe
  - Filtern der „richtigen“ Daten aus den Quellen
  - Bereitstellung der Datenfiles im gewünschten Format zum gewünschten Zeitpunkt am gewünschten Ort
  - Kontinuierliche Datenversorgung des DWH
- Prinzip: Producer - Consumer
  - Quelle informiert über Änderungen
  - DWH konsumiert Änderungen

# ETL - Transformation

---

- Aufgabe
  - Umwandlung der Daten in eine „DWH-gerechte“ Form
- **Form follows Function**
  - Quellen: hoher Transaktionsdurchsatz
  - DWH: spezifische statistische Analysen
- Arten von Transformationen
  - Schematransformationen
  - Datentransformationen
- Transformationen möglich **an zwei Stellen**
  - Transformation der Quell-Extrakte in Load-Files
  - Transformation von Staging-Area nach Basis-DB

# Beispiel

## Handelshaus, Daten einer Woche, 1 Filiale

Laden mit voller Qualitätskontrolle	10 min
Laden mit partieller Datenverbesserung	2 min
Nur Laden	45 sec

## Handelshaus, Daten einer Woche, 2000 Filiale

Laden mit voller Qualitätskontrolle	330h = 14d
Laden mit partieller Datenverbesserung	67 h = 2,8d
Nur Laden	25h = 1d

# Inhalt dieser Vorlesung

---

- Multidimensionales Datenmodell (MDDM)
- Klassifikationsschema, –stufen, -hierarchie
- Dimensionen und Granularität
- Beispiel

# MDDM Grundidee

---

- Unterscheidung von
  - **Fakten** (Measures) – Gemessene Werte
  - **Dimensionen** – Beschreibung der Messwerte in Raum, Zeit, Organisation, ...
  - **Klassifikationshierarchien** – Dimensionen haben hierarchische Struktur
- Metapher: Würfel (Cube) bzw. Hypercube
  - Fakten: **Punkte im multidimensionalen Raum**
  - Klassifikationshierarchien: Achsenbeschriftung / Koordinaten in unterschiedlichem Verfeinerungsgrad
- Analyse durch **Operationen auf dem Cube**
  - Dimensionen ausblenden / einblenden
  - Auswahl von Subwürfeln (Flächen, Punkten, ...)
  - Hierarchiestufe vergrößern/verfeinern

# Beispiel

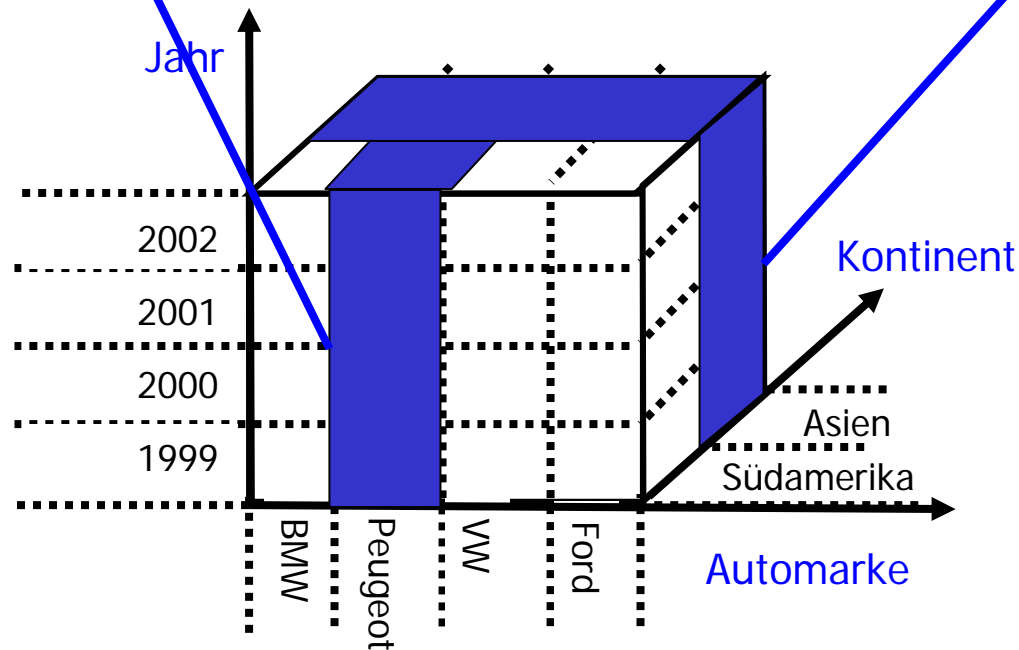
---

- Verkäufe von Autos pro Marke, Kontinent und Jahr gemessen in Euro
  - Fakten
    - Verkäufe in Euro
  - Dimensionen
    - Automarke
    - Kontinent
    - Jahr

