Data Warehousing und Data Mining

OLAP Operationen

Ein konzeptionelles MDDM

Aggregierbarkeit

Ulf Leser

Wissensmanagement in der Bioinformatik

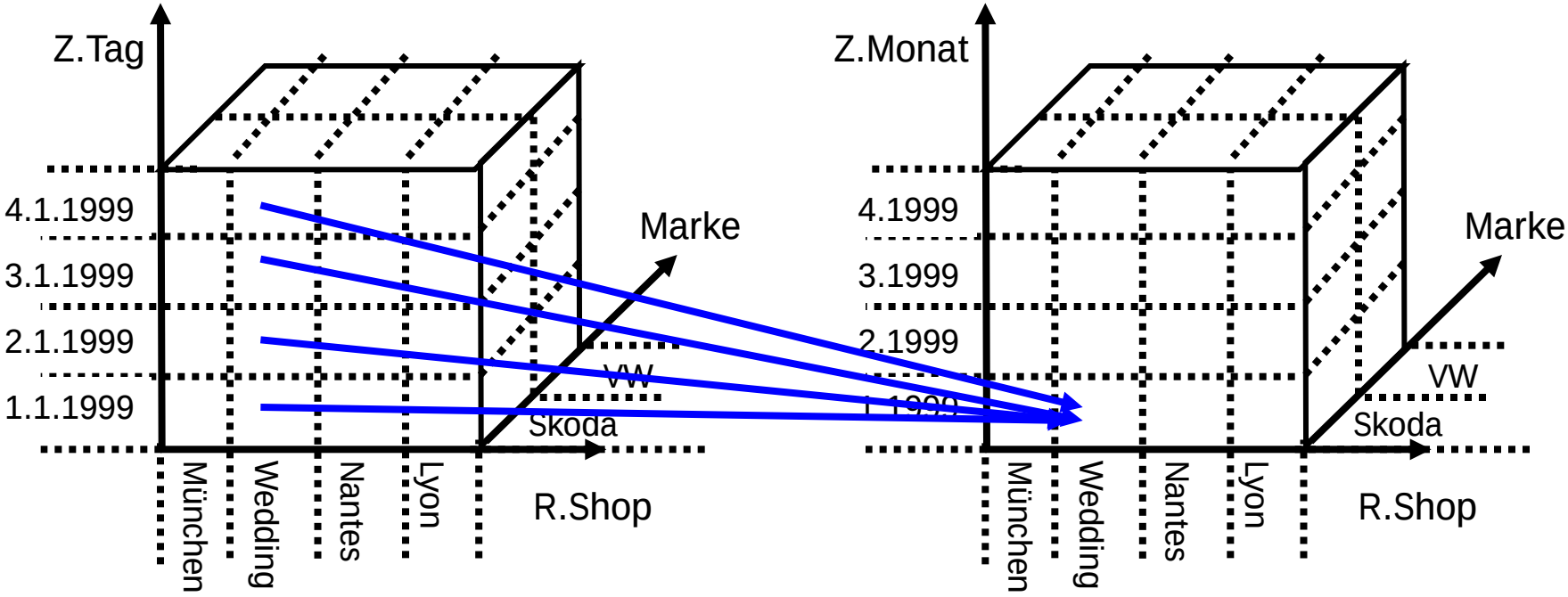
Inhalt dieser Vorlesung

- Operationen im multidimensionalen Datenmodell
  - Aggregation
  - Verfeinerung
  - Weitere Operationen
- ME/R: Graphische multidimensionale Datenmodellierung
- Aggregierbarkeit

OLAP Operationen

- MDDM entspricht der Sprache des Betriebswirts
  - Objekte seiner täglichen Welt (Kunden, Waren, etc.)
  - Keine Aufsplittung in Relationen etc.
- Operationen auf dem MDDM sollen betriebswirtschaftliche Analysen unterstützen
  - Schnelle und intuitive Navigation durch multidimensionale Daten
  - Implementierung kommt später
- Grundoperationen
  - Aggregation (Roll-Up): Granularität (Abstraktion) wird erhöht
  - Verfeinerung (Drill-Down): Granularität wird erniedrigt
  - Wahl der Dimensionen, Klassifikationspfade und -stufen
  - Aggregation immer bzgl. einer Funktion f
    - SUM, AVG, MEDIAN, ...

Übersicht



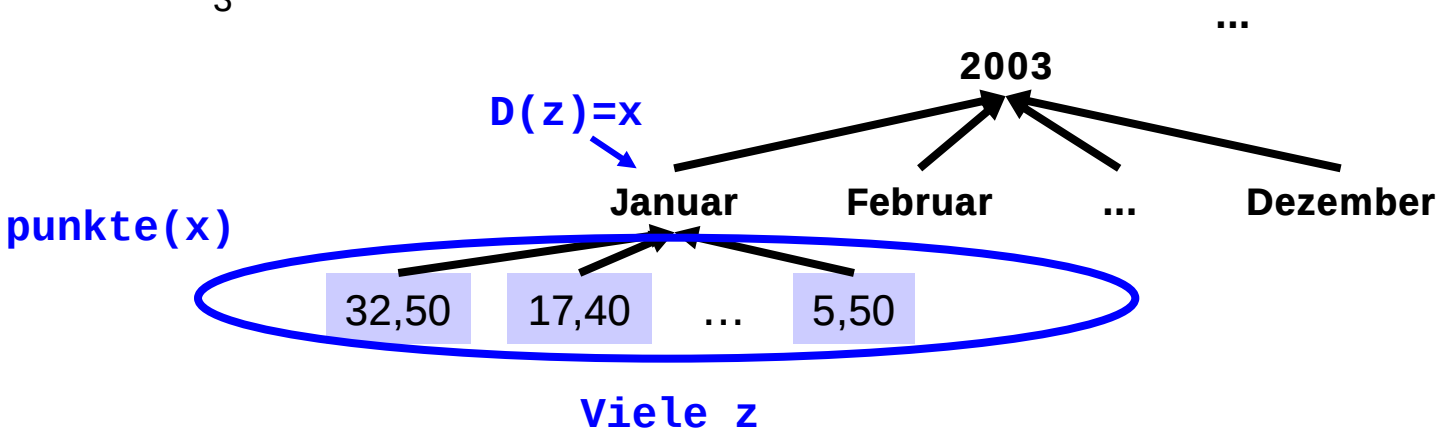

```
(1.101,  1.1.1999, Wedding, Skoda)
(  129,  8.1.1999, Wedding, Skoda)
(  225, 23.1.1999, Wedding, Skoda)
(1.540,  4.2.1999, Wedding, Skoda)
(2.500,  5.2.1999, Nantes, Skoda)
…
```

Rollup Tag→Monat

```
(1.455, 1.1999, Wedding, Skoda)
(4.040, 2.1999, Nantes, Skoda)
…
```

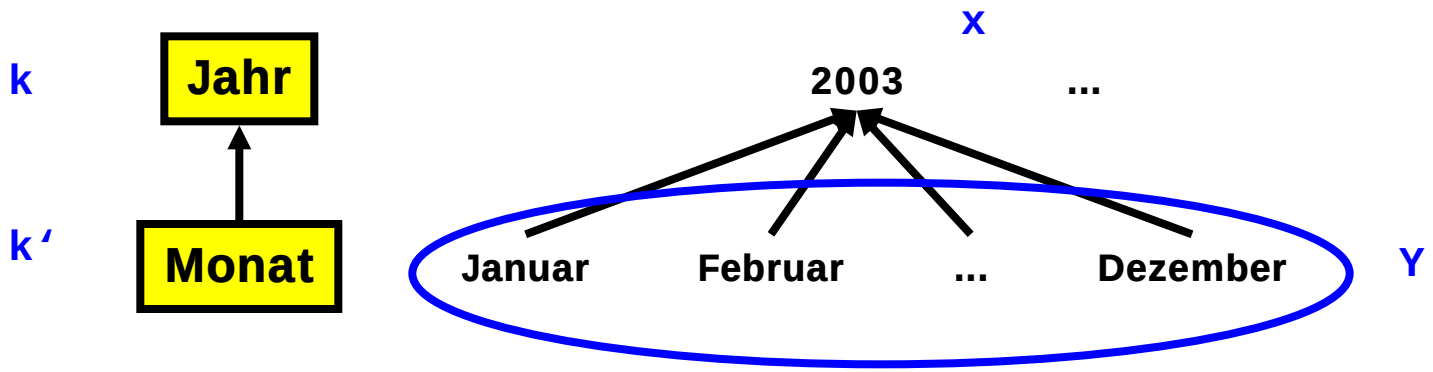
Datenpunkte und Würfelkoordinaten

- Definition
  Gegeben ein Würfel W, eine Dimension D aus W und ein Klassifikationsknoten x aus D. Sei z ein Fakt (Datenpunkt) und D(z) seine Koordinate bzgl. D. Dann gilt
  - *z liegt in x, $z \in x$, gdw. D(z)=x oder $D(z) \in$ nachfahren(x)*
  - *punkte(x)=$\{z \mid z \in x\}$*
- Erweiterung auf mehrere Koordinaten durch Schnitt
  - punkte(x,y,z)=punkte(x)$\cap$punkte(y)$\cap$punkte(z) mit $x \in D_1$, $y \in D_2$, $z \in D_3$



