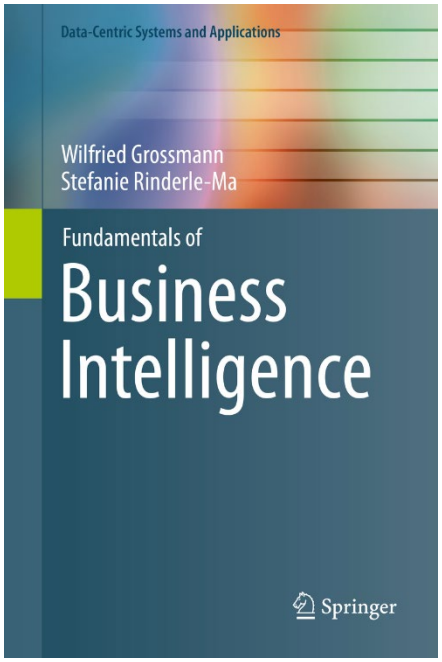


Chapter 3:
Data Provisioning

Contents

1 Introduction

- „*It's all about the data. […] But data doesn't come to you*…"[1]
- Data collection, extraction, and integration is often the most complex and expensive tasks in a BI project
- According to Bernstein and Haas[2]
  - "*information integration is thought to consume about 40% of their budget*"
  - "*the market for data integration and access software […] was about $2.5 billion in 2007 and is expected to grow to $3.8 billion in 2012*"

[1] http://mashable.com/2009/12/23/marketing-data/

[2] P. A. Bernstein and L. M. Haas, „Information integration in the enterprise", *Commun. ACM*, 51(9):72–79 (2008)

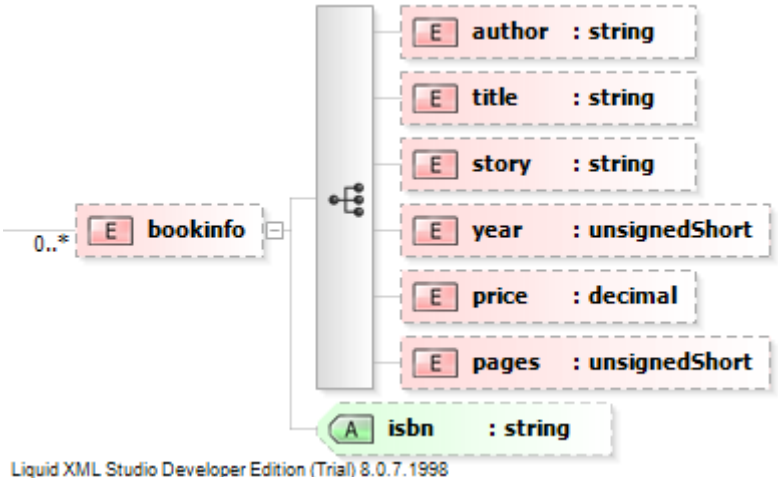
1 Introduction

- In addition: more and more data is available
- According to Chauduri et al.[3], we face „*very large amounts of data arising from sources such as customer transactions in banking, retail as well as in e-businesses, RFID tags for inventory tracking, email, query logs for Web sites, blogs, and product reviews*"
- On top, "*Real-world Data is Dirty*" according to Hernandes and Stolfo[4] therefore data quality is of utmost importance
- Crucial: Keep an eye on your analysis goals!
- In summary, we have to
  - collect / select
  - extract
  - clean, and
  - integrate data

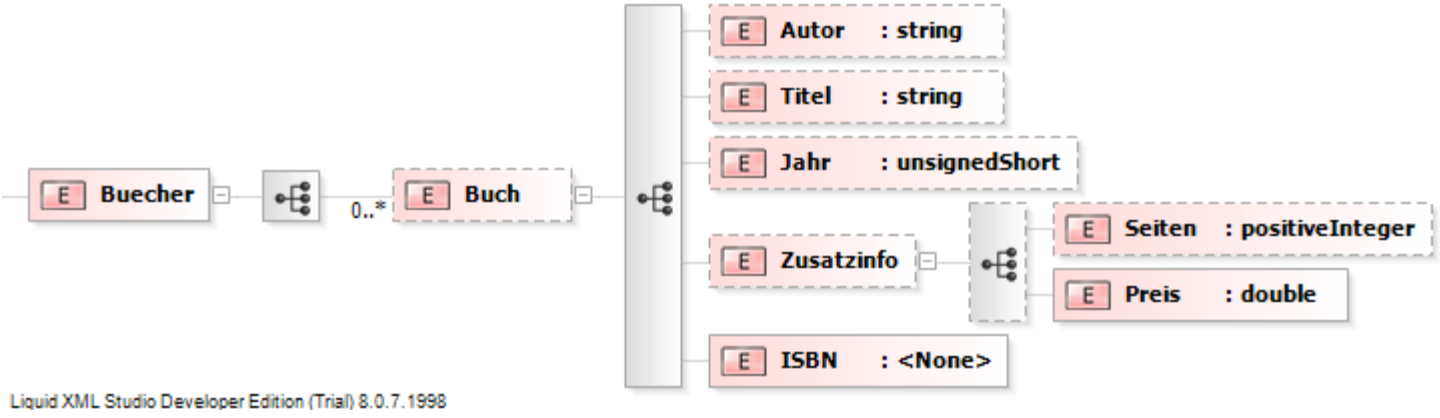
[3]S. Chaudhuri, U. Dayal, V. Narasayya, „An overview of business intelligence technology", *Communications of the ACM*, 54:88 (2011)

[4]Hernandez and S. J. Stolfo, „Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem", *Data Mining and Knowledge Discovery* 2(1):9–37 (1998)

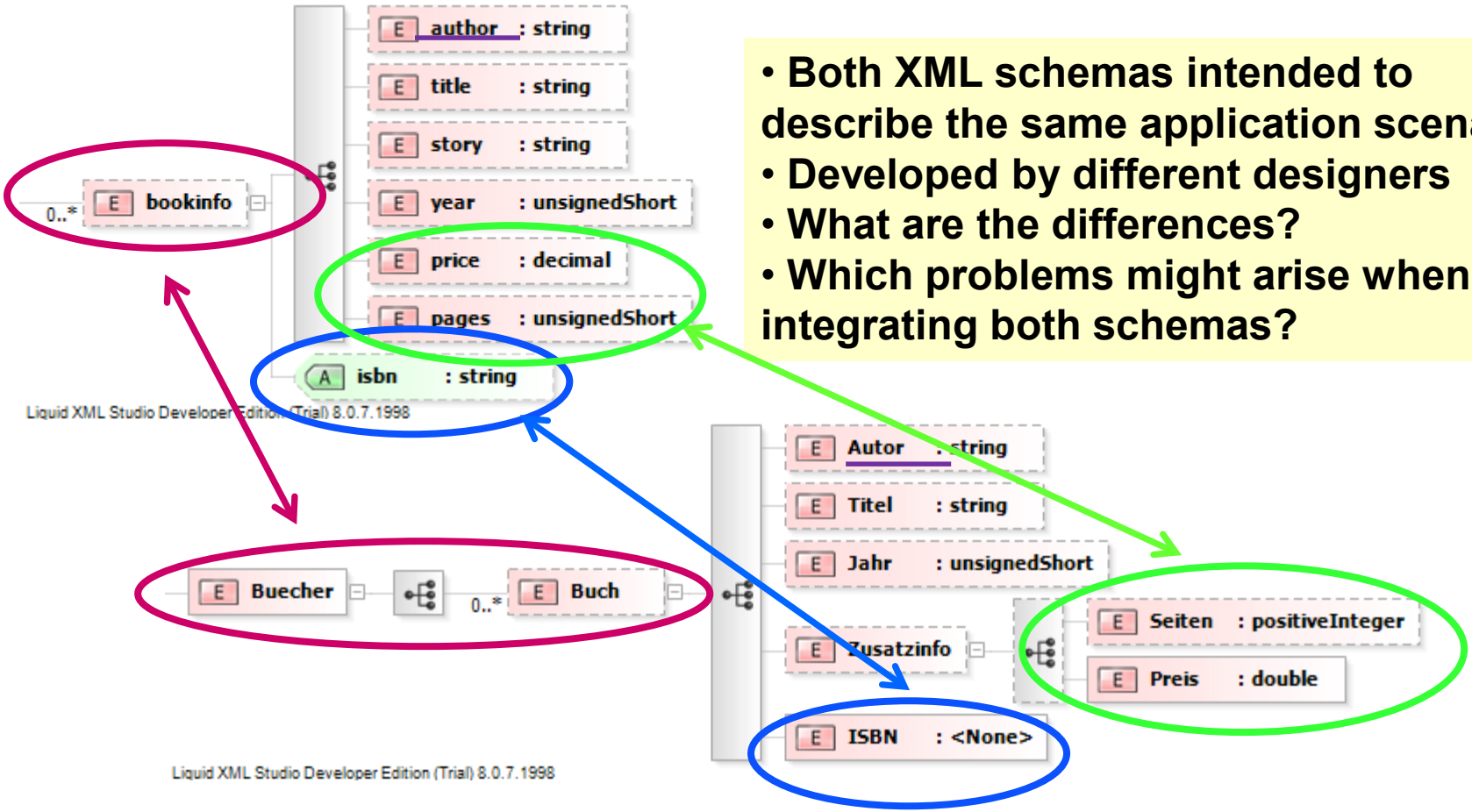
– Example problem: integration at schema level



- **Both XML schemas intended to describe the same application scenario**
- **Developed by different designers**
- **What are the differences?**
- **Which problems might arise when integrating both schemas?**



1 Introduction



- **Both XML schemas intended to describe the same application scenario**
- **Developed by different designers**
- **What are the differences?**
- **Which problems might arise when integrating both schemas?**

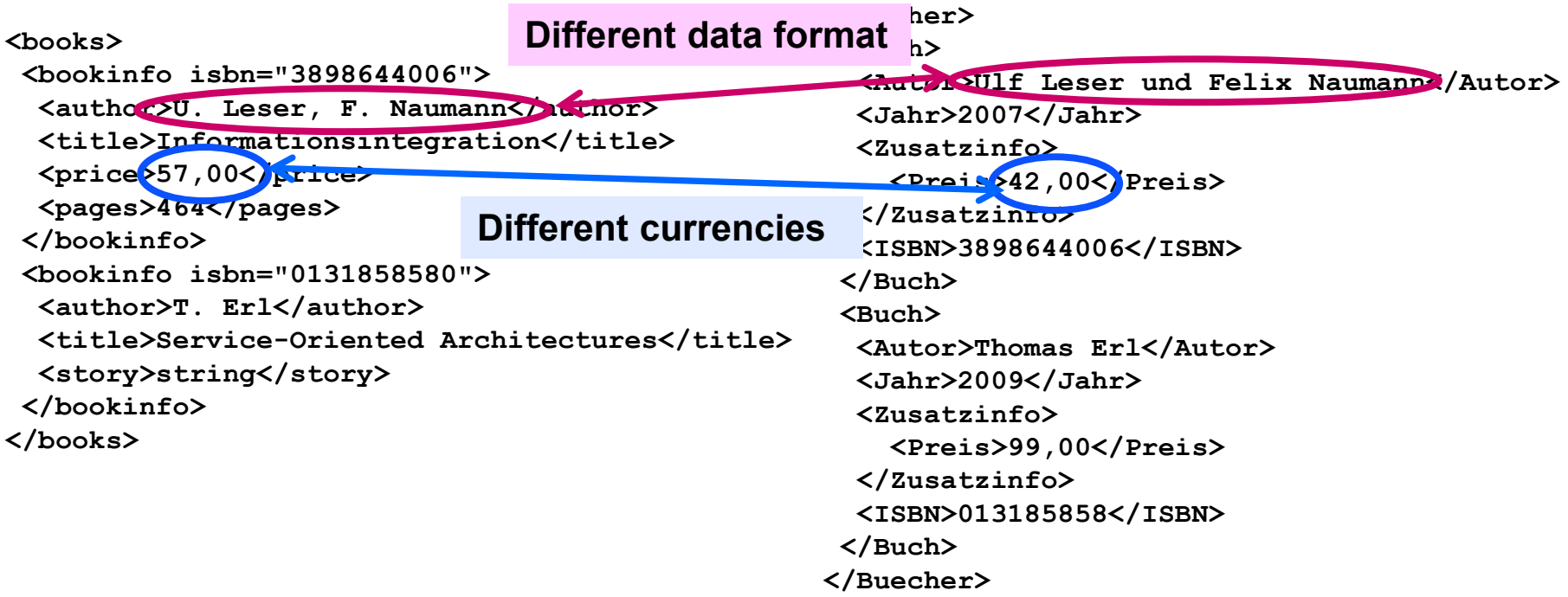
- In addition: Data Integration

```
<books>
 <bookinfo isbn="3898644006">
  <author>U. Leser, F. Naumann</author>
  <title>Informationsintegration</title>
  <price>57,00</price>
  <pages>464</pages>
 </bookinfo>
 <bookinfo isbn="0131858580">
  <author>T. Erl</author>
  <title>Service-Oriented Architectures</title>
  <story>string</story>
 </bookinfo>
</books>
```