# Beispiel: Auswahl (Slicing)

Verkäufe von Peugeot  
pro Jahr und Kontinent

Verkäufe in Asien  
pro Jahr und Marke

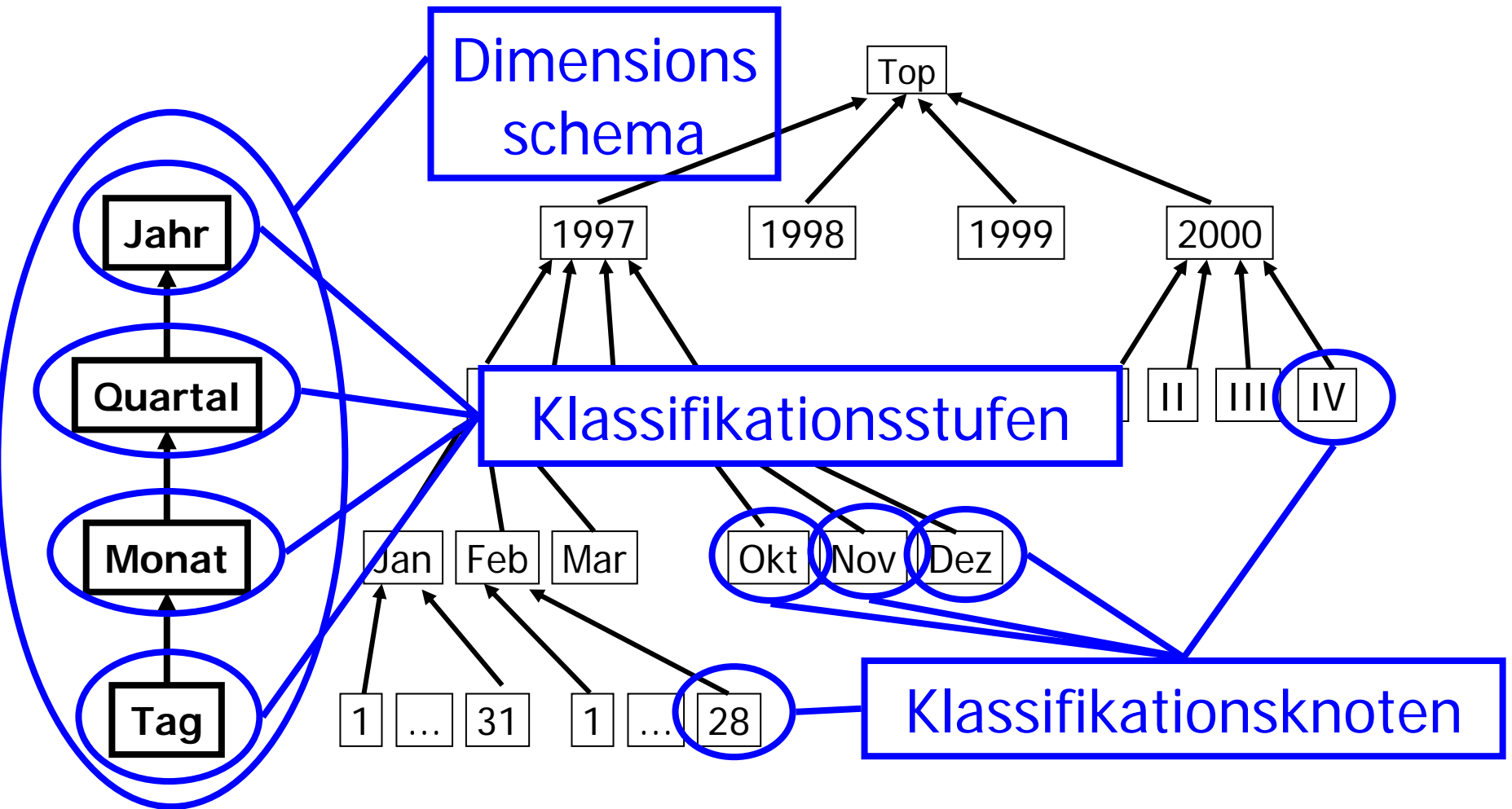


# Dimensionen

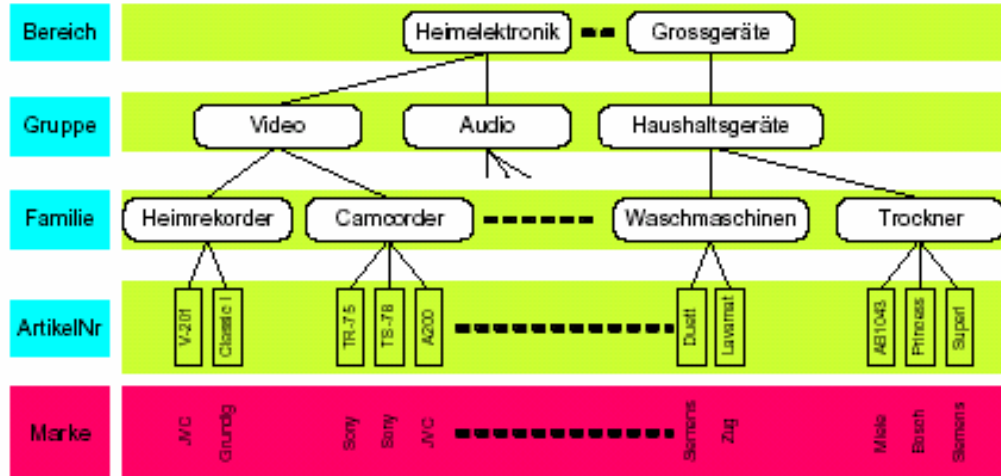
---

- Eindeutige Strukturierung des Datenraums
- Hoffentlich orthogonal
  - **Abhängigkeiten zwischen Dimensionen** bereiten an vielen Stellen Probleme – später
- Jede Dimension hat ein **Schema**
  - Zeit: Tag, Woche, Jahr, ...
    - Achtung: Es sind individuelle Tage, Monate, ... gemeint. Als nicht „1 = der erste jedes Monats“, sondern „1 = 1.1.2006“ etc.
  - Region: Landkreis, Land, Staat, ...
  - Produkt: Produktgruppe, Produktklasse, Produktfamilie, ...
- ... und **Werte**
  - (1/1/99, 2/1/99, ..., 31/1/99,...31/12/07), (1/99, ... 52/07), (1900, ..., 2003)
  - (...), (Berlin, NRW, Department-1, ...), (BRD, F, ...)

# Dimension



# Produktthierarchie

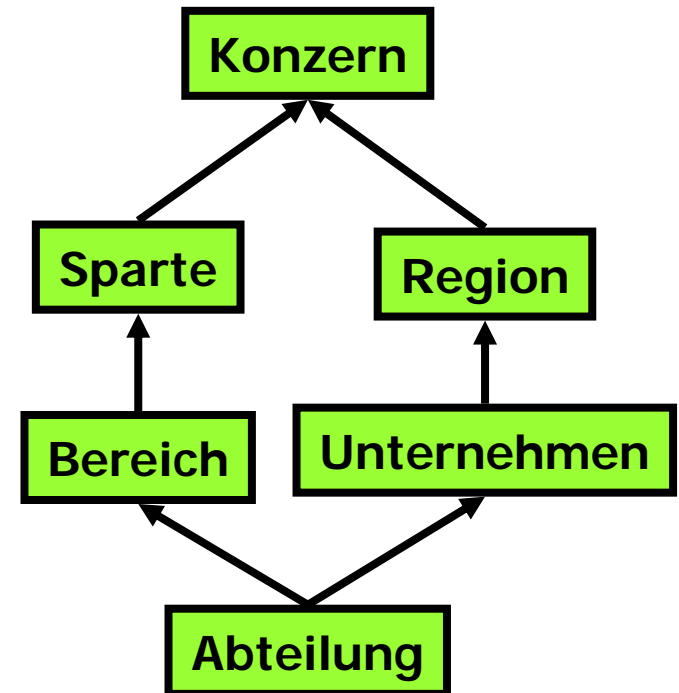


Aus: Geppert, ETZ Zürich, Vorlesung „Data Warehouse“

- Elemente einer Stufe können **geordnet** sein
  - Geordnet: Zeit
  - Ungeordnet: Produkte