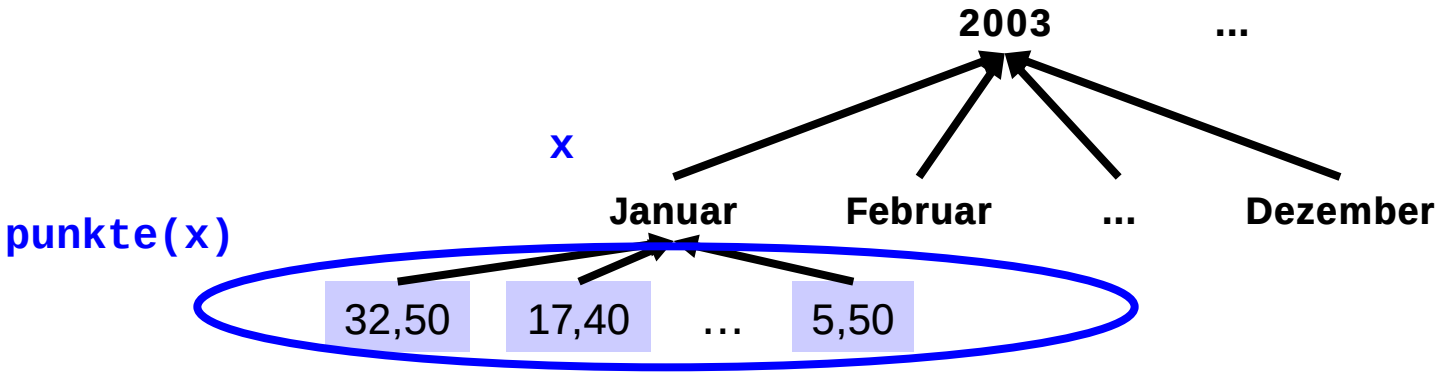
Aggregation in Hierarchien

- Definition. *Gegeben*
  - *Dimension D, Pfad P={$k_0 \rightarrow ... \rightarrow k' \rightarrow k \rightarrow ... \rightarrow TOP$} in D*
  - *x ein Klassifikationsknoten der Klassifikationsstufe k aus P*
  - *Aggregationsfunktion f , Measure F*
  - *Sei Y=kinder(x) die Menge {$y_1,...,y_n$} von Klassifikationsknoten von k', von denen x funktional abhängt*
  - *Die Aggregation von Y nach x bzgl. f und F bezeichnet die Berechnung des aggregierten Faktes $F(x)=f(F(y_1),...., F(y_n))$*
  - *Die Aggregation von k' nach k bzgl. f und F bezeichnet die Berechnung von F(x) für alle $x \in$ knoten(k)*
    - Das schreiben wir kurz als F(k)



Startpunkt

- Sei $P=\{k_0 \rightarrow \ldots \rightarrow k_n \rightarrow TOP\}$
- Man berechnet F(TOP) aus $F(k_n)$, $F(k_n)$ aus $F(k_{n-1})$, etc.
- Wie berechnet man $F(k_0)$?
- Definition. *Gegeben*
  - *Dimension D, Pfad $P=\{k_0 \rightarrow \ldots \rightarrow TOP\}$ in D*
  - *Ein Klassifikationsknoten $x_0 \in knoten(k_0)$ und $punkte(x_0)=\{z_1,\ldots,z_n\}$*
  - *Aggregatfunktion f , Measure F*
  - *Dann gilt: $F(x_0)=f(F(z_1),\ldots,F(z_n))$*
    - *Mit F(z) als dem Wert des Measures F von Datenpunkt z*



Berechnung

- Rekursive Definition ist anwendbar für alle Klassifikationsknoten bzw. –stufen eines Pfades
  - Also können wir F(x) für beliebige Klassifikationsstufen x aus den Datenpunkten berechnen
- Aber
  - Es ist schneller, die Aggregation aus der nächst-feineren Stufe zu berechnen
    - Weniger Werte
    - Das setzt voraus, dass man die auch hat – Präaggregation
    - Geht nicht für alle Aggregatfunktionen - Summierbarkeit

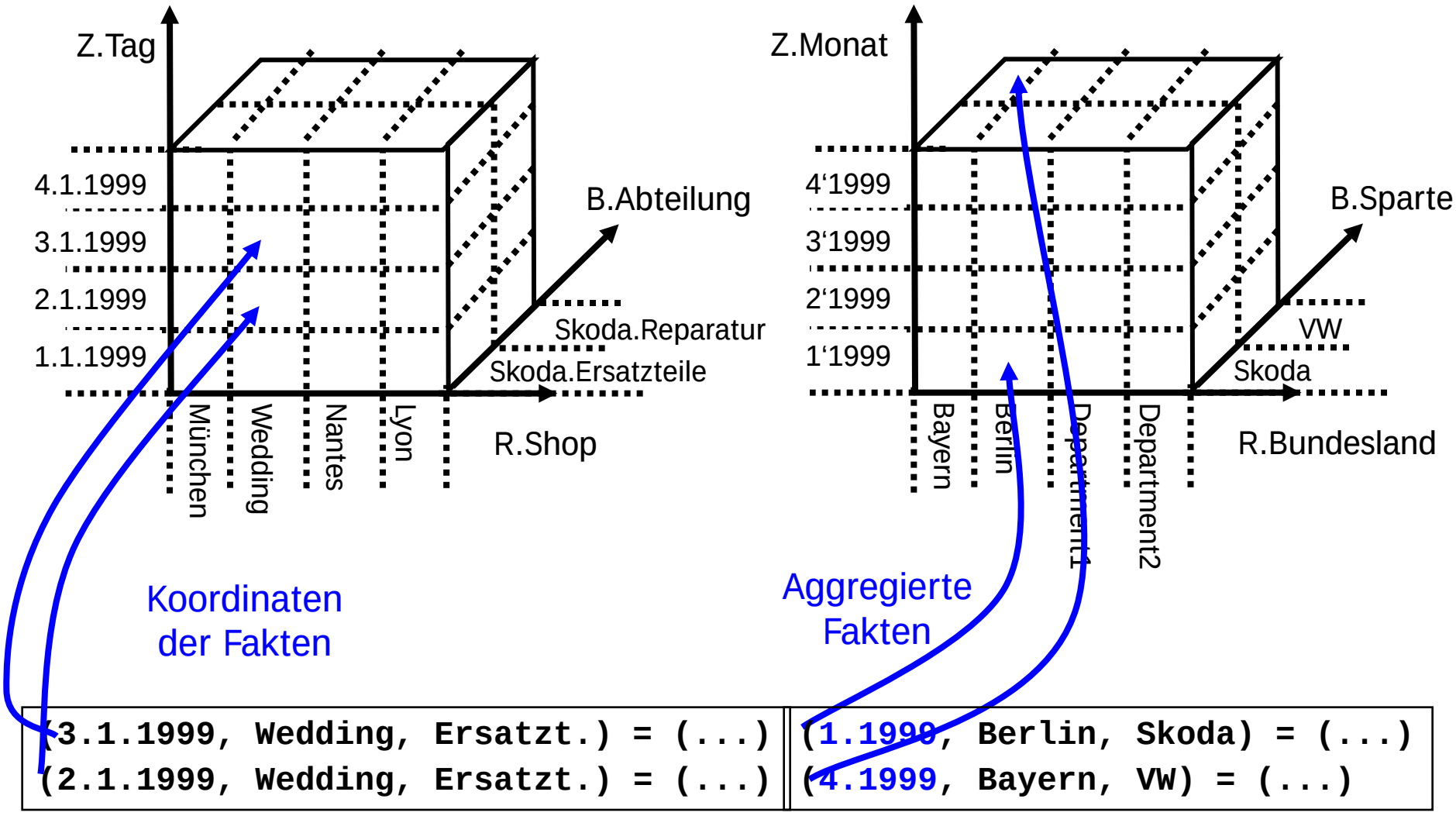
Würfelinhalt

- Ein Würfel W=(G,F)
  - Granularität G=($D_1.k_1$,...,$D_n.k_n$)
  - Measures F={F_1,...,F_m} mit Aggregationsfunktionen f_1,...,f_m
- Schreibweise für Zellen eines Würfels
  - W(x_1, x_2,..., x_n) = (F_1,...,F_m)
  - x_i sind die Koordinaten im Würfel bzgl. G: $x_i \in$ knoten(k_i)
  - Pro Zelle in W gibt es m (evt. aggregierte) Measures F_1,..., F_m
- Betrachten wir den einfachen Fall: n=m=1
  - Dimension D mit Knoten x, ein Measure F mit Aggregatfunktion f
  - Dann: W(x) = F(x) = f({F(y)|y$\in$kinder(x)})
- Bemerkung
  - Auf unterster Ebene ist kinder(x)=punkte(x)

Würfelinhalt, allgemeiner Fall

- Ein Würfel W=(G,F)
  - Granularität $G=(D_1.k_1,...,D_n.k_n)$
  - Measures $F=\{F_1,...,F_m\}$ mit Aggregationsfunktionen $f_1,...,f_m$
- Allgemeiner Fall
  - $W(x_1,...,x_n) = (F_1,...,F_m) =$
    $(f_1(F_1(punkte(x_1) \cap punkte(x_2) \cap ... \cap punkte(x_n))),$
    $f_2(F_2(punkte(x_1) \cap punkte(x_2) \cap ... \cap punkte(x_n))),$
    $...$
    $f_m(F_m(punkte(x_1) \cap punkte(x_2) \cap ... \cap punkte(x_n))))$
    mit $x_i \in knoten(k_i)$
- Bemerkung
  - Jedes Measure wird prinzipiell gesondert aggregiert

Beispiel



(3.1.1999, Wedding, Ersatzt.) = (...)	(1.1999, Berlin, Skoda) = (...)
(2.1.1999, Wedding, Ersatzt.) = (...)	(4.1999, Bayern, VW) = (...)

