Data Warehousing

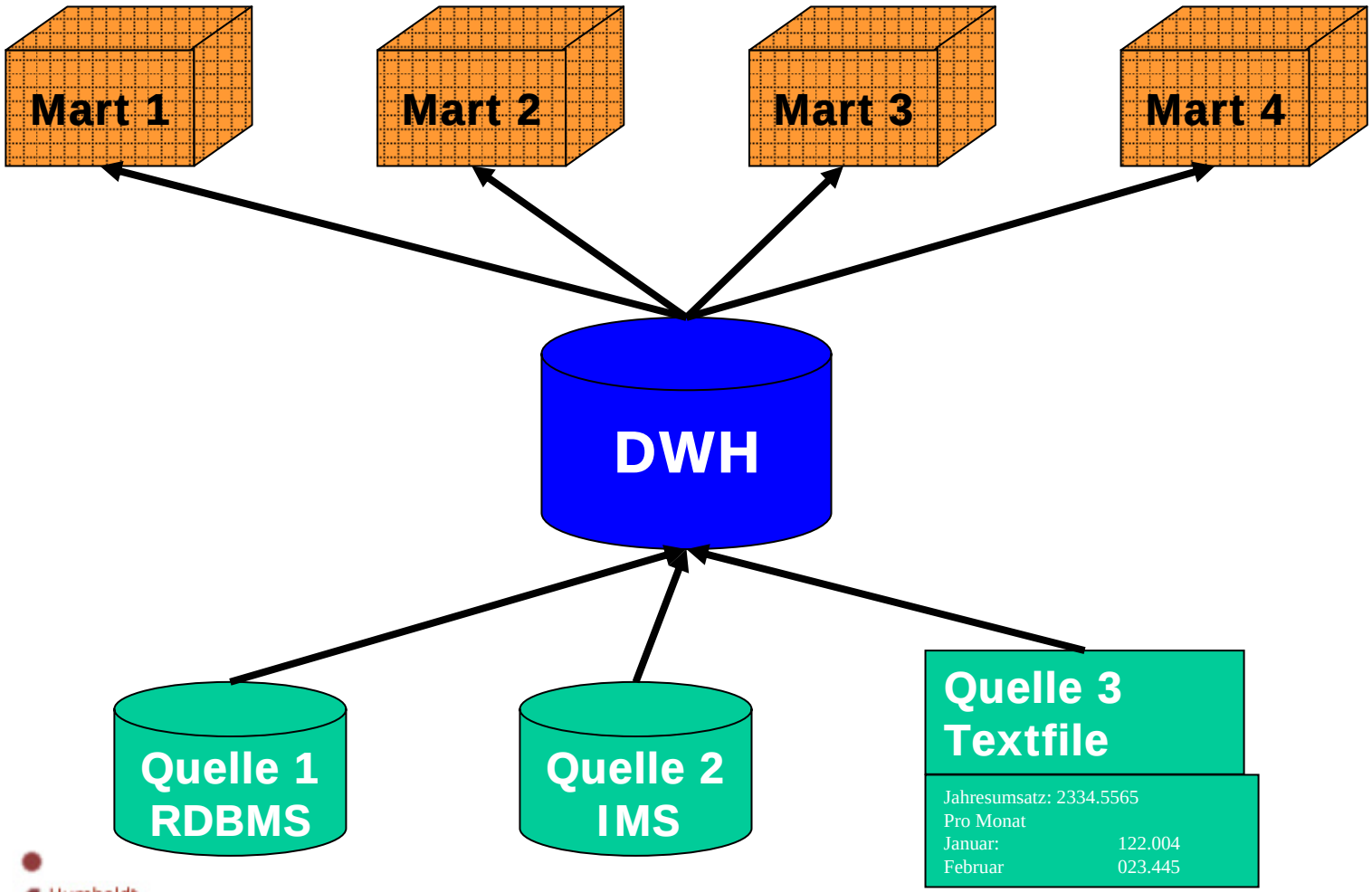
Modellierung im DWH

Das multidimensionale Datenmodell

Ulf Leser

Wissensmanagement in der Bioinformatik

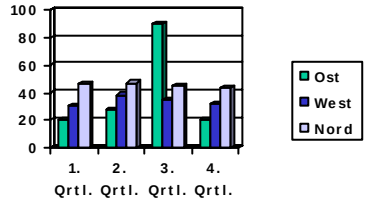
Zusammenfassung: Hubs and Spokes



DWH Architektur

Data Warehouse Manager

Analysewerkzeuge



Metadaten

Quelle 1 RDBMS

Quelle 2 IMS

Staging Area

Staging Area

Cube

Mart 2

Mart 1

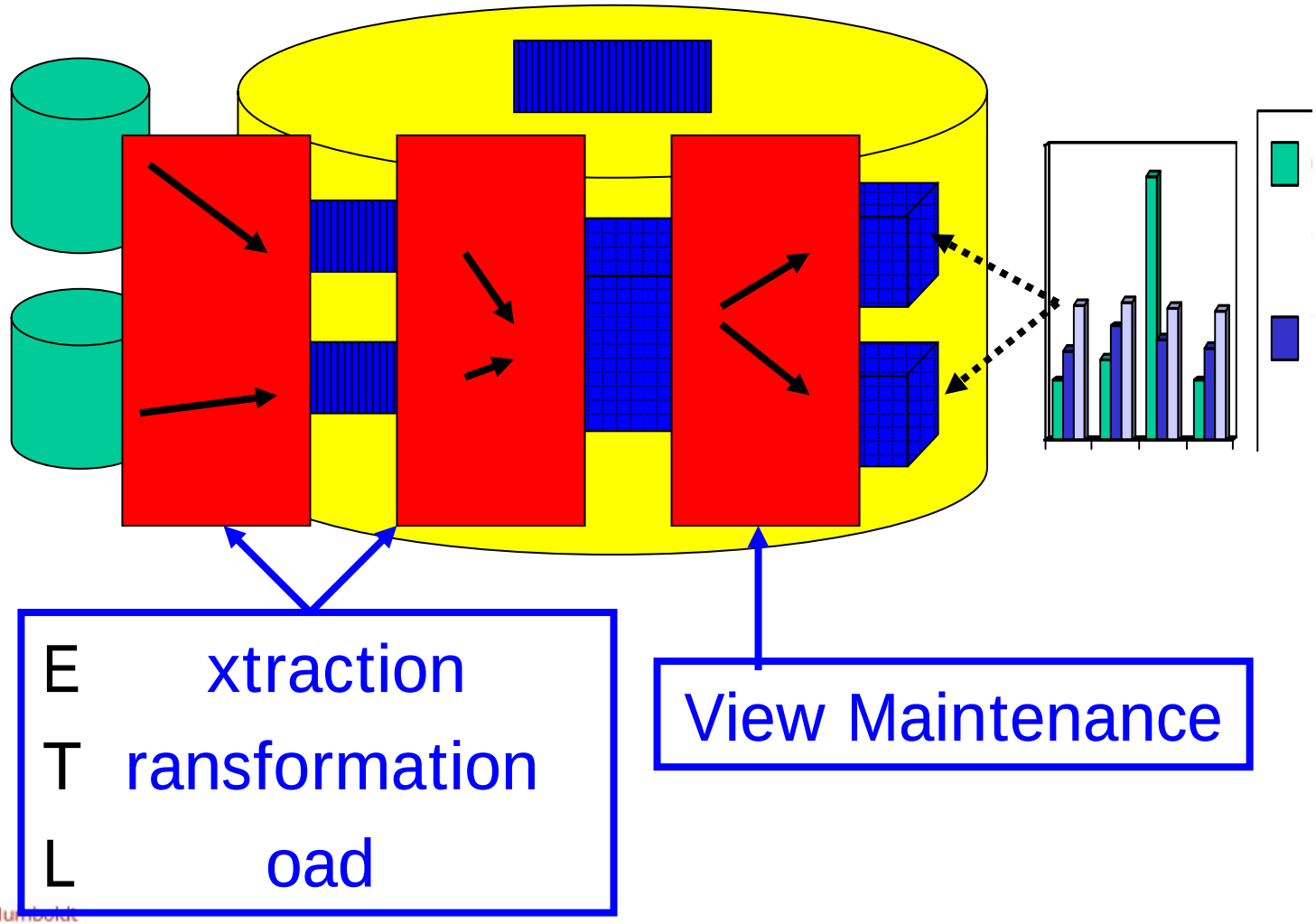
Datenquellen

Basisdaten

Abgeleitete Sichten

Arbeitsbereich

DWH Prozesse



Transformationen in Staging Area

- Benutzte SQL
- Effiziente mengenorientierte Berechnungen möglich
- Vergleiche über Zeilen hinaus möglich
  - Schlüsseleigenschaft, Namensduplikaterkennung, ...
- Vergleiche mit Daten in Basisdatenbank möglich
  - Duplikate, Plausibilität (Ausreißer), Konsistenz (Artikel-Ids)
- Typisch: Tagging von Datensätzen durch Prüf-Regeln

```
UPDATE sales SET price=price/MWST;
UPDATE sales SET cust_name=
  (SELECT cust_name FROM customer WHERE id=cust_id);
...
UPDATE sales SET flag1=FALSE WHERE cust_name IS NULL;
...
INSERT INTO DWH
  SELECT * FROM sales WHERE f1=TRUE & f2=TRUE & ...
```

BULK Uploads

- Für große Datenmengen einzige ausreichend performante Schnittstelle
- Kritischer Prozess
  - Load-Vorgänge blockieren i.d.R. die komplette DB (Schreibzugriff auf komplette Tabelle)
  - Konsistenz, Trigger, ICs i.d.R. deaktiviert
  - Indexaktualisierung
  - Update oder Insert ? (Upsert!)
- Performance von LOAD oft limitierender Faktor

Inhalt dieser Vorlesung

- Das Multidimensionale Datenmodell (MDDM)
  - Grundidee
  - Formale Definition der Modellelemente
  - 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 in unterschiedlichem Verfeinerungsgrad
- Analyse durch Operationen auf dem Cube
  - Dimensionen ausblenden / einblenden
  - Auswahl von Subwürfeln (Flächen, Punkten, ...)
  - Hierarchiestufe vergröbern/verfeinern