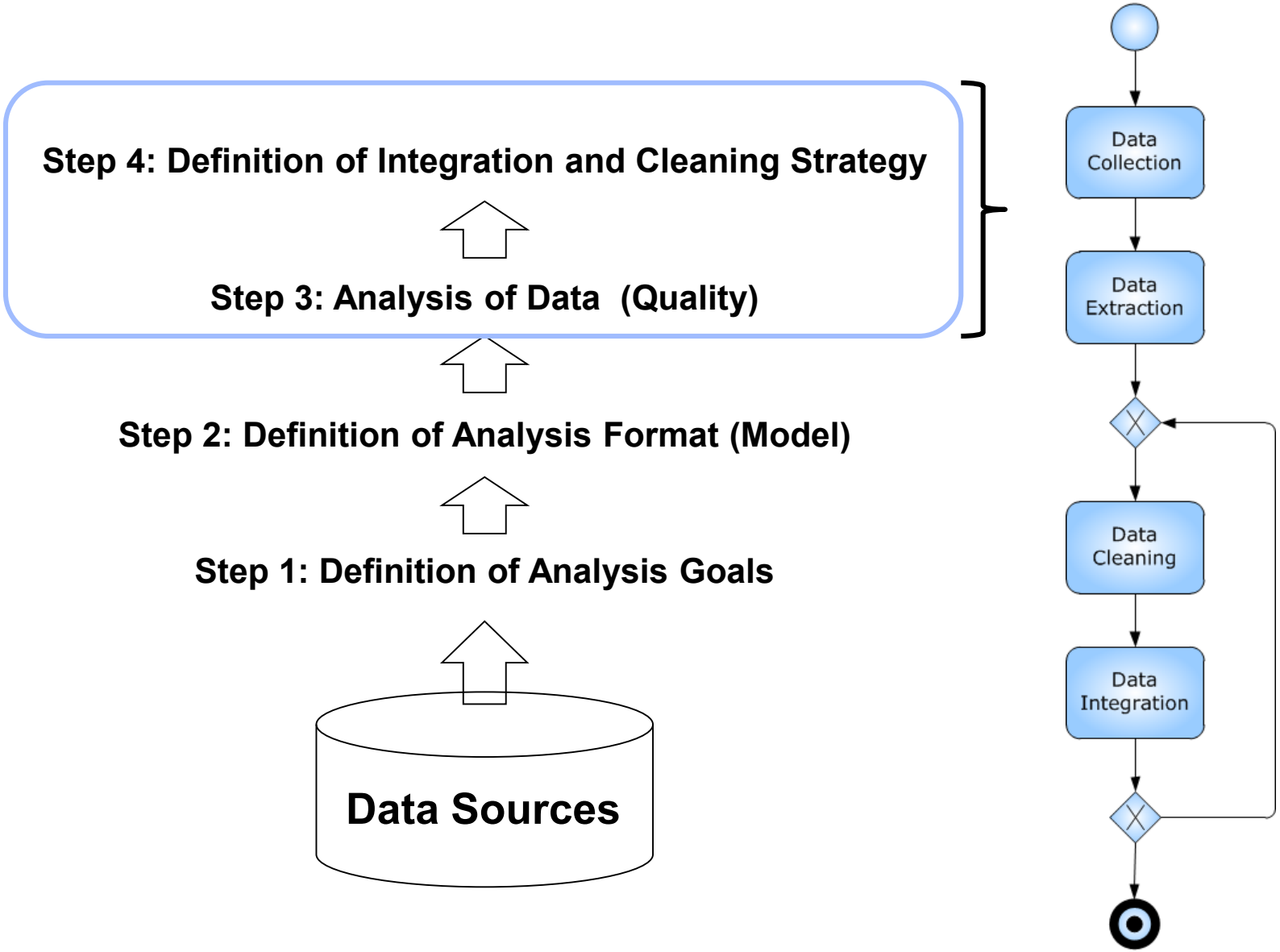
```
<Buecher>
 <Buch>
  <Autor>Ulf Leser und Felix Naumann</Autor>
  <Jahr>2007</Jahr>
  <Zusatzinfo>
    <Preis>42,00</Preis>
  </Zusatzinfo>
  <ISBN>3898644006</ISBN>
 </Buch>
 <Buch>
  <Autor>Thomas Erl</Autor>
  <Jahr>2009</Jahr>
  <Zusatzinfo>
    <Preis>99,00</Preis>
  </Zusatzinfo>
  <ISBN>013185858</ISBN>
 </Buch>
</Buecher>
```

Which problems arise at data level?

- In addition: Data Integration!!



- These structural problems can be SOMEHOW solved.
- Even harder: semantical problems → Example?


Step 4: Definition of Integration and Cleaning Strategy
Step 3: Analysis of Data (Quality)
Step 2: Definition of Analysis Format (Model)
Step 1: Definition of Analysis Goals
Data Sources
Data Collection
Data Extraction
Data Cleaning
Data Integration

Contents

Goals

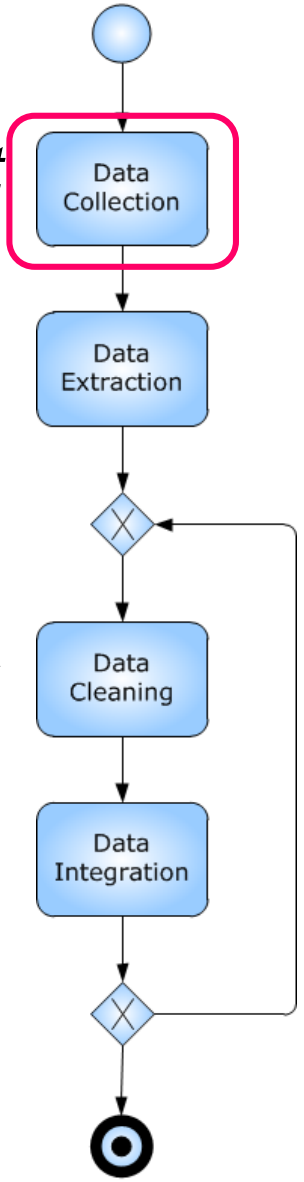
This chaper aims at conveying approaches, techniques, and tools to build an integrated data basis for an BI project, in particular:

- Understanding challenges in obtaining and integrating data
- Learning basic techniques of data extraction
- Understanding challenges and learn techniques for improving data quality
- Getting to know different data integration formats
- Understanding how to determine a data integration strategy
- Understanding challenges and learn techniques for data integration in different target formats
- Getting to know use cases from different domains

Contents

3 Data Extraction

- Remember: „*It‘s all about the data. […] But data doesn’t come to you…*”[1]
- In practice different situations
- Data sources are already existing (and accessible) → assumed in literature, practically not always the case
- Nonetheless, the relevant sources have to be selected
- Necessary data is collected „on-demand“ (or in the right format)
- Conclusio 1: **Data collection** is an active task



[1] http://mashable.com/2009/12/23/marketing-data/

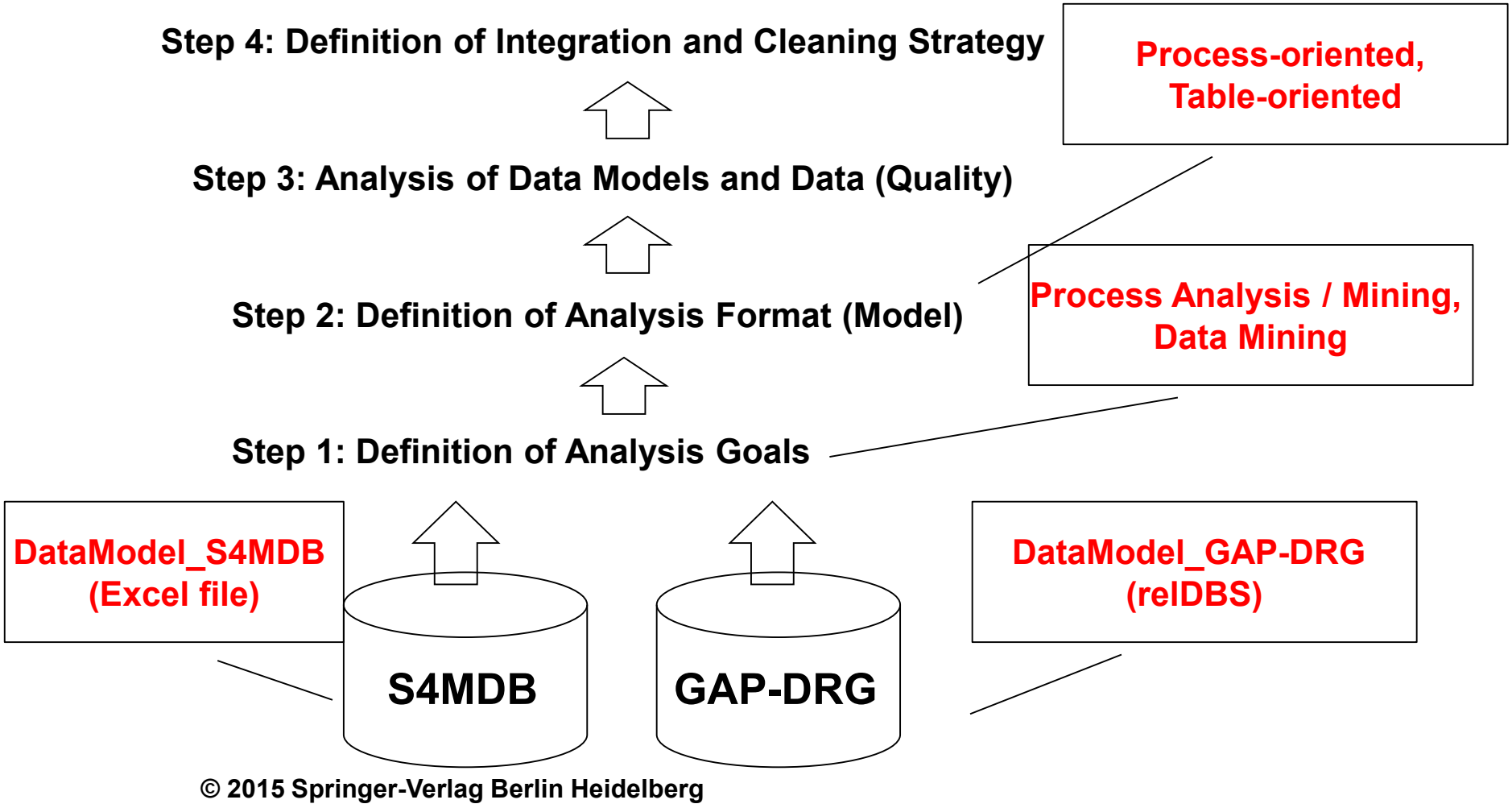
Conclusio 1: **Data collection** is an active task

- Identification of relevant data sources
- Clarification of issues such as data access (particularly, if external data sources are to be accessed)
- **Use Case 1: Patient treatment processes**
- EBMC^2 project[5]: co-funded by University of Vienna and Medical University of Vienna
  - Formalizing medical guidelines for skin cancer treatment
  - Mining and analysis of real-world treatment processes
  - In particular regarding their compliance with the guidelines
  - Selected Key Performance Indicators:
    - Survival time
    - Health status of a specific group of persons
    - Cost effectiveness of certain health policies

[5]R. Dunkl, M. Binder, W. Dorda, K. A. Fröschl, W. Gall, W. Grossmann, K. Harmankaya, M. Hronsky, S. Rinderle-Ma, C. Rinner, S. Weber: On Analyzing Process Compliance in Skin Cancer Treatment: An Experience Report from the Evidence-Based Medical Compliance Cluster (EBMC2). Int'l Conf. on Advanced Information Systems Engineering (CaISE 2012), pp. 398-413 (2012)

- Balance between:
  - What data sources do we need (to fulfill a certain analysis goal) and
  - Which data sources are actually available and accessible (privacy, data ownership, data access costs, etc.)
- Available data sources:
  - detailed data collection of clinical Cutaneous Melanoma (CM) stage IV protocols (Stage IV Melanoma Database, **S4MDB**, for short)
  - administrative data of the Main Association of Austrian Social Security Institutions comprising a billing-oriented view of medical patient treatments (**GAP-DRG**)

