

Business Intelligence


04 Data Warehouse Modeling II

& First Short Introduction to Data Mining

Prof. Dr. Bastian Amberg
(summer term 2024)

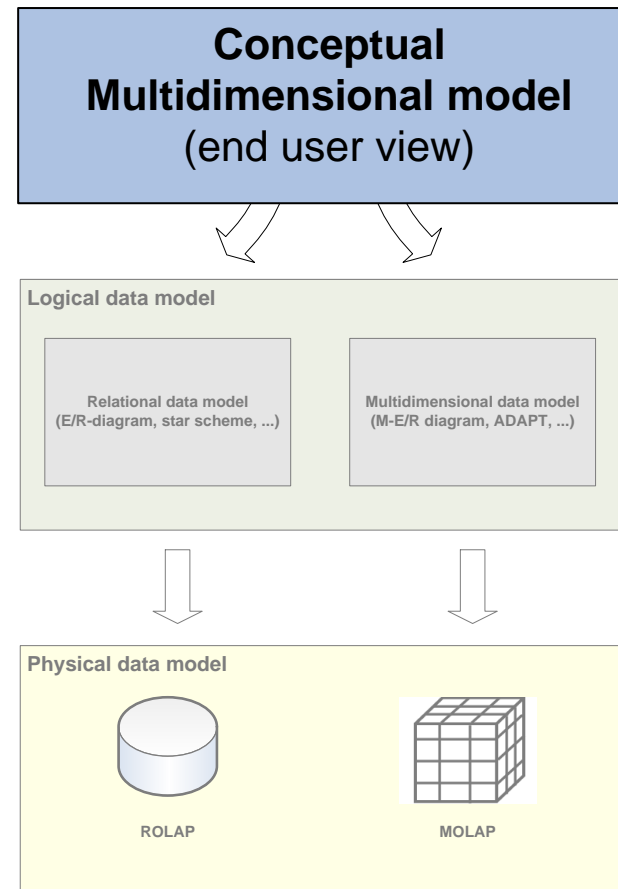
10.5.2024

Schedule

		Wed., 10:00-12:00		Fr., 14:00-16:00 (Start at 14:30)		Self-study
Basics	W1	17.4.	(Meta-)Introduction	19.4.		Python-Basics Chap. 1
	W2	24.4.	Data Warehouse – Overview & OLAP	26.4.	[Blockveranstaltung SE Prof. Gersch]	Chap. 2
	W3	1.5.		3.5.	Data Warehouse Modeling I 	Chap. 3
	W4	8.5.	Data Warehouse Modeling I & II	10.5.	Data Mining Introduction	
Main Part	W5	15.5.	CRISP-DM, Project understanding	17.5.	Python-Basics-Online Exercise	Python-Analytics Chap. 1
	W6	22.5.	Data Understanding, Data Visualization	24.5.	No lectures, but bonus tasks 1.) Co-Create your exam 2.) Earn bonus points for the exam	Chap. 2
	W7	29.5.	Data Preparation	31.5.		
	W8	5.6.	Predictive Modeling I	7.6.	Predictive Modeling II (10:00 -12:00)	BI-Project Start
	W9	12.6.	Fitting a Model I	14.6.	Python-Analytics-Online Exercise	
	W10	19.6.	Guest Lecture	21.6.	Fitting a Model II	
	W11	26.6.	How to avoid overfitting	28.6.	What is a good Model?	
Deepening	W12	3.7.	Project status update Evidence and Probabilities	5.7.	Similarity (and Clusters) From Machine to Deep Learning I	
	W13	10.7.		12.7.	From Machine to Deep Learning II	
	W14	17.7.	Project presentation	19.7.	Project presentation	End
Ref.					Klausur 1.Termin ~ 22.7. bis 3.8. Klausur 2.Termin ~ 23.9. bis 5.10.	Projektbericht

- ✓ Online Analytical Processing (OLAP)
- ✓ How can multidimensional data models be **developed** and **stored**?

1. Identify **facts** and **dimensions**
2. Create a **conceptual** data model
3. Derive a **logical** data model from the semantic model
4. Derive a **physical** data model from the logical model

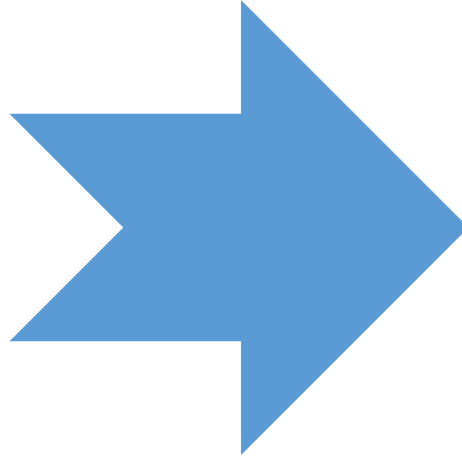


*Dimension = finite **set of categories** which are semantically related to each other with respect to business matters*

*Categories of one dimension represent a **different levels of aggregation** of the associated business *measures* (facts)*

For example, dimension: “date”

Four categories: day \Rightarrow month \Rightarrow quarter \Rightarrow year



Continuation Data Warehouse Modeling

Basic Elements of
multidimensional modeling

Conceptual modeling

Logical modeling

Physical modeling

Conceptual Modeling

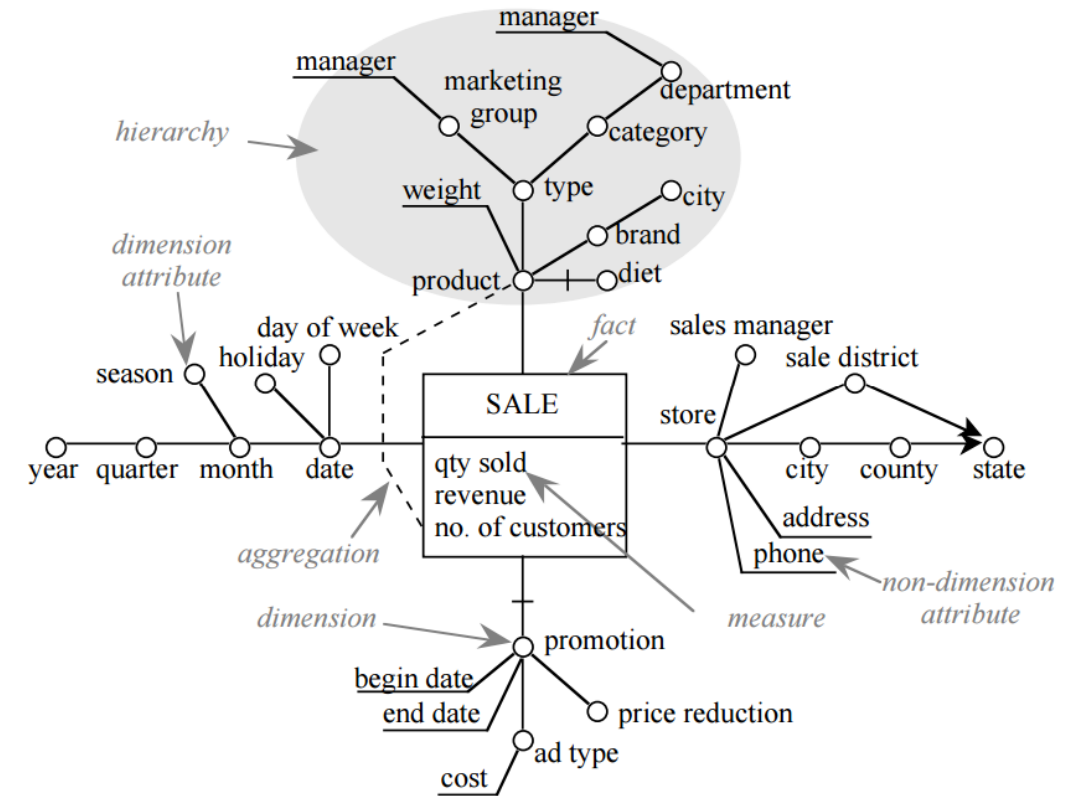
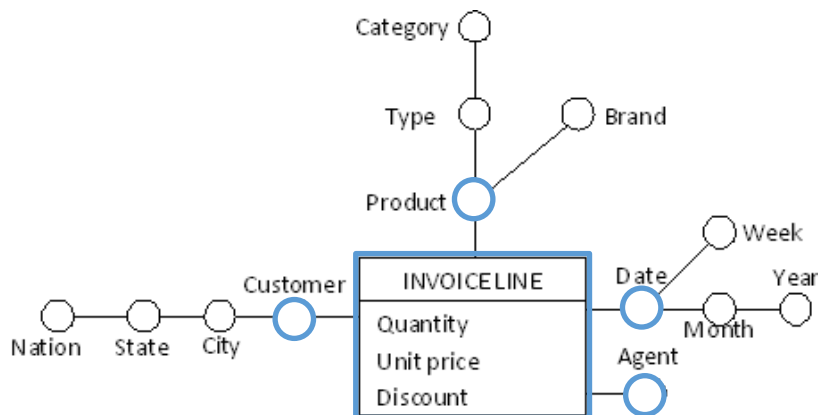
Dimensional fact model (or fact scheme)