# Formale Definition

- Ziel
  - Operationen auf einem MDDM exakt definieren
  - Aus dem Modell muss man ersehen können, **welche Verdichtungen semantisch sinnvoll sind** und welche nicht
    - Tools bieten dann nur die sinnvollen Operationen an
  - Optimierer können die Informationen für schnellere Anfragen benutzen
  - Multidimensionale Modelle grafisch spezifizieren
    - Mit E/R nicht erreichbar



# Klassifikationsschema

---

- Definition

Ein *Klassifikationsschema*  $K$  (einer Dimension  $D$ ) ist ein Quadrupel  $(K_s, \rightarrow_s, K_k, \rightarrow_k)$  mit

- $K_s$  ist die Menge von *Klassifikationsstufen*  $\{k_0, \dots, k_n\}$
- „ $\rightarrow_s$ “ ist eine Halbordnung auf  $K_s$  mit größtem Element  $\text{top}(K_s)$ 
  - D.h.:  $\forall k \in K_s: k \rightarrow_s \text{top}(K_s)$
- $K_k$  ist die Menge von *Klassifikationsknoten*  $\{n_0, \dots, n_m\}$
- Jeder Klassifikationsknoten  $n$  ist genau einer Klassifikationsstufe  $k$  zugeordnet.
  - $\text{stufe}(n) = k$
  - $\text{knoten}(k) = \{n \mid n \in K_k \wedge \text{stufe}(n) = k\}$
- „ $\rightarrow_k$ “ ist die Halbordnung auf  $K_s$  übertragen auf  $K_k$ 
  - Sei  $k, l \in K_k$ . Es gilt  $k \rightarrow_s l$  gdw.  $\text{stufe}(k) \rightarrow_s \text{stufe}(l)$

- Bemerkung

- Eine Klassifikationsstufe hat mehrere Klassifikationsknoten, aber jeder Klassifikationsknoten ist genau einer Klassifikationsstufe zugeordnet
- Wir benutzen i.d.R. einfach  $\rightarrow$  für  $\rightarrow_k$  oder  $\rightarrow_s$

