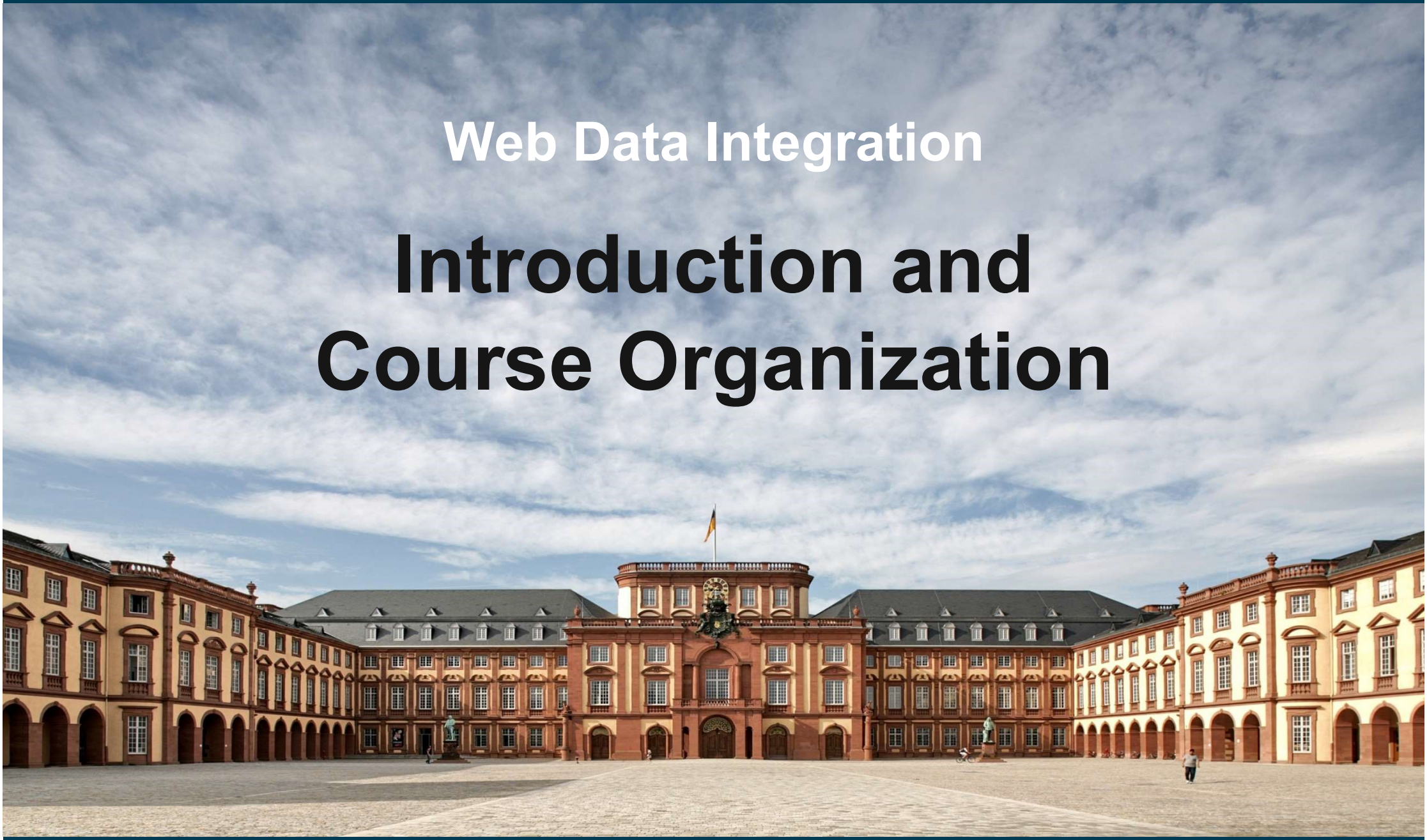
UNIVERSITY OF MANNHEIM

Web Data Integration

Introduction and Course Organization



Hallo

- **Prof. Dr. Christian Bizer**
- Professor for Information Systems V
- Research Interests:
  - Web-based Systems
  - Large-Scale Data Integration
  - Data and Web Mining
  - Large Language Models
- Room: B6, 26 - B1.15
- eMail: chris@informatik.uni-mannheim.de

- Will teach the lecture (IE670)

Hallo

- **M. Sc. Wi-Inf. Alexander Brinkmann**
- Graduate Research Associate
- Research Interests:
  - Data Search using Deep Learning
  - Information Extraction using LLMs
- Room: B6, 26, C 1.04
- eMail: alexander.brinkmann@uni-mannheim.de

- Will teach the exercises and will supervise student projects (IE683)

Hallo

- **M. Sc. Wi-Inf. Ralph Peeters**
- Graduate Research Associate
- Research Interests:
  - Entity Matching using Deep Learning
  - Product Data Integration using LLMs
- Room: B6, 26, C 1.04
- eMail: ralph.peeters@uni-mannheim.de

- Will teach the exercises and will supervise student projects (IE683).

Outline

1. Course 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 Organization

The Lecture (IE670)

- introduces the principal 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 the number of participants, registration via Portal2
- 3 ECTS (offline exam: 60 minutes, Thursday, 12.12.24)

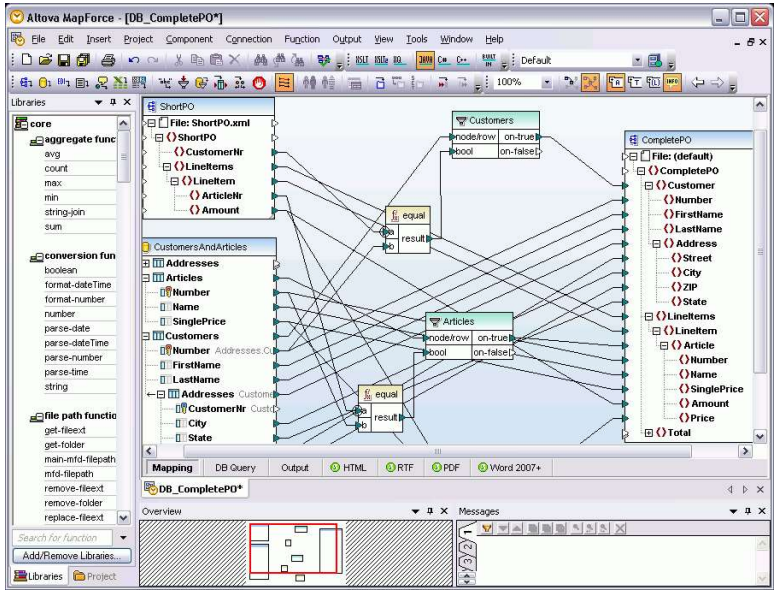
The Student Projects (IE683)

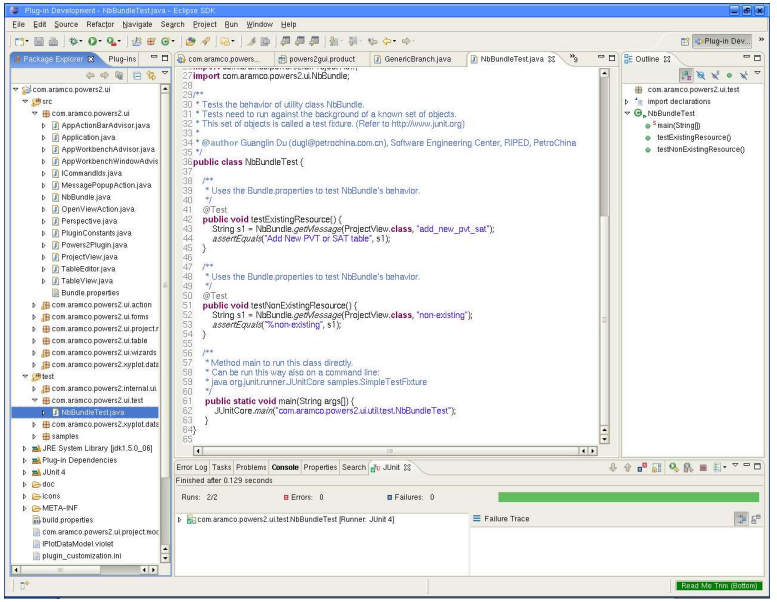
- teams of **five students** realize a data integration project including
  1. data gathering
  2. schema mapping and data translation
  3. identity resolution
  4. data quality assessment and data fusion
- teams write a 12-page report about their project, present project results
- you may choose their own application domain and data sets
  - minimum 3 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, products, restaurants, companies, geographic information
- restricted to 60 participants, registration via Portal2
- 3 ECTS (70 % written project report, 30 % presentation of project results)

The Exercise

In the exercise sessions, Ralph and Alex give you an introduction to tools that you can use for your projects. You experiment with the tools along the use case of integrating data about films.

