Web Data Integration

Data Quality Assessment and Data Fusion
The Data Integration Process

- Data Collection
- Schema Mapping
  - Data Translation
- Identity Resolution
- Data Quality Assessment
  - Data Fusion
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 information on the Web is partly wrong, biased, outdated, incomplete, and inconsistent.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 of Monaco

Different claims:

- en.wikipedia.org: 2.02 sq km
- www.state.gov: 1.95 sq km
- www.atlapedia.com: 1.94 sq km
- military.countrywatch.com: 2 sq km

Different unit of measurement 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)

Source: Peter Bunemann
### Example: Population of Monaco in 2004

<table>
<thead>
<tr>
<th>Year</th>
<th>Value</th>
<th>Metainformation</th>
<th>Webpage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>29,972</td>
<td></td>
<td><a href="http://worldatlas.com/webimage/countrys/europe/mc.htm">http://worldatlas.com/webimage/countrys/europe/mc.htm</a></td>
</tr>
<tr>
<td>2003</td>
<td>32,130</td>
<td>(July 2003 est.)</td>
<td><a href="http://www.greenfacts.org/studies/climate_change/index.htm">http://www.greenfacts.org/studies/climate_change/index.htm</a></td>
</tr>
<tr>
<td>2003</td>
<td>30,000</td>
<td></td>
<td><a href="http://www.tlfq.ulaval.ca/axl/europe/monaco.htm">http://www.tlfq.ulaval.ca/axl/europe/monaco.htm</a></td>
</tr>
<tr>
<td>2000</td>
<td>32,020</td>
<td>(C 2000-05-03)</td>
<td><a href="http://www.citypopulation.de/Monaco.html">http://www.citypopulation.de/Monaco.html</a></td>
</tr>
<tr>
<td>2000</td>
<td>31,693</td>
<td>(July 2000 est.)</td>
<td><a href="http://www.mapquest.com/atlas/?region=monaco">http://www.mapquest.com/atlas/?region=monaco</a></td>
</tr>
<tr>
<td>2000</td>
<td>31,842</td>
<td></td>
<td><a href="http://www.fact-index.com/m/mo/monaco.html">http://www.fact-index.com/m/mo/monaco.html</a></td>
</tr>
<tr>
<td>2000</td>
<td>31,842</td>
<td></td>
<td><a href="http://en.wikipedia.org/wiki/Monaco">http://en.wikipedia.org/wiki/Monaco</a></td>
</tr>
<tr>
<td>2000</td>
<td>31,700</td>
<td>(e2000m)</td>
<td><a href="http://www.library.uu.nl/wesp/populstat/Europe/monacoc.htm">http://www.library.uu.nl/wesp/populstat/Europe/monacoc.htm</a></td>
</tr>
</tbody>
</table>
Analysis: Population of Monaco in 2004

- The 2004 figures vary by 17% !!
- For deciding which value to trust, we would need to know
  1. the original source of the data
  2. the timeliness of the data
  3. the methods used to produce the data
- Most sites copy from the CIA World Factbook
  - The US state department does not – and contradicts itself!
- Only two sites publish provenance information about the original source.
- No evidence is given for how the estimates were made.
  - The last official census appears to have been taken in 1990!

Source: Peter Bunemann
Definition: Data Conflict

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

<table>
<thead>
<tr>
<th>DB1</th>
<th>DB2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chris Miller</td>
<td>Christian Miller</td>
</tr>
<tr>
<td>12/20/1982</td>
<td>2/20/1982</td>
</tr>
<tr>
<td>Bardon Street, Melville</td>
<td>7 Bardon St., Melville</td>
</tr>
</tbody>
</table>

Reasons for data conflicts:

- Data creation: Typos, measurement errors, erroneous OCR
- Data representation: Different coding of values (“Mrs.” vs. “2”)
- Data currency: Different points in time, missing updates
- Data semantics: Different definitions of concepts (like population)
- Data integration: Wrong data translation or identity resolution
- Actual disagreement of data providers: Subjective attributes (like cuteness)
Handling Data Conflicts