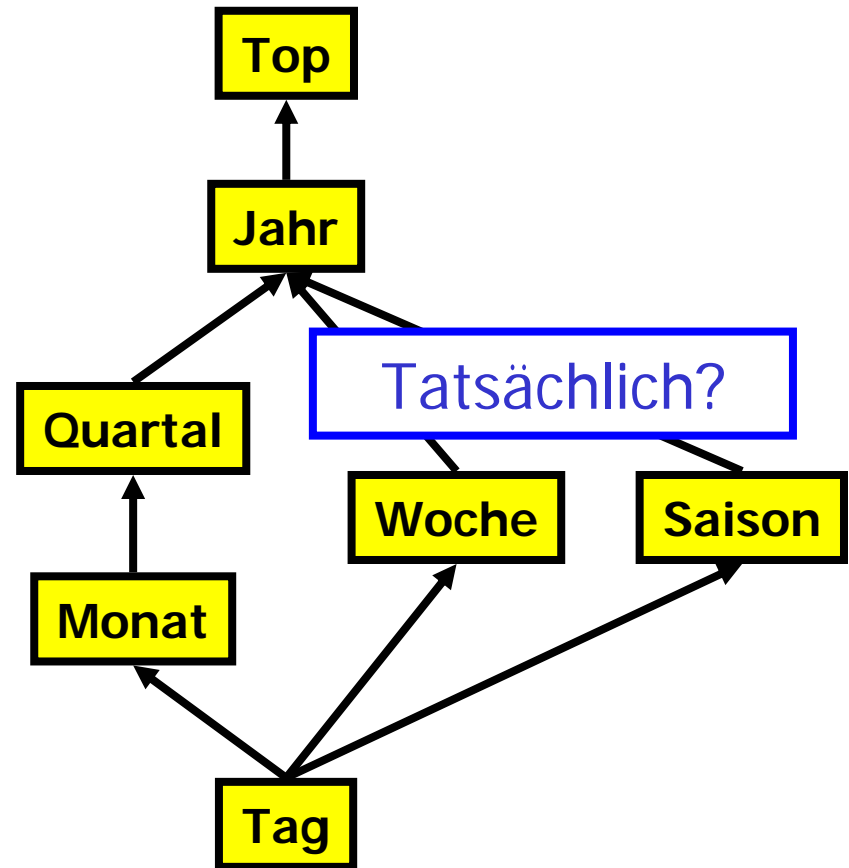
# Erläuterung

---

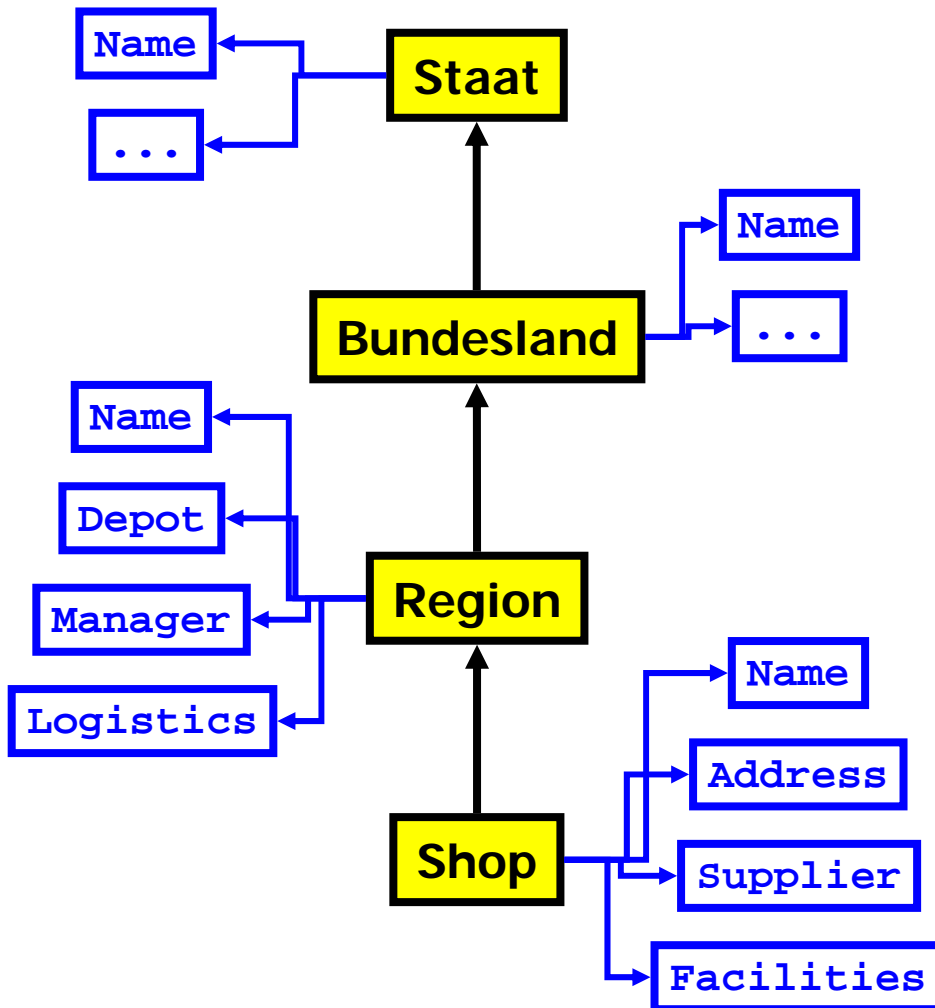
- Die Klassifikationsstufen sind die **Schemaelemente** der Dimension
- Klassifikationsknoten sind die Instanzen der Schemaelemente
- Das größte Element der Stufenhierarchie ist meistens artifiziell – es steht für „alles“
  - Wir nennen es TOP
- Interpretation von „→“
  - **Funktionale Abhängigkeit**
  - (Später: Aggregierbarkeit)
  - Tag bestimmt Monat bestimmt Jahr bestimmt TOP
    - 21.12.2003 → 12.2003 → 2003 → TOP
  - Produkt → Produktfamilie → Produktgruppe → TOP
    - “Asus M2400N” → Notebooks → Büroelektronik → TOP
- Beachte: Halbordnung ist immer **zyklusfrei**

# Beispiel

- Ordnung
  - Tag → Monat
  - Monat → Quartal
  - Quartal → Jahr
  - Tag → Woche
  - Woche → Jahr
  - Alle → Top
- Keine Ordnung
  - Quartal ? Woche
  - Monat ? Woche
- Transitivität
  - Tag → Jahr



# Knotenattribute



- Jede Klassifikationsstufe hat eine Menge von Attributen, die **Knotenattribute**
  - Teil des Schemas des Klassifikationsschemas
- Jeder Klassifikationsknoten hat Werte für seine Knotenattribute

# Klassifikationspfade

---

- Definition

Ein *Klassifikationspfad*  $P$  in einem Klassifikationsschema  $K$  mit Klassifikationsstufen  $K_s$  ist eine Menge  $\{p_0, \dots, p_m\}$  mit

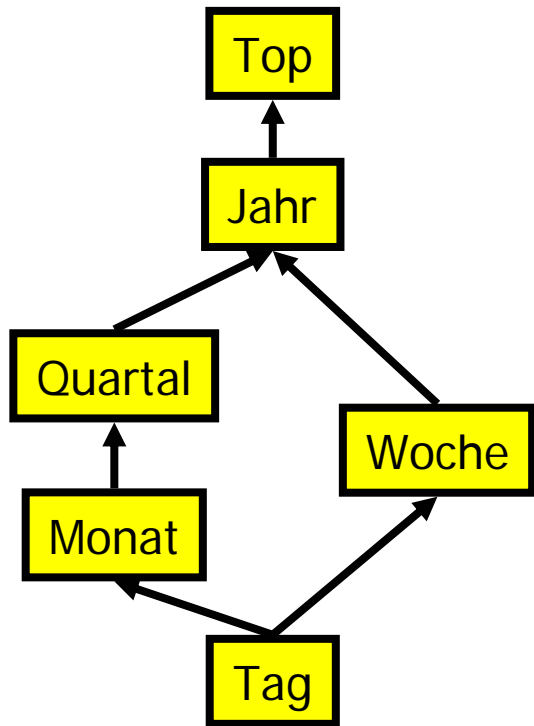
- $\{p_0, \dots, p_m\} \subseteq K_s$
- $p_m = \text{top}(K_s)$
- $\forall p_i, 1 \leq i \leq m: p_{i-1} \rightarrow p_i$  und  $\nexists q: p_{i-1} \rightarrow q \rightarrow p_i$
- Die Länge des Pfades  $P$  ist  $|P|=m+1$
- Der *Klassifikationslevel* von  $p_i$  in  $P$  ist  $i$

- Bedeutung

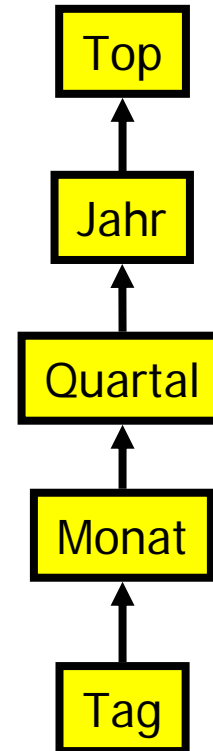
- Ein Pfad ist eine zusammenhängende und vollständig geordnete Teilmenge von  $K_s$
- Jeder Pfad beinhaltet das größte Element TOP
- *Verdichtung* werden wir später entlang von Klassifikationspfaden definieren
  - Und damit entlang funktionaler Abhängigkeiten

