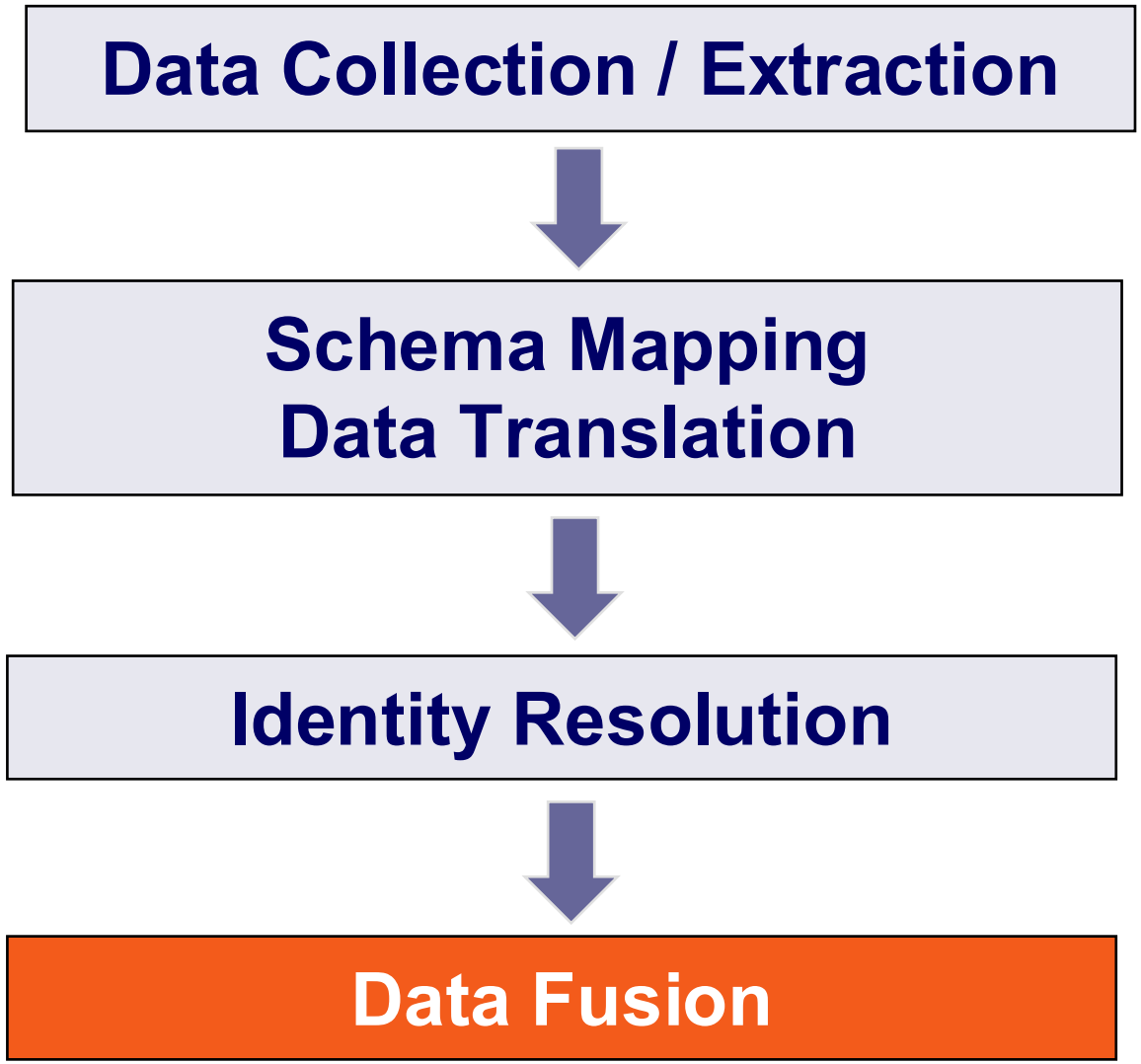
Web Data Integration

Data Fusion

The Data Integration Process



Final Exam HWS2026 (IE670, 3 ECTS)

- Date and Time
  - Thursday, 17.12.2026, time and room will be announced
- Format
  - 6 open questions that show that you have understood the content of the lecture, 5 points per question, 60 minutes time
  - All lecture slides are relevant, including
    - pro and cons of web data publication mechanisms
    - XML syntax and DTDs, XPath or SPARQL query (one question)
    - data discovery methods, FAIR data principles, profiling, data quality
    - strength and weaknesses of schema matching methods + data samples
    - blocking, entity matching, learning entity matching models
    - strength and weaknesses of different similarity metrics
    - data fusion, conflict resolution methods, evaluation measures
  - We want precise answers, not all you know about the topic.
  - Three example questions and answers are provided on the course webpage.

Outline

1. Introduction to Data Fusion
2. Slot Filling and Conflict Resolution
3. Conflict Resolution Functions
4. Evaluation of Fusion Results
5. Case Studies

1. Introduction

Information providers on the Web have

- different levels of knowledge
- different views of the world
- different intentions

Therefore,

1. information on the Web is partly wrong, biased, outdated, incomplete, and inconsistent.
2. every piece of information on the Web needs to be considered as a claim by somebody at some point in time and not as a fact.
3. the information consumer needs to make up her mind which claims to use for a certain task

Definition: Data Conflict

Multiple records that describe the same real-world entity provide different values for the same attribute.

Chris Miller	12/20/1982	Bardon Street, Melville
Christian Miller	2/20/1982	7 Bardon St., Melwille

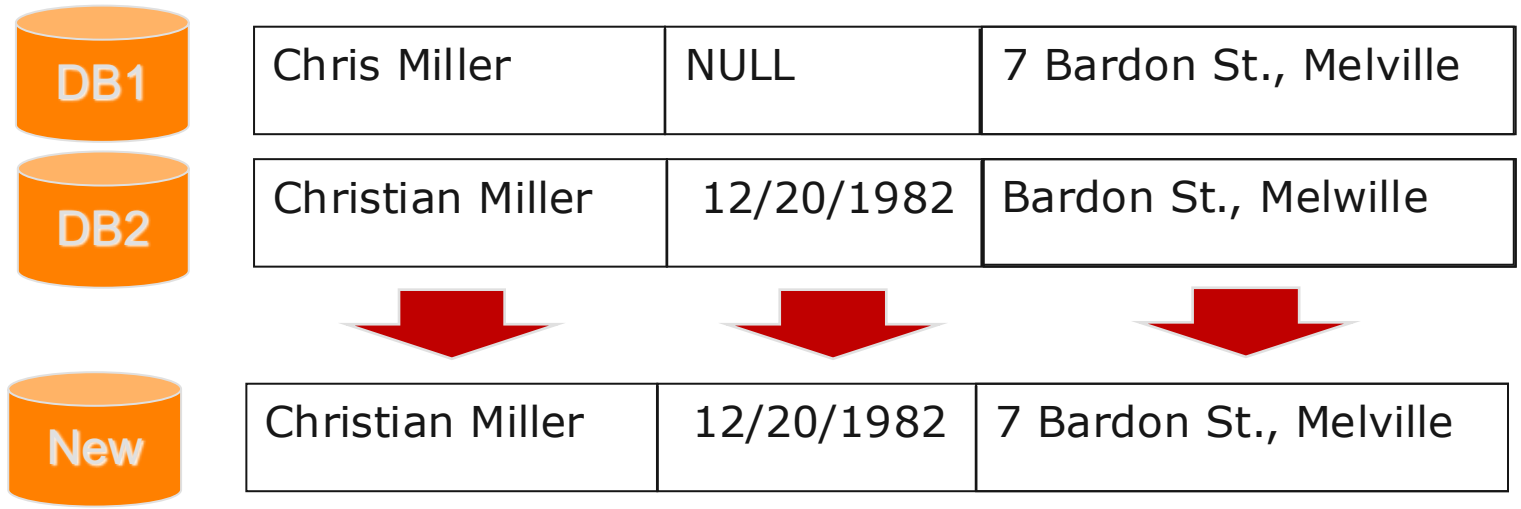
Correspondence

Reasons for data conflicts:

1. Data creation: Typos, measurement errors, erroneous information extraction
2. Data currency: Different points in time, missing updates
3. Data semantics: Different definitions of concepts (like population or GDP)
4. Data representation: Different coding of values (“Mrs.” vs. “2”)
5. Data integration: Wrong data translation or identity resolution
6. Actual disagreement of data providers: Subjective attributes

5. Data Fusion

Given multiple records that describe the same real-world entity, create a single record while resolving conflicting data values.