Categories of a dimension arranged in a non-cyclic graph, directed between all-category and log-categories

Categories can have an arbitrary number of (non recursive) **relations** between each other

Several **aggregation paths** (e.g., sum/count/mean) may be included in the graph

Hierarchies are discrete attributes and define the granularities of facts (i.e., product -> type -> category)



Non- & Cross-dimensional attributes

Dimensional fact model

There may be **various types of relations** between individual categories of one (or more) dimension(s)

Categories having 1:1 relations can be summarized into a single category

When previous categories become non-dimensional, **descriptive attributes**

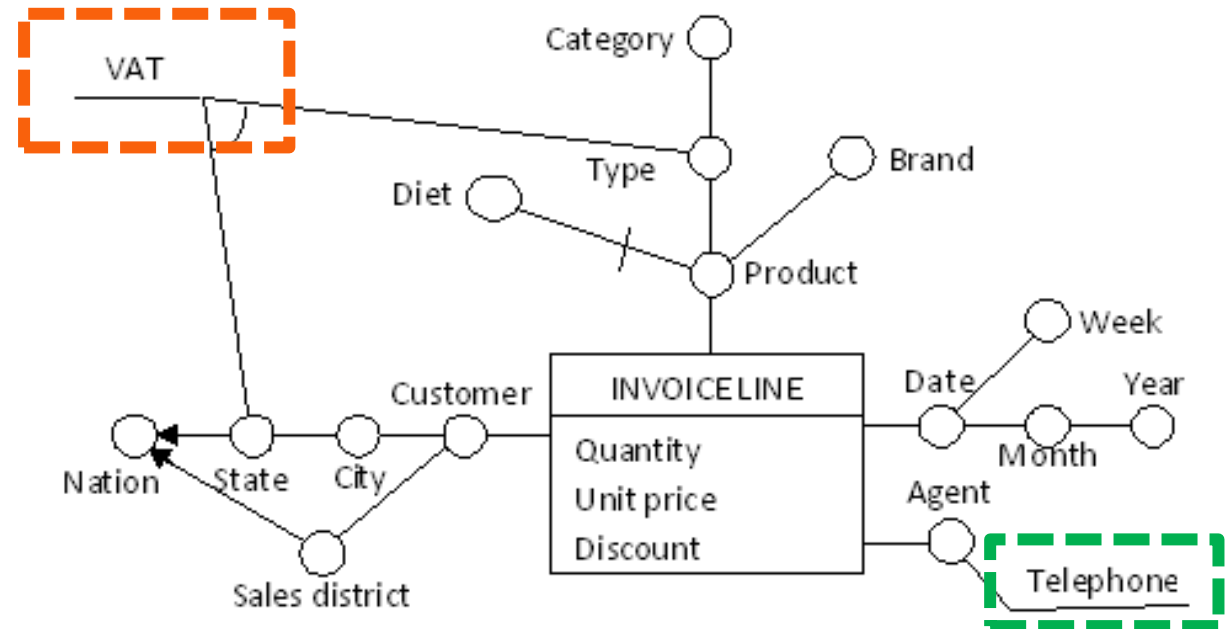
Non-dimensional attributes provide additional information and have 1:1 relations with the corresponding categories (phone-No. cannot be aggregated)

OLAP-compliant queries encompassing non dimensional attributes are generally not supported

Categories having **1:N** or **N:M** relations

When value is defined by multiple categories (e.g., product type and state. For instance, VAT for water or books in Germany vs. USA) they become **cross-dimensional attributes**

N:M - example



1:N - example

Nation \leftarrow State \leftarrow City \leftarrow Customer
1:N 1:N 1:N

1:1 - example

Exercise

Conceptual Modeling (Dimensional fact model)

Design a **conceptual model** for the local Food Company:

Conny's Corner Shop

Your managers need to keep themselves up to date on the number of items in the company's inventory. They especially want to keep an eye on their products with regards to location, and time.

- Conny's Corner Shop sells a range of snacks and beverages. Both categories have different types of products, such as juices and water, as well as prezels and crackers.
- The products are sold under different brands.
- Products have different package types, sizes, and weights.
- They store products in different stores across Europa

Show your conceptual model as a dimensional fact model. Make reasonable assumptions if necessary.

Ref.

10 Min.

Identifying fact groups

Single Star Scheme

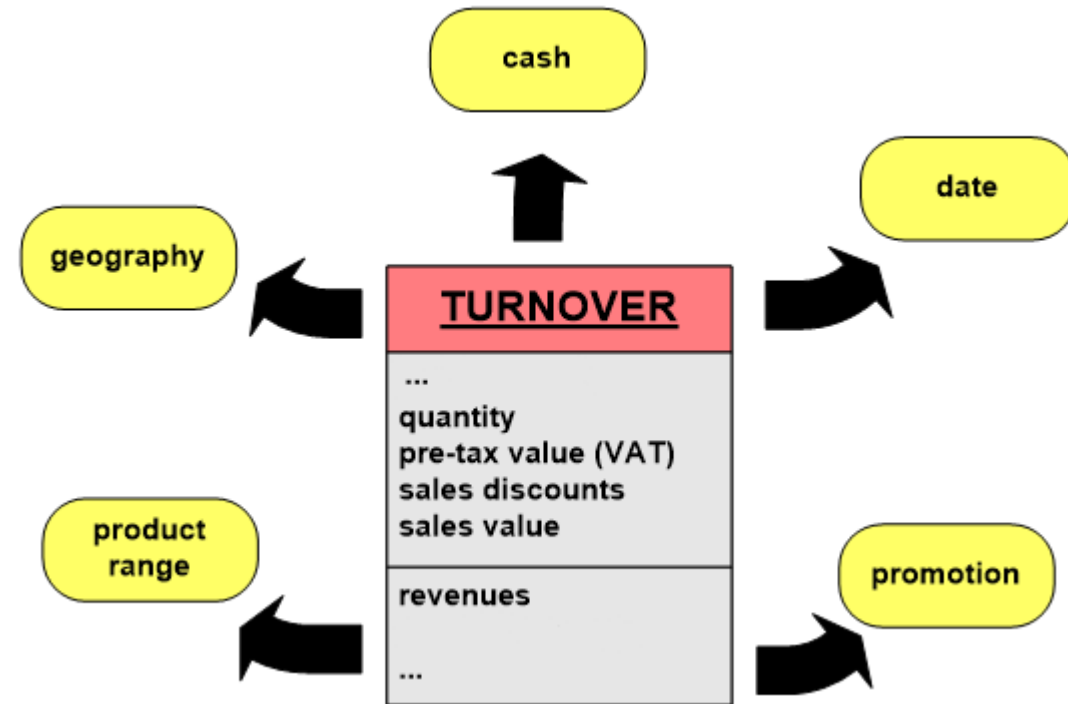
A graphical overview should be created for each fact group

single star scheme

Distinction between materialized facts and derived facts should be drawn

High level of abstraction:
aggregation formulas are commonly not modeled

Remember: **Fact group** = set of facts featuring *a common set* of dimensions



Combining fact groups

Multiple Star Scheme

A data warehouse data model encompasses a number of fact groups (**multiple star scheme**)
The sets of associated dimensions of different fact groups may overlap.

