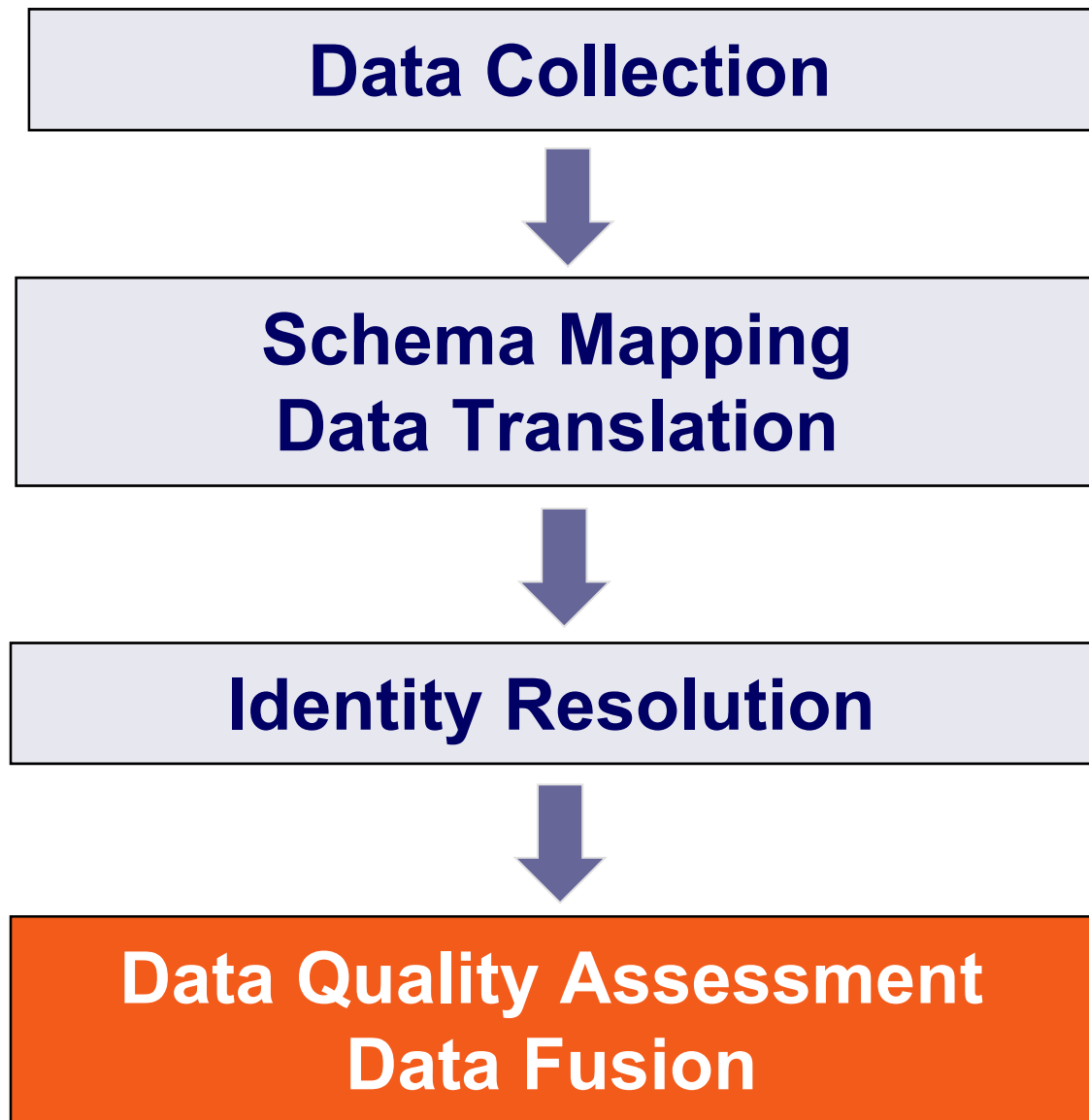


Web Data Integration

Data Quality Assessment and Data Fusion



The Data Integration Process



Outline

1. Introduction
2. Data Provenance
3. Data Quality Assessment
4. Data Fusion
5. References

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 information on the Web is partly wrong, biased, outdated, incomplete, and inconsistent.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**, not as a fact.
3. the information consumer needs to make up her mind which claims to use for a certain task.



Example: Area and Population of Monaco

Area: Different claims and different conversions

en.wikipedia.org	2.02 sq km	0.78 sq miles
www.state.gov	1.95 sq km	0.8 sq miles
www.atlapedia.com	1.94 sq km	1 sq mile

(1.95 sq km = 0.753 sq miles)



Population in 2004: Different claims and sparse meta-information

Year	Value	Meta-information	Webpage
2004	32,270	(July 2004 est.)	http://www.cia.gov/cia/publications/factbook/geos/mn.html
2003	32,130	(est.)	http://www.greenfacts.org/studies/climate_change/index.htm
2003	30,000		http://www.tlfq.ulaval.ca/axl/europe/monaco.htm
2000	31,842		http://en.wikipedia.org/wiki/Monaco

Source: Peter Bunemann

Definition: Data Conflict

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

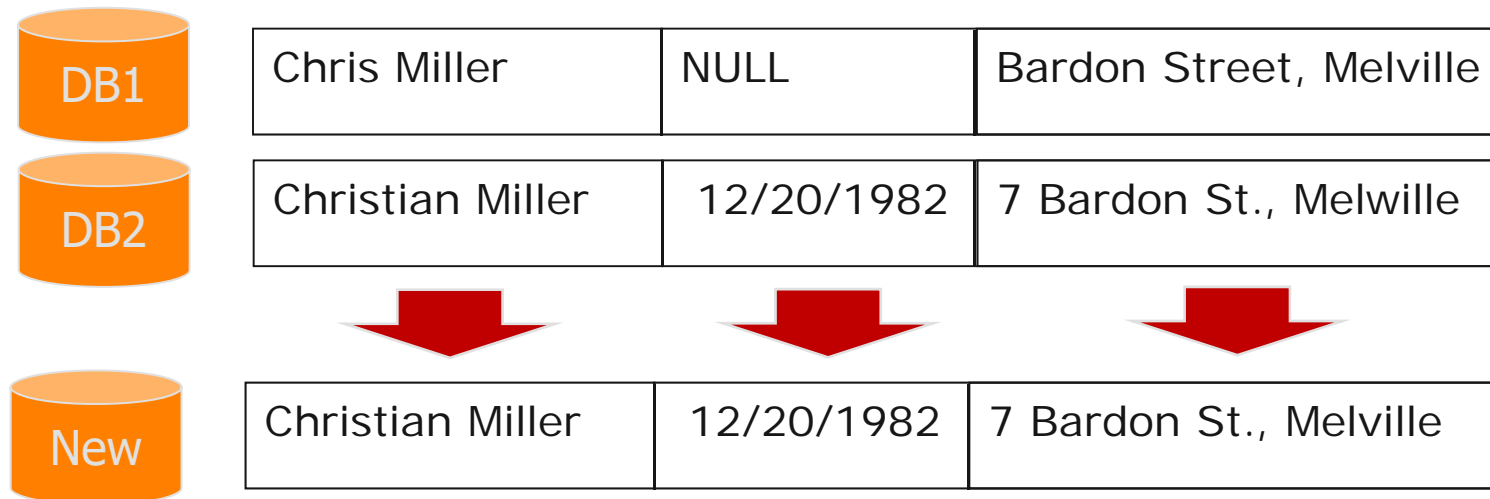
DB1	Chris Miller	12/20/1982	Bardon Street, Melville
DB2	Christian Miller	2/20/1982	7 Bardon St., Melville

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)
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 (like cuteness)

Definition: Data Fusion

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



- **Goal:** Create a high quality record.
- But what does high data quality actually mean?

Data quality is a multi-dimensional construct which measures the **fitness for use of data for a **specific task**.**

Fitness for use

1. has **many dimensions**
 - accuracy, timeliness, completeness, understandability, ...
2. is **task-dependent**
 - you verify information more tightly when you invest 1 million €
3. is **subjective**
 - some people are more paranoid than others

Data Quality Assessment

– Content-based Metrics

- use information to be assessed itself as quality indicator
- examples: constraints and consistency rules, statistical outlier detection

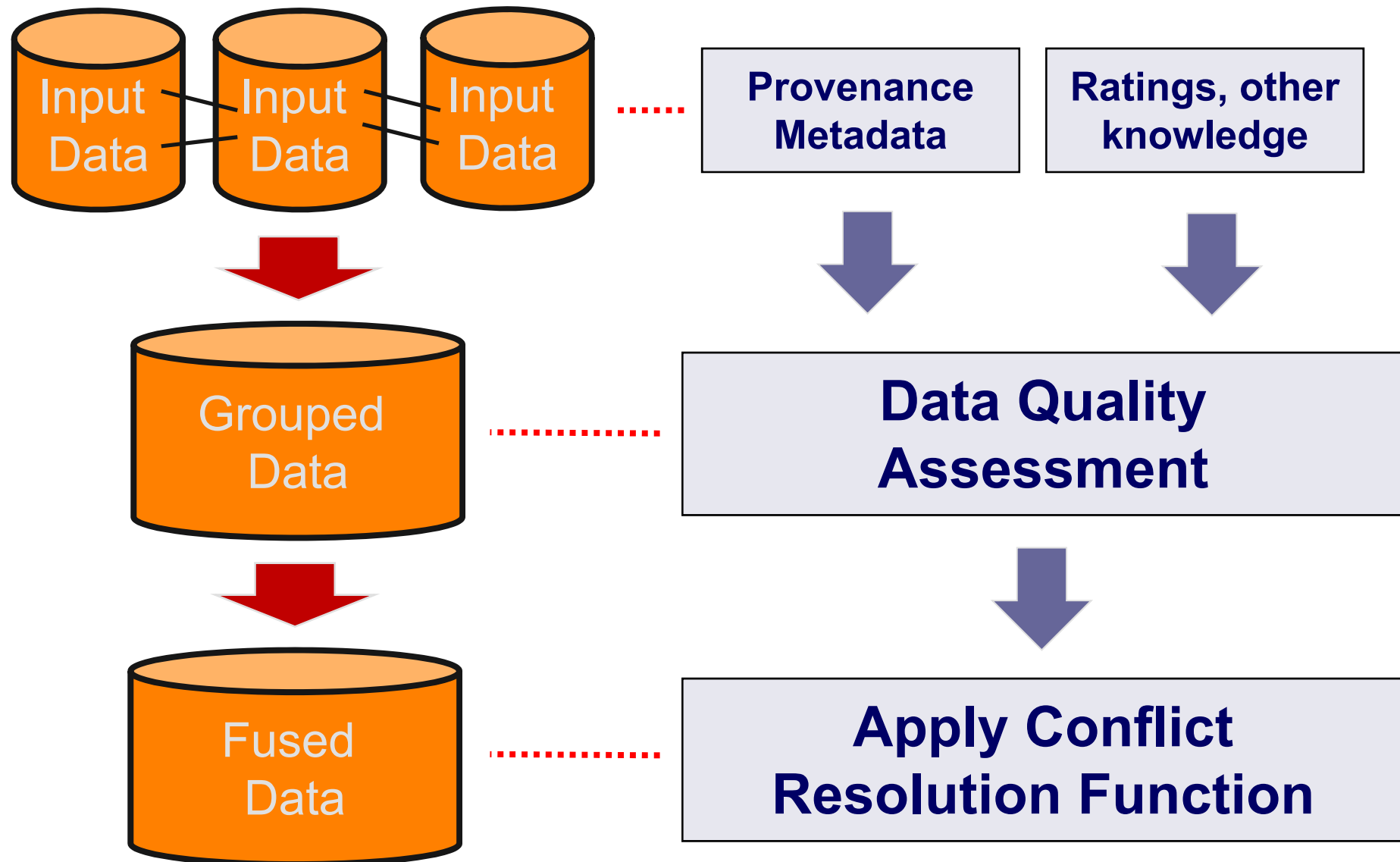
– Provenance-based Metrics

- employ provenance meta-information about the circumstances in which information was created as quality indicator
- examples: “Disbelieve everything a vendor says about its competitor” or “Do not use information that is older than one week”

– Rating-based Metrics

- rely on explicit or implicit ratings about information itself, information sources, or information providers
- examples: “Only read news articles with at least 100 Facebook likes”, “Accept recommendations from a friend on restaurants, but distrust him on computers”, “Prefer content from websites with a high PageRank”

Summary: Elements of the Data Fusion Process



2. Data Provenance

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.

