

# Web Data Integration

# Introduction and

# Course Outline



- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web-based Systems
  - Large-Scale Data Integration
  - Data and Web Mining
- Room: B6, 26 - B1.15
- Consultation: Wednesday 13:30-14:30
- eMail: [chris@informatik.uni-mannheim.de](mailto:chris@informatik.uni-mannheim.de)



# Hallo

- M. Sc. Wi-Inf. Anna Primpeli
- Graduate Research Associate
- Research Interests:
  - Semantic Annotation of Web Pages
  - Product Data Integration
  - Identity Resolution
- Room: B6, 26, C 1.04
- eMail: [anna@informatik.uni-mannheim.de](mailto:anna@informatik.uni-mannheim.de)
- Will teach exercise group 2 and will supervise the student projects.



# Hallo

- M. Sc. Wi-Inf. Oliver Lehmberg
- Graduate Research Associate
- Research Interests:
  - Web Table Integration
  - Knowledge Base Extension
  - Network Analysis
- Room: B6, 26, C 1.04
- eMail: [oli@informatik.uni-mannheim.de](mailto:oli@informatik.uni-mannheim.de)
- Will teach exercise group 1 and will supervise the student projects.



# Introduction and Course Outline

1. Course Outline and Organization
2. What is Data Integration?
3. Application Areas
4. Types of Heterogeneity
5. The Data Integration Process
6. Data Integration Architectures
7. The Data Integration Software Market

# 1. Course Outline and Organization

# The Lecture

- introduces the principle methods of data integration
- discusses how to evaluate data integration results
- presents practical examples of how the methods are applied
- Topics
  1. Introduction to Data Integration
  2. Structured Data on the Web
  3. Data Exchange Formats
  4. Schema Mapping and Data Translation
  5. Identity Resolution
  6. Data Quality and Data Fusion
- no restriction on number of participants
- lecture is concluded with written exam
- 3 ECTS

# The Student Projects

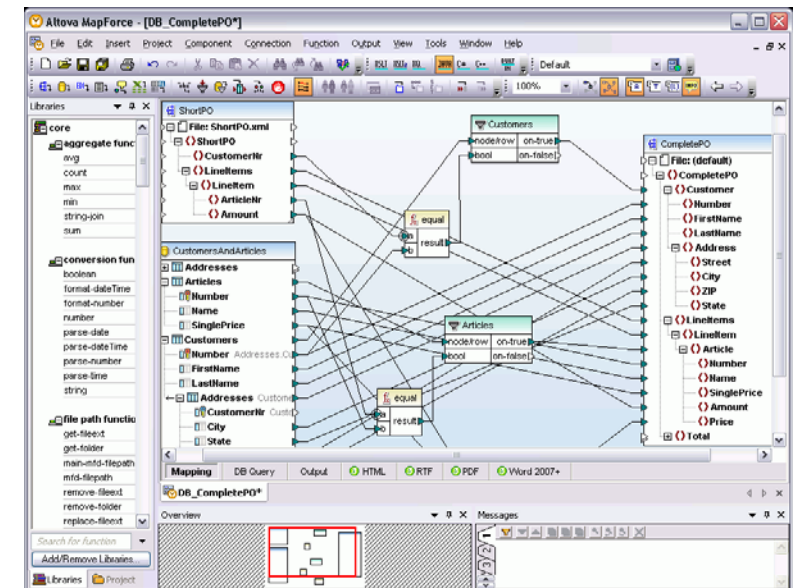
- Teams of **five students** realize a data integration project including
  1. data gathering
  2. schema mapping and data translation
  3. identity resolution
  4. data quality assessment and data fusion
- Teams will use data integration tools and will extend Java projects which implement basic integration methods
- Teams write 12 page report about their project, present project results
- You may choose their own application domain and data sets
  - minimum 4 data sets with a good degree of overlap in attributes and instances
- In addition, we will propose some suitable data sets from the domains of
  - films and actors, products and e-shops, restaurants, geographic information
- The number of participants in the projects is restricted to 60 (30 + 30)
- You need to register via Portal 2 until August 29th for the projects.
- 3 ECTS (70 % written project report, 30 % presentation of project results)

# Tools for Your Projects

Anna and Oliver give you an introduction to tools that you can use for your projects. They give you exercises to experiment with the tools along the use case of integrating data about films.

## 1. Data Translation

- Altova MapForce
- graphical data mapping and conversion tool

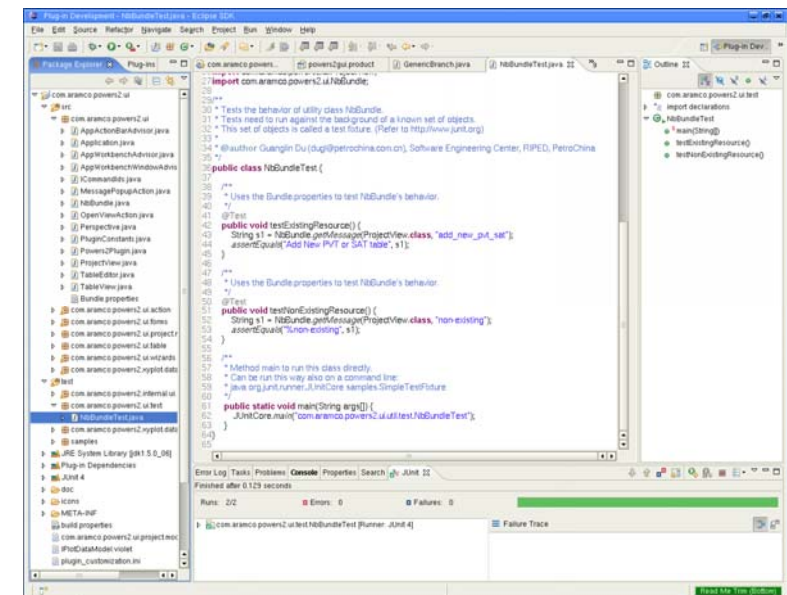


## 2. Identity Resolution

- Java framework Winte.r which implements the necessary methods

## 3. Data Fusion

- Java framework Winte.r which implements the necessary methods



# Schedule

Week	Wednesday	Thursday
5.9.2018	Lecture: Introduction to Web Data Integration	Lecture: Structured Data on the Web
12.9.2018	Lecture: Data Exchange Formats	Lecture: Data Exchange Formats
19.9.2018	Lecture: Schema Mapping	Lecture: Schema Mapping
26.9.2018	Project: Introduction to Student Projects	Tool Intro: MapForce
3.10.2018	- Holiday -	Project Work: Data Translation
10.10.2018	Project: Feedback about Project Outlines	Lecture: Identity Resolution
17.10.2018	Lecture: Identity Resolution	Tool Intro: Winte.r Identity Resolution
24.10.2018	Project Work: Identity Resolution	Project Work: Identity Resolution
31.10.2018	Project Work: Identity Resolution	- Holiday -
7.11.2018	Lecture: Data Fusion	Lecture: Data Fusion
14.11.2018	Tool Intro: Winte.r Data Fusion	Project Work: Data Fusion
21.11.2018	Project Work: Data Fusion	Project Work: Data Fusion
28.11.2018	Project Work: Data Fusion	Project Work: Data Fusion
5.12.2018	Presentation of project results	Presentation of project results
17.12.2018	Final Exam	

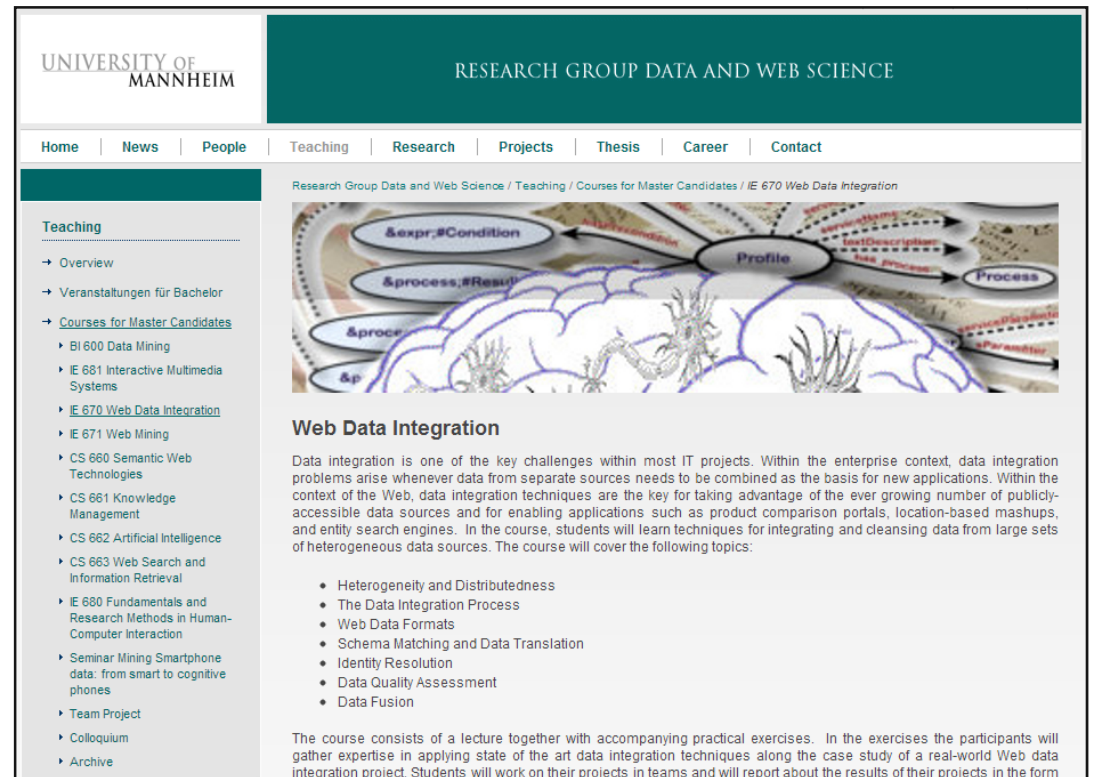
# Course Organization

## – Course Webpage

- <http://dws.informatik.uni-mannheim.de/en/teaching/courses-for-master-candidates/ie-670-web-data-integration/>
- The lecture slides will be published on this webpage.
- Project-related material will be provided in ILIAS.