- Goal: Create a single high-quality record.
- Two basic fusion situations: Slot Filling and Conflict Resolution

5.1 Slot Filling and Conflict Resolution

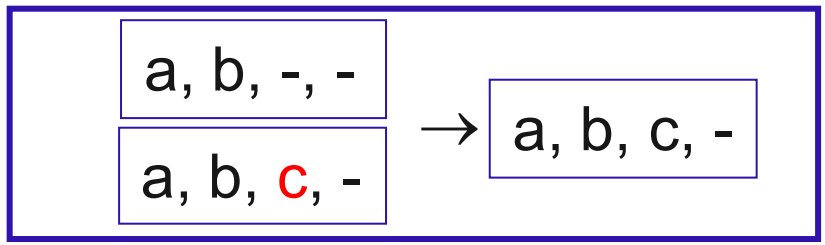
Slot Filling: Fill missing values (NULLs) in one dataset with corresponding values from other datasets.

Result: increased dataset density

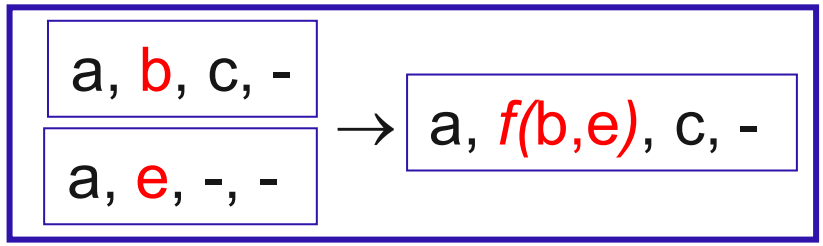
Conflict Resolution: Resolve contradictions between records by applying a conflict resolution function (heuristic).

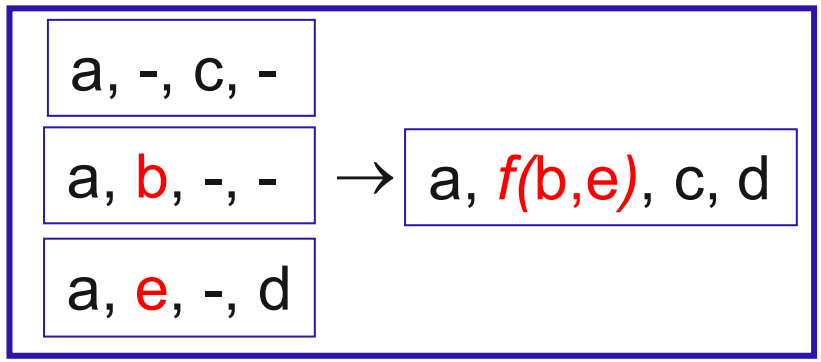
Result: increased data quality

Complementary records

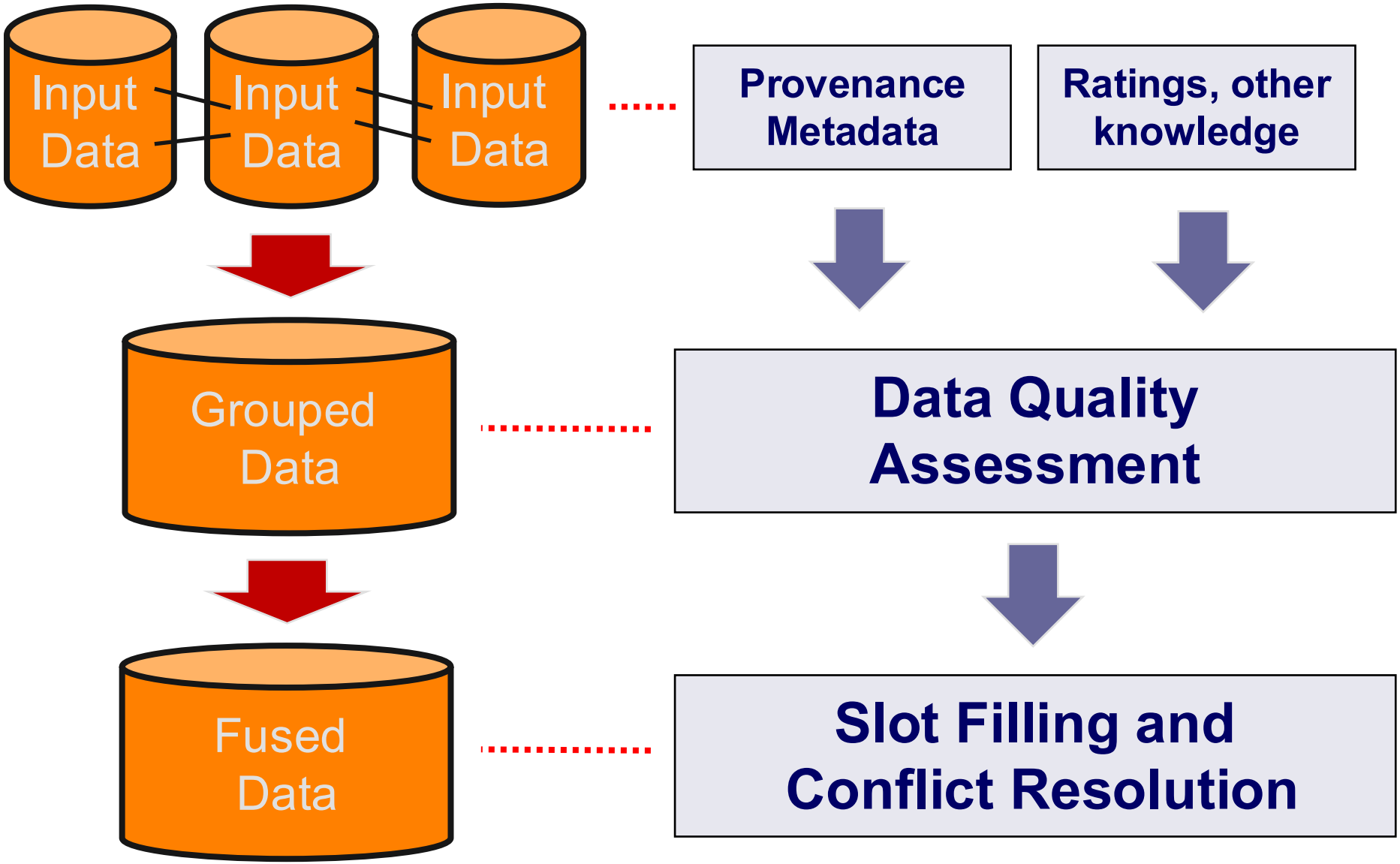


Conflicting records





Summary: Elements of the Data Fusion Process



Cluster Size Distribution, Matching Errors, and Data Fusion

– Records are clustered using the correspondences that were discovered during identity resolution. Example with **3** data sources:

Cluster Size	Frequency
1	4256
2	939
3	503
4	75
5	14
35	3
61	1

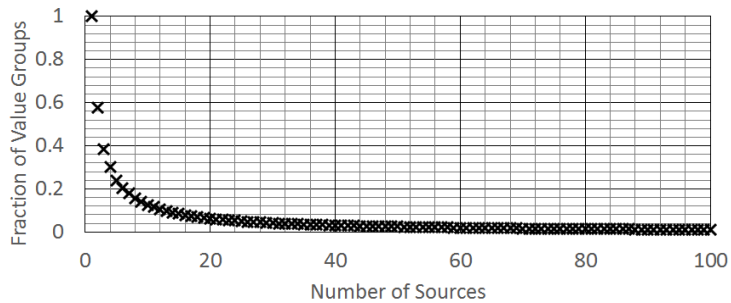
No slot filling possible as single records with no overlap

Slot filling and conflict resolution allow the generation of improved records

Large cluster size indicates matching errors or duplicates in data sources

– Cluster size distribution from matching web tables to DBpedia

- Out of 33.3 million web tables, 949,970 tables contain at least one matching row
- 42% of the clusters have a size of 2
- 16% of the clusters have a size of 3
- 39% of the clusters have a size of at least 4
- 13% of the clusters have a size of at least 11



Ritze, et al.: Profiling the Potential of Web Tables for Augmenting Cross-domain Knowledge Bases. WWW 2016.