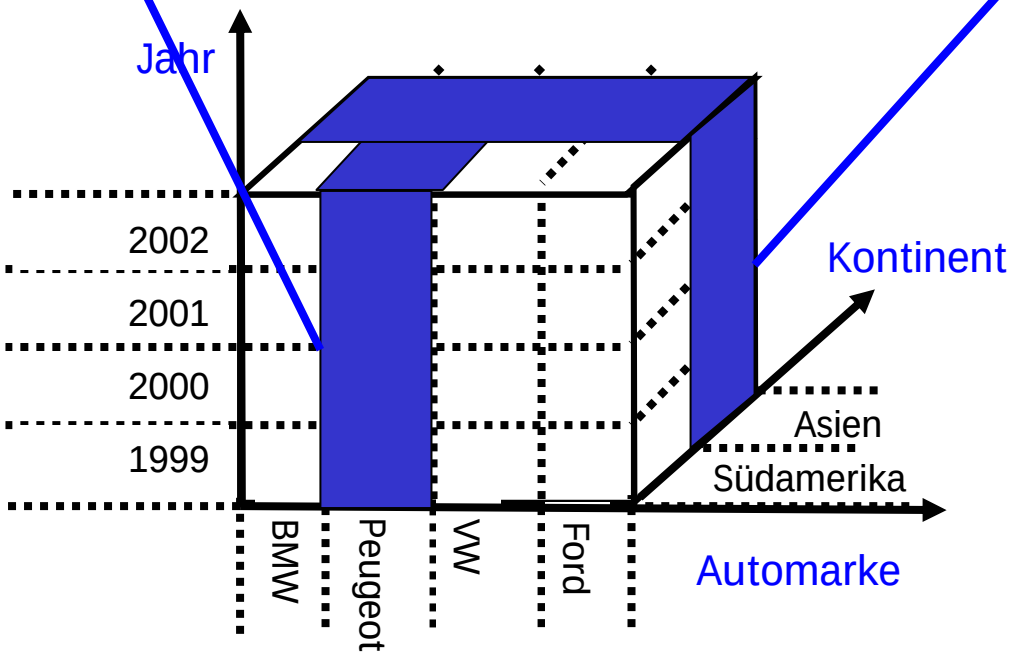
Beispiel

- 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

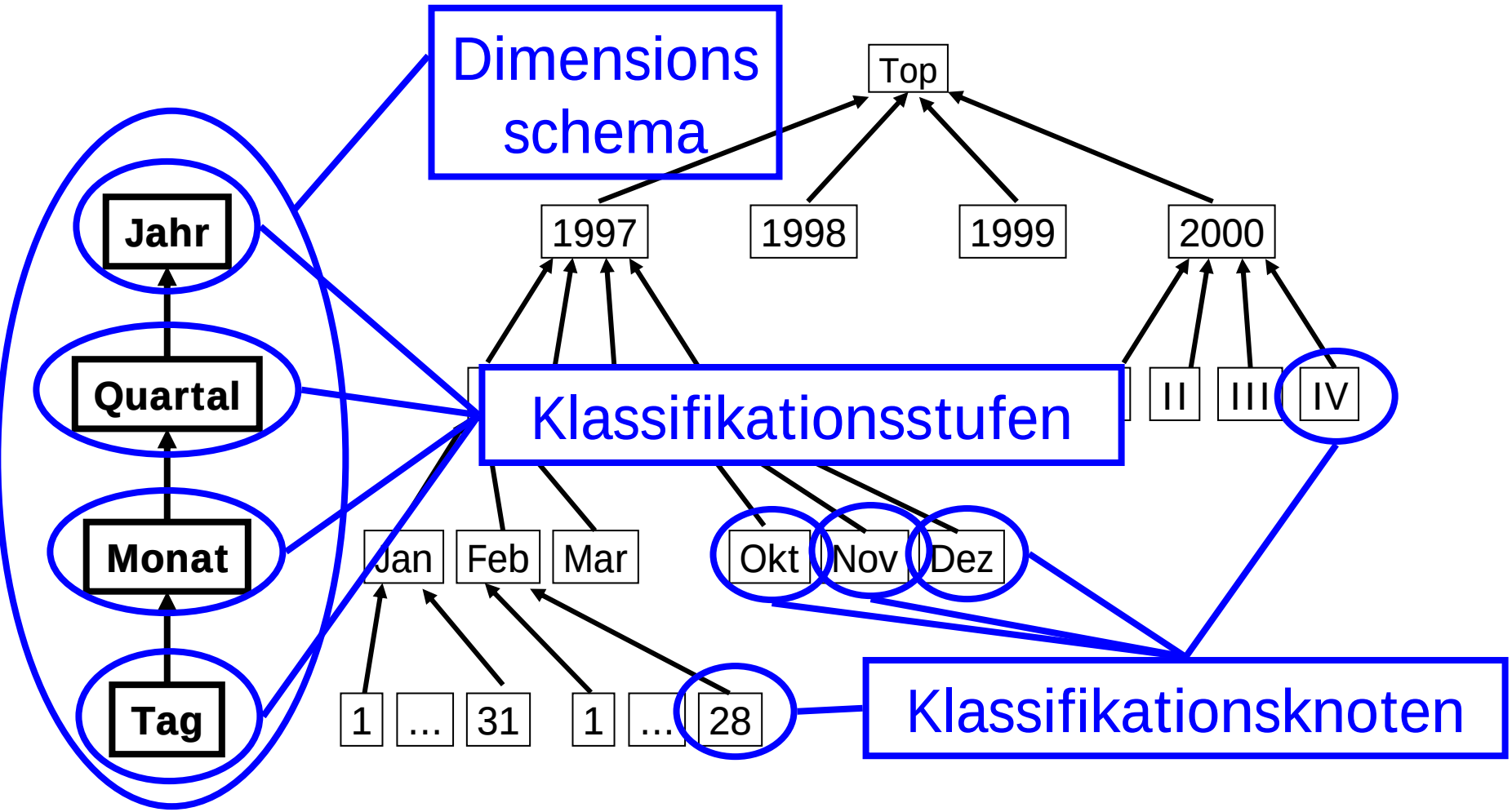
Verkäufe in Asien pro Jahr und Marke



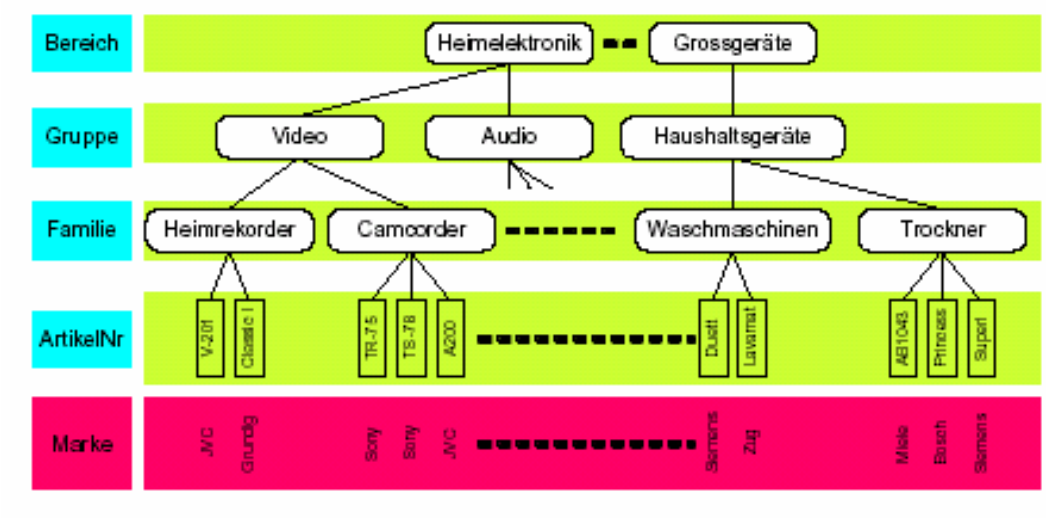
Dimensionen

- Eindeutige Strukturierung des Datenraums
- Hoffentlich orthogonal
  - Abhängigkeiten zwischen Dimensionen bereiten an vielen Stellen Probleme – später
- Jede Dimension hat ein Schema
  - Tag, Woche, Jahr
  - Landkreis, Land, Staat
  - Produktgruppe, Produktklasse, Produktfamilie
- ... und Werte
  - (1, 2, 3, ..., 31), (1, ... 52), (1900, ..., 2003)
  - (...), (Berlin, NRW, Department-1, ...), (BRD, F, ...)

Dimension


Dimensions schema
Top
1997
1998
1999
2000
Jahr
Quartal
Monat
Tag
Klassifikationsstufen
II
III
IV
Jan
Feb
Mar
Okt
Nov
Dez
1
...
31
1
...
28
Klassifikationsknoten