# Beispielpfade

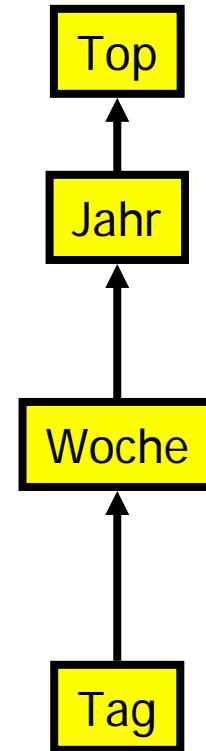
Klassifikationsschema



Pfad 1

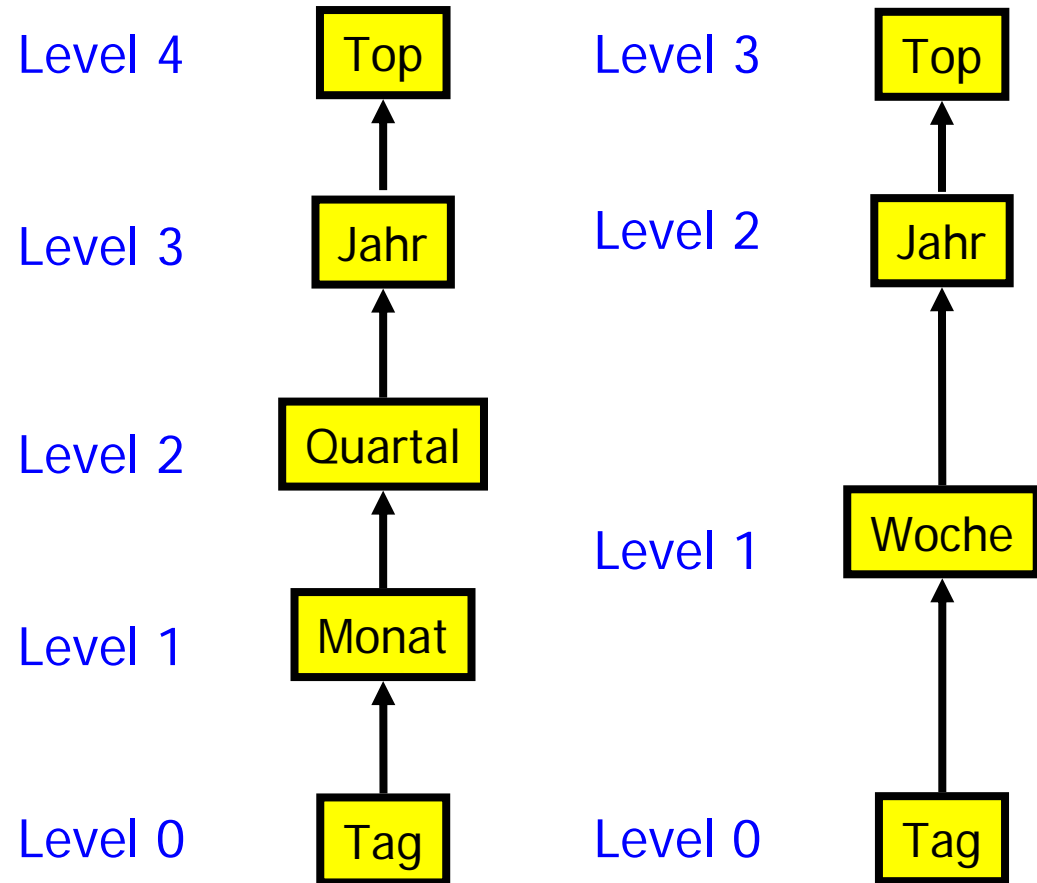


Pfad 2



# Klassifikationsstufen und Pfade

- Der Klassifikationslevel einer Stufe ist nur eindeutig in einem Pfad
- Der Level des größten Elements TOP ist nicht konstant



# Klassifikationshierarchie

---

- Definition

- Die *Klassifikationshierarchie*  $H$  zu einem Klassifikationsschema  $K$  mit Pfad  $P$  ist der Baum mit Knoten  $N$  und Kanten  $E$  wie folgt:

$$N = \bigcup_{p_i \in P} \text{knoten}(p_i)$$

$$E = \left\{ (n, m) \left| \begin{array}{l} n, m \in N \wedge n \rightarrow m \wedge \\ \exists j : n \in \text{knoten}(p_j) \wedge m \in \text{knoten}(p_{j+1}) \end{array} \right. \right\}$$

- Bemerkungen

- Klassifikationshierarchie = Knotenhierarchie in einem Pfad
- Jede Klassifikationshierarchie ist **balanciert**: Alle Pfade Wurzel-Blatt haben die Länge  $|P|$

# Inhalt dieser Vorlesung

---

- Multidimensionales Datenmodell (MDDM)
- Klassifikationsschema, –stufen, -hierarchie
- Dimensionen und Granularität
- Beispiel

# Dimension

---

- Definition

Eine *Dimension*  $D=(K, \{P_1, \dots, P_j\})$  besteht aus

- Einem Klassifikationsschema  $K$
- Einer Menge von Pfaden  $P_i$  in  $K$

- Bemerkungen

- $D$  muss nicht alle Pfade enthalten, die es in  $K$  gibt
  - Designentscheidung
- Theoretisch müssen nicht alle Klassifikationsstufen von  $K$  in einem Pfad enthalten sein
  - Aber man wird seine Pfade so wählen, dass dies doch gilt

- Schreibweise

- $D.k$  bezeichnet die Klassifikationsstufe  $k$  aus  $D$
- Ein  $D.k$  kann in mehreren Pfaden vorkommen

# Granularität

---

- Definition

*Gegeben eine Menge  $U$  von Dimensionen  $D_1, \dots, D_n$ .  
Eine **Granularität  $G$  über  $U$**  ist eine Menge  $\{D_1.k_1, \dots, D_n.k_n\}$  für die gilt*