5.2 Conflict Resolution Functions

- Conflict resolution functions are attribute-specific
  - you select or learn a specific function for each attribute that should be fused

$$V_F = f(V_A, M_A, B)$$

Fused Value

Input Values

Meta-Information

Background Knowledge

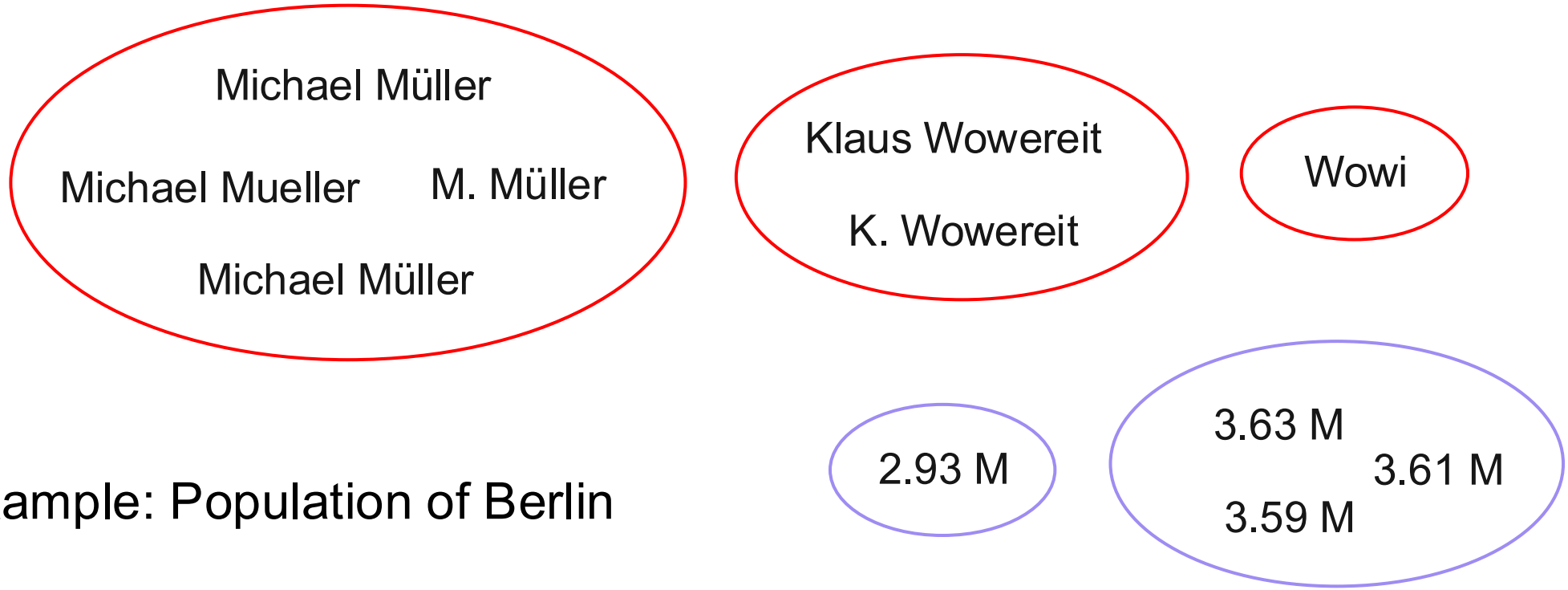
- There is a wide range of different functions (heuristics) that fit different requirements
- Functions differ regarding the data types, they can be applied for
  - numerical values (e.g. population of a place)
  - nominal values / text (e.g. name of a person)
  - sets of values (e.g. actors performing in a movie)
- Two main categories of conflict resolution functions
  1. Content-based functions that rely only on the data values to be fused
  2. Metadata-based functions that rely on provenance data, ratings, or quality scores

Content-based Conflict Resolution Functions

Function	Explanation	Use Case
Average, Median	Calculate average/median of all values	Population, ratings
Longest, Shortest	Choose longest / shortest value	First name
Max, Min	Take maximal, minimal value	Number of children
Vote	Majority decision (one vote per site or page?)	Mayor of city
Clustered Vote	Choose centroid / medoid of largest cluster	Population of city
Weighted Vote	Weight sources, for instance according to the fraction of true values they provided	Address of a shop
Union	Union of all values (A ∪ B ∪ C)	Product Reviews
Intersection	Intersection of all values (A ∩ B ∩ C)	Movie Actors
IntersectionKSources	Values must appear in at least k sources	Movie Actors
MostComplete	Choose value from record that is most complete	Postal addresses
MostAbstract, MostSpecific	Use a taxonomy / ontology	Location
Random	Fallback: Choose random value	

Clustered Vote: Which Values are Similar Enough?

- Approach:
  1. Calculate similarity of values
     - using an appropriate similarity function (see slide set Identity Resolution)
  2. Treat similar values as equal (equal if similarity above threshold)
- Example: Mayor of Berlin