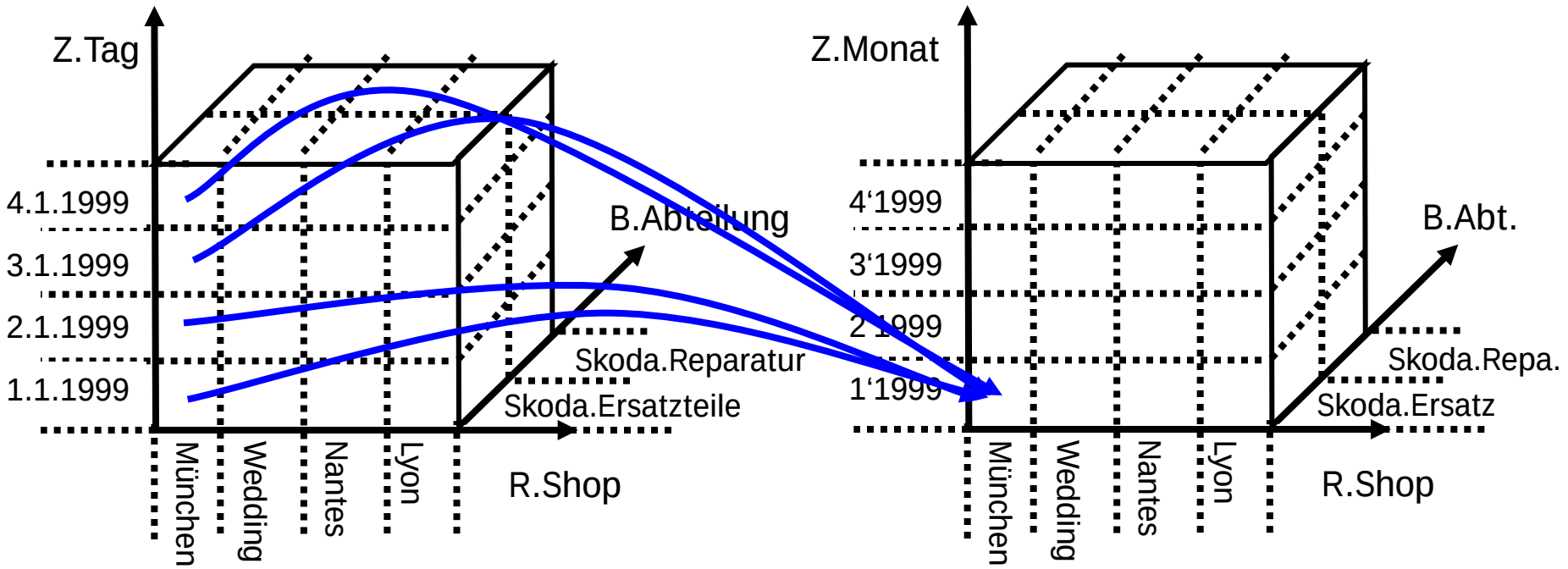
Operationen auf Würfeln

- OLAP Operationen überführen einen Würfel W=(G,F) in einen Würfel W‘=(G‘,F‘)
- Dabei gilt
  - Aggregation: G < G‘
  - Verfeinerung: G > G‘
- Eine einfache Operation verändert nur die Klassifikationsstufe einer Dimension in G
  - Komplexe Operationen können auf natürliche Weise aus einfachen Operationen durch Verkettung zusammengesetzt werden
  - Wir betrachten im folgenden nur einfache Operationen

OLAP Operation: Aggregation (Roll-Up)

- Definition
  - *Gegeben ein Würfel $W=(G,F)$ mit $G=(D_1.k_1,\ldots, D_i.k_i,\ldots,D_n.k_n)$ und $F=(F_1,..,F_m)$*
  - *Sei $P=\{k_0\rightarrow\ldots\rightarrow k_{i''}\rightarrow k_i\rightarrow k_{i'}\rightarrow\ldots\rightarrow TOP\}$ ein Pfad in D_i*
  - *Die einfache Aggregation in W entlang P mit Aggregatfunktion f überführt W in*
    - $W' = ((D_1.k_1,\ldots, D_i.k_{i'},\ldots, D_n.k_n), (F_{1'}, \ldots, F_{m'}))$
    - $(F_{1'}, \ldots, F_{m'}) = (f_1(F_1(punkte(k_1) \cap \ldots punkte(k_{i'}) \cap \ldots punkte(k_n))),$ $\ldots,$ $f_m(F_m(punkte(k_1) \cap \ldots punkte(k_{i'}) \cap \ldots punkte(k_n))))$

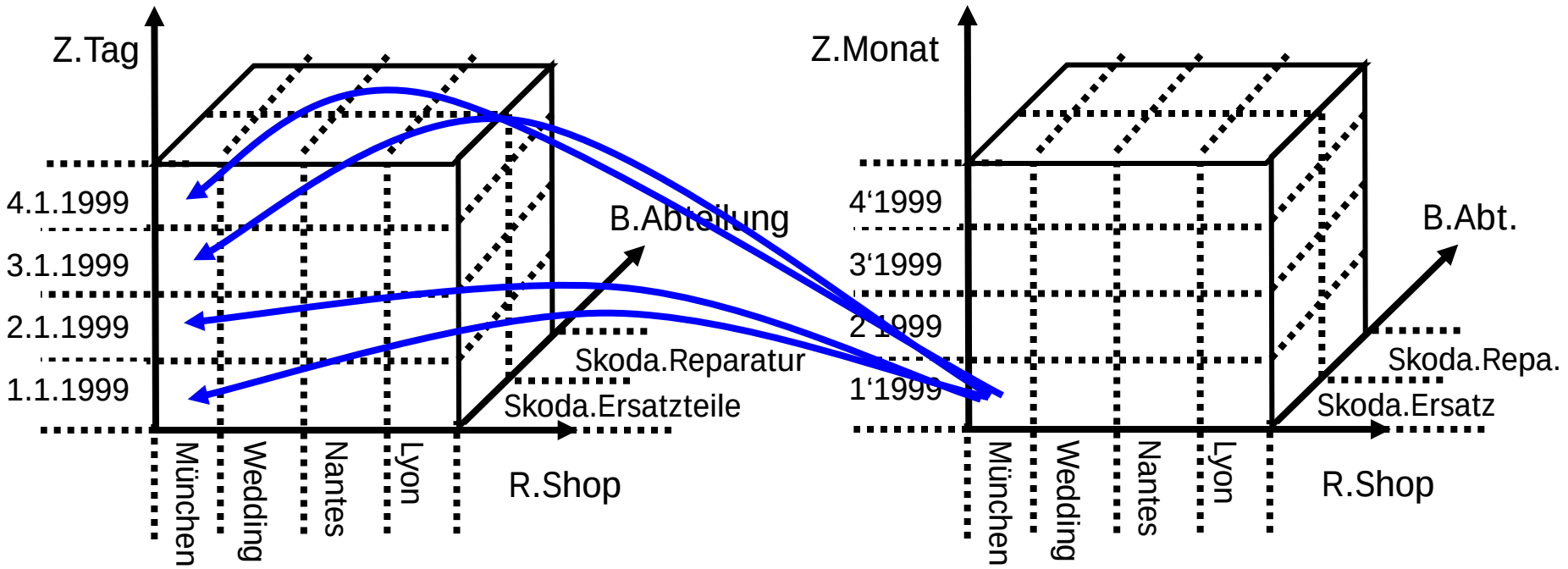
Beispiel Aggregation



Verfeinerung (Drill-Down)

- Definition
  - *Gegeben ein Würfel W=(G,F) mit G=($D_1.k_1$,…, $D_i.k_i$,…,$D_n.k_n$) und F=(F_1,..,F_m)*
  - *Sei P={$k_0 \rightarrow \ldots \rightarrow k_{i"} \rightarrow k_i \rightarrow k_{i'} \rightarrow \ldots \rightarrow TOP$} ein Pfad in D_i*
  - *Die einfache Verfeinerung in W entlang P mit Aggregatfunktion f überführt W in*
    - *W' = (($D_1.k_1$,…, $D_i.k_{i"}$,…, $D_n.k_n$), ($F_{1"}$, …, $F_{m"}$))*
    - *Mit ($F_{1"}$, …, $F_{m"}$) = …*
- Bemerkung
  - Ein „Aufbrechen“ einmal aggregierter Daten ist für die meisten Aggregatfunktionen nicht möglich
  - Stattdessen: Zugriff auf Daten (Datenpunkte / Prä-aggregate)