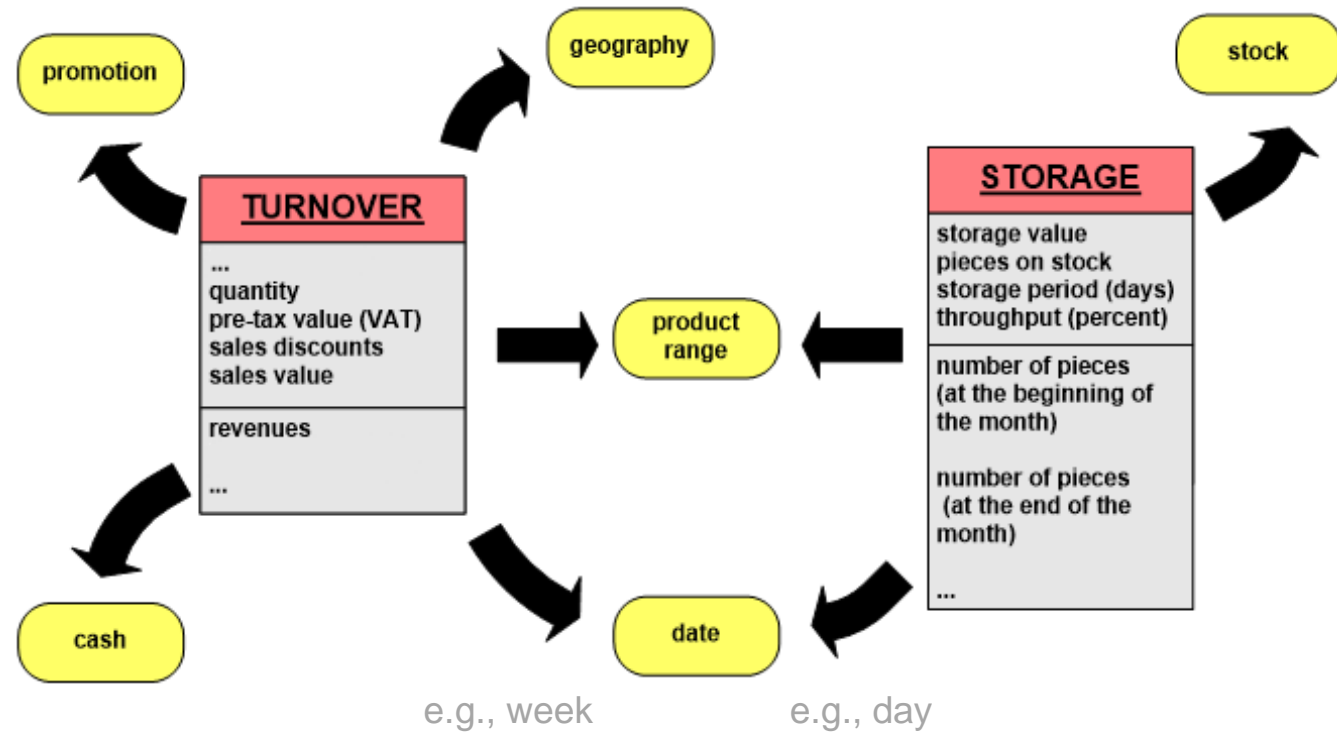
Different fact groups may use different log-categories with respect to one common dimension

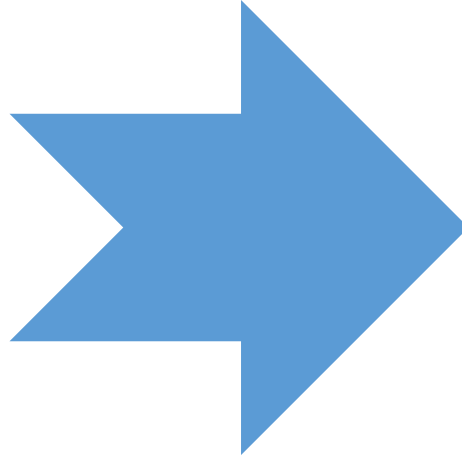
e.g., actual / debit values:

Actual facts: log-category of dimension date: day

Target facts: log-category of dimension date: month

Category used as log-category should be specified in the model (at least if standard log-category is not used)





Continuation Data Warehouse Modeling

Basic Elements of
multidimensional modeling

Conceptual modeling

Logical modeling

Physical modeling

Basic tasks with logical modeling

Logical modeling is about adapting the general conceptual schema to the applied database technology

Relational database technology

⇒ ER diagrams, ...

Multidimensional database technology

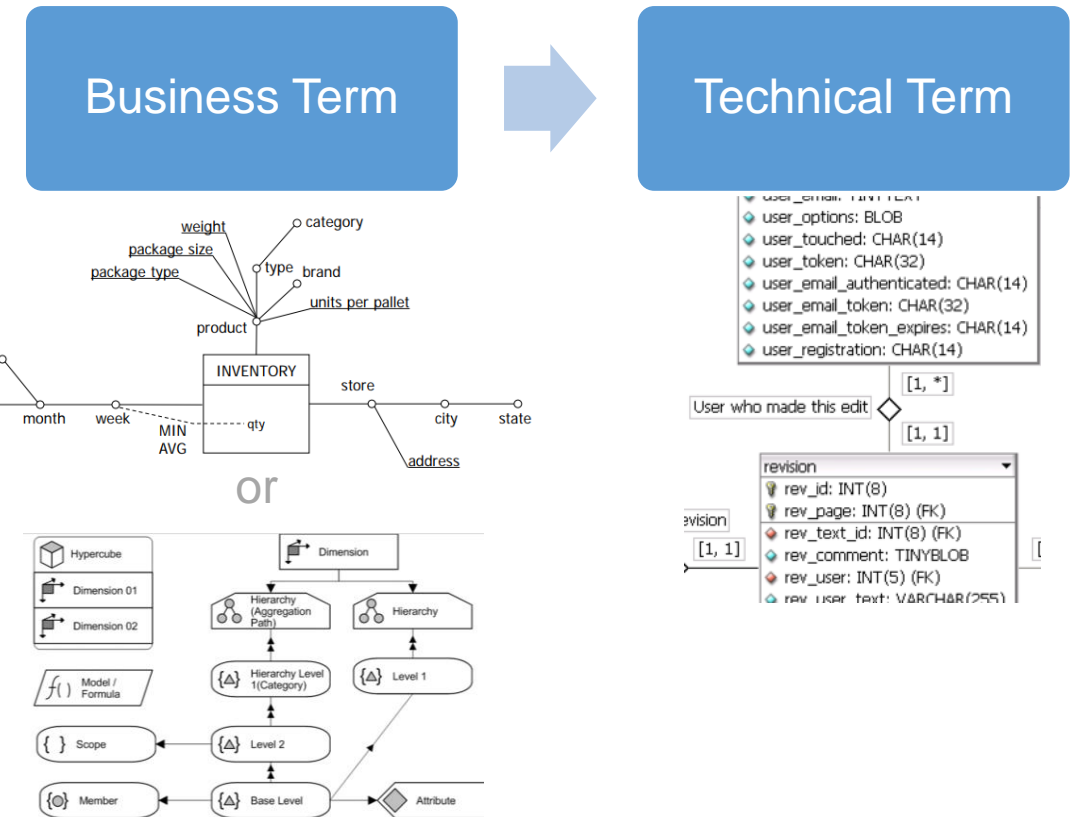
⇒ M-ER diagrams, ADAPT diagrams, ...