1. Data Translation
   - Altova MapForce
   - graphical mapping and data translation tool
2. Identity Resolution
   - Winte.r Data Integration Framework
   - provides matching methods
3. Data Fusion
   - Winte.r Data Integration Framework
   - provides conflict resolution methods



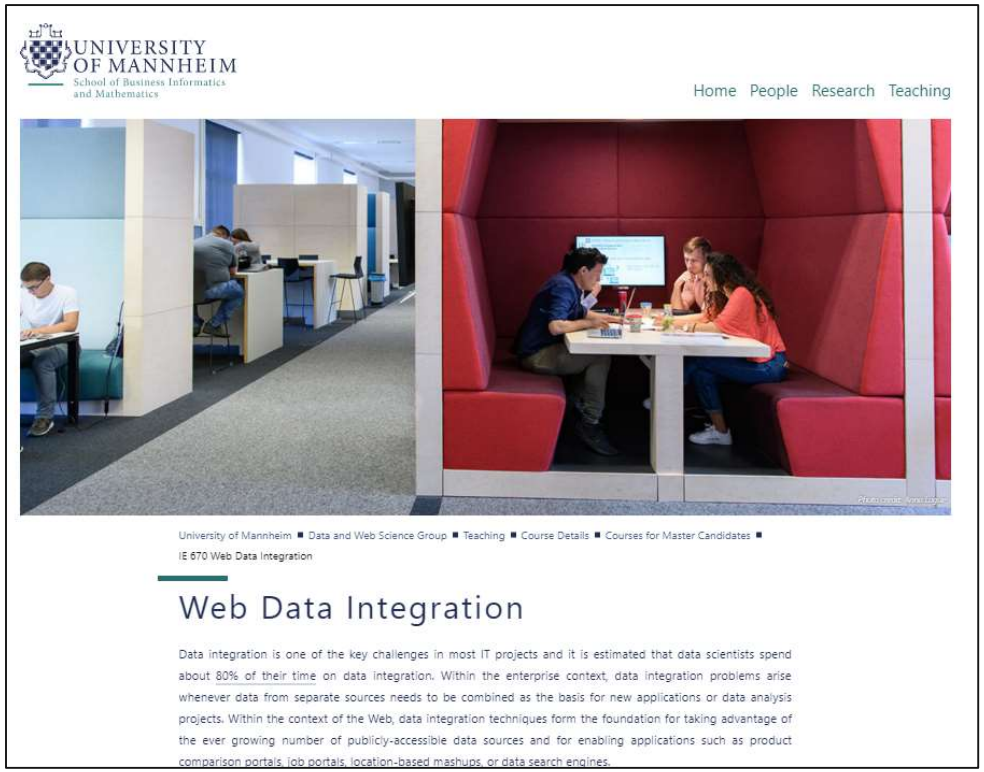


Schedule

Week	Wednesday	Thursday
04.09.2024	**Lecture:** Introduction to Web Data Integration	- no lecture -
11.9.2023	**Lecture:** Structured Data on the Web	**Lecture:** Data Exchange Formats
18.9.2024	**Lecture:** Data Exchange Formats	**Lecture:** Schema Mapping
25.9.2024	**Lecture:** Schema Mapping	**Project:** Introduction to Student Projects
02.10.2024	**Exercise:** Introduction to MapForce	**Coaching:** Schema Mapping
09.10.2024	**Project:** Feedback about Project Outlines	**Lecture:** Identity Resolution
16.10.2024	**Lecture:** Identity Resolution	**Exercise:** Identity Resolution
23.10.2024	**Project Work:** Identity Resolution	**Coaching:** Identity Resolution
30.10.2024	**Project Work:** Identity Resolution	**Coaching:** Identity Resolution
06.11.2024	**Lecture**: Data Quality and Data Fusion	**Lecture:** Data Quality and Data Fusion
13.11.2024	**Exercise:** Data Quality and Data Fusion	**Project Work:** Data Quality and Data Fusion
20.11.2024	**Project Work:** Data Quality and Fusion	**Coaching:** Data Quality and Fusion
27.11.2024	**Project Work:** Data Quality and Fusion	**Coaching:** Data Quality and Fusion
04.12.2024	**Presentation of Project Results**	**Presentation of Project Results**
12.12.2024	Final Exam	

Course Organization

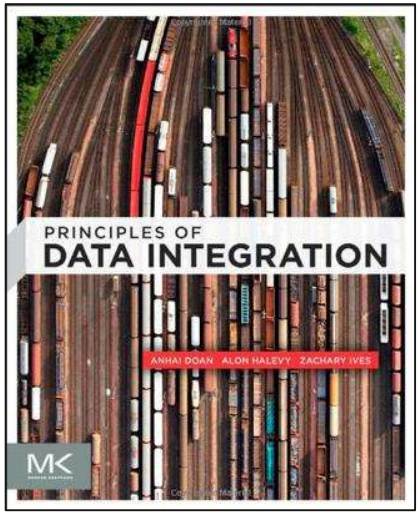
- Course Webpage
  - https://www.uni-mannheim.de/dws/teaching/course-details/courses-for-master-candidates/ie-670-web-data-integration/
  - The lecture slides are published on this webpage (as part of schedule).
  - Exercise materials will be provided on this webpage.

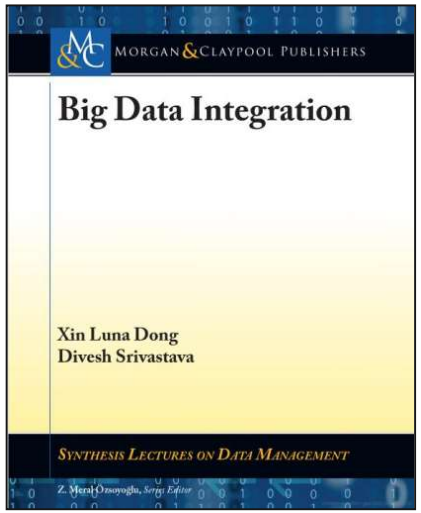


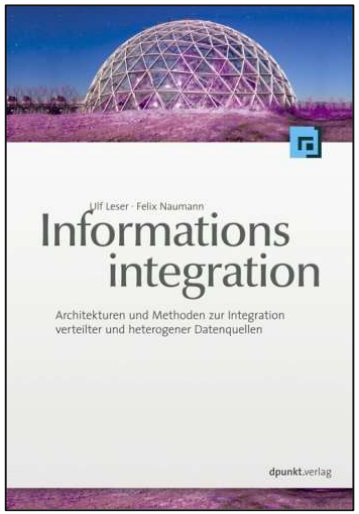
- Time and Location
  - Wednesday, 15:30 to 17:00. B6 A101
  - Thursday, 13:45 to 15:15. B6 A101
  - Start: 4.9.2024

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**. Dpunkt Verlag, 2007. (several copies in the library, Video lecture at https://www.tele-task.de/series/1293/)
4. Peter Christen: **Data Matching**. Springer, 2012.







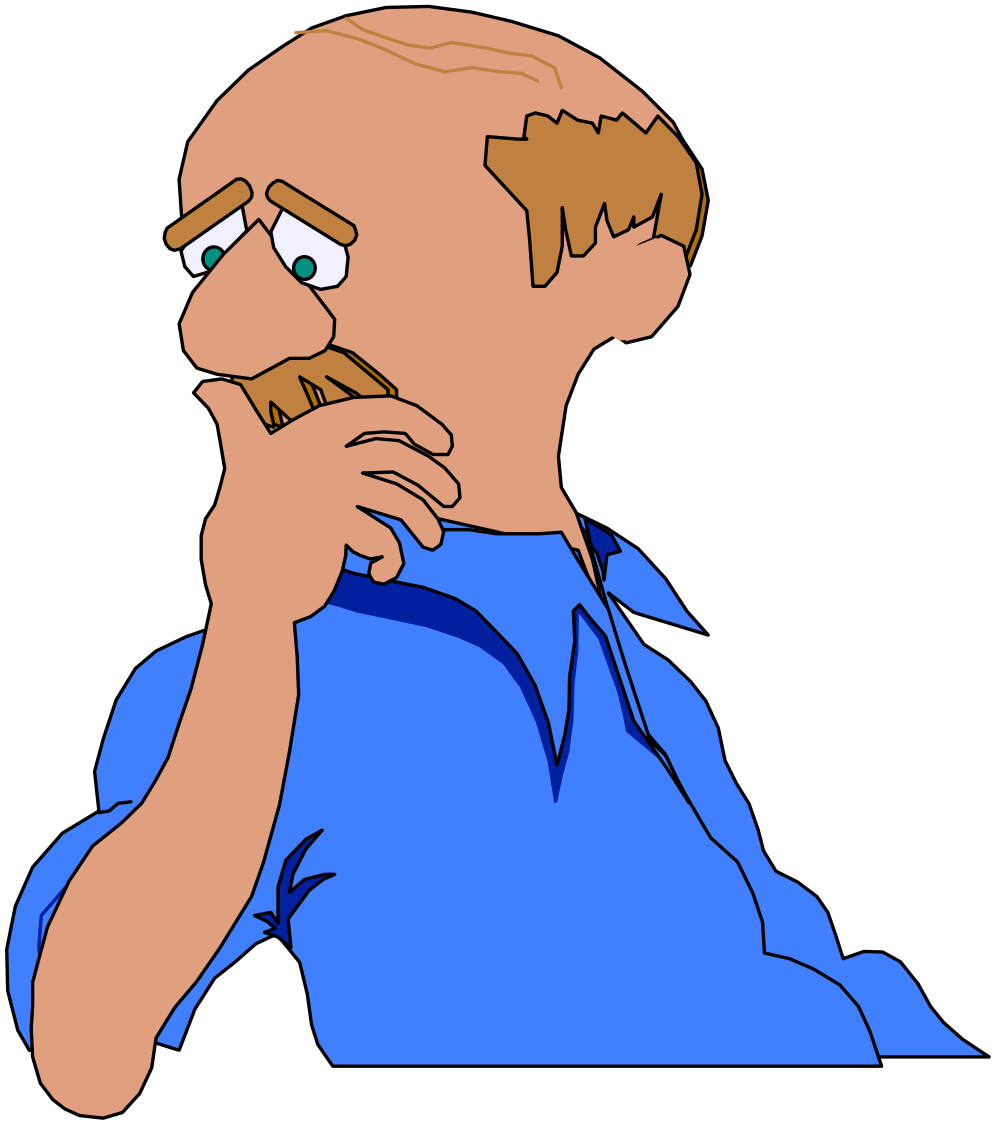
Credits

The slide set of this lecture builds on slides from:

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

Lots of thanks to all of you!

