

Data Warehousing und Data Mining

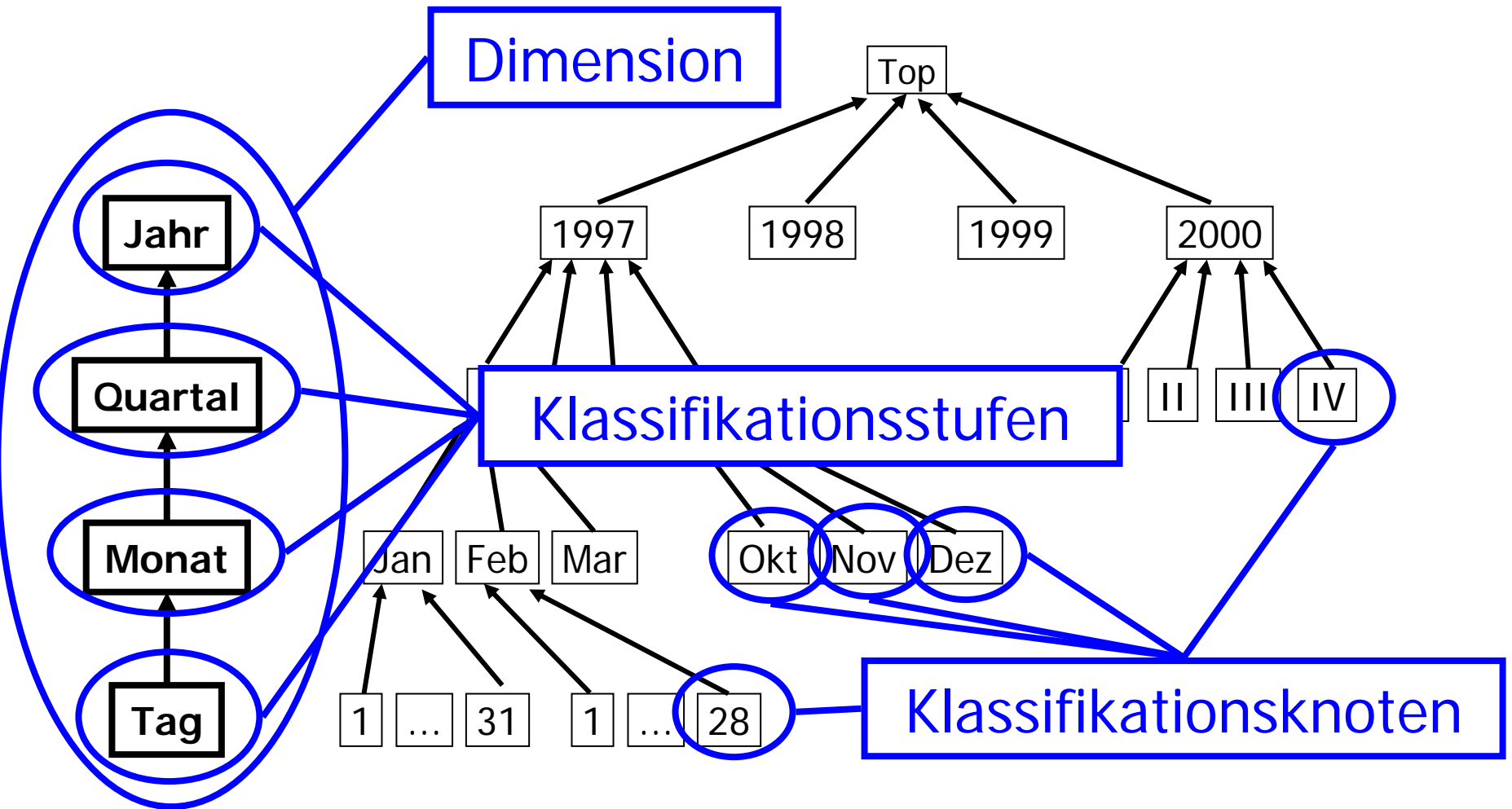
OLAP Operationen
Graphische MDDM

Ulf Leser

Wissensmanagement in der
Bioinformatik



Dimension



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 $k \rightarrow_k l$ 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

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

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

Halbordnung auf Granularitäten

- Definition

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

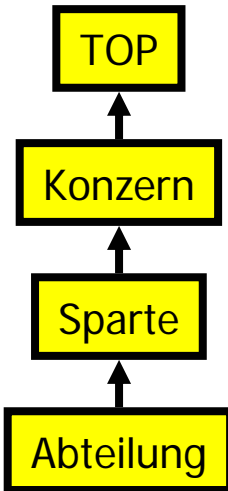
- Sei $G_1 = \{D_1.k_1^1, \dots, D_n.k_n^1\}$ und $G_2 = \{D_1.k_1^2, \dots, D_n.k_n^2\}$
- Es gilt $G_1 \leq G_2$ genau dann wenn
 - $\forall i: D_i.k_i^1 \rightarrow D_i.k_i^2$
 - (Dies impliziert als Möglichkeit auch $D_i.k_i^1 = D_i.k_i^2$)

- Benutzung

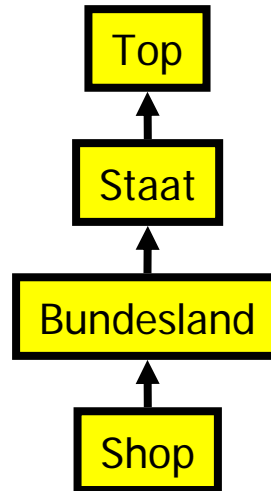
- Beschreibung der Transformation von Granularitäten
- Anfrageoptimierung: **Wiederverwendung von Aggregaten**

Beispiel

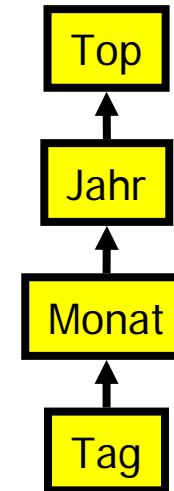
Bereich



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

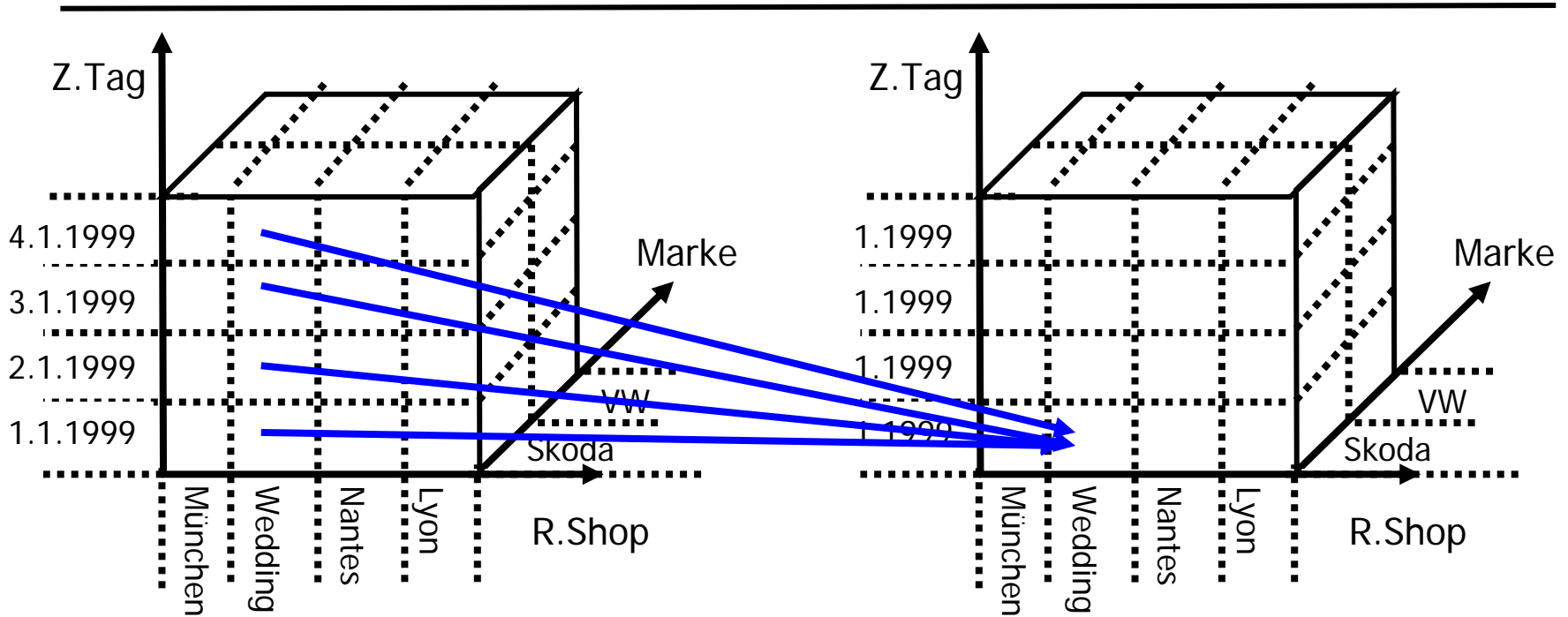
Inhalt dieser Vorlesung

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

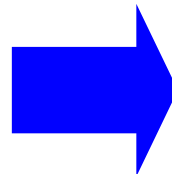
OLAP Operationen

- MDDM entspricht der Sprache des Betriebswirts
 - Umgang mit Objekten seiner täglichen Welt (Kunden, Waren, etc.)
 - Keine technischen Details
- Operationen auf dem MDDM müssen Analysevorgänge abbilden bzw. unterstützen
 - Schnelle und intuitive Manipulation der Daten
 - Abbildung der Aufgaben von Betriebswirten
- Die wichtigsten Operationen
 - Aggregation (Roll-Up): Granularität wird erniedrigt
 - Verfeinerung (Drill-Down): Granularität wird erhöht
 - Aggregation bzgl. einer Funktion f wie SUM, AVG, MEDIAN, ...
 - Und viele mehr, die später kommen ...