M-ER: Multidimensional Entity-Relationship

ADAPT: Application Design for Analytical Processing Technologies

(Both also usable as starting point for relational DB design)

Star Scheme is the standard way of logical modeling concerning **relational DBMS**



Excursus: [Genealogy of Relational Database Management Systems](#)



Image: [AutumnSnow \(2008\)](#) | Wikimedia (cc by sa 3.0)

Properties of star scheme models

A number of fact tables are associated with a set of dimension tables each (via **unique keys**)

One dimension is mapped to exactly one relation

Primary key of a fact table consists of the keys of the associated dimensions

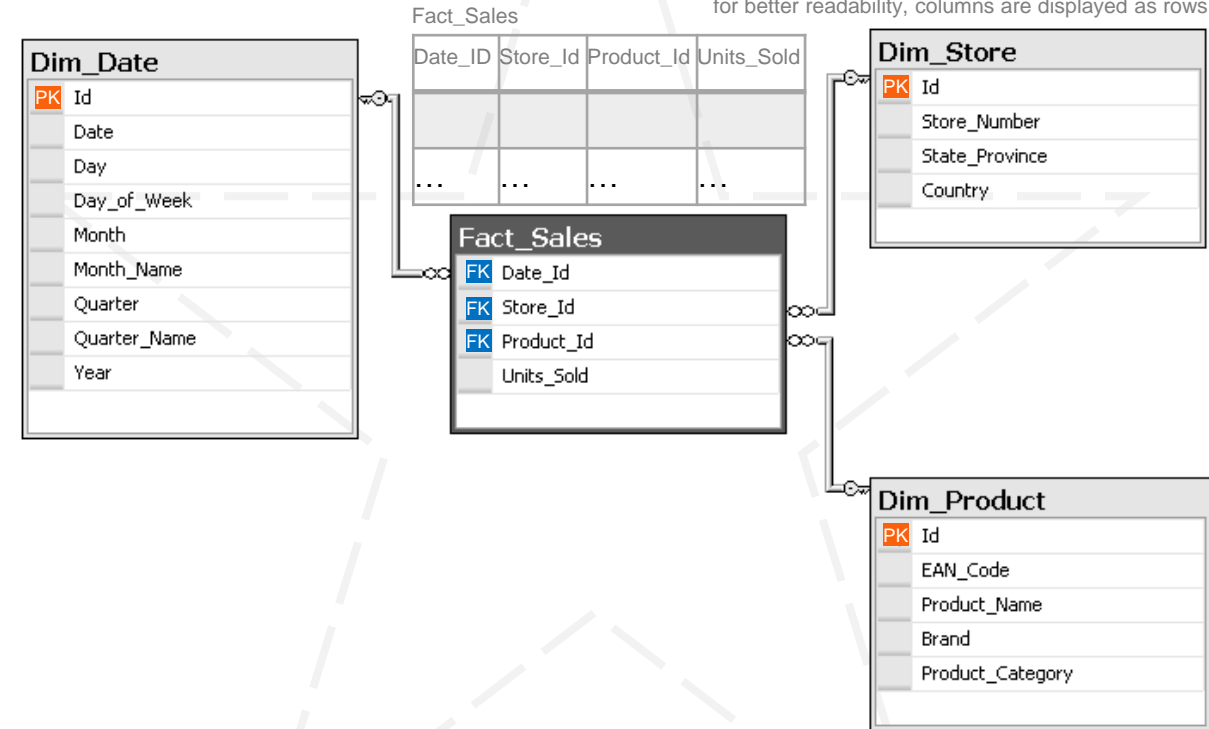
Keys of the dimension tables are usually “artificially” created (e.g. serial numbers)

1:m type relations between dimensions and facts

Typical “multi-join” query for star schemes:

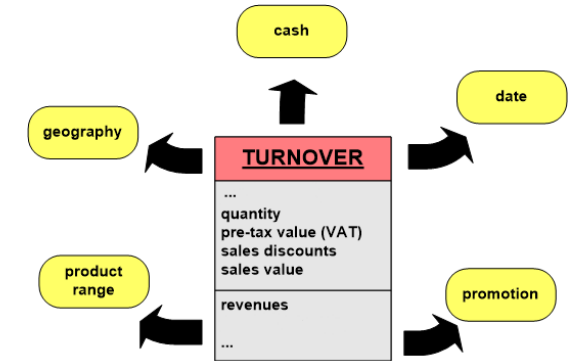
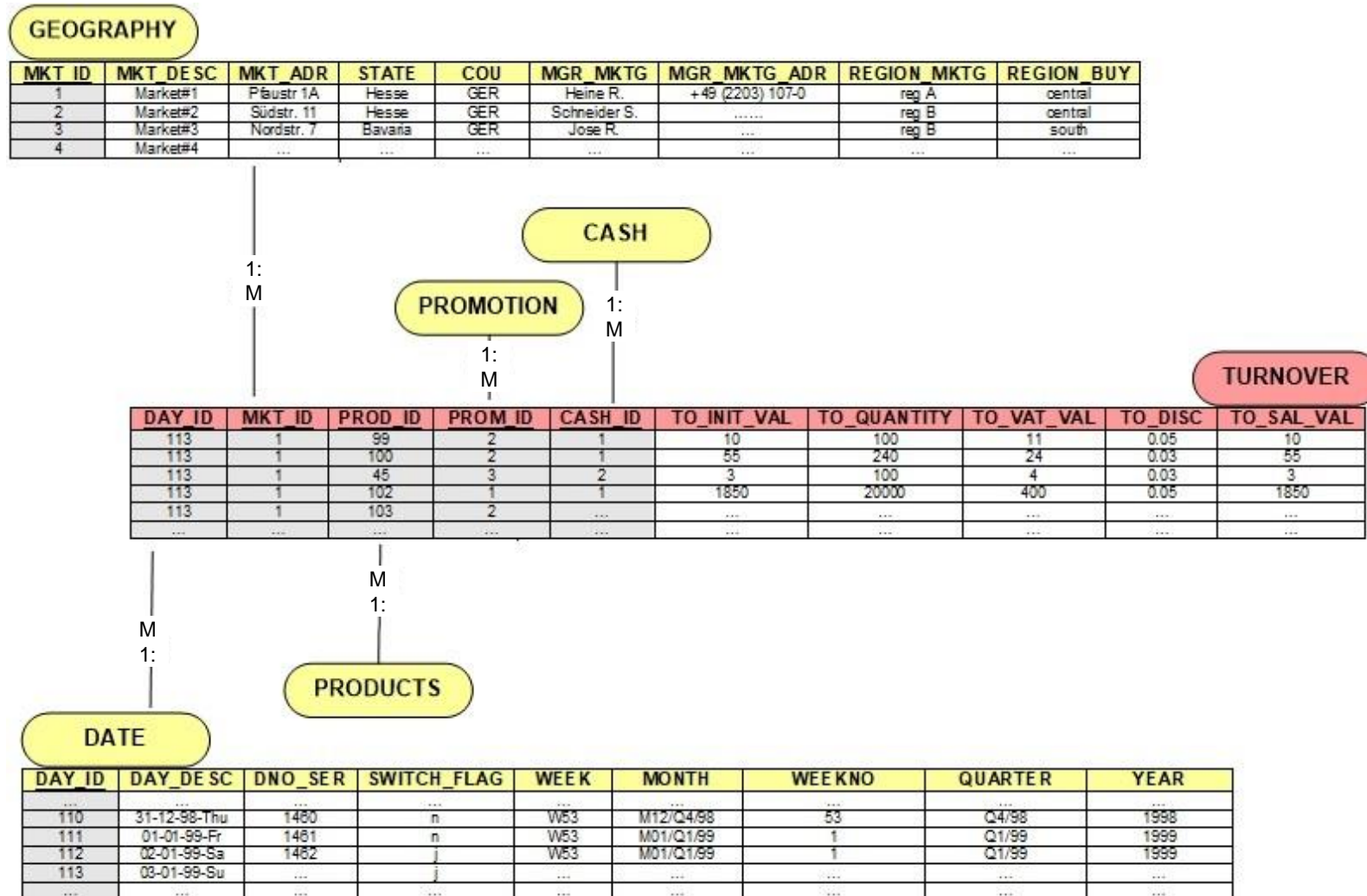
```
SELECT
  D1.State_Province, D2.Month,
  SUM (F1.Units_Sold)
FROM Dim_Store D1, Dim_Date D2, Fact_Sales F1
WHERE
  D1.Id = F1.Store_Id AND
  D2.Id = F1.Date_Id
GROUP BY D1.State_Province, D2.Month
```

Result?