Questions about the Course Organization?



2. What is Data Integration?

- Databases and machine learning frameworks 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

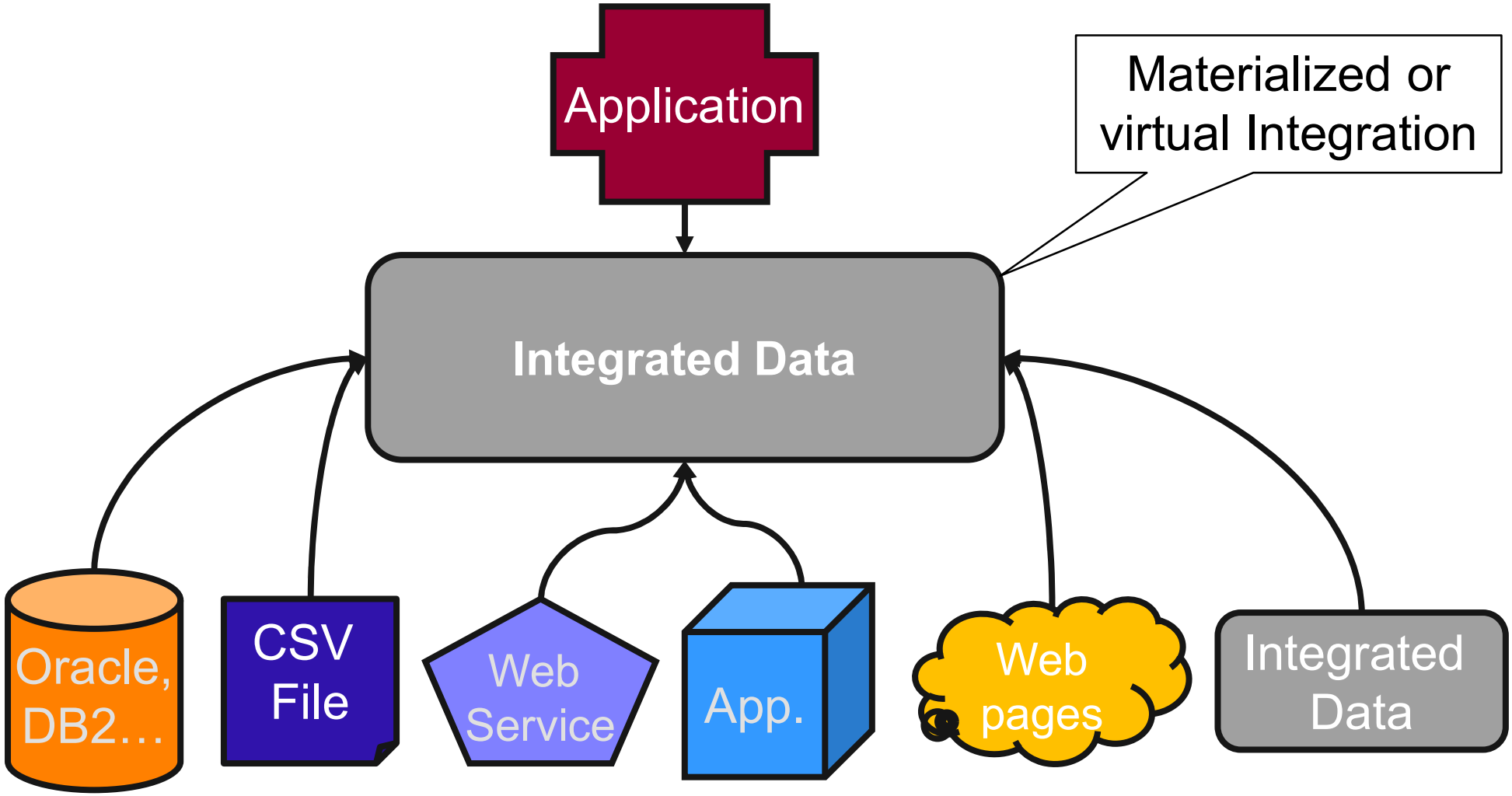
- In reality, applications often need to work with data from multiple independently created data sources
  1. different sources use different data models
  2. different sources use different schemata
  3. different sources describe the same real-world entity
  4. different sources provide conflicting data about a single entity
  5. different sources provide different limited query interfaces to their data

Definition of Data Integration

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

- The integrated data 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 each 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

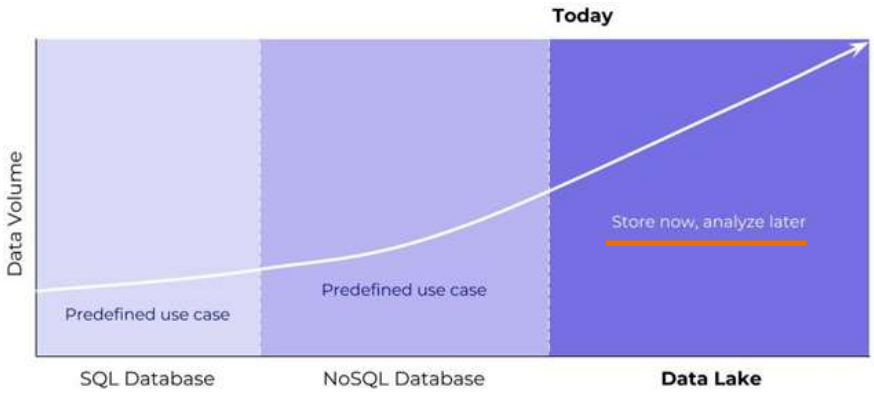
Overview: Traditional Data Integration



Data Lake Scenario

- Data lakes
  - are repositories of raw data in different formats
  - collect or generate metadata about datasets
  - provide a common access interface

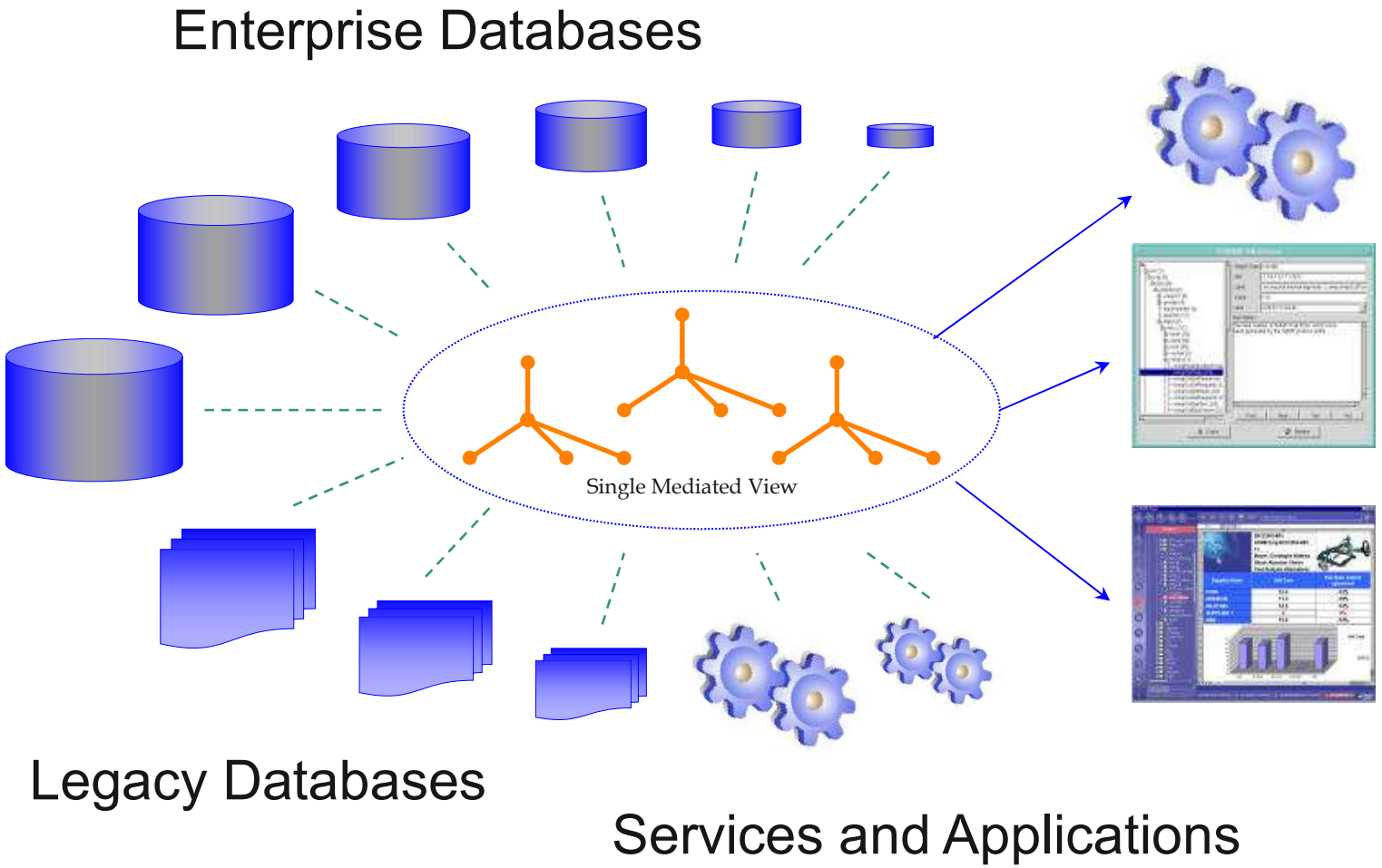
- different, not yet known use cases
- are used in a schema-on-read fashion: Pay-as-you-go integration
- target users: data scientists



