universität wien

DICE – A Data Integration and Cleansing Environment

Wilfried Grossmann, Christoph Moser

Content

- Introduction
- Framework for Unified View
- A Procedural Format
- Business Modelling Techniques
- Data Preparation Techniques

Introduction

- Business Case, Idea:
  - An Insurance Company is interested in developing a new product
  - Goal: Increase sales by the new product
- Define a Business Plan

Introduction – Perspectives of Business

- Perspectives of the business plan
  - **Production and organisation oriented perspective**
  - **Customer oriented perspective**
  - Financial considerations

Introduction – Perspectives of Business

- A customer oriented perspective focuses on the marketing aspects of the business plan:
  - Analyse potential of the new product based on
    - Internal Data: Behaviour of already existing customers
    - External Data: Business environment, potential of new customers
  - Monitor execution of the process from external point of view (customer behaviour)
  - Modify / Optimize the marketing activities

Introduction – Perspectives of Business

- This perspective focuses on the following activities:
  - Definition of a business process model
  - Implementation the business process model
  - Monitoring execution of the business process model from internal point of view:
    - Monitoring without strategic considerations
    - Monitoring taking a global perspective
  - Changing / Optimization of the business process model according to results

Introduction – Perspectives of Business

- The different perspectives require
  - Different kinds of data organized in different ways
    - Data about the structure of the company
    - Data about already existing customers
    - Data of the monitoring activities
  - Different kinds of models
    - Process models
    - Business Analytics models
  - Different kind of analysis techniques
    - Process Analysis Techniques
    - Business Analytics Techniques

Introduction – Perspectives of Business

- Integration at different levels
  - Top level
  - **Operational level**
- Main topic of the lecture is a unified view on the perspectives at the operational level
  - How can Process Modelling and Business Intelligence support each other?

Framework for Unified Approach

- A unified approach has to
  - Look at the process from different perspectives (see Introduction)
  - Identify different types of processes
  - Identify the different views on data about the process
  - Define precisely goals and sub-goals
  - Define a method for organisation of analysis

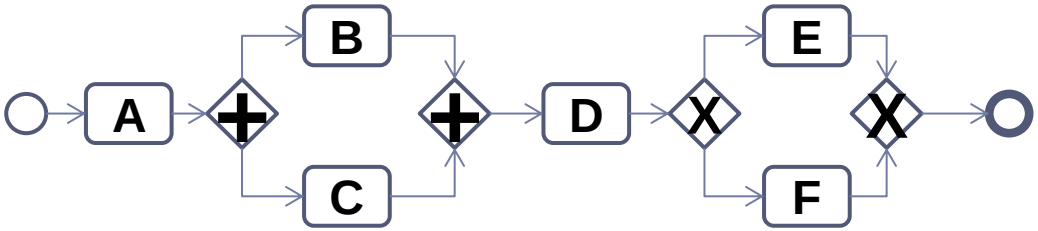
Framework for Unified Approach – Process Types

- In general we can understand all business activities as a process in time
- Business Process:

  A collection of related and structured activities necessary for delivering a certain good or service to customers,
  together with possible response activities of customers
- The addition of the customers gives the main distinction between the processes

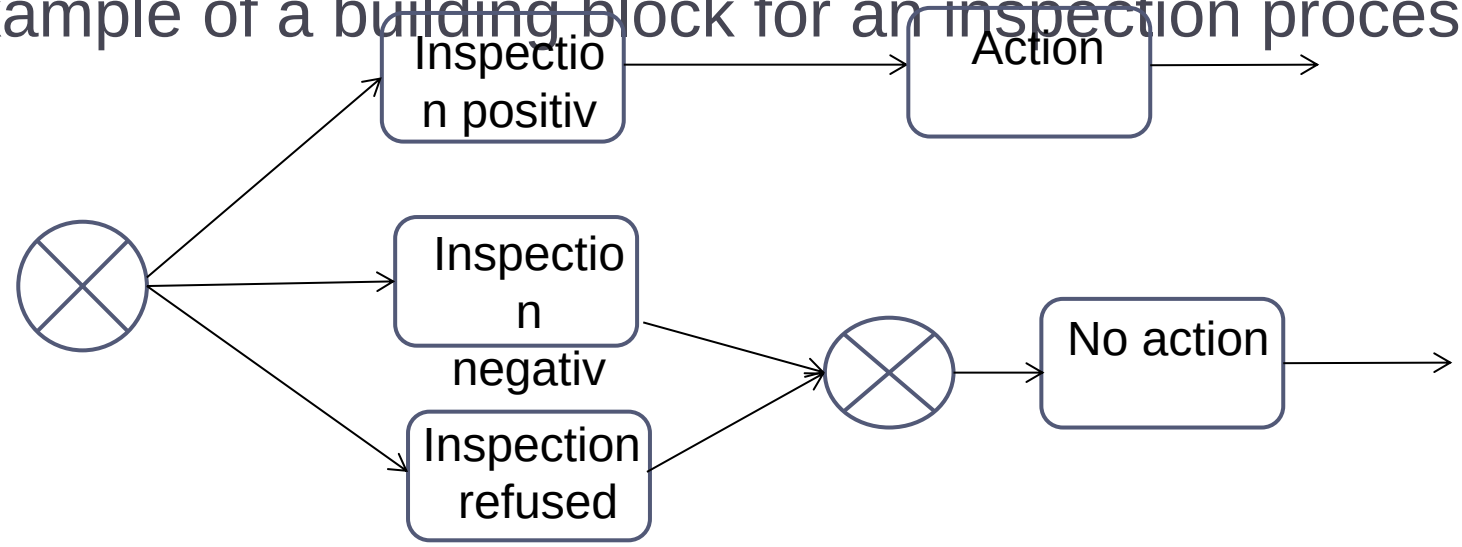
Framework for Unified Approach – Process Types