PK: Primary Key
FK: Foreign Key

Simple star scheme



Category or descriptive attribute?
Hierarchies?
Performance for aggregation functions?

Simple star schemes

Pros and Cons

PRO

Simple, intuitive model

Not many physical join-operations needed

Not many physical tables needed

⇒ maintenance easy

⇒ Extract, Transform, Load (ETL) easy

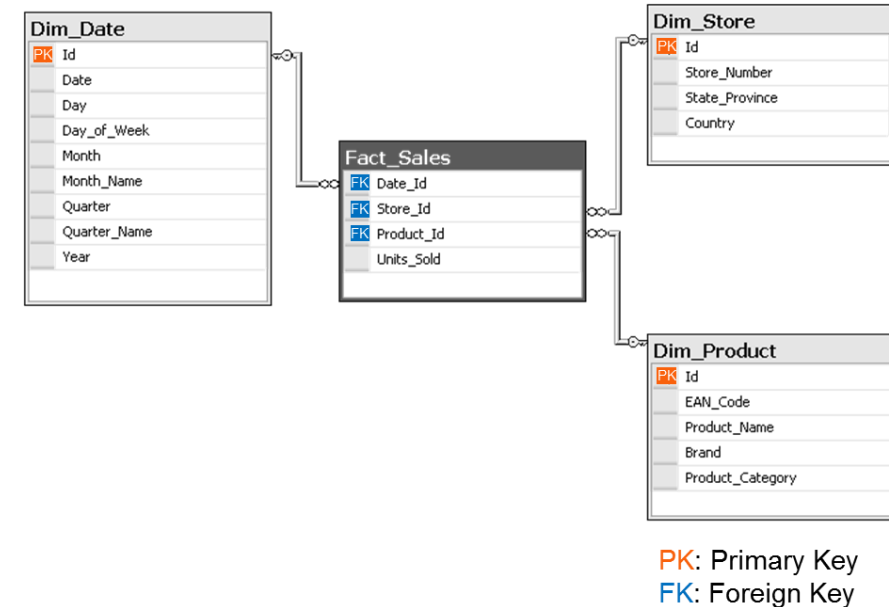
CON

Large dimension tables may cause **bad response times**

Creation of materialized views (aggregated summary tables) difficult
⇒ wrong aggregate values may be calculated due to **multiple counting of entries**

Changes of the conceptual model cause extensive reorganization efforts on the tables ⇒ versioning of meta data required

Redundancy



Variants of the star scheme

Snowflake scheme:

Normalization of the dimensional tables of a star scheme (see also example on next slide)

Dimensional tables are broken down into several small lookup-tables \Rightarrow no double entries

Consolidated star scheme:

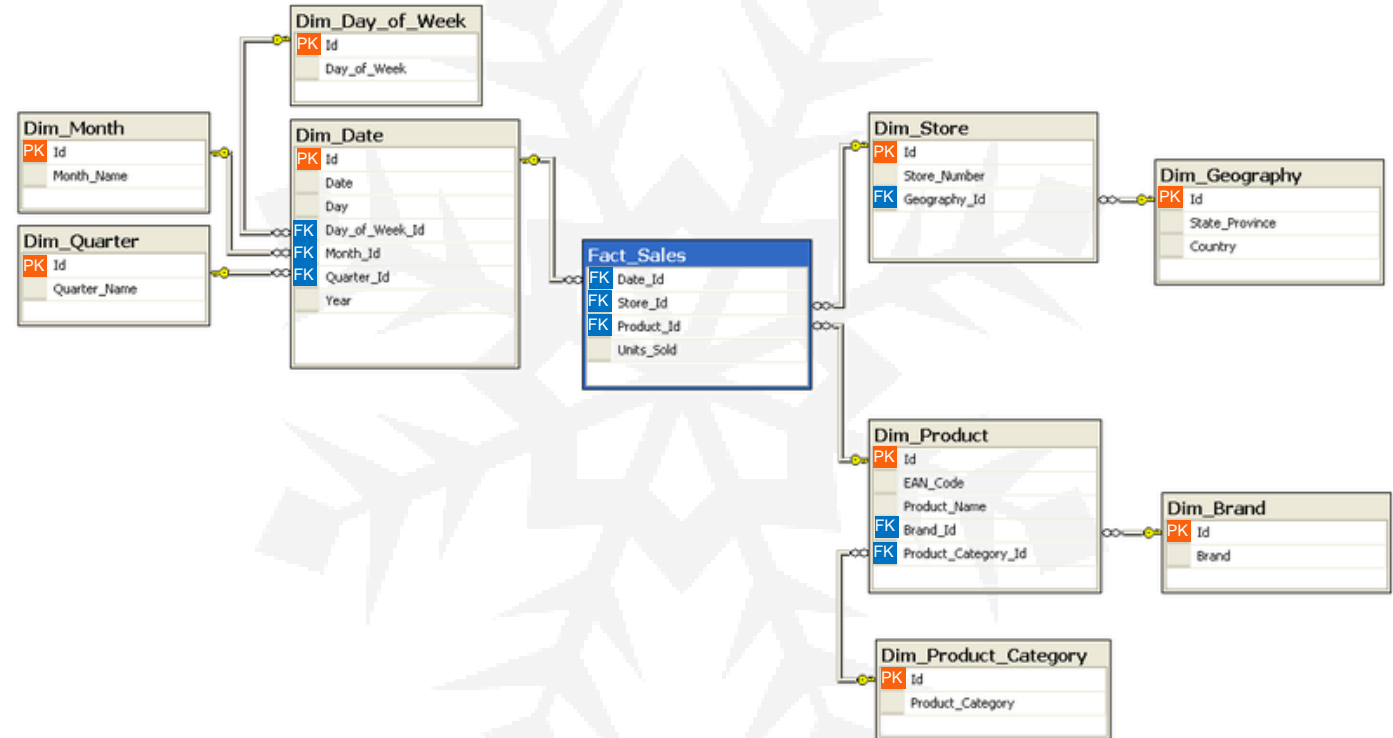
Basic idea: storage of aggregate values **within the fact tables** (“in-table aggregation”)

Requires level attributes within the dimensional tables

Provides good response times \Rightarrow calculations displaced to ETL-procedure (extract, transform, load)