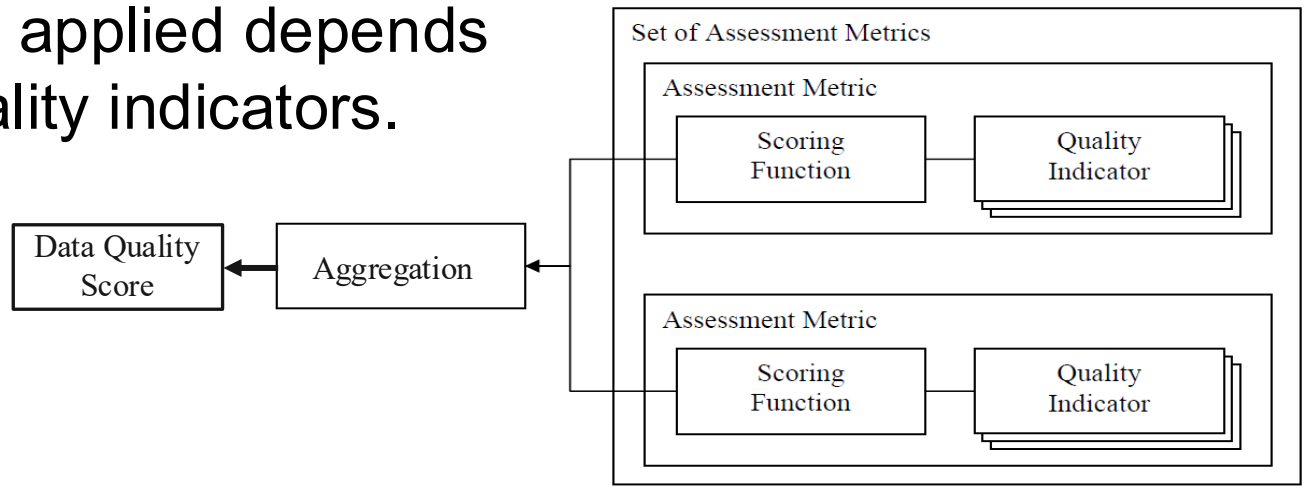
Example: Population of Berlin

Metadata-based Conflict Resolution Functions

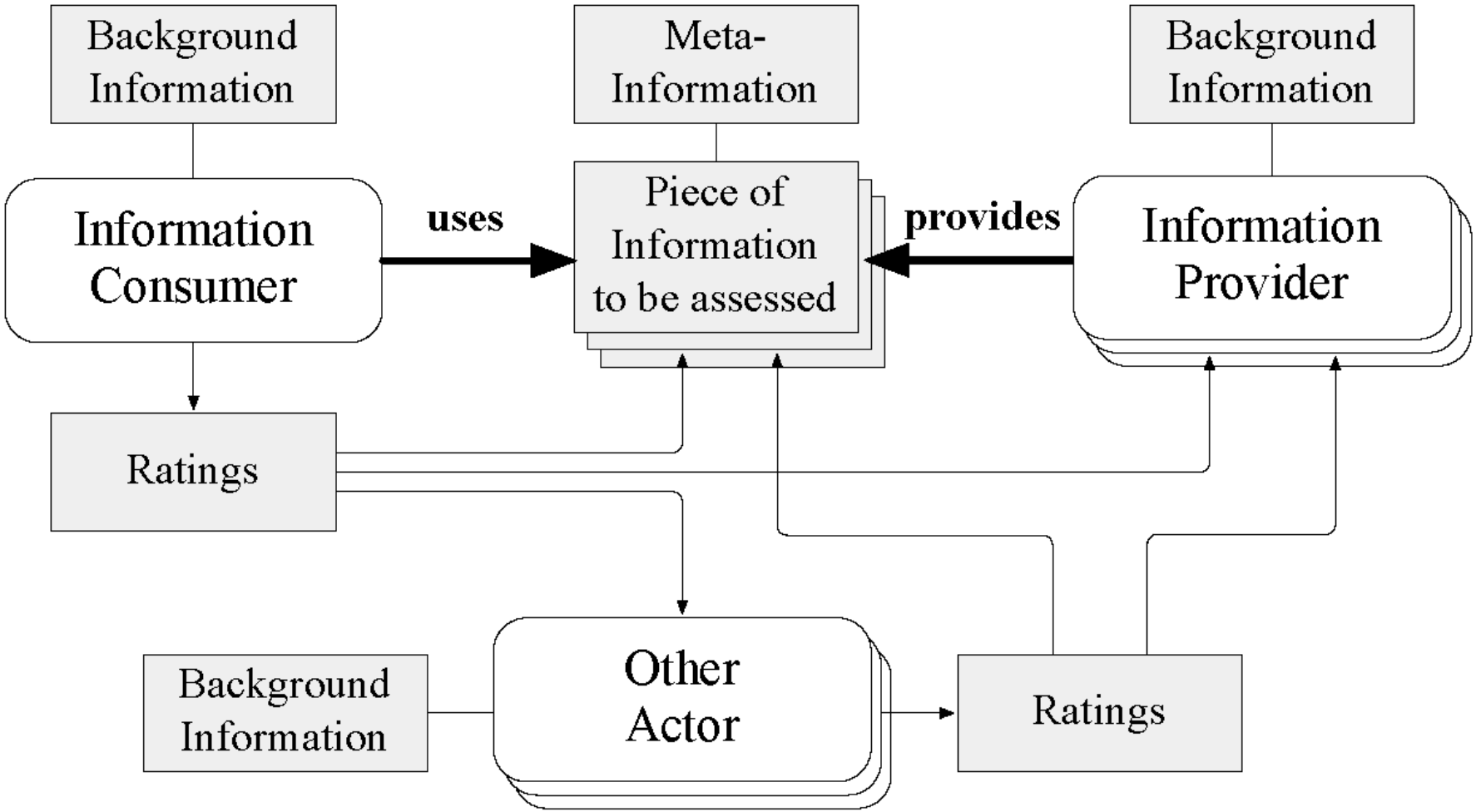
Function	Explanation
FavorSources	Take first non-null value in particular order of sources Example: Use Eurostat for GDP, alternatively use Wikipedia
MostRecent	Choose most recent (up-to-date) value Example: Address, phone number
MostActive	Choose value that is most often accessed/edited Example: Prefer Wikipedia page with more edits
FavorSources basedOnRatings	Calculate quality of sources from ratings, take value from source with highest score or all values from sources with scores above specific threshold
MaxIQ	Choose the value with the highest quality score. Score might cover multiple quality dimensions, e.g. timeliness and believability of a source
ClusteredVoteAfter Filtering	Filter values using quality scores and apply clustered vote afterwards
IntersectionAfterFiltering	Values of values from sources having a high quality score
…	..

How to Determine Data Quality Scores?

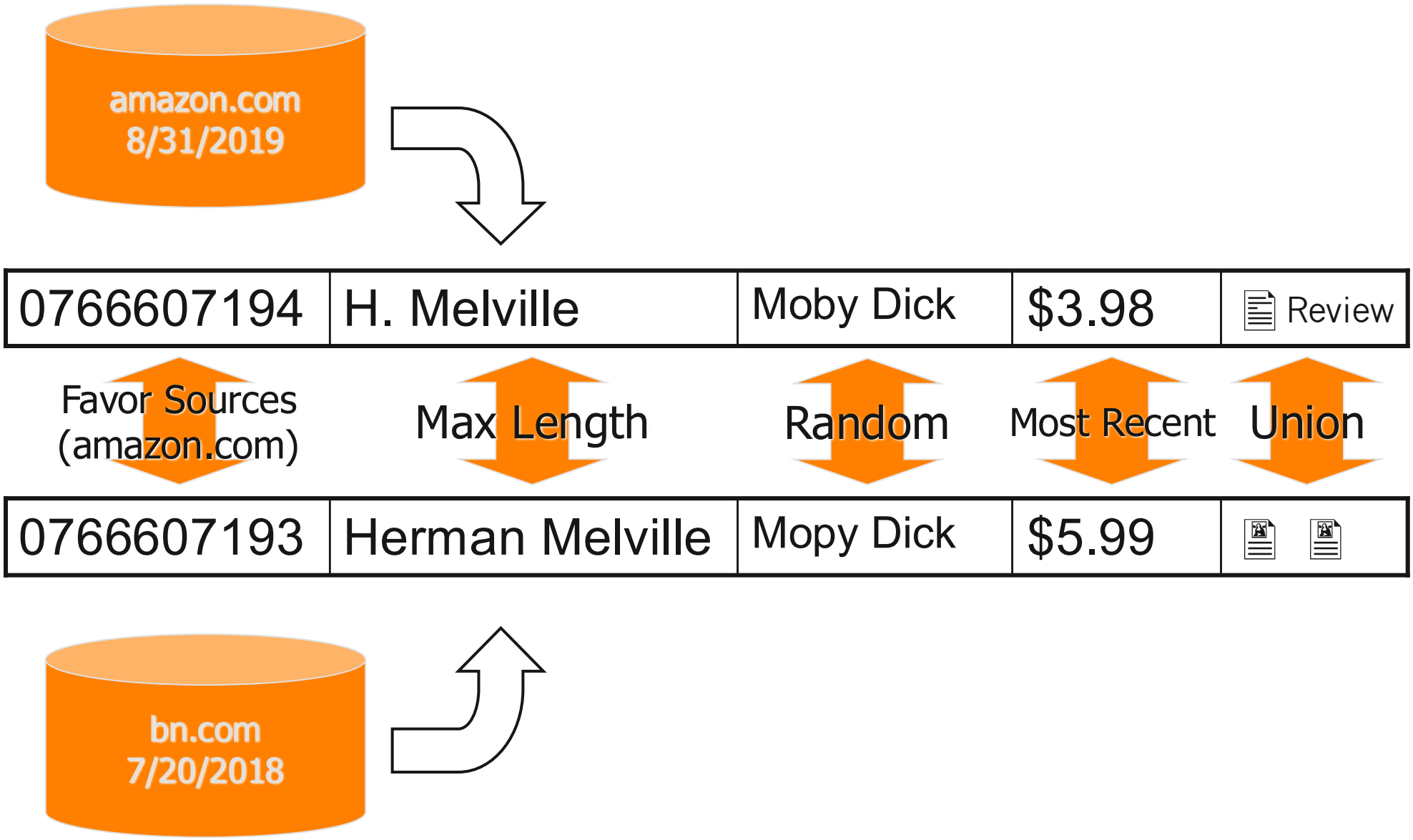
- Data quality is a multi-dimensional construct which measures the "fitness for use" of data for a specific task.
  - accuracy, timeliness, completeness, format consistency
- Which quality dimensions matter depends on the task at hand.
- A wide range of heuristics can be used to assess data quality
  - constraint testing, outlier detection, rating-based scoring, PageRank
  - see slide set on Data Discovery for details
- Which heuristics can be applied depends on the availability of quality indicators.



Quality Indicators in the Web Context



Example: Complete Data Fusion Strategy



5.3 Evaluation of Fusion Results

1. Data Centric Evaluation Measures
   - Density
   - Consistency
2. Ground Truth Based Evaluation Measures
   - Accuracy

Density

Density measures the fraction of non-NULL values.

$$density_{Column} = \frac{|non{-}NULL\ values\ in\ column|}{|rows\ in\ table|}$$

$$density_{Table} = \frac{|non{-}NULL\ values\ in\ table|}{|columns|*|rows|}$$

- As a result of schema integration, translated tables often contain many null values (empty columns)
- We are interested in the density increase after fusion
  1. Measure density of table A or column C_1
  2. Fuse table A with table B
  3. Measure density of resulting table A' or column C_1'

Consistency

A data set is consistent if it is free of conflicting information.

$$consistency_{Column} = \frac{|non\text{-}conflicting\ values\ in\ column|}{|real\text{-}world\ entities\ described|}$$

$$consistency_{Table} = \frac{|non\text{-}conflicting\ values\ in\ table|}{|columns| * |real\text{-}world\ entities\ described|}$$

Measurement:

1. Group records that refer to same real-world entity
   - using correspondences discovered by identity resolution
2. Calculate fraction of non-conflicting attribute values
   - same attribute value is provided by all data sources

Accuracy

Accuracy: Fraction of correct values selected by the conflict resolution function.