Beispiel Verfeinerung

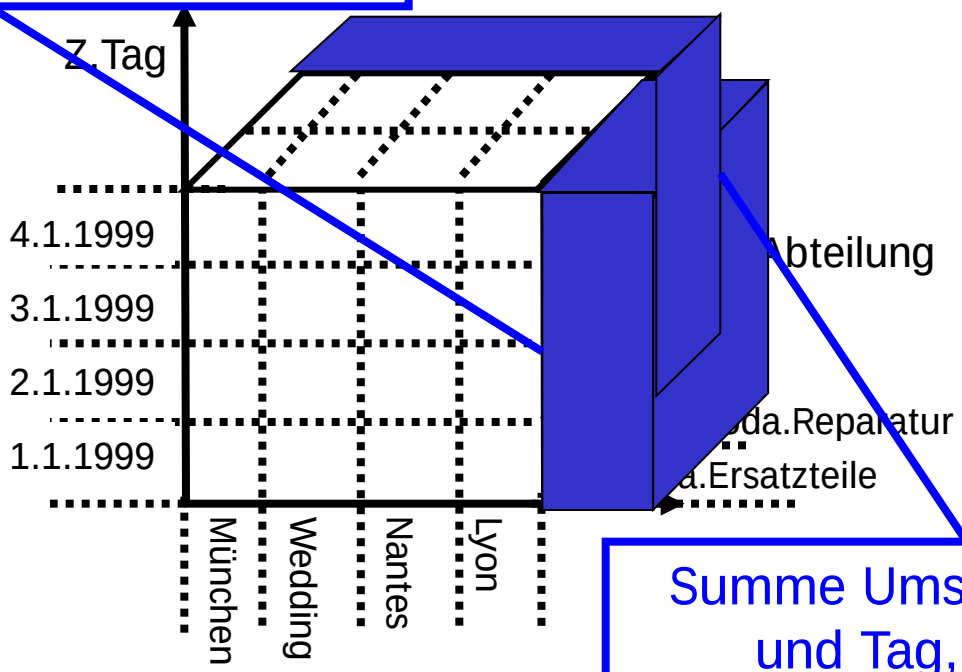


Auf einen Blick

- MDDM: Klassifikationsstufen, Knoten, Pfade, Dimensionen
- Die Daten bestehen aus verschiedenen Measures, jeweils mit Koordinaten in den jeweiligen Dimensionen
- Koordinaten (Dimensionen) sind hierarchisch gegliedert (entlang der Pfade); jeder Datenpunkt hat eine Koordinate in jeder Klassifikationsstufe
  - Die Koordinaten in einem Pfad sind voneinander abhängig
- Würfel stellen aggregierte Daten gemäß der Würfelgranularität dar
- Aggregation und Verfeinerung sind (nur) entlang der Pfade möglich und verändern die Granularität und damit den Aggregationslevel der Fakten

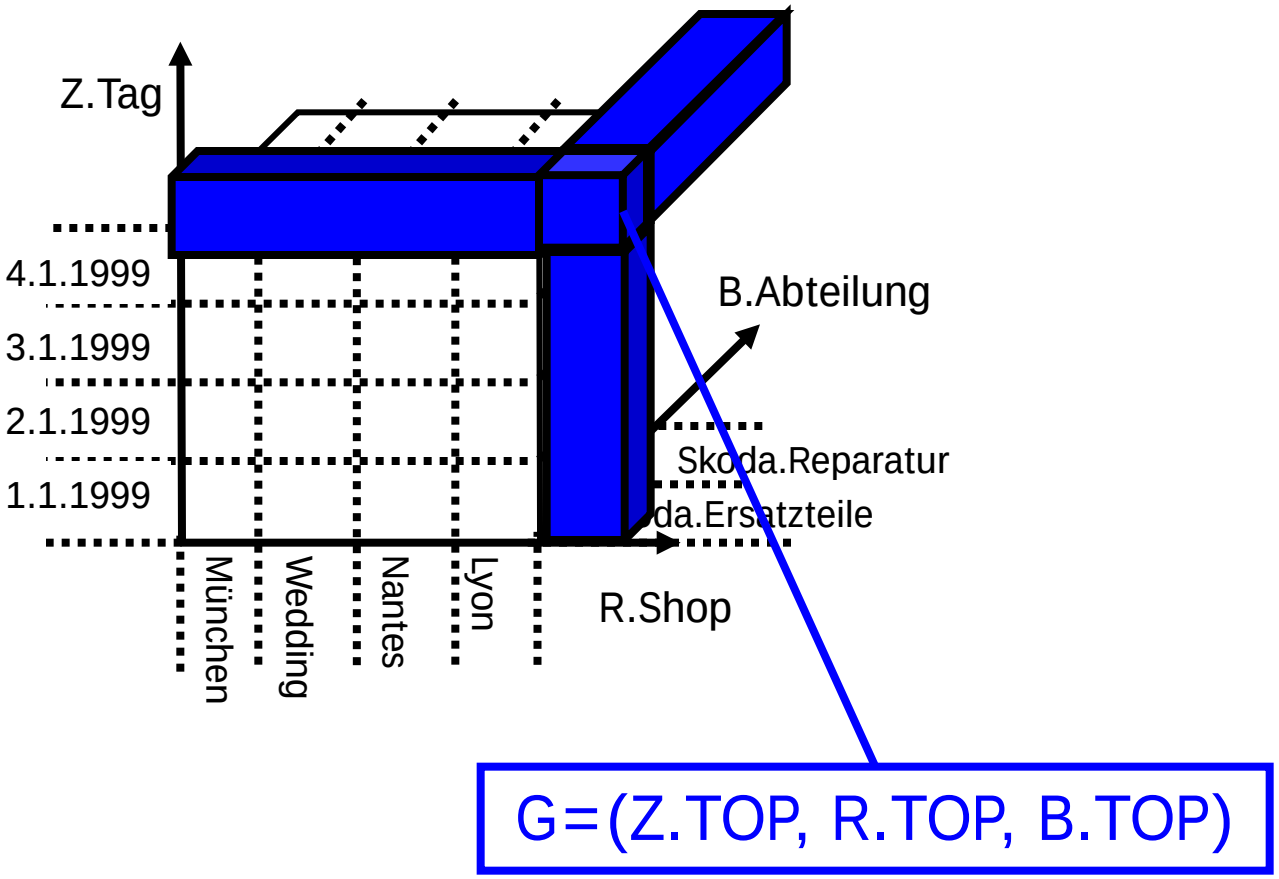
Aggregation bis TOP

Summe Umsatz pro Tag und Abteilungen über alle Shops



Summe Umsatz pro Shop und Tag, über alle Abteilungen

... in mehreren Dimensionen



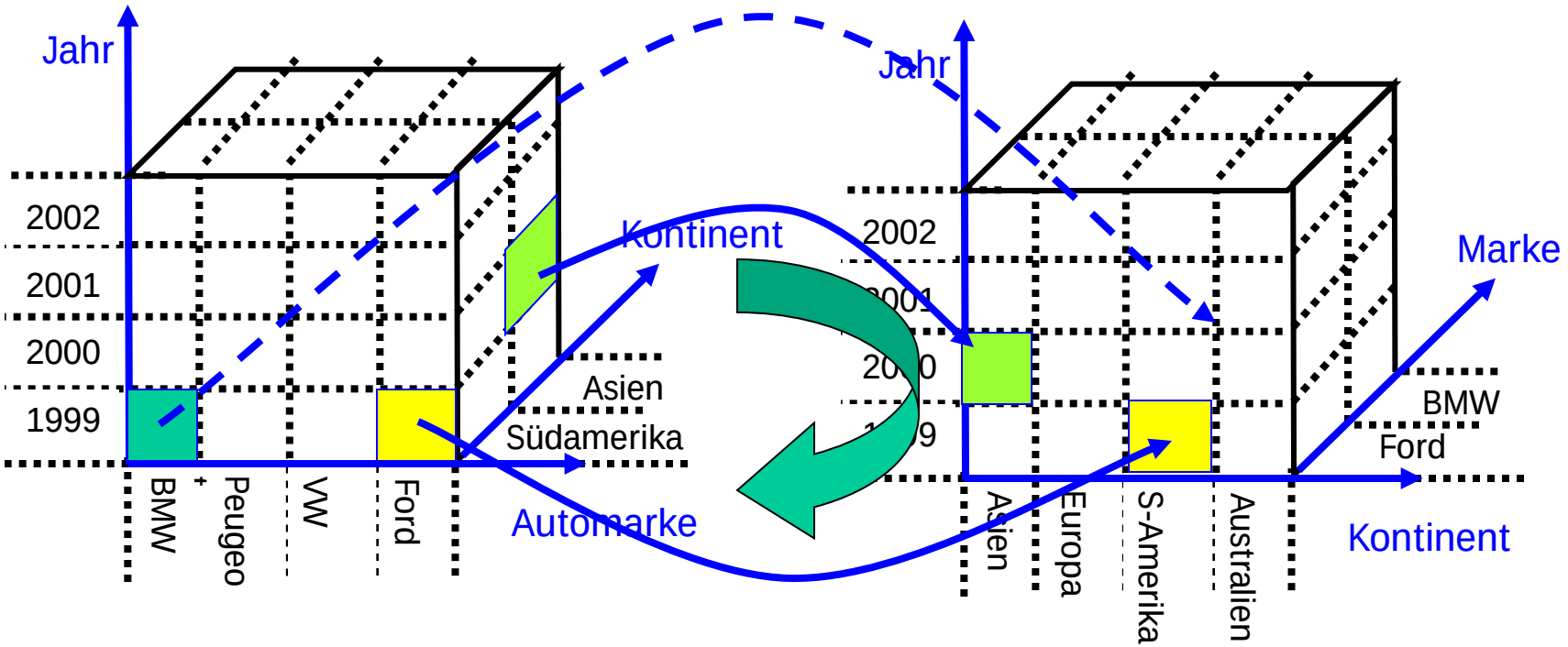
Inhalt dieser Vorlesung

- Operationen im multidimensionalen Datenmodell
  - Aggregation
  - Verfeinerung
  - Weitere Operationen
- ME/R: Graphische multidimensionale Datenmodellierung
- Aggregierbarkeit

Weitere Operationen

- Aggregation: Technisch interessanteste OLAP Operation
- Weitere führen keine Berechnungen durch, sondern selektieren und visualisieren Ausschnitte des Würfels
  - Hat keinen Einfluss auf die Granularität, aber auf die Datenbasis
  - Vermischung von Präsentations- und Datenschicht
- In der Literatur werden diverse weitere Operationen angeführt (und tw. widersprüchlich definiert)
  - Drill-Across: Zugriff über mehrere Würfel hinweg
  - Drill-Through: Zugriff auf Daten „unter“ dem Würfel
  - ...