Source: W3C PROV Specification

Provenance information = Important data quality indicator

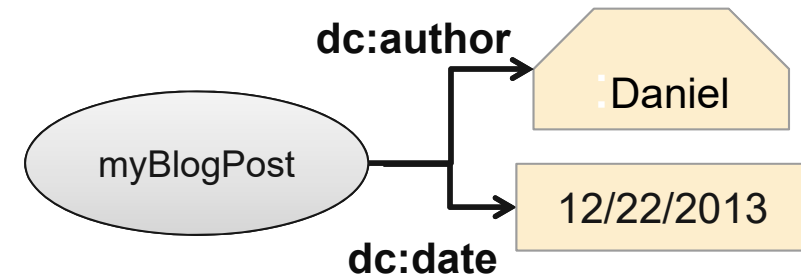
Outline of this Subsection

1. Simple Attribution versus Full Provenance Chains
2. Publishing Provenance Information on the Web
3. Representing Provenance Metadata together with Integrated Data

2.1 Simple Attribution versus Full Provenance Chains

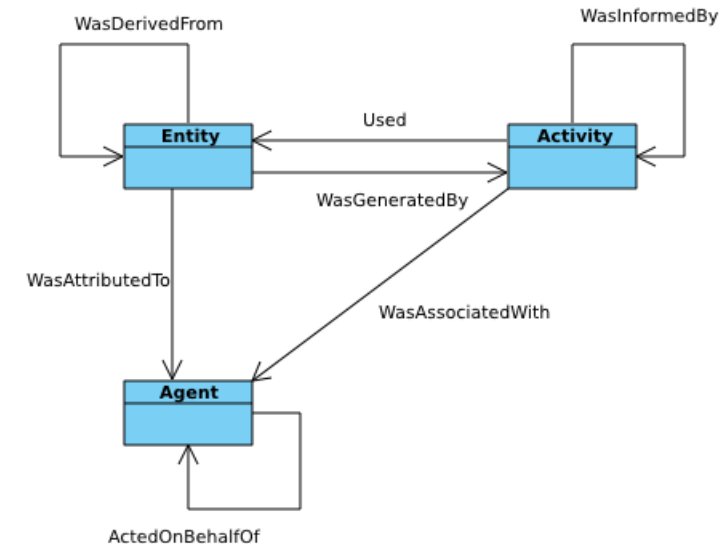
1. Simple Attribution:

- state **who** created a document/data item and **when** it was created
- standard: Dublin Core vocabulary



2. Full Provenance Chains

- Describe the **full process** of data creation / reuse / integration / aggregation
- standard: W3C PROV Specification
- alternative name: Data Lineage (explain why something is in a query result)



– Factors for the decision between both alternatives:

- Will the users be interested in all the details?
 - Yes for law suits and investing. No for deciding which chewing gum to buy
- Can target applications understand/reason about all details?

Application Area in which Provenance Matters

- Web-based Information Systems
 - display origin of data or documents
 - filter, rank or fuse data or documents using provenance as quality indicator
 - explain result creation process
- Science
 - it is important to know how the results of a publication were obtained
 - trace scientific workflows
- Law
 - explain which information was used for a conclusion (e.g. for court cases)
 - prove that information could legally be used for something (e.g. deporting people)
 - license / attribution of documents or data
 - remixing music or films

2.2 Publishing Provenance Information on the Web

In the context of the Web, you always know the **URL** from which you downloaded things. Some sites also give you **Last-Modified** information.

HTTP-Response

```
HTTP/1.1 200 OK
Date: Mon, 18 Jan 2018 20:54:26 GMT
Server: Apache/1.3.6 (UNIX)
Last-Modified: Mon, 06 Dec 2017 14:06:11 GMT
Content-length: 6345
Content-Type: text/html

<html>
  <head><title>CB CD-Shop</title></head>
  <body><h1>Willkommen beim CB CD-Shop</h1> ....
```

Which vocabularies should websites use to publish more detailed provenance information?

- The Dublin Core vocabulary defines terms for representing **simple attribution** information
 - creator, contributor, publisher, date, rights, format, language, ...
- The terms are used in different technical contexts
 - HTML, Linked Data, proprietary library formats
 - Example of a Linked Data document:



http://dbpedia.org/data/Alec_Empire

Metadata and Licensing Information

```
<http://dbpedia.org/data/Alec_Empire>
  rdfs:label "RDF document describing Alec Empire" ;
  rdf:type foaf:Document ;
  dc:publisher <http://dbpedia.org/resource/DBpedia> ;
  dc:date "2017-07-13"^^xsd:date ;
  dc:rights <http://en.wikipedia.org/wiki/WP:GFDL> .
```

The Document Content

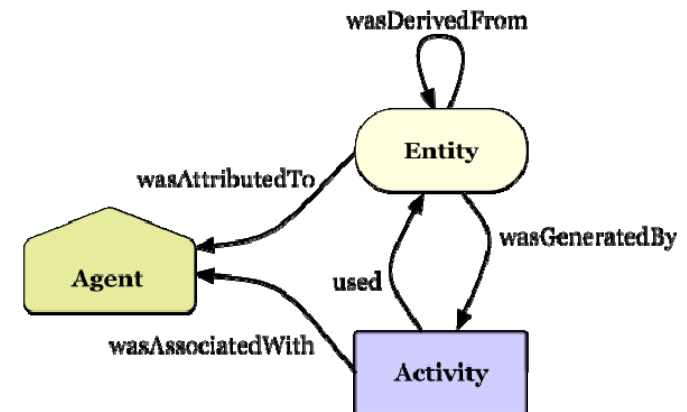
```
<http://dbpedia.org/resource/Alec_Empire>
  foaf:name "Empire, Alec" ;
  rdf:type foaf:Person ;
  rdfs:comment "Alec Empire (born May 2, 1972) is a German musician..."@en ;
  ...
```

W3C PROV

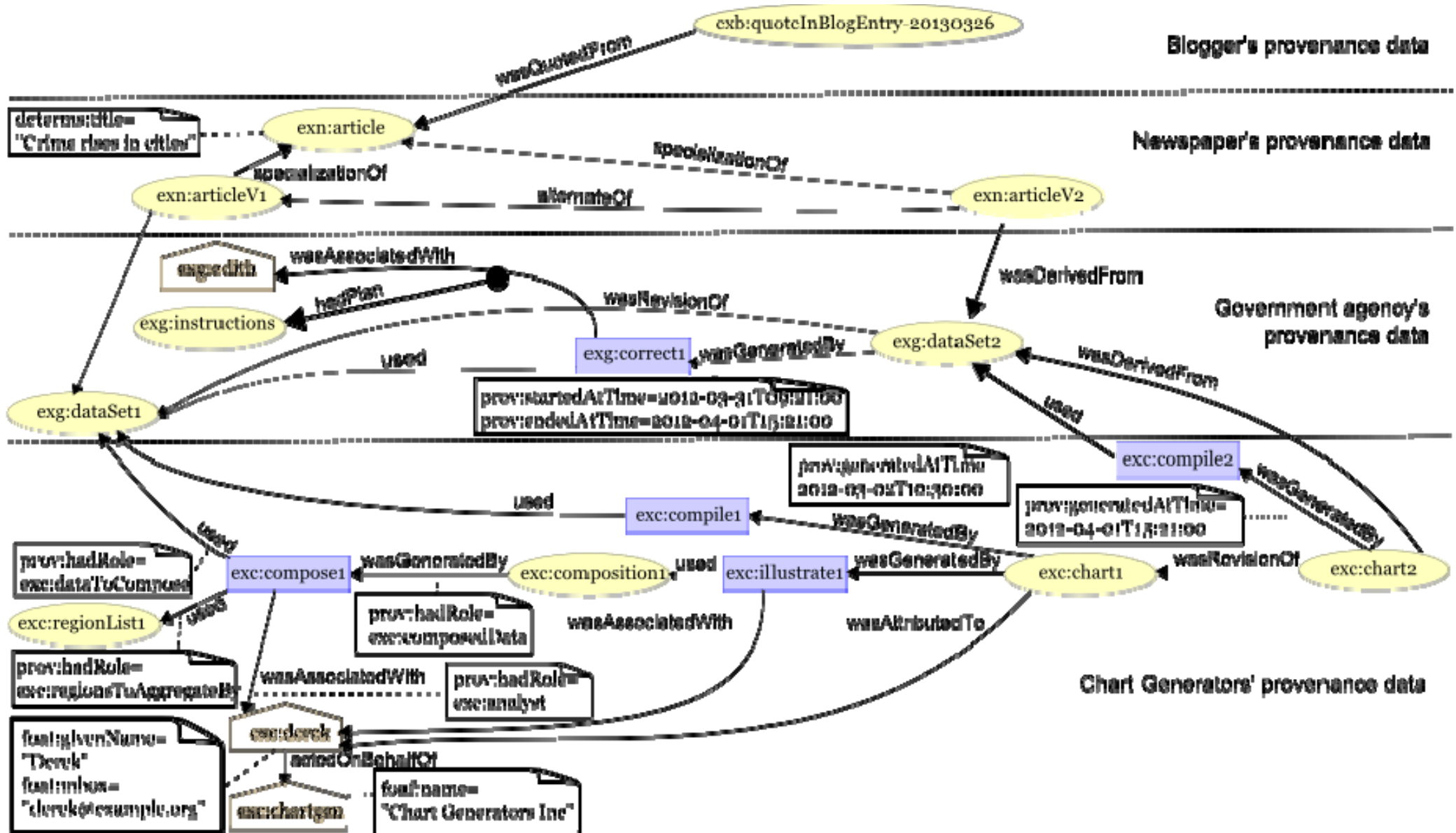
- The W3C PROV vocabulary defines terms for representing **complex provenance chains**
- Example of a PROV XML document:



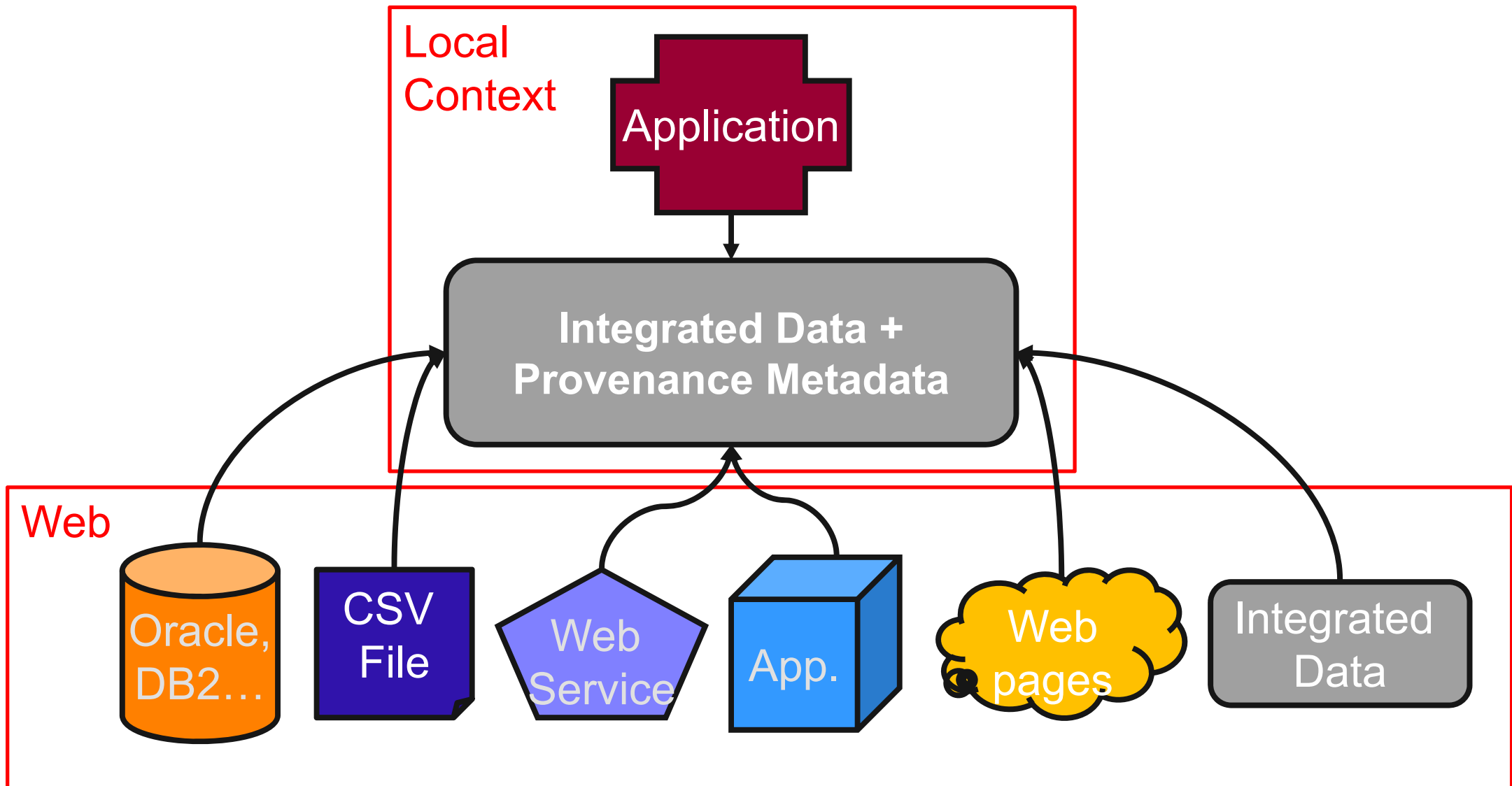
```
<prov:document>
  <!-- Entities -->
  <prov:entity prov:id="exn:article">
    <dct:title>Crime rises in cities</dct:title>
  </prov:entity>
  <!-- Agents -->
  <prov:agent prov:id="exc:derek">
    <prov:type>prov:Person</prov:type>
    <foaf:givenName>Derek Smith</foaf:givenName>
    <foaf:mbox>mailto:derek@example.org</foaf:mbox>
  </prov:agent>
  <!-- Activities -->
  <prov:activity prov:id="exc:compile1"/>
  <!-- Usage and Generation -->
  <prov:wasGeneratedBy>
    <prov:entity prov:ref="exn:article"/>
    <prov:activity prov:ref="exc:compile1"/>
  </prov:wasGeneratedBy>
  <!--Agent's Responsibility -->
  <prov:wasAssociatedWith>
    <prov:activity prov:ref="exc:compile1"/>
    <prov:agent prov:ref="exc:derek"/>
  </prov:wasAssociatedWith>
  ...
```



More Complex Example: W3C PROV



2.3 Representing Provenance Metadata together with Integrated Data



Relational Data Model

- Alternative 1: Record-Level Provenance (coarse grained, fast queries)
- Alternative 2: Value-Level Provenance (fine grained, but slow queries)
- Alternative 3: Employ special database engine which implements extended relational data model with a pointer to provenance information for each attribute value (e.g. Stanford Trio Database)

Physicians with **Record-Level Provenance**

<u>Key</u>	Name	Street	ProvID
1425	Dr. Mark Smith	14 Main Street	001
4217	Mark Smith	12 Main St.	002
...