Produkthierarchie

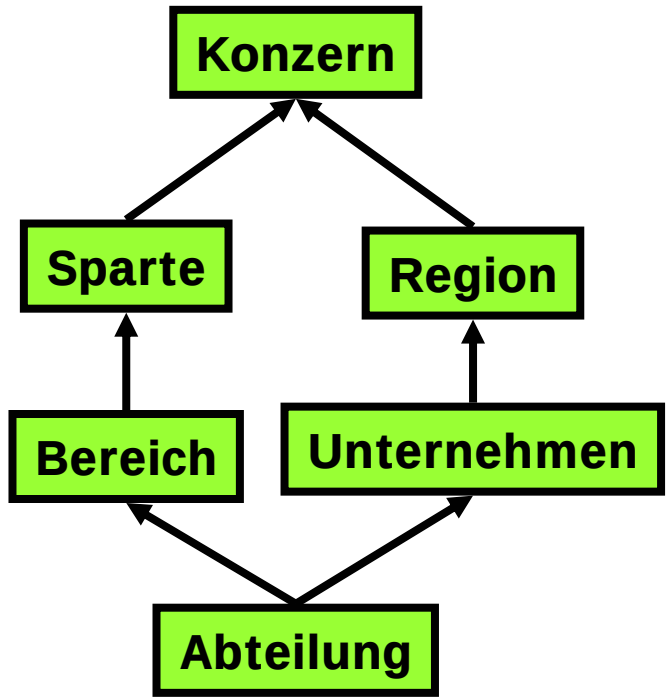


Aus: Geppert, ETZ Zürich, Vorlesung „Data Warehouse“

- 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

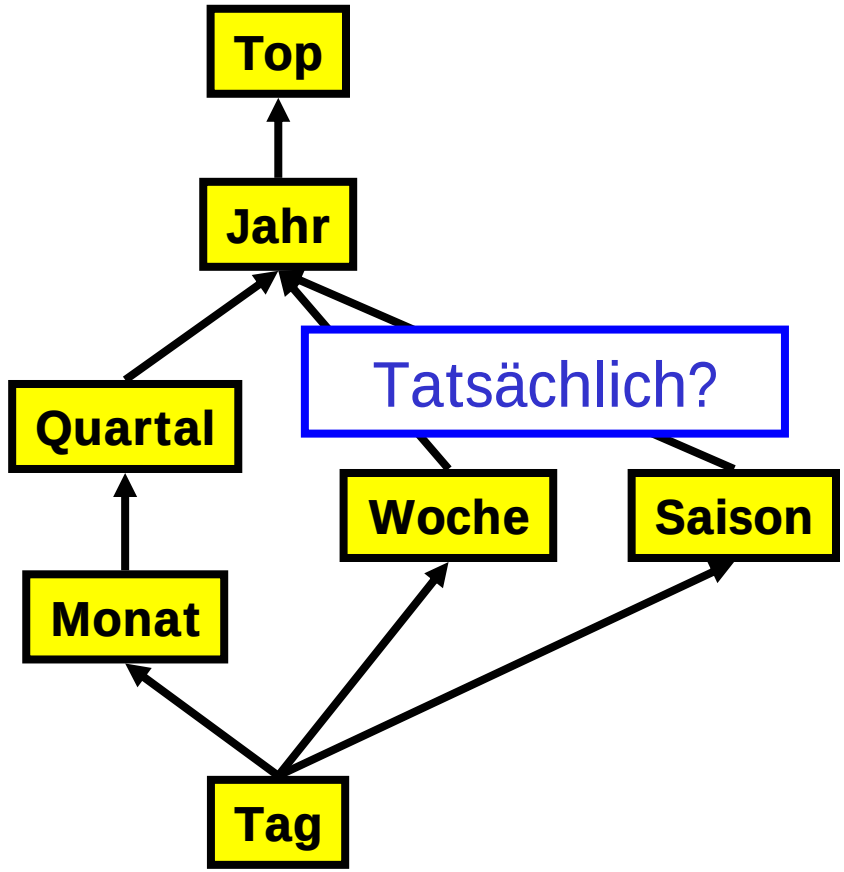
- *Definition*
  Ein Klassifikationsschema K (einer Dimension D) ist ein Quadrupel $(K_k, \rightarrow_k, K_s, \rightarrow_s)$ *mit*
  - K_s *ist die Menge von Klassifikationsstufen* $\{k_0, \ldots k_n\}$
  - *„*$\rightarrow_s$*“ ist eine Halbordnung auf* K_s *mit größtem Element* $top(K_s)$
    - *D.h.:* $\forall k \in K_s: k \rightarrow_s top(K_s)$
  - K_k *ist die Menge von Klassifikationsknoten* $\{n_0, \ldots n_m\}$
  - *Jeder Klassifikationsknoten n ist genau einer Klassifikationsstufe k zugeordnet*
    - *stufe(n) = k*
    - $knoten(k) = \{n | n \in K_k \wedge stufe(n)=k\}$
  - *„*$\rightarrow_k$*“ ist die Halbordnung auf* K_s *übertragen auf* K_k
    - $k,l \in K_s, k \rightarrow_s l \Rightarrow \forall n \in knoten(k), m \in knoten(l): n \rightarrow_k m$
- 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

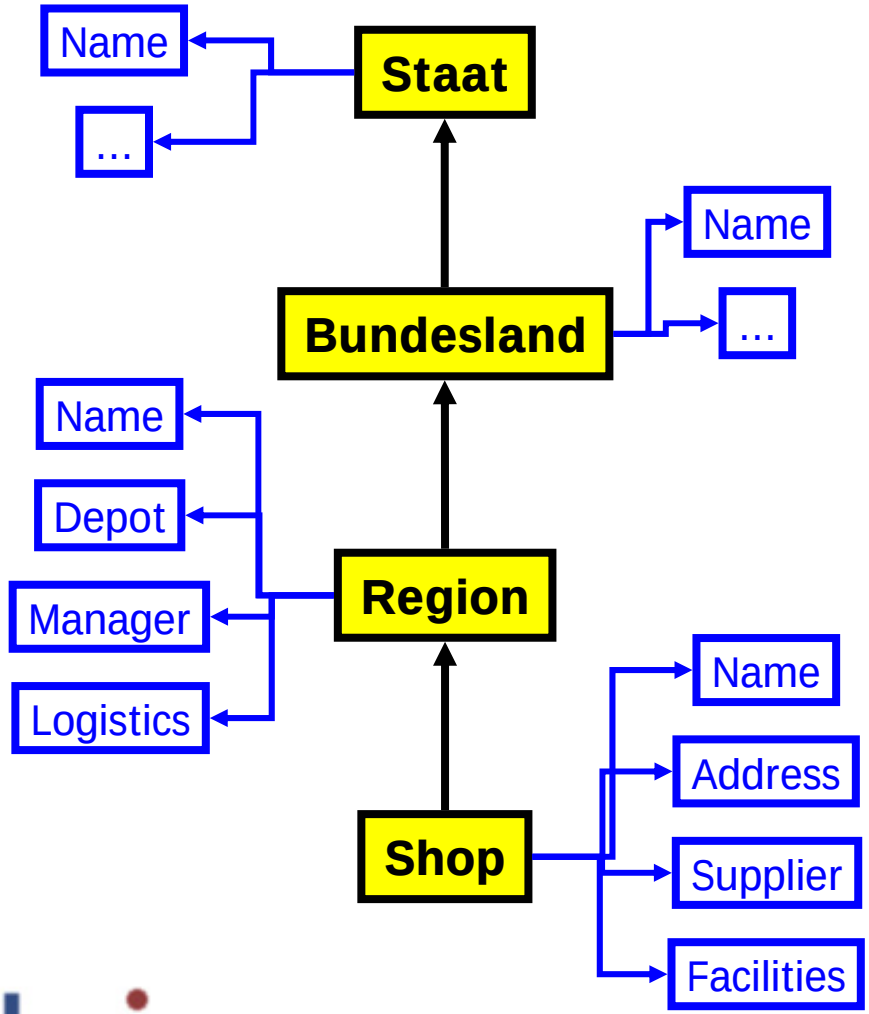
- Die Klassifikationsstufen sind die Schemaelemente der Dimension
- Das größte Element ist artifiziell – es steht für „alles“, also die Verdichtung in einen einzelnen Wert
  - Wir nennen es TOP
- Interpretation von „→“
  - Funktionale Abhängigkeit
  - Aggregierbarkeit
  - Tag bestimmt Monat bestimmt Jahr bestimmt TOP
    - 21.12.2003 → 12.2003 → 2003 → TOP
  - Produkt → Produktfamilie → Produktgruppe → TOP
    - “Asus M2400N“ → Notebooks → Büroelektronik → TOP
- Beachte: Halbordnung ist immer zyklusfrei
- Klassifikationsknoten sind die Instanzen der Schemaelemente