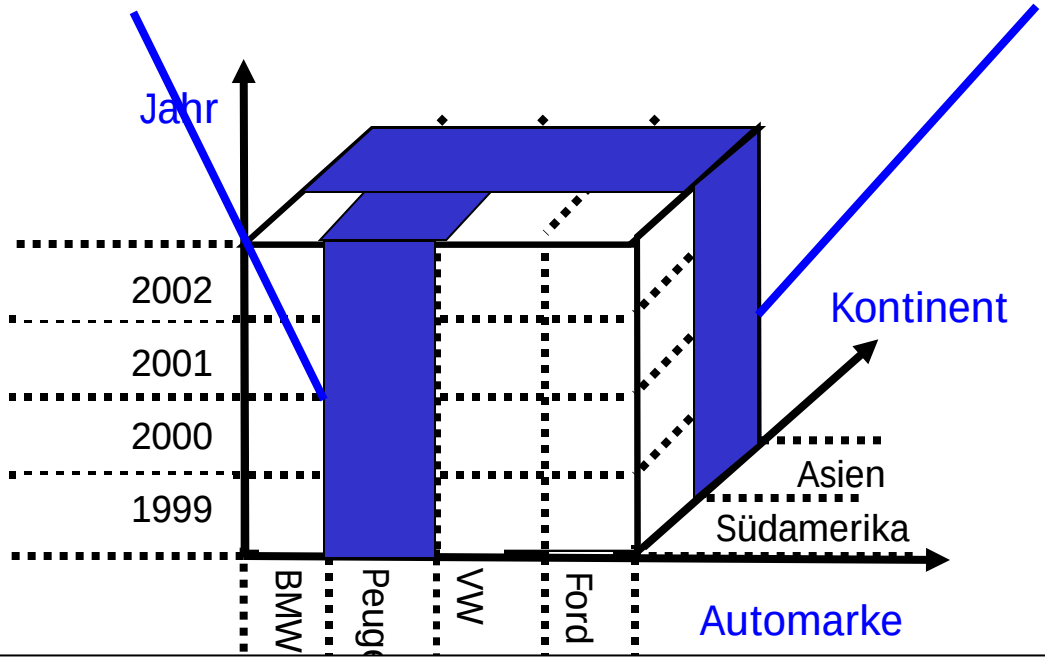
Rotation (Pivoting)



Selektion einer Scheibe (Slicing)

Verkäufe von Peugeot pro Jahr und Kontinent

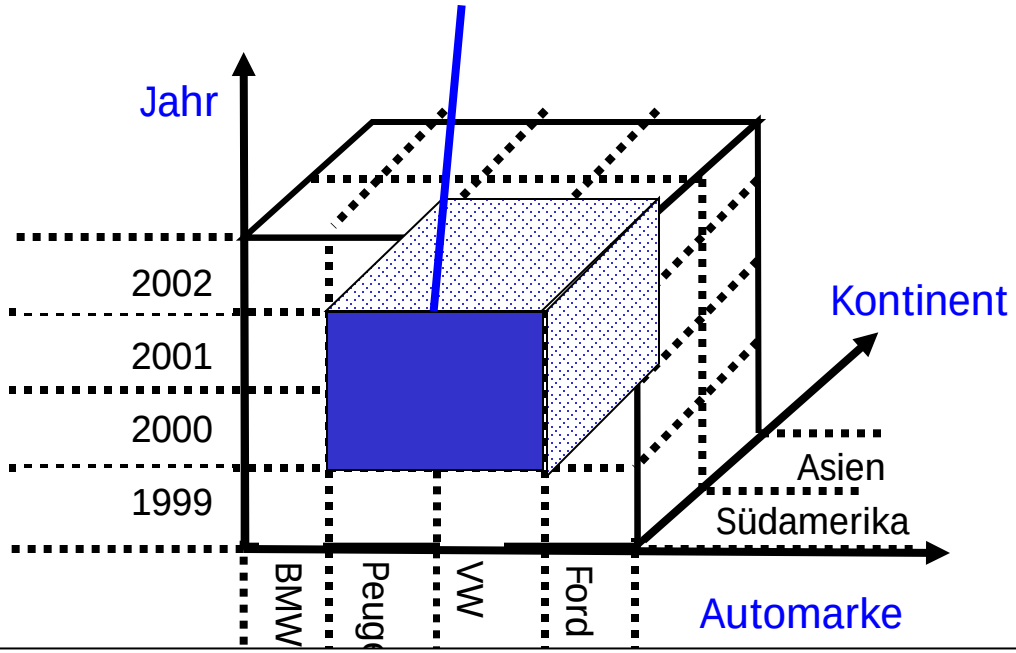
Verkäufe in Asien pro Jahr und Marke



Formal: Granularität verliert eine Dimension, punkte aller Stufen enthalten immer Schnitt mit festem Wert in dieser Dimension

Auswahl von Unterwürfeln (Dicing)

Verkäufe von (Peugeot, VW) in (2000, 2001) pro Kontinent

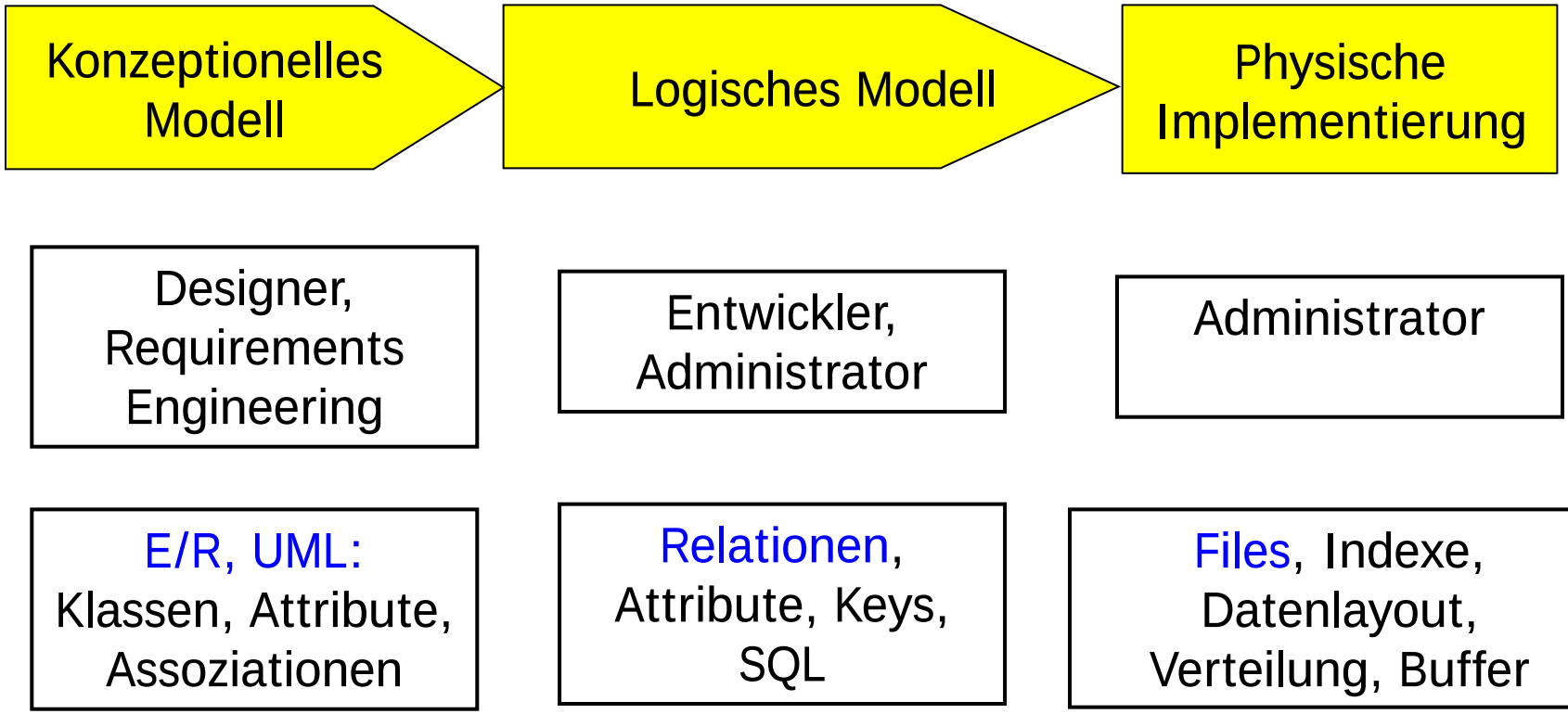


Formal: Granularität verliert mehrere Dimensionen, punkte aller Stufen enthalten immer Schnitt mit festen Werten in diesen Dimensionen