$$accuracy = \frac{|correct\ values|}{|all\ values|}$$

$$error\ rate = 1 - accuracy$$

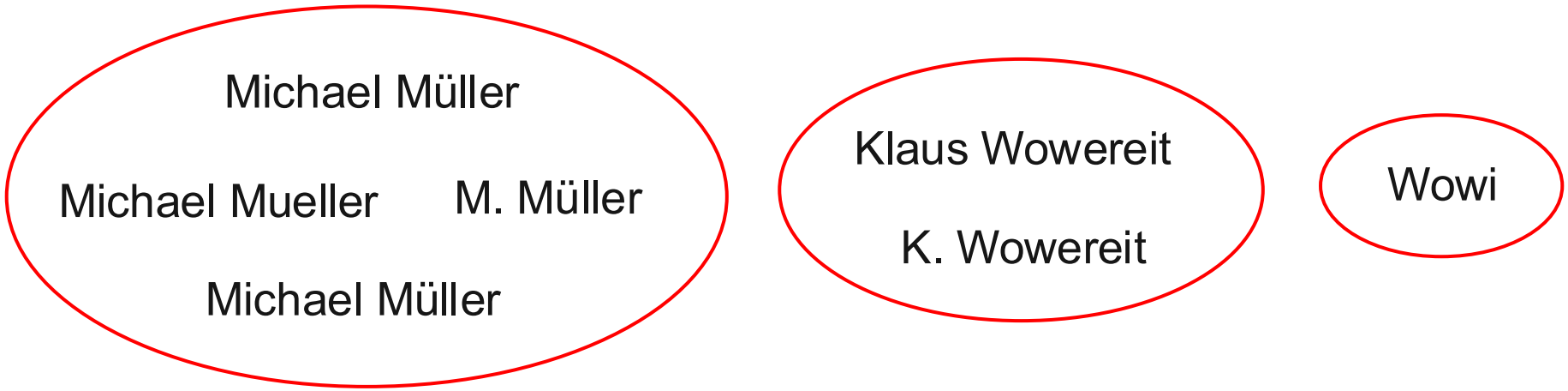
Assessment:

1. Gather Ground Truth
   - Manually determine correct values for a subset of the records
   - Alternative: Use/buy correct data from external provider
   - Can be tricky as this requires you or external provider to know the truth!
2. Compare values after fusion with true values

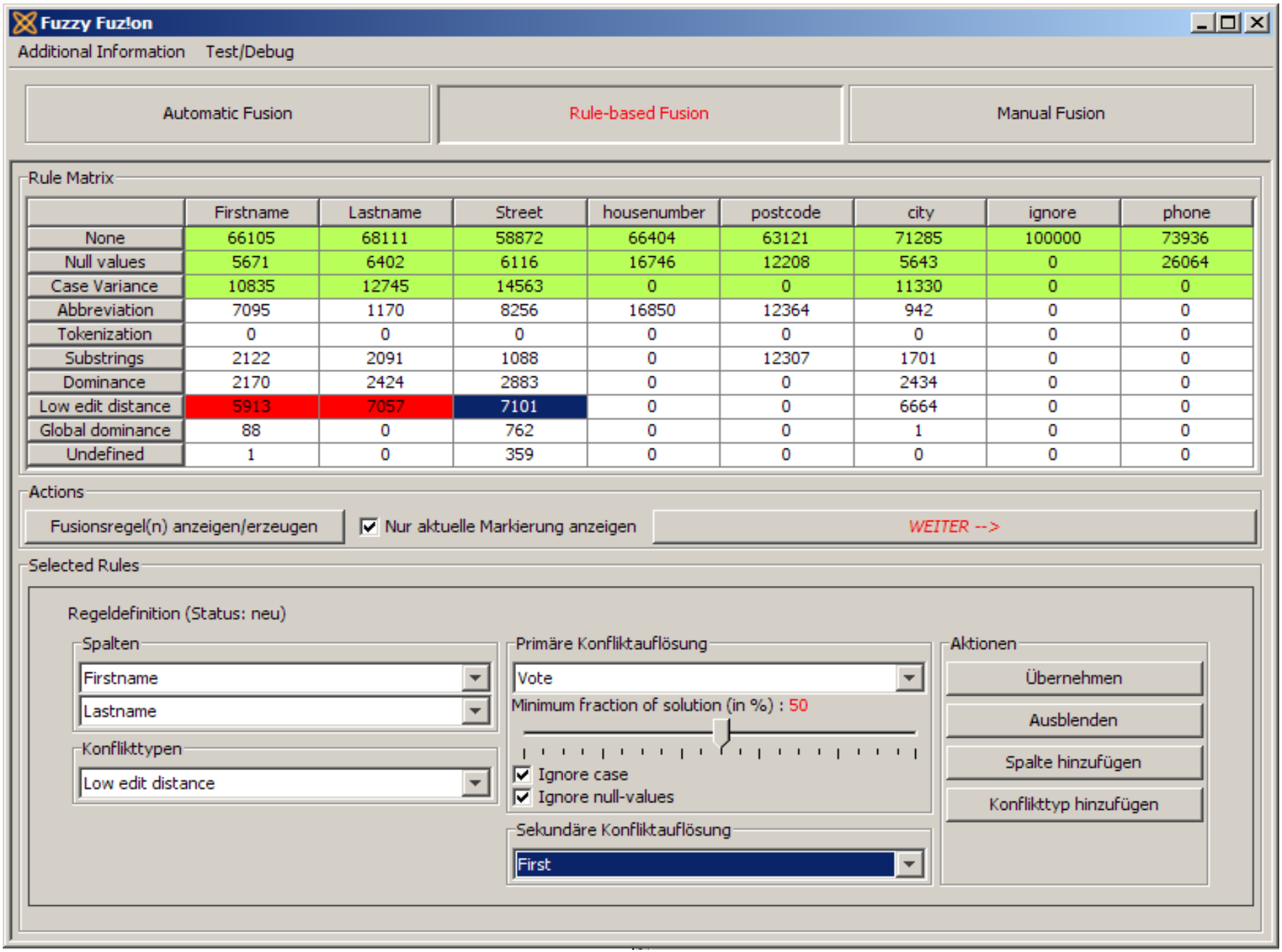
Gao, et al.: Efficient Knowledge Graph Accuracy Evaluation. VLDB Endowment, 2019.

How to Treat Similar Values?

- Treatment of similar values matters for calculating consistency and accuracy.
- Approach:
  1. Calculate similarity of values
     - using an appropriate similarity function (see slide set Identity Resolution)
  2. Treat similar values as equal (equal if similarity above threshold)
- Example: Mayor of Berlin