3 Data Extraction



3 Data Extraction

Patient	Id	GivenName	Surname	BirthDate

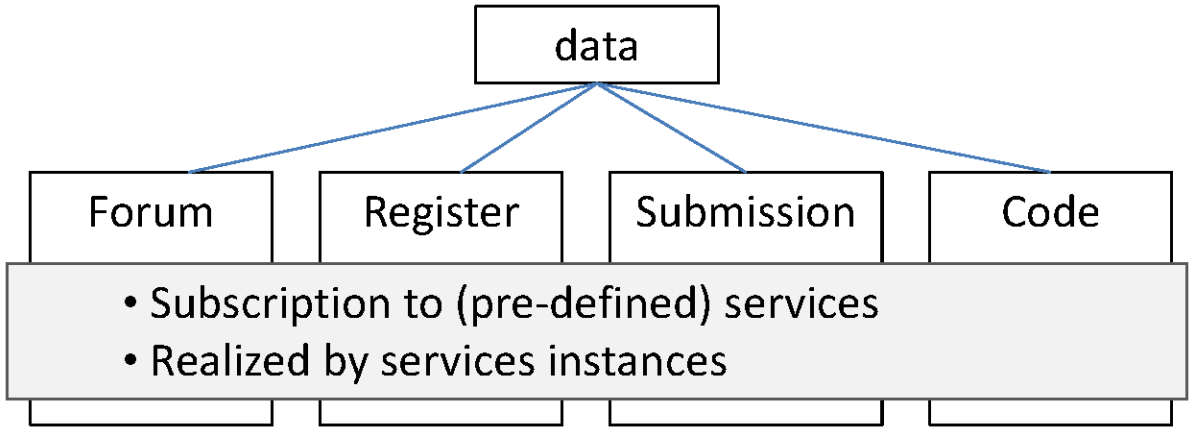
Treatment	Id	Code	Label

HospitalStay	Id	PatientId	Admission	Discharge

StayTreatment	Id	TreatId	StayId	made

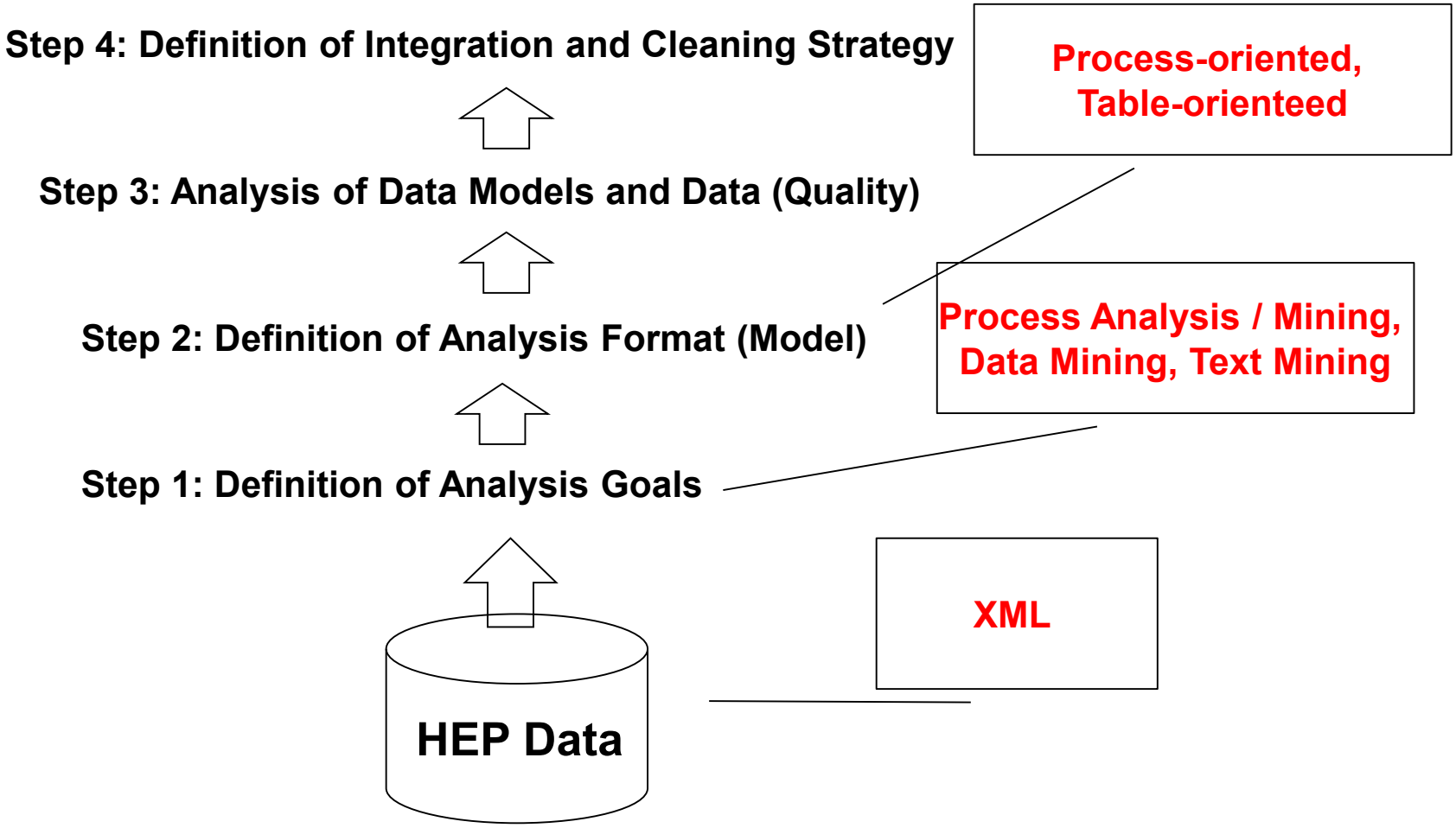
S4MDB

- Use Case 2: Higher-Education Data (HEP)
- Data source for practical project in Summer 2013
- Collected from service-oriented learning platform CEWebs



- Main analysis questions:
  - Analysis of learning processes
  - Mining of reference processes
- Selected key performance indicators:
  - Success of learning techniques (e.g., forum)
  - Flexibility degree (i.e., analyzing deviations from reference process)

3 Data Extraction

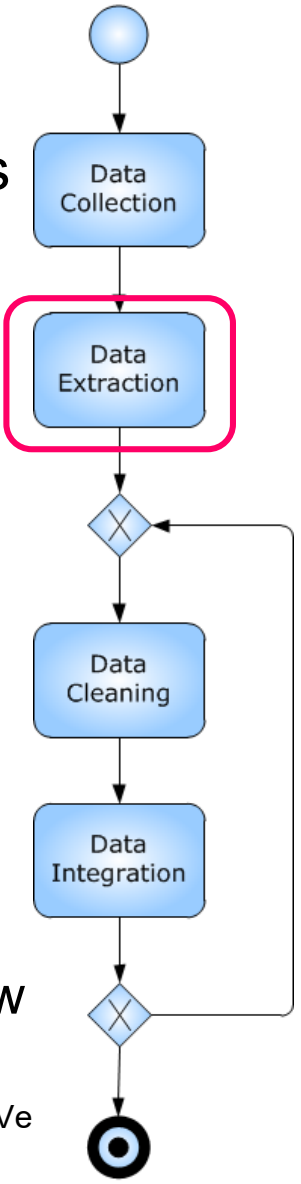


Further Use Cases, taken from Business Process Intelligence Challenge

- BPIC 2014: IT-Management:
  - Rabobank Group ICT
  - Implementation of frequent software releases managed by ITIL processes
  - Analysis of underlying change processes to predict the workload faced by Service Desks and IT Operations
- BPIC 2015: Municipalities (NL) - Building Permits
  - Collection of building permit application data by several municipalities
  - Understand the processes and roles of the participants, and differences in the execution between municipalities
- BPIC 2016: Customer Contacts
  - Employee Insurance Agency (NL)
  - Focus on Customers' utilization of various communication channels
  - Analysis of the customer behavior

3 Data Extraction

- After selecting and / or collecting data sources, data has to be extracted
- Data extraction is a rather technical question:
- Classically: ETL (Extraction – Transformation – Load)
- Access to heterogeneous data sources
  - Depends on the type of data source
  - Important: do we need the the entire data (or fragments) OR do we need a data update (delta file)?
  - Example (relational) databases: offer access by query language (SQL), but also by logging
  - Example legacy systems: do not offer any support → many approaches for determining snapshot deltas, e.g., by Window algorithm[6]



[6]W.J. Labio, H. Garcia-Molina: Efficient Snapshot Differential Algorithms for Data Warehousing. In Proc. Ve Large Databases, pp. 63 - 74 (1996)

Commercial Tools:

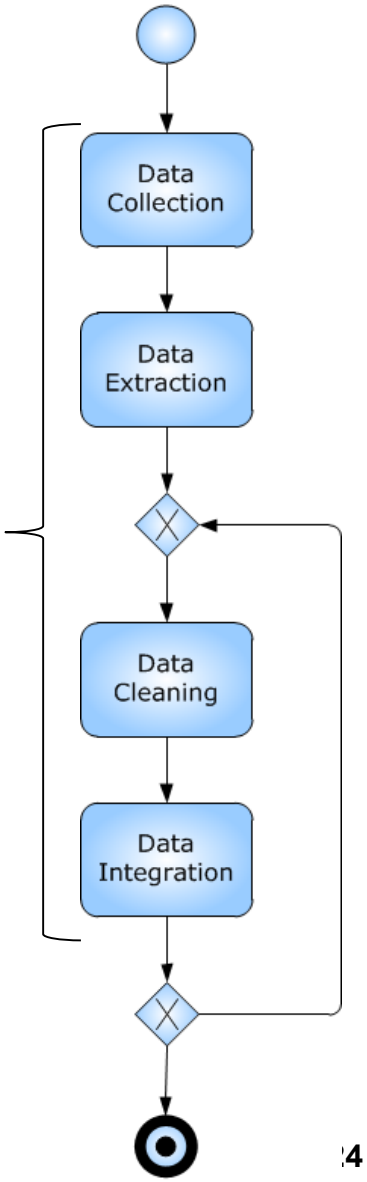
- SQL Server Integration Services (included in Microsoft SQL Server product line)
- Oracle Data Integrator
- SAP BusinessObjects Data Integrator
- SAS Data Integration Server

Open Source / Dual-licensed

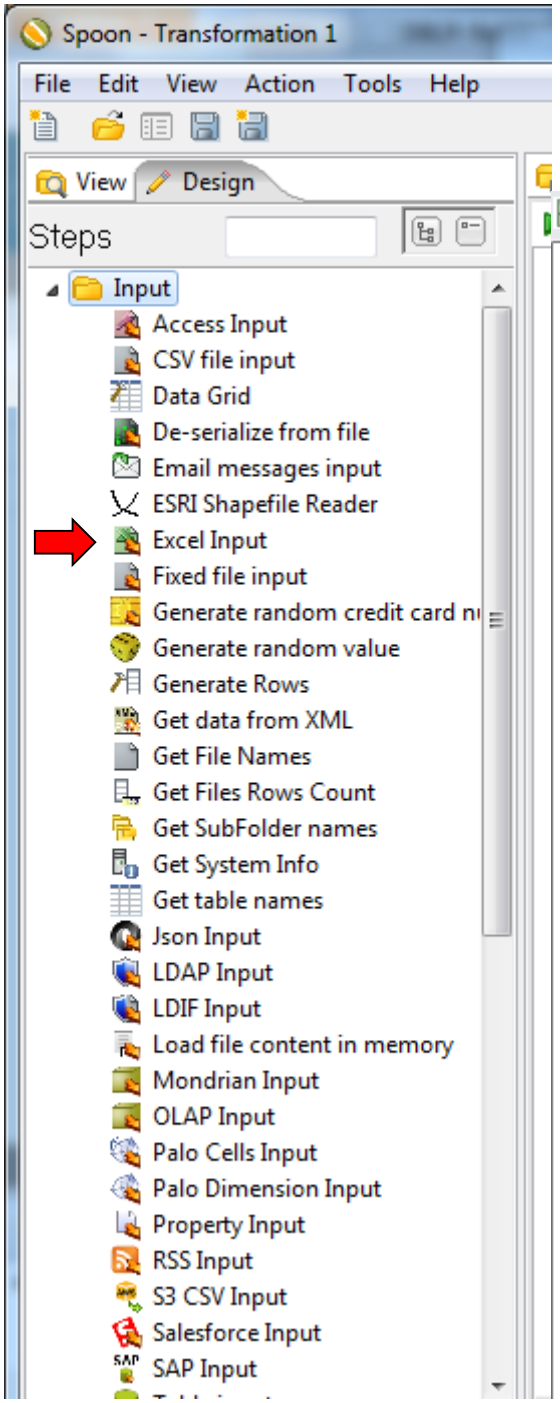
- Pentaho
- Talend Open Studio

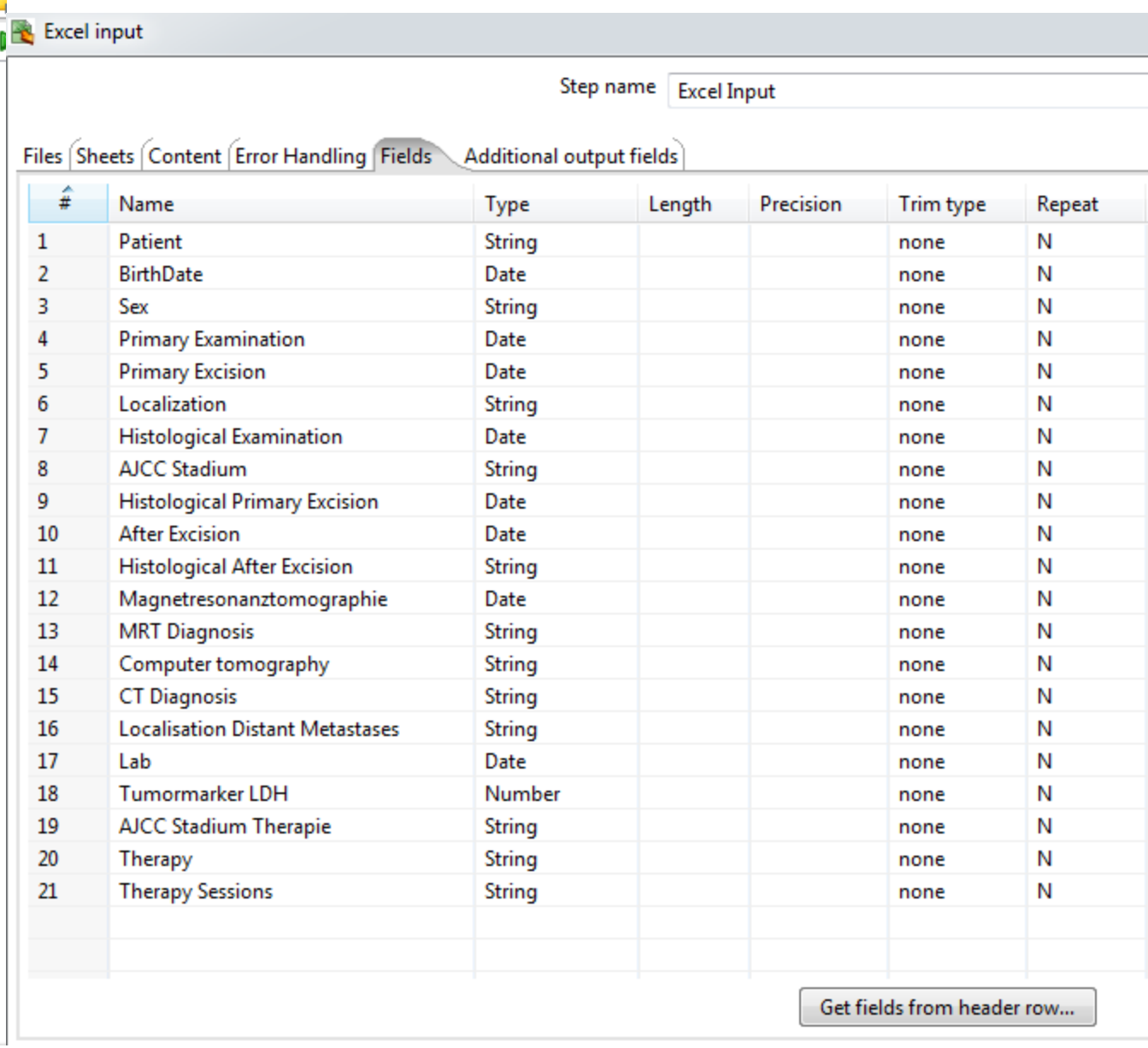
Pentaho

- Commercial: http://www.pentaho.com/
- Open source: http://community.pentaho.com/
- Product family:
  - BI Server und Admin Tool: BI Server 3.7.0
  - **Data Integration: Spoon 4.1.0**
  - Data Analysis: PAT - Pentaho Analysis Tool 0.8
  - Reporting: Report Designer 3.7.0
  - Data Mining: Weka 3.6