Inhalt dieser Vorlesung

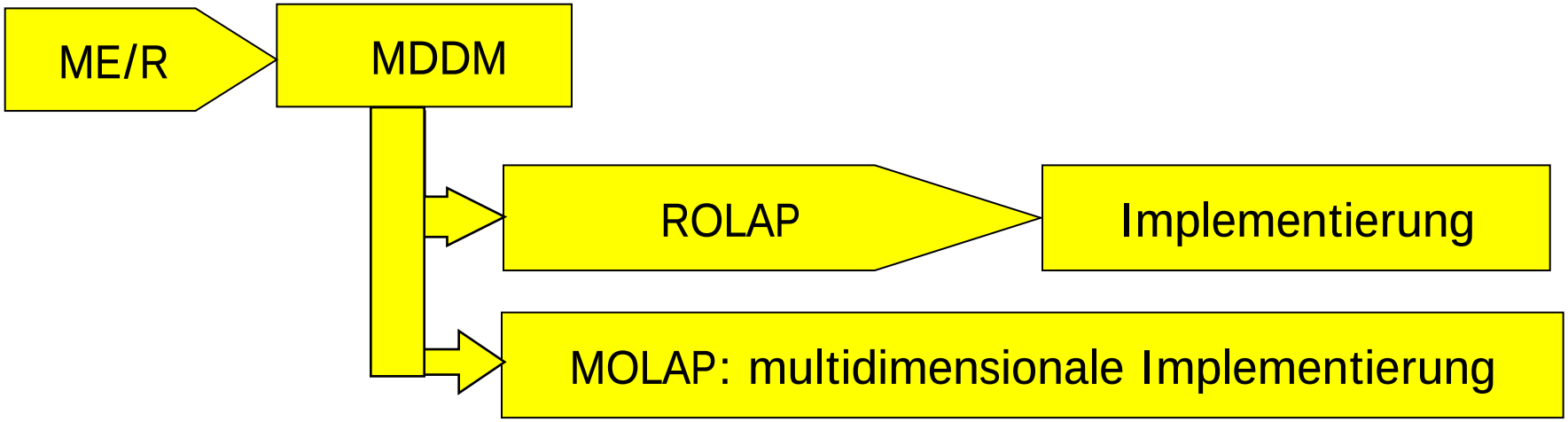
- Operationen im multidimensionalen Datenmodell
- ME/R: Graphische multidimensionale Datenmodellierung
- Aggregierbarkeit

Datenbankentwurf



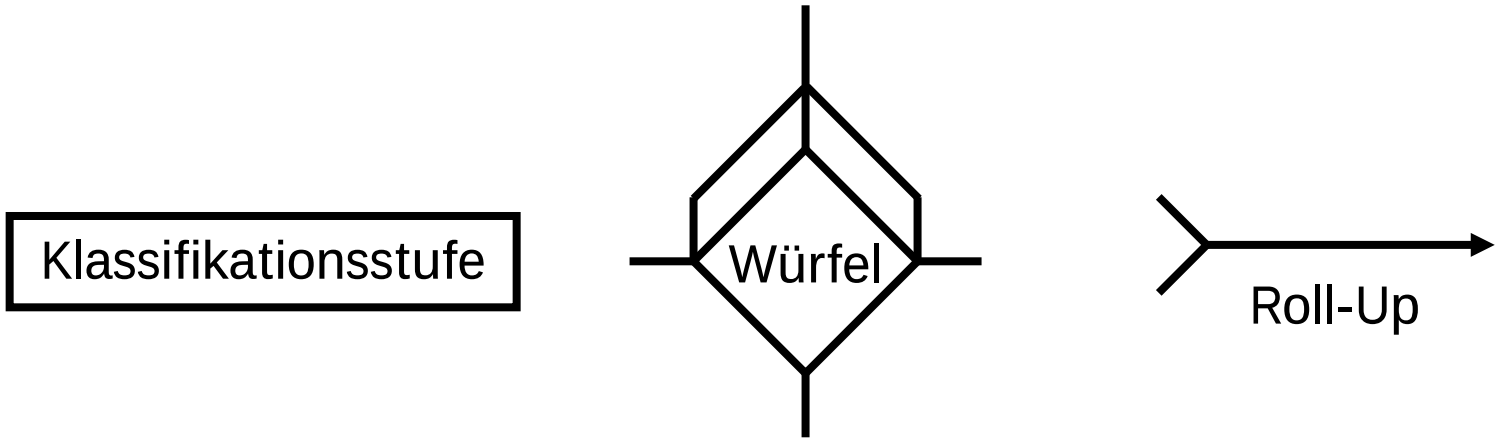
ME/R [SPHD98]

- Unser MDDM ist ein logisches, formales Modell
  - Fakten, Klassifikationsstufen, Dimensionen, ...
- Ein konzeptionelles Modell fehlt noch
  - E/R nicht ausreichend – keine Hierarchien, Dimensionen ...
- ME/R: Erweitertes E/R Modell