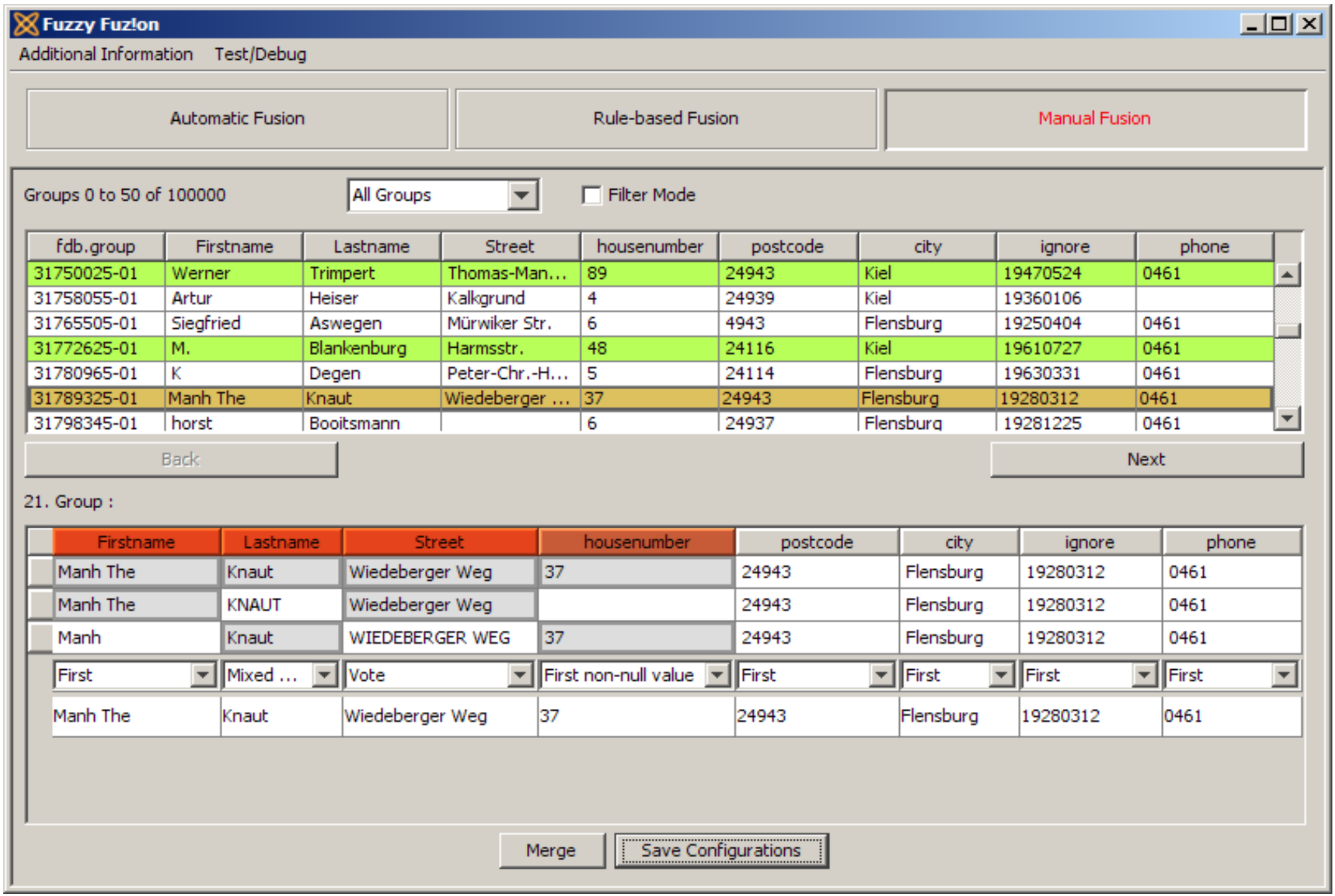
5. Case Study: Data Fusion Tool: Fuz!on



	Firstname	Lastname	Street	housenumber	postcode	city	ignore	phone
None	66105	68111	58872	66404	63121	71285	100000	73936
Null values	5671	6402	6116	16746	12208	5643	0	26064
Case Variance	10835	12745	14563	0	0	11330	0	0
Abbreviation	7095	1170	8256	16850	12364	942	0	0
Tokenization	0	0	0	0	0	0	0	0
Substrings	2122	2091	1088	0	12307	1701	0	0
Dominance	2170	2424	2883	0	0	2434	0	0
Low edit distance	5913	7057	7101	0	0	6664	0	0
Global dominance	88	0	762	0	0	1	0	0
Undefined	1	0	359	0	0	0	0	0

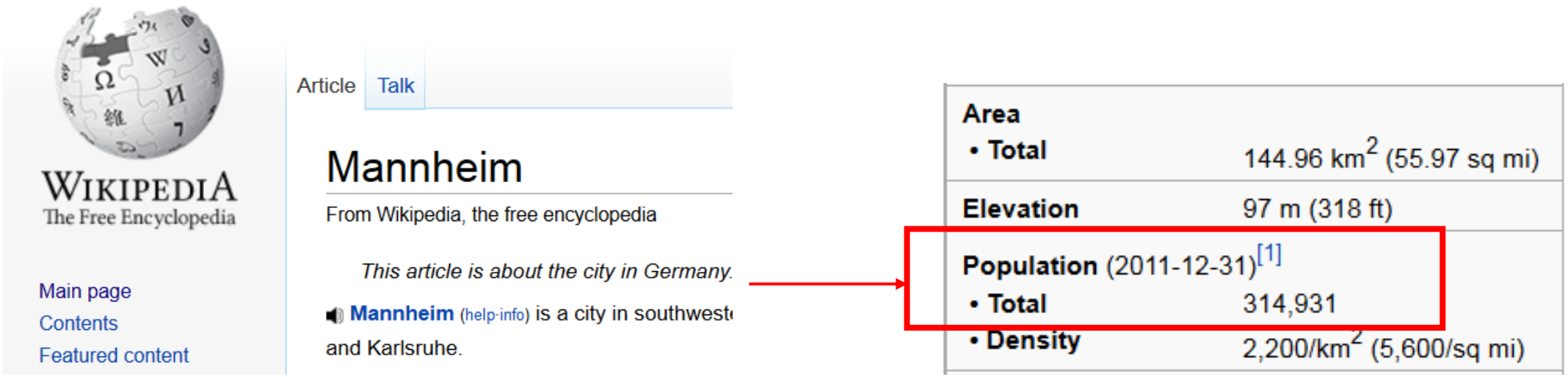
Prototype developed at Hasso Plattner Institute

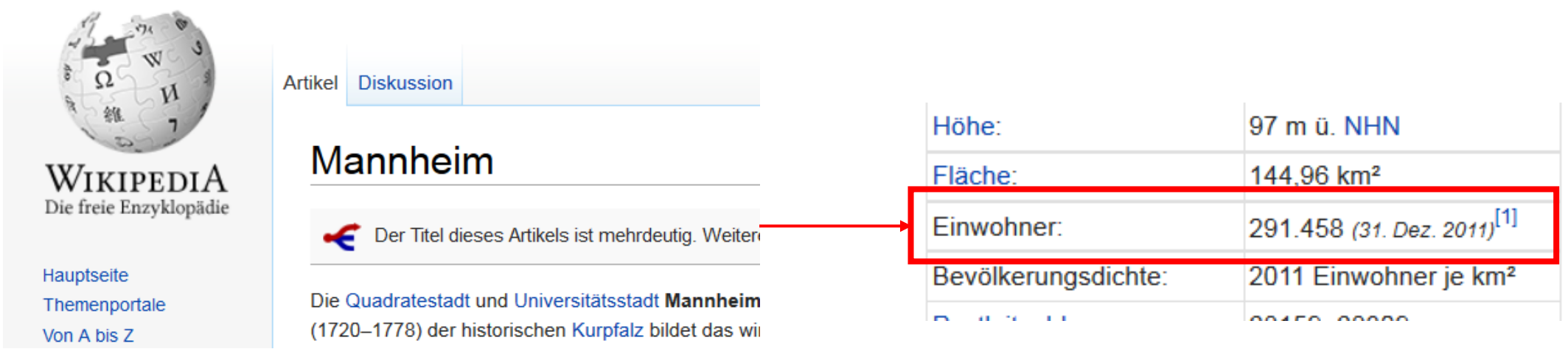
Manual Fusion of Record Groups in Fuz!on



Case Study: DBpedia Cross-Language Data Fusion

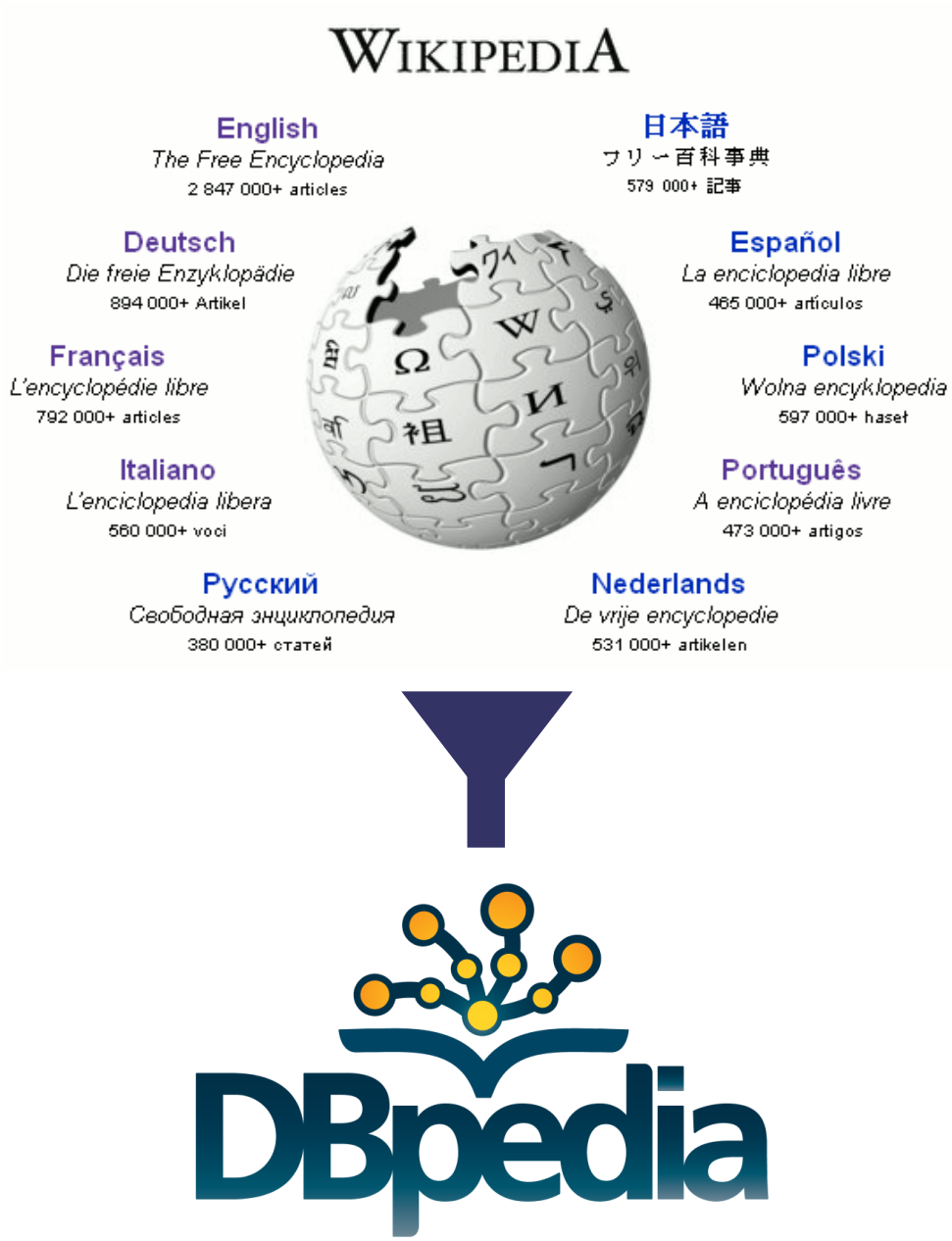
- Infoboxes in different Wikipedia editions contain conflicting values.
- Which value to prefer?