- *$k_i$  ist eine Klassifikationsstufe in  $D_i$*
- *Es gibt **keine funktionalen Abhängigkeiten** zwischen den Klassifikationsstufen  $D_1.k_1, \dots, D_n.k_n$*

- Bemerkungen

- Zweite Bedingung ist immer erfüllt, wenn keine funktionalen Abhängigkeiten zwischen Dimensionen bestehen
  - Beispiel: Nicht gleichzeitig Dimensionen Zeit und „Fiskalisches Jahr“ in einer Granularität betrachten

# Erläuterung

---

- Mit einer Granularität legt man fest, in welcher **Detailstufe** Daten beschrieben werden
  - Festlegung für jede Dimension
  - Eine Granularität ist damit ein Hyperwürfel in einer bestimmten Auflösung
- **OLAP-Operationen** manipulieren Granularitäten
  - Insbesondere Navigation entlang von Pfaden
  - Herausschneiden von Dimensionen (= Granularität auf TOP)
- Durch die Anordnung der Stufen in Pfaden sind Granularitäten **halb-geordnet**

# Halbordnung auf Granularitäten

---

- Definition

*Auf der Menge aller Granularitäten zu einer Menge  $U$  von Dimensionen ist eine **Halbordnung** „ $\leq$ “ wie folgt definiert*

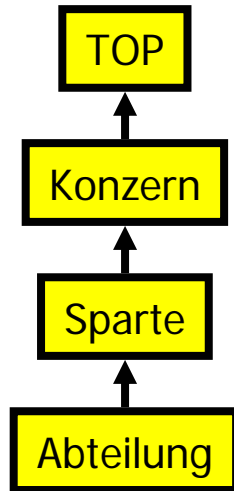
- Sei  $G_1 = \{D_1.k_1^1, \dots, D_n.k_n^1\}$  und  $G_2 = \{D_1.k_1^2, \dots, D_n.k_n^2\}$
- Es gilt  $G_1 \leq G_2$  genau dann wenn
  - $\forall i: D_i.k_i^1 \rightarrow D_i.k_i^2$
  - (Dies impliziert als Möglichkeit auch  $D_i.k_i^1 = D_i.k_i^2$ )

- Benutzung

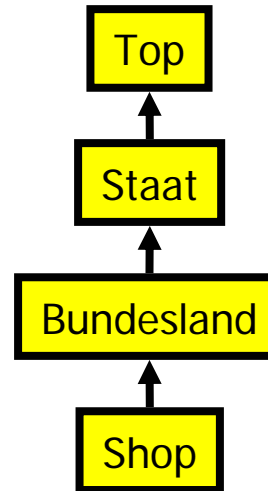
- Beschreibung der Transformation von Granularitäten
- Anfrageoptimierung: **Wiederverwendung von Aggregaten**

# Beispiel

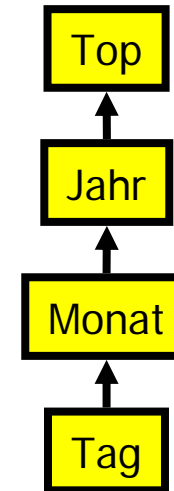
## Bereich



## Region



## Zeit



$(B.Sparte, R.Shop, Z.Tag) \leq (B.Sparte, R.Shop, Z.Monat) \leq (B.Sparte, R.Top, Z.Monat) \leq (B.Top, R.Top, Z.Top)$

$(B.Sparte, R.Staat, Z.Tag) ? (B.Konzern, R.Shop, Z.Tag)$

# Würfelschema und Würfel

---

- Definition

Ein *Würfelschema*  $WS$  ist ein Tupel  $(G, F)$  mit

- Einer Granularität  $G$
- Einer Menge  $F$  von Fakten mit  $|F|=m$

- Ein *Würfel*  $W$  ist eine Instanz eines Würfelschema  $(G, F)$

$$W = \text{dom}(G) \times \text{dom}(F)$$

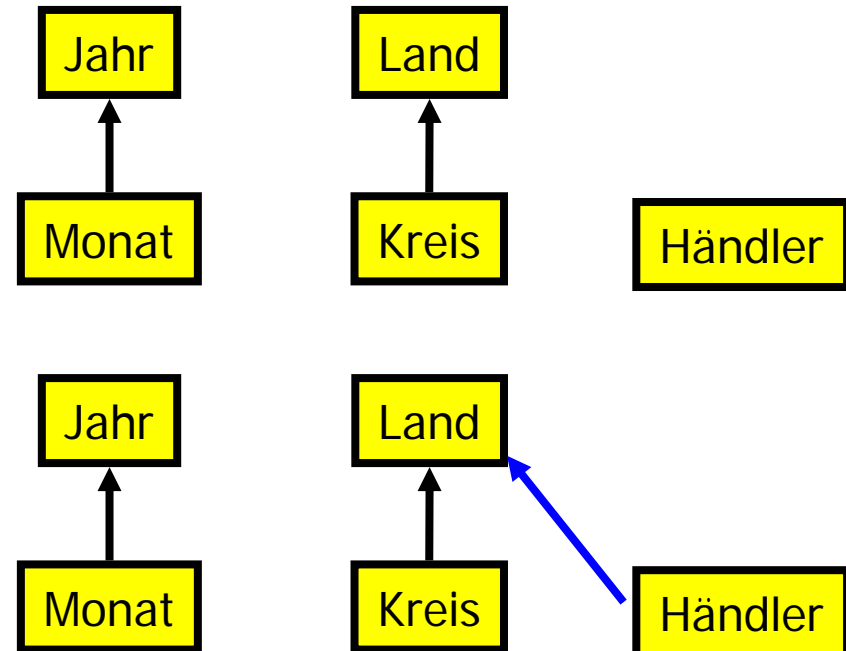
$$= \text{knoten}(D_1.k_1) \times \dots \times \text{knoten}(D_n.k_n) \times \text{dom}(F_1) \times \dots \times \text{dom}(F_m)$$

- Bemerkung

- Die Werte  $\text{dom}(G)$  geben die **Koordinaten** der Werte  $\text{dom}(F)$  an
- Verhältnis Würfelschema zu Würfel ist wie Relationenschema zu Relation