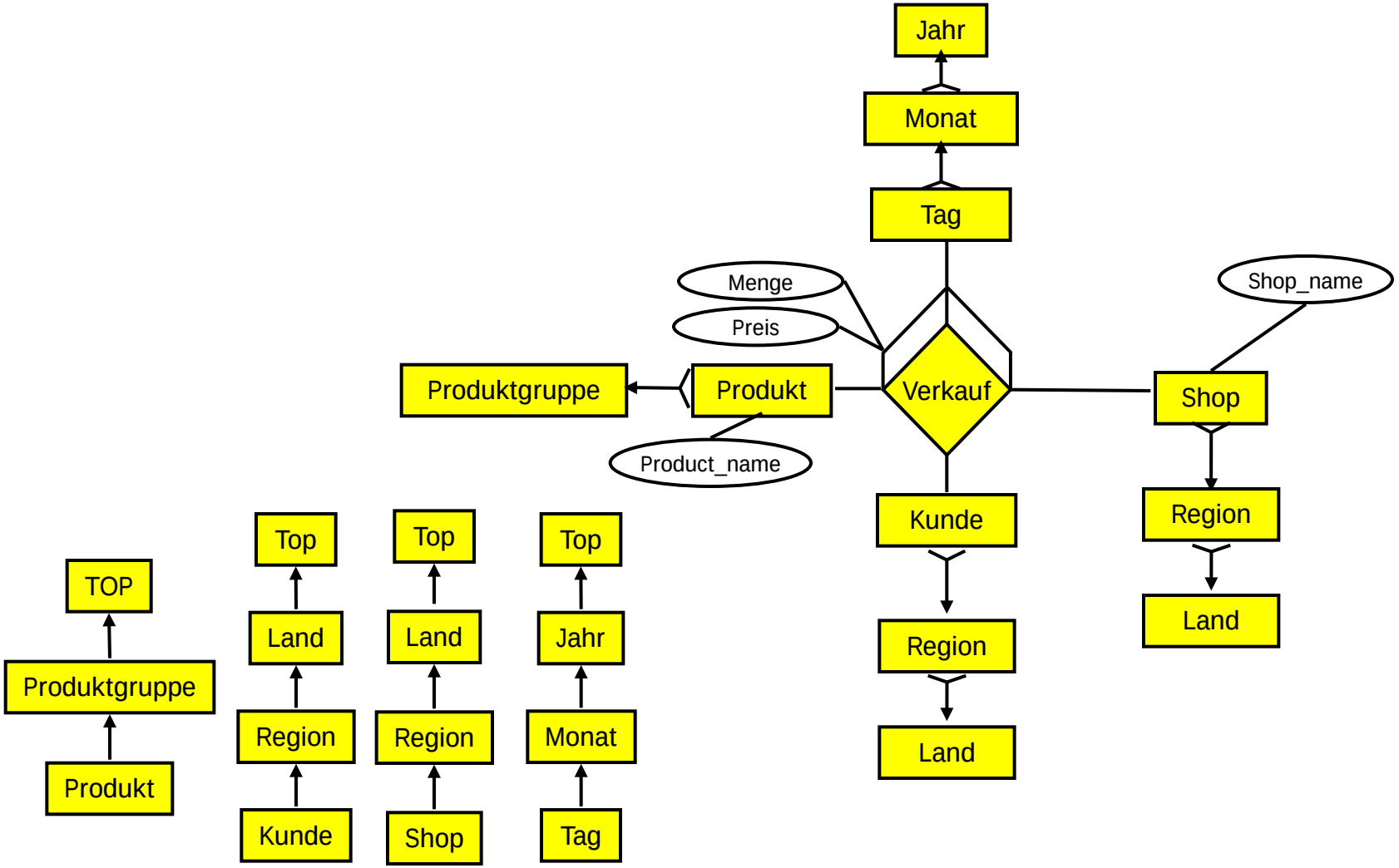


ME/R Elemente und Notation

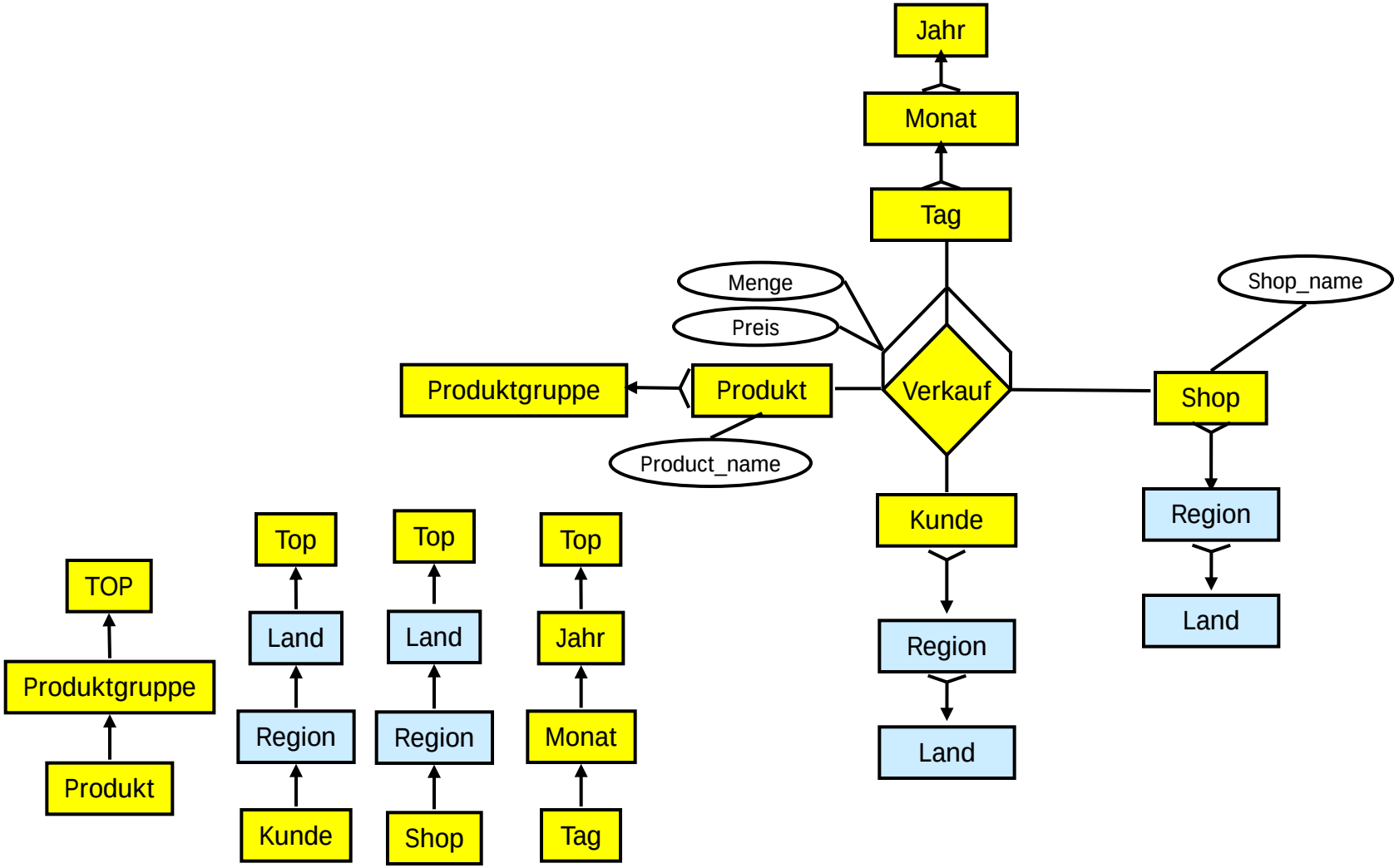
- Ausgangspunkt: Das klassische E/R Modell
- Was fehlt?
  - Klassifikationsstufen (und deren Instanzen, die K-Knoten)
  - Würfel (Granularitäten sind nicht Teil des Datenmodells)
  - Halbordnung zwischen Klassifikationsstufen (Roll-Ups)
  - Dimensionen werden in ME/R nicht gesondert definiert
- Constraint: Keine Zyklen in den ROLL-UP Beziehungen



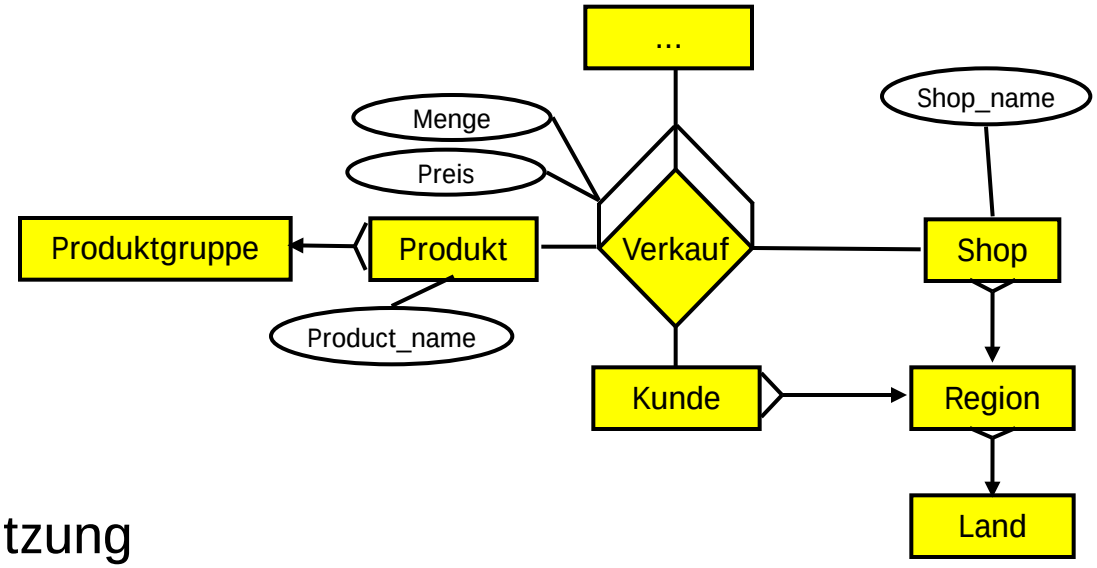
Beispiel


Jahr
Monat
Tag
Menge
Preis
Shop_name
Produktgruppe
Produkt
Verkauf
Shop
Product_name
Kunde
Region
Land
Region
Land
TOP
Produktgruppe
Produkt
Top
Land
Region
Kunde
Top
Land
Region
Shop
Top
Jahr
Monat
Tag

Beispiel


Jahr
Monat
Tag
Menge
Preis
Shop_name
Produktgruppe
Produkt
Verkauf
Shop
Product_name
Kunde
Region
Land
Region
Land
TOP
Produktgruppe
Produkt
Top
Land
Region
Kunde
Top
Land
Region
Shop
Top
Jahr
Monat
Tag

„Wiederverwendung“ von Stufen



- Umsetzung
  - Kunden-Regionen müssen gleich organisiert sein wie Shop-Regionen
  - Fremdschlüssel in Kunde und Shop verweisen auf PK von Region
- ... aber nicht unproblematisch (in MDDM war das verboten)
  - Vermischung von Dimensionen
  - Eventuell implizite Abhängigkeiten
    - Wenn Kunden in mehr Regionen wohnen dürfen als es Shops gibt
    - Aggregation über Shops bis auf Region gibt viele leere Regionen

ME/R Bewertung

- Minimale, konservative Erweiterung
- Mehrere Würfel sind möglich
- Klare Semantik durch Metamodell in Extended E/R
- Greift eher kurz
  - Keine Eigenschaften von Fakten (Summierbarkeit)
  - Keine non-standard Klassifikationshierarchien
    - Z.B. unbalancierte Hierarchien – siehe später
  - Keine expliziten Dimensionen
- Toolunterstützung?

Weitere Ansätze

- mUML
  - UML-Erweiterung basierend auf UML Metamodell (Constraints, Stereotypen) -> Werkzeugunterstützung vorhanden
  - Klassifikationsstufe: <<Dimensional class>>
  - Fakten: <<Fakt class>>
  - Würfel: <<Dimension>>
  - Hierarchie: <<Roll-Up>>
  - ...
- Multi-dimensional Modelling Language (MML)
- ...
- Keine der Methoden hat sich (bisher) durchgesetzt

Inhalt dieser Vorlesung

- Operationen im multidimensionalen Datenmodell
- ME/R: Graphische multidimensionale Datenmodellierung
- Aggregierbarkeit
  - Klassen von Aggregatfunktionen
  - Überlappungsfreiheit und Vollständigkeit in Hierarchien
  - Typverträglichkeit

Aggregierbarkeit

- Verdichtung (Aggregation) geht nicht immer
- Probleme sind
  - Numerische versus kategoriale Fakten
    - Numerisch: Umsatz, Verkäufe, Messwerte, ...
    - Kategorial: Geschlecht, Kundensegment,...
      - Kann man als Dimension modellieren, muss man aber nicht
  - Snapshot-Fakten
  - Nicht alle Aggregatfunktionen sind hierarchisch anwendbar

<table>
<tr><th>Level 0</th><th>Level 1</th><th>Level 2</th></tr>
<tr><td>1</td><td rowspan="2">Avg: 1</td><td rowspan="3">Avg: 5,5 (?)</td></tr>
<tr><td>1</td></tr>
<tr><td>10</td><td>Avg: 10</td></tr>
</table>

Klassen von Aggregatfunktionen [LS97]

- Definition
  Geg. Menge X und eine disjunkte Partitionierung (X_1 ,X_2, …, X_n) von X. Eine Aggregatfunktion f heißt:
  - *distributiv gdw $\exists g$: $f(X) = g(f(X_1), f(X_2), \ldots f(X_n))$*
  - *algebraisch gdw f(X) berechenbar aus fester Menge von g‘s*
    - *Deren Zahl und Art unabhängig von X ist*
  - *holistisch gdw f(X) kann nur aus den Grundelementen von X berechnet werden*
    - Menge von g‘s nur durch |X| begrenzt
- Bemerkungen
  - X: Klassifikationsknoten, (X_1 ,X_2, …, X_n) seine Kindern
  - Bei holistischen Funktionen ist Prä-“aggregation“ unmöglich, bei algebraischen speicherplatzintensiv

Beispiele

Distributiv	
Algebraisch	
Holistisch	

Median

Sum

Max

STDDEV

Percentile

Top-N

AVG

Count

Highest-Frequency

Rank

Beispiele

Distributiv	Sum, Count (Sum von counts), Max, Min, ...
Algebraisch	Avg (mit g_1=Sum und g_2=Count), STDDEV, TopN, ...
Holistisch	Median, Rank, Percentile, Highest Frequency, ...