Übersicht



(1.101, 1.1.1999, Wedding, Skoda)
 (129, 2.1.1999, Wedding, Skoda)
 (225, 3.1.1999, Wedding, Skoda)
 (1.540, 4.1.1999, Wedding, Skoda)
 (2.500, 5.1.1999, Nantes, Skoda)
 ...



(3.000, 1.1999, Wedding, Skoda)
 (2.500, 2.1999, Nantes, Skoda)
 ...

Rollup Tag→Monat

Datenpunkte und Würfelkoordinaten

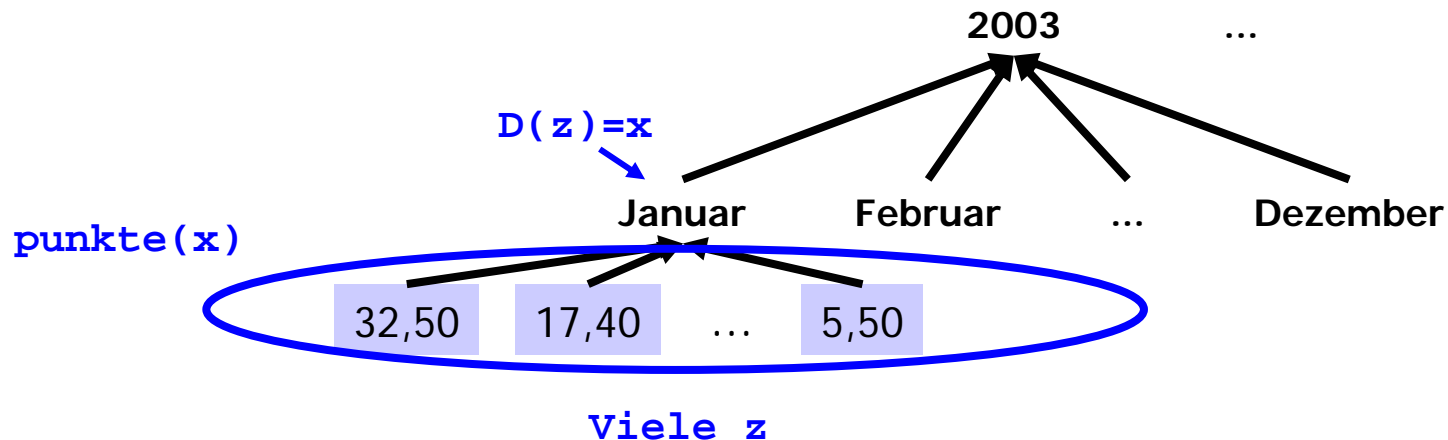
- Definition

Gegeben ein Würfel W , eine Dimension D und ein Klassifikationsknoten x aus D . Sei z ein Datenpunkt (Fakt) und $D(z)$ seine Koordinate bzgl. D . Dann gilt

- *z liegt in x gdw. $D(z)=x$ oder $D(z) \in \text{nachfahren}(x)$*
- *$\text{punkte}(x)$ ist die Menge aller Datenpunkte z mit $D(z)=x$*

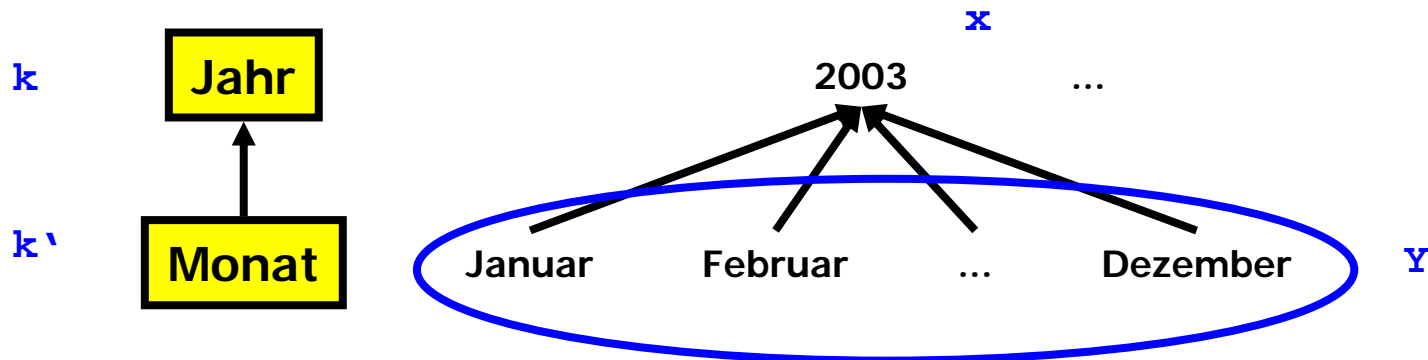
- Bemerkung

- Erweiterung auf mehrere Koordinaten durch Schnittmengenbildung
 - $\text{punkte}(x,y,z) = \text{punkte}(x) \cap \text{punkte}(y) \cap \text{punkte}(z)$
mit x in D_1 , y in D_2 , z in D_3



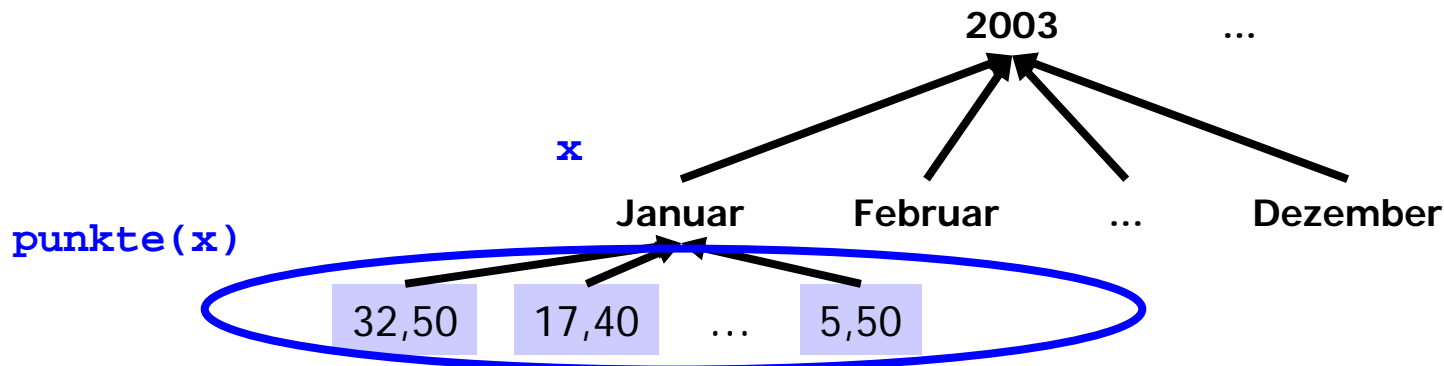
Aggregation in Hierarchien

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



Startpunkt

- Sei $P = \{k_0 \rightarrow \dots \rightarrow k_n \rightarrow \text{TOP}\}$
- Man berechnet $F(\text{TOP})$ aus $F(k_n)$, $F(k_n)$ aus $F(k_{n-1})$, etc.
- **Wie berechnet man aber $F(k_0)$?**
 - Aus den einzelnen Datenpunkten
- **Definition**
 - *Gegeben*
 - Dimension D , Pfad $P = \{k_0 \rightarrow \dots \rightarrow \text{TOP}\}$ in D
 - Ein Klassifikationsknoten $x \in \text{knoten}(k_0)$
 - Aggregatfunktion f , Measure F
 - Sei $\text{punkte}(x) = \{z_1, \dots, z_n\}$ die Menge aller Datenpunkte mit $D(z_i)$ in x
 - Dann berechnet sich das aggregierte Fakt $F(x)$ als $F(x) = f(F(z_1), \dots, F(z_n))$
 - Mit $F(z)$ als dem Wert des Measures F von Datenpunkt z



Berechnung

- Die letzte Definition ist prinzipiell abwendbar für alle Klassifikationsknoten bzw. –stufen eines Pfades
 - Also können wir $F(x)$ für x aus beliebiger Klassifikationsstufe aus den Datenpunkten berechnen
- Aber
 - Es schneller, die Aggregation aus der nächst-feineren Stufe zu berechnen
 - Weniger Werte - weniger Arbeit
 - Das setzt voraus, dass man die auch hat – Präaggregation
 - Geht auch nicht für alle Aggregatfunktionen - Summierbarkeit
 - OLAP Operationen wandern immer nur eine Stufe in einem Pfad herauf oder herab – deshalb definieren wir Aggregation von Ebene zu Ebene

Wo sind wir?