3. Application Areas of Data Integration

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

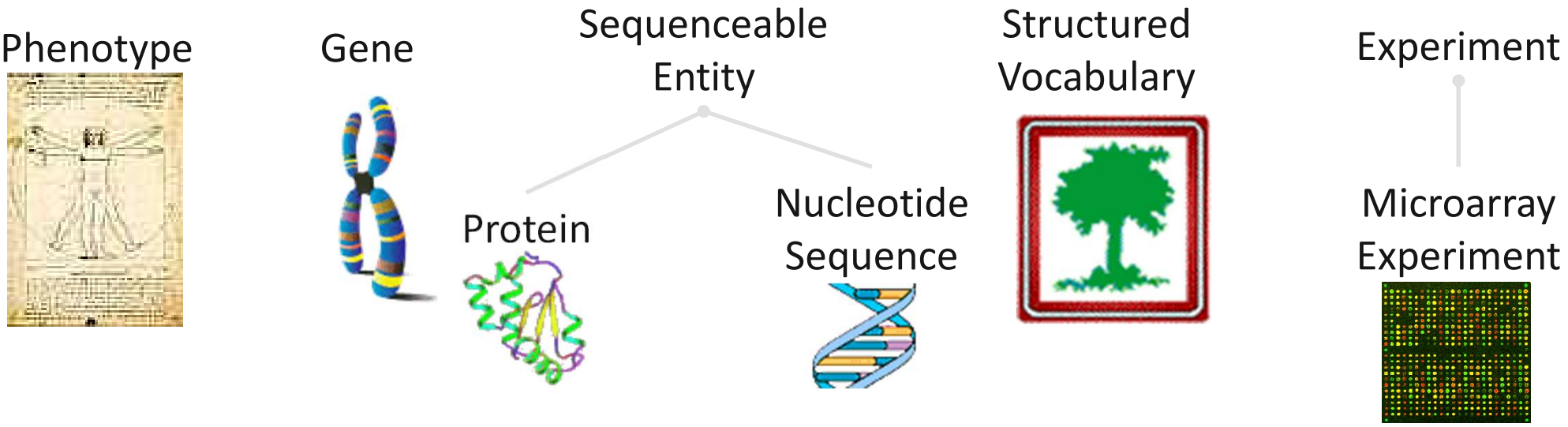
Application Area: Business

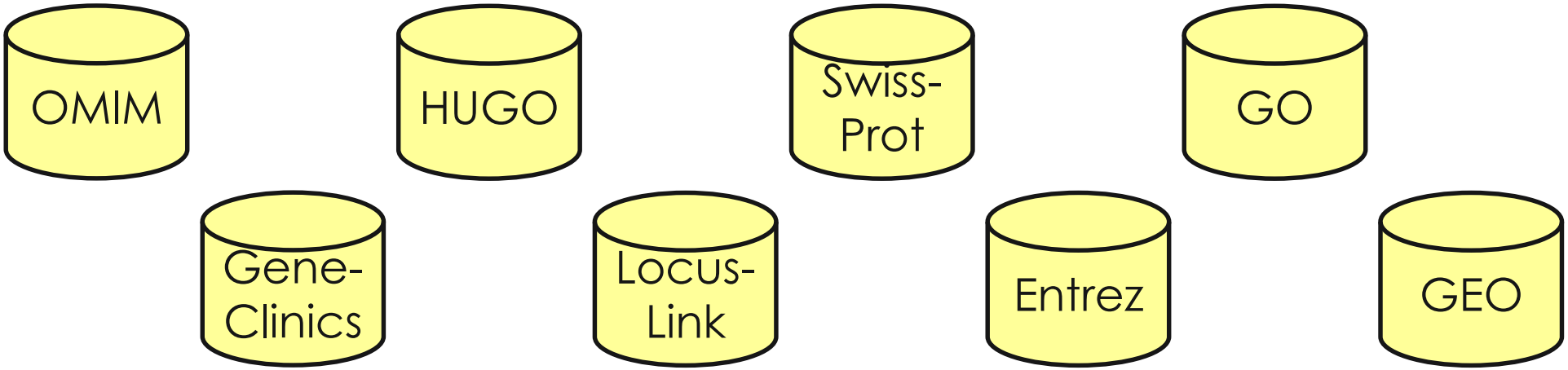


- CRM
- SCM
- Company Mergers
- …

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 mine unknown amounts of data from various sources to identify or rate individuals.

- GPS location data
- cell phone calls
- online profiles (Facebook)
- web browsing behavior
- credit card transactions
- intelligence from other agencies
- …

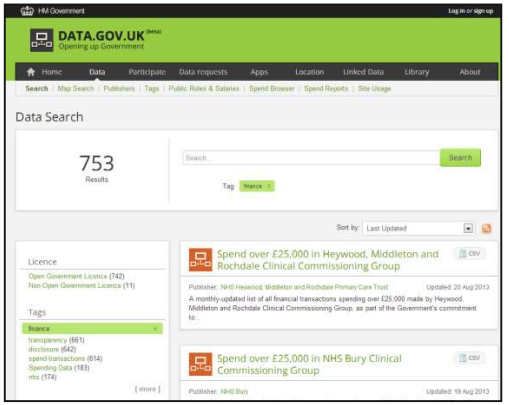
Application Area: Data Journalism

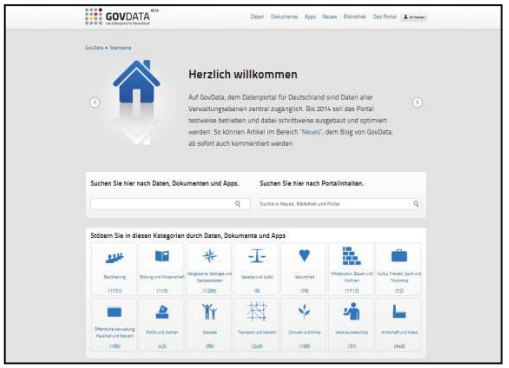
- Government data is increasingly published under open licenses on the Web
- Journalists discover stories by combining data from different sources

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

Members of parliament
- donations / membership in company boards
- voting behavior

Panama Papers
- ownership information about company networks
- discussable financial transactions





Application Area: The Web

for instance, online shopping

amazon

KATSUKI BOOKS

Shelf Books.com

Quilt-Books.Com

COMPU-BOOKS

Afrikan Books.com

NGHOSI books

Providence-Books

USBORNE CHILDREN BOOKS

God Gave His Word

КНИГАРНЯ bookshop online

BORDERS.com

CACTUS-BOOKS.COM

metro-Books

BARNES & NOBLE www.bn.com

StoneCreekBooks.com

Professional Books Inc.

ASTROLOGY-BOOKS.COM

SeeMe4Books

HALF PRICE COMPUTER BOOKS

DANCE BOOKS

acma BOOKS .COM

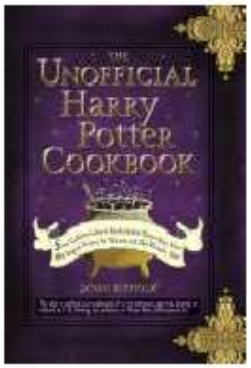
NORTH 49 BOOKS

funny face Bookstore

Just Great Books

Comparison Shopping

Google harry potter books 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

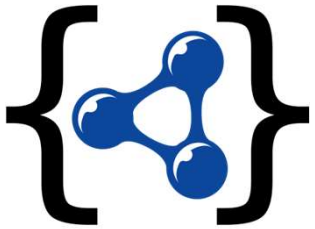
Sellers	Seller Rating	Details	Base Price	Total Price	
MovieMars.com	★★★★☆ (42)	Free shipping	$20.92		Shop »
ValoreBooks.com	No rating	No tax	$3.24 $3.95 shipping	$7.19	Shop »
Overstock.com	★★★★★ (5,886)		$12.93		Shop »

Structured Data on the Web (Topic of the next lecture)

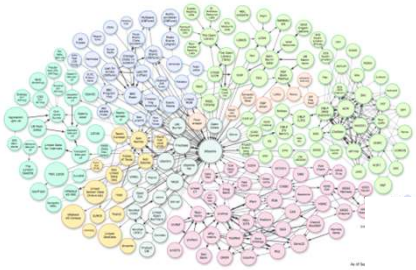
More and more Websites

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

JSON-LD



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.