## – Time and Location

- Wednesday, 15:30 to 17:00.  
Building: B6, Room: A 101
- Thursday, 10:15 to 11:45.  
Building: B6, Room: A 101 and A305
- Start: 5.9.2018



The screenshot shows the website of the University of Mannheim, Research Group Data and Web Science. The header includes the university logo and the group name. A navigation bar lists: Home, News, People, Teaching, Research, Projects, Thesis, Career, Contact. The 'Teaching' section is active, displaying a list of courses for master candidates, including 'IE 670 Web Data Integration'. To the right, a diagram illustrates the 'Web Data Integration' process, showing a central 'Profile' node connected to various data sources and processes. Below the diagram, the text explains that data integration is a key challenge in IT projects, particularly in the context of the Web, and lists topics such as Heterogeneity and Distributedness, The Data Integration Process, Web Data Formats, Schema Matching and Data Translation, Identity Resolution, Data Quality Assessment, and Data Fusion. The course description states that it consists of a lecture with practical exercises and a case study of a real-world Web data integration project.

UNIVERSITY OF MANNHEIM


RESEARCH GROUP DATA AND WEB SCIENCE

Home | News | People | Teaching | Research | Projects | Thesis | Career | Contact

Teaching

- Overview
- Veranstaltungen für Bachelor
- Courses for Master Candidates
  - ▶ BI 600 Data Mining
  - ▶ IE 681 Interactive Multimedia Systems
  - ▶ IE 670 Web Data Integration
  - ▶ IE 671 Web Mining
  - ▶ CS 660 Semantic Web Technologies
  - ▶ CS 661 Knowledge Management
  - ▶ CS 662 Artificial Intelligence
  - ▶ CS 663 Web Search and Information Retrieval
  - ▶ IE 680 Fundamentals and Research Methods in Human-Computer Interaction
  - ▶ Seminar Mining Smartphone data: from smart to cognitive phones
  - ▶ Team Project
  - ▶ Colloquium
  - ▶ Archive

Research Group Data and Web Science / Teaching / Courses for Master Candidates / IE 670 Web Data Integration



**Web Data Integration**

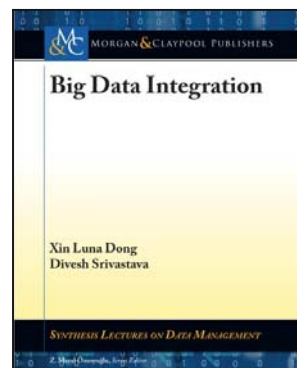
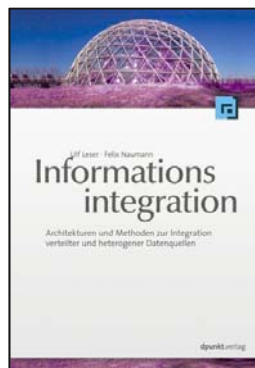
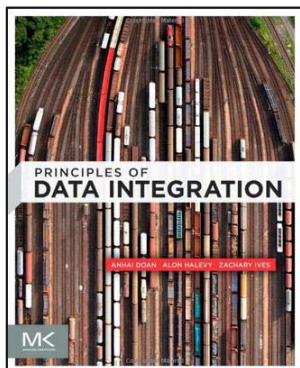
Data integration is one of the key challenges within most IT projects. Within the enterprise context, data integration problems arise whenever data from separate sources needs to be combined as the basis for new applications. Within the context of the Web, data integration techniques are the key for taking advantage of the ever growing number of publicly-accessible data sources and for enabling applications such as product comparison portals, location-based mashups, and entity search engines. In the course, students will learn techniques for integrating and cleansing data from large sets of heterogeneous data sources. The course will cover the following topics:

- Heterogeneity and Distributedness
- The Data Integration Process
- Web Data Formats
- Schema Matching and Data Translation
- Identity Resolution
- Data Quality Assessment
- Data Fusion

The course consists of a lecture together with accompanying practical exercises. In the exercises the participants will gather expertise in applying state of the art data integration techniques along the case study of a real-world Web data integration project. Students will work on their projects in teams and will report about the results of their projects in the form

# Literature and Credits

1. AnHai Doan, Alon Halevy, Zachary Ives: **Principles of Data Integration**. Morgan Kaufmann, 2012. (Online access via the library)
2. Xin Luna Dong, Divesh Srivastava: **Big Data Integration**, Morgan & Claypool, 2015 (Online access via the library)
3. Ulf Leser, Felix Naumann: **Informationsintegration**. DBunkt Verlag, 2007. (Several copies in the library, PDF version at <https://www.dpunkt.de/openbooks/informationsintegration.pdf>, Video lecture at <http://www.tele-task.de/archive/series/overview/892/>)
4. Jérôme Euzenat, Pavel Shvaiko: **Ontology Matching**. Springer, 2014.
5. Felix Naumann: **An Introduction to Duplicate Detection**. Morgan & Claypool, 2012. (Online access via the library)
6. **Lecture videos** from HWS2015 on DWS page.



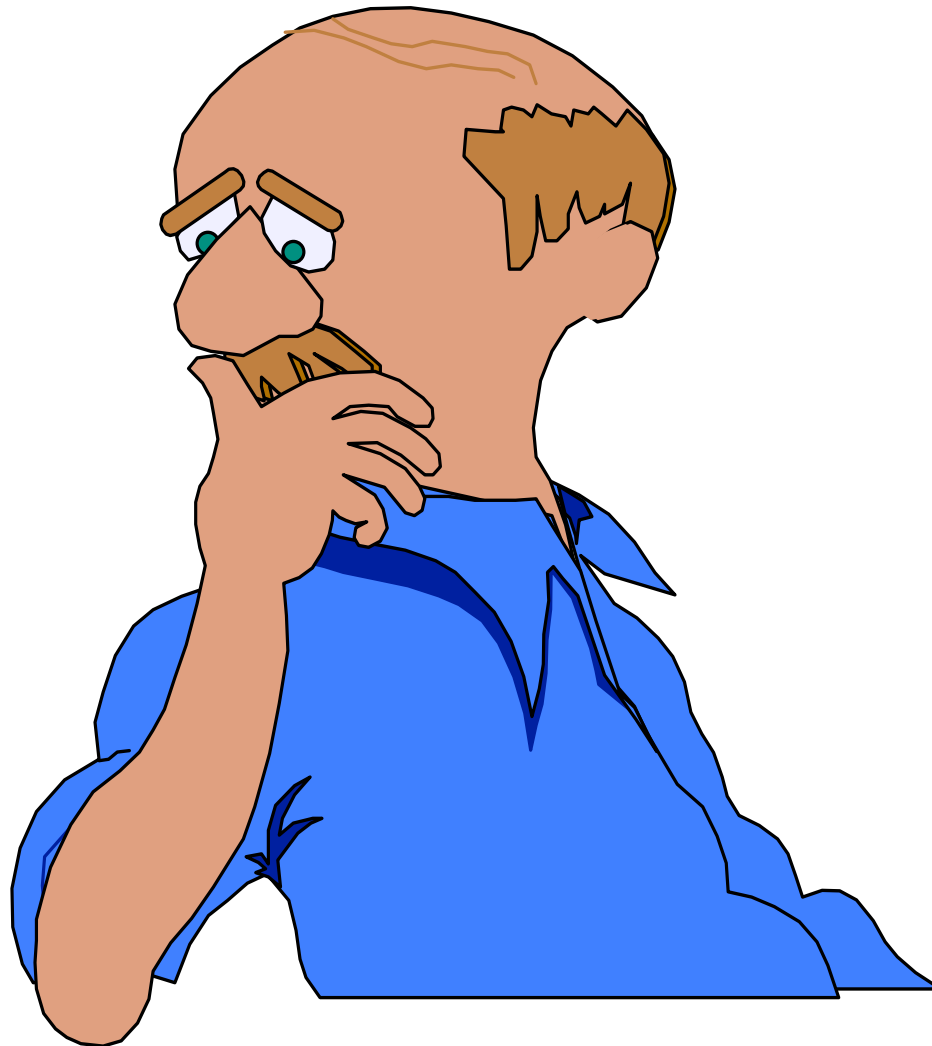
## Credits

The slide set of this lecture builds on slides from:

- Ulf Leser, Felix Naumann
- AnHai Doan, Alon Halevy, Zachary Ives

Lots of thanks to all of you!

# Questions about the Course Organization?



# Introduction to Data Integration

1. Course Outline and Organization
2. What is Data Integration?
3. Application Areas
4. Types of Heterogeneity
5. The Data Integration Process
6. Data Integration Architectures
7. The Data Integration Software Market

## 2. What is Data Integration?

- Databases and data mining tools are great: They let us manage and analyze huge amounts of data.
  1. **Assuming** you've put it all into a single schema.
  2. **Assuming** the database doesn't contain duplicate records.
  3. **Assuming** that data is current and contains no data conflicts.
- In reality, applications often need to work with data from multiple independently created data sources.
  1. Different sources use different data models.
  2. Different sources use different schemata.
  3. Different sources describe the same real-world entity.
  4. Different sources provide conflicting data about a single entity.
  5. Different sources provide different limited query interfaces to their data.