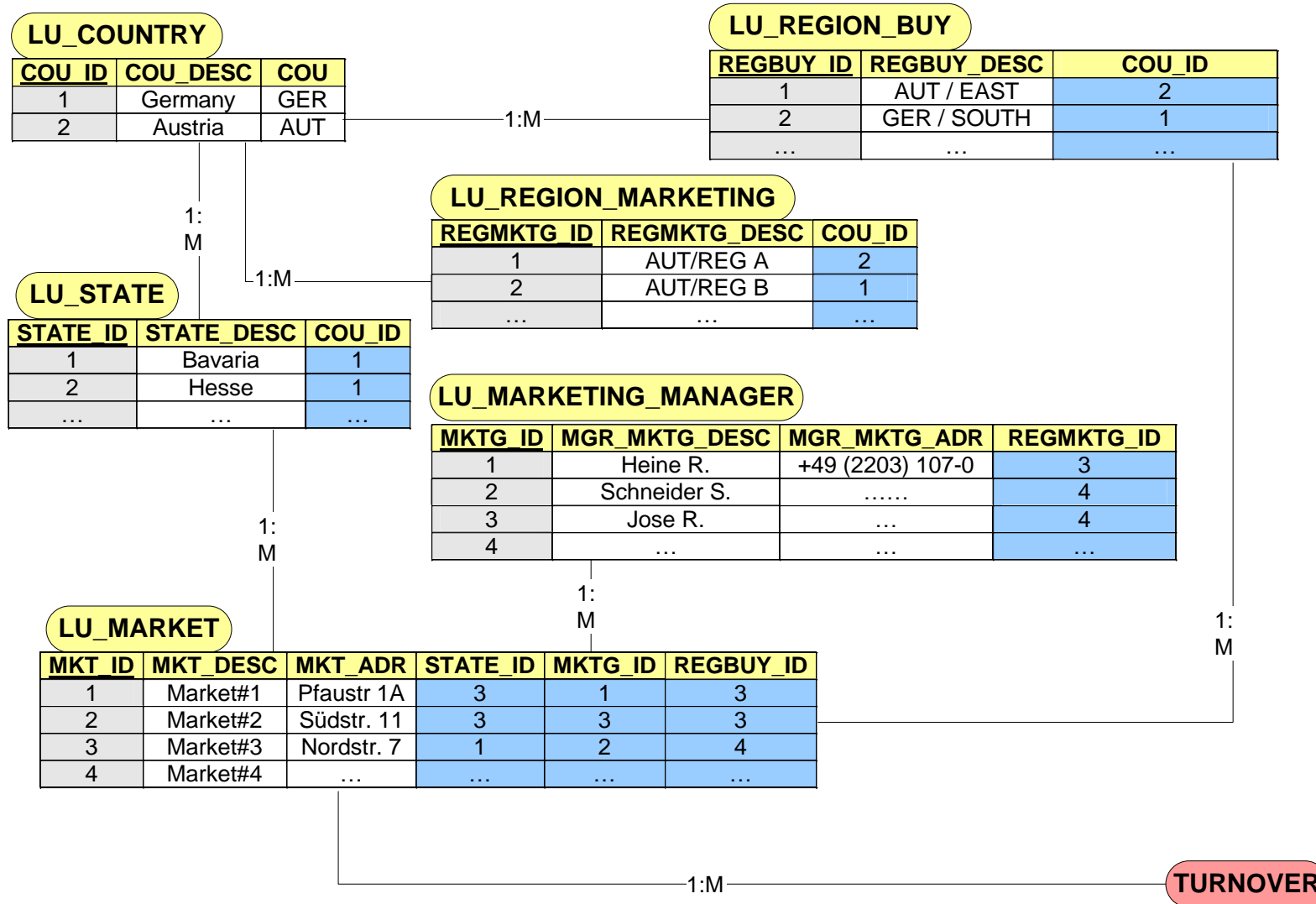
Fact constellation scheme:

Variant of consolidated star schemes (avoids level-attributes, uses additional fact tables for storage of aggregated values)



PK: Primary Key
FK: Foreign Key

Snowflake scheme



GEOGRAPHY

Hierarchies?
Storage of aggregated values?

Snowflake scheme

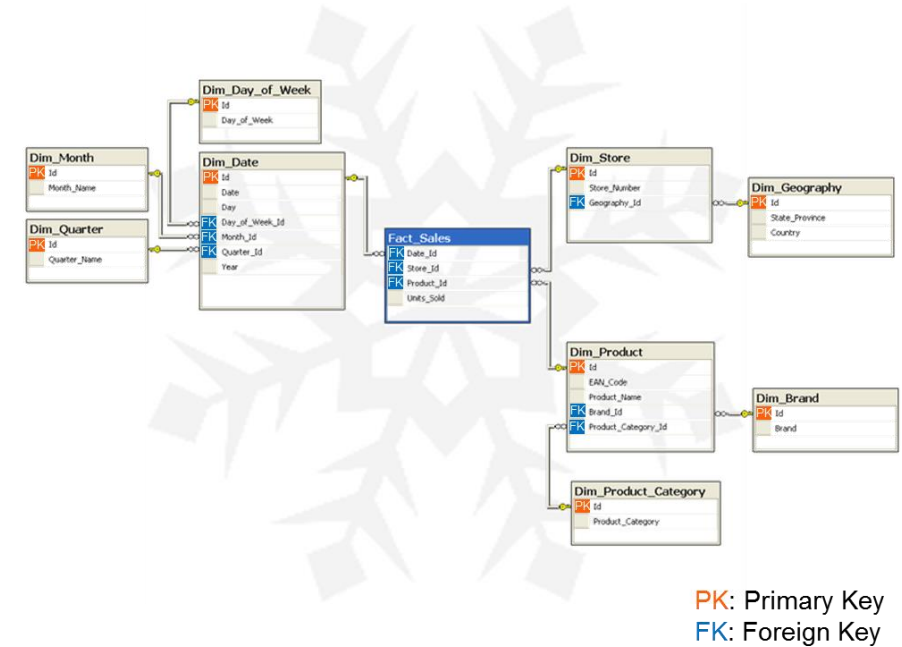
Pros and Cons

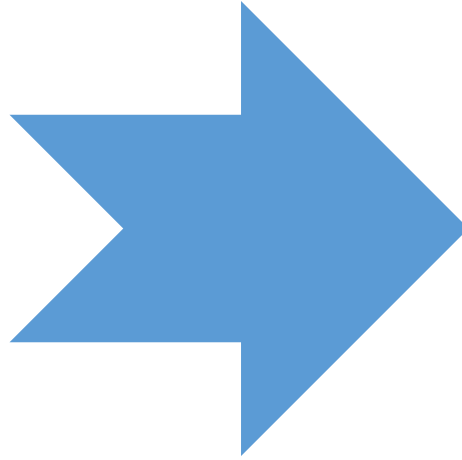
PRO

- Good support of materialized views
- Browsing can be easily implemented based on snowflake schemes
- No redundancy within the dimension tables

CON

- More physical join operations required
- More physical tables required
- Higher level of complexity
 - ETL-process
 - Maintenance
 - SQL-queries





Continuation Data Warehouse Modeling

Basic Elements of
multidimensional modeling

Conceptual modeling

Logical modeling

Physical modeling

Physical modeling

Database architectures

Physical implementation of logical schemata in a database system.

Relational database

usually denormalized structure for storage (star scheme)

„virtual cube“

Examples: MS Access, see
Genealogy of RDBMS for more

.....
.....
.....
.....

Z.B. Absatz,
Dimensionen:
Monat (3),
Kunde (3),
Artikel (3)

Multidimensional database

(precalculated) data stored in multidimensional data structures on OLAP server

cube on server

Examples: IBM Cognos TM1,
Oracle Hyperion/Essbase, Jedox

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
K1	K2	K3	K1	K2	K3	K1	K2	K3	K1	K2	K3	K1	K2	K3	K1	K2	K3	K1	K2	K3	K1	K2	K3	K1	K2	K3
Artikel 1			Artikel 2			Artikel 3			Artikel 1			Artikel 2			Artikel 3			Artikel 1			Artikel 2			Artikel 3		
Monat 1									Monat 2									Monat 3								

Client-based files

small extracts of data are held on clients

cube on client

Select and fine tune the used DBMS technology

Building the database using the specific data definition language (e.g.: SQL-DDL-commands)

Decide on how to index, partition, denormalize and partly pre-aggregate the data.

```
CREATE TABLE employees (  
  id INT(6) AUTO_INCREMENT PRIMARY KEY,  
  first_name VARCHAR (50) not null,  
  last_name VARCHAR (75) not null,  
  age INT(3) not null,  
  dateofbirth DATE not null )
```

Physical modeling

OLAP architectures

OLAP - architectures		STORAGE		
PROCESSING	SQL	RDBMS ROLAP (thin client)	MDBMS -	client-based -
	multidimensional server engine	ROLAP (thin client)	MOLAP (thin client)	-
	client multidimensional engine	ROLAP (fat client)	MOLAP (fat client)	DOLAP (fat client)

MOLAP = multidimensional OLAP
(data in cubes)

ROLAP = relational OLAP

DOLAP = desktop OLAP

Ref. e.g., Jukic et al. (2008)

Recent developments
in data warehousing,
see article "Paradigmenwechsel:
Data Warehouses für die Cloud"
[Kursmaterial > Readings/Übungen](#)

Fragen?