- Wir können entlang eines Pfades in einer Dimension für ein Measure mit einer Aggregatfunktion die aggregierten Fakten für höherstufige Knoten berechnen
- Zwei Möglichkeiten
 - Aggregierte Measures von Knoten einer Stufe aus aggregierten Measures der nächstkleineren Stufe berechnen
 - Verlangt Präaggregation
 - Aggregierte Measures immer direkt aus den Werten aller Datenpunkte mit den entsprechenden Koordinaten berechnen
 - Ist eine Anfrage über dem Cube höchster Granularität

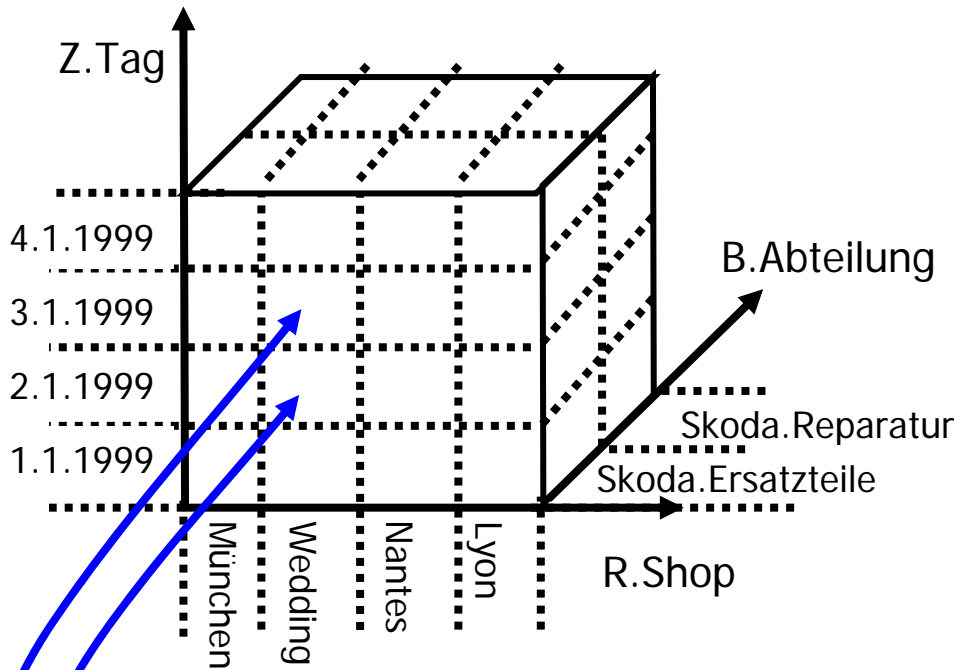
Würfelinhalt

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

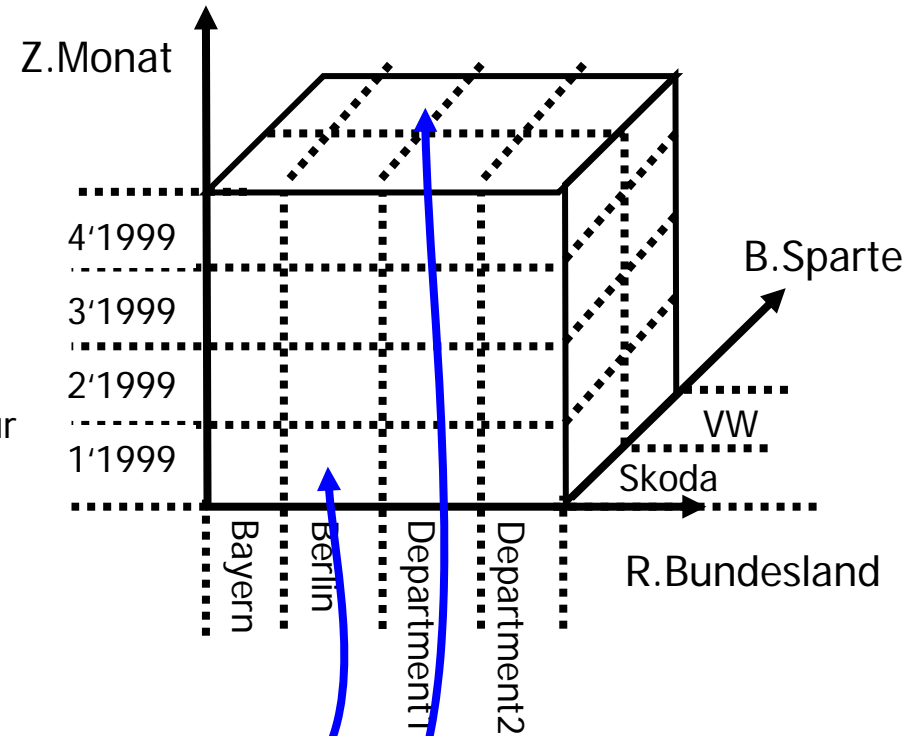
Würfelinhalt, allgemeiner Fall

- Ein Würfel $W = (G, F)$
 - Granularität $G = (D_1.k_1, \dots, D_n.k_n)$
 - Menge Measures $F = \{F_1, \dots, F_m\}$ mit Aggregatfunktionen f_1, \dots, f_m
- Allgemeiner Fall
 - $W(x_1, \dots, x_n) = (F_1, \dots, F_m) =$
 $(F_1(\text{punkte}(x_1) \cap \text{punkte}(x_2) \cap \dots \cap \text{punkte}(x_n)),$
 $F_2(\text{punkte}(x_1) \cap \text{punkte}(x_2) \cap \dots \cap \text{punkte}(x_n)),$
 \dots
 $F_m(\text{punkte}(x_1) \cap \text{punkte}(x_2) \cap \dots \cap \text{punkte}(x_n)))$
 - mit Koordinaten $x_i \in \text{knoten}(k_i)$
- Bemerkung
 - Die Schnittmengenbildung berechnet die Menge von Punkten in der Würfelzelle mit Koordinaten x_1, \dots, x_n
 - Jedes Measure wird gesondert aggregiert

Beispiel



Koordinaten
der Fakten



Aggregierte
Fakten

(3.1.1999, Wedding, Ersatzt.) = (...)
(2.1.1999, Wedding, Ersatzt.) = (...)

(1.1999, Berlin, Skoda) = (...)
(4.1999, Bayern, VW) = (...)