Physicians with **Value-Level Provenance**

<u>Key</u>	<u>Attribute</u>	Value	ProvID
1425	Name	Dr. Mark Smith	001
1425	Name	Mark Smith	002
1425	Street	14 Main Street	001
...

Provenance Table

ProvID	Source	Date
001	www.mark-smith.com	12/6/2017 18:42:12
002	www.doc-find.com	12/1/2017 12:21:54
...

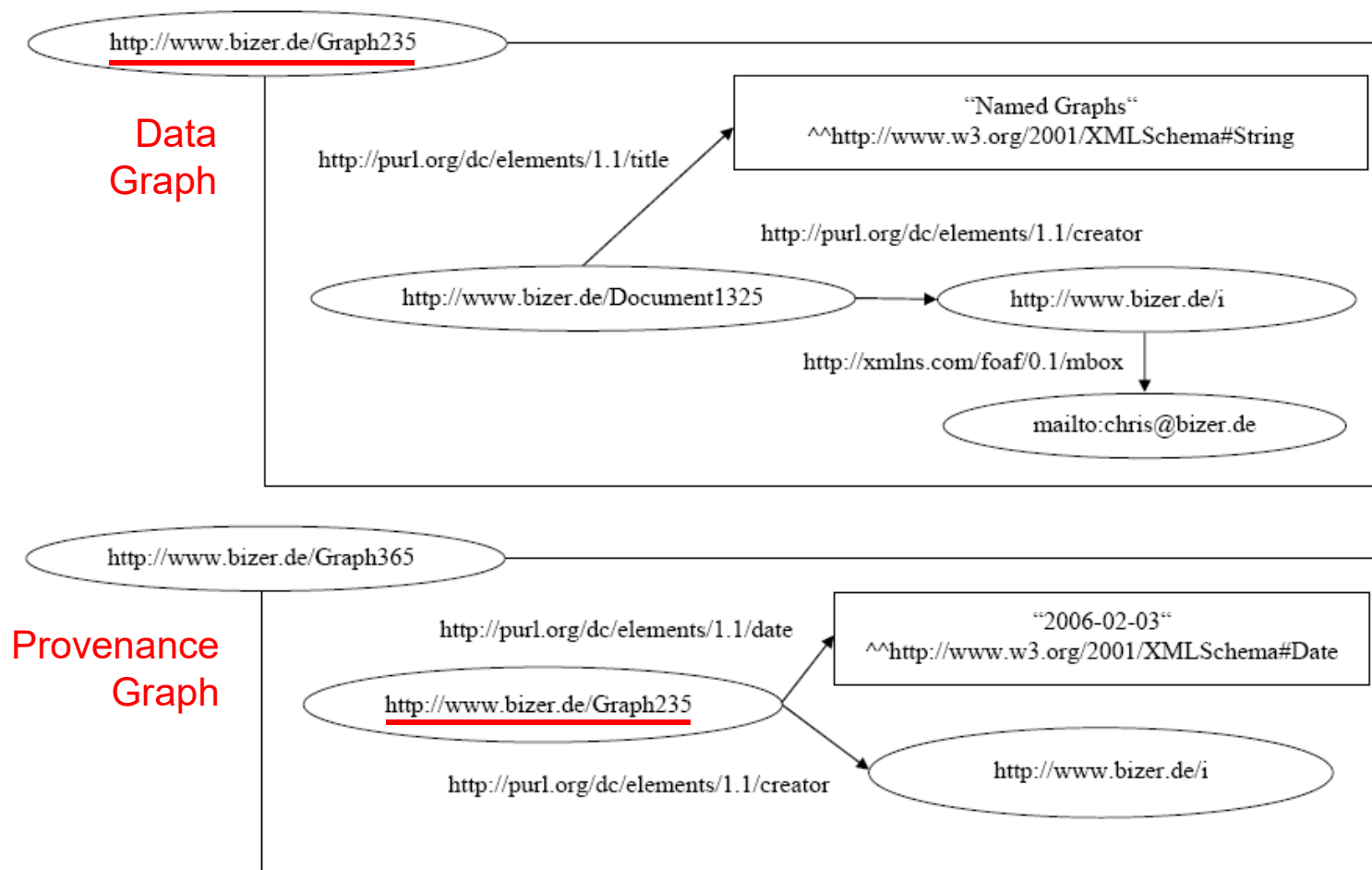
XML Data Model

Represent provenance using multiple **value elements** and references to **provenance elements**.

```
<physician>
  <name>
    <value prov="prov01">Dr. Mark Smith</value>
    <value prov="prov02">Mark Smith</value>
  </name>
  <address>
    <street>
      <value prov="prov01">14 Main Street</value>
      <value prov="prov02">12 Main St.</value>
    </street>
    <city> ... </city>
  </address>
</physician>
<provenance id= "prov01">
  <source>http://www.marksmith.com/index.htm</source>
  <date>06 Nov 2017 14:06:11 GMT</date>
</provenance>
<provenance id= "prov02">
  ...
```


RDF Data Model

- Group triples into **Named Graphs** (= set of triples that is identified by a URI)
- Provide provenance information by talking about a graph in another graph
- Named Graphs can be queried using the SPARQL keyword GRAPH



Carroll, Bizer, Hayes, Stickler:
Named Graphs. Journal of
Web Semantics, 2005.

3. Data Quality

Data quality is a multi-dimensional construct which measures the “fitness for use” of data for a specific task.

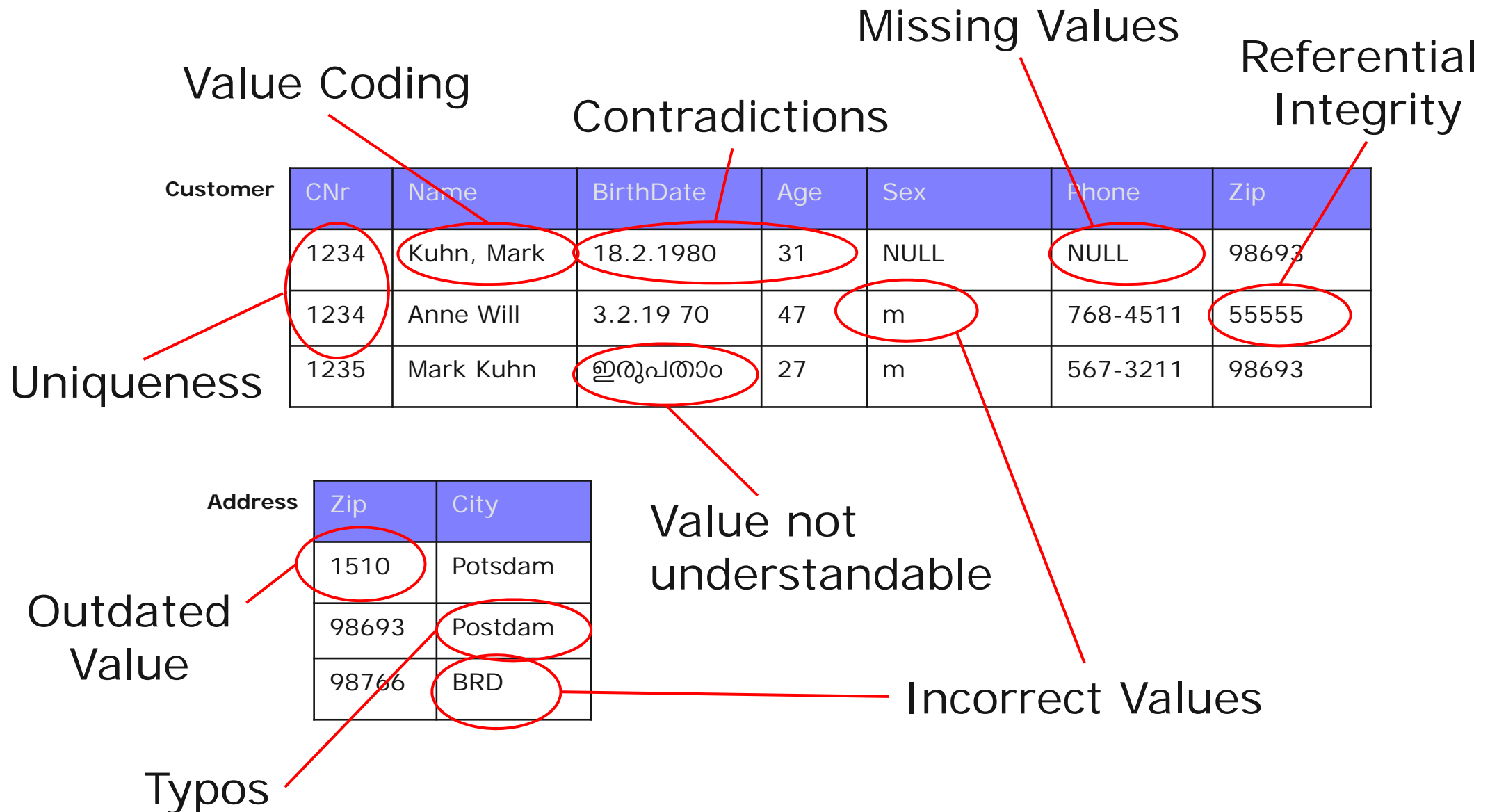
- Which quality dimensions matter depends on the task
- The required level of quality depends on the task and the user

Outline of this Subsection

3.1 Data Quality Dimensions

3.2 Data Quality Assessment

Different Types of Data Quality Problems



Data Quality in the Enterprise and Web Context

– Enterprise Context

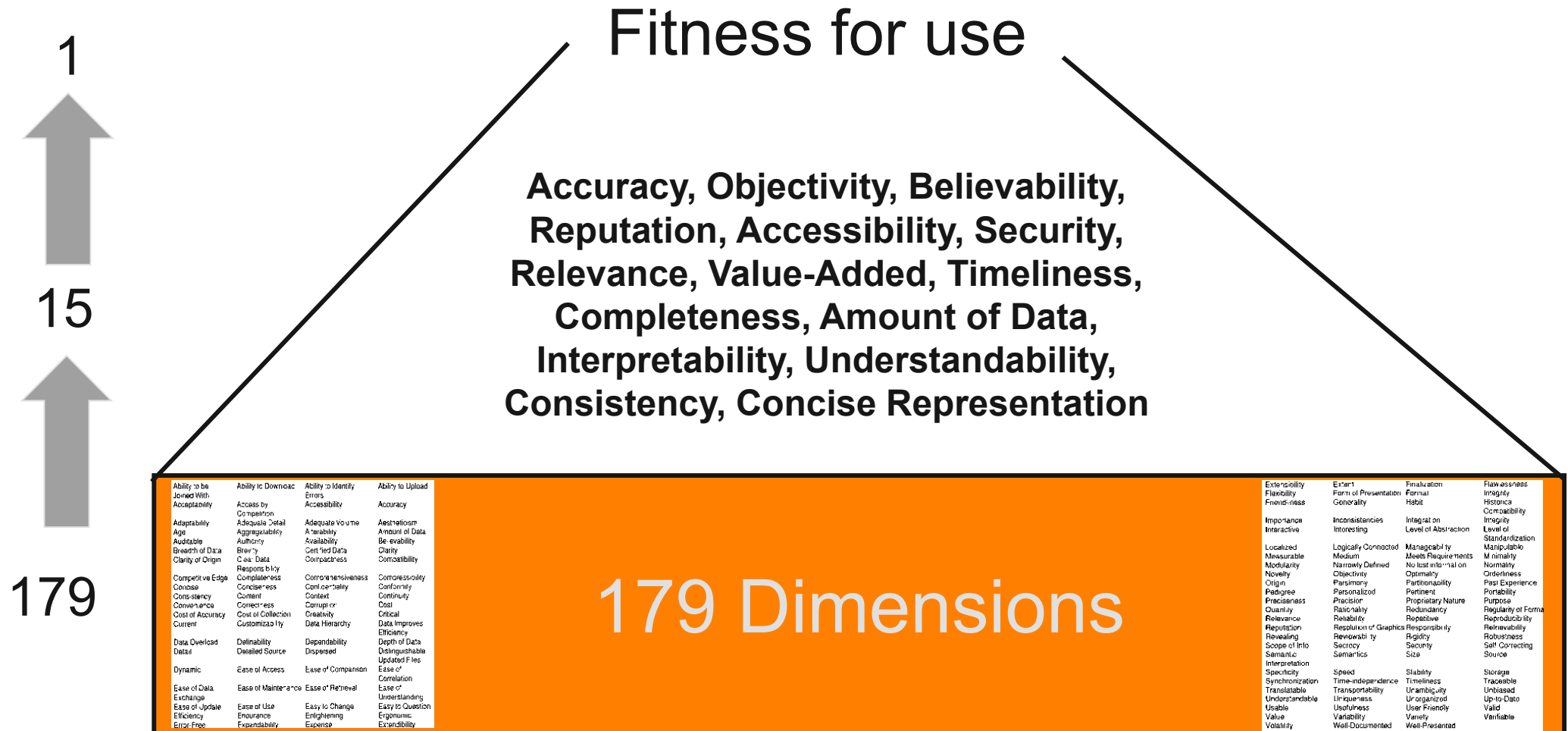
- the goal is to establish **procedures and rules** that guarantee high quality data production, quality monitoring, and regular data cleansing
- pioneering research by MIT Total Data Quality Management (TDQM) program
- consequences of low data quality:
 - A.T. Kearny: 25%-40% of the operational costs result from low data quality as low quality data leads to wrong management decisions
 - US postal service: out of 100.000 mass-letters, 7.000 cannot be delivered because of wrong address
 - SAS: Only 18% of all German companies trust their data