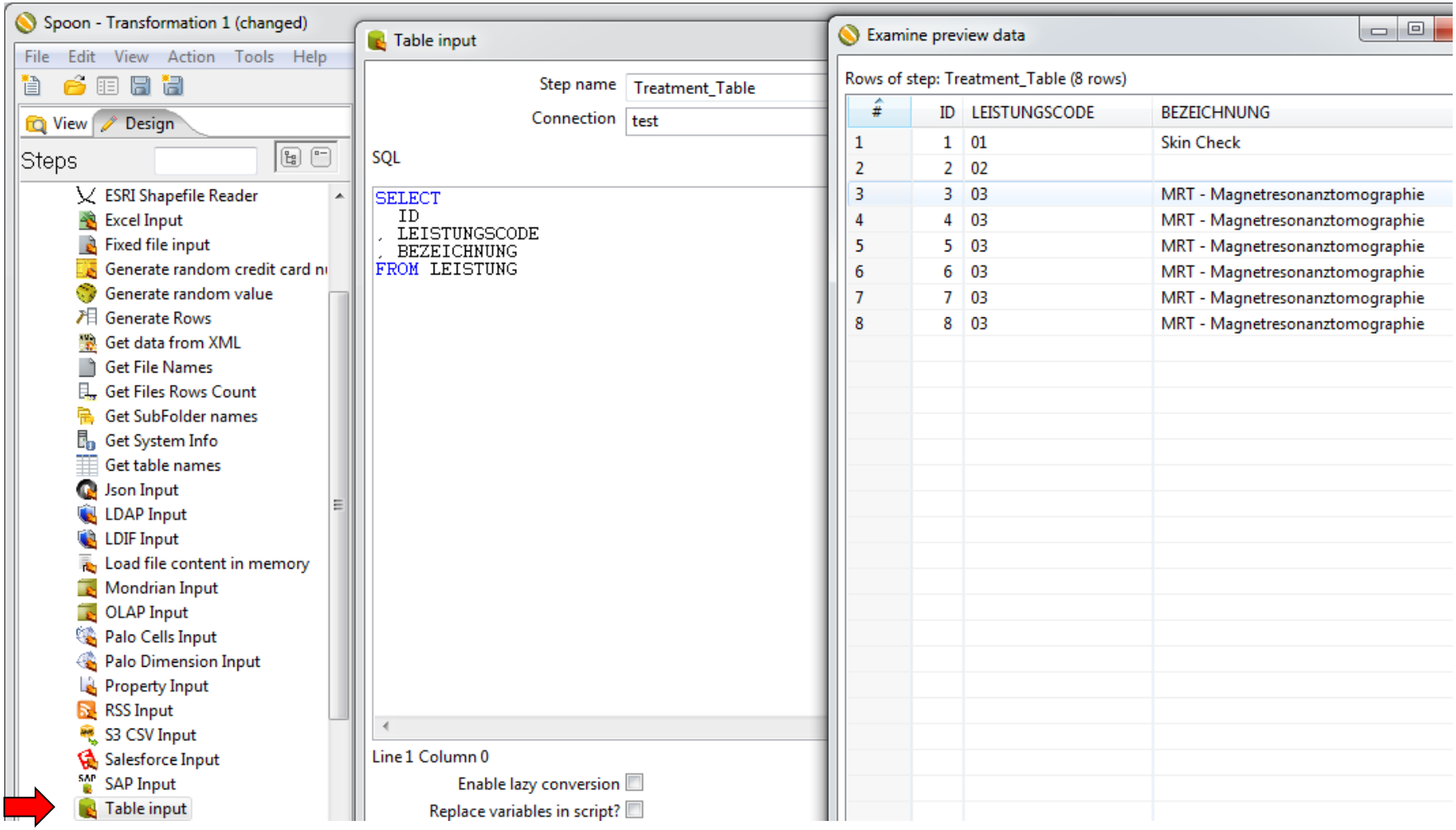
Connecting EBMC² Data Sources in Pentaho





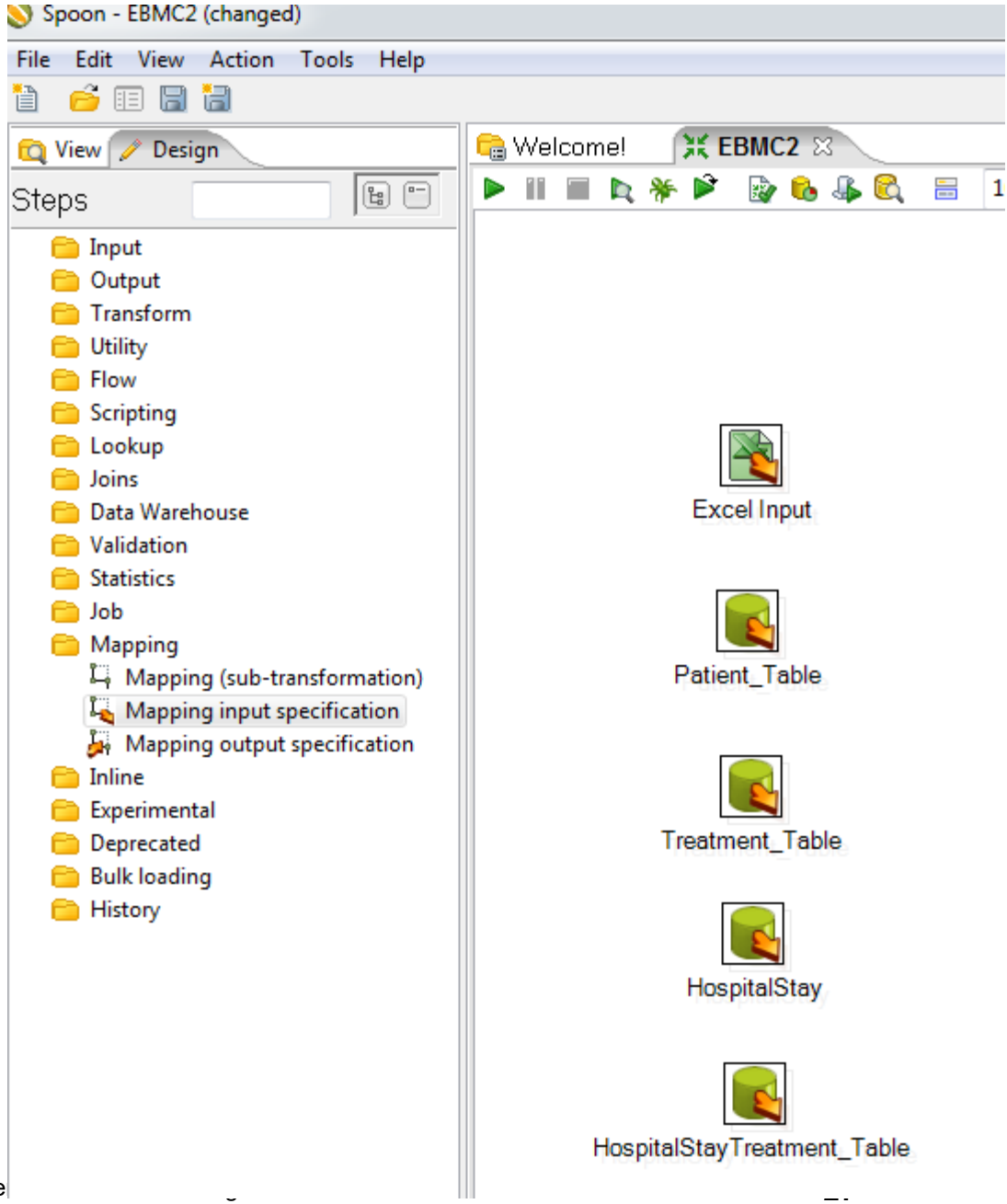
#	Name	Type	Length	Precision	Trim type	Repeat
1	Patient	String			none	N
2	BirthDate	Date			none	N
3	Sex	String			none	N
4	Primary Examination	Date			none	N
5	Primary Excision	Date			none	N
6	Localization	String			none	N
7	Histological Examination	Date			none	N
8	AJCC Stadium	String			none	N
9	Histological Primary Excision	Date			none	N
10	After Excision	Date			none	N
11	Histological After Excision	String			none	N
12	Magnetresonanztomographie	Date			none	N
13	MRT Diagnosis	String			none	N
14	Computer tomography	String			none	N
15	CT Diagnosis	String			none	N
16	Localisation Distant Metastases	String			none	N
17	Lab	Date			none	N
18	Tumormarker LDH	Number			none	N
19	AJCC Stadium Therapie	String			none	N
20	Therapy	String			none	N
21	Therapy Sessions	String			none	N

Connecting EBMC2 Data Sources in Pentaho



Pentaho

- Definition of Integration Workflow
- Input definition
- Output depends on integration strategy
- Integration Strategy depends on analysis goals
- Everything is defined manually
- Workflow is documented
- Can be replayed
- Changes in the data sources can be (semi-)automatically treated



- New Trend: Managing **big data**
  - Computational sciences
  - Cloud computing
  - Data from social networks
  - Sensors

3 Data Extraction

According to Beyer[7] challenges are

- *Data volume*:
  - Data becomes „too big“ for (relational) databases → Big Tables, NoSQL
  - “*Too much volume is a storage issue, but too much data is also a massive analysis issue*.”[7] → MapReduce, BigQuery
- *Data velocity:*
  - Data extraction during runtime
  - Continuous data streams (e.g., produced by sensors)
- *Data Variety:*
  - Structured versus unstructured data
  - Cross-sectional vs. event-based data
  - Text, images, videos

[7]Beyer, Mark. "Gartner Says Solving 'Big Data' Challenge Involves More Than Just Managing Volumes of Data". Gartner. http://www.gartner.com/it/page.jsp?id=1731916

Data volume

- NoSQL databases, not based on tables as basic data structures, instead:
  - Document-stores (→ data variety)
  - Graph databases
  - Key-Value storage systems
- Commercial solutions:
  - Google's BigTable: https://cloud.google.com/bigtable/
  - Amazon's Dynamo: https://aws.amazon.com/de/dynamodb/
  - Facebook's Cassandra
- Open Source solutions:
  - Apache Hadoop
  - Key-Value storage systems

Graph databases

- Before RDBMS: CODASYL and IMS databases (still running in many enterprises!)
- The data is represented as graph structure
- Queries navigate on the graph structure
- In principle well suited for handling large data sets: WHY?

– Example sones GraphDB (sones.de)
  - Combining object-oriented aspected and graph database
  - Basic structure: graph G:=(V, E) with V set of vertices and E set of edges
  - Definintion in Graphical Query Language (GQL):

```
CREATE VERTEX TYPE Person

ATTRIBUTES (SET<Person> Debitors,
SET<Person> Buddies, String name)

INCOMINGEDGES (Person.Debitors ows,
Person.Buddies friendOf)
```

Definition of the vertex types

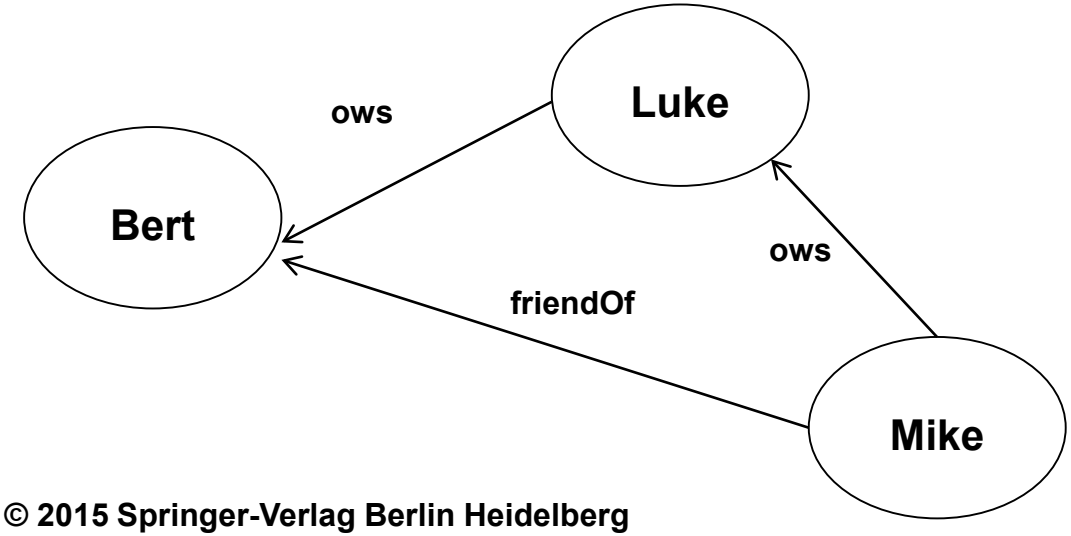
Set definition → object-orientation

Definition of the edges

INSERT INTO Person Values (name='Bert')

INSERT INTO Person VALUES (name = "Luke", Debitors = SETOF(name = "Bert"))

INSERT INTO Person VALUES (name = "Mike", Debitors = SETOF(name = "Luke"), Buddies = SETOF(name = "Bert"))