Beispiel Halbordnung

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



Knotenattribute



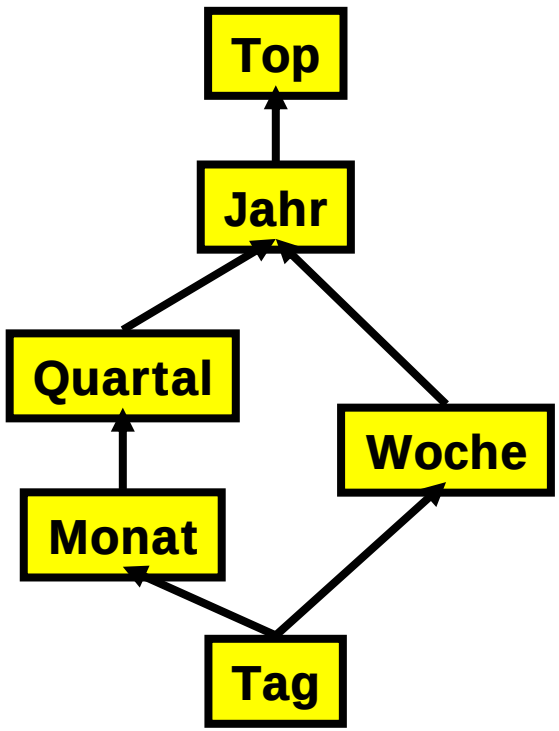
- Jede Klassifikationsstufe hat eine Menge von Attributen, die Knotenattribute
  - Teil des Schemas des Klassifikationsschemas
- Die Klassifikations-knoten haben Werte für diese Knotenattribute

Klassifikationspfade

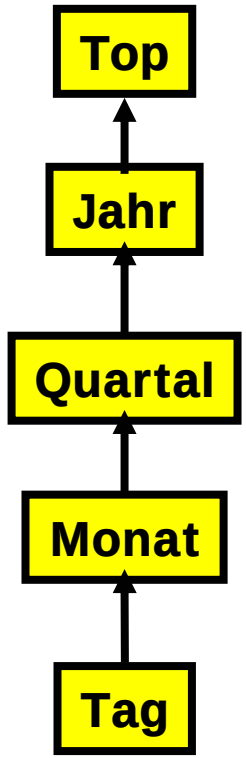
- *Definition*
  Ein Klassifikationspfad P in einem Klassifikationsschema K mit Klassifikationsstufen K_s ist eine Menge $\{p_0, ..., p_m\}$ mit
  - $\{p_0, ..., p_m\} \subseteq K_s$
  - $p_m = top(K_s)$
  - $\forall p_i , 1 \leq i \leq m: p_{i-1} \rightarrow p_i$ und $\nexists q: p_{i-1} \rightarrow q \rightarrow p_i$
  - *Die Länge des Pfades P ist $|P|=m+1$*
  - *Der Klassifikationslevel von p_i in P ist i*
- Bedeutung
  - Ein Pfad ist damit eine voll geordnete Teilmenge von K_s
  - Jeder Pfad beinhaltet das größte Element TOP
  - Verdichtung wird nur entlang von Klassifikationspfaden definiert
    - Und damit entlang funktionaler Abhängigkeiten