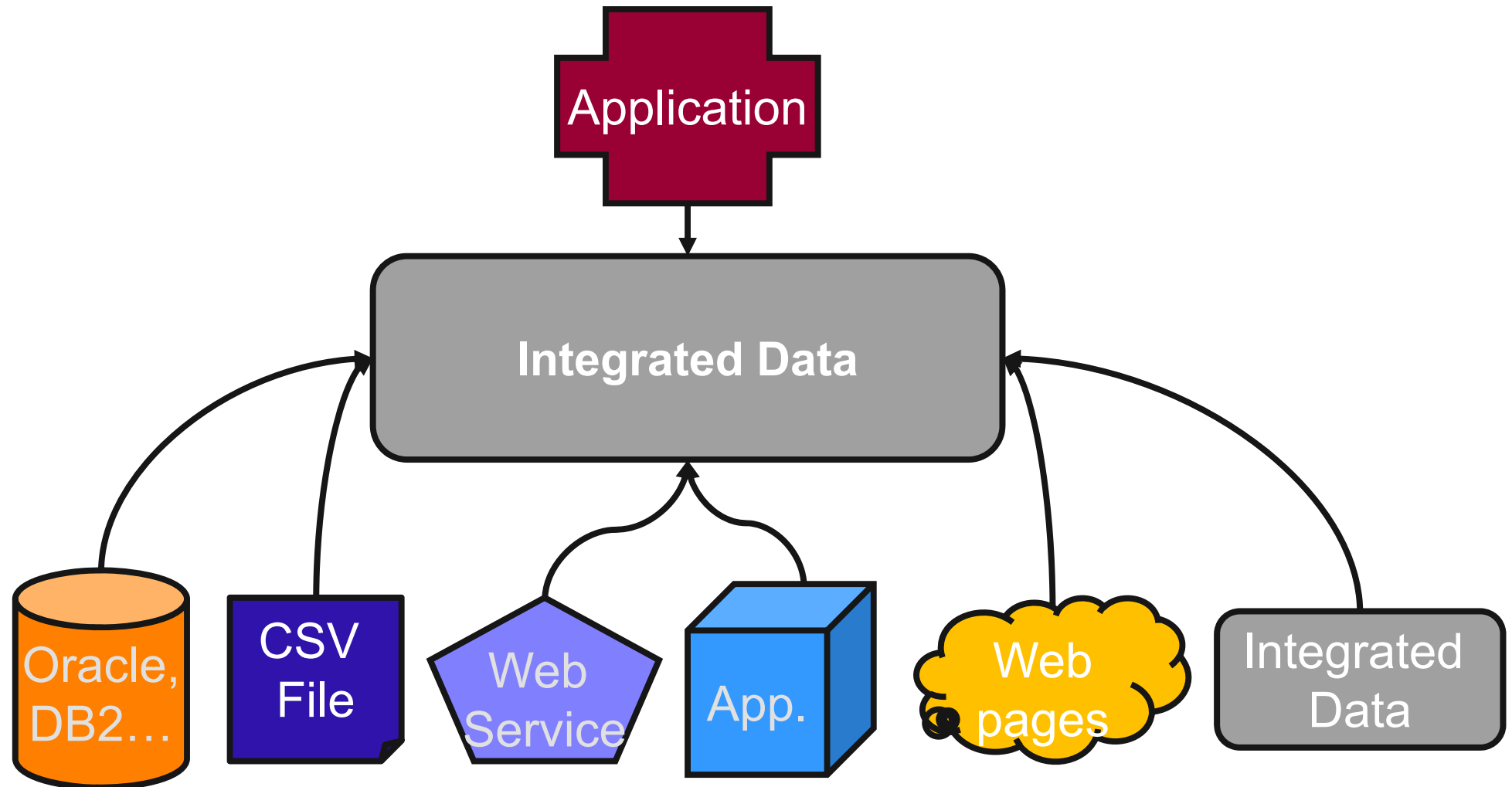


# What is Data Integration?

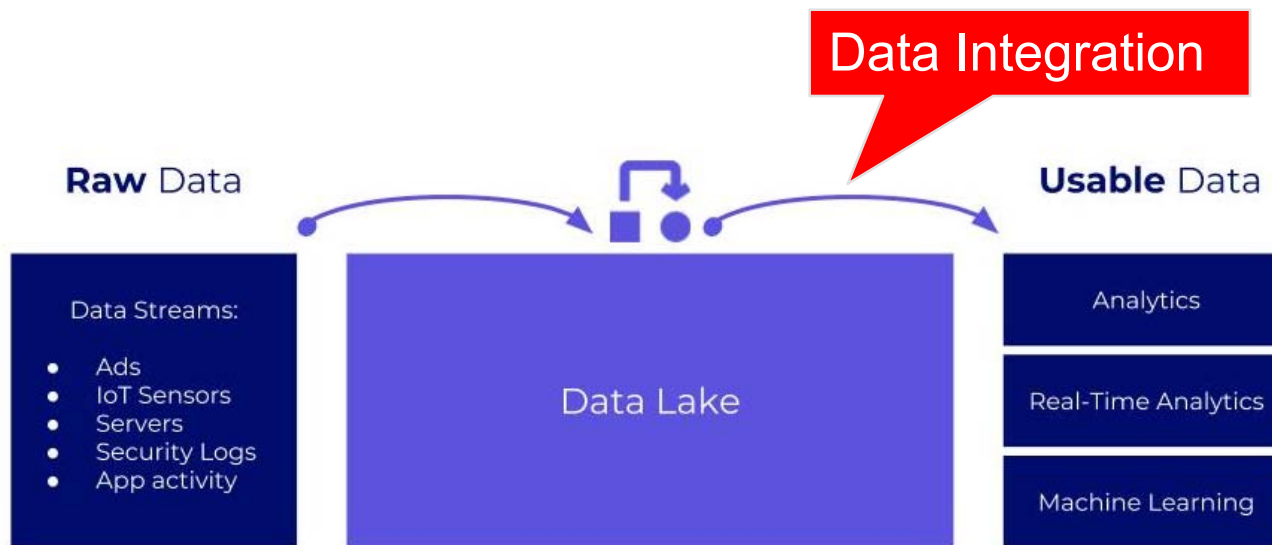
**Data integration is the process of consolidating data from a set of heterogeneous data sources into a single uniform data set or view on the data.**

- The integrated data set should:
  1. Correctly and completely represent the content of all data sources.
  2. Use a single data model and a single schema.
  3. Only contain a single representation of every real-world entity.
  4. Not contain any conflicting data about single entities.
- To achieve this, data integration needs to resolve various types of **heterogeneity** that exist between data sources.