– Web Context

- large number of data sources, but no possibility to influence data providers
- thus, focus on **identifying the high-quality subset** of the available data
- challenge: Quality indicators often sparse and unreliable

3.1 Data Quality Dimensions

As part of the MIT Total Data Quality Management (TDQM) program, [Wang/Strong1996] asked managers which data quality dimensions matter for their tasks:



Category	IQ Criteria	TDQM	MBIS	Weikum	DWQ	SCOUG	Chen
Content-related Criteria	Accuracy	Yes	Yes	Yes	Yes	Yes	Yes
	Documentation					Yes	
	Relevancy	Yes	Yes		Yes		Yes
	Value-Added	Yes				Yes	
	Completeness	Yes	Yes	Yes	Yes	Yes	Yes
	Interpretability	Yes			Yes		
Technical Criteria	Timeliness	Yes	Yes	Yes	Yes	Yes	Yes
	Reliability			Yes			
	Latency			Yes			Yes
	Performability			Yes		Yes	
	Response time		Yes	Yes			Yes
	Security	Yes		Yes	Yes		
	Accessibility	Yes	Yes	Yes	Yes	Yes	
	Price		Yes	Yes		Yes	
	Customer Support					Yes	
Intellectual Criteria	Believability	Yes	Yes	Yes	Yes	Yes	
	Reputation	Yes	Yes		Yes		
	Objectivity	Yes					
Instantiation related Criteria	Verifiability			Yes			
	Amount of data	Yes	Yes				Yes
	Understandability	Yes	Yes				
	Concise represent.	Yes					
	Consistent represent.	Yes	Yes	Yes	Yes	Yes	

Source: Felix Naumann

Content-related Data Quality Dimensions

...concern the actual data.

- **Accuracy**

- is the extent to which data is correct, reliable, and certified free of error [WS96]

- **Timeliness**

- is the extent to which the age of the data is appropriate for the task at hand [WS96]

- **Completeness**

- is the extent to which data is not missing and is of sufficient breadth, depth, and scope for the task at hand [WS96]

- **Interpretability**

- is the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear [WS96]

- **Documentation**

- is the amount and usefulness of documents with metadata

- **Relevancy (or relevance)**

- is the extent to which data is applicable and helpful for the task at hand [WS96]

- **Value-Added**

- is the extent to which data is beneficial and provides advantages from its use [WS96]

Technical Data Quality Dimensions

...concern software and hardware used to access the data.

- **Accessibility** (or availability)

- Is the extent to which data are available or easily and quickly receivable [WS96]

- Latency

- is the amount of time in seconds from issuing the query until the first data item reaches the user

- Response time

- measures the delay in seconds between submission of a query by the user and reception of the complete response from the IS

- **Price**

- is the amount of money a user has to pay for a query
- is the extent to which the cost of collecting appropriate data is reasonable [WS96]

- Security

- is the extent to which access to data is restricted appropriately to maintain its security [WS96]

Intellectual Data Quality Dimensions

...concern subjective aspects.

- **Believability**

- is the extent to which data is regarded as true, real, and credible [WS96]

- Objectivity

- is the extent to which data is unbiased, unprejudiced, and impartial [WS96]

- Reputation

- is the extent to which data is trusted or highly regarded in terms of its source or content [WS96]

- Reliability

- is the degree to which the user can trust the information

Instantiation-related Data Quality Dimensions

...concern the representation of data.

- **Amount of data**

- is the extent to which the quantity or volume of available data is appropriate [WS96]

- **Representational conciseness**

- is the extent to which data is compactly represented without being overwhelming [WS96]

- **Representational consistency**

- is the extent to which data is always represented in the same format and are compatible with previous data [WS96]

- **Understandability (ease of understanding)**

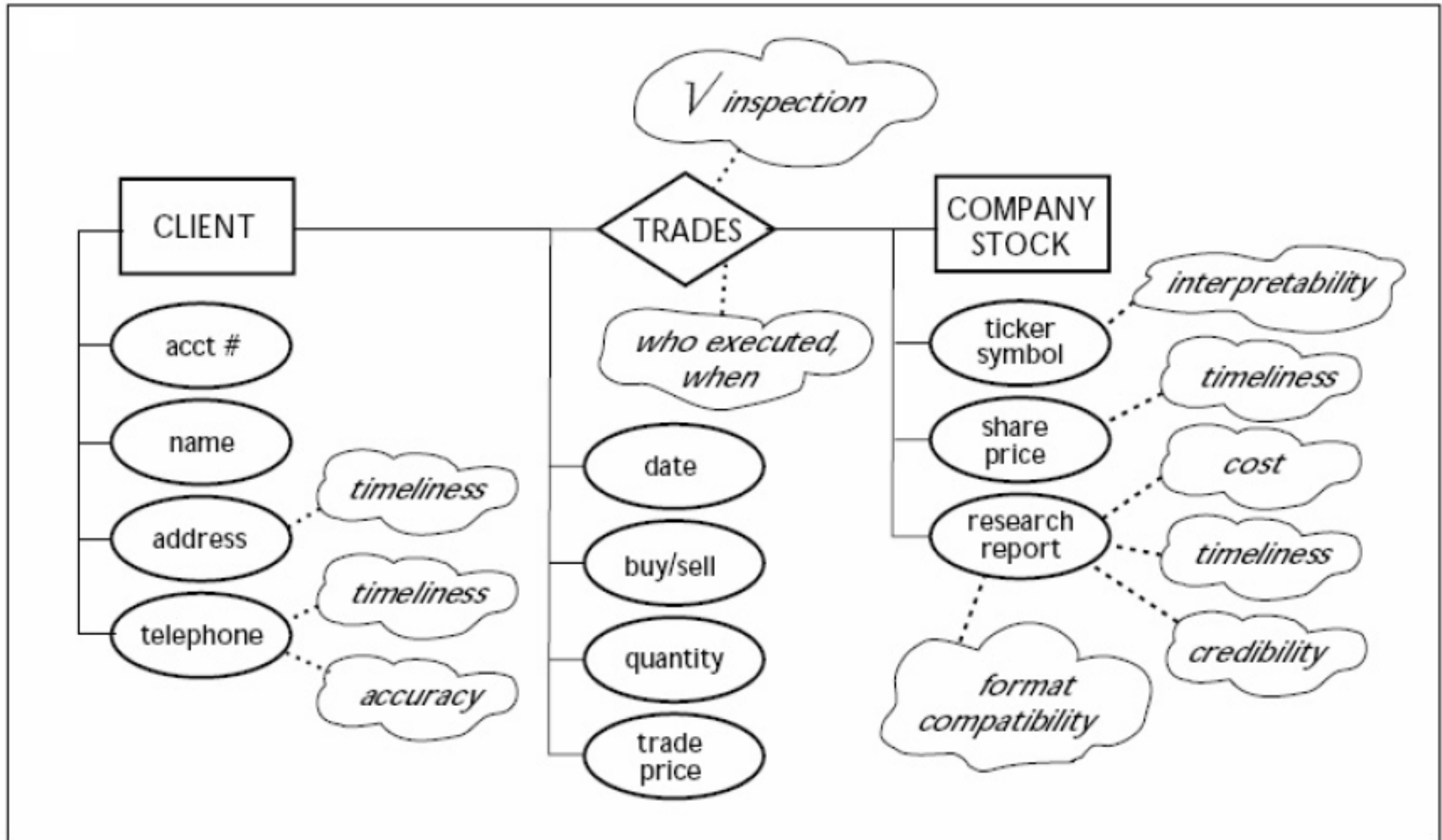
- is the extent to which data are clear without ambiguity and easily comprehended [WS96]

- **Verifiability** (traceability, lineage)

- is the extent to which data are well documented, verifiable, and easily attributed to a source [WS96]

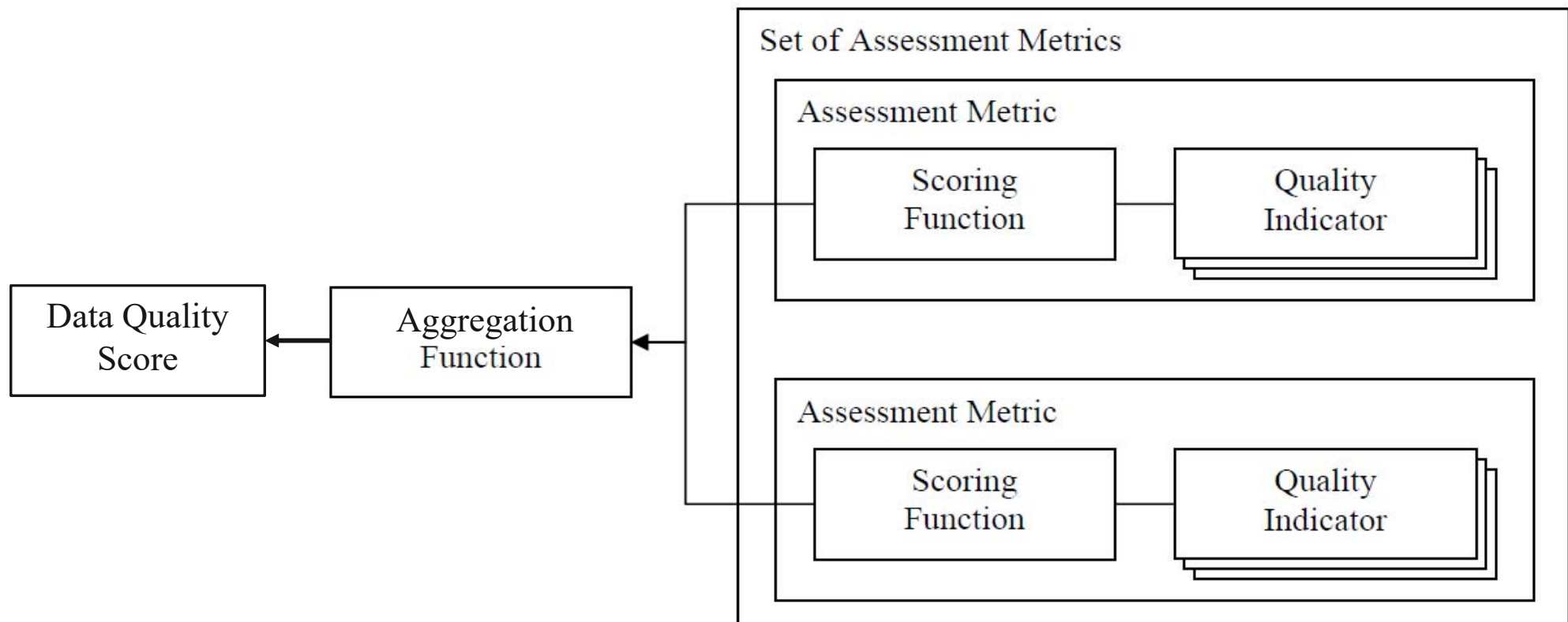
Relevancy of Data Quality Dimensions

Which quality dimensions matter depends on the task at hand.