Operationen auf Würfeln

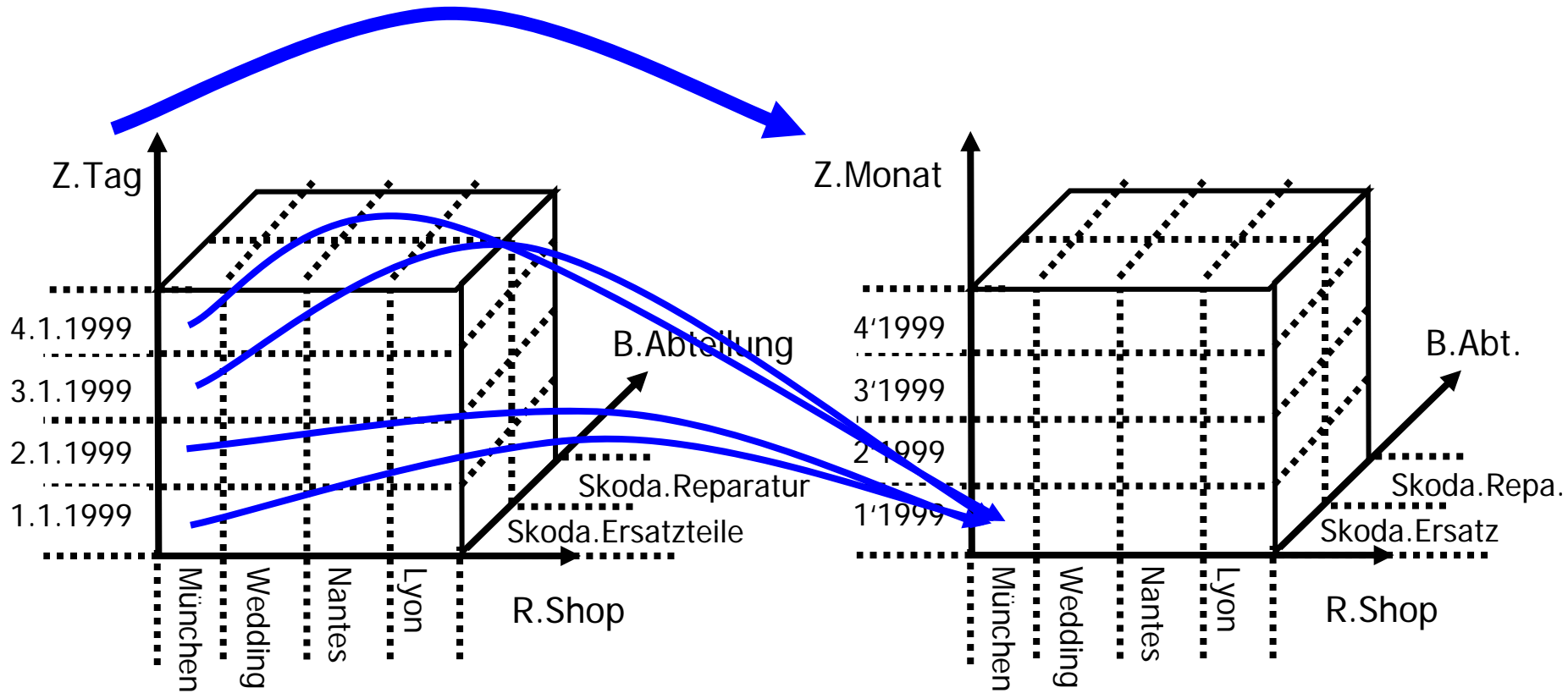
- **OLAP Operationen** überführen einen Würfel $W=(G,F)$ in einen Würfel $W'=(G',F')$
- Dabei gilt
 - Aggregation: $G < G'$
 - Verfeinerung $G > G'$
- Eine **einfache Operation** verändert nur die Klassifikationsstufe einer Dimension in G
 - Komplexe Operationen können auf natürliche Weise aus einfachen 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 $G=(D_1.k_1, \dots, D_i.k_i, \dots, D_n.k_n)$ und $F=(F_1, \dots, F_m)$
- Sei $P=\{k_0 \rightarrow \dots \rightarrow k_{i''} \rightarrow k_i \rightarrow k_{i'} \rightarrow \dots \rightarrow TOP\}$ ein Pfad in D_i
- Die *einfache Aggregation in W entlang P mit Aggregatfunktion f* überführt W in
 - $W' = ((D_1.k_1, \dots, D_i.k_{i''}, \dots, D_n.k_n), (F_{1'}, \dots, F_{m'}))$
 - $(F_{1'}, \dots, F_{m'}) = (F_1(\text{punkte}(k_1) \cap \dots \text{punkte}(k_{i'}) \cap \dots \text{punkte}(k_n)), \dots, F_m(\text{punkte}(k_1) \cap \dots \text{punkte}(k_{i'}) \cap \dots \text{punkte}(k_n)))$

Beispiel Aggregation



Verfeinerung (Drill-Down)

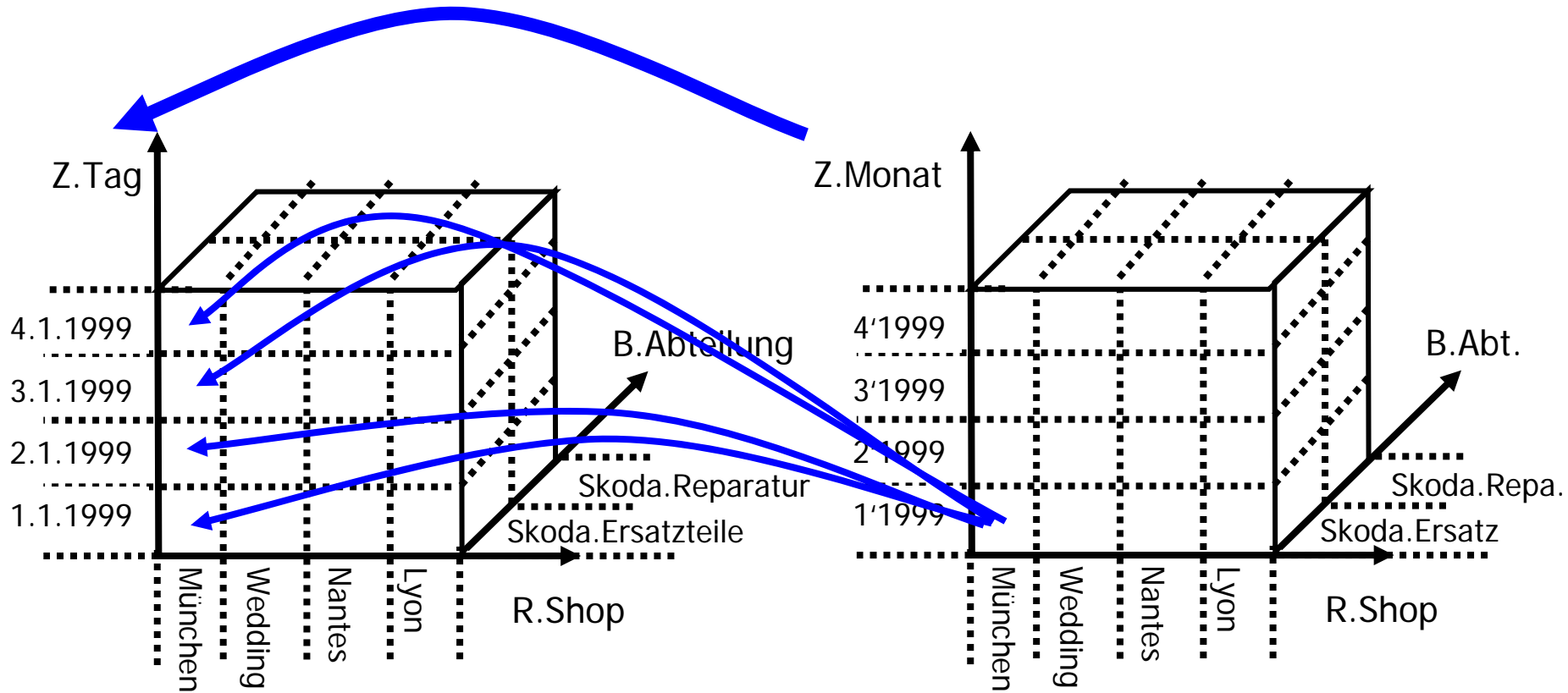
- Definition

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

- Bemerkung

- Bei Verfeinerung ist ein Zugriff auf die Datenbank (Datenpunkte oder Präaggregate) notwendig
- Ein „Aufbrechen“ einmal aggregierter Daten ist für die meisten Aggregatfunktionen nicht möglich

Beispiel Verfeinerung

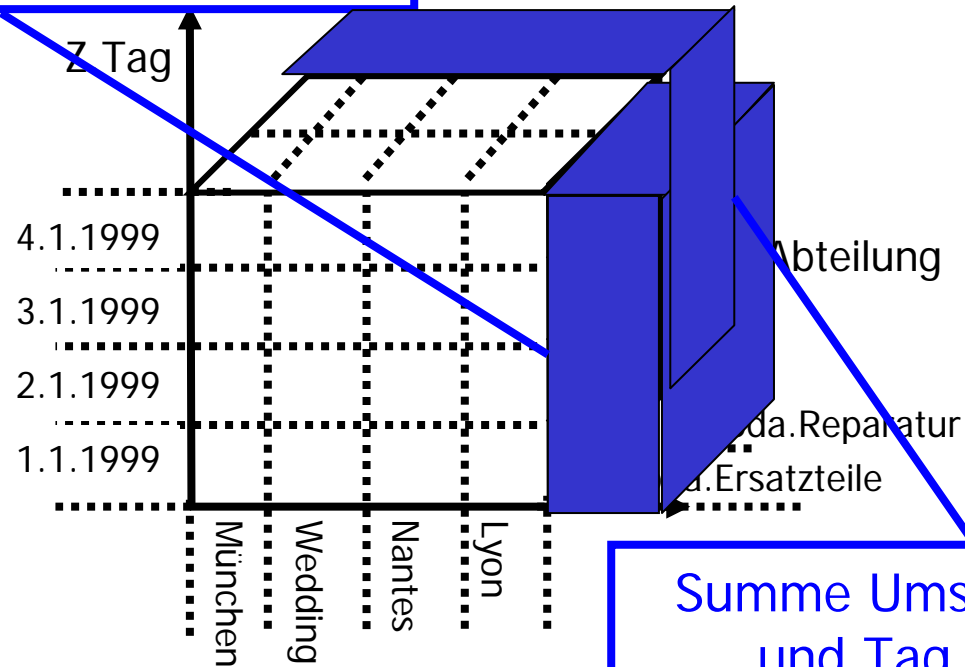


Auf einen Blick

- Das MDDM gibt Klassifikationsstufen, Knoten, Pfade und Dimensionen vor
- 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
- OLAP Operationen (Aggregation, Verfeinerung) sind entlang der Pfade möglich
- Sie verändern die Granularität und damit den Aggregationslevel der Fakten