✓ Data Warehouse Modeling

Exercise

Conceptual Modeling and Logical Model (Star Scheme)

Design a **logical model** for the local Food Company:

Conny's Corner Shop

Your managers need to keep themselves up to date on the company's situation. They especially want to keep an eye on their products with regards to location, and time.

- Conny's Corner Shop sells beverages and snacks. Both categories have different types of products, such as juices and water, as well as prezels and crackers. The products are either branded as strictly vegan or organic (bio).
- Products are sold in different cities, regions and EU-countries
- Most managers are interested in quantities, income and discounts of sales.

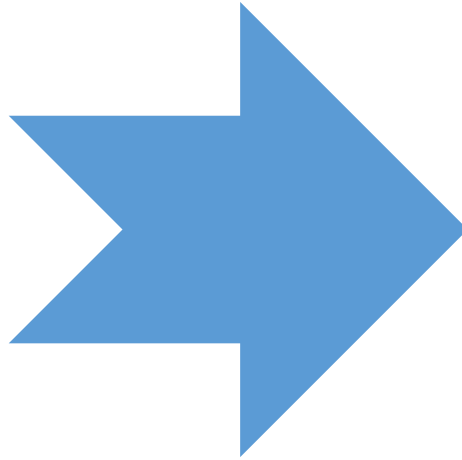
Create a simplified Star Scheme (logical model) with tables/relations.
Use your conceptual model as input.

10 Min.

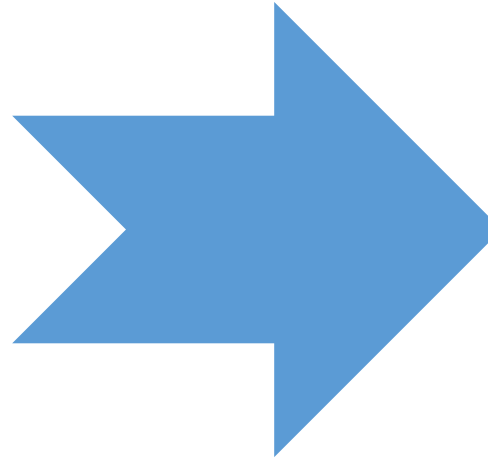
Outlook next lesson

Data Mining Introduction

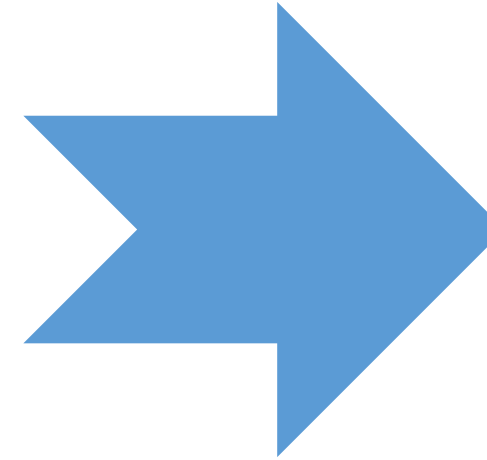
*Where are the limits of
the handling of data
considered so far?*



(1) The need for data mining



(2) From business problems to data mining tasks



(3) Supervised vs. unsupervised methods

“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”

(Hand, Mannila, Smyth (2001), Principles of Data Mining)



The need for data mining

“80% of the knowledge of interest in a business context can be extracted from data using **conventional tools**.” (Lusti)