Beispielpfade

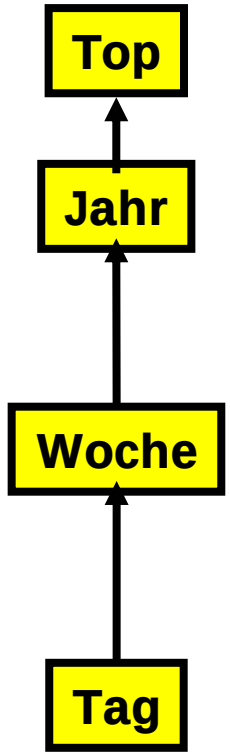
Klassifikationsschema



Pfad 1

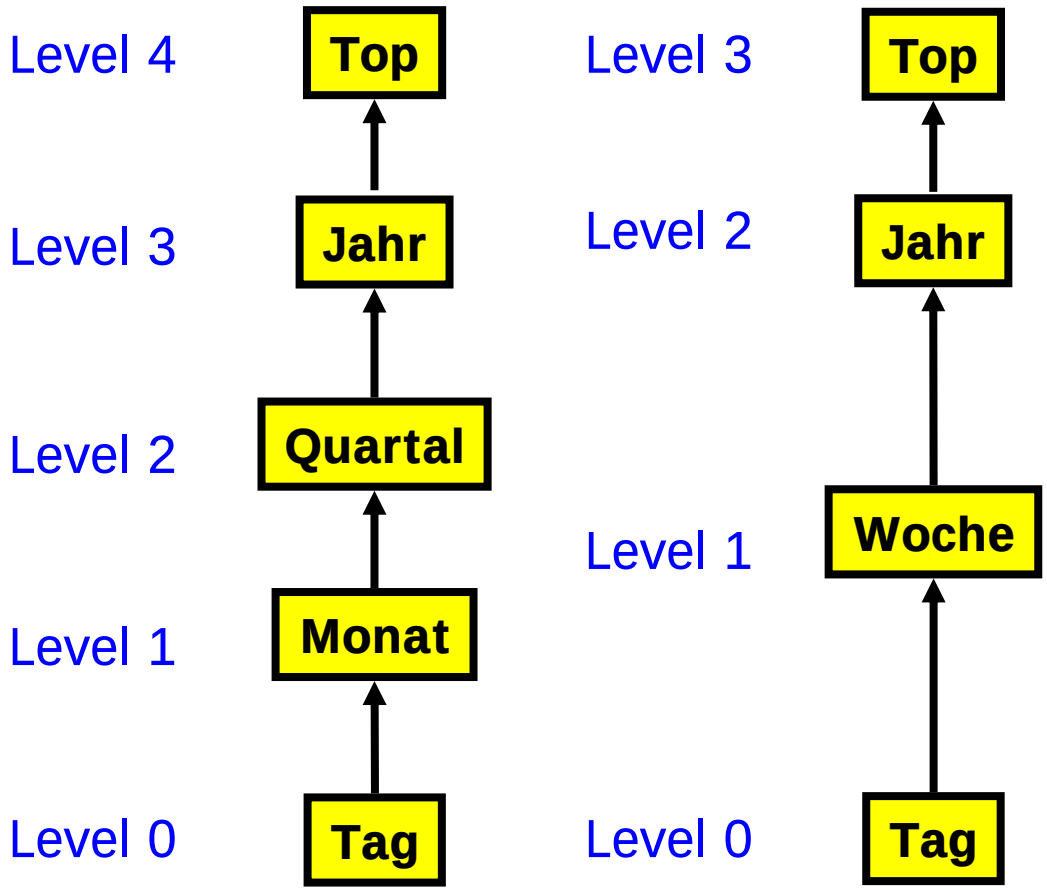


Pfad 2



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:*

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

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

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

Dimension

- *Definition*
  Eine Dimension $D=(K, \{P_1,\ldots,P_j\})$ 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
  - Nicht alle Klassifikationsstufen von K müssen in einem Pfad enthalten sein
    - Aber man wird seine Pfade so wählen, dass doch
- Schreibweise
  - D.k bezeichnet eine Klassifikationsstufe k aus D
  - Ein D.k kann in mehreren Pfaden vorkommen

Granularität

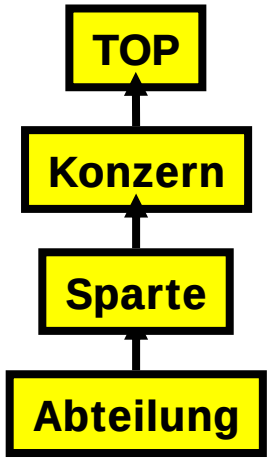
- *Definition*
  Gegeben eine Menge U von n Dimensionen $D_1,\ldots, D_n$.
  Eine Granularität G über U ist eine Menge $\{D_1.k_1,\ldots ,D_n.k_n\}$ für die gilt
  - *k_i ist eine Klassifikationsstufe in D_i*
  - *Es gibt keine funktionalen Abhängigkeiten zwischen den Klassifikationsstufen $D_1.k_1,\ldots, D_n.k_n$*
- Bemerkungen
  - Abkürzung: Lässt man in U eine Dimension D_i weg, meint dies implizit D_i.TOP
  - Zweite Bedingung ist immer erfüllt, wenn keine funktionalen Abhängigkeiten zwischen Dimensionen bestehen
    - Beispiel: Nicht gleichzeitig Dimensionen Zeit und „Fiskalisches Jahr“ in einer Granularität betrachten
  - Mit einer Granularität legt man fest, in welcher Detailstufe Fakten in den einzelnen Dimensionen beschrieben werden
    - Eine Granularität ist ein Hyperwürfel in einer bestimmten Auflösung
    - Operationen navigieren zwischen Granularitäten

Halbordnung auf Granularitäten

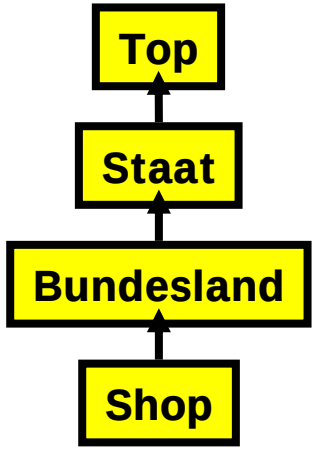
- *Definition*
  Auf der Menge aller Granularitäten zu einer Menge U von Dimensionen ist eine Halbordnung „≤“ wie folgt definiert
  - *Sei $G_1=\{D_1^1.k_1^1,\ldots ,D_n^1.k_n^1\}$ und $G_2=\{D_1^2.k_1^2,\ldots ,D_n^2.k_n^2\}$*
  - *Ordne die Dimensionen in G_1‘ und G_2‘ beliebig, aber gleich*
  - *Es gilt $G_1 \leq G_2$ genau dann wenn*
    - *$\forall i: D_i^1.k_i^1 \rightarrow D_i^2.k_i^2$ oder $D_i^1.k_i^1 = D_i^2.k_i^2$*
- Benutzung
  - Beschreibung der Transformation von Granularitäten
  - Anfrageoptimierung: Wiederverwendung von Aggregaten

Beispiel

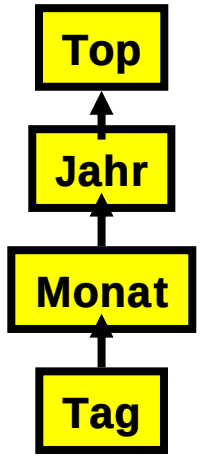
Bereich



Region



Zeit




```
(B.Sparte, R.Shop, Z.Tag)
≤ (B.Sparte, R.Shop, Z.Monat)
≤  (B.Sparte, R.Top, Z.Monat)
≤  (B.Top, R.Top, Z.Top)
```

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

Würfelschema

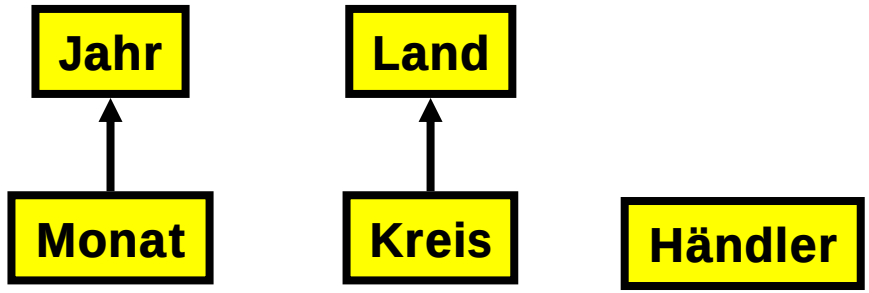
- *Definition*
  Ein Würfelschema WS=(G,F) besteht aus
  - *Einer Granularität G*
  - *Einer Menge F von unterschiedlichen Fakten F_i*
- *Definition*
  Ein Würfel W ist eine Instanz eines Würfelschema (G,F)

$$W = dom(G) \times dom(F)$$
$$= knoten(D_1.k_1) \times ... \times knoten(D_n.k_n) \times dom(F_1) \times ... \times dom(F_m)$$