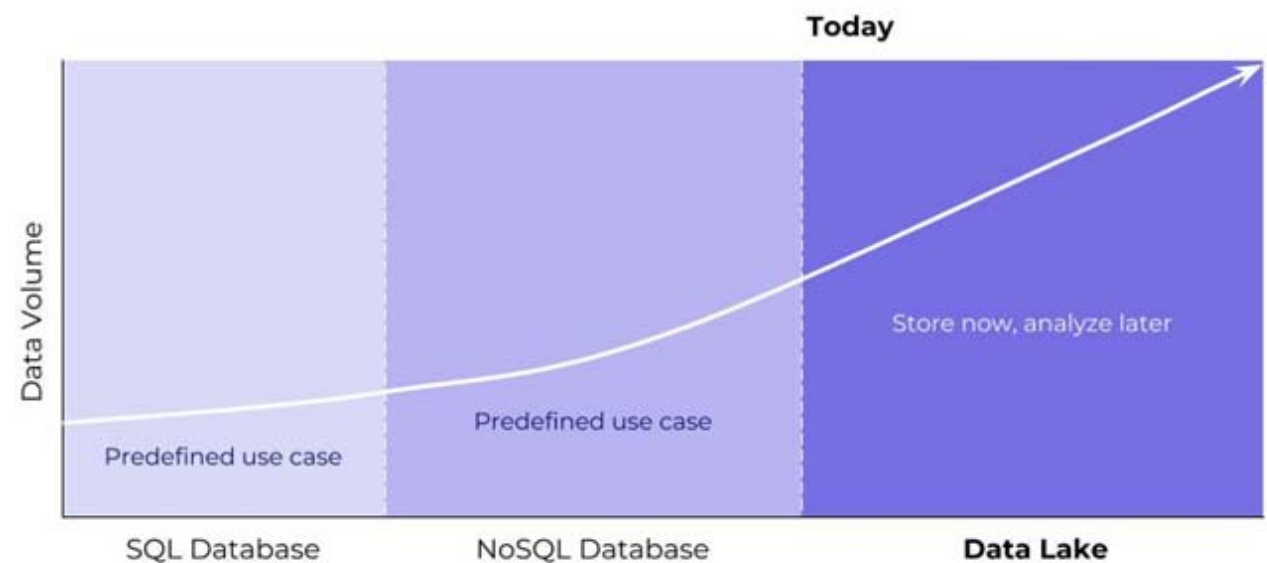
# Overview: Data Integration



# Big Data Integration: Draining the Data Lake



**Data Lake:** Unintegrated pool of potentially relevant raw data.

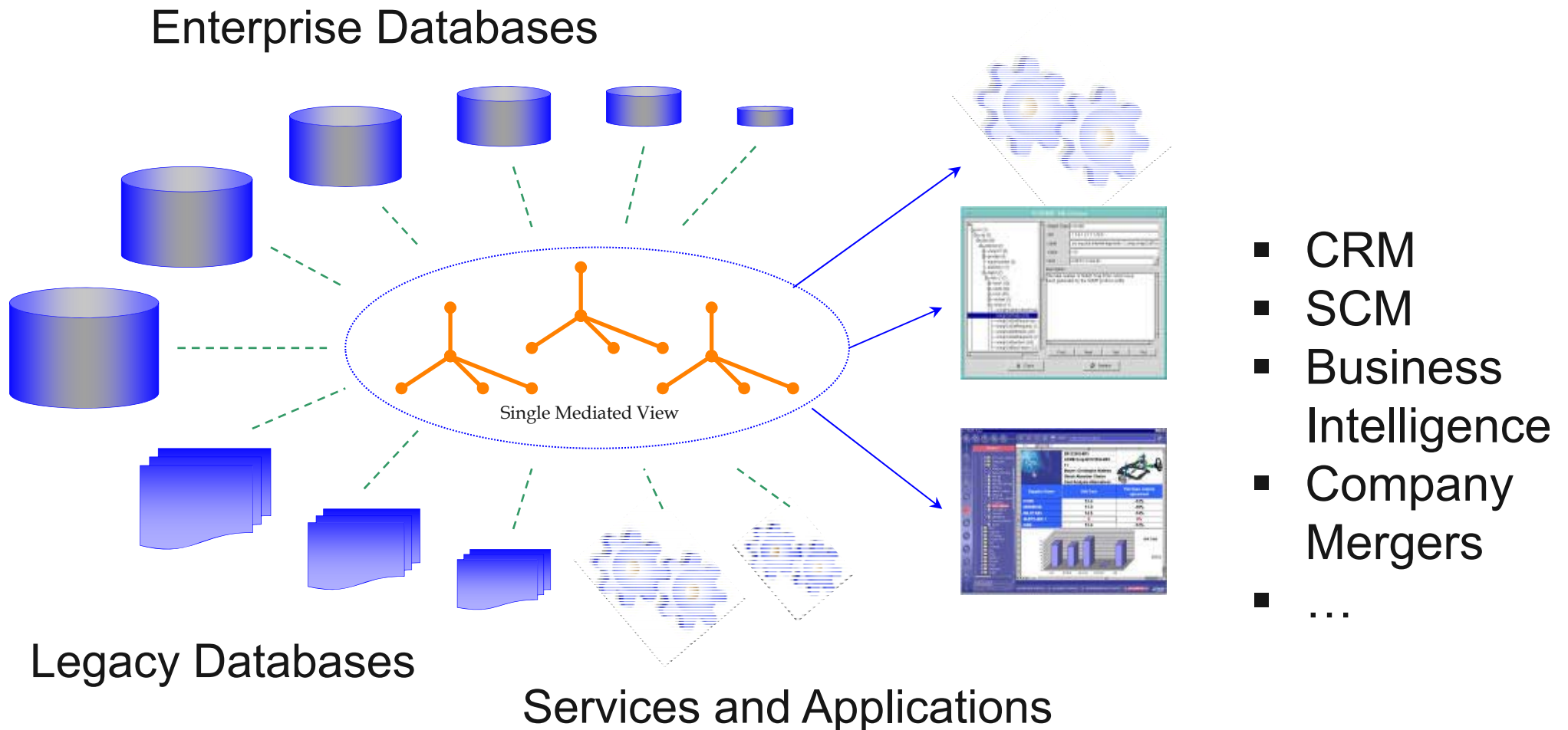


Source: <https://www.kdnuggets.com/2018/06/why-data-lake-matters.html>

# 3. Application Areas

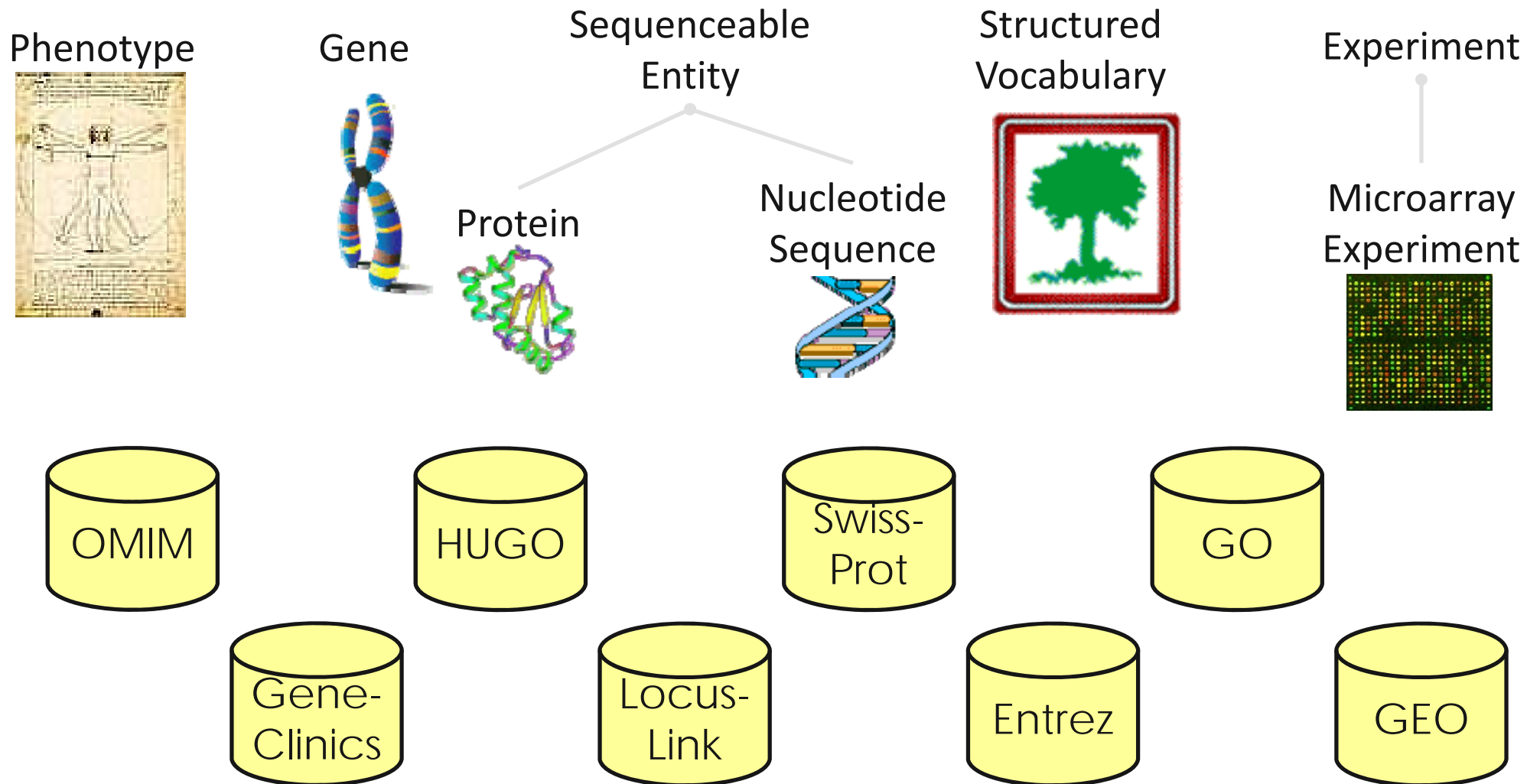
1. Business
2. Science
3. Government
4. The Web
5. .... pretty much every application area

# Application Area: Business



Oracle estimate: 50% of all IT \$\$\$ are spent here!

# Application Area: Science



Hundreds of biomedical data sources available; growing rapidly!

# Application Area: Government

Law enforcement agencies integrate data from various sources in order to identify suspects.

- Cell phone calls
- Location data
- Online profiles (Facebook)
- Web browsing behavior
- Credit card transactions
- Intelligence from other agencies
- ...



# Application Area: Data Journalism

- Government data is increasingly published under open licenses on the Web.
- Journalists discover stories by combining data from different sources.

## EU subsidies

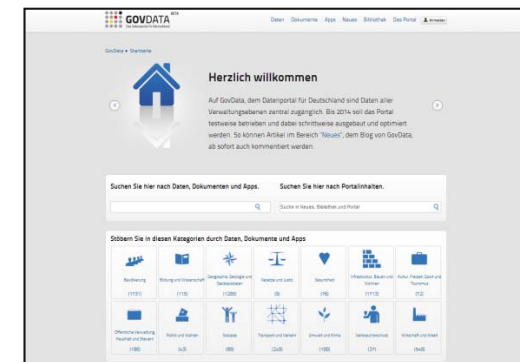
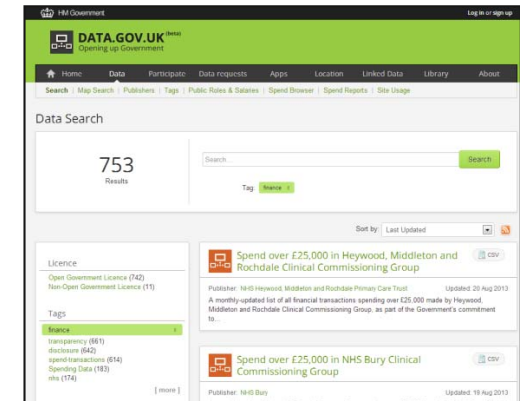
- received for renovating a ship
- received for scraping the same ship

## Members of parliament

- donations/membership in supervisory boards
- voting behavior

## Panama Papers

- ownership information about company networks
- discussable financial transactions



# Application Area: Online Shopping



Quilt-Books.Com



COMPU-BOOKS

amazon

KATSUKI  
BOOKS 日本書店



Providence-Books  
www.providence-books.com



NGHOSI  
books

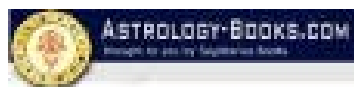
КНИГАРНЯ  
bookshop online



CACTUS-BOOKS.COM  
THE BEST COLOR FOR PAPER

metro-Books

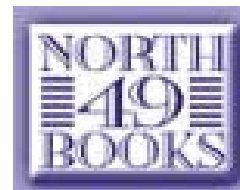
BARNES  
& NOBLE  
www.bn.com



SeeMe4Books

amazon.com

HALF PRICE COMPUTER BOOKS  
DANCE BOOKS



Just Great Books

# Comparison Shopping

[SIGN IN](#)

## The Unofficial Harry Potter Cookbook: From Cauldron Cakes to Knickerbocker Glory--More Than 150 Magical Recipes for Muggles and Wizards [Book]

**\$3** [online](#)

[Write a review](#)[Add to Shortlist](#)

By Dinah Bucholz - Adams Media - 2010 - Hardback - 256 pages - ISBN 1440503257

Bangers and mash with Harry, Ron, and Hermione in the Hogwarts dining hall. A proper cuppa tea and rock cakes in Hagrid's hut. Cauldron cakes and pumpkin juice on the Hogwarts Express. With this cookbook, dining a la Hogwarts is as easy as Banoffi Pie! With more than 150 easy-to-make ... [more »](#)

[Online stores](#)[Reviews](#)[Details](#)

**Online stores** [set your location](#)

☐ Free shipping ☐ Refurbished / used

Sponsored ⓘ

Sellers ▾	Seller Rating	Details	Base Price	Total Price	
<a href="#">MovieMars.com</a>	★★★★★ (42)	Free shipping	\$20.92		<a href="#">Shop »</a>
<a href="#">ValoreBooks.com</a>	No rating	No tax	\$3.24 \$3.95 shipping	\$7.19	<a href="#">Shop »</a>
<a href="#">Overstock.com</a>	★★★★★ (5,896)		\$12.92		<a href="#">Shop »</a>

# The Deep Web is accessible via HTML Forms

YAHOO! hotjobs®

Home Job Search My Searches My Jobs My Resumes Career T

## Search for Jobs Across the Web

Keyword(s)   
(e.g. Job title, company, occupation)

City & State or Zip

☒ Include surrounding cities

Job Category

Job Search Saved Searches

## Find a job

Advanced search | Search help

Enter keywords

Select an employer

All

Select a job type

All

SEARCH

USAJOBS® USAJOBS is the official job site of the United States Federal Government. It's your one-stop source for Federal jobs and employment opportunities.

"WORKING FOR AMERICA"

Search Jobs

My USAJOBS

Info Center

Veterans

Forms

Basic Search | Agency Search | Series Search | Advanced Search

## Keyword Search ?

(e.g.: Job Title, Agency Name, Vacancy Announcement #, Control #) [More Tips](#)

## Location Search ?

For multiple selections, hold down **Ctrl** (**Command** for Macs) while clicking selections.