- Closed processes: mainly from the production perspective
  - Process has a definite start and end



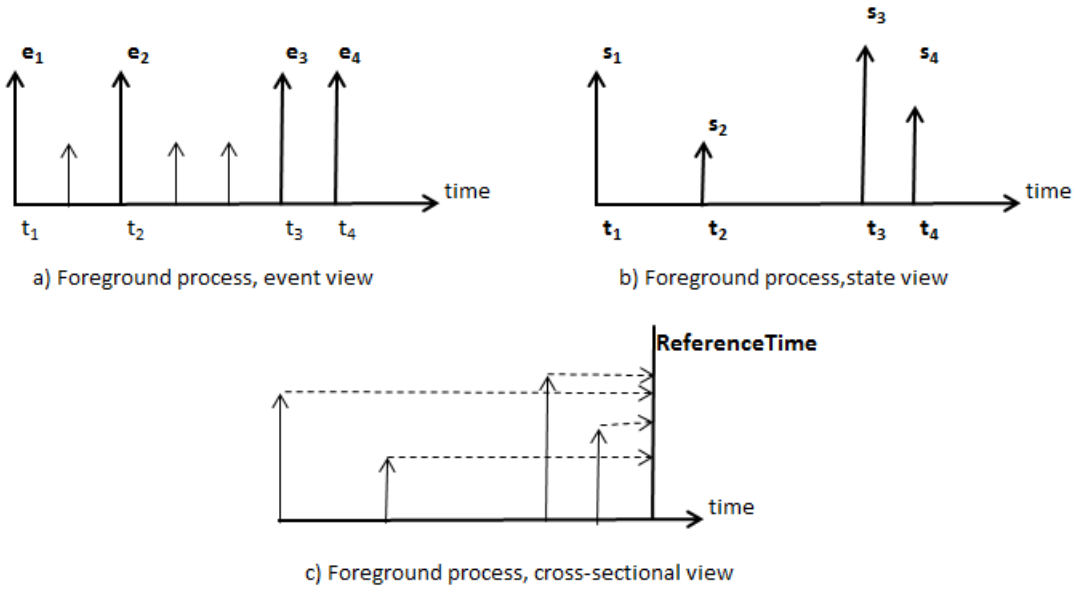
Framework for Unified Approach – Process Types

- Open processes: mainly from the customer perspective
  - Process develop more like a tree according to (customer) decisions, often censored e.g. have an open end
  - Example of a building block for an inspection process



Framework for Unified Approach – Process Views

- Note: In general we cannot observe all events in a process
- Observed events are called “foreground process”
- Schematic representation of views on foregr



Framework for Unified Approach – Goal formulation

- Usually one starts with a rather vague goal formulation
- For making the goal formulation more precise two different specifications are used
  - Key Performance Formulation (KPI)
  - Analytical Goal Formulation

Framework for Unified Approach – Goal formulation

- KPIs maybe of different type, e. g.
  - Quantitative Indicators describing some aspect of the process by numbers
  - Practical Indicators that interface with business processes
  - Directional Indicators specifying change in performance
  - Financial Indicators used for performance management
- Usually KPIs depend on a number of factors so called Influential Factors
- Precise formulation and evaluation of a KPI depends on knowledge how influential factors affect the KPI

Framework for Unified Approach – Goal formulation

- Analytical Goals are formulation of the goals in such way that the impact of influential factors on a KPI can be analysed in the framework of a model for the business process
- Examples of Analytical goals:
  - KPIs can be predicted by the influential factors
  - KPIs can be diversified according to profitability of customers
  - Identification of existing processes
  - Compliance of process instances with a predefined business process