- Highest Frequency
  - Merke Werte und jeweilige Frequenz: ((v_1,f_1), (v_2,f_2), ...)
  - Merge zweier Sets möglich
  - Man braucht also nicht |X| Platz, sondern nur |distinct(X)|
  - Aber: Keine feste Grenze für Platzbedarf, da Anzahl unterschiedlicher Werte nicht fest

Aggregierbarkeit

- Wann darf man Werte hierarchisch aggregieren?
  - Klar: Nur numerische Fakten
  - Subtiler: Art der Aggregatfunktion beachten
- Das reicht im Allgemeinen nicht als Bedingung
  - Summe der Lagerbestände pro Produkt über Jahre?
  - Gesamtsumme Studenten als Summe über Studenten pro Studiengang?
- Weitere notwendige Kriterien
  - Typverträglich von Fakt und Aggregatfunktion
  - Überlappungsfreiheit der Zuordnung von Klassifikationsknoten
  - Vollständigkeit der Zerlegung pro Klassifikationslevel

Typen von Fakten

- Flow (Ereignis zu Zeitpunkt T)
  - Verkäufe, Umsatz, Lieferungen, diplomierte Studenten, ...
- Stock (Zustand zu Zeitpunkt T)
  - Lagerbestand, eingeschriebene Studenten, Einwohnerzahlen, ...
- Value-per-Unit (Eigenschaft zu Zeitpunkt T)
  - Preis, Einkaufspreis, Währungskurs, ...

Typverträglichkeit

	Stock	Flow	Value-per-Unit
MIN/MAX	√	√	√
SUM	Zeit: nein Sonst: √	√	Nie
AVG	√	√	√

- Zuordnung muss beim Design erfolgen
  - Metadaten – Beschreibung der Measures

Beispiel

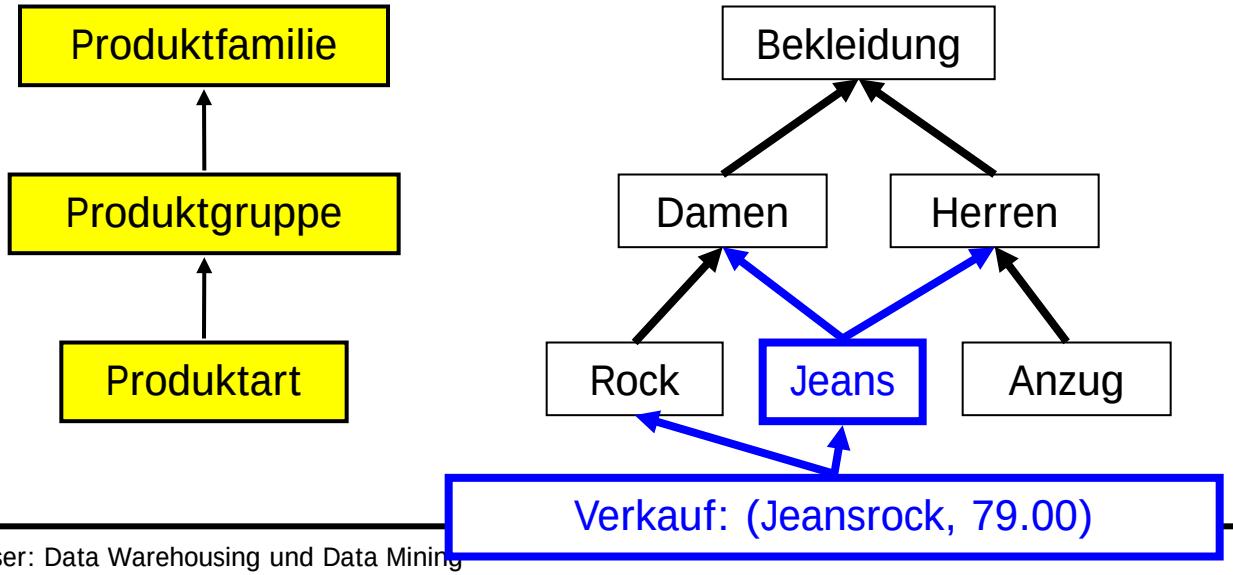
Aktuelle Studentenzahl nach Jahr der Einschreibung und Studiengang

	1994	1995	1996	Gesamt
Informatik	15	17	13	28
BWL	10	15	11	21
Gesamt	25	31	23	49

- Wie kann das stimmen?
  - Wie lange dauern Studiengänge?
  - Studenten nur in einem Studiengang eingeschrieben?

Überlappungsfreiheit

- Definition.
  Eine Klassifikationshierarchie ist überlappungsfrei gdw
  - *jeder Klassifikationsknoten mit Level i höchstens einem Klassifikationsknoten in Level i+1 zugeordnet ist*
  - *jeder Datenpunkt höchstens einem Klassifikationsknoten mit Level 0 zugeordnet ist*



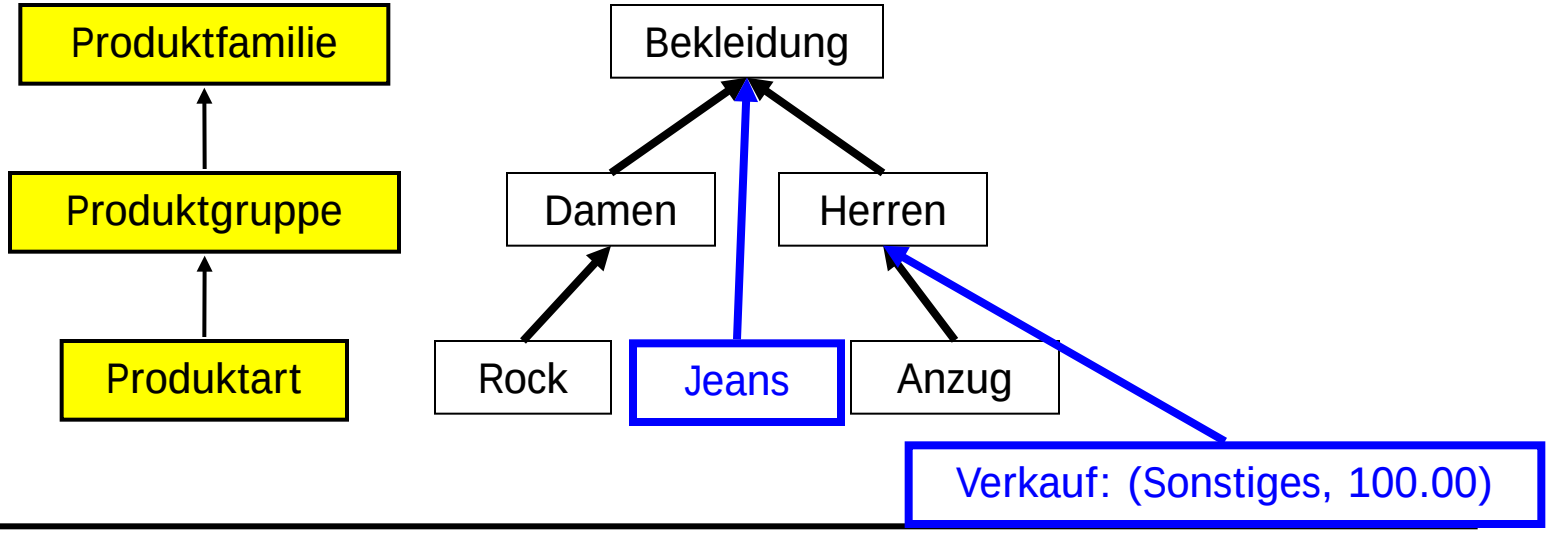
Vollständigkeit

- Definition

  Eine Klassifikationshierarchie ist vollständig gdw

  – *jeder Klassifikationsknoten mit Level i mindestens einem Klassifikationsknoten in Level i+1 zugeordnet ist*

  – *jeder Datenpunkt mindestens einem Klassifikationsknoten mit Level 0 zugeordnet ist*



Was tun?

- MDDM verlangt vollständige, überlappungsfreie Klassifikationshierarchien
- In der Praxis gibt es oft Abweichungen
- Typische Abhilfen
  - Artifizielle Klassifikationsknoten
    - Virtuelle Knoten: „Others“, „Rest“, „Nicht zugewiesen“
    - Knotenaufspaltung: „Herrenjeans“, „Damenjeans“
  - Gewichtete Zuordnung
    - Türkei ist 10% Europa, 90% Asien

Literatur

- Lenz, H.-J. and Shoshani, A. (1997). "Summarizability in OLAP and Statistical Databases". 9th International Conference on Scientific and Statistical Database Management, Olympia, Washington. pp 132-143.
- Vassiliadis, P. (1998). "Modeling Multidimensional Databases, Cubes, and Cube Operations". 10th International Conference on Scientific and Statistical Database Management, Capri, Italy. pp 53-62.
- Sapia, C., Blaschka, M., Höfling, G. and Dinter, B. (1998). "Extending the E/R Model for the Multidimensional Paradigm". Workshop on Data Warehousing and Data Mining, Singapore. pp 105-116.

Selbsttest

- Nennen Sie die wichtigsten OLAP Operationen
- Modellieren Sie die folgende Situation ... mit ME/R
- Was ist eine holistische Aggregationsfunktion? Was macht eine Aggregationsfunktion holistisch?
- Ist die folgende Funktion ... algebraisch, distributiv oder holistisch? Warum?
- Nehmen wir einen Würfel mit drei Dimensionen an, jeweils ein Pfad mit drei Stufen, und eine Granularität dazu mit feinster Stufe in jeder Dimension. Wie viele einfache, wie viele komplexe Roll-Ups gibt es?