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, XPath XQuery, 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

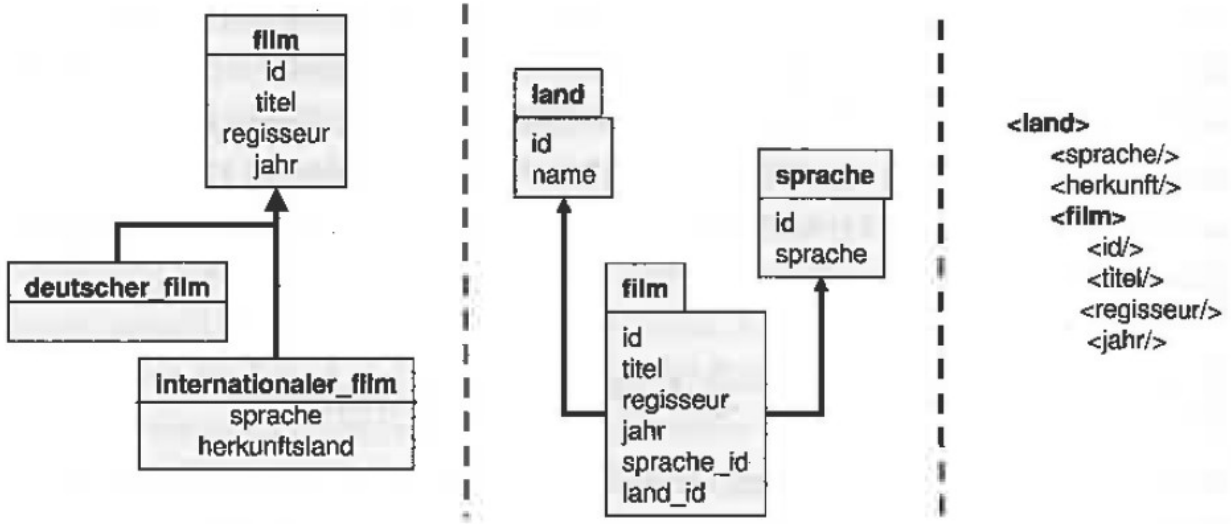
- Synonymous values
  - 1GB versus 1000MB ➔ Semantic heterogeneity
- Structural differences
  - First name: Chris, last name: Bizer versus name: Chris Bizer ➔Structural heterogeneity

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. Graph data models (property graphs, RDF)
4. Object-oriented data model



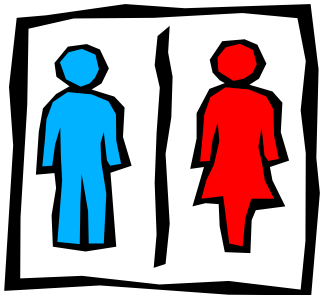
Object-oriented relational XML

Structural Heterogeneity

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

1. Normalized versus Denormalized
2. Below 1NF Attributes versus Multiple Attributes
3. Nested versus Foreign Key Relationship
4. Alternative Modeling
   - Attribut vs. Value
   - Relation vs. Attribute
   - Relation vs. Value
   - Example: See next slide …

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.

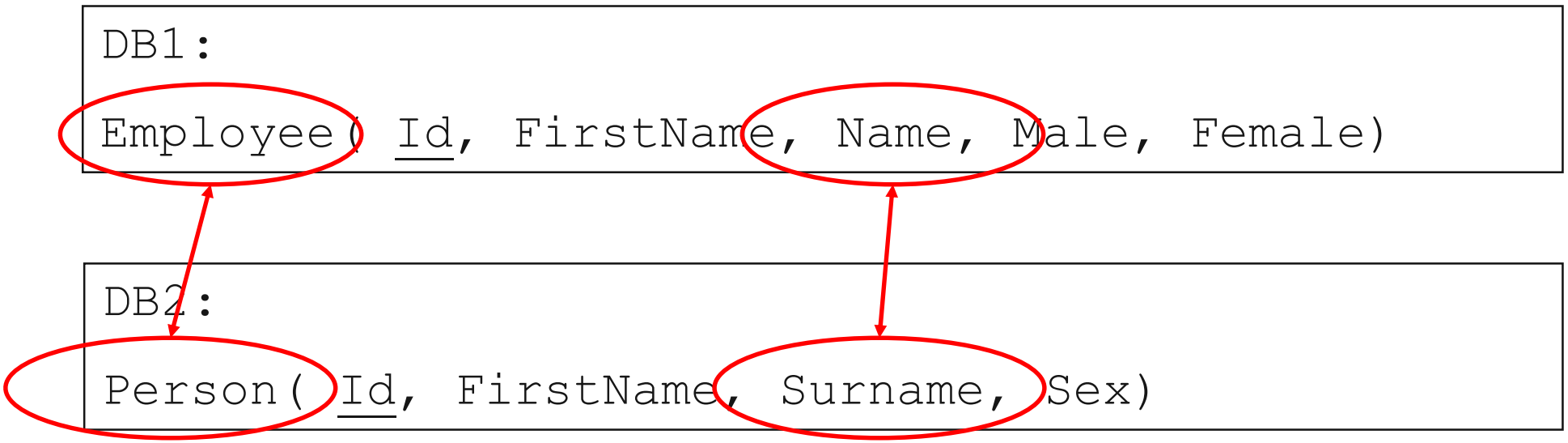
1. Naming Conflicts
   - synonyms, homonyms, slightly deviating concepts
2. Object Identity / Duplicates
   - multiple data sources as well as multiple records within one data source may describe the same real-world entity
   - Which “Marie Müller” does a record describe?
3. Data Conflicts
   - conflicting data about the same real-world entity in different data sources as well as within different records in the same data source

Main focus of this course!

Naming Conflicts: Synonyms

Different words having the same meaning.

1. Synonymous schema element names:

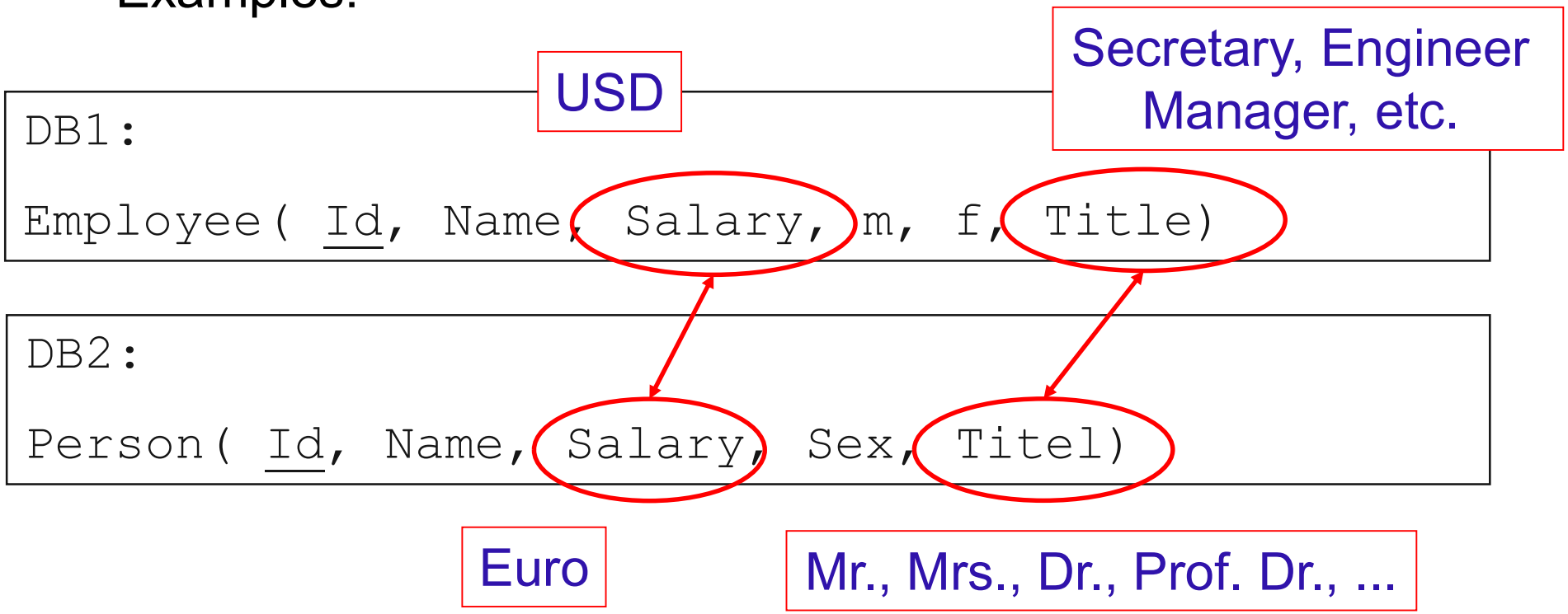


2. Synonymous attribute values / surface forms:
   - Different value coding schemas: Manager vs. 2
   - Different spellings / abbreviations: Kantstr. vs. Kantstraße vs. Kantstrasse
   - Different units of measurement: 1 GB vs. 1000 MB

Naming Conflicts: Homonyms

Same words having different meanings.

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



Object Identity / Duplicates

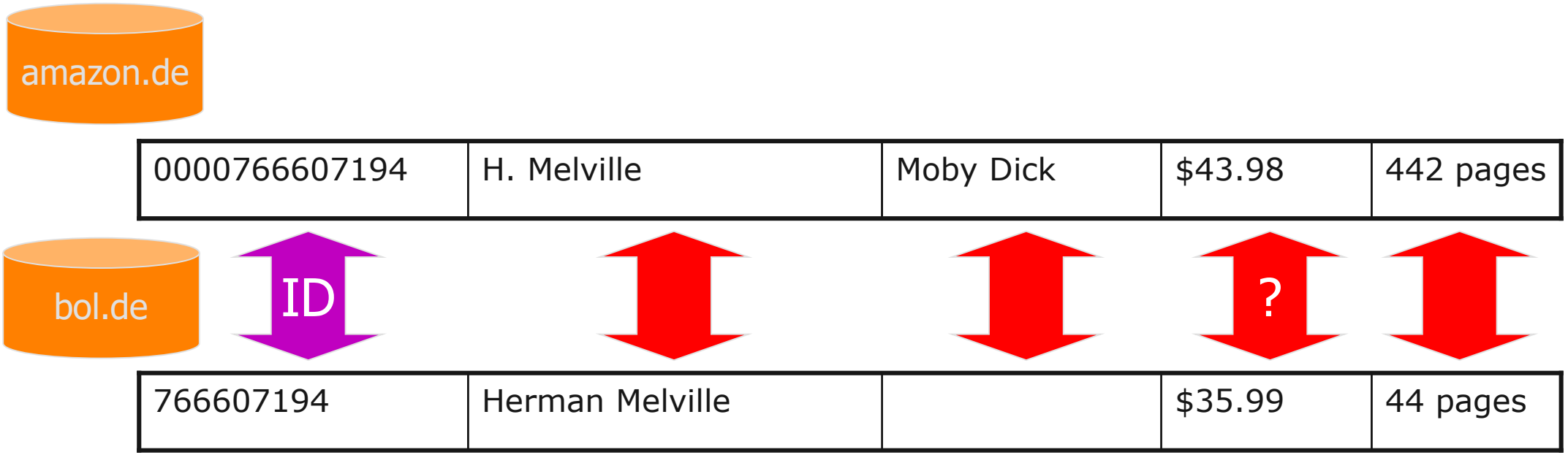
Problem: The same real-world entity is often represented