3.2. Data Quality Assessment

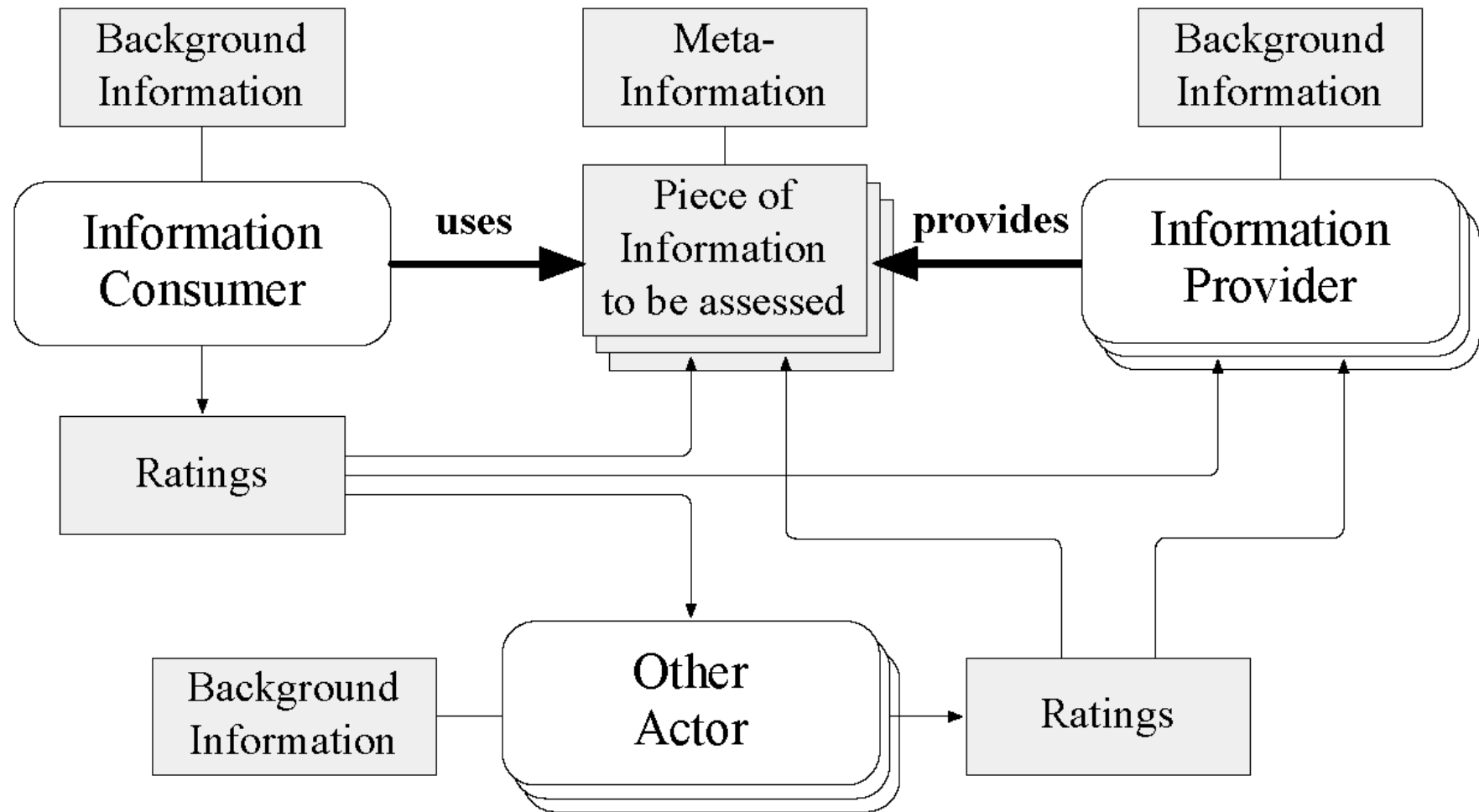
Various domain-specific heuristics are used to measure data quality.



The **applicability** of specific heuristics depends on


1. Availability of quality indicators (like provenance information or ratings)
2. Quality of quality indicators (fake ratings, sparse provenance information)


Quality Indicators in the Web Context



Examples of Data Quality Assessment Metrics

Number of DQ classification schemata containing specific quality dimensions:

We will look at 

<i>Dimension</i>	<i>Count</i>
Accuracy	7
Timeliness	7
Completeness	6
Relevancy	5
Availability	5
Rep. Consistency	4
Amount of Data	4
Interpretability	3
Rep. Conciseness	3
Security	2
Objectivity	2
and  Believability	2
Understandability	2
Verifiability	2
Response Time	2
Consistency	2
Reputation	1

Source: Bizer, 2007

Assessing Data Accuracy

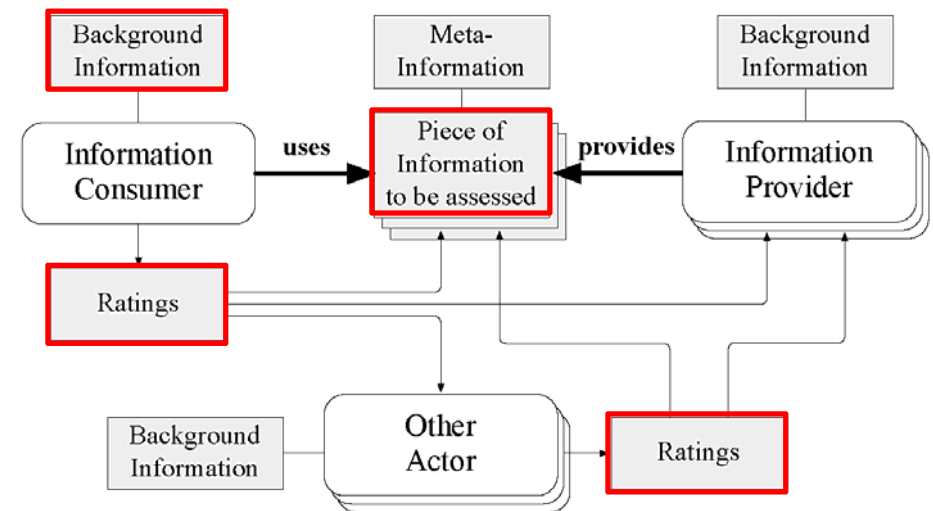
– Definition:

- Accuracy is the extent to which data is correct, reliable, and free of error
- Also called: **Truth Discovery**

– Assessment Methods:

1. Outlier detection
2. Constraint testing
3. Lookup tables / reference data
4. Expert- or user ratings

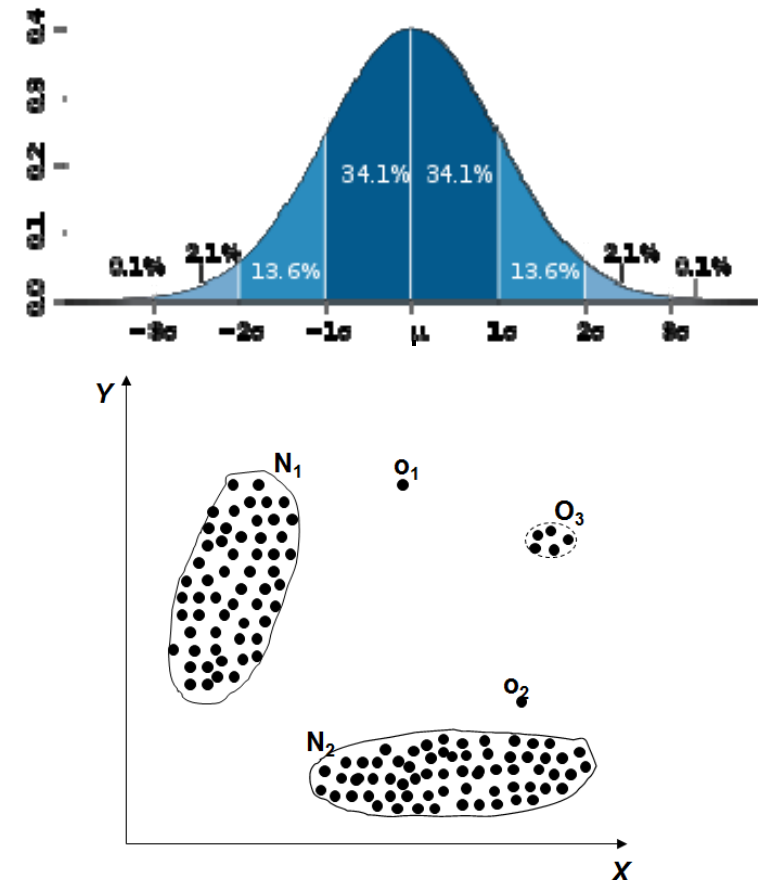
– Relevant quality indicators:



Outlier Detection

An outlier is a individual data instance that is anomalous with respect to the rest of the data.

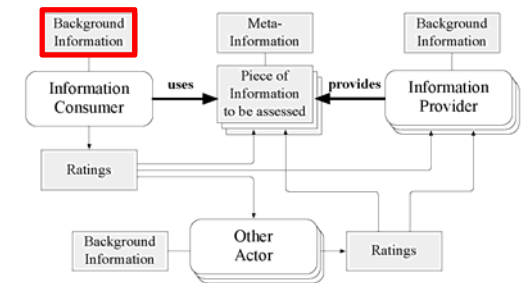
- Outliers can be considered as errors and be assigned a low quality score
- Techniques
 - statistical distributions, clustering, classification
- Challenges
 - the exact notion of an outlier is different for different application domains
 - an individual may be a outlier w.r.t. a single attribute or a combination of multiple attributes
 - normal behaviour keeps evolving over time
 - natural outliers: Population of Mexico City
- More information
 - Chandola, et al.: Anomaly Detection: A Survey. ACM Computing Surveys, 2009.



Constraint Testing

Match data against constraints and consistency rules in order to detect errors.

- Examples of consistency rules
 - if person is in middle school, then age is (likely) below 25
 - if area code is 131, then the city should be Edinburgh
- Examples of constraints
 - books must have at least one author
 - the age of humans should be between 0 and 130
 - disbelieve everything a vendor says about its competitor.
- Rule and constraint acquisition
 - define rules and constraints manually
 - or learn from examples e.g. using association analysis (see lecture Data Mining)
- More information
 - Fan, Geerts: Foundations of Data Quality Management. Morgan & Claypool, 2012.



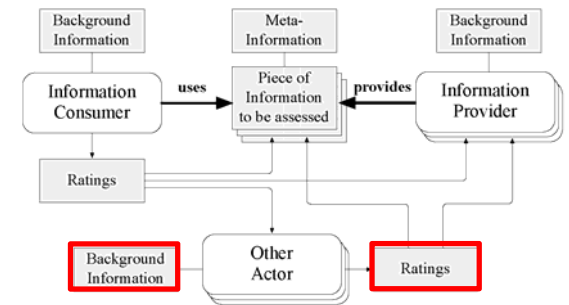
Ratings

Data is often filtered or ranked based on ratings provided by users or experts.

- Various scoring functions exist
 - practical systems often use simple, easily understandable functions

- Challenges:

1. Motivate users to rate
 - data
 - data providers
 - data sources
2. Quality of the ratings
 - fake ratings
 - clueless raters



A screenshot of a TripAdvisor page for 'The Place Luxury Boutique Villas' in Koh Tao, Thailand. The page shows the hotel's location, contact info, and a list of reviews. Two reviews are visible: one by 'ExplorerAsh' (5 stars, 'It's so nice you don't need to leave') and one by 'pcgrowler' (5 stars, 'Could be perfect!'). Each review includes a photo, the reviewer's name, location, and a 'Was this review helpful?' button.

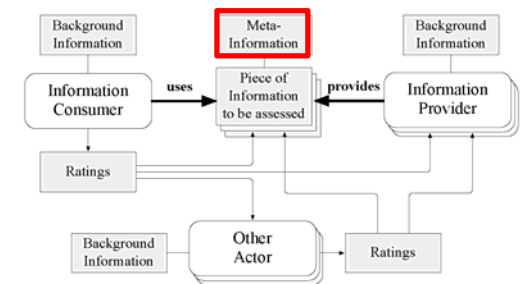
Implicit Ratings

- Events potentially interpretable as positive ratings
 - clicks, page views
 - time spent on some page
 - items bought, ...
- Advantage
 - large amounts of implicit ratings can be collected constantly by the application
 - collection of ratings does not require additional effort from the user
- Problem
 - one cannot be sure whether the user behavior is correctly interpreted
 - for example, a user might not like all the books he or she has bought; the user also might have bought a book for someone else
- More details: Web Mining Lecture – Chapter: Recommender Systems

Assessing Data Timeliness

The assessment of the timeliness of data usually requires provenance data.