Framework for Unified Approach – Goal formulation

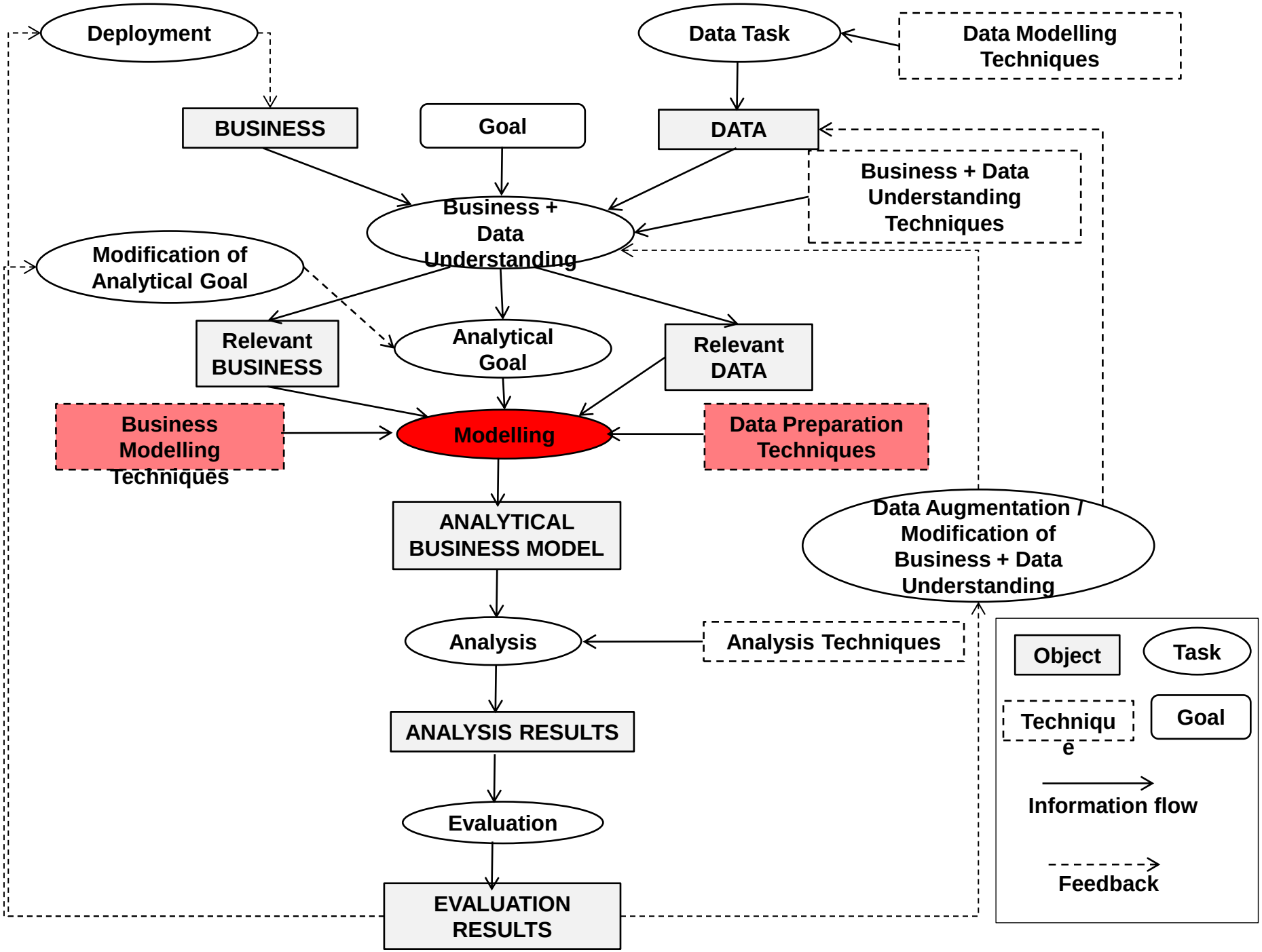
- Usually a goal has to be segmented in a number of sub-goals, oriented on different business perspectives
- Example:
  - The goal “increase sales by new product” has to be segmented into sub-goals like
    - Development of a process model
    - Identification of possible customers
    - Financial considerations

Framework for Unified Approach – Method Format

- Achieving an analytical goal needs a method which combines
  - Empirical knowledge about the business process with knowledge about the business
  - Using this knowledge for defining an appropriate model
  - Knowing how to apply analytical techniques for answering the analytical goal within the model
  - Evaluation of analysis results in context of the business
- The method can be oriented on existing method formats like CRISP or L* format

Framework for Unified Approach – Method Format

- There exist a number of formats which define such methods
  - CRISP from a more customer oriented perspective, oriented towards data in the cross-sectional view
  - L* from a more production oriented perspective oriented towards data in the event view
- The different sub-goals require a cyclic format
- A proposal for combining the essential part of the different formats is shown on the next slide