3 Data Extraction

Queries:

FROM Person

SELECT name, Debitors, Buddies

Selection of Result Mike


```
      "Properties": {
        "name": "Mike"
      }
    },
    {
      "Edges": [
        {
          "HyperEdgeView": {
            "Debitors": [
              {
                "SingleEdge": [
                  {
                    "Properties": []
                  },
                  {
                    "TargetVertex": [
                      {
                        "Properties": {
                          "VertexTypeID": "-9223372036854775782",
                          "VertexID": "-9223372036854775807"
                        }
                      },
                      {
                        "Edges": []
                      }
                    ]
                  }
                ]
              }
            ]
          }
        },
        {
          "HyperEdgeView": {
            "Buddies": [
              {
                "SingleEdge": [
                  {
                    "Properties": []
                  },
                  {
                    "TargetVertex": [
```

- *Key-Value* storage systems
- according to Agrawal et al.[8], they are
  - adopted by various enterprises.
  - Data analysis: MapReduce paradigm
  - open-source implementation Hadoop
  - widespread adoption in industry and academia
  - Solutions to improve Hadoop systems' usability and performance

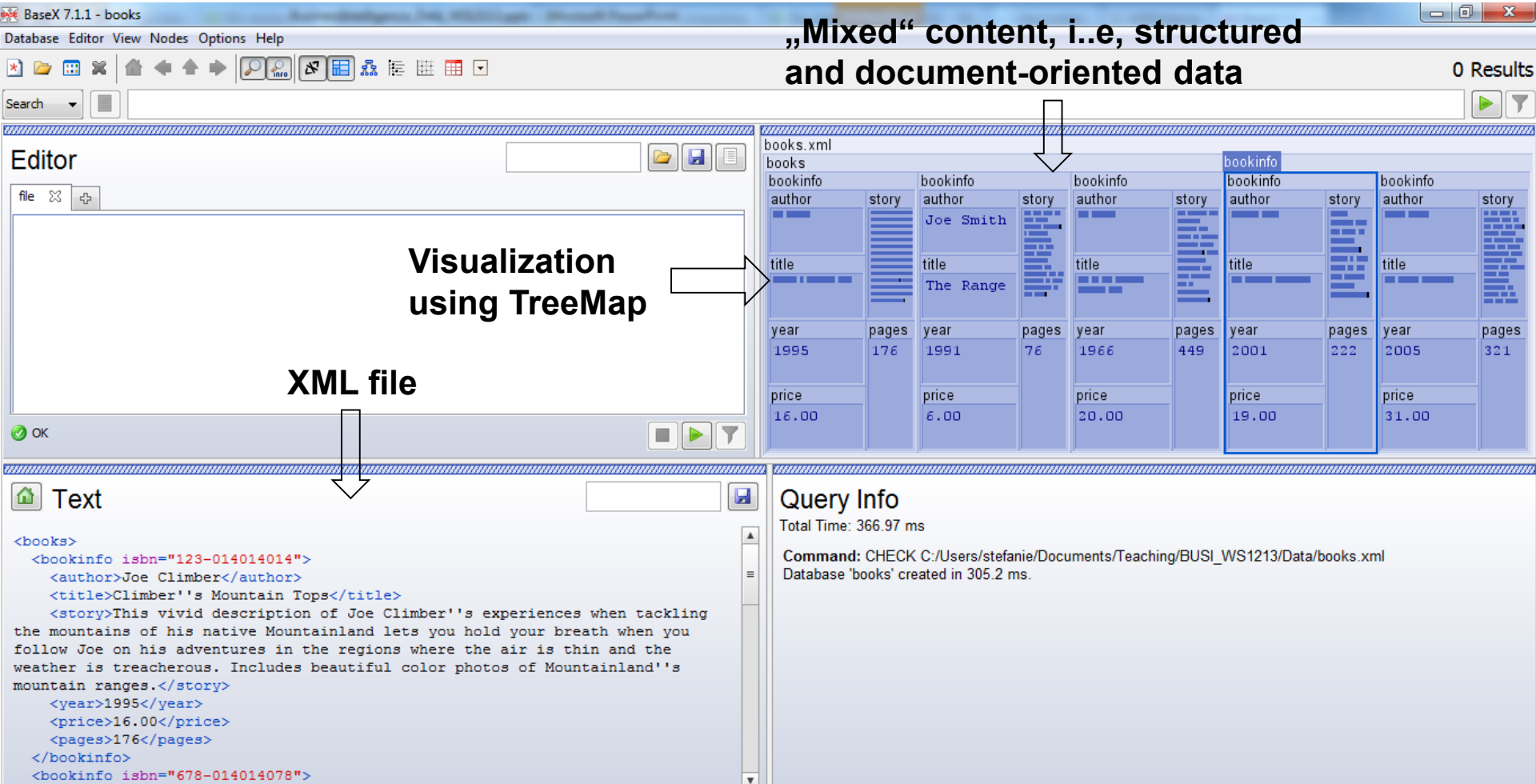
[8]Divyakant Agrawal, Sudipto Das, and Amr El Abbadi. 2011. Big data and cloud computing: current state and future opportunities. In *Proceedings of the 14th International Conference on Extending Database Technology* (EDBT/ICDT '11), ACM, pp. 530-533 (2011)

Data Variety

- Document-stores
  - Ready for storing unstructured data
  - 1st possibility: XML extensions on relational DBMS (SQLXML standard)
    - Example DB2 Express
      - New type XML
      - Can be queried using Xpath
  - By contrast: storing documents as CLOB, however, limited query functionalities (retrieval)
- 2nd possibility: XML databases
  - Example BaseX (http://basex.org/)
  - Stores XML files containing structured and unstructured, i.e., document-oriented content

3 Data Extraction

- XML databases continued
- BaseX:



Summary:

- Data variety / data heterogenity is an old and new problem
- Data extraction is a technical question, however, thoughts on data quality and later integration strategy are crucial
- Myriad of tools offer support
- However, definition and implementation of data cleaning and integration strategies (including mapping and definition of target formats) is manual job
- Tools support the definition, documentation of the process as well as support maintenance in case of changes

Summary:

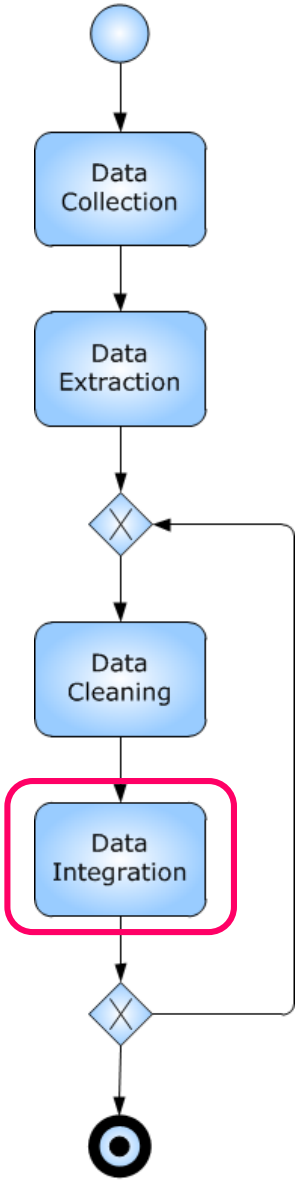
- New challenges mainly in data velocity, i.e., just-in-time data extraction becomes necessary
- Big data volume has led to looking for NoSQL databases such as Graph databases, Key/Value stores, document databases
- By contrast: extensions of RDBMS, Big Tables, etc.
- After discussion of data extraction techniques, crucial to discuss integration formats and data quality issues

Contents

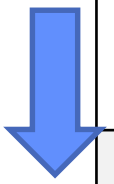
4 From transactional data towards analytical data

Very important:

- Integration format and
- Analytical format
- Not necessarily the same, but possible
- Analytical format depends on analysis questions and key performance indicators defined before
- Integration format depends on results of data extraction step + analytical format
- Also connected with data quality issues



4 From transactional data towards analytical data



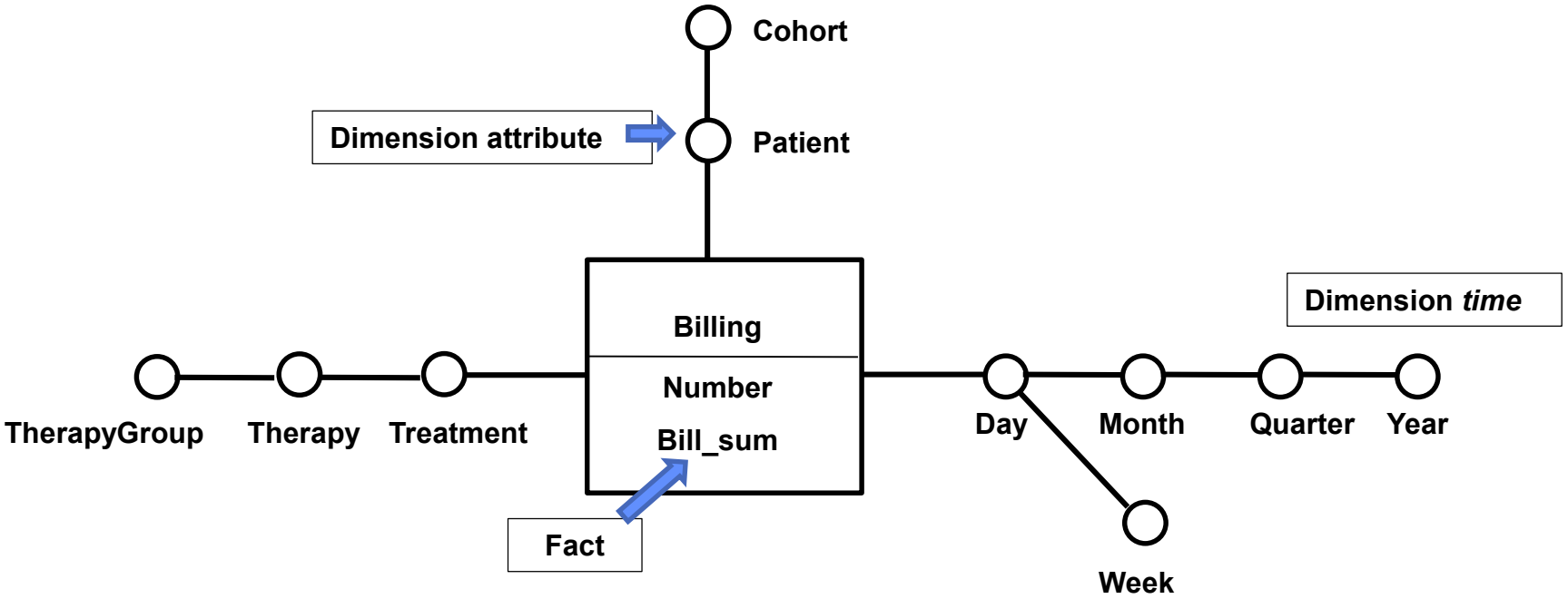
Integration / Analysis Formats		Structured Data Formats			Unstructured Data
		Flat (e.g., relational, CSV)	Hierarchical (e.g., XML)	Hybrid (e.g., XES)	Text
Table Formats	Flat	Contains / generates (mapping)	Generates (mapping)	Contains	Mining and generation
	Multidimensional	Generates (mapping & aggregation)		Generates (mapping & aggregation)	
Log		Generates (mapping & transformation)	Generates (mapping & transformation)	Contains or generates (transformation)	

Table Data

- Multi-dimensional table structures as usually used in Data Warehouse Systems
- Metaphor: Cube (but not necessarily three-dimensional)
- Basic components:
  - Data to be analyzed are called *facts* (e.g., profit)
  - Data can be analyzed along different *dimensions* (e.g., time, location)
  - Dimensions are divide into different *granularity* levels (e.g., day → month → year).
  - Facts can be *aggregated* along the dimensions (e.g., profit is aggregated from *profit per day* to *profit per year*)
  - Aggregation functions: *sum, count, min, max, average*