- **within multiple data sources.**
- **by multiple records within the same data base.**

- Relevant for: Product data, customer contact 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/chairs.
- Reasons for duplicates in the same data base:
  - different people enter data without identity checks
  - same entity observed several times
  - no consistent global IDs in input data (ISBN, GTIN, EAN, DUNS, …)

Data Conflicts

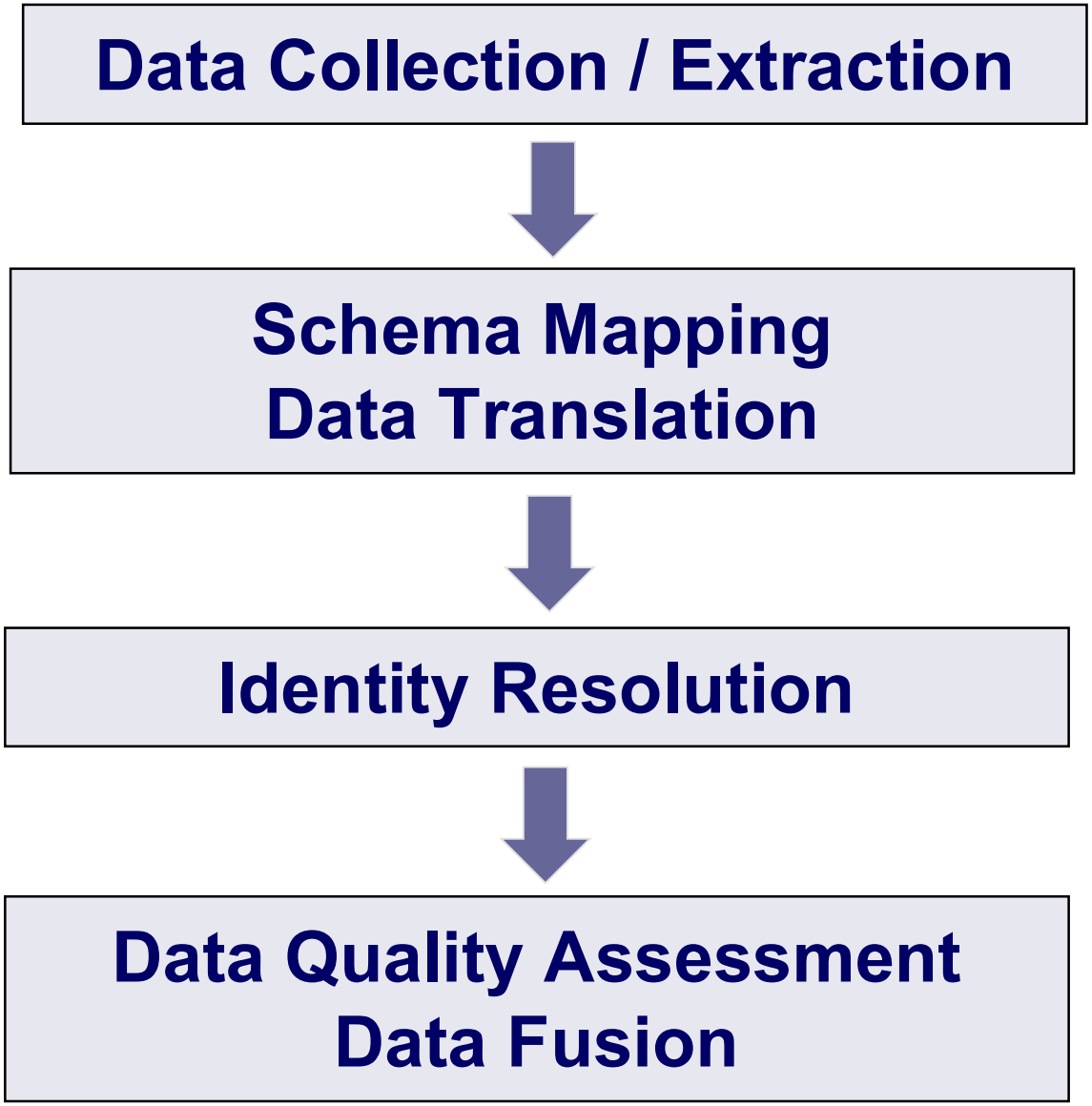
Problem: Two duplicate records contain different values for the same attribute.



Reasons for data conflicts

1. Errors: Typos and other errors when data is entered or matched
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

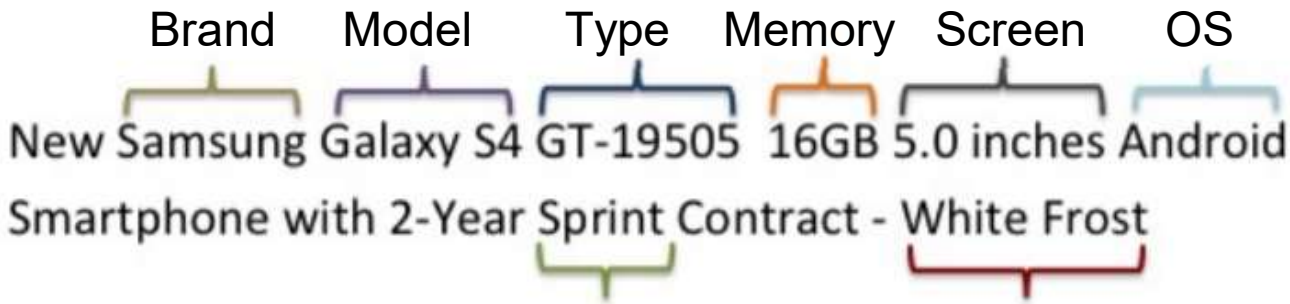
Goal: Resolve technical and data model heterogeneity so that data from all sources can be accessed / gathered and is represented in the same data model.

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

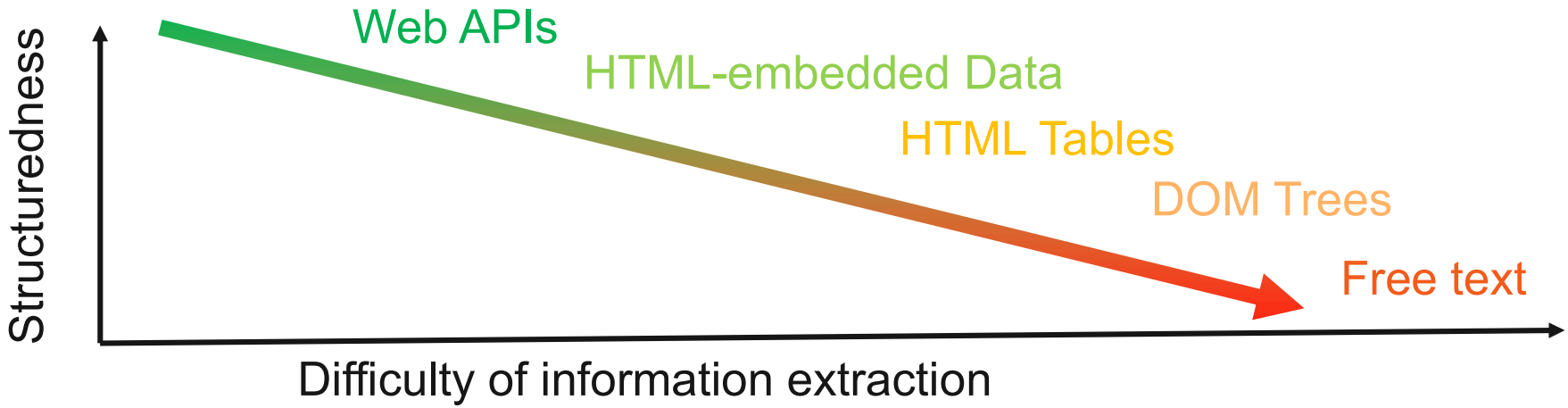
Information Extraction

Goal: Automatic extraction of structured information from unstructured or semi-structured content.