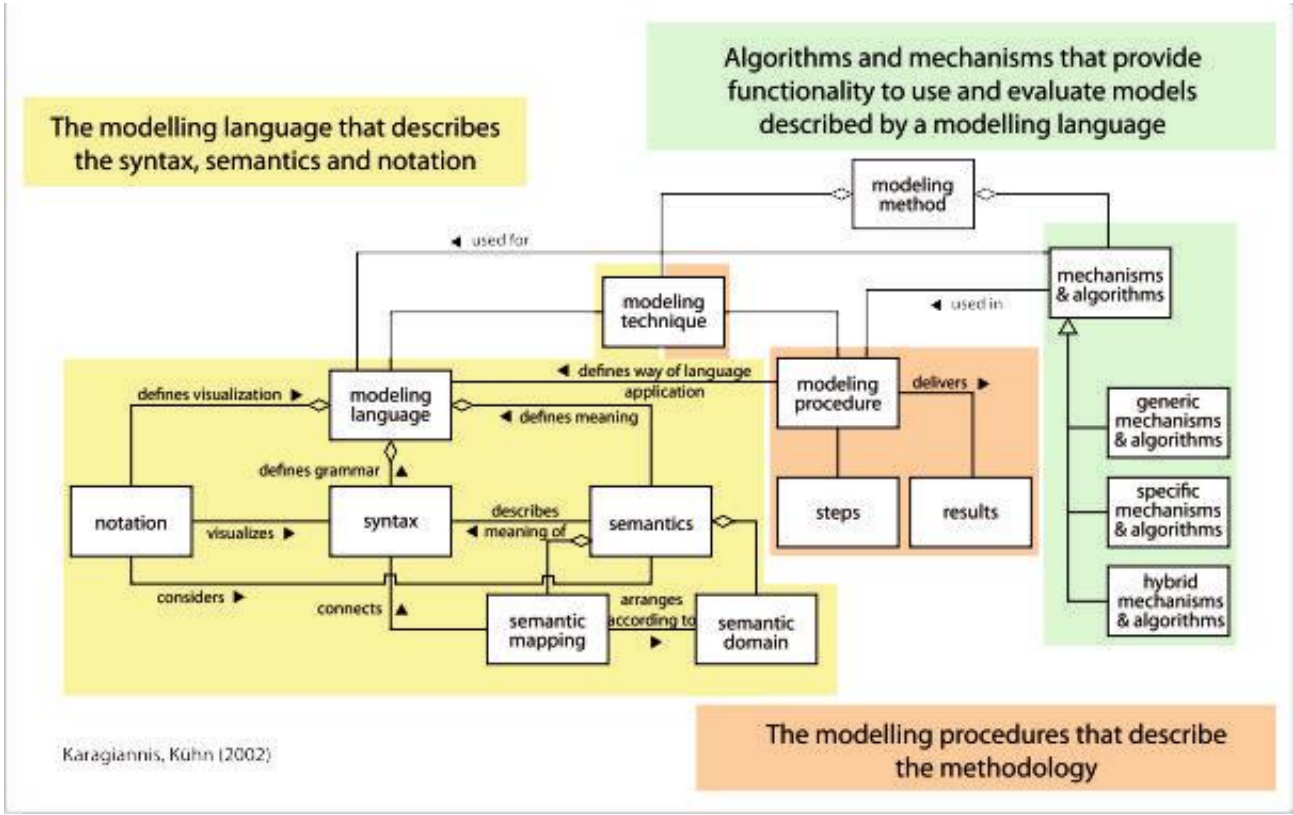

Deployment
Data Task
Data Modelling Techniques
BUSINESS
Goal
DATA
Business + Data Understanding Techniques
Business + Data Understanding
Modification of Analytical Goal
Relevant BUSINESS
Analytical Goal
Relevant DATA
Business Modelling Techniques
Modelling
Data Preparation Techniques
ANALYTICAL BUSINESS MODEL
Data Augmentation / Modification of Business + Data Understanding
Analysis
Analysis Techniques
ANALYSIS RESULTS
Evaluation
EVALUATION RESULTS
Object
Task
Technique
Goal
Information flow
Feedback

Framework for Unified Approach – Method Format

- We will focus in the following on the modelling task
- Main input from Business and Data Understanding Task:
  - An appropriate view on the business relevant for the analysis goal
  - An appropriate view on existing data about the enterprise and the business environment available in internal data sources (e.g. a warehouse) and external data sources
  - An analytical goal which can be answered hopefully using knowledge about the business and empirical information

Modelling Techniques

- Modelling methods should be understood as defined by Karagiannis and Kühn



Modelling Techniques

- Due to the fact that the modelling language should support models for different analytical goals we will specify some details of the language and the methods
- First of all note that for different analytical goals there exist “mechanics and algorithms” formulated in different languages, e.g.
  - Graph oriented languages for specification of business processes in the production perspective or organisational perspective
  - Statistical languages for answering questions in the customer perspective

Modelling Techniques

- Each language has different basic model elements which have a well defined meaning (semantic) within the language, rather independent of any application
  - In a graph oriented language a graph with labelled edges the labels may be interpreted rather generic as distance or similarity
  - In a statistical oriented language an equation may be interpreted as regression or as probability of a certain class

Modelling Techniques

- This semantic of the model language allows formulation of generic questions, which can be solved by specific methods, e.g.
  - Finding a shortest path in a graph
  - Finding the solution of parameters in a regression equation which fits best to the data
- Consequently we should think about the language in terms of model structures comprising the language, the basic model elements, and the generic questions in the language together with the analysis techniques for answering this questions

Modelling Techniques

- Taking the empirical character of the information into account a modelling procedures has to treat the following three topics:
  - Definition of a model configuration, i.e. am admissible expression in the model structure, which allows formulation and answering the analysis goal as question about properties of the model configuration
  - Connection of model configuration with data, i.e. the empirical information about instances of the business process defines input and output for the configuration
  - Definition of variability for handling blurred data