- reporting
- query-languages (SQL, QBE, ...)
- OLAP and spreadsheets

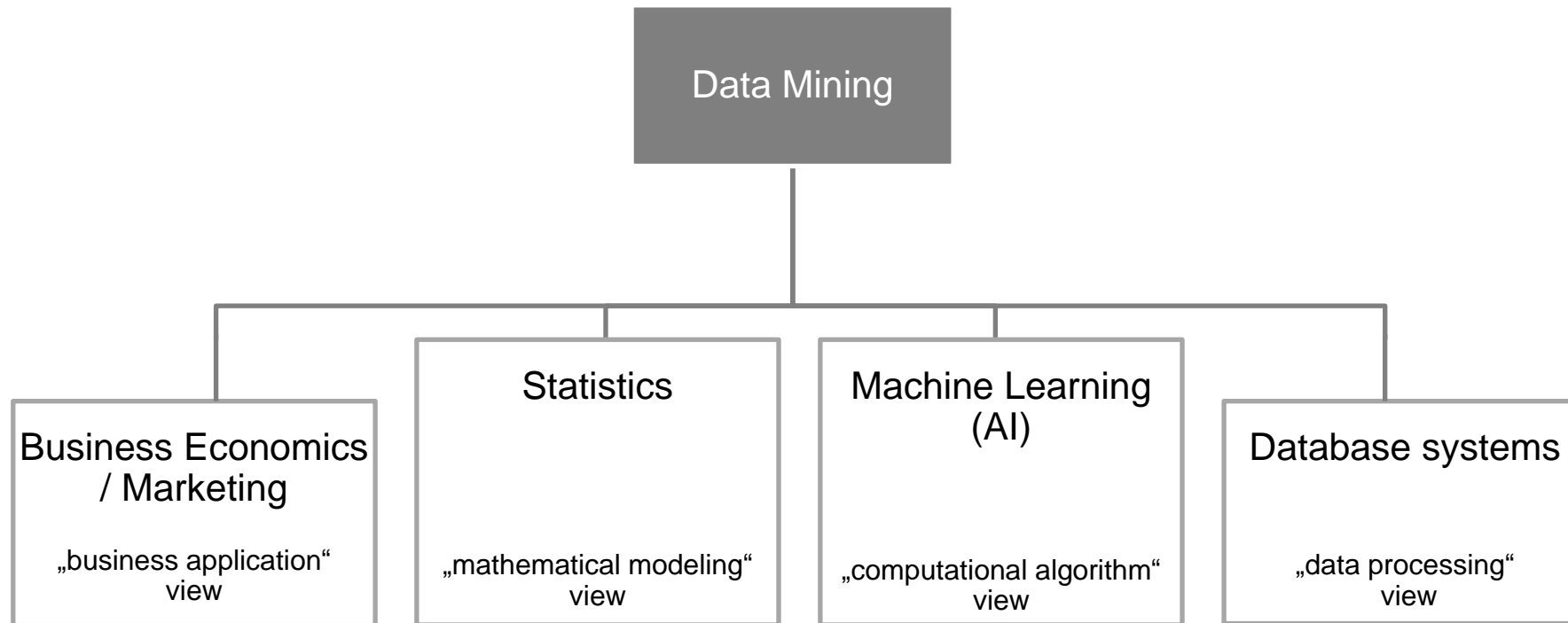


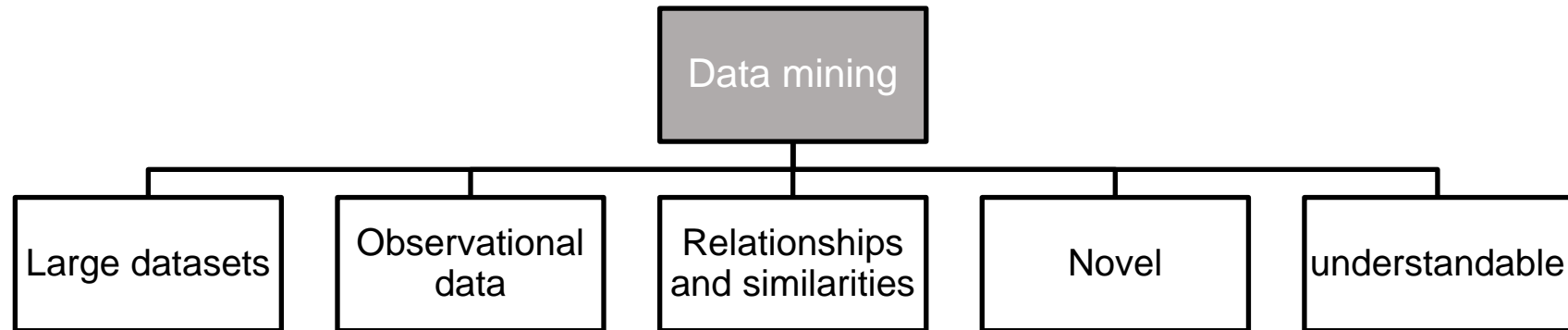
Phase ID	Description	Results	Start Date	End Date	Resource
1.0 Phase	Initial software requirement on the business requirement and develop the product.	RFR			Product Manager
1.0.1 Milestone	Review the Business Requirement	RFR			
1.0.1.1 Task	Review the Product Requirement	Product Requirement			
1.0.1.2 Task	Prepare and Review the Product Requirement	Product Requirement			
1.0.2 Milestone	Develop the Product	RFR			
1.0.2.1 Task	Review the Product Structure	Product Structure			
1.0.2.2 Task	Prepare Product Structure	RFR			
1.0.3 Phase	Develop the user interface and building the user interface	RFR			
1.0.3.1 Milestone	Review the User Interface	RFR			
1.0.3.1.1 Task	Review the User Interface	User Interface			
1.0.3.1.2 Task	Review the User Interface Structure	User Interface Structure			
1.0.3.1.3 Task	Review the User Interface	Review the User Interface			
1.0.3.2 Task	Develop the User Interface and User Interface	User Interface			
1.0.3.3 Task	Prepare Product Structure	Product Structure			
1.0.3.4 Milestone	Review the Building Product for Review	RFR			

Disadvantages of conventional tools:

- Often, merely simple questions can be answered
- OLAP: query-focused and low complexity of analysis
e.g., performance changes are visible, but what about the overall context/ reasons?
- Automation of knowledge discovery is difficult
hypothesis needed
- Only small amounts of data may be handled (esp. spreadsheets)
but exploding amount of raw data available

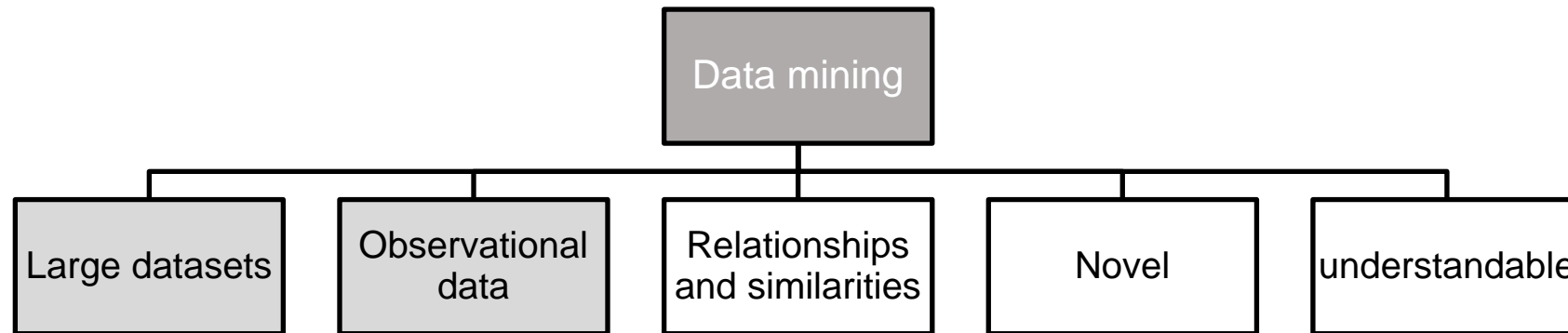
Major roots of data mining





“Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.”

(Hand, Mannila, Smyth (2001), Principles of Data Mining

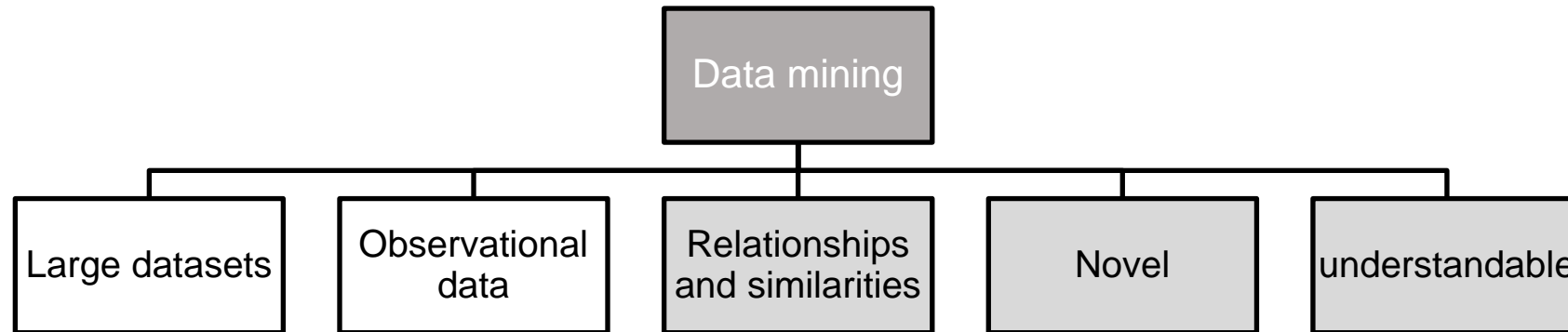


Often large datasets:

- Small datasets \Rightarrow exploratory data analysis in statistics
- Large datasets (as they exist in DWHs) provoke new problems
 - Storage and access of data
 - Runtime issues
 - Determination of representativeness of data
 - Difficulty to decide whether an apparent relationship is merely a chance occurrence or not

Observational data:

- Data often collected for some other purpose than data mining
- Objectives of the data mining exercise play no role in data collection strategy
 - e.g., DWH data relying on an airline reservation system or a bank account administration system
- opposite: experimental data (as it is used quite often in statistics)



Relationships and summaries:

often referred to as **models** or **patterns**
e.g., linear equations, tree structures, clusters,
patterns in time series, ...

Understandable:

Novelty is not sufficient to qualify
relationships worth finding
Simple relationships may be preferred to
complicated ones

Novel:

Novelty should be measured relative to users prior
knowledge

Exercise: Data Mining vs. OLAP

Typical questions

Fragestellung	Data Mining	OLAP
Kundenwert	Welche Kunden bieten uns das größte Deckungsbeitragspotenzial?	Wer waren letztes Jahr unsere 10 besten Kunden?

Kahoot-Fragen
www.kahoot.it
(über Smartphone oder Laptop)
PIN folgt

(Diese Folie ist nach der Vorlesung mit Lösungen verfügbar)

Fragen?

- Data Mining Introduction
(will be continued in the next lesson)

Todo for next Week

Support Conny's Corner Shop by creating a (or finishing the) simplified Star Scheme (logical model) with tables/relations.

See exercise on slide 23

- Hand, David J., Heikki Mannila, and Padhraic Smyth. *Principles of data mining*. MIT press, 2001.
- Vaisman, A., & Zimányi, E. (2014). *Data Warehouse Systems*. Springer, Heidelberg

Recommended reading (for next lessons)

Data Mining

Provost, F. Chapter 2
Fawcett, T.

Berthold et al. Chapters 1, B, C

Lusti, M. Data Warehousing und Data Mining (Chapter 6)

Hand, D. et al.: Principles of Data Mining (esp. Chapters 1, 5, 6 and 11)