### Possible Conflict Handling Strategies from Bleiholder and Naumann [2006]

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Classification</th>
<th>Short Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass it on</td>
<td>ignoring</td>
<td>Escalates conflicts to user or application</td>
</tr>
<tr>
<td>Consider all possibilities</td>
<td>ignoring</td>
<td>Creates all possible value combinations</td>
</tr>
<tr>
<td>Take the information</td>
<td>avoiding, instance based</td>
<td>Prefers values over null values</td>
</tr>
<tr>
<td>No Gossiping</td>
<td>avoiding, instance based</td>
<td>Returns only consistent tuples</td>
</tr>
<tr>
<td>Trust your friends</td>
<td>avoiding, metadata based</td>
<td>Takes the value of a preferred source</td>
</tr>
<tr>
<td>Cry with the wolves</td>
<td>resolution, instance based, deciding</td>
<td>Takes the most often occurring value</td>
</tr>
<tr>
<td>Roll the dice</td>
<td>resolution, instance based, deciding</td>
<td>Takes a random value</td>
</tr>
<tr>
<td>Meet in the middle</td>
<td>resolution, instance based, mediating</td>
<td>Takes an average value</td>
</tr>
<tr>
<td>Keep up to date</td>
<td>resolution, metadata based, deciding</td>
<td>Takes the most recent value</td>
</tr>
</tbody>
</table>

- Main Types of Strategies:
  - Ignoring
  - Avoiding
  - Resolution
**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
  - by applying conflict handling strategies to type *avoiding* or *resolution*

- But what does high data quality 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 check information more tightly when you invest 1 million Euro

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: “Use only information that has been published in the last week” or “Disbelieve everything a vendor says about its competitor”

- **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

- Provenance Metadata
- Ratings, other knowledge
- Data Quality Assessment
- Apply Conflict Resolution Function
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.

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

Source: W3C PROV Specification
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 2017 20:54:26 GMT
Server: Apache/1.3.6 (UNIX)
Last-Modified: Mon, 06 Dec 2016 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?
Dublin Core

- 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

```xml
# 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 "2013-07-13"^^xsd:date ;

# 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 ;
  ...
```
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</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>
...</prov:document>
```
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**

<table>
<thead>
<tr>
<th>Key</th>
<th>Name</th>
<th>Street</th>
<th>ProvID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1425</td>
<td>Dr. Mark Smith</td>
<td>14 Main Street</td>
<td>001</td>
</tr>
<tr>
<td>4217</td>
<td>Mark Smith</td>
<td>12 Main St.</td>
<td>002</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Provenance Table**

<table>
<thead>
<tr>
<th>ProvID</th>
<th>Source</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Physicians with **Value-Level Provenance**

<table>
<thead>
<tr>
<th>Key</th>
<th>Attribute</th>
<th>Value</th>
<th>ProvID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1425</td>
<td>Name</td>
<td>Dr. Mark Smith</td>
<td>001</td>
</tr>
<tr>
<td>1425</td>
<td>Name</td>
<td>Mark Smith</td>
<td>002</td>
</tr>
<tr>
<td>1425</td>
<td>Street</td>
<td>14 Main Street</td>
<td>001</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
XML Data Model

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