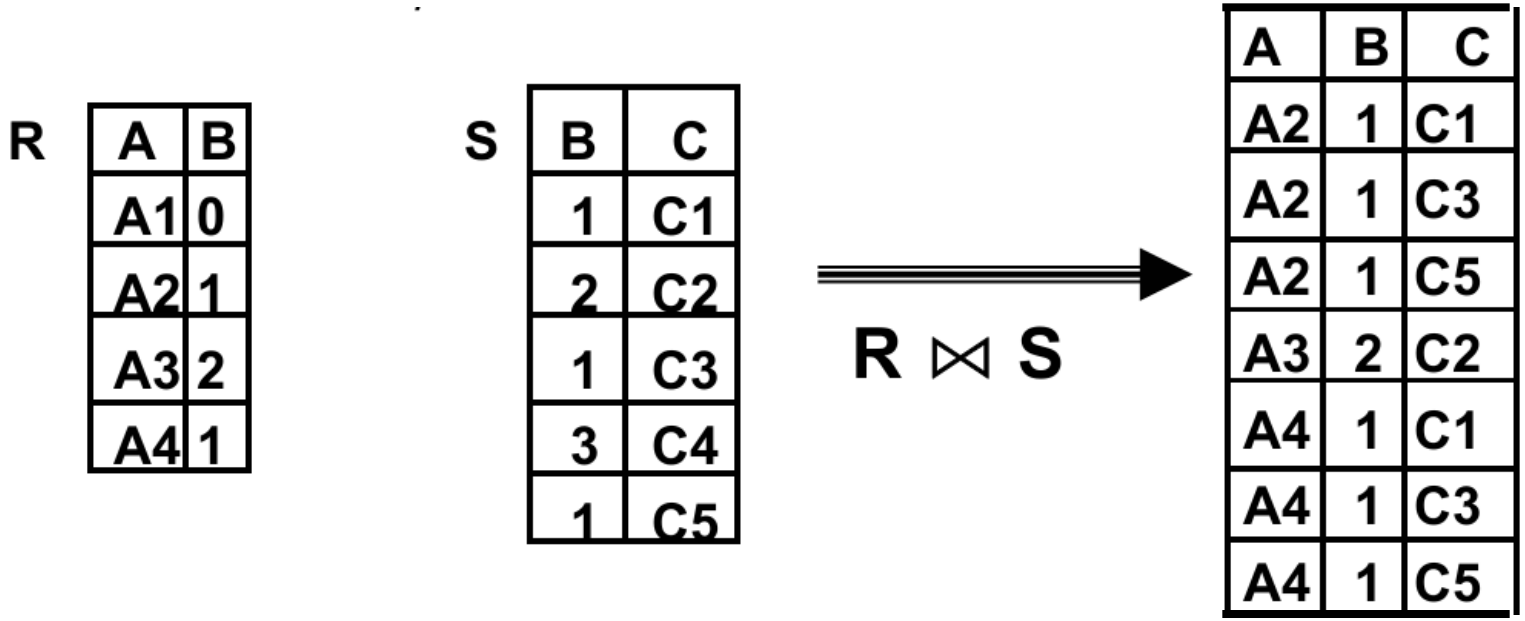
Data Preparation Techniques

- Data preparation techniques have to organize the data in such a way that data could be used as input and output for the model configuration
- Important activities are:
  - Data Transformations
  - Data Integration
  - Data Quality

Typical transformation task patterns

- Selection
- Projection
- Group by
- Reclassification
- Join
- Numeric transformation (Count, Sum, Avg, Variance, Median, Max , Min etc.)

Join


R
A B
A1 0
A2 1
A3 2
A4 1
S
B C
1 C1
2 C2
1 C3
3 C4
1 C5
R ⋈ S
A B C
A2 1 C1
A2 1 C3
A2 1 C5
A3 2 C2
A4 1 C1
A4 1 C3
A4 1 C5

Examples of join algorithms

- **Nested-loop join …** *The simplest algorithm. It may not be efficient.*
- **Blocked nested-loop join** *… Improved nested-loop join.*
- **Index join** *… Look up index to find matching tuples.*
- **Sort-merge join** … *Both relations are sorted on the join attributes first. Then, the relations are scanned in the order of the join attributes.*
- **Hash join …** *Depends on roughly equally sized buckets.*

Example - nested-loop join

```
FOR EACH r IN R DO
    FOR EACH s IN S DO
        IF ( r.B=s.B) THEN
OUTPUT (r ⋈ s)
```

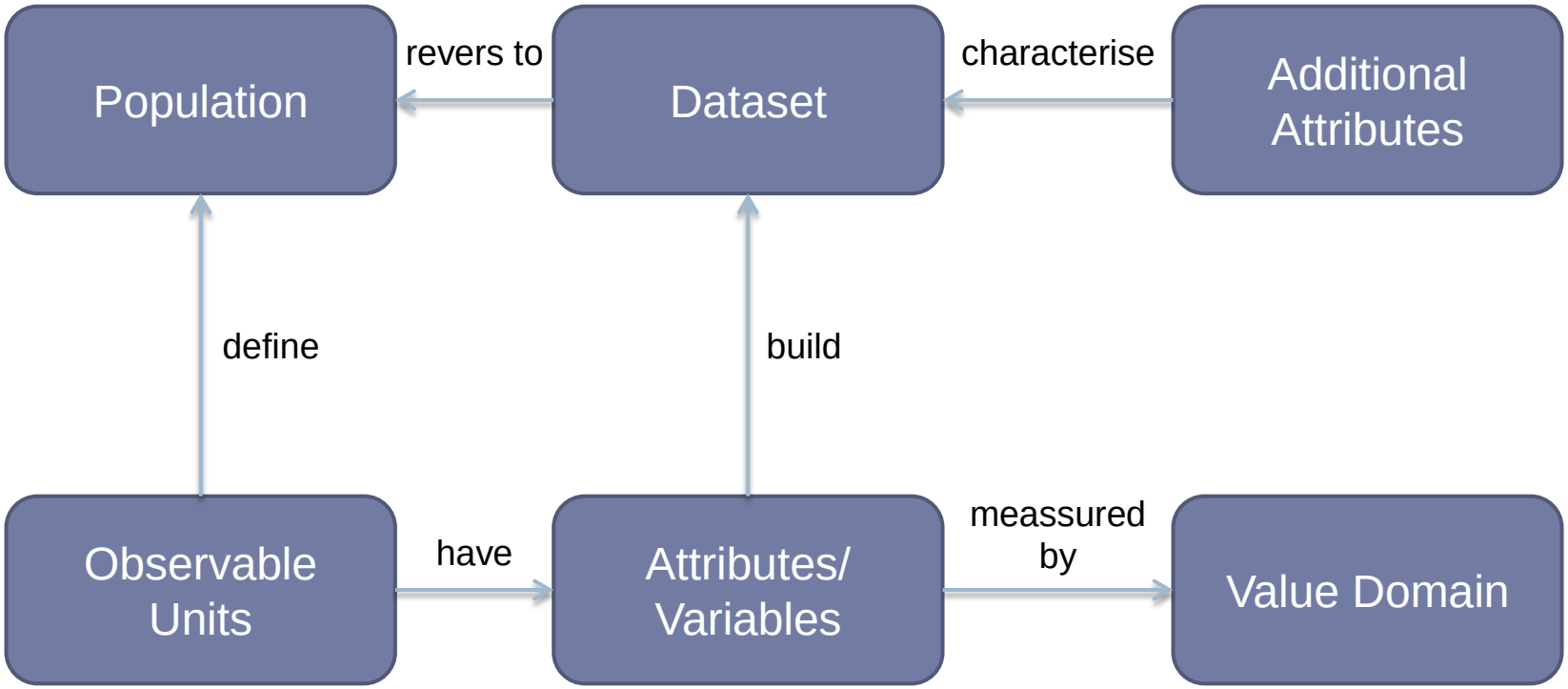
Source: Ulf Leser: Datenbanksysteme II, 2005