- Provenance metadata
 - HTTP Last-Modified
 - dc:date
- Fallbacks if no timestamps are available
 - propagate timestamps to data without timestamps
 - e.g. two tables provide same profit for a company, only one table has a timestamp
 - Zhang, Chakrabarti: InfoGather+, SIGMOD 2013.
 - use rules instead of timestamps
 - Number of children: Prefer higher value, as number of children of a person usually grows
 - Employee salaries: Prefer higher values, as salaries usually do not decrease



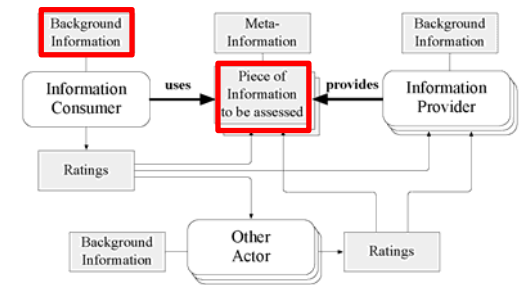
Assessing Data Completeness

– Definition:

- The extent to which data is not missing and is of sufficient breadth, depth, and scope for the task at hand
- **Density**: Fraction of attributes filled
- **Coverage**: Fraction of real-world objects represented

– Assessment:

- Density
 - Sample data source and calculate density from sample
- Coverage
 - hard to calculate as overall number of real-world objects is unknown in many cases: Countries: OK; Products or people: Problematic
 - fallback: Prefer data sources that describe more entities



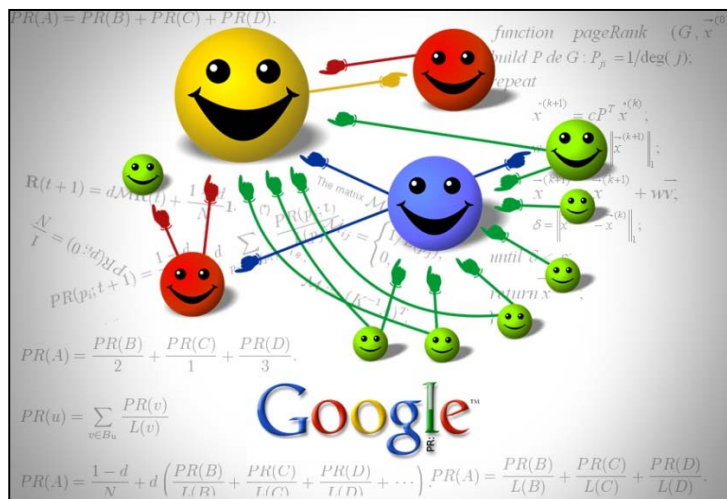
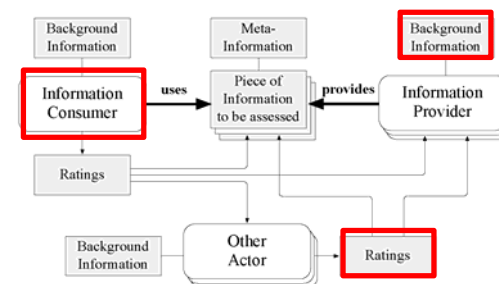
Assessing Data Relevancy

– Definition:

- The extent to which data is applicable and helpful for the task at hand

– Assessment:

- Example: TripAdvisor
 - Filter reviews based on background information about information provider
- Example: Google
 - Rank webpages based on search terms and PageRank score



31 reviews from our community [Write a Review](#)

Traveler rating

Excellent	0
Very good	2
Average	0
Poor	3
Terrible	3

Trip type

- [Family reviews \(8\)](#)
- [Couples reviews \(12\)](#)
- [Solo travel reviews \(1\)](#)
- [Business reviews \(0\)](#)
- [Friends reviews \(6\)](#)

Your selections Families ☐

8 reviews sorted by [Date](#) [Rating](#)

salman_10
geneva switzerland
1 review
2 helpful votes

"nightmare in koh tao"
Reviewed January 10, 2010

We had booked three nights at the black tip resort and it was terrible. The staff is always in a bad moon, never polite or helpfull. The beach was full of garbage and gazoil . we could not swimm and it was worst everyday. The restaurant staff would never smile or say hi . Well, having travel many times in...

Was this review helpful? 2 [Problem with this review?](#)

Assessing Believability / Trustworthiness

– Definition

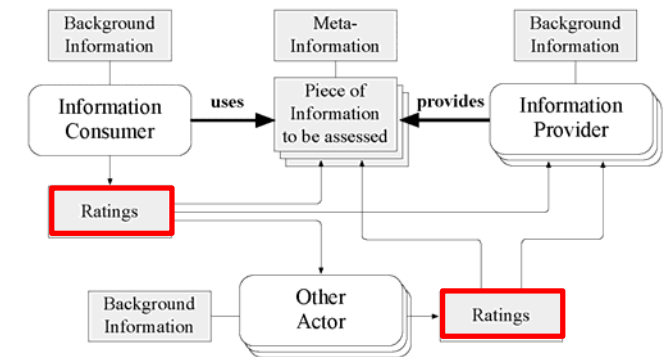
- The extent to which data is regarded as true, real, and credible.
- Subjective dimension which depends on the individual user

– Assessment:

- Individual experience with the data
- Fallbacks:
 - corporate guidance about sources
 - trust networks

– Explanations about the data quality assessment process

- in order to trust data, the users must understand why the system regards data to be high quality
- Tim Berners-Lee's “Oh, yeah?”-button



Prototype: The WIQA - Browser

- Enables users to employ different quality assessment policies
- Can explain assessment results

The screenshot shows the WIQA Browser interface in Mozilla Firefox. The browser window title is "WIQA Browser - Mozilla Firefox". The address bar shows the URL: <http://127.0.0.1:1978/piggy-bank/e1eb9ba7fe10653332021055d7562c83/default?command=browse&policyURI=Information+from+German+analysts&=&%40lwq.project.Proj>. The page title is "WIQA Browser" and the timestamp is "19.07.2006 14:35:50".

The main content area displays a search result for the URI `urn:ISIN:DE0007236101`. The result is titled "positive analyst report" and includes a snippet of text: "Siemens agrees partnership with Novell unit SUSE. Siemens Business Services (SBS), the IT services arm of German technology conglomerate Siemens, said on Tuesday it had agreed a partnership deal with Novell's (nasdaq: NOVI) Windows software such as Microsoft (nasdaq: MSFT - news - asking more and more for open-source platforms, SBS partner status. SBS is one of Europe's top 10 exclusive province of a few dedicated enthusiasts, supported by U.S. giant International Business Machines (nyse: IBM - news - people), among others. Its advocates, who include big businesses and government departments, argue it is cheaper, simpler and more secure than Windows."

A callout box labeled "Oh, yeah? Button" points to a button labeled "Oh, yeah? Button" located below the text snippet.

The right sidebar contains a "Policy Selection Panel" with a search bar and a list of policies. The selected policy is "Information from German analysts". The list of policies includes:

- ☐ is a
- ☐ name
- ☐ discussion forum posting
- ☐ emitted by
- ☐ positive analyst report
- ☐ negative analyst report

Below the list, the selected policy is displayed: "Policy: Information from German analysts". A list of criteria for this policy is shown:

- ☐ Information from German analysts
- ☐ Information from positively rated information providers
- ☐ New information from highly rated analysts
- ☐ Only German or English information
- ☐ Accept only information from Deutsche Bank
- ☐ More positive Ratings
- ☐ TidalTrust rating above 5
- ☐ Asserted by two different analysts
- ☐ Asserted by analysts with at least 3 positive ratings
- ☐ Accept everything

At the bottom of the sidebar, there are logos for "simile" and "Simile".

Explanation about an Assessment Decision

The screenshot shows a Mozilla Firefox browser window titled "WIQA Browser - Mozilla Firefox". The address bar displays a URL: `http://127.0.0.1:1978/piggy-bank/e1eb9ba7fe10653332021055d7562c83/default?command=browse&policyURI=Information+from+German+analysts&=&%40lwq.project.Proj`. The page content includes a header "WIQA Browser" with a timestamp "19.07.2006 14:35:50". Below this, there is a section "1 filter criterion" with a bullet point "is a: Share (remove) [add more]". A table with two columns, "Order" and "Commands", is visible. The main content area shows "2 items sorted by name [A to Z]". The first item is "urn:ISIN:DE0007236101", which is a "Share" and a "positive analyst report". The second item is "urn:ISIN:US4581401001", which is a "Share" and a "negative analyst report". A "Show Referers" button is present at the bottom of the list. An "Explanation" window is open over the first item, titled "EXPLANATION" and "WIQA Browser". It contains the following text:

The Triple:

Siemens Share positive analyst report Siemens agrees partnership with Novell unit SUSE. Siemens Business Services (SBS), the IT services arm of German technology conglomerate Siemens <SIEGN.DE>, said on Tuesday it had agreed a partnership deal with Novell's (nasdaq: NOVL - news - people) newly acquired unit SUSE Linux. Linux software is open-source, meaning it can be freely copied and modified, unlike proprietary software such as Microsoft (nasdaq: MSFT - news - people) Windows. In the past months clients have been asking more and more for open-source platforms, SBS said in a statement which said SUSE would have premier partner status. SBS is one of Europe's top 10 information technology service providers. Linux, once the exclusive province of a few dedicated enthusiasts, is now seen as the only serious rival to Windows and is supported by U.S. giant International Business Machines (nyse: IBM - news - people), among others. Its advocates, who include big businesses and government departments, argue it is cheaper, simpler and more secure than Windows.

fulfils the policy:

Use only information which has been asserted by German analysts.

because:

- it is stated in the document **Information from Peter Smith**, which is asserted by the German analyst **Peter Smith**.

The window also has a "Close" button and a "Fertig" button at the bottom.

The triple:

- Siemens AG has positive analyst report: "As Siemens agrees partnership with Novell unit SUSE ..."

fulfills the policy:

- Accept only information that has been asserted by people who have received at least 3 positive ratings.

because:

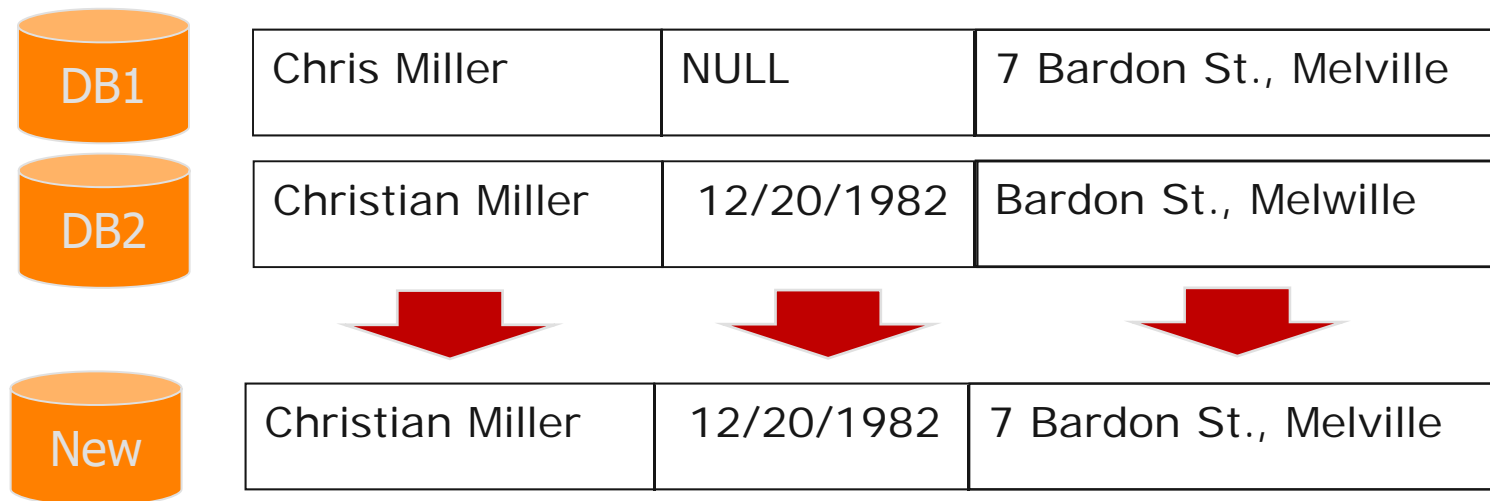
- it was asserted by Peter Smith and
- Peter Smith has received positive ratings from
 - Mark Scott who works for Siemens.
 - David Brown who works for Intel.
 - John Maynard who works for Financial Times.

Summary

- Data quality assessment is essential for Web data integration as **errors accumulate**:
 1. Quality of the external data sources (everybody can publish on the Web)
 2. Quality of the integration process (wrong mappings, wrong identity resolution)
- Many data quality problems only become visible when we integrate data from multiple sources
- A wide **range of different quality assessment heuristics** can be used
 - content-based, provenance-based, rating-based metrics
- The **applicability** of the heuristics depends on
 - the availability of quality indicators (like provenance information or ratings)
 - quality of quality indicators (fake ratings, coarse grained provenance)
- Many systems only try to assess the accuracy and the timeliness of Web data and ignore the other quality dimensions

4. Data Fusion

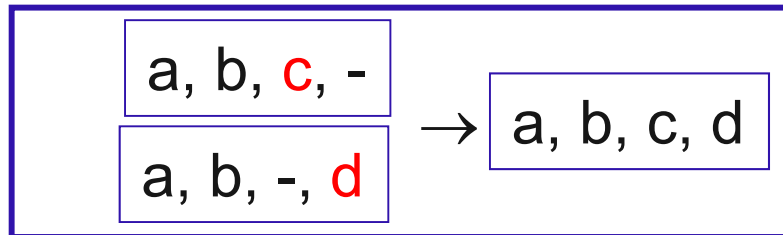
Given multiple records that describe the same real-world entity, create a single record while resolving conflicting data values.



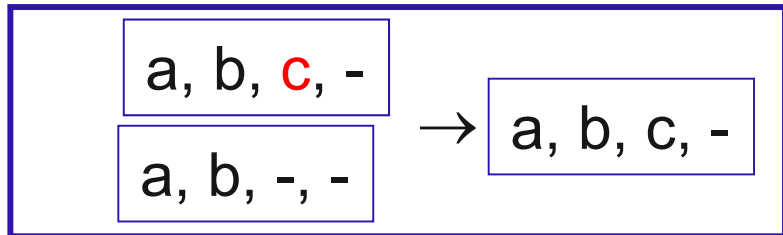
- Goal: Create a single high quality record.
- Two basic fusion situations: Uncertainty and Contradiction

4.1 Uncertainty and Contradiction

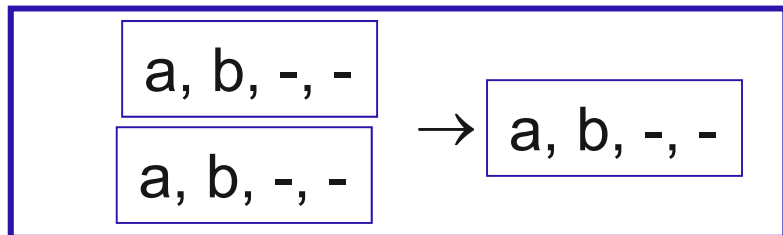
NULL values indicate that a source is uncertain about a value.