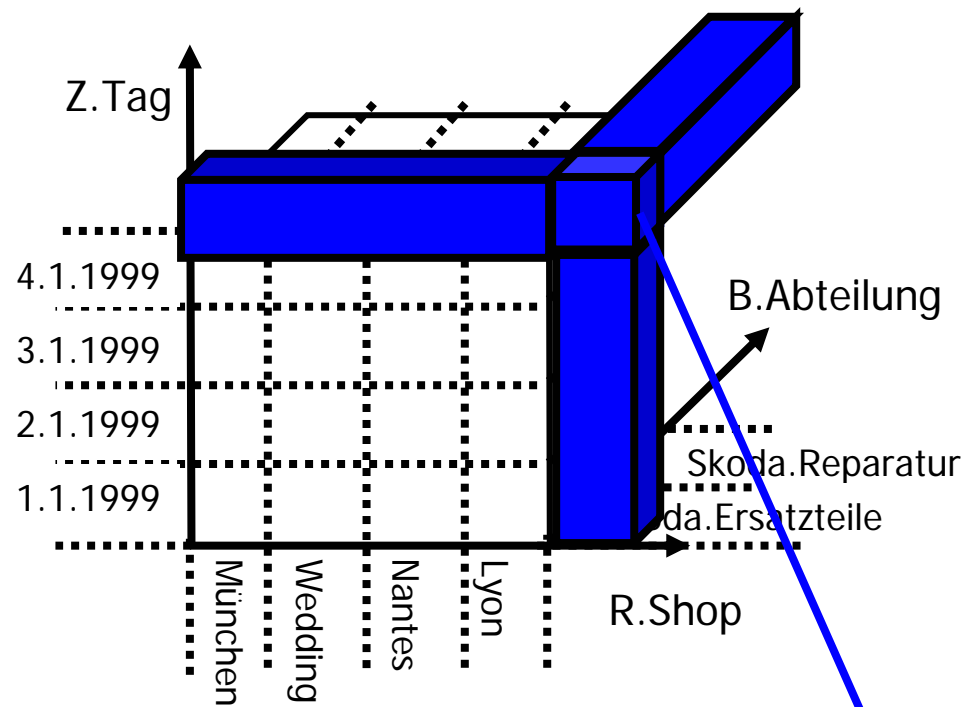
Aggregation bis TOP

Summe Umsatz pro Tag und
Abteilungen über alle Shops



Summe Umsatz pro Shop
und Tag, über alle
Abteilungen

... in mehreren Dimensionen



$$G = (Z.TOP, R.TOP, B.TOP)$$

Inhalt dieser Vorlesung

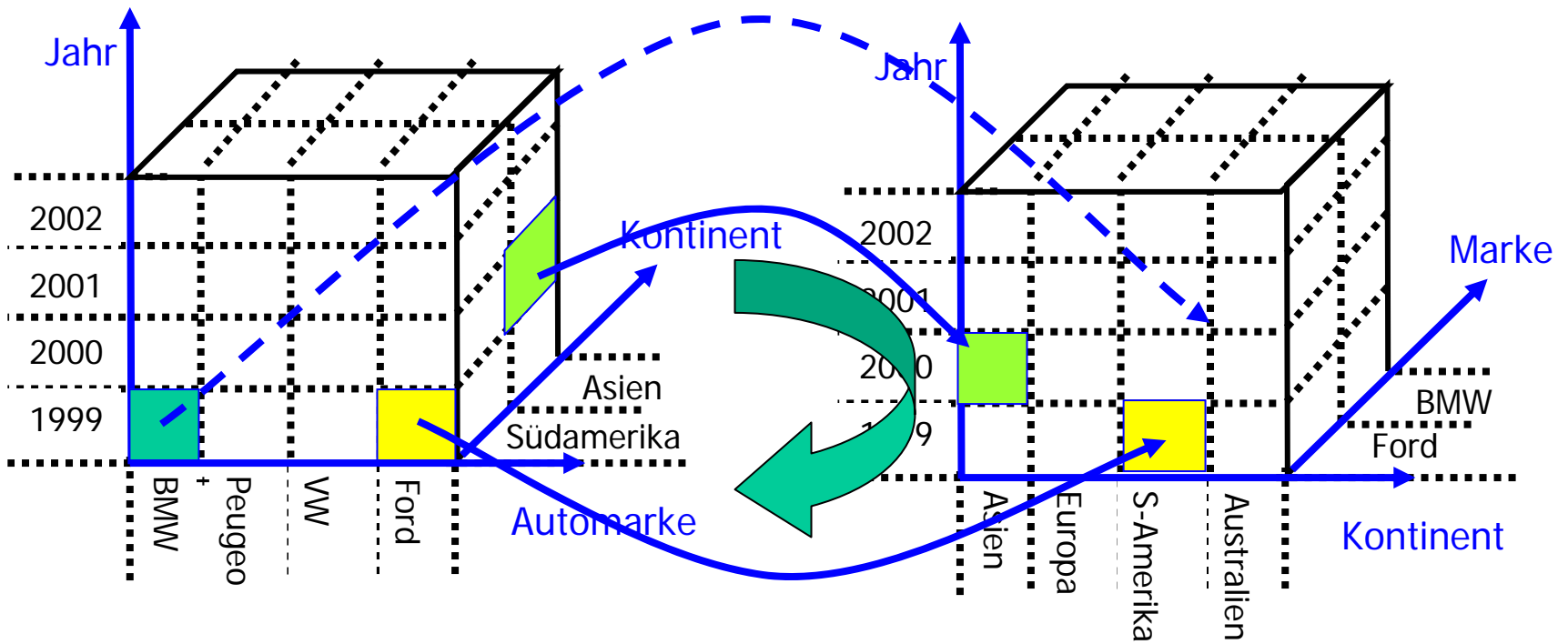
- Operationen im multidimensionalen Datenmodell
 - Aggregation
 - Verfeinerung
 - Weitere Operationen
- ME/R: Graphische multidimensionale Datenmodellierung
- Summierbarkeit

Weitere Operationen

- Aggregation ist die spannendste OLAP Operation
- Die restlichen führen keine Berechnungen durch, sondern **selektieren bzw. ändern** Ausschnitte des Würfels
 - Hat keinen Einfluss auf die Granularität
 - Teilweise wird nur die Darstellung (Visualisierung) beeinflusst, ohne Daten zu ändern
 - Vermischung von Präsentations- und Datenschicht
- In der Literatur werden noch **weitere Operationen** angeführt (und tw widersprüchlich definiert)
 - Drill-Across: Zugriff über mehrere Würfel hinweg
 - Drill-Through: Zugriff auf Daten „unter“ dem Würfel
 - ...

Rotation (Pivoting)

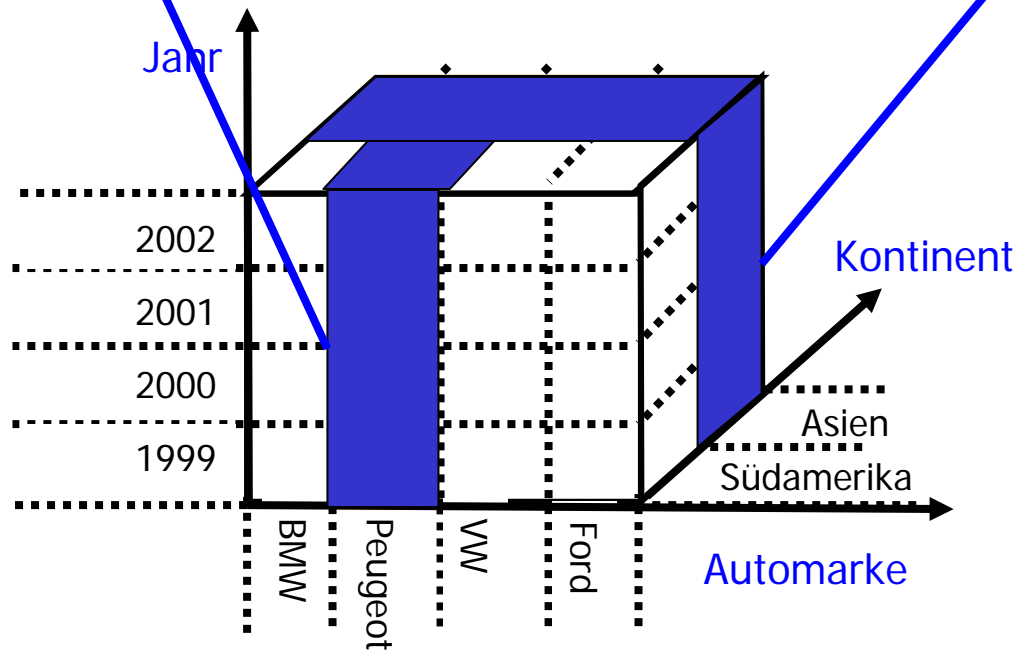
Unterschiedliche Sichtweisen auf einen Datenbestand



Selektion einer Scheibe (Slicing)