Example - sort-merge join

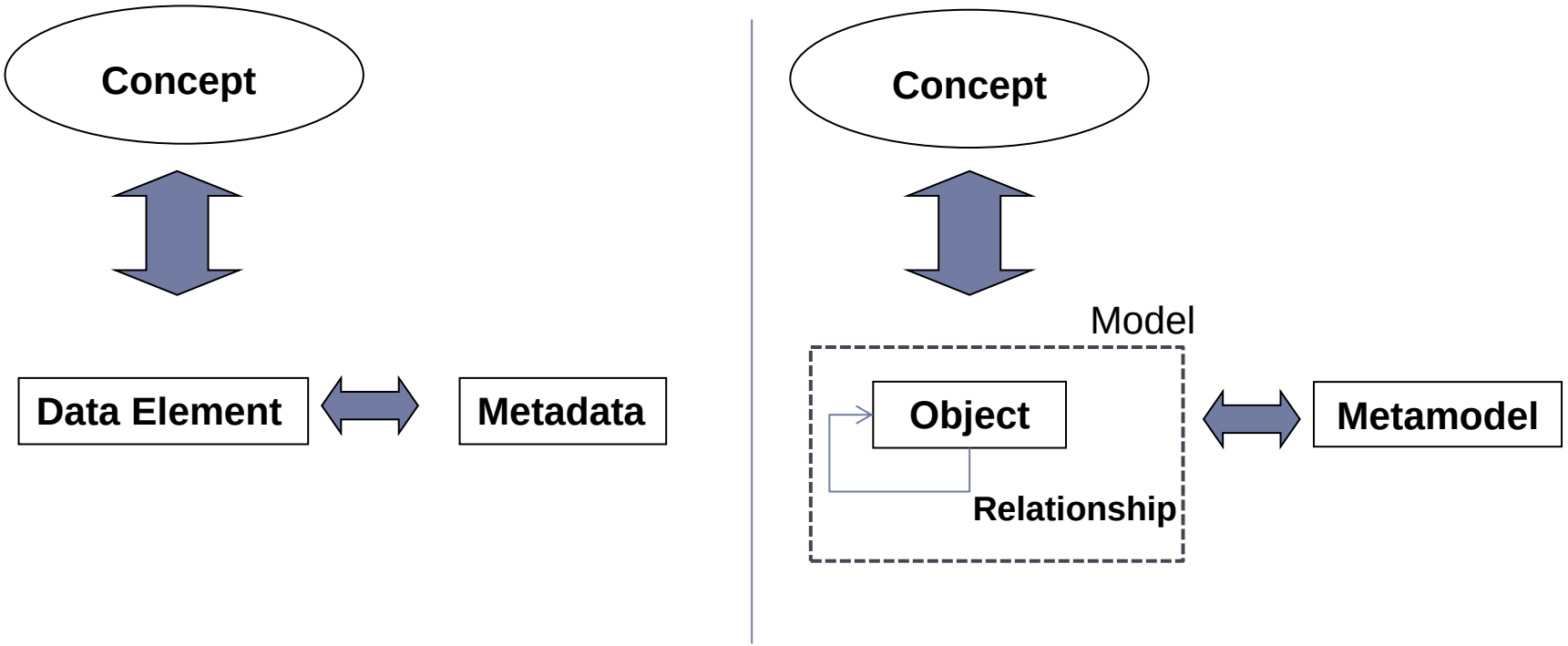
```
r := first (R); s := first (S);
WHILE NOT EOR (R) and NOT EOR (S) DO
      IF r[B] < s[B] THEN r := next (R)
      ELSEIF r[B] > s[B] THEN s := next (S)
      ELSE/* r[B] = s[B]*/
           b := r[B]; B := ∅;
           WHILE NOT EOR(S) and s[B] = b DO
                 B := B ∪ {s};
                 s = next (S);
           END DO;
           WHILE NOT EOR(R) and r[B] = b DO
                 FOR EACH e in B DO
                       OUTPUT (r,e);
                 r := next (R);
           END DO;
END DO;
```

What is missing in both algorithms?