Cross-Lingual Data in DBpedia

- DBpedia extracts structured data from Wikipedia in **119 languages**.
- DBpedia contains **lots of data conflicts**, inherited from Wikipedia.
- **Identity resolution is solved** by Wikipedia inter-language links.
- **Schema heterogeneity problem is solved** by community-created mappings from infoboxes to DBpedia ontology.



Goal: Fuse Data between different Language Editions

Which value to prefer?

- maximum?
- average?
- most frequent?

data itself

- from the specific language edition?
- most recent?
- inserted by most trusted author?
- edited most times?

prove nance

- combination of the above?

Population of Mannheim in 8 DBpedia language editions

```
Mannheim populationTotal
                "314,931"@en
                "291,458"@de
                "311,969"@eu
                "311,342"@fr
                "308,676"@nl
                "309,795"@pt
                "313,174"@ru
                "310,000"@sl
```

Provenance Metadata from the Wikipedia Revision Dumps

- We extract provenance metadata from the Wikipedia revision dumps of the Top10 languages
  - File size of revision dumps: > 6 TByte for English, >2 TByte for German
- Extracted metadata
  - Last edit timestamp of a fact
  - Number of edits of a fact
  - Author of the last edit
    - Author edit count
    - Author registration date

Provenance metadata

```
ru:Mannheim:populationTotal

        lastedit        2011-12-22T00:50:21Z
        propeditcnt     3
        autheditcnt     1136639
        authregdate     2009-12-18T02:08:09Z

nl:Mannheim:populationTotal

        lastedit        2007-12-09T16:41:06Z
        propeditcnt     1
        autheditcnt     73
        authregdate     2007-04-05T08:54:19Z
```

Learning Conflict Resolution Functions

- Ground Truth: Geonames, public geographical database
- Learning: Choose function with smallest mean absolute error with respect to validation subset of the ground truth.
- Tested conflict resolution functions
  1. *Maximum*
  2. *Average*
  3. *English* – prefer values from English DBpedia
  4. *Vote* – choose the most frequent value
  5. *MostRecent* fact – last edit timestamp
  6. *MostActive* fact – number of edits of a property
  7. *MostActive* author – author edit count
  8. *MostSenior* author – author registration date

DBpedia Case Study: Results

Property	Dataset	Count	Learned Fusion Function	Error, %	Error, %, en.dbpedia
populationTotal	cities1000-Germany *	7330	Vote (most frequent value)	0.3029	0.6796
populationTotal	cities1000-Netherlands	493	Maximum Value	2.1933	3.5714
populationTotal	countries	243	Maximum Value	2.1646	6.3485
country	cities1000-Italy	1078	Vote	0.0000	1.2060
country	cities1000-Brazil	1119	Max author edit count	9.8302	30.9205
country	cities1000-Germany	7638	Vote	0.0131	0.6415

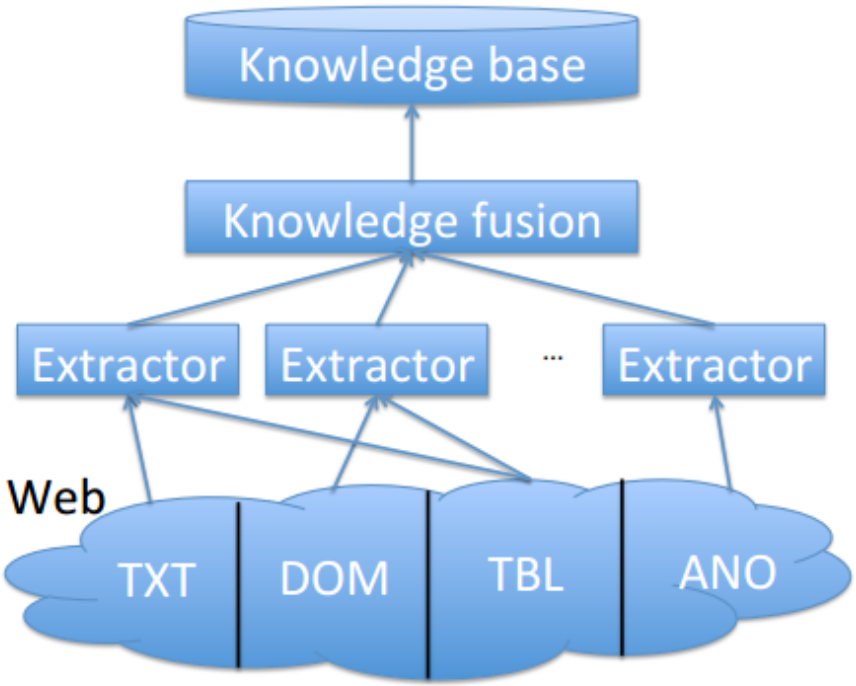
** "cities1000" are cities with population >1000*

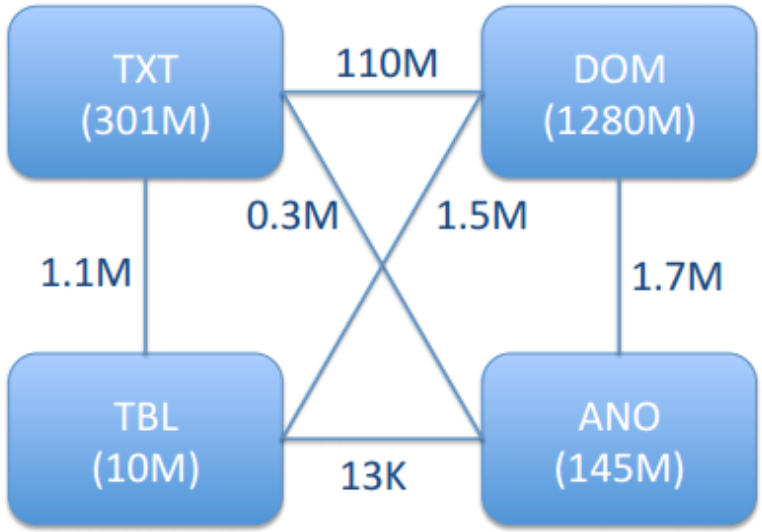
- **Error:** Mean absolute percentage error between chosen value and ground truth
- **Error en.dbpedia:** Mean absolute percentage error between value in English DBpedia and gold standard

Volha Bryl, Christian Bizer: Learning Conflict Resolution Strategies for Cross-Language Wikipedia Data Fusion. 4th Workshop on Web Quality @ WWW 2014.