----- Select all -----

US

AK

AK-Aleutian Islands

AK-Anchorage

## Job Category Search ?

For multiple selections, hold down **Ctrl** (**Command** for Macs) while clicking selections.

----- SELECT ALL -----

Accounting, Budget and Finance

Biological Sciences

Business, Industry, and Procurement

Copyright, Patent, and Trademark

## Salary Range ?

from  to

OR

## Pay Grade (GS) ?

from  to

# Structured Data on the Web

## More and more Websites

- semantically markup the content of their HTML pages
- publish structured data in addition to HTML pages

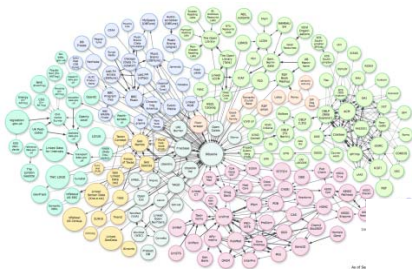
**Microformats**



**RDFa**



**Linked Data**



programmableweb

**Web APIs**



**Microdata**



# 4. Types of Heterogeneity

**We distinguish five types of heterogeneity:**

1. Technical Heterogeneity
2. Syntactical Heterogeneity
3. Data Model Heterogeneity
4. Structural Heterogeneity
5. Semantic Heterogeneity

**The goal of data integration is to bridge all these types of heterogeneity.**

Data source autonomy is the main reason for heterogeneity:

- Data sources independently decide how to store things and how to provide access
- Agreeing on standards partly reduces heterogeneity

# Technical Heterogeneity

**Technical heterogeneity comprises all differences in the means to access data, not the data itself.**

Level	Possibilities
Communication Protocol	HTTP, ODBC/JDBC, SOAP
Data Exchange Format	XML, JSON, CSV, RDF, HTML, binary data
Query Language	Full query language: SQL, SPARQL Canned queries: Web APIs, Web Forms Download of complete data set dumps
Additional Restrictions	Number of queries Cost per query / data set Access rights

# Syntactical Heterogeneity

**Syntactical heterogeneity comprises all differences in the encoding of values.**

Level	Possibilities
Character format	ASCII versus Unicode
Number format	Little endian versus big endian
Delimiter format	Tab-delimited versus Comma-separated values

Syntactical heterogeneity does not comprise

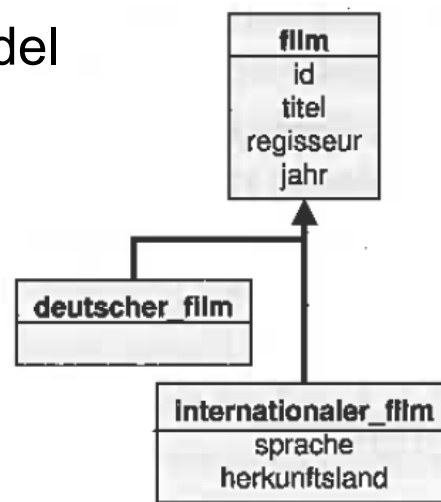
- Synonymous values
  - 1GB versus 1000MB → Semantic heterogeneity
- Structural differences
  - First name: Chris, last name: Bizer versus name: Chris Bizer  
→ Structural heterogeneity

# Data Model Heterogeneity

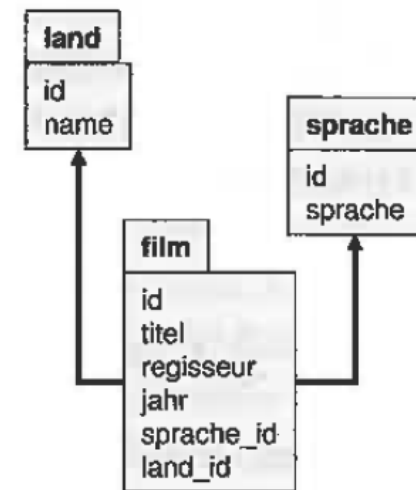
Data model heterogeneity comprises differences in the **data model** that is used to represent data.

Data Models:

1. Relational data model
2. XML data model
3. Object-oriented data model
4. RDF graph data model



Object-oriented



relational

```
<land>
  <sprache/>
  <herkunft/>
  <film>
    <id/>
    <titel/>
    <regisseur/>
    <jahr/>
```

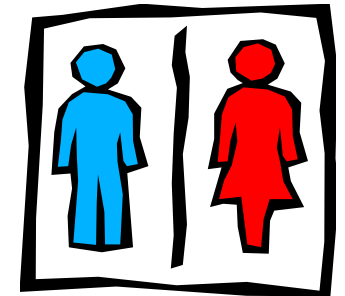
XML

# Structural Heterogeneity

**Structural heterogeneity comprises differences in the way different **schemata** represent the same part of reality.**

1. Alternative Modeling
  - Relation vs. Attribute
  - Attribut vs. Value
  - Relation vs. Value
2. Normalized vs. Denormalized
3. Nested vs. Foreign Key Relationship

# Example: Alternative Modelling



```
Man( Id, Firstname, Surname)  
Woman( Id, Firstname, Surname)
```

Relation vs. Attribute

```
Person( Id, Firstname,  
        Surname, Male,  
        Female)
```

Relation vs. Value

```
Person( Id, Firstname,  
        Surname, Sex)
```

Attribute vs. Value

# Semantic Heterogeneity

**Semantic heterogeneity comprises differences concerning the **meaning** of data and schema elements.**

## 1. Naming Conflicts

- Synonyms, homonyms, slightly deviating concepts

## 2. Object Identity / Duplicates

- Multiple data sources as well as multiple records within one data source may describe the same real-world entity.
- Which “Franz Müller” does a record describe?

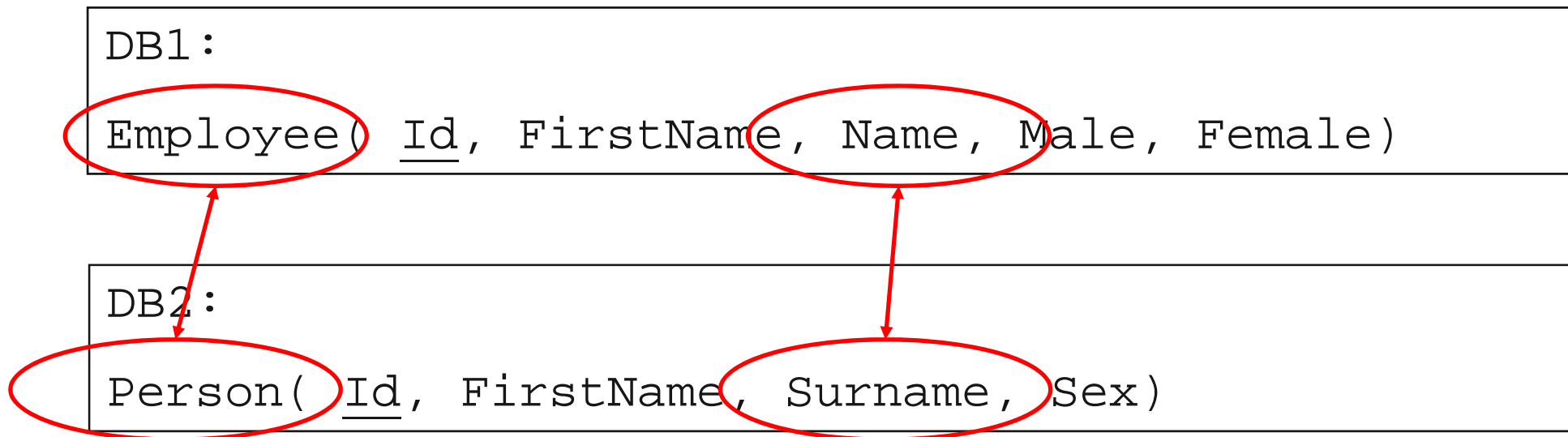
## 3. Data Conflicts

- Conflicting data about the same real-world entity in different data sources as well as within different records in the same data source.

# Semantic Heterogeneity: Synonyms

**Different words having the same meaning.**

## 1. Synonymous schema element names:



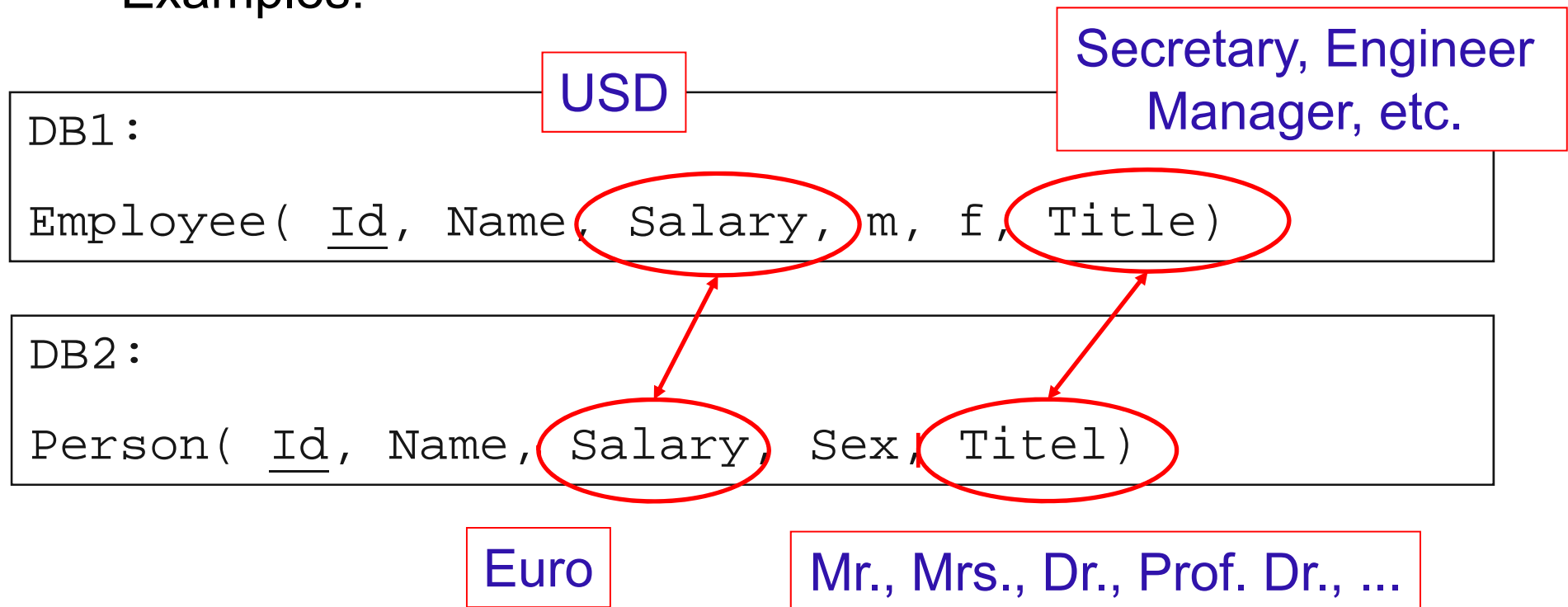
## 2. Synonymous attribute values:

- Different value coding schemas: Manager vs. 2
- Different spellings / abbreviations: Kantstr. vs. Kantstrasse vs. Kant Str.
- Different units of measurement: 1 GB vs. 1000 MB

# Semantic Heterogeneity: Homonyms

**Same words having different meanings.**

- Reason: Different people (in different situations) associate different meanings with the same word.
- Examples:



# Problem: Precision of Concept Definitions

## Business question: How many employees has IBM?

- Definition of Employee:
  - Temporary employees?
  - Students writing master theses?
  - External consultants?
  - Positions in organization chart or currently employed people?
- Definition of IBM
  - Which global region? Which business unit?
  - Include companies that are partly owned by IBM?
- Which point in time?
- How to count people that work part-time?

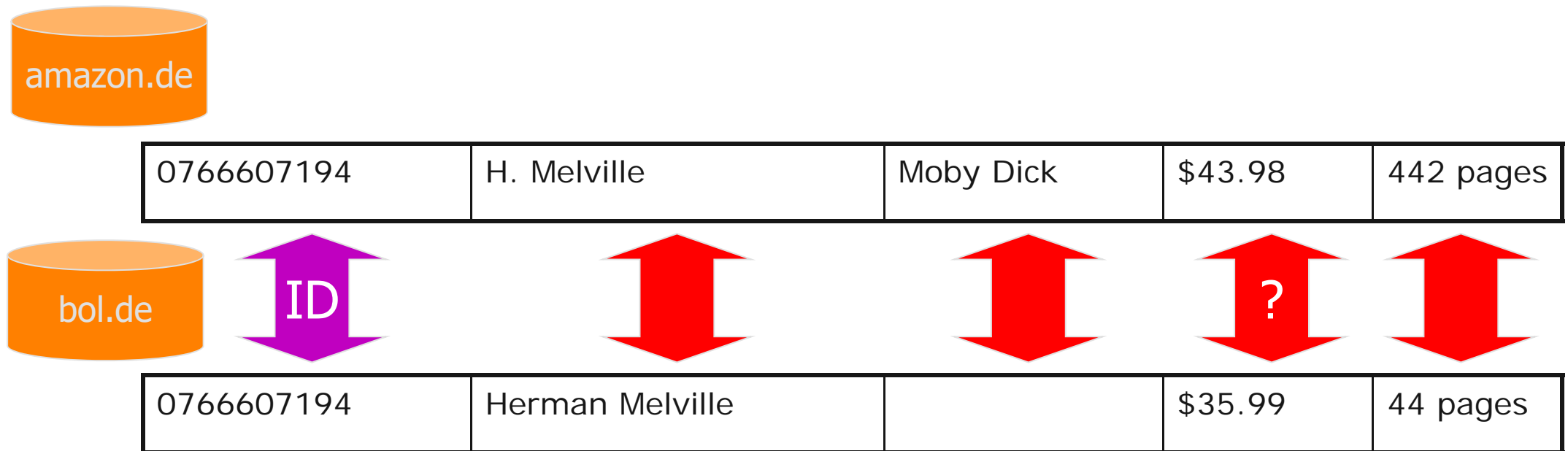
# Semantic Heterogeneity: Object Identity / Duplicates

**Problem: The same real-world entity is often represented**

- **within multiple data sources.**
  - **by multiple records within the same data base.**
- 
- Relevant for: Product data, customer data, scientific data, ...
  - Business question: How much hardware did we sell to the University of Mannheim?
  - Problem: CRM database likely contains multiple records referring to the university itself as well as the different faculties/professors.
  - Reasons for duplicates in the same data base:
    - Different people entered data without identity checks
    - Same entity observed several times
    - No consistent global IDs in input data (ISBN, IBAN, URL, EAN, ...)

# Semantic Heterogeneity: Data Conflicts

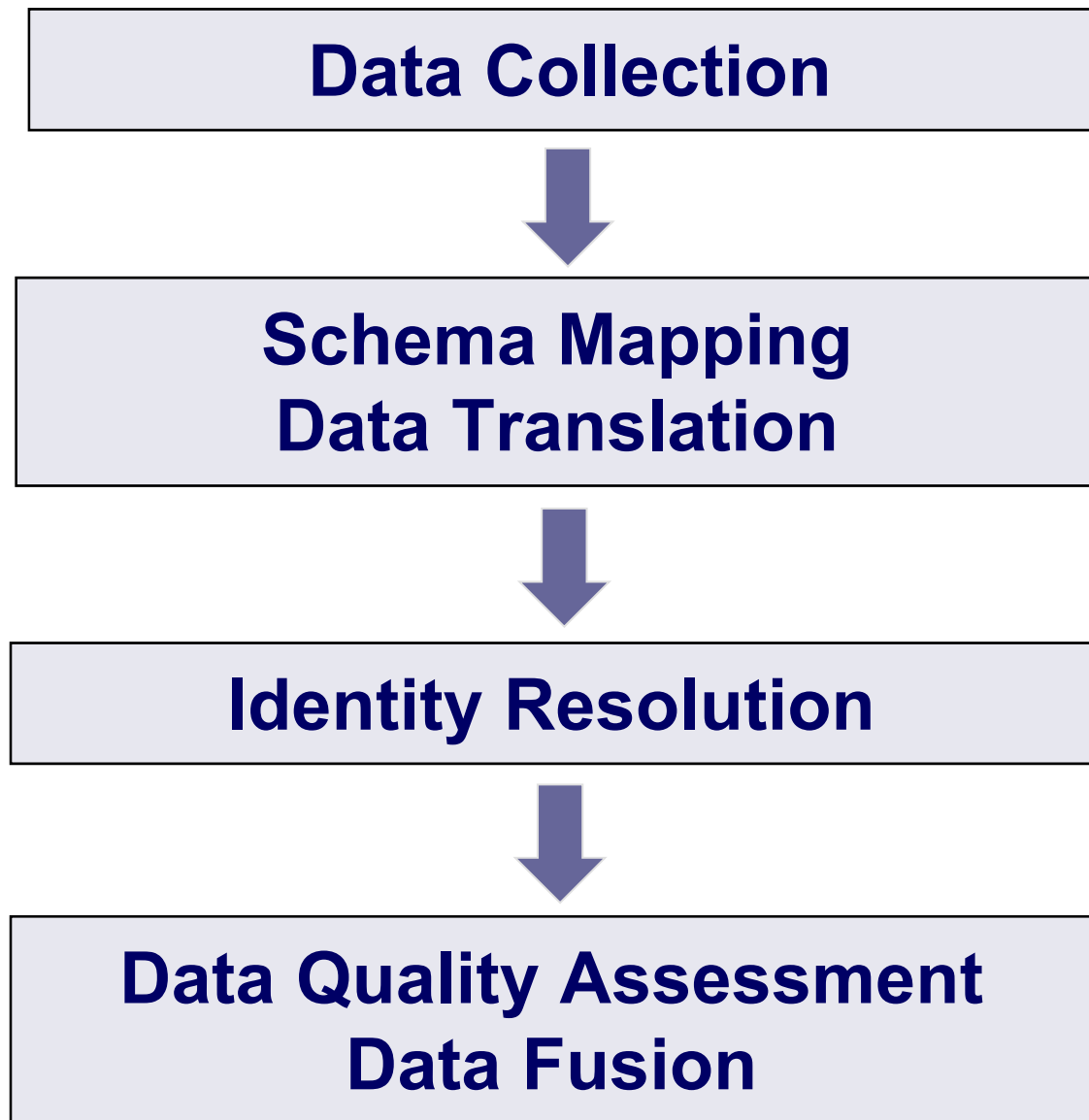
**Problem: Two duplicate records contain different values for the same attribute.**



## Reasons for data conflicts

1. **Errors:** Typos and other errors when data is entered.
2. **Outdated data:** One source/record is older than the other one.
3. **Disagreement:** Different sources actually disagree on the correct value / the truth.