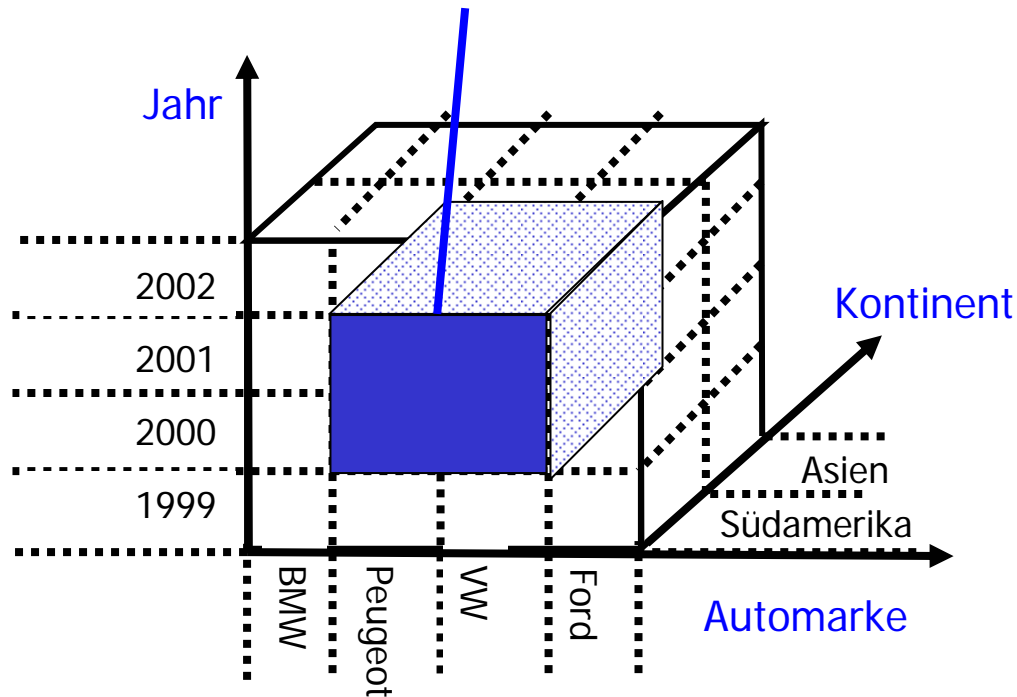
Verkäufe von Peugeot pro Jahr und Kontinent

Verkäufe in Asien pro Jahr und Marke



Auswahl von Unterwürfeln (Dicing)

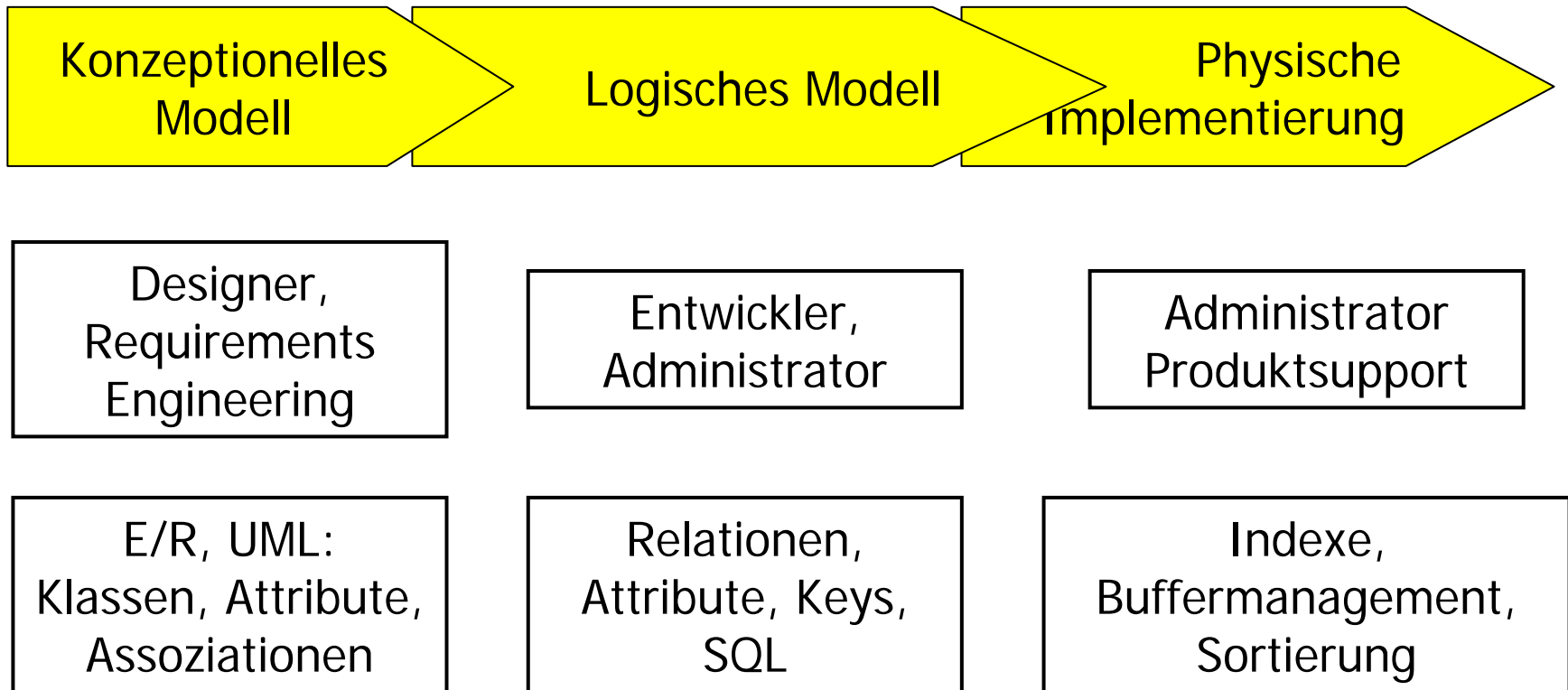
Verkäufe von (Peugeot, VW) in
(2000, 2001) pro Kontinent



Inhalt dieser Vorlesung

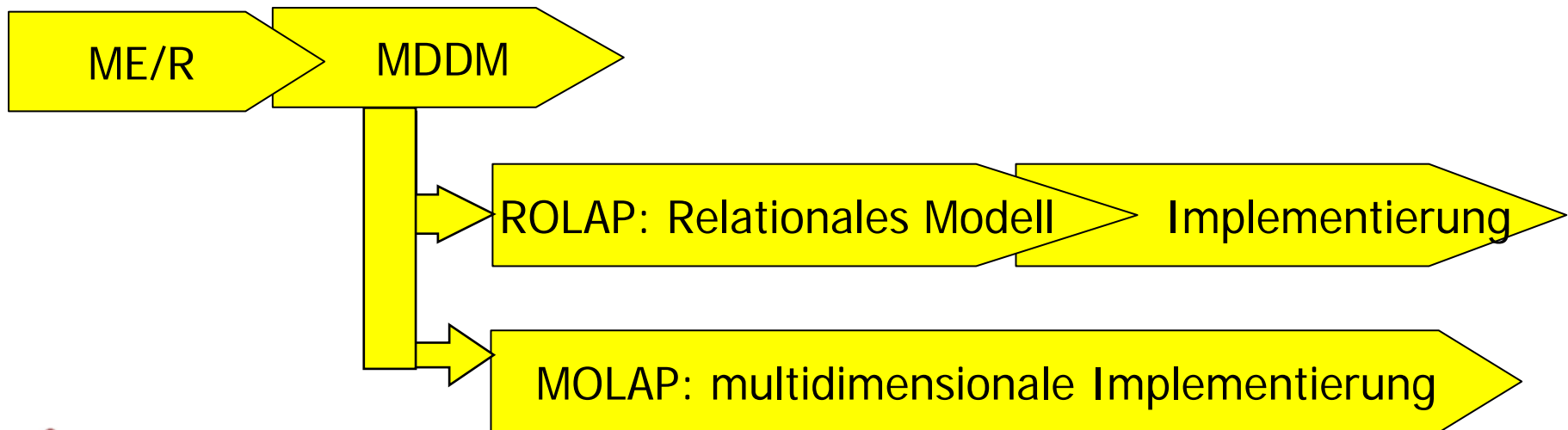
- Operationen im multidimensionalen Datenmodell
- ME/R: Graphische multidimensionale Datenmodellierung
- Summierbarkeit

Datenbankentwurf



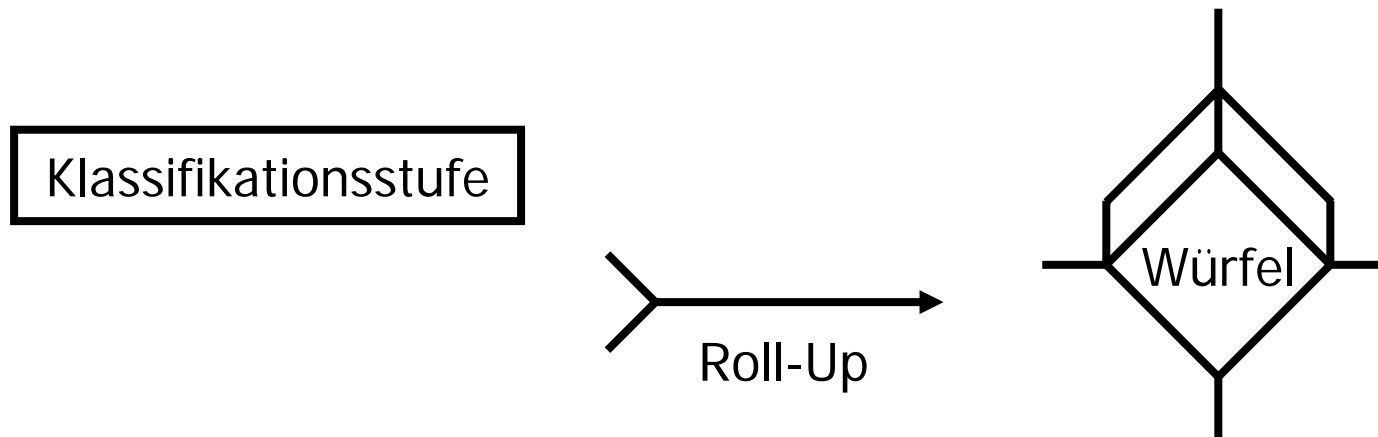
ME/R [SPHD98]