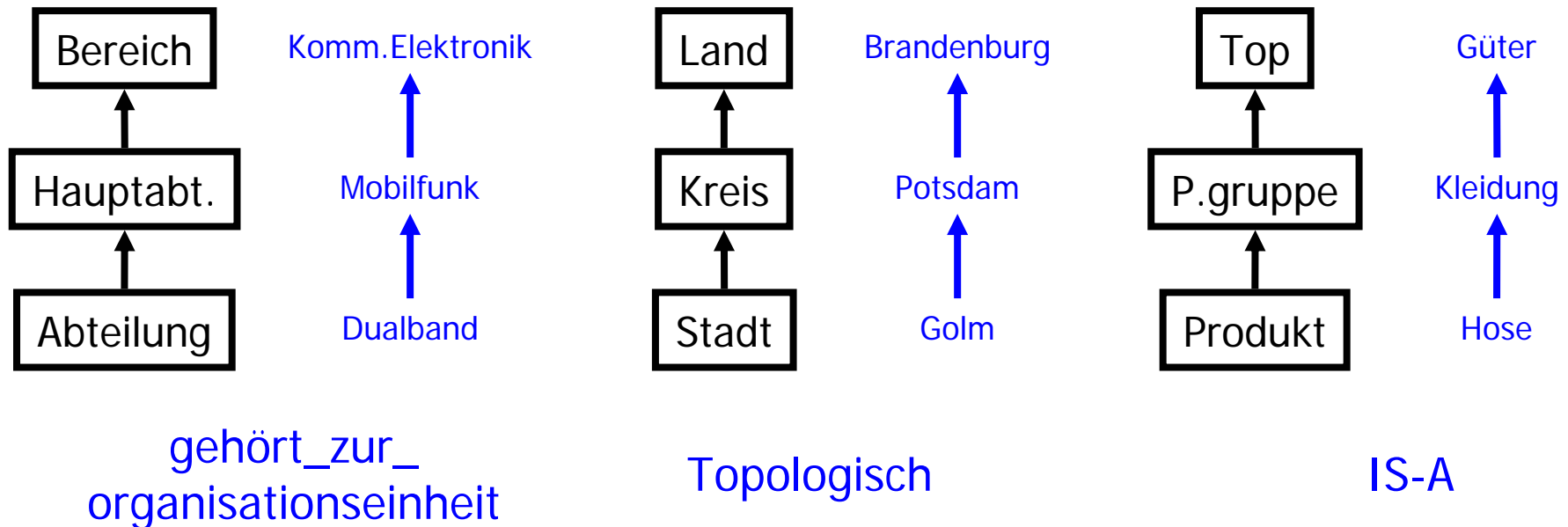
# Kein Würfelschema

- Autoverkäufe pro Zeit (Monat, Jahr), Händler und Region (Kreis, Land)
- Drei Dimensionen
  - Monat → Jahr
  - Händler
  - Kreis → Land
- Aber: EU Recht !
  - Händler → Land
  - Damit können wir keine Granularität bauen



# Semantik von Kanten

- Die Hierarchie von Klassifikationsstufen wird durch funktionale Abhängigkeiten bestimmt
- Das beinhaltet keine Bestimmung der **Semantik der Kanten**



# Inhalt dieser Vorlesung

---

- Multidimensionales Datenmodell (MDDM)
- Klassifikationsschema, –stufen, -hierarchie
- Dimensionen und Granularität
- Beispiel

# Ein längeres Beispiel

---

- Wir bauen ein DWH zur Verwaltung von Lagerbeständen
- Wir haben viele Lagerhäuser (international verteilt)
- Eine „Messung“ ist der Zugang oder Abgang von Produkten
- Jede Messung erzeugt **zwei Fakten**
  - Bestand und Delta von Artikeln
- **Klassifikationsschema K**
  - Zeit
    - Klassifikationsstufen: Monat, Quartal, Woche, Jahr
  - Ort
    - Klassifikationsstufen: Region, Land
  - Produkt
    - Klassifikationsstufen: Artikel, Artikelgruppe, Bereich

# Klassifikationsschema

---

- Halbordnung
  - Top ← Jahr
  - Jahr ← Quartal
  - Quartal ← Monat
  - Jahr ← Woche
  - Top ← Land
  - Land ← Region
  - Top ← Bereich
  - Bereich ← Artikelgruppe
  - Artikelgruppe ← Artikel
- Struktur der Dimensionen

# Pfade

---

- $P_1$ : Top  $\leftarrow$  Jahr  $\leftarrow$  Quartal  $\leftarrow$  Monat
- $P_2$ : Top  $\leftarrow$  Jahr  $\leftarrow$  Woche
- $P_3$ : Top  $\leftarrow$  Land  $\leftarrow$  Region
- $P_4$ : Top  $\leftarrow$  Bereich  $\leftarrow$  Artikelgruppe  $\leftarrow$  Artikel
  
- Entlang der Pfade sind Verdichtungen im Modell sinnvoll

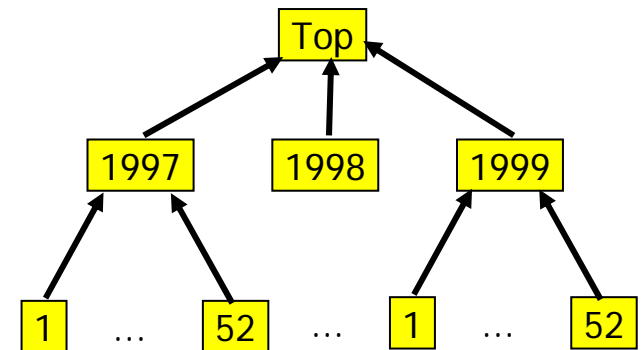
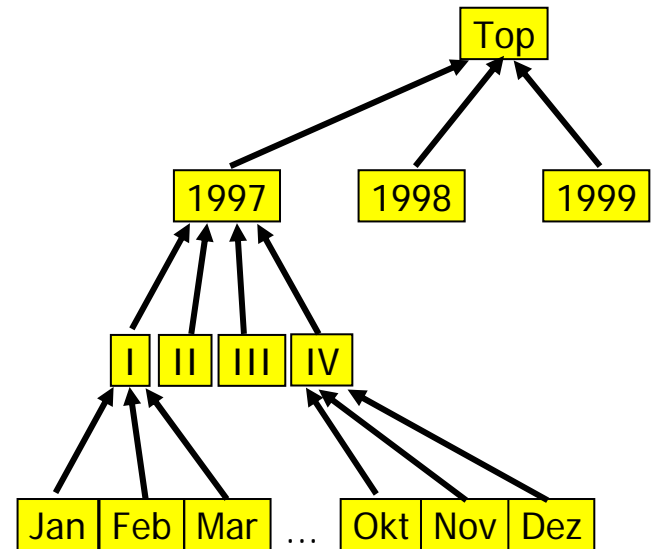
# Klassifikationsknoten