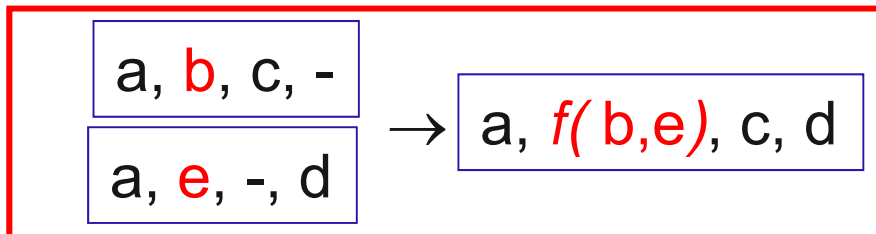
Complementary records \Rightarrow Take information



Subsumed records \Rightarrow Remove subsumed



Identical records \Rightarrow Remove duplicate



Conflicting records \Rightarrow Apply Conflict Resolution Function

4.2 Conflict Resolution Functions

- Conflict Resolution Functions are attribute-specific
 - you define a specific function for each attribute that should be fused
- There is a wide range of different functions that fit different requirements
- Functions differ in regard to the data types, they can be applied for
 - numerical values (e.g. population of a place)
 - nominal values (e.g. name of a person)
 - value sets (e.g. actors performing in a movie)
- Two main groups of conflict resolution functions
 1. Functions that rely only on the data values to be fused
 2. Functions that rely on provenance data, ratings, or information quality scores

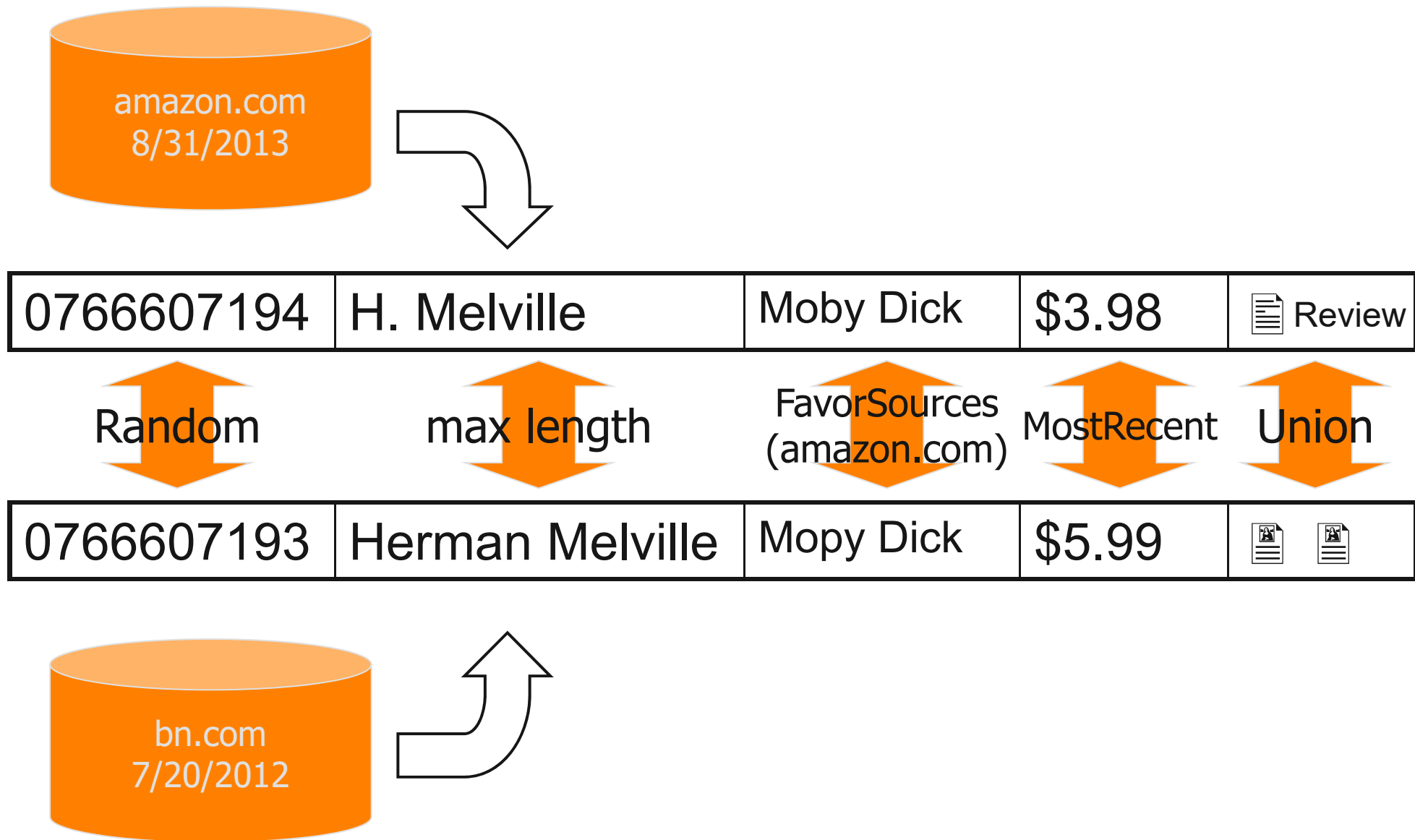
Functions that rely on the Data Values

Function	Explanation	Use Case
Vote	Majority decision (one vote per site or page?)	Capital city
ClusteredVote	Choose centroid / medoid of largest cluster	Population of city
Average, Median	Calculate average/median of all values	Rating
Longest, Shortest	Choose longest / shortest value	First name
Max, Min	Take maximal, minimal value	Number of children
Union	Union of all values ($A \cup B \cup C$)	Product Reviews
Intersection	Intersection of all values ($A \cap B \cap C$)	Movie Actors
IntersectionKSources	Values must appear in at least k sources	Movie Actors
ChooseDepending	Choose depending on value of other attribute (See: Fan Geerts, 2012)	city & zip, e-mail employer
MostComplete	Choose value from record that is most complete	People's addresses
MostAbstract,MostSpecific	Use a taxonomy / ontology	Location
Random	Fallback: Choose random value	

Functions that rely on Provenance, Ratings or IQ Scores

Function	Explanation
Favor Sources	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, NumChildren
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
TopkIQ	Choose the top K values with the highest quality scores
....

Example: Complete Conflict Resolution Heuristic



4.3 Evaluation of Data Fusion Results

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

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 normalization, translated data sets 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.

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

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

Measurement:

1. Combine multiple tables using entity correspondences
 - group records that refer to same real-world entity
2. Calculate fraction of non-conflicting attribute values
 - same attribute value is provided by all data sources

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

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

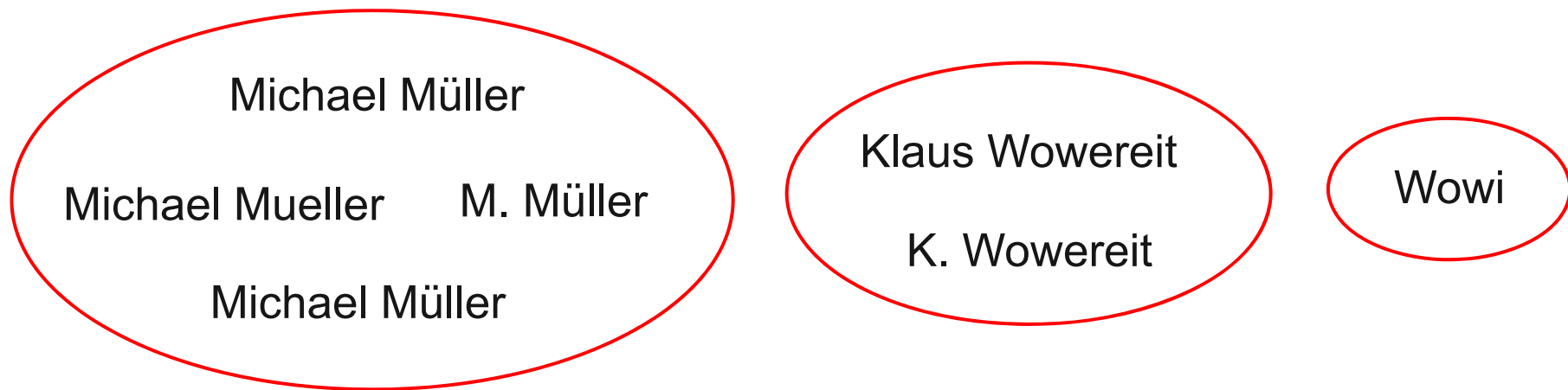
$$error\ rate = 1 - accuracy$$

Measurement:

1. Build 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 selected by fusion function with true values

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 Chapter Identity Resolution)
 2. Treat all values above threshold as equal
- Example: Mayor of Berlin



4.4. Example Data Fusion Tool: Fuz!on

Fuzzy Fuz!on [Additional Information] [Test/Debug]

Automatic Fusion | **Rule-based Fusion** | Manual Fusion

Rule Matrix

	Firstname	Lastname	Street	houzenumber	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

Actions

Fusionsregel(n) anzeigen/erzeugen ☒ Nur aktuelle Markierung anzeigen **WEITER -->**

Selected Rules

Regeldefinition (Status: neu)

Spalten
Firstname
Lastname

Konflikttypen
Low edit distance

Primäre Konfliktauflösung
Vote
Minimum fraction of solution (in %) : 50


☒ Ignore case
☒ Ignore null-values

Sekundäre Konfliktauflösung
First

Aktionen
Übernehmen
Ausblenden
Spalte hinzufügen
Konflikttyp hinzufügen

Prototype
developed at
Hasso Plattner
Institute

Manual Fusion of Record Groups in Fuz!on

 **Fuzzy Fuz!on**

Additional Information Test/Debug

Automatic Fusion

Rule-based Fusion

Manual Fusion

Groups 0 to 50 of 100000

All Groups

☐ Filter Mode

fdb.group	Firstname	Lastname	Street	houzenumber	postcode	city	ignore	phone
31750025-01	Werner	Trimpert	Thomas-Man...	89	24943	Kiel	19470524	0461
31758055-01	Artur	Heiser	Kalkgrund	4	24939	Kiel	19360106	
31765505-01	Siegfried	Aswegen	Mürwiker Str.	6	4943	Flensburg	19250404	0461
31772625-01	M.	Blankenburg	Harmsstr.	48	24116	Kiel	19610727	0461
31780965-01	K	Degen	Peter-Chr.-H...	5	24114	Flensburg	19630331	0461
31789325-01	Manh The	Knaut	Wiedeberger ...	37	24943	Flensburg	19280312	0461
31798345-01	horst	Booitsmann		6	24937	Flensburg	19281225	0461

Back

Next

21. Group :


	Firstname	Lastname	Street	houzenumber	postcode	city	ignore	phone
	Manh The	Knaut	Wiedeberger Weg	37	24943	Flensburg	19280312	0461
	Manh The	KNAUT	Wiedeberger Weg		24943	Flensburg	19280312	0461
	Manh	Knaut	WIEDEBERGER WEG	37	24943	Flensburg	19280312	0461
	<div>First</div>	<div>Mixed ...</div>	<div>Vote</div>	<div>First non-null value</div>	<div>First</div>	<div>First</div>	<div>First</div>	<div>First</div>
	Manh The	Knaut	Wiedeberger Weg	37	24943	Flensburg	19280312	0461

Merge

Save Configurations

4.5 Case Study: DBpedia Cross Language Data Fusion

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



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Featured content](#)

Article [Talk](#)


Mannheim

From Wikipedia, the free encyclopedia

This article is about the city in Germany.

► **Mannheim** ([help](#)·[info](#)) is a city in southwest and Karlsruhe.