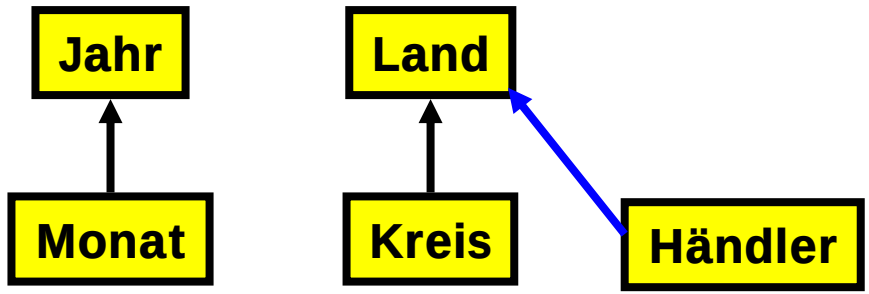
- Bemerkung
  - Die Werte dom(G) geben die Koordinaten der Werte dom(F) an
  - Verhältnis Würfelschema zu Würfel ist wie Relationenschema zu Relation

Kein Würfelschema

- Autoverkäufe pro Zeit (Monat, Jahr) , Händler und Region (Kreis, Land)
- Drei Dimensionen
  - Monat → Jahr
  - Händler
  - Kreis → Land

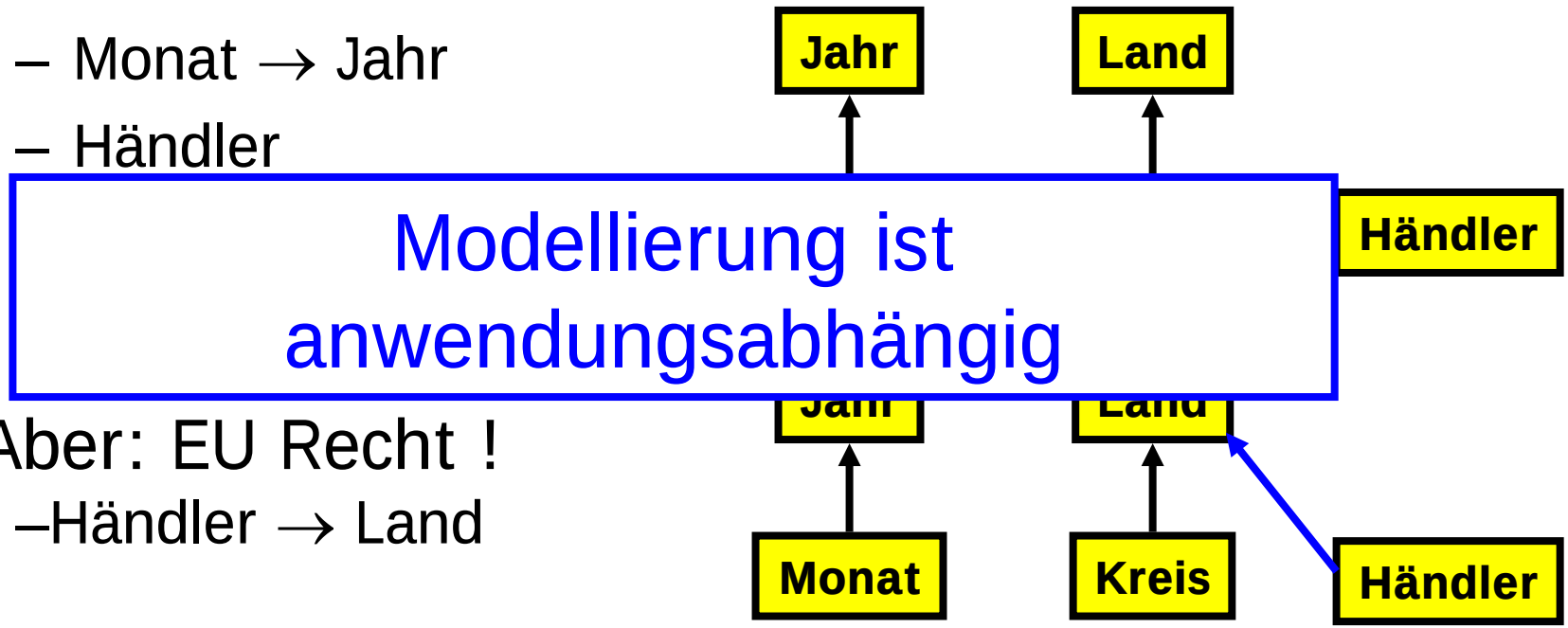


- Aber: EU Recht !
  - Händler → Land
  - Damit können wir keine Granularität bauen



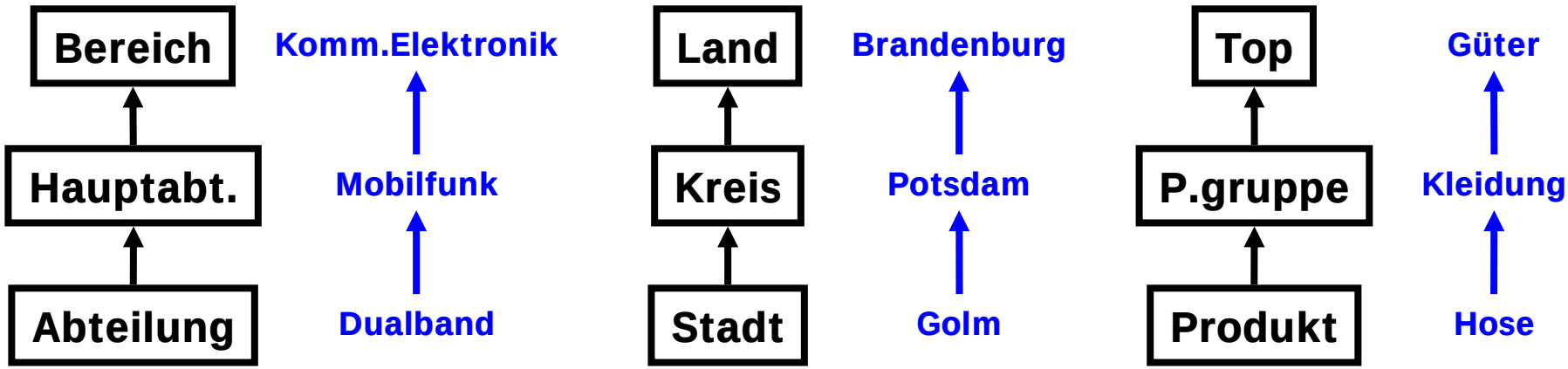
Kein Würfelschema

- Autoverkäufe pro Zeit (Monat, Jahr) , Händler und Region (Kreis, Land)
- Drei Dimensionen
  - Monat → Jahr
  - Händler
- Aber: EU Recht !
  - Händler → Land