4 From transactional data towards analytical data

Queries in different classification directions and aggregation levels



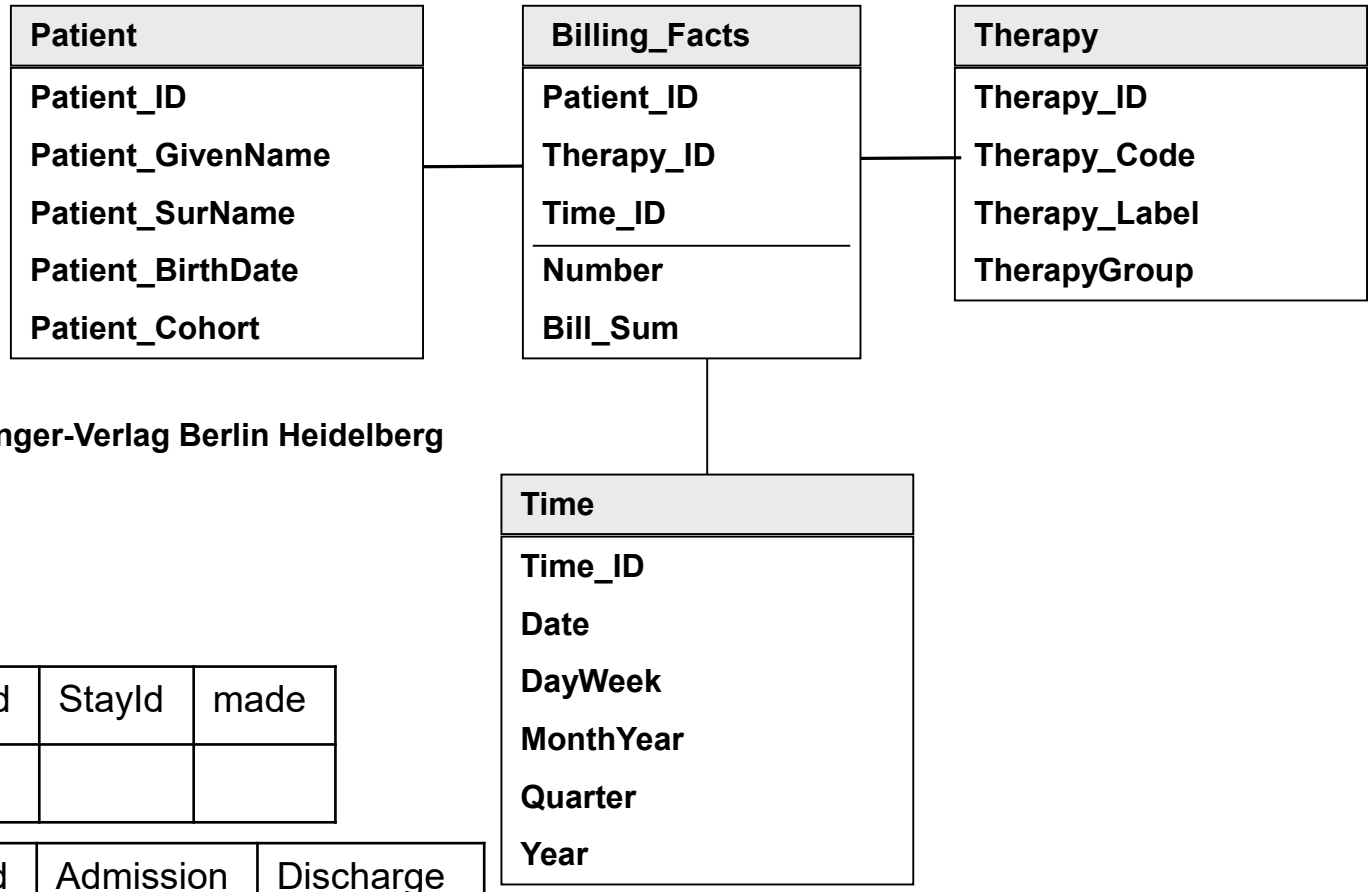
Report with aggregated facts

Billing		MRT	XRay	SkinCheck	SUM
2001	Cohort1	100	20	120	240
	Cohort2	50	30	40	120
	SUM	150	50	160	360
2002	Cohort1	110	30	100	240
	Cohort2	40	40	40	120
	SUM	150	70	140	360
2003	Cohort1	100	100	30	230
	Cohort2	10	10	40	60
	SUM	110	110	70	290
SUM		410	230	370	1010

Storage of multi-dimensional structures:

- Mapping to relational structures (ROLAP)
  - *Snowflake schema* (cf. e.g., Levene and Loizou[9]):
    - Fact table referencing dimension tables of lowest granularity
    - For each classification level one dimension table
    - Example: billing_facts references dim_time_day, dim_time_day references dim_time_month, dim_time_month references dim_time_year, etc.
    - Normalized
    - Possibly long join „chains“ when applying OLAP operations
  - *Star schema* (cf. e.g., Levene and Loizou[9]) One fact table
    - One table per dimension
    - De-normalized
- Multidimensional storage (MOLAP)
- Hybrid storage (HOLAP)

[9]Mark Levene, George Loizou, Why is the snowflake schema a good data warehouse design?, Information Systems 28(3): 225-240 (2003)



Database structure

StayTreatment	Id	TreatId	StayId	made

HospitalStay	Id	PatientId	Admission	Discharge

Treatment	Id	Code	Label

Patient	Id	GivenName	Surname	BirthDate

Star Schema:

- **De-normalized**
- **Compact**
- **Question: How do we get the billing facts in?**

4 From transactional data towards analytical data

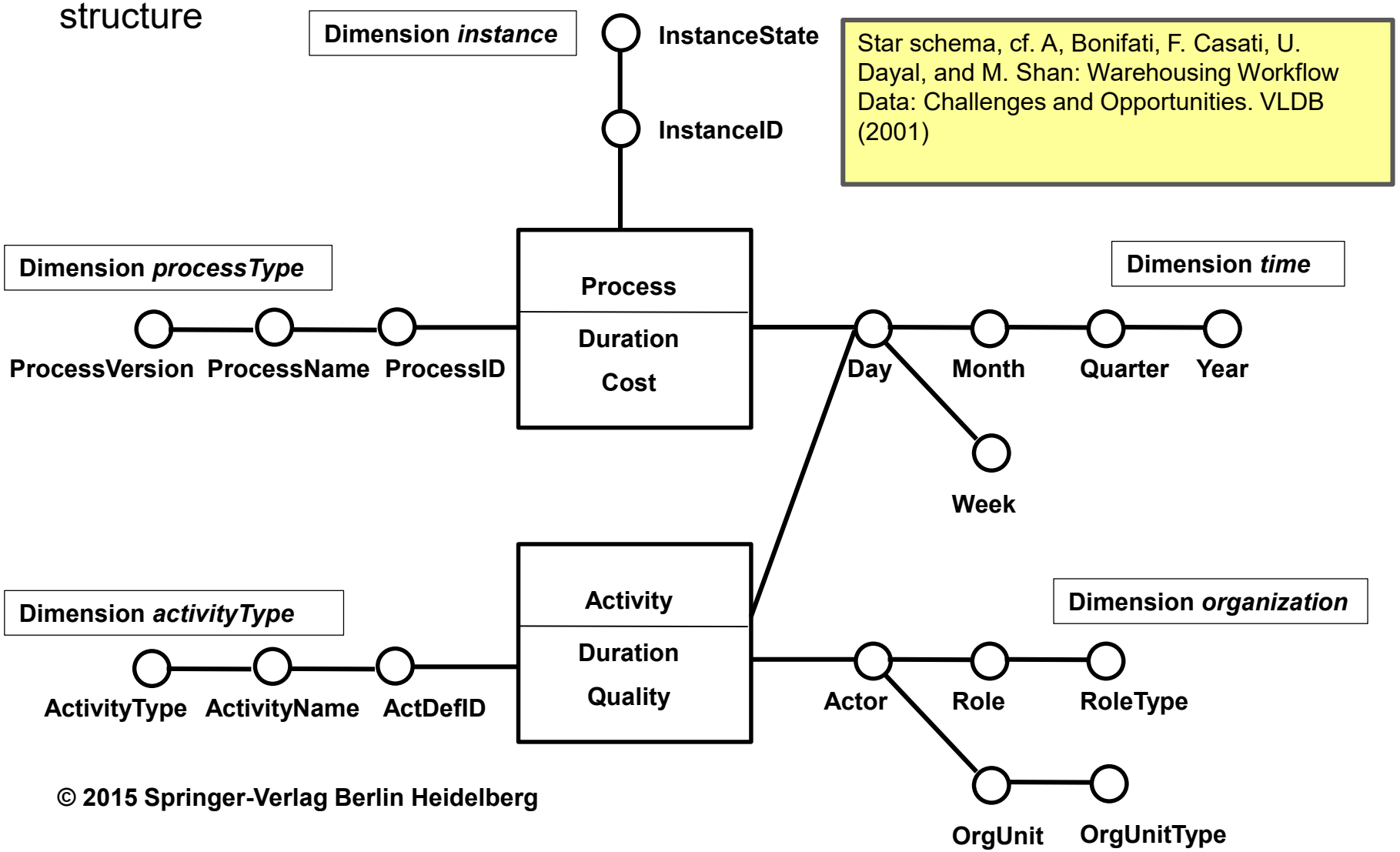
- create database starschemas4dbm
- create table s_patient(patient_id int not null primary key, patient_givenname varchar(20), patient_surname varchar(30), patient_birthdate date, cohort int)
- create table s_therapy(therapy_id int not null primary key, therapy_code varchar(30), therapy_label varchar(50), therapygroup varchar(50))
- create table s_time(time_id int not null primary key, date date, dayweek varchar(20), monthyear varchar(20), quarter int, year int)
- create table billing_facts(patient_id int not null, therapy_id int not null, time_id int not null, number int, bill_sum float, foreign key (patient_id) references s_patient(patient_id), foreign key (therapy_id) references s_therapy (therapy_id), foreign key (time_id) references s_time(time_id), primary key (patient_id, therapy_id, time_id))

4 From transactional data towards analytical data

- Basis is always the data at the lowest granularity level
  - Example entry in billing_fact table: (p177, t244, t855, 1, 20.5)
  - Example entry in time table: (t855, 2012-03-02 14:52:00, 5_9, 3_2012, 2012)
- Aggregation along dimensions by applying OLAP operations:
  - Specification fo adequate aggregation functions (e.g., SUM, AVG)
  - Drill-up / drill-down (roll-up / roll-down)
  - Slice and dice
  - Can results in pre-defined data warehouses / data marts
  - Example: number of patients in March 2012 (using SUM)
  - Pentaho: Reporting
- Application of Data Mining techniques
  - Pentaho: WEKA (http://www.cs.waikato.ac.nz/ml/weka/)

4 From transactional data towards analytical data

– Process Warehousing: describing process-oriented data as multi-dimensional structure



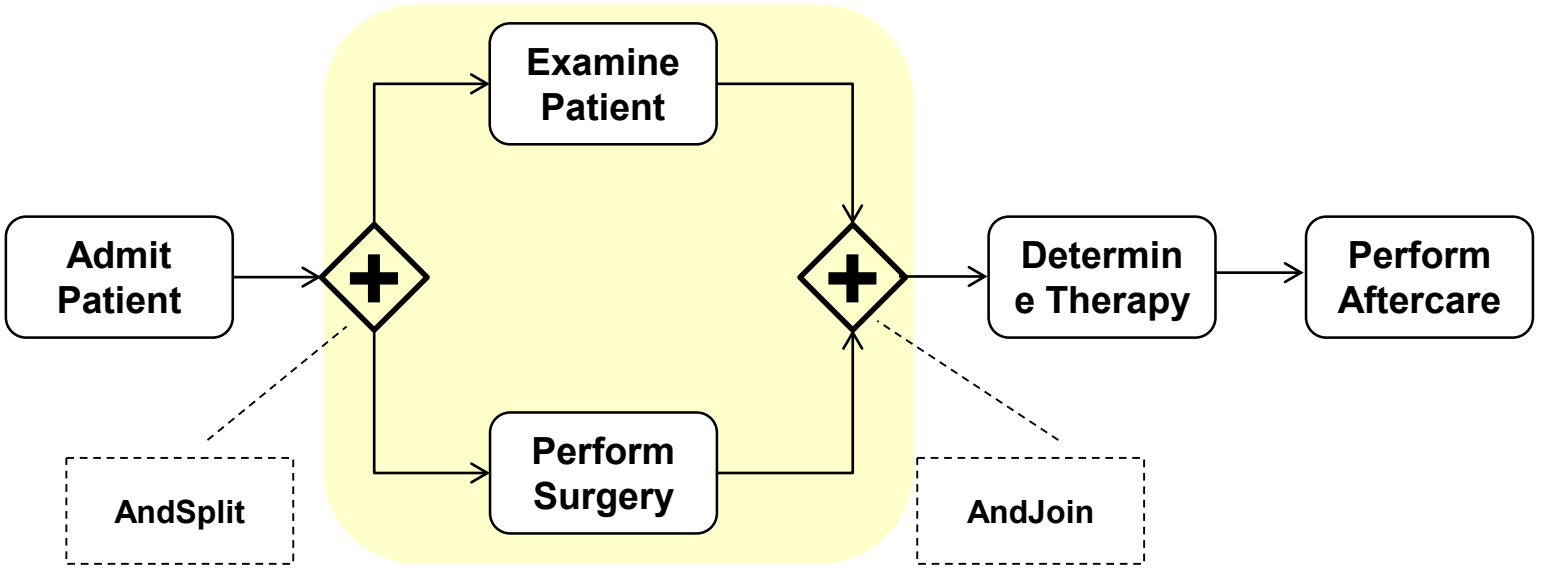
Log Data: Collection of events recorded during runtime of an information system, examples:

- Database logs: record transaction states, e.g., BOF, EOF → fundament for restoring database after crash (recovery)
- Process logs: record events produced during execution of a process-oriented application (e.g., a workflow)
- Other logs: record event logs produced by an ERP system

– Basis for process-oriented analysis

- Compliance checking and monitoring
- Process mining
- Process conformance checking
- Process performance analysis

- Log formats used in the process community:
  - MXML: http://www.processmining.org/tools/mxmlib
  - XES (eXtensible Event Stream): http://www.xes-standard.org/
    - → Combines table and log data



a) MXML

```
<AuditTrailEntry>
  <WorkflowModelElement>5
PerformSurgery</WorkflowModelElement>
  <EventType>start</EventType>
  <Timestamp>2012-10-
02T13:56:36.075+01:00</Timestamp>
  <Originator>unknown</Originator>
</AuditTrailEntry>

<AuditTrailEntry>
  <WorkflowModelElement>5
PerformSurgery</WorkflowModelElement>
  <EventType>complete</EventType>
  <Timestamp>2012-10-
02T13:56:36.078+01:00</Timestamp>
  <Originator>unknown</Originator>
</AuditTrailEntry>

<AuditTrailEntry>
  <WorkflowModelElement>4
ExaminePatient</WorkflowModelElement>
  <EventType>start</EventType>
  <Timestamp>2012-10-
02T13:56:36.080+01:00</Timestamp>
  <Originator>unknown</Originator>
</AuditTrailEntry>
```

b) XES

```
 <event>
  <string key="org:resource" value="unknown"/>
  <date key="time:timestamp" value="2012-10-
02T14:56:36.075+02:00"/>
  <string key="concept:name" value="5
PerformSurgery"/>
  <string key="lifecycle:transition" value="start"/>
</event>


<event>
  <string key="org:resource" value="unknown"/>
  <date key="time:timestamp" value="2012-10-
02T14:56:36.078+02:00"/>
  <string key="concept:name" value="5
PerformSurgery"/>
  <string key="lifecycle:transition" value="complete"/>
</event>


<event>
  <string key="org:resource" value="unknown"/>
  <date key="time:timestamp" value="2012-10-
02T14:56:36.080+02:00"/>
  <string key="concept:name" value="4
ExaminePatient"/>
  <string key="lifecycle:transition" value="start"/>
</event>
```