Area	
• Total	144.96 km ² (55.97 sq mi)
Elevation	97 m (318 ft)
Population (2011-12-31) ^[1]	
• Total	314,931
• Density	2,200/km ² (5,600/sq mi)



WIKIPEDIA
Die freie Enzyklopädie

[Hauptseite](#)
[Themenportale](#)
[Von A bis Z](#)

Artikel [Diskussion](#)

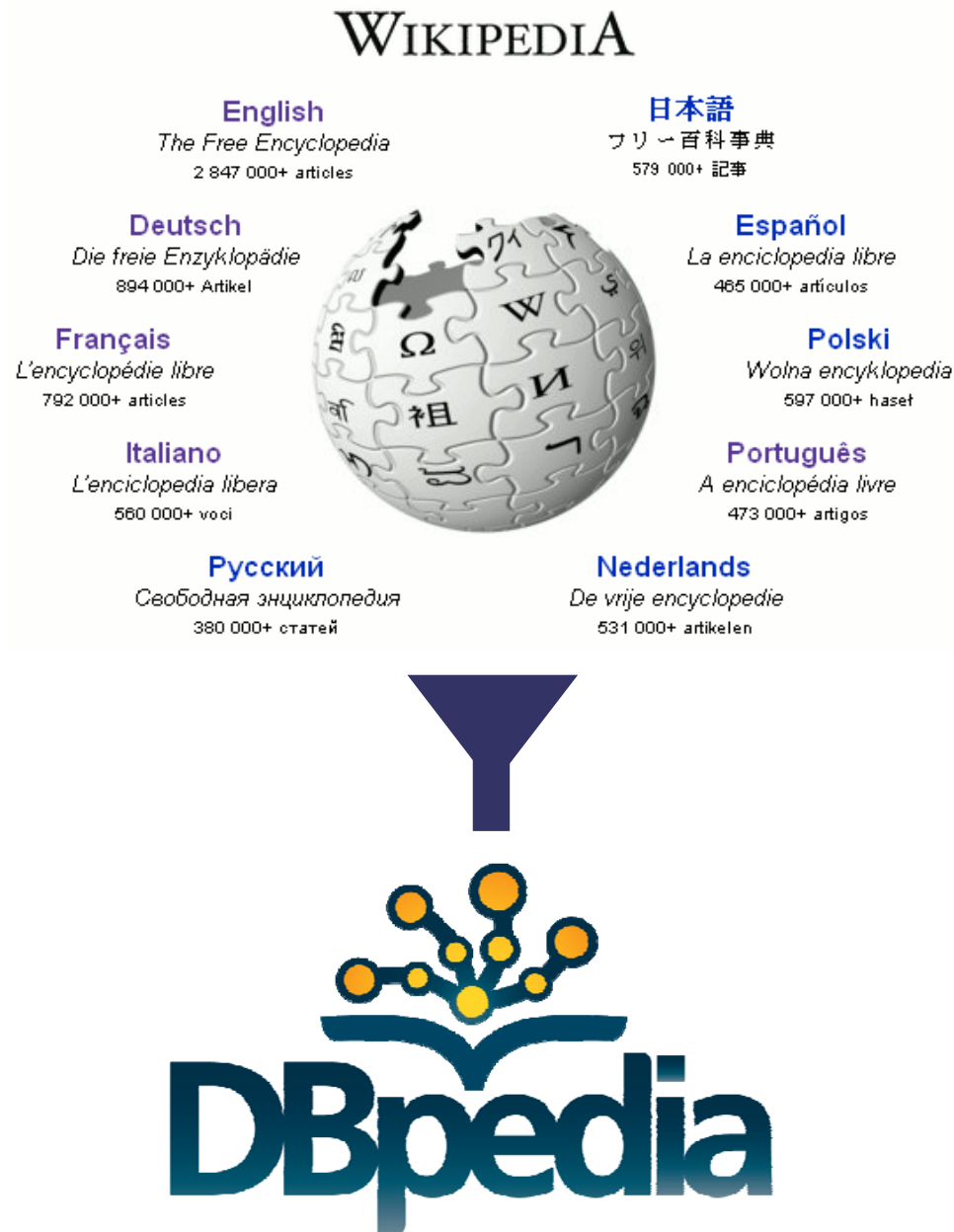
Mannheim

Der Titel dieses Artikels ist mehrdeutig. Weiter

Die [Quadratstadt](#) und [Universitätsstadt](#) **Mannheim** (1720–1778) der historischen [Kurpfalz](#) bildet das wi

Höhe:	97 m ü. NHN
Fläche:	144,96 km ²
Einwohner:	291.458 (31. Dez. 2011) ^[1]
Bevölkerungsdichte:	2011 Einwohner je km ²

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

**data
itself**

**prove
nance**

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 gold standard.
- 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

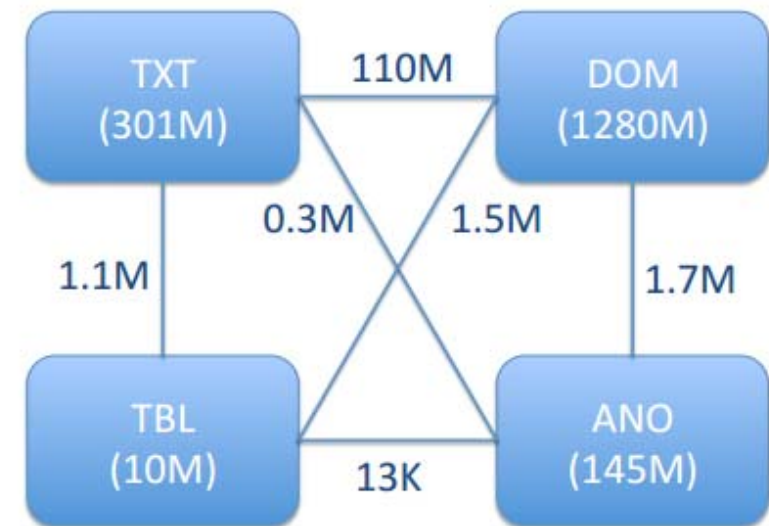
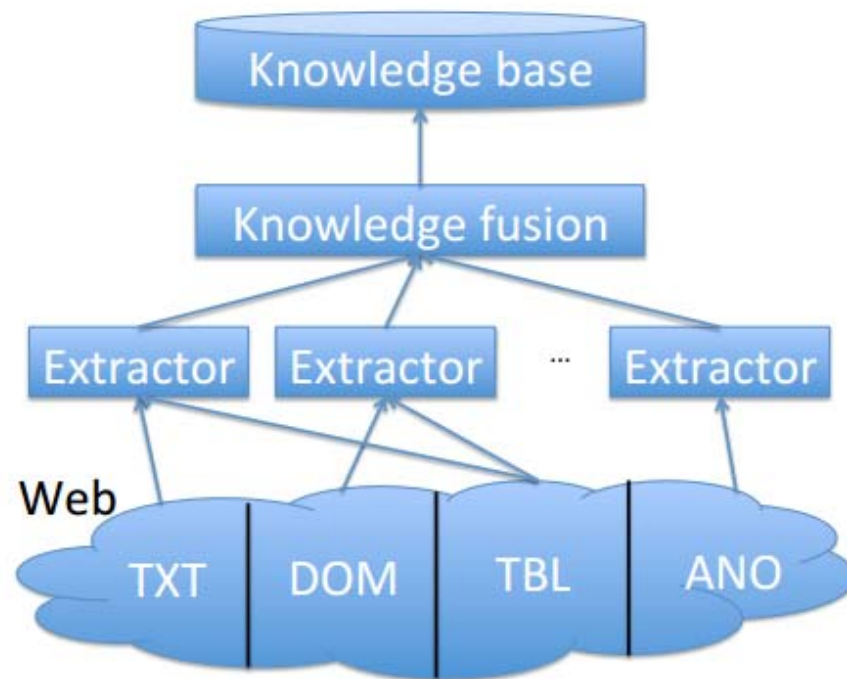
* “cities1000” are cities with population >1000

- **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.

4.6 Case Study: Google Knowledge Vault

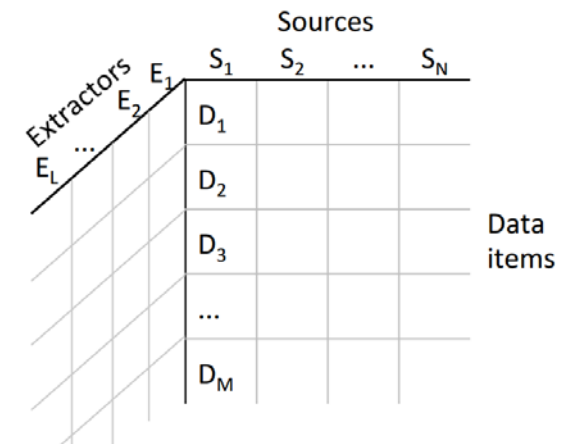
- uses 12 different extractors to extract 6.4 billion triples (1.6 billion unique triples) from 1 billion page Web crawl
- extracted data is fused to extend the Freebase knowledge base



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

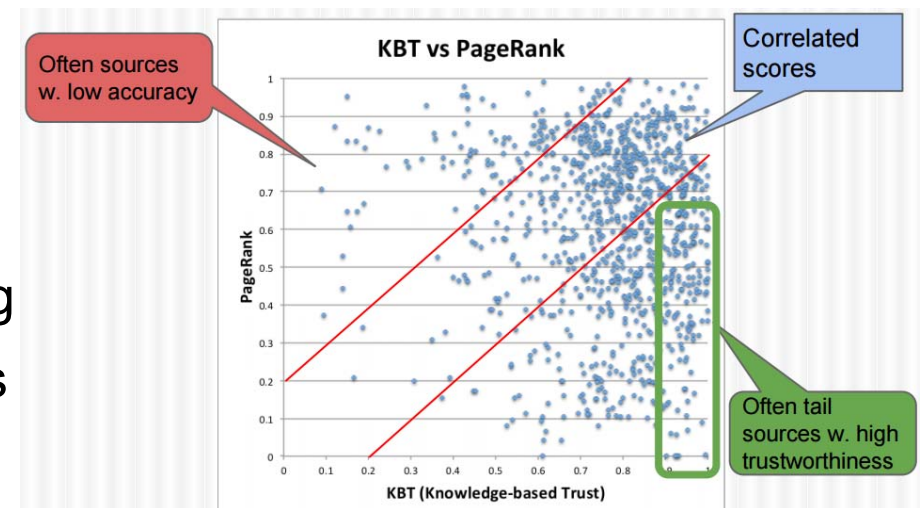
Google Knowledge Vault

- 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

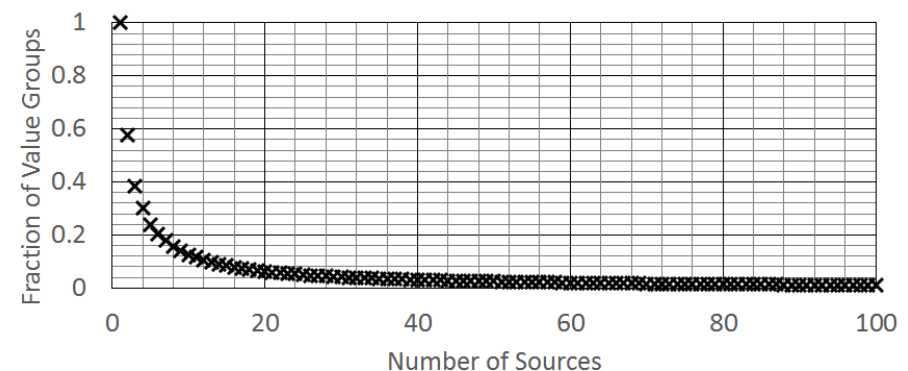
- determine trustworthiness of a data source by comparing its content with a knowledge base (ground truth)
- result: Better than PageRank in identifying
 - tail websites with high trustworthiness
 - gossip websites



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.

Summary: Data Fusion

- Data Fusion addresses **uncertainty** (missing values) as well as **contradictions** (data conflicts)
- Appropriate conflict resolution function depends on
 - data type of the values
 - availability of quality-related metadata
 - availability of overlapping data
- 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 entities (some village)
 - example: Web tables matched to DBpedia
 - conflict resolution function cannot do anything in 40% of the cases as there is only a single value ☹



5. References

– Provenance

- Dublin Core Metadata Element Set. <http://dublincore.org/documents/dces/>, 2012.
- Gil, Miles: PROV Model Primer, <http://www.w3.org/TR/prov-primer/>, 2013.
- Doan, Halevy, Ives: Principles of Data Integration. Chapter 14, Morgan Kaufmann, 2012.
- Cui, Widom: Lineage tracing for general data warehouse transformations. VLDB Journal, 2003.

– Data Quality

- Wang, Strong: Beyond accuracy: What data quality means to data consumers. JMIS, 1996.
- Naumann, Rolker: Assessment Methods for Information Quality Criteria. Conference on Information Quality, 2000.
- Rahm & Do: Data Cleaning: Problems and Current Approaches, IEEE Bulletin, 2000.
- Ziawasch et al.: Detecting data errors: where are we and what needs to be done? VLDB 2016.
- Fan, Geerts: Foundations of Data Quality Management. Morgan & Claypool, 2012. (Focus: integrity rules)
- Chandola et al.: Anomaly Detection: A Survey. ACM Computing Surveys, 2009.
- Aggarwal: Managing and Mining Uncertain Data. Springer, 2010. (Focus: probabilistic models)

References

– Data Fusion

- Bleiholder, Naumann: Data Fusion. ACM Computing Surveys, 2008.
- Li , Gao, Meng, et al.: Survey on Truth Discovery. arXiv, 2015.
- Leser, Naumann: Informationsintegration. Chapter 8.3, dpunkt Verlag, 2007.
- Dong, Srivastava: Big Data Integration. Chapter 4. Morgan & Claypool, 2015.
- Ganti, Das Sarma: Data Cleaning: A Practical Perspective. Morgan & Claypool, 2013.
- Bleiholder, Naumann: Conflict Handling Strategies in an Integrated Information System. IIWeb, 2006.
- Luna Dong & Felix Naumann: Data Fusion. Tutorial at VLDB 2009.
Slides: http://dc-pubs.dbs.uni-leipzig.de/files/dataFusion_vldb.pdf
- Luna Dong, et al.: From Data Fusion to Knowledge Fusion. VLDB 2014.
- Theo Rekatsinas: Tutorial Data Integration and Machine Learning. SIGMOD 2018.
Chapter ML for DF. https://thodrek.github.io/di-ml/sigmod2018/slides/05_MLforDF.pdf

– Data Fusion Evaluation Datasets

- Luna Dong: Data Sets for Data Fusion Experiments
<http://lunadong.com/fusionDataSets.htm>