Semantik von Kanten

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



Part-Of Topologisch IS-A

Ein längeres Beispiel

- Wir bauen ein DWH zur Verwaltung von Lagerbeständen
- 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 ← Jahr ← Quartal ← Monat
- P_2: Top ← Jahr ← Woche
- P_3: Top ← Land ← Region
- P_4: Top ← Bereich ← Artikelgruppe ← Artikel

➢ Entlang der Pfade sind Verdichtungen im Modell sinnvoll

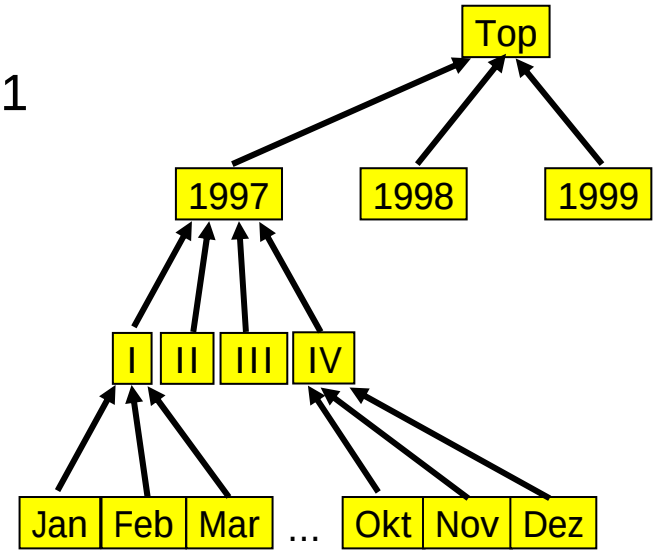
Klassifikationsknoten

- 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
  - ...

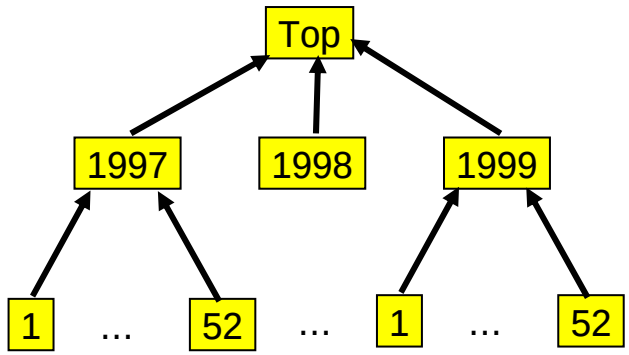
➢ Alle möglichen Ausprägungen der Klassifikationsstufen

Klassifikationshierarchien 1

- Klassifikationshierarchie zu P_1

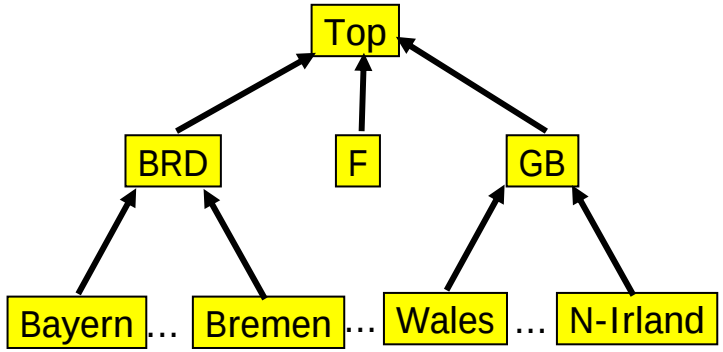


- Klassifikationshierarchie zu P_2

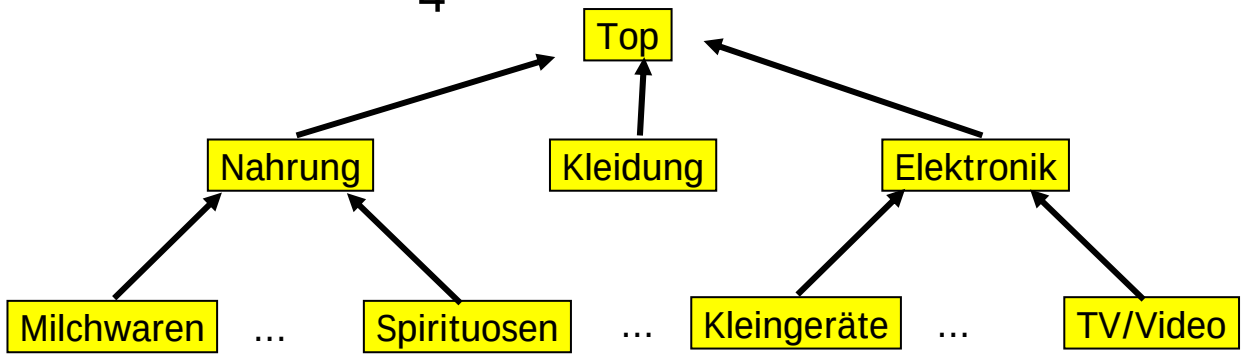


Klassifikationshierarchien 2

- Klassifikationshierarchie P_3



- Klassifikationshierarchie P_4



Dimensionen

- Dimension ZEIT
  - (K, {P_1, P_2})
  - Umfasst Monat, Quartal, Woche, Jahr
- Dimension ORT
  - (K, {P_3})
  - Umfasst Region, Land
- Dimension PRODUKT
  - (K, {P_4})
  - Umfasst Artikel, Artikelgruppe, Bereich

➢ Dimensionen enthalten mehrere Pfade und damit Klassifikationsstufen

Granularität, Würfel

- Mögliche Granularitäten
  - G_1 = (Zeit.Woche, Ort.Land, Produkt.Artikel)
  - G_2 = (Zeit.Jahr, Ort.Gebiet, Produkt.TOP)
  - Halbordnung:
    - (Zeit.Woche, Ort.Gebiet, Produkt.Artikel)
    - ≤ (Zeit.Jahr, Ort.Gebiet, Produkt.Bereich)
    - ≤ (Zeit.Jahr, Ort.Top, Produkt.Bereich)
    - ≤ (Zeit.ZOP, Ort.Top, Produkt.Top)
- Würfelschema
  - Granularität plus Menge von Fakten (F_1=Bestand, F_2=Delta)
- ➢ Würfel: Instanz des Würfelschemas
- ➢ Operationen auf Würfeln verändern die Granularität
- ➢ Ziel: Nur sinnvolle Operationen zulassen

Zusammenfassung