## 5. The Data Integration Process



# 5.1 Data Collection

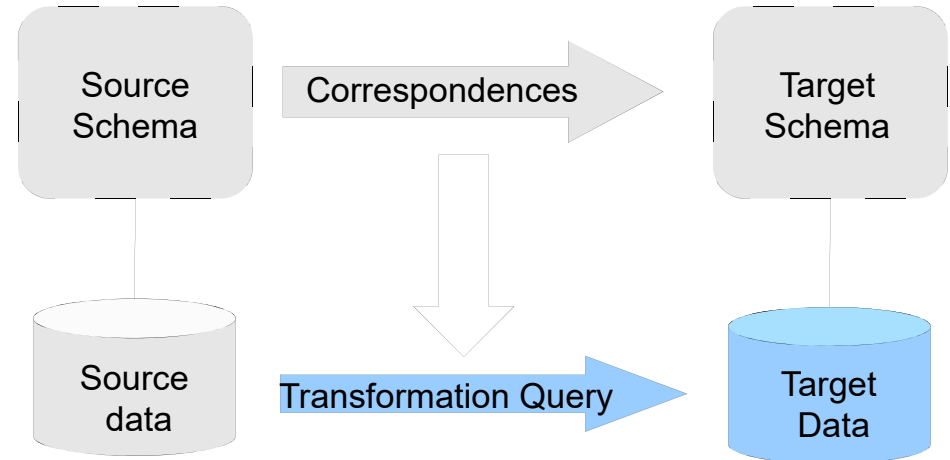
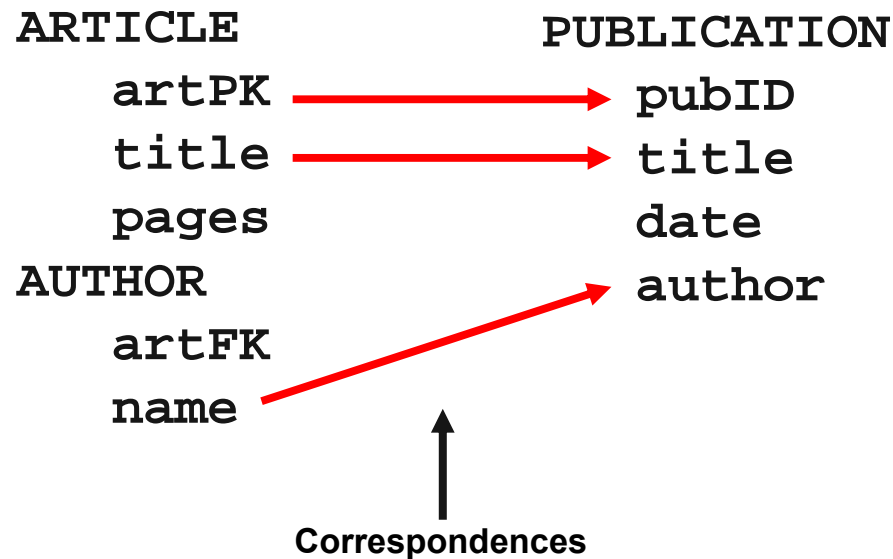
**Goal: Resolve technical and data model heterogeneity so that data from all sources can be accessed / gathered and represented in the same data model.**

- Using middleware libraries that provide
  - different communication protocols (HTTP, ODBC, ...)
  - readers for different data exchange formats (XML, RDF, JSON, ...)
  - for querying remote data sources using different query languages (SQL, SPARQL, ...)
  - for crawling remote data sources (HTML pages, Web APIs, Linked Data)
  - for translating data between different data models (XML-2-Relational, ...)

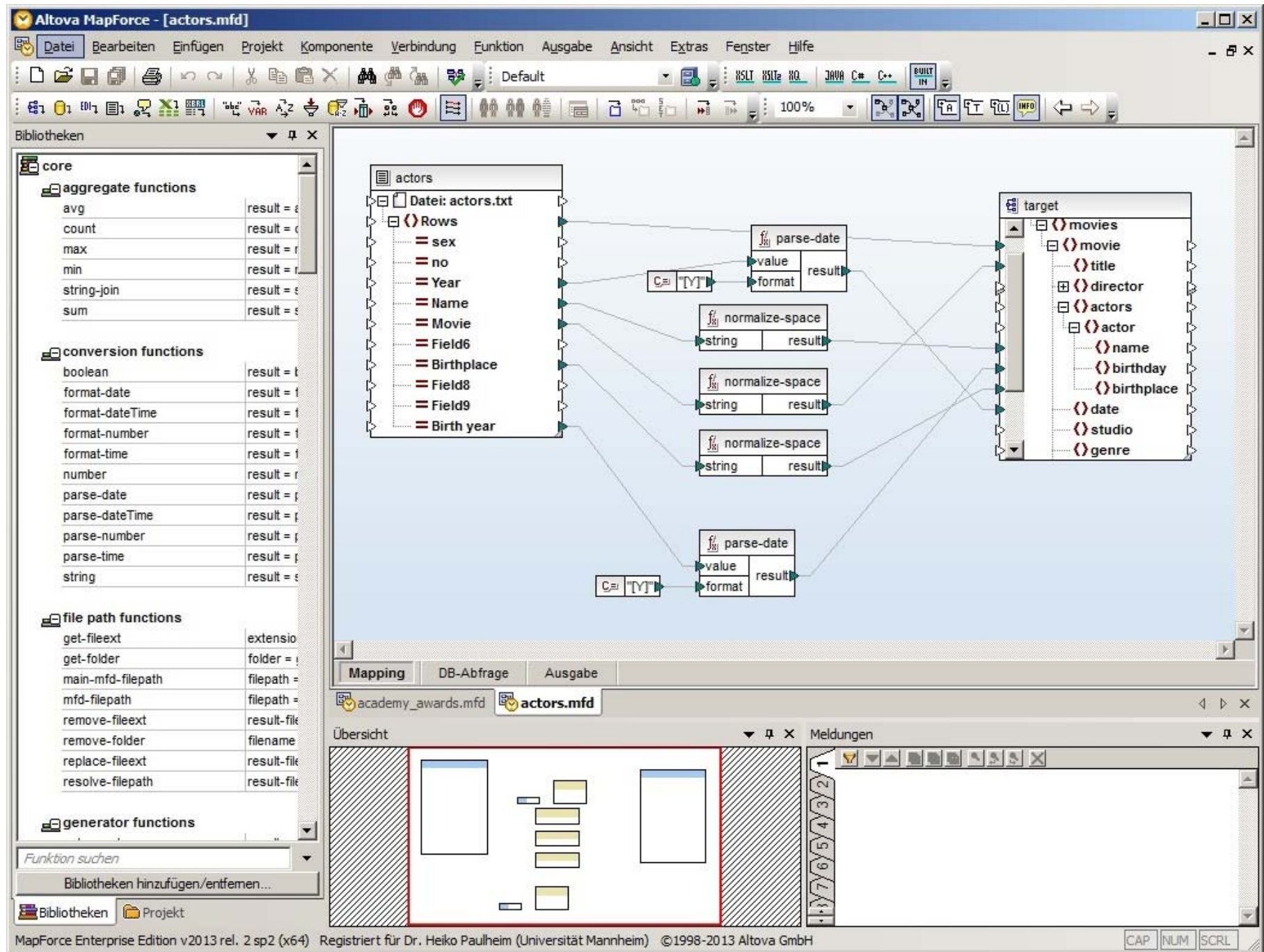
## 5.2 Schema Mapping and Data Translation

**Goal: Resolve structural and schema-related semantic heterogeneity by**

- 1. finding correspondences between the elements of the different schemata.**
- 2. translate data to a single target schema based on these correspondences.**



## Example: Defining Correspondences



## 5.3 Identity Resolution

**Goal: Resolve semantic heterogeneity by identifying all records in all data sources that describe the same real-world entity.**

### ■ Other names for the task:

■ Duplicate Detection, Record Linkage, Entity Matching

### ■ Basic Approach:

1. Compare records using a combination of different similarity metrics
2. If similarity is above threshold → Consider records to describe the same real-world entity

DB1	CID1243	Chris Miller	12/20/1982	Bardon Street, Melville	32 sales
DB2	34	Christian Miller	2/20/1982	7 Bardon St., Melville	24 sales
DB3	427859	Chris Miller	12/14/1973	7 Bardon St., Madison	13 sales

# Example: Combining different Similarity Metrics

Silk Workbench

Workspace: Cora

Editor: linkcora

Generate Links

Reference Links

Learn

About

Export as Silk-LS

Help

Precision = 0.98 | Recall = 0.20 | F-measure = 0.33



## Property Paths

Source: cora

Restriction: ?a ?p ?o .

(custom path)

?a/<http://test.org/author>

?a/<http://test.org/title>

?a/<http://test.org/date>

Target: cora

Restriction: ?b ?p ?o .

(custom path)

?b/<http://test.org/author>

?b/<http://test.org/title>

?b/<http://test.org/date>

## Transformations

Lower case

Merge

Numeric reduce

Regex replace

Remove blanks

## Comparators

Jaccard

Jaro distance

Jaro-Winkler distance

Levenshtein distance

Normalized Levenshtein distance

## Aggregators

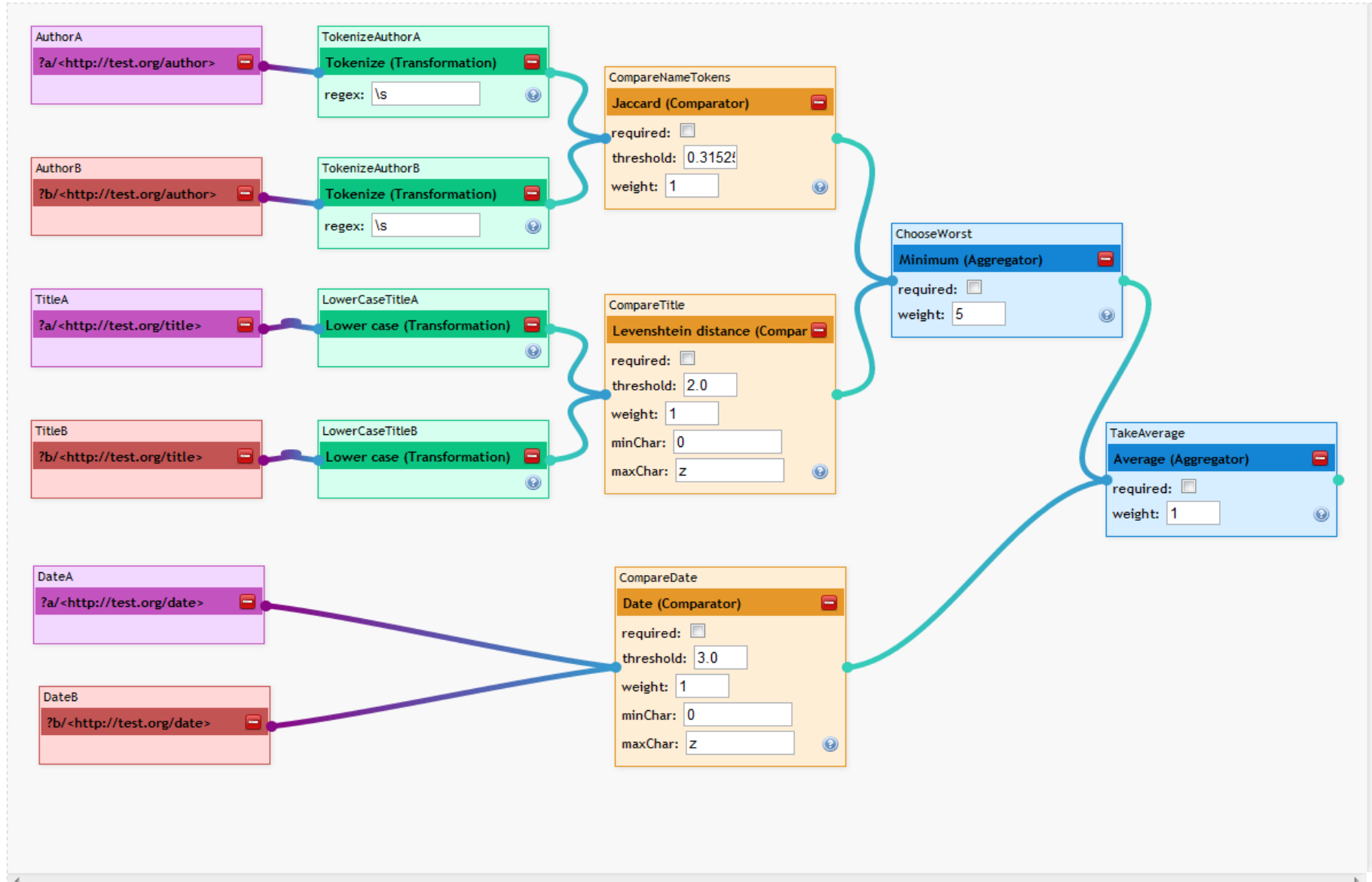
Average

Euclidian distance

Geometric mean

Maximum

Minimum



## 5.4 Data Fusion

**Goal: Resolve data conflicts by combining attribute values of duplicate records into a single consolidated description of an entity.**