```xml
<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">
  ...
</provenance>
```
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

---

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

<table>
<thead>
<tr>
<th>Customer</th>
<th>CNr</th>
<th>Name</th>
<th>BirthDate</th>
<th>Age</th>
<th>Sex</th>
<th>Phone</th>
<th>Zip</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1234</td>
<td>Kuhn, Mark</td>
<td>18.2.1980</td>
<td>31</td>
<td>NULL</td>
<td>NULL</td>
<td>98693</td>
</tr>
<tr>
<td></td>
<td>1234</td>
<td>Anne Will</td>
<td>3.2.1970</td>
<td>47</td>
<td>m</td>
<td>768-4511</td>
<td>55555</td>
</tr>
<tr>
<td></td>
<td>1235</td>
<td>Mark Kuhn</td>
<td>27</td>
<td>m</td>
<td>567-3211</td>
<td>98693</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Address</th>
<th>Zip</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1510</td>
<td>Potsdam</td>
</tr>
<tr>
<td></td>
<td>98693</td>
<td>Postdam</td>
</tr>
<tr>
<td></td>
<td>98766</td>
<td>BRD</td>
</tr>
</tbody>
</table>

Uniqueness
Contradictions
Missing Values
Referential Integrity
Value Coding
Incorrect Values
Outdated Value
Typos
Value not understandable
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 spare 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:

Fitness for use

Accuracy, Objectivity, Believability, Reputation, Accessibility, Security, Relevance, Value-Added, Timeliness, Completeness, Amount of Data, Interpretability, Understandability, Consistency, Concise Representation

179 Dimensions
Focus: The Complete Data Usage Situation

User
- Relevancy
- Understandability
- Believability
- ...

Process
- Accessibility
- Response Time
- Price
- ...

Data
- Accuracy
- Timeliness
- Completeness
- ...

Diagram showing the process of data usage with focus on various aspects such as accuracy, timeliness, completeness, accessibility, response time, price, relevancy, understandability, and believability.
<table>
<thead>
<tr>
<th>Category</th>
<th>IQ Criteria</th>
<th>TDQM</th>
<th>MBIS</th>
<th>Weikum</th>
<th>DWQ</th>
<th>SCOUG</th>
<th>Chen</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content-related Criteria</strong></td>
<td>Accuracy</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Documentation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Relevancy</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Value-Added</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Completeness</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interpretability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Technical Criteria</strong></td>
<td>Timeliness</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Reliability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Latency</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Performability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Response time</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Security</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accessibility</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Price</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Customer Support</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Intellectual Criteria</strong></td>
<td>Believability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Reputation</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Objectivity</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Instantiation related Criteria</strong></td>
<td>Verifiability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amount of data</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Understandability</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Concise represent.</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Consistent represent.</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

- **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 presentation 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
**3.2. Data Quality Assessment**

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

Applicability of 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

- **Information Consumer**
  - Background Information
  - Uses
  - Ratings

- **Meta-Information**
  - Piece of Information to be assessed

- **Information Provider**
  - Background Information
  - Provides
  - Ratings

- **Other Actor**
  - Background Information
  - Ratings
Examples of Data Quality Assessment Metrics

Number of DQ classification schemata containing specific quality dimensions:

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

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:
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 an 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
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**
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
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 relies on 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
  - 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
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
Explanation about an Assessment Decision

The Triple:

Siemens Share positive analyst report Siemens agrees partnership with Novell, new SUSE. Siemens Business Service (SBS) The IT service arm of German technology conglomerate Siemens – SBS – SBS). 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 has been adding 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 ten 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 (ibm: IBM – news – people), among others, its advocates, who include big businesses and government departments, argue it is cheaper, simpler and more secure than Windows.

fulfills 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.
Example Explanation

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.
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.

<table>
<thead>
<tr>
<th></th>
<th>DB1</th>