Import / transformation event-based data directly into process-oriented log formats (MXML, XES)

- Challenges:
  - distributed sources
  - different format
- → information integration problem
- Minimum data requirements
  - case ID
  - events (START / END); order relevant (time stamps or ordered log)
- Additionally: performers, general data
- Useful tools:
  - ProM Import
  - DISCO (http://fluxicon.com/disco/)

4 Unstructured Data

- So far: structured data, i.e., data follows a data model
- Unstructured data → mostly text
- Estimation: about 85% of the data is unstructured; cf. https://www.business-standard.com/article/technology/-85-of-world-s-data-is-unstructured-106100301029_1.html
- Text is often analyzed using text mining
- Pre-processing becomes often necessary
  - Stemming
  - Removing stopwords
- → see part on text mining

Contents

Schema integration:

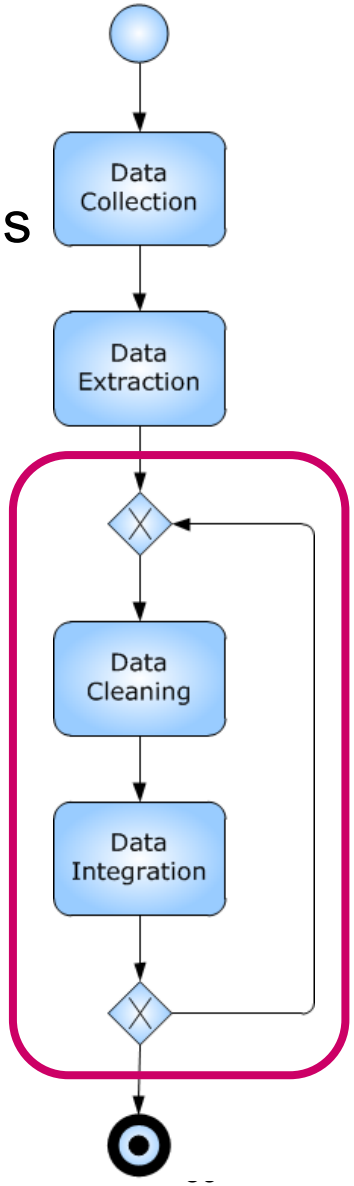
- Given participating schemata S_1, ... ,S_n
- Unite them into one integrated schema S* with (based on Batini et al.[10]):
- *Completeness*: no information loss with respect to the entities contained within schemata S_i, i = 1,...,n
- *Validity*: S* should reflect a real-world scenario that can be seen as a union of the real-world scenarios reflected by S_i, i = 1,...,n
- *No contradictions* within S*
- *Minimality*: no redundancies, every entity contained in S_i, i = 1,...,n should occur just once in S*
- *Understandability*: the transformation and integration steps should be documented in order to enable the traceability and reproducibility of the result

[10]Batini, Carlo, Maurizio Lenzerini, and Shamkant B. Navathe. "A comparative analysis of methodologies for database schema integration." *ACM computing surveys (CSUR)* 18, no. 4 (1986): 323-364.

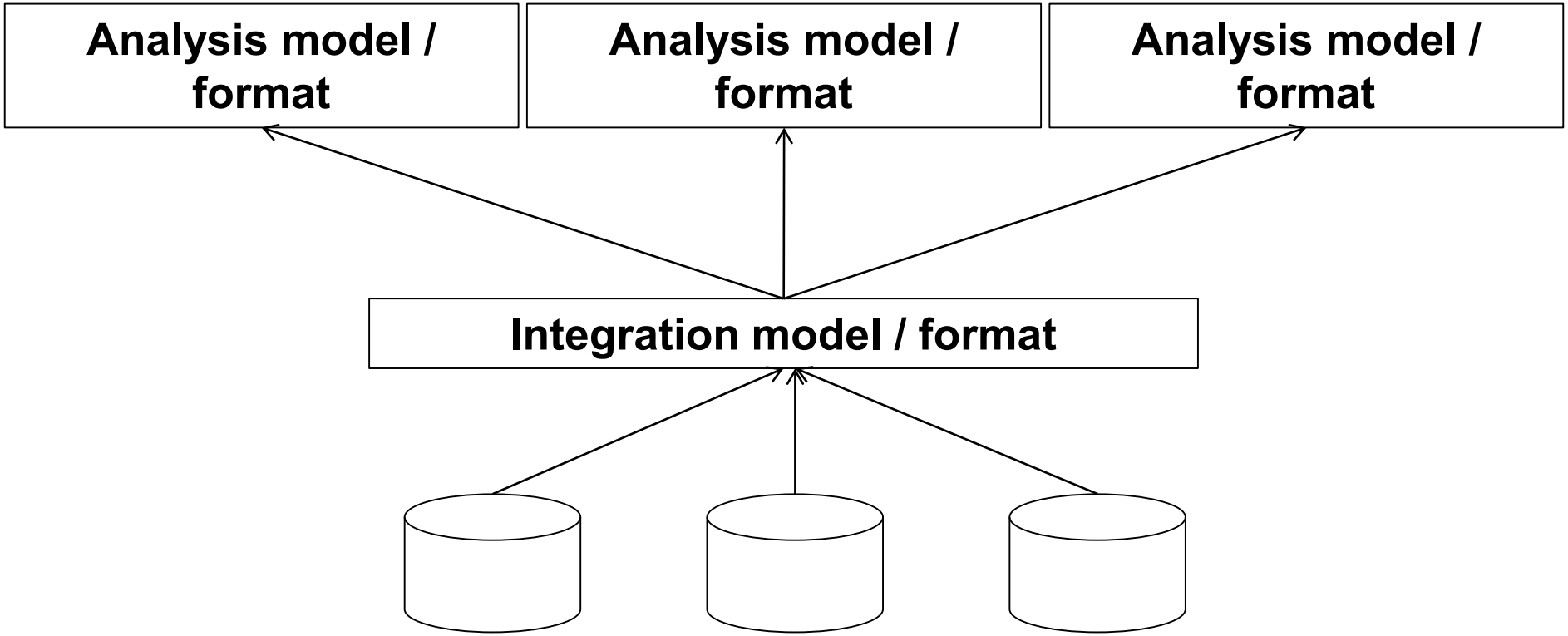
- General schema integration steps
  - pre-integration
  - schema comparison
  - schema conforming
  - schema merging and restructuring.
  - possibly iterative
- Schema matching and mapping as techniques for comparison and conforming of the participating schemas

Questions of pre-integration

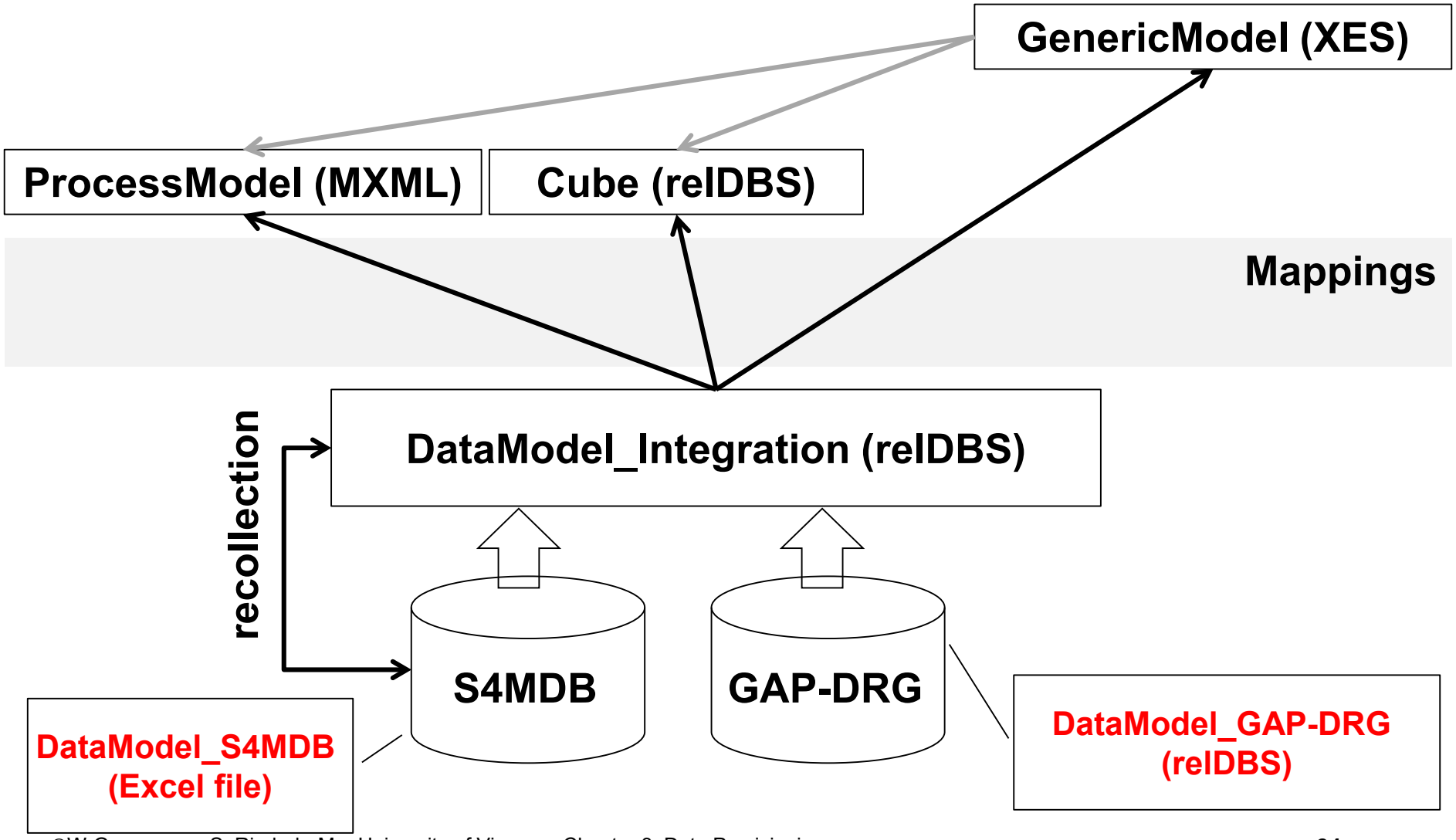
- Choosing the target format(s) depends on the analysis goals
- Questions for the integration:
  - Direct integration into target format (for direct import see last slides)
  - Integration into „intermediate“ integration model
- Reasons for choice?