Case Study: Google Knowledge Vault

- uses 12 different extractors to extract 6.4 billion triples (1.6 billion unique triples) from 1 billion page Web crawl
- extracted data is fused to extend the Freebase knowledge base

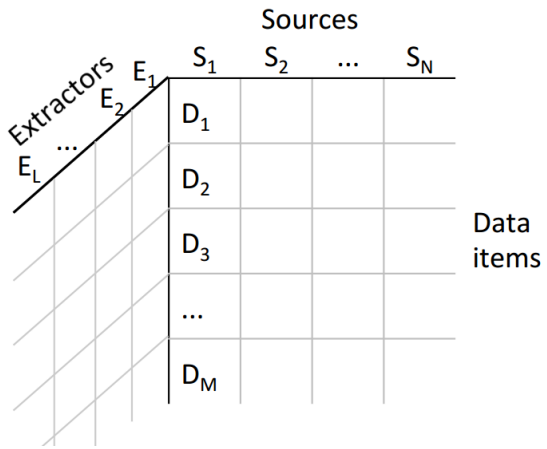




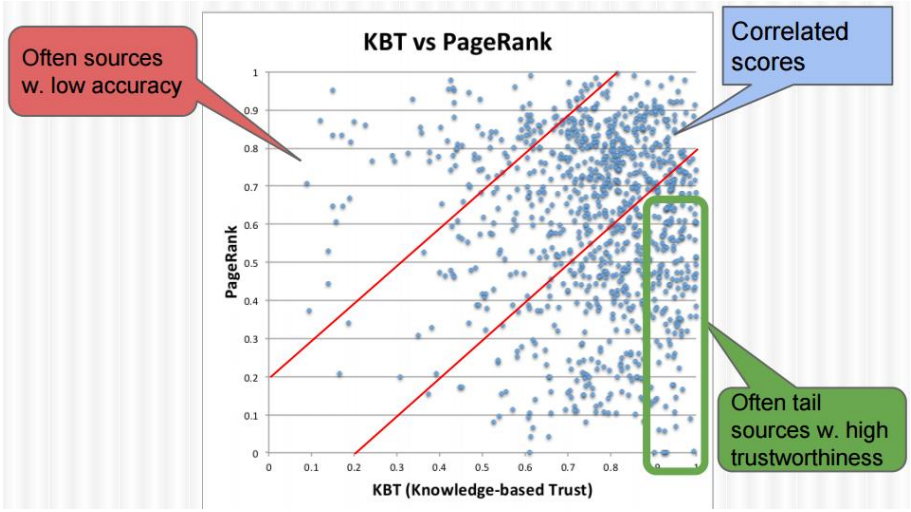
Luna Dong, et al.: From Data Fusion to Knowledge Fusion. VLDB 2014.

Google Knowledge Vault

- uses probabilistic model to iteratively determine quality of triples, sources, and extractors
- result: 90 million triples with p>0.9 that were not in Freebase before



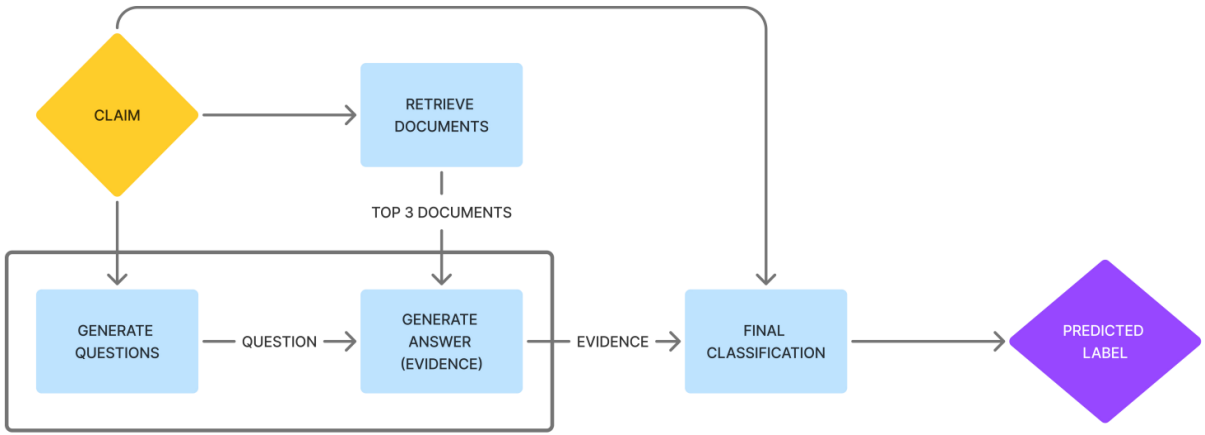
- Knowledge-based Trust
  - determine trustworthiness of a data source by comparing its content with a knowledge base (ground truth)
  - result: better than PageRank in identifying
    - tail websites with high trustworthiness
    - gossip websites



Luna Dong, et al.: Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. SIGKDD 2014.
Luna Dong, et al.: Knowledge-based Trust: Estimating the Trustworthiness of Web Sources. VLDB 2015.

Excurse: RAG-based Fact Checking

- Fact checking: Verification of the factual accuracy of claims (often by politicians)
  - FactCheck.org, *Washington Post* publish claim reviews on the Web
- Current Trend: Automated fact checking using RAG Pipelines



Claim: *The USA has succeeded in reducing greenhouse emissions in previous years.*
Date: 2020.11.2 ***Speaker:*** Morgan Griffith

Q1: What were the total gross U.S. greenhouse gas emissions in 2007?
A1: In 2007, total gross U.S. greenhouse gas emissions were 7,371 MMT.

Q2: When did greenhouse gas emissions drop in US?
A2: In 2017, total gross U.S. greenhouse gas emissions were 6,472.3 MMT, or million metric tons, carbon dioxide.

Q3: Did the total gross U.S. greenhouse gas emissions rise after 2017?
A3: Yes. After 3 years of decline, US CO2 emissions rose sharply last year. Based on preliminary power generation, natural gas, and oil consumption data, we estimate emissions increased by 3.4% in 2018.

Verdict: Conflicting Evidence/Cherrypicking.

Dmonte, et al.: Claim Verification in the Age of Large Language Models: A Survey. Arxiv, 2025.

AVeriTeC Shared Task 2024

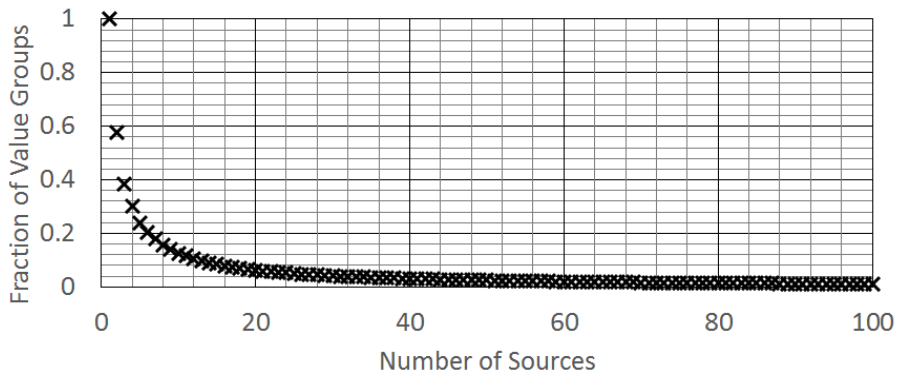
- Testset 1200 claim reviews from 50 fact-checking organizations
  - collected from the Web using schema.org ClaimReview annotations
  - possible labels: supported, refuted, conflicting evidence, not enough evidence
- Success Rate of best systems 64% (correct label and right evidence)

Team Name	Evidence	QG	Retrieval	QA	Veracity
TUDA_MAI	KS	GPT-4o	gte_base_en_v1.5	GPT-4o	GPT-4o
HUMANE	KS	Llama-3-8b	BM25 SFR-embedding-2 Llama-3.1-70b	-	Llama-3.1-70b
CTU AIC	KS	GPT-4o	mxbai-large-v1	GPT-4o	GPT-4o
Dunamu-ML	KS	GPT-4	BM25	GPT-4	GPT-4
Papelo	Google	T5-large GPT-4o	-	GPT-4o	GPT-4o

Schlichtkrull, et al.: The Automated Verification of Textual Claims (AVeriTeC) Shared Task. ACL 2024.
Ge, et al.: Resolving Conflicting Evidence in Automated Fact-Checking: A Study on Retrieval-Augmented LLMs. ArXiv:2505.17762, 2025

Summary: Data Fusion

- Data Fusion addresses missing values (slot filling) as well as contradictions (conflict resolution)
- Appropriate conflict resolution function depends on
  - data type of the values
  - availability of quality-related metadata
  - availability of overlapping data (e.g. for voting)
- On the Web, we often encounter long-tailed distributions
  - lots of overlapping data for head entities (New York)
  - hardly any data to fuse for tail entitles (some village)
  - example: Web tables matched to DBpedia



6. References

- Data Fusion
  - Bleiholder, Naumann: Data Fusion. ACM Computing Surveys, 2008.
  - Li, Gao, Meng, et al.: Survey on Truth Discovery. SIGKDD Explorations, 2016.
  - Dong, Srivastava: Big Data Integration. Chapter 4. Morgan & Claypool, 2015.
  - Rekatsinas: Tutorial Data Integration and Machine Learning. SIGMOD 2018, Chapter ML for DF.
  - Dong & Naumann: Data Fusion. Tutorial at VLDB 2009.
  - Aggarwal: Managing and Mining Uncertain Data. Springer, 2010.
- Data Fusion Evaluation Datasets
  - Dong: Data Sets for Data Fusion Experiments http://lunadong.com/fusionDataSets.htm
- PyDI Wiki Page on Data Fusion
  - https://github.com/wbsg-uni-mannheim/PyDI/blob/main/docs/wiki/DataFusion.md