- Example of below 1NF data:



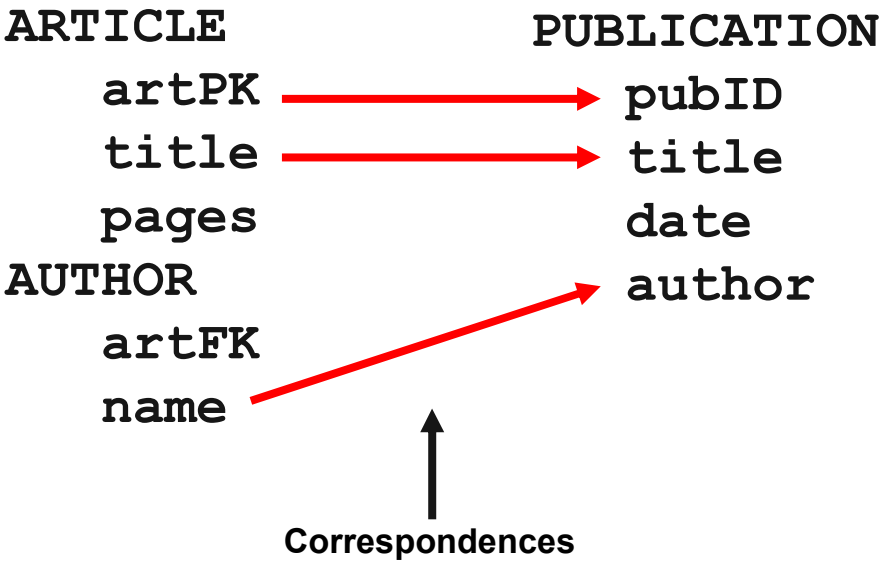
- Successful application area of LLMs
- The difficulty of the extraction depends on the structuredness

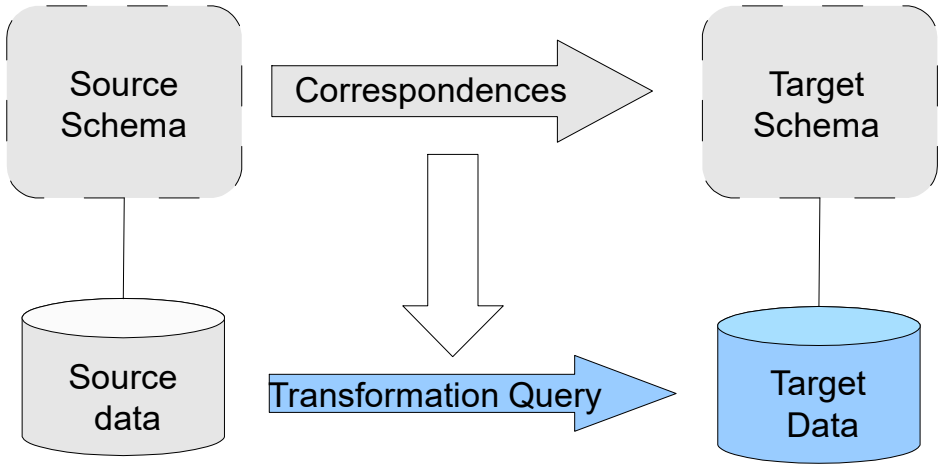


5.2 Schema Mapping and Data Translation

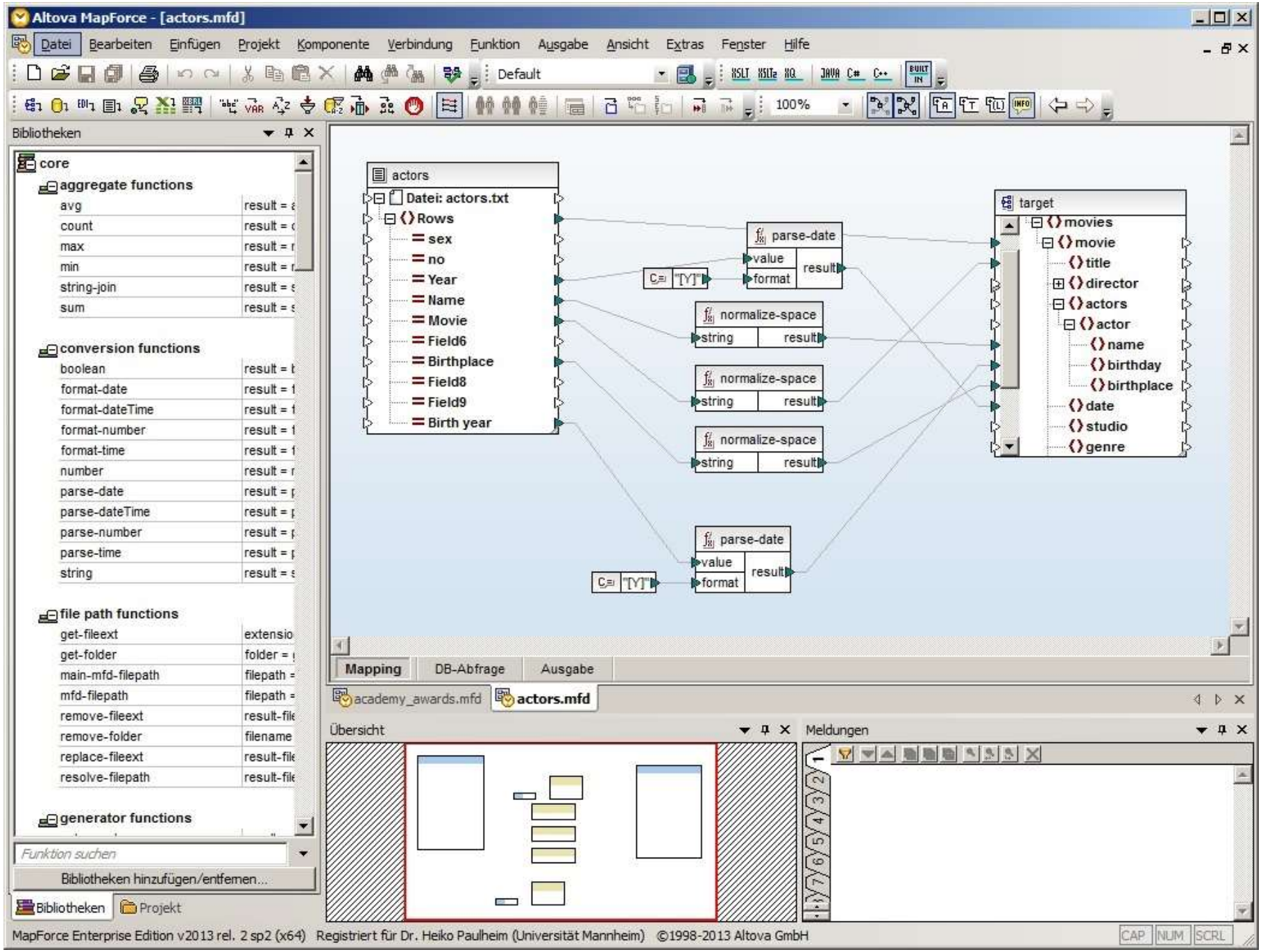
Goal: Resolve structural and schema-related semantic heterogeneity by

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





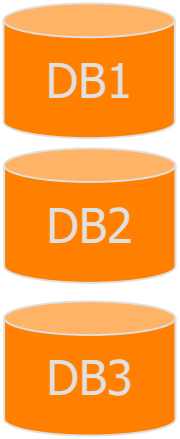
Example: Defining Correspondences



5.3 Identity Resolution

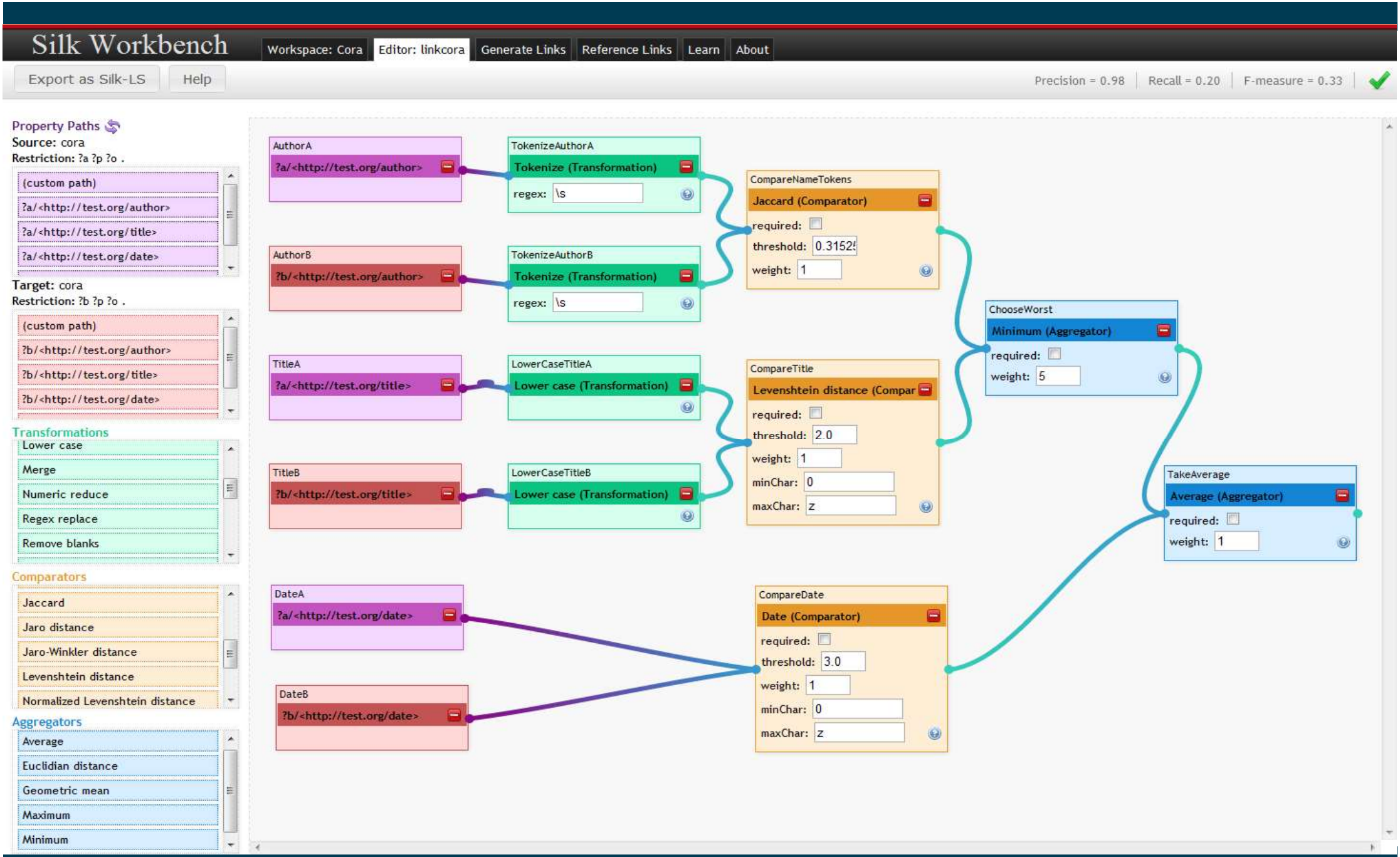
Goal: Identifying all records in all data sources that describe the same real-world entity.