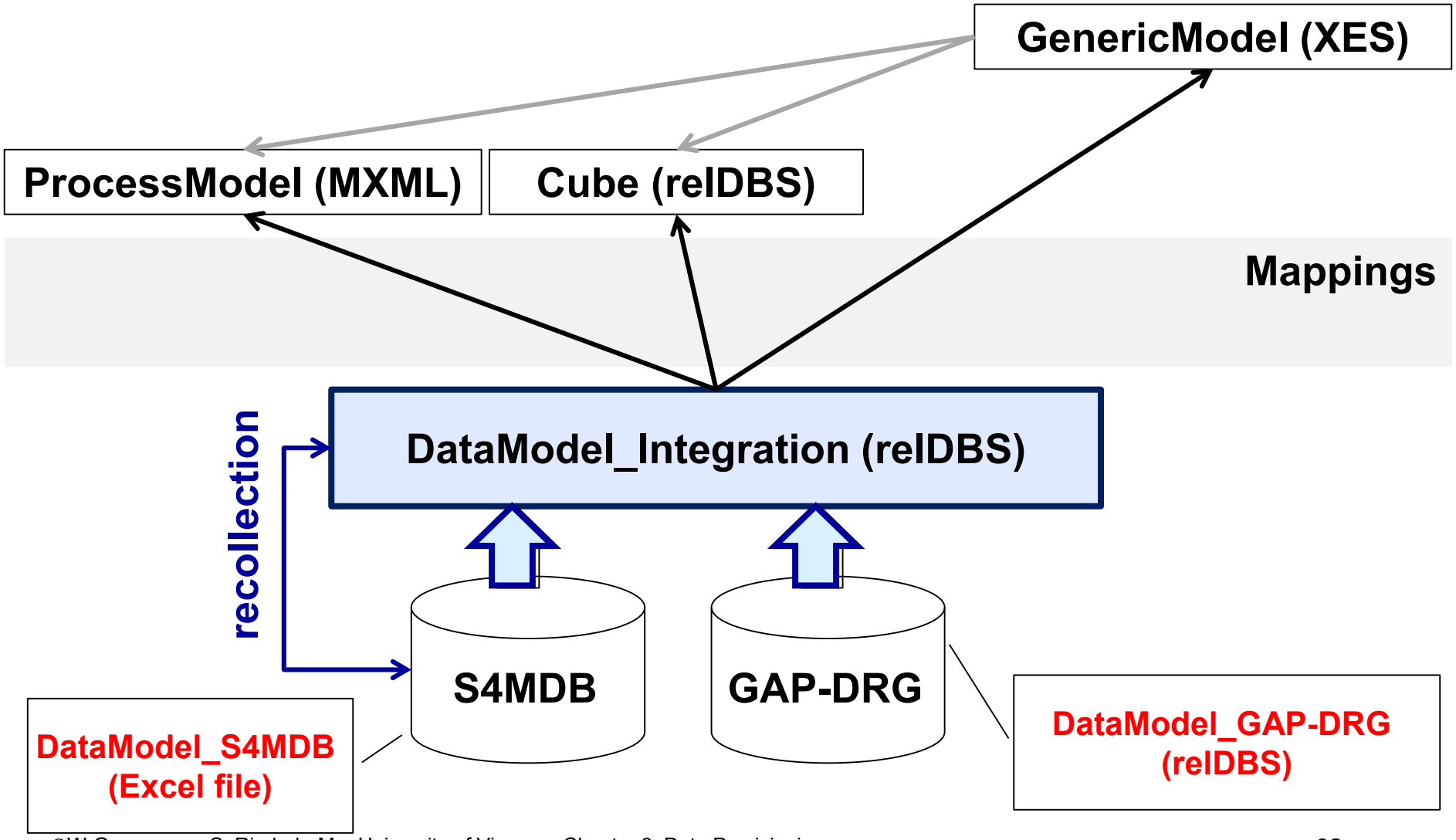
5 Schema and data integration

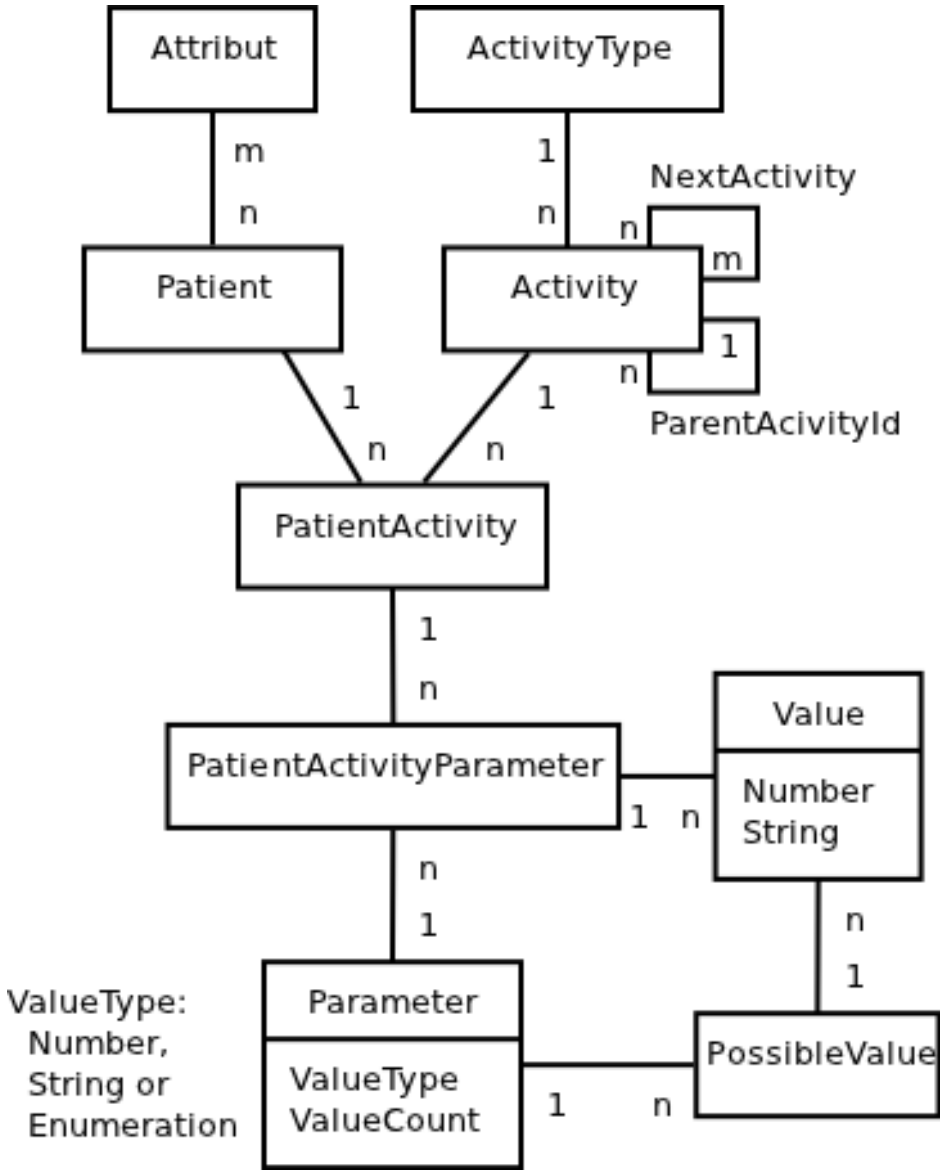


5 Schema and data integration



5 Schema and data integration





DataModel_Integration (relDBS)

- **Intermediate Integration Model**
- **Format: RDBMS**

- General Problems:
  - Heterogeneous data sources
  - Heterogeneous schemas
  - Heterogeneous data
- Consequence: several conflicts
  - Semantic
  - Descriptive
  - Heterogeneity
  - Structural

– Schema mapping

- Goal: based on two schemas as input a mapping between elements of these schemas that are semantically corresponding should be found
- Formally acc. to Bellahsene et al.[11]:

> Let S* and T* be two relational schemas. Then a mapping between S* and T* is defined as (S, T, m) where S is a relation in S* and T is a relation in T* and m is a set of attribute correspondences between S and T.

> At instance level: Let D_S and D_T be instances of S and T. Then D_S and D_T **satisfy** mapping m if for $\forall$ t_s in D_S $\exists$ t_t in D_T such that $\forall$ attribute correspondences (s, t) $\in$ m, the value of attribute s in t_s is the same value of attribute t in t_t.

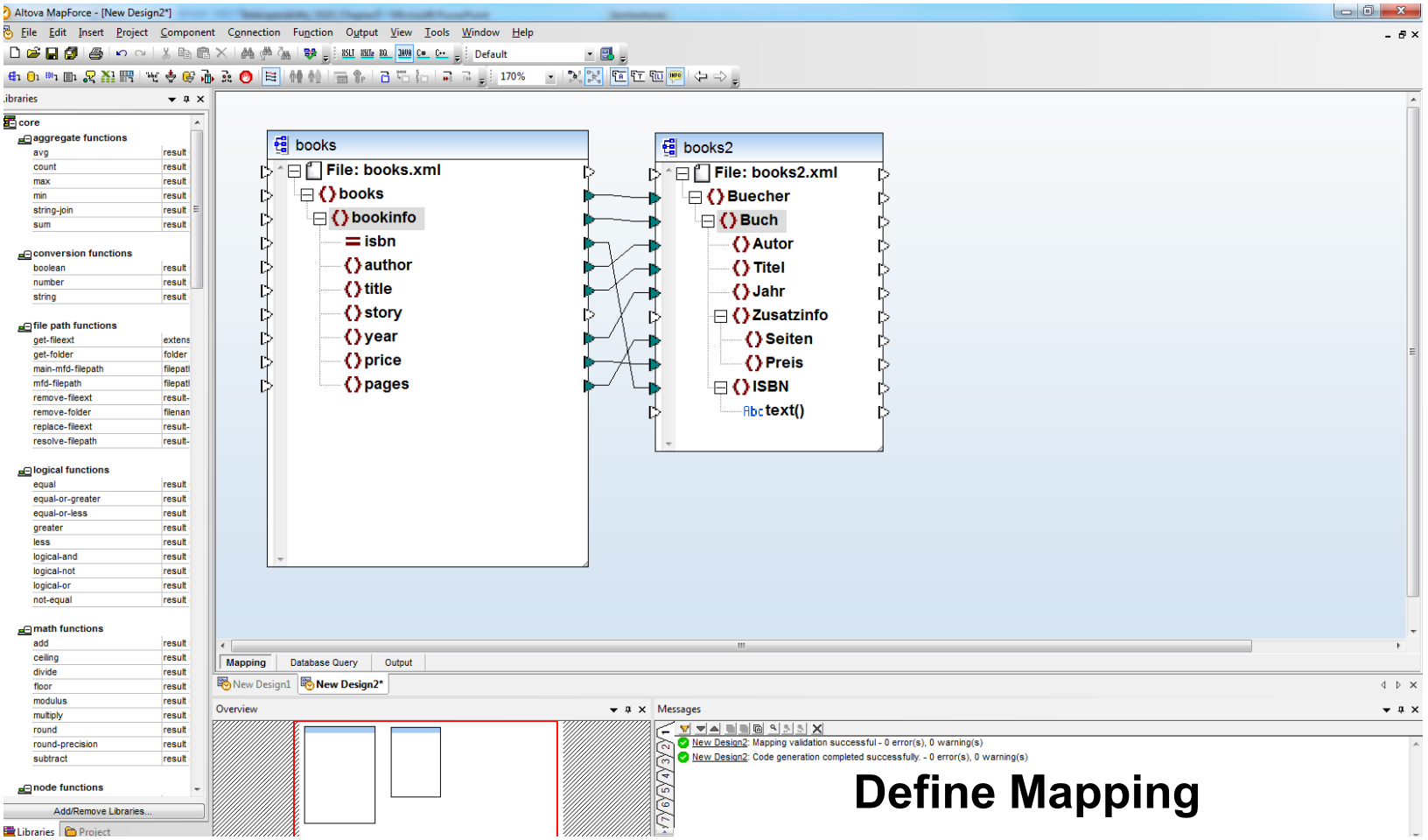
[11]Z. Bellahsene, A. Bonifati, E. Rahm: Schema Matching and Mapping. Springer (2011)

- Schema mapping, cf. Rahm and Do[12]
  - Manual task → errorprone and tedious
  - General algorithm:
    - Given two schemas S and T (two relations) with attribute sets A and B
    - Core idea:
      - Build cross product A × B between all attributes from A and B
      - For each pair calculate similarity
        - E.g., regarding attribute name
        - E.g., regarding stored data
      - Choose a mapping
      - Most similar pairs until threshold
      - In addition: consider constraints

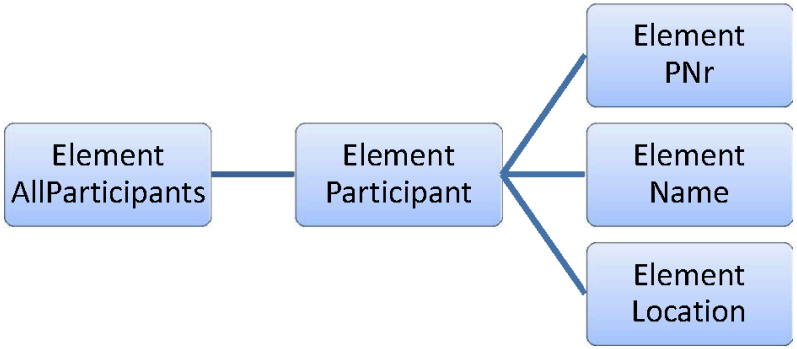
[12]Erhard Rahm, und Hong Hai Do. *Data Cleaning: Problems and Current Approaches*. IEEE Data Engineering Bulletin, 23(4):3-13 (2000)

5 Schema and data integration

Altova Mapforce, http://www.altova.com/mapforce.html



a) XML Schema (structure tree):



b) Relational Schema and Data
Database: AllParticipants

Participant		
PNr	**Name**	**Location**
171	Huber	Vienna
194	Brown	London

c) Publishing Statement in SQL/XML:

```
SELECT XMLSERIALIZE(
  XMLDOCUMENT(
    XMLELEMENT( NAME "All_Participants",
      XMLAGG(
        XMLELEMENT( NAME "Participant",
          XMLFOREST(T.PNr, T.Name, T.Location)))))
AS CLOB VERSION '1.0' INCLUDING XMLDECLARATION)
FROM Participant;
```

d) Result XML Document:

```
<All_Participants>
   <Participant>
      <PNR>171</PNR>
      <NAME>Huber</NAME>
      <Location>Vienna</Location>
   </Participant>
   <Participant>
      <PNR>194</PNR>
      <NAME>Brown</NAME>
      <Location>London</Location>
   </Participant>
</All_Participants>
```

Problems with data, cf. Leser and Naumann[13]

- Data errors
  - Different formats
  - Errors (e.g., typos)
  - Inconsistencies (e.g., zip code does not match city)
- Duplicates
- Data quality
  - Credibility
  - Relevance
- Completeness
  - Are all real world objects considered?
  - Do all attributes have values?

[13]U. Leser, F. Naumann: Information Integration. dpunkt (2007)

Dealing with data errors, cf. Leser and Naumann[13]

- Profiling:
  - Statistical analysis of the data, typically on numeric values
  - Pattern analysis
- Assessment:
  - Stating certain conditions on the data values, e.g., weight < 100 kg
- Stating measures:
  - Fixing data errors
  - Removing error sources
- Monitoring:
  - Controlling data quality

[13]U. Leser, F. Naumann: Information Integration. dpunkt (2007)

Data normalization cf. Leser and Naumann[13]

- (De-)capitalization
- Abbreviations / spelling: Str., street, Straße ←→ strasse, ...
- Stemming
- Names
- Formats:
  - Date: 18 February 2005, 18.02.2005, 2/18/05
  - Coding: 1: female, 2: male
  - Preciseness, field length, digits
  - Scales: grades, temperature, currency, etc.
- Good support by commercial systems (SQL)
- In addition: outlier detection, detection of duplicates

[13]U. Leser, F. Naumann: Information Integration. dpunkt (2007)

- Conflicts at data level
- Example

a) XML Document 1:

```
<All_Participants>
   <Participant>
      <PNR>244</PNR>
      <NAME>Smith, S.</NAME>
      <FEE>135.00</FEE>
      <Location>New York</Location>
   </Participant>
</All_Participants>
```

b) XML Document 2:

```
<All_Participants>
   <Participant>
      <PNR>244</PNR>
      <NAME>Sam Smith</NAME>
      <FEE>99.80</FEE>
      <Location>New York</Location>
   </Participant>
</All_Participants>
```

Contents

6 Summary & outlook

- Dealing with data is THE prerequisite of your BI project
- Often a complex and expensive task
- Calculate enough time and manpower
- Include the domain experts
- Document every step of the integration process (→ tool support)
- Always keep an eye on your analysis goals
- Also keep in mind maintenance issues (your data sources might be changing!)

6 Summary & outlook