- Logisches MDDM
 - Fakten, Klassifikationsstufen, Dimensionen, ...
- Warum ein graphisches Modell für MDDM?
- E/R nicht ausreichend
 - Keine Repräsentation von MDDM Konzepten
- **ME/R: Erweitertes E/R Modell**

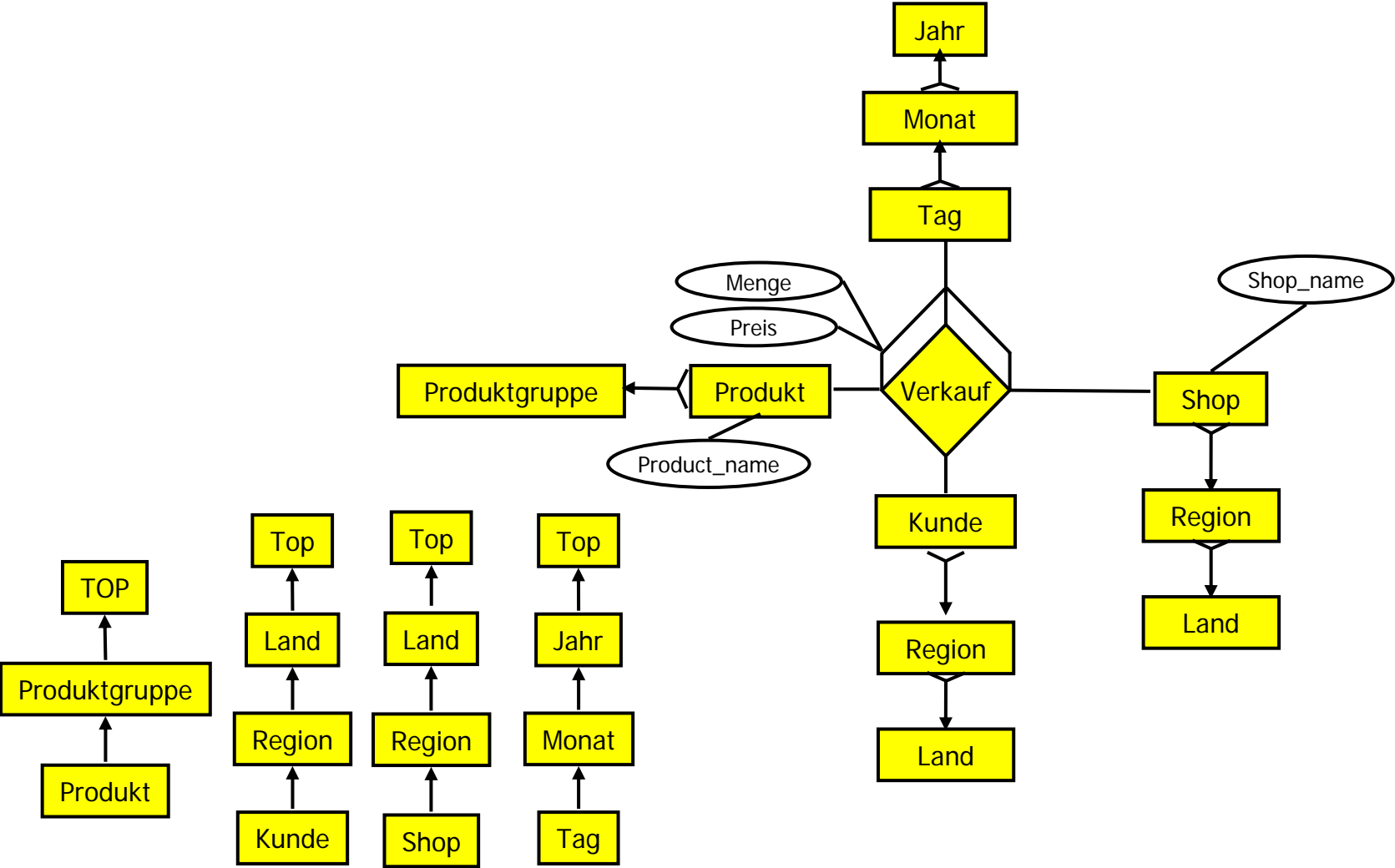


ME/R Elemente und Notation

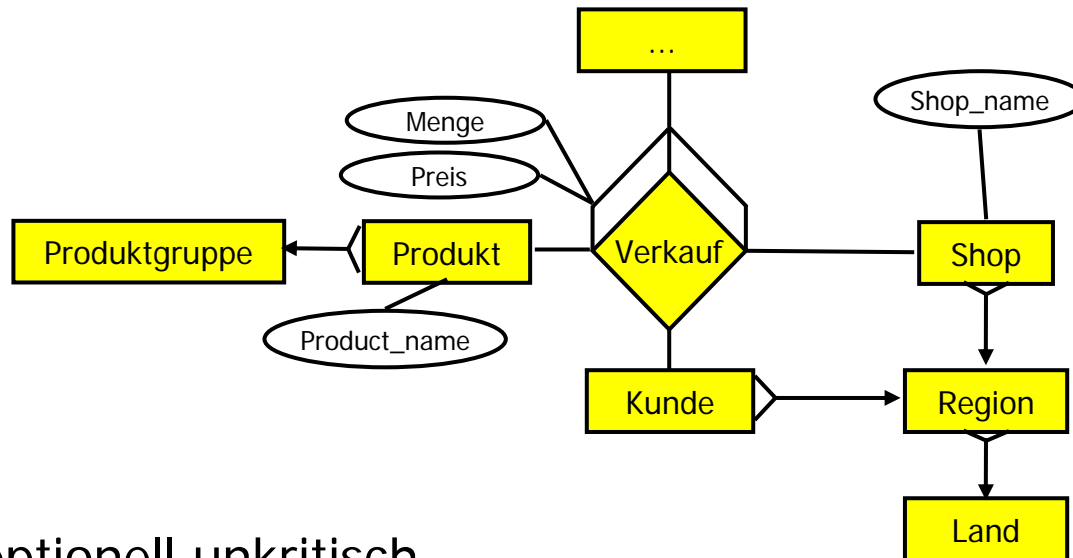
- Neue Elemente
 - Klassifikationsstufe (Dimension Level)
 - Würfel (Fact)
 - Halbordnung zwischen Klassifikationsstufen (Roll-Ups)
- Constraint: **Keine Zyklen** in den ROLL-UP Beziehungen



Beispiel



„Wiederverwendung“ von Stufen



- Konzeptionell unkritisch
 - Fremdschlüssel in Kunde und Shop verweisen auf PK von Region
- ... aber **nicht unproblematisch**
 - Aufbrechen der klaren Dimensionsstruktur
 - Wenn Kundenregionen (nach Vertrieb) anders organisiert sind als Shopregionen (nach Logistik) – IC schwierig zu überwachen
 - Wenn Kunden in mehr Regionen wohnen dürfen als es Shops gibt – „verborgenes“ Integritätsconstraint
 - Aggregation über Shops bis auf Region gibt viele leere Regionen

ME/R Bewertung

- Minimale, konservative Erweiterung
- Mehrere Würfel sind möglich
- Flexible Zuordnung Dimensionen – Würfel
- Klare Semantik durch Metamodell in Extended E/R
- Greift eher kurz
 - Keine Eigenschaften von Fakten (Summierbarkeit)
 - Keine Besonderheiten von Klassifikationspfaden
 - Übersprungene Stufen, unendliche Hierarchien
- Keine Übersetzungsmethode definiert
- Toolunterstützung?

Weitere Ansätze

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

Inhalt dieser Vorlesung

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

Summierbarkeit

- Eines der Hauptziele des MDDM
 - Verdichtung von Daten entlang der Klassifikationspfade
- Aber das **geht nicht immer gut**
 - **Numerische versus kategorische** Fakten
 - Numerisch: Umsatz, Verkäufe, Messwerte, ...
 - Kategorisch: Geschlecht, Kundensegment, ...
 - Kann man als Dimension modellieren, muss man aber nicht
 - Nicht alle **Aggregatfunktionen sind hierarchisch anwendbar**
 - Summe der Umsätze aller Tage → Umsatz des Monats
 - Summe der Umsätze aller Monate → Umsatz des Jahres

Level 0	Level 1	Level 2
1	Summe: 2	Summe: 12 Avg: ?
1	Avg: ?	
10	Summe: 10 Avg: ?	