Statistical metadata to ensure lineage



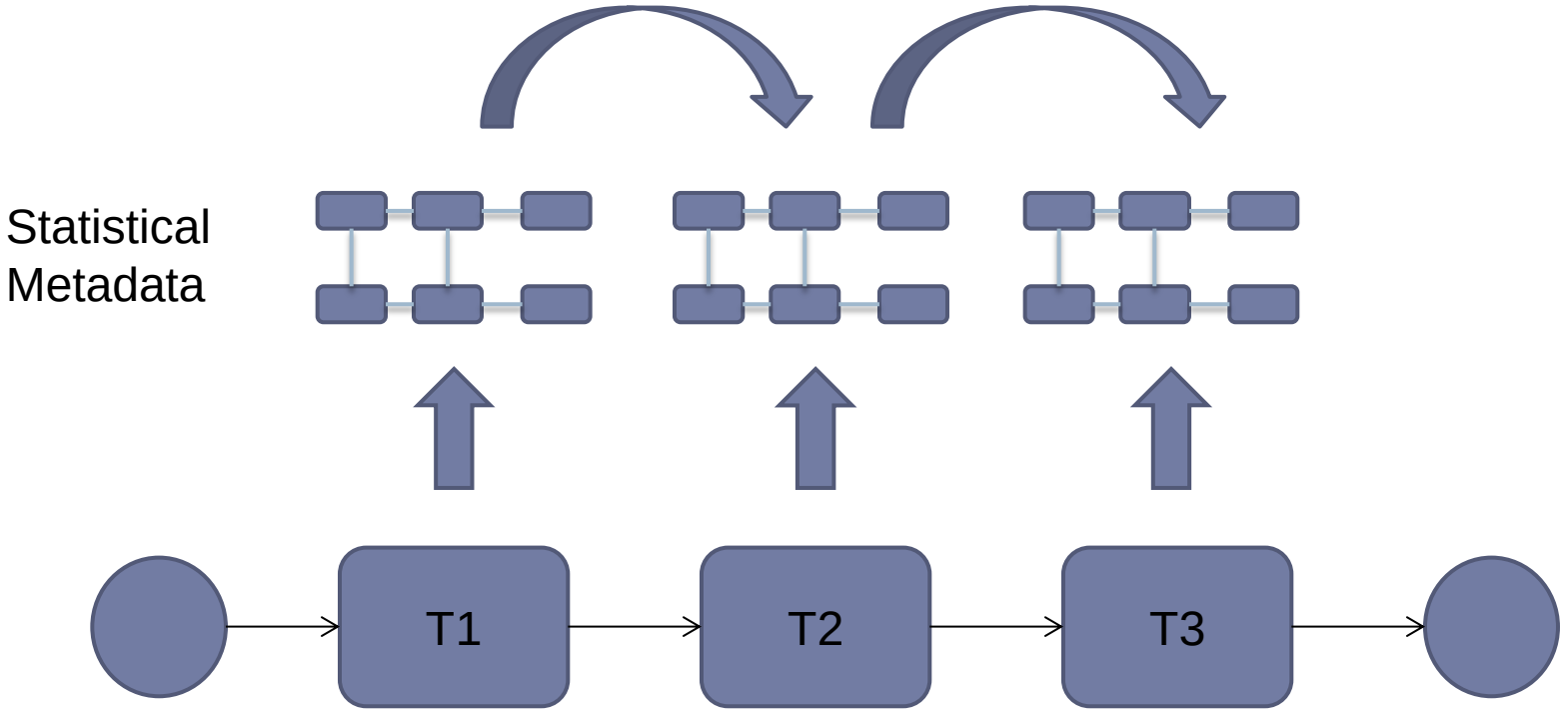
Difference between Metadata and Metamodel



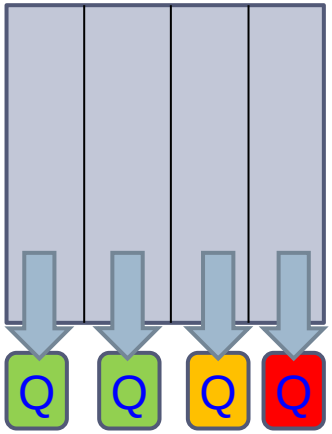
Metadata:
Data which describes other data

Metamodel:
Model which describes other model.

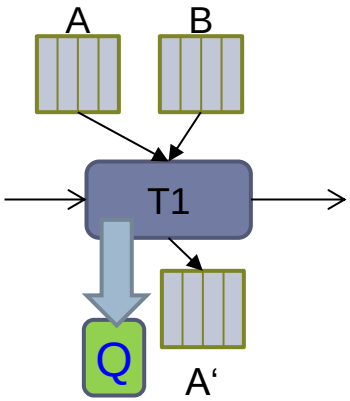
Ideally: Automatic update of statistical metadata



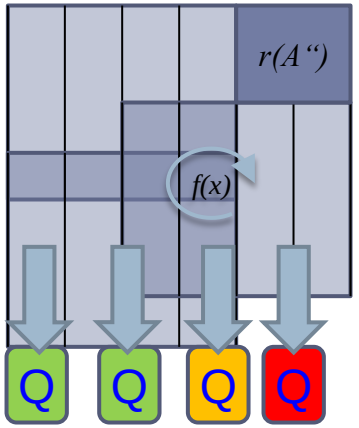
Quality Dimension (Additional Attributes)