### ■ Basic Approach:

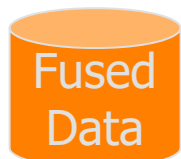
1. Assess the **quality** of data sources / records / values
  - Quality dimensions: timeliness, reputation of source, ...
2. Apply a **conflict resolution function** to choose most promising values or to correct values.
  - Example functions: highest estimated quality, voting, average, ...



CID1243	Chris Miller	12/20/1982	Bardon Street, Melville	32 sales
---------	--------------	------------	-------------------------	----------



34	Christian Miller	2/20/1982	7 Bardon St., Melville	24 sales
----	------------------	-----------	------------------------	----------



	Christian Miller	12/20/1982	7 Bardon Street, Melville	56 sales
--	------------------	------------	---------------------------	----------

# 6. Data Integration Architectures

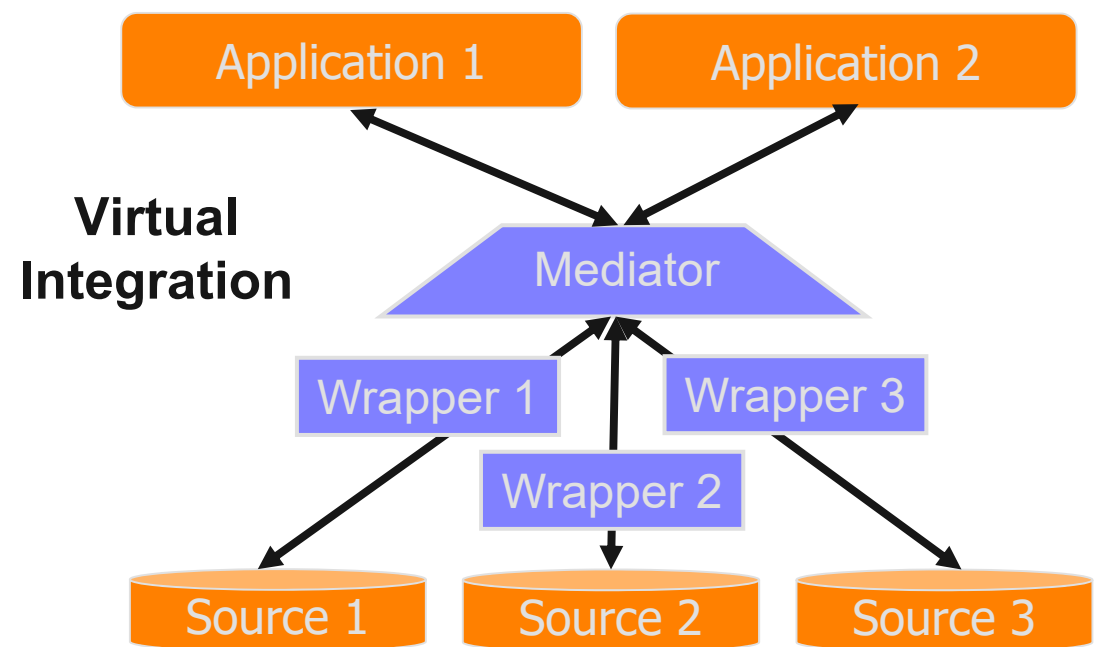
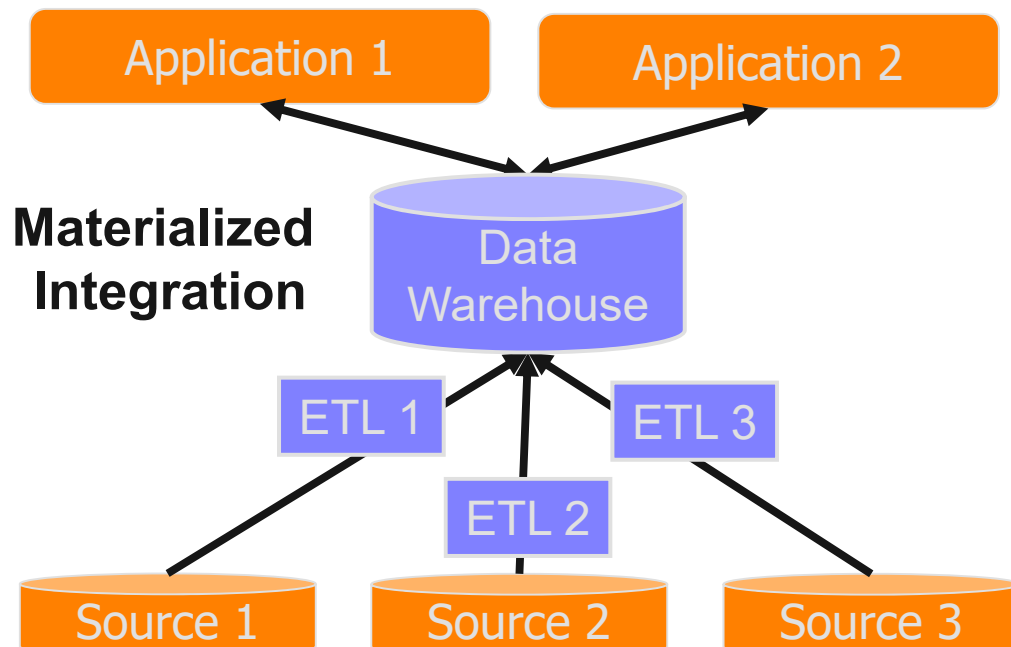
## 1. Materialized Integration

- integrate sources by bringing the data into a single physical database (**data warehouse**).

## 2. Virtual Integration

- leave the data at the sources and access it at query time via wrappers (**integrated view**).

## 3. Numerous intermediate architectures



# Materialized versus Virtual Integration

	Materialized Integration	Virtual Integration
Data currency	Low (regular updates)	High (always current)
Storage requirements	High (copy all data locally)	Low (data remains in sources)
Query processing time	Low (local query processing)	High (slow network traffic)
System Complexity	Low (like normal DB)	High (planning of distributed queries)
Query Expressiveness	High (like normal DB)	Low (as sources might be restricted)
Workload on data source	Can be planned	Hard to plan
Identity Resolution / Data Fusion	possible	difficult (often too slow)

- Rule of thumb: Virtual integration not applicable
  - if 5+ data sources need to be joined.
  - identity resolution and data fusion are important.
- This course illustrates data integration through the **materialized architecture**.

# 7. The Data Integration Software Market

- Market size 2013:  
2.3 billion US\$ (growth: 9.4%)
- Tools for specific tasks
  - Altova Map Force
- Comprehensive solutions covering the complete data integration process
  - Informatica Plattform
  - IBM InfoSphere Information Server
  - SAP Data Services, SAP Data Hub
  - Microsoft SQL Server Integration Services
  - Pentaho Data Integration
- New challengers aiming at Big Data integration
  - Tamr Data Unification Platform



Source: Gartner, Magic Quadrant for Data Integration Tools. Beyer, Thoo, Selvage, August 2017.

# Getting an Impression of the Tools



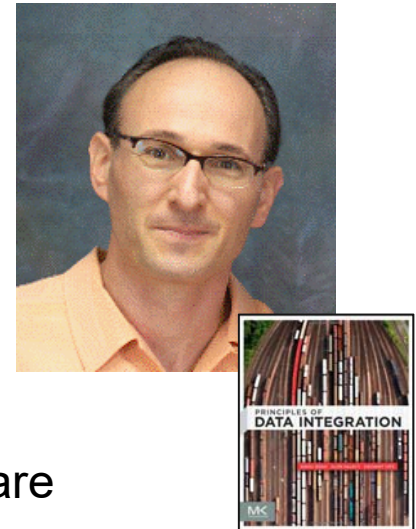
Video tutorials on YouTube

- **SAP Data Hub**  
<https://www.youtube.com/watch?v=CjLc4eDNpso>
- **SAP Information Steward**  
<https://www.youtube.com/watch?v=xrnrtWXI3nc>
- **Informatica PowerCenter**  
<https://www.youtube.com/watch?v=u6oLXidGoqs>
- **Microsoft SQL Server Integration Services**  
<https://www.youtube.com/watch?v=0ikNnenDyNw>

# Setting Expectations

## Alon Halevy: "Data Integration is AI-Complete"

- Meaning that completely automated solutions are unlikely.
- Reasons:
  1. System Level: Managing different platforms, distributed query processing
  2. Logical reasons: Schema and data heterogeneity
  3. Social reasons: Locating relevant data, convincing people to share (data fiefdoms)



### Goal 1:

- Reduce the effort needed to set up an integration application.

### Goal 2:

- Enable the system to perform gracefully with uncertainty (e.g., on the web)

# Summary

- Goal of Data Integration: Abstract away the fact that data comes from multiple sources in varying schemata
- The problem occurs everywhere: Handling it is curial for many applications in business, science, government, and Web
- Architectures range from warehousing to virtual integration
- Regardless of the architecture, bridging heterogeneity is the key issue
- Goal: Reduce the human effort involved