---

- Jahr
  - 1997, 1998, 1999
- Quartal
  - I, II, III, IV (pro Jahr)
- Woche
  - 1-52 (pro Jahr)
- Monate
  - 1-3 (pro Quartal I), 4-6 (pro Quartal II), ...
- Land
  - Deutschland, Frankreich, Großbritannien, ...
- Region
  - Bayern, Berlin, ..., Departament1, Departament2, ...
- Bereich
  - Kleidung, Nahrung, Elektronik, ...
- Artikelgruppe
  - Oberbekleidung, Unterbekleidung, Spirituosen, Kindernahrung, Kleingeräte, TV/Video, ...
- Artikel
  - ...
- **Alle möglichen Ausprägungen der Klassifikationsstufen**

# Klassifikationshierarchien 1

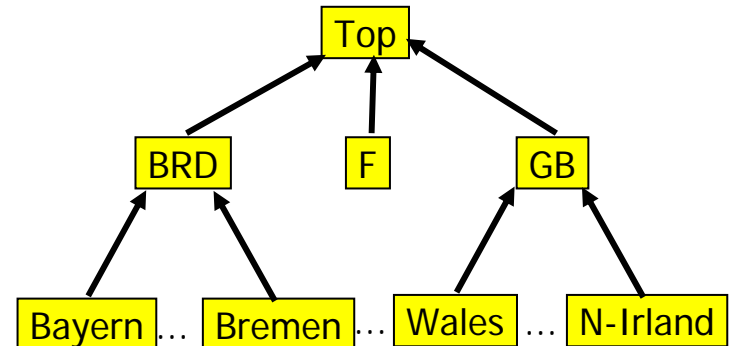
- Klassifikationshierarchie zu  $P_1$ 
  - $P_1$ : Top  $\leftarrow$  Jahr  $\leftarrow$  Quartal  $\leftarrow$  Monat
- Klassifikationshierarchie zu  $P_2$ 
  - $P_2$ : Top  $\leftarrow$  Jahr  $\leftarrow$  Woche



# Klassifikationshierarchien 2

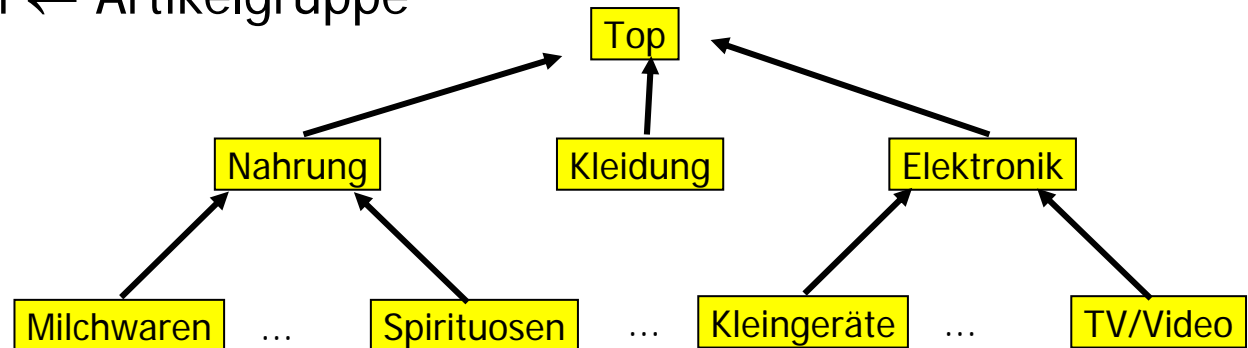
- Klassifikationshierarchie  $P_3$

- $P_3$ : Top  $\leftarrow$  Land  $\leftarrow$  Region



- Klassifikationshierarchie  $P_4$

- $P_4$ : Top  $\leftarrow$  Bereich  $\leftarrow$  Artikelgruppe  $\leftarrow$  Artikel



# Dimensionen

---

- Dimension ZEIT
  - ({Monat, Quartal, Woche, Jahr}, {P<sub>1</sub>, P<sub>2</sub>})
- Dimension ORT
  - ({Region, Land}, {P<sub>3</sub>})
- Dimension PRODUKT
  - ({Artikel, Artikelgruppe, Bereich}, {P<sub>4</sub>})
- Dimensionen enthalten mehrere Pfade

# Granularität, Würfel

---

- Mögliche Granularitäten

- $G_1 = (\text{Zeit.Woche}, \text{Ort.Land}, \text{Produkt.Artikel})$
- $G_2 = (\text{Zeit.Jahr}, \text{Ort.Gebiet}, \text{Produkt.TOP})$
- Halbordnung:
  - $(\text{Zeit.Woche}, \text{Ort.Gebiet}, \text{Produkt.Artikel})$
  - $\leq (\text{Zeit.Jahr}, \text{Ort.Gebiet}, \text{Produkt.Bereich})$
  - $\leq (\text{Zeit.Jahr}, \text{Ort.Top}, \text{Produkt.Bereich})$
  - $\leq (\text{Zeit.ZOP}, \text{Ort.Top}, \text{Produkt.Top})$

- Würfelschema

- Granularität plus Menge von Fakten ( $F_1 = \text{Bestand}$ ,  $F_2 = \text{Delta}$ )

- Würfel: Instanz des Würfelschemas

- Operationen auf Würfeln verändern die Granularität
- Ziel: Nur sinnvolle Operationen zulassen