- **Other names for the task:**
  - **Entity Matching, Data Matching, Duplicate Detection, Record Linkage**
- **Basic Approach:**
  1. **Compare records using a combination of attribute-specific similarity metrics**
  2. **If record are similar enough ➔ consider records to describe the same real-world entity**



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

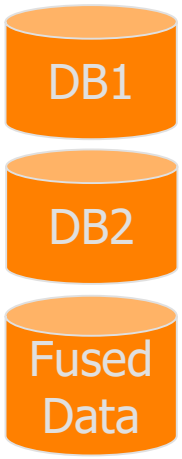
Example: Combining different Similarity Metrics



5.4 Data Fusion

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

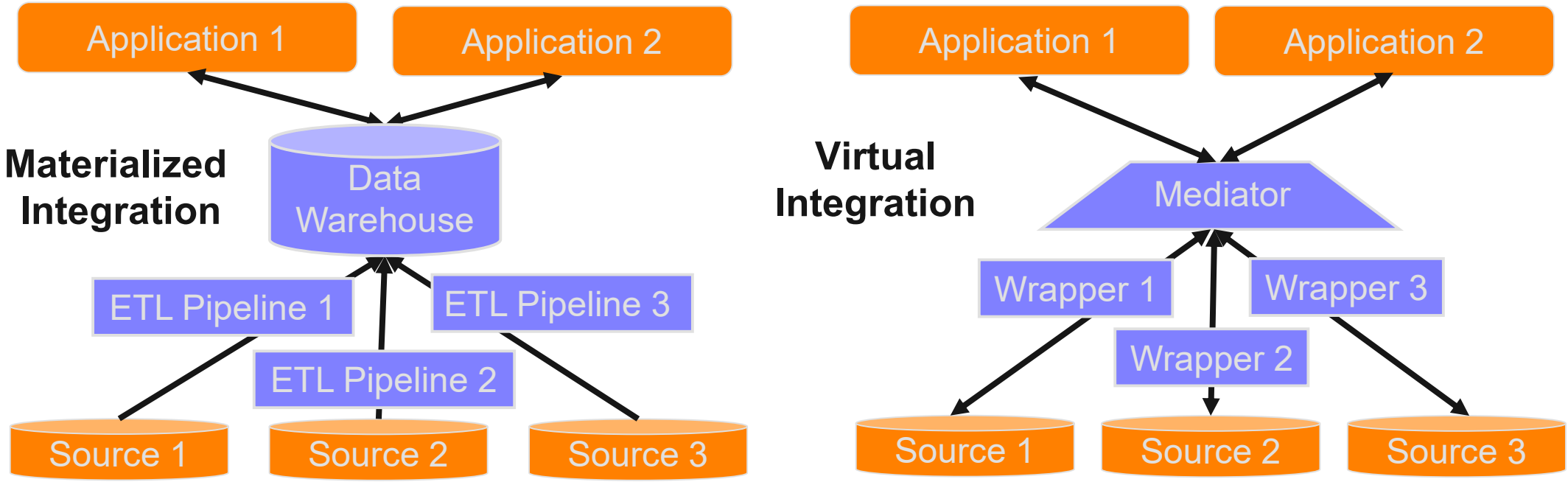
- **Basic Approach:**
  1. **Assess the quality of data sources / records / values**
     - **Quality dimensions: timeliness, reputation of source, …**
  2. **Apply a conflict resolution function to choose most promising values or to correct values**
     - **Example functions: highest estimated quality,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



EAN1243	Chris Miller	12/20/1982	Bardon Street, Melville	32 sales
34	Christian Miller	2/20/1982	7 Bardon St., Melwille	24 sales
EAN1243	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. **Data Lake Architectures**



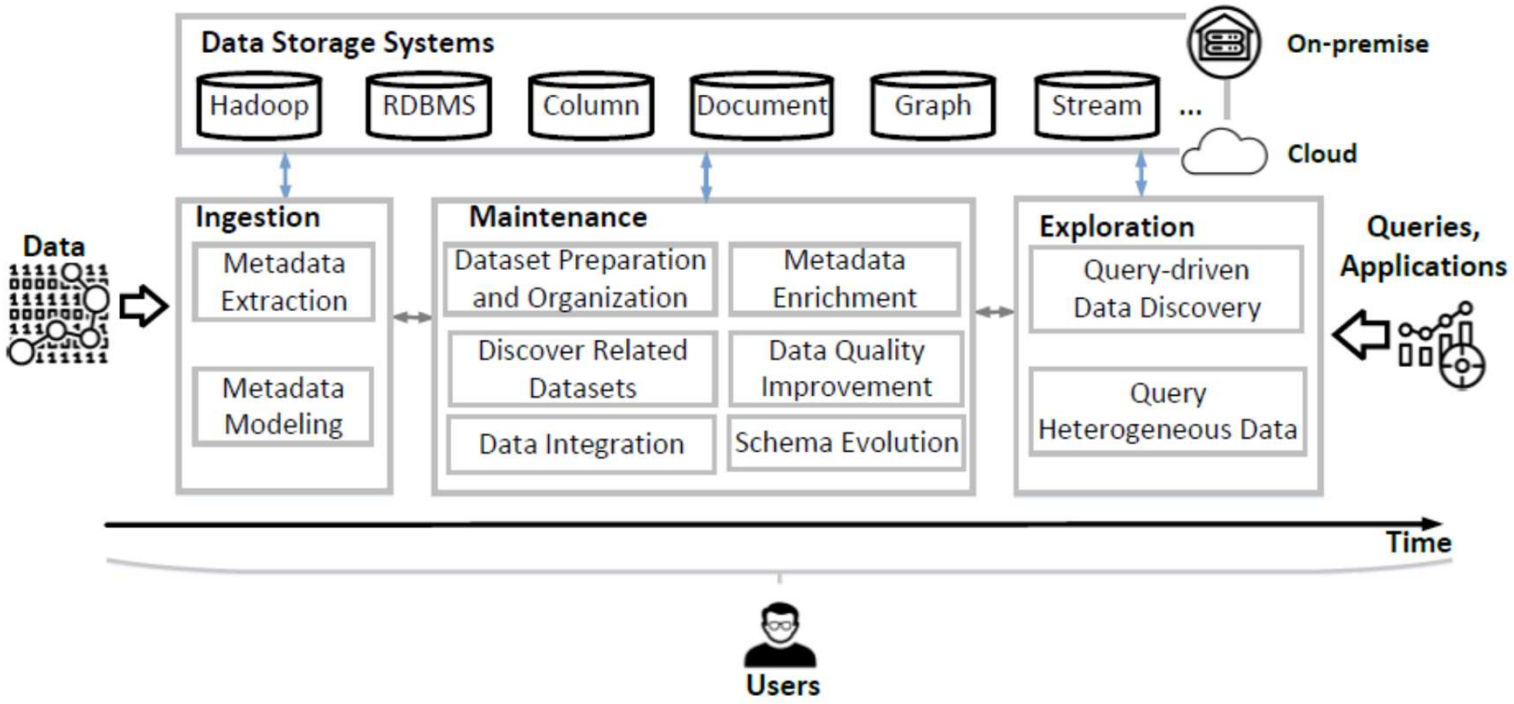
Materialized versus Virtual Integration

	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)
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.

Components of Data Lake Management Systems

Data Lakes provide for pay-as-you-go data integration



Criteria	Data Warehouses	Data Lakes
Data ingestion	ETL	Load-as-is
Ingested data format	Structured	Heterogeneous (structured, semi-structured, and unstructured)
Data storage	Relational databases	Hadoop, Relational databases, NoSQL data stores, etc
Data access	SQL queries (OLTP, OLAP)	Different query languages (e.g., SQL, Cypher), programming languages (e.g., Java, Python, R)

Hai, Rihan, Christoph Quix, and Matthias Jarke: Data Lake Concept and Systems: A Survey. arXiv:2106.09592, 2021.

7. The Data Integration Software Market

- Market size 2017:
  7.45 billion US$ (growth: 14.4%)
- Tools for specific tasks
  - Altova Map Force for schema mapping
- Comprehensive solutions covering the complete data integration process
  - Informatica Plattform
  - IBM InfoSphere Information Server
  - SAP Data Hub, SAP Vora
  - Microsoft SQL Server Integration Services
  - Talend Data Integration
- Cloud-based data lake solutions
  - Amazon AWS Glue, Databricks Lakehouse, Microsoft Fabric

Figure 1: Magic Quadrant for Data Integration Tools

Source: Zaidi, et al.: Gartner Report - Magic Quadrant for Data Integration Tools. December 2023.

Getting an Impression of the Tools

Video tutorials on YouTube

- **Informatica PowerCenter**
  https://www.youtube.com/watch?v=u6oLXidGoqs
- **SAP Data Hub**
  https://www.youtube.com/watch?v=CjLc4eDNpso
- **Amazon AWS Glue**
  https://www.youtube.com/watch?v=jwGGd-kUaLo
- **Microsoft Fabric**
  https://www.youtube.com/watch?v=-f0XIVEP7bE
- **Tamr Unify**
  https://www.youtube.com/watch?v=7jz740cdtDE

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 the Web
- Architectures range from warehousing over virtual integration to data lakes
- Regardless of the architecture, bridging heterogeneity is the key issue
- Goal: Reduce the human effort involved
- Current trend: AI / human co-work