Quality of input datasets



Feasibility check

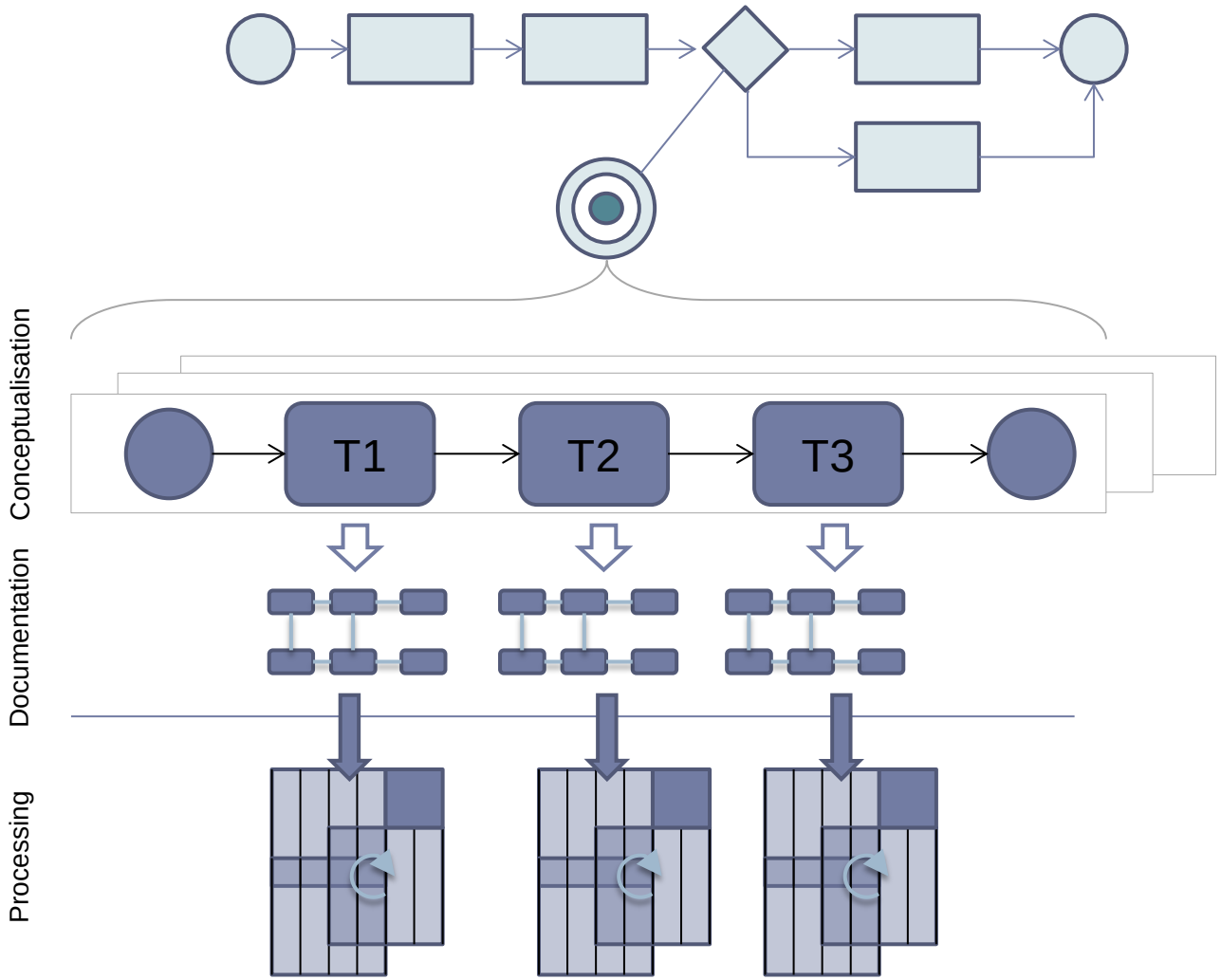


Resulting quality

Some typical quality criteria

- Freshness
  - Currency *(delivery time vs. extraction time)*
  - Timeliness *(delivery time vs. time of last update)*
- Accuracy
  - Semantic *(Closeness to Value)*
  - Syntactic *(Closeness to Structure)*
- Precision *(the exactness of measurement or description)*
- Completeness

Tying it all together



Business Process

BA KPI

DIBA Workflow

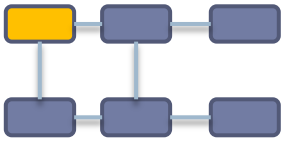
DIBA Metadata

Datasets

Showcase

Campaign Management at an insurance company

„Identify high potential customers with cross-selling opportunities for life insurances“



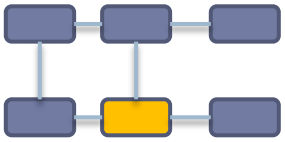
Population

Initial population: all customers
Target population: customers with potential for cross-selling…

Definitions

Campaign Management at an insurance company

„Identify **high potential** customers with cross-selling opportunities for life insurances“



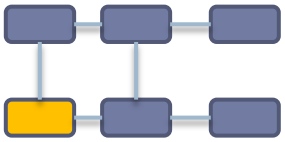
Attribute
(= KPI)

customers with high gross-margin
(premium vs. claims)

Definitions

Campaign Management at an insurance company

„Identify high potential **customers** with cross-selling opportunities for life insurances“



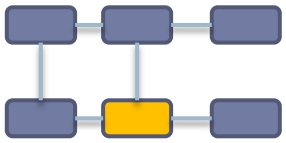
Observable Unit

owner of at least one insurance contract issued by our insurance company

Definitions

Campaign Management at an insurance company

„Identify high potential customers with **cross-selling opportunities** for life insurances“



Attribute

Customers who do not own a life insurance

Online DEMO


OMiLAB
DICE 1.0
Modelling Toolkit
powered by ADOxx
www.adoxx.org
© Copyright BOC Information Technologies Consulting AG, Vienna 2016.
ADOxx, the BOC Management Office, ADONIS:Community as well as
ADOSCORE, ADONIS, ADOLOG and ADOIT are registered trademarks of
the BOC Group.