Beispiel

Level 0	Level 1	Level 2
1	Summe: 2	Summe: 12 Avg: $11/2 \neq 12/3$
1	Avg: 1	
10	Summe: 10 Avg: 10	

Klassen von Aggregatfunktionen [LS97]

- Definition

Gegeben eine Menge X und eine Partitionierung (X_1, X_2, \dots, X_n) von X . Eine Aggregatfunktion f heißt:

- *distributiv* gdw $\exists g: f(X) = f(g(X_1), g(X_2), \dots, g(X_n))$
- *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
 - Die Menge von g 's ist dann nur durch die Größe von X begrenzt
 - Keine Präaggregation möglich

- Bemerkungen

- X entspricht einem Klassifikationsknoten, (X_1, X_2, \dots, X_n) seinen Kindern
- Die Definition verallgemeinert offensichtlich für Hierarchien

Beispiele

Distributiv	
Algebraisch	
Holistisch	

Beispiele

Distributiv	Summe, Count , Max, Min, ...
Algebraisch	AVG (mit G_1 =SUM und G_2 =CNT) STDDEV, MaxN, ...
Holistisch	MEDIAN, RANK, PERCENTILE Highest Frequency , ...

- COUNT: Man wendet SUM (als g) auf den Counts an
- Highest Frequency
 - Merke Werte und jeweilige Frequenz: $((v_1, f_1), (v_2, f_2), \dots)$
 - Merge zweier Sets möglich
 - Aber: Keine **feste Grenze für Platzbedarf**, da Anzahl unterschiedlicher Werte nicht fest

Summierbarkeit / Aggregierbarkeit

- Wann darf man Werte überhaupt hierarchisch summieren?
 - Natürlich nur numerische Fakten
 - Art der Aggregatfunktion beachten
- Das reicht nicht als Bedingung
 - Summe der Lagerbestände pro Produkt über Jahre?
 - Gesamtsumme Studenten als Summe über Studenten pro Studiengang?
- Weitere notwendige Kriterien für Aggregierbarkeit entlang eines Klassifikationspfades
 - **Überlappungsfreiheit** der Zuordnung von Klassifikationsknoten
 - **Vollständigkeit** der Zerlegung pro Klassifikationslevel
 - **Typverträglich** von Fakt und Aggregatfunktion

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?

Summen lassen sich nicht aus den Einzelwerten berechnen

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, Herstellungskurs, Währungskurs, ...

Typverträglichkeit

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

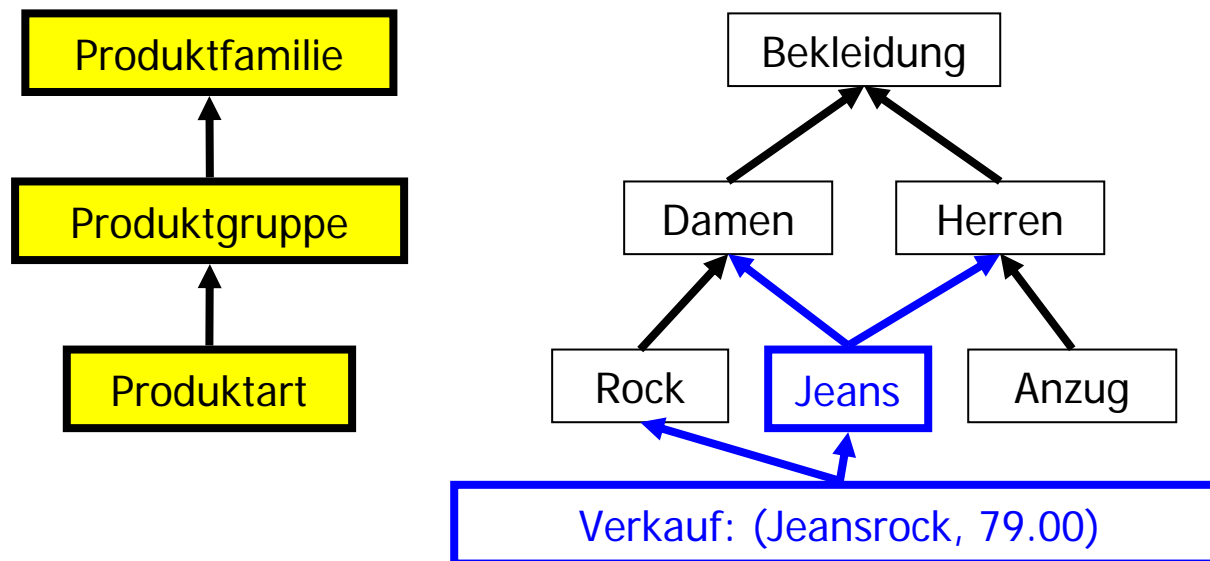
- Erkennen von Problemen
 - Nicht automatisch möglich
 - Metadaten – Beschreibung der Measures notwendig

Überlappungsfreiheit

- Definition

*Eine Klassifikationshierarchie ist **überlappungsfrei** gdw*

- *Jeder Klassifikationsknoten mit Level i ist höchstens einem Klassifikationsknoten in Level $i+1$ zugeordnet*
- *Jeder Datenpunkt ist höchstens einem Klassifikationsknoten mit Level 0 zugeordnet*

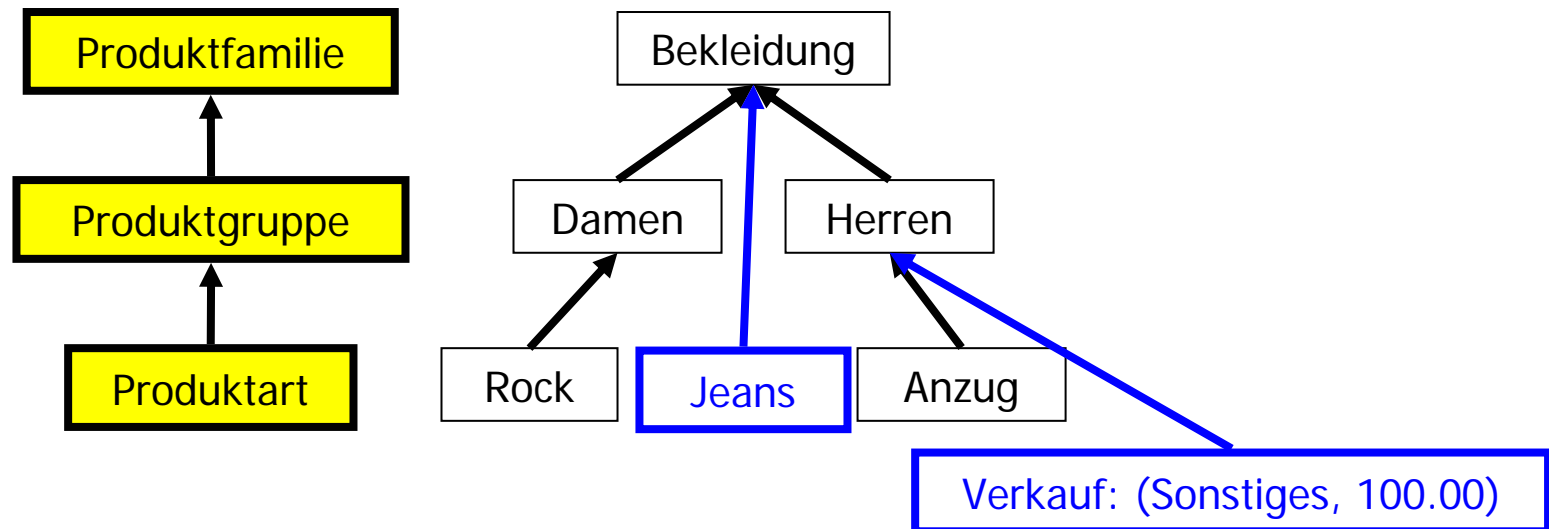


Vollständigkeit

- Definition

*Eine Klassifikationshierarchie ist **vollständig** gdw*

- *Jeder Klassifikationsknoten mit Level i ist mindestens einem Klassifikationsknoten in Level $i+1$ zugeordnet*
- *Jeder Datenpunkt ist mindestens einem Klassifikationsknoten mit Level 0 zugeordnet*



Was tun?

- Wir sind bisher immer von vollständigen und überlappungsfreien Klassifikationsschemas ausgegangen
 - Summierbarkeit ist ein wichtiger Grund dafür
- Das entspricht nicht immer der **betrieblichen Realität**
- Was tun wenn nicht?
 - **Neue Klassifikationsknoten**
 - Artifizielle Klassifikationsknoten: „Others“, „Rest“, „Nicht zugewiesen“
 - Knotenaufspaltung: „Herrenjeans“, „Damenjeans“
 - **Gewichtete Zuordnung**
 - Türkei ist 10% Europa, 90% Asien

Literatur

- [SL97] 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.
- [Vas98] Vassiliadis, P. (1998). "Modeling Multidimensional Databases, Cubes, and Cube Operations". 10th International Conference on Scientific and Statistical Database Management, Capri, Italy. pp 53-62.
- [SPHD98] 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.