<th>DB2</th>
<th>New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Chris Miller</td>
<td>Christian Miller</td>
<td>Christian Miller</td>
</tr>
<tr>
<td>DOB</td>
<td>NULL</td>
<td>12/20/1982</td>
<td>12/20/1982</td>
</tr>
<tr>
<td>Address</td>
<td>7 Bardon St., Melville</td>
<td>Bardon St., Melville</td>
<td>7 Bardon St., Melville</td>
</tr>
</tbody>
</table>

- **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**: Take information
- **Subsumed records**: Remove subsumed
- **Identical records**: Remove duplicate
- **Conflicting records**: Apply Conflict Resolution Function

**Examples:**

- **Complementary records**
  - `a, b, c, -` → `a, b, c, d`
  - `a, b, - , d` → `a, b, c, -`

- **Subsumed records**
  - `a, b, c, -` → `a, b, c, -`
  - `a, b, -, -` → `a, b, -, -`

- **Conflicting records**
  - `a, b, c, -` → `a, b, f(b,e), c, d`
  - `a, e, -, d`
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

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

<table>
<thead>
<tr>
<th>Function</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favor Sources</td>
<td>Take first non-null value in particular order of sources</td>
</tr>
<tr>
<td></td>
<td>Example: Use Eurostat for GDP, alternatively use Wikipedia</td>
</tr>
<tr>
<td>MostRecent</td>
<td>Choose most recent (up-to-date) value</td>
</tr>
<tr>
<td></td>
<td>Example: Address, NumChildren</td>
</tr>
<tr>
<td>MostActive</td>
<td>Choose value that is most often accessed/edited</td>
</tr>
<tr>
<td></td>
<td>Example: Prefer Wikipedia page with more edits</td>
</tr>
<tr>
<td>FavorSources basedOnRatings</td>
<td>Calculate quality of sources from ratings, take value from source with highest score or all values from sources with scores above specific threshold</td>
</tr>
<tr>
<td>MaxIQ</td>
<td>Choose the value with the highest quality score. Score might cover multiple quality dimensions, e.g. timeliness and believability of a source</td>
</tr>
<tr>
<td>TopkIQ</td>
<td>Choose the top K values with the highest quality scores</td>
</tr>
<tr>
<td>….</td>
<td>….</td>
</tr>
</tbody>
</table>
Example: Complete Conflict Resolution Heuristic

- **amazon.com**
  - 8/31/2013
  - 0766607194
  - H. Melville
  - Moby Dick
  - $3.98
  - Review

- **bn.com**
  - 7/20/2012
  - 0766607193
  - Herman Melville
  - Moby Dick
  - $5.99

| Random | max length | FavorSources (amazon.com) | MostRecent | Union |
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.

\[
density_{column} = \frac{|\text{non-NULL values in column}|}{|\text{rows in table}|}
\]

\[
density_{table} = \frac{|\text{non-NULL values in table}|}{|\text{columns}|*|\text{rows}|}
\]

- 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 \(C_1'\)
A data set is consistent if it is free of conflicting information.

\[
\text{consistency}_{\text{Column}} = \frac{|\text{non-conflicting values in column}|}{|\text{real-world entities described}|}
\]

\[
\text{consistency}_{\text{Table}} = \frac{|\text{non-conflicting values in table}|}{|\text{columns}|*|\text{real-world entities described}|}
\]

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

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

\[
\text{accuracy} = \frac{|\text{correct values}|}{|\text{all values}|}
\]

\[
\text{error rate} = 1 - \text{accuracy}
\]

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

Prototype developed at Hasso Plattner Institute

![Fuzzy Fuz!on](image)

<table>
<thead>
<tr>
<th>Rule Matrix</th>
<th>Firstname</th>
<th>Lastname</th>
<th>Street</th>
<th>housenumber</th>
<th>post code</th>
<th>city</th>
<th>ignore</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>66105</td>
<td>66111</td>
<td>56872</td>
<td>66404</td>
<td>63121</td>
<td>71285</td>
<td>100000</td>
<td>73936</td>
</tr>
<tr>
<td>Null values</td>
<td>5671</td>
<td>6402</td>
<td>616</td>
<td>16746</td>
<td>12208</td>
<td>5643</td>
<td>0</td>
<td>26064</td>
</tr>
<tr>
<td>Case Variance</td>
<td>10835</td>
<td>12745</td>
<td>14563</td>
<td>0</td>
<td>0</td>
<td>11330</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>7095</td>
<td>1170</td>
<td>8256</td>
<td>16850</td>
<td>12364</td>
<td>942</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tokenization</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Substrings</td>
<td>2122</td>
<td>2091</td>
<td>1088</td>
<td>0</td>
<td>12307</td>
<td>1701</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Dominance</td>
<td>2170</td>
<td>2424</td>
<td>2883</td>
<td>0</td>
<td>2434</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Low edit distance</td>
<td>5913</td>
<td>7032</td>
<td>7101</td>
<td>0</td>
<td>0</td>
<td>6664</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Global dominance</td>
<td>88</td>
<td>0</td>
<td>762</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Undefined</td>
<td>1</td>
<td>0</td>
<td>359</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Actions**

- Fusionsregel(n) anzeigen/erzeugen
- Übersicht
- Nur aktuelle Markierung anzeigen

**Selected Rules**

- Spalten
  - Firstname
  - Lastname
- Konflikttypen
  - Low edit distance
- Primäre Konfliktlösung
  - Vote
  - Minimum fraction of solution (in %): 50
- Sekundäre Konfliktlösung
  - First
Manual Fusion of Record Groups in Fuz!on

### Fuzzy Fuz!on

**Manual Fusion**

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

**21. Group:**

<table>
<thead>
<tr>
<th>Firstname</th>
<th>Lastname</th>
<th>Street</th>
<th>housenumber</th>
<th>postcode</th>
<th>city</th>
<th>ignore</th>
<th>phone</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manh The</td>
<td>Knaut</td>
<td>Wiedeberger Weg</td>
<td>37</td>
<td>24943</td>
<td>Flensburg</td>
<td>19280312</td>
<td>0461</td>
</tr>
<tr>
<td>Manh The</td>
<td>KNAUT</td>
<td>Wiedeberger Weg</td>
<td>37</td>
<td>24943</td>
<td>Flensburg</td>
<td>19280312</td>
<td>0461</td>
</tr>
<tr>
<td>Manh The</td>
<td>Knaut</td>
<td>WIEDEBERGER WEG</td>
<td>37</td>
<td>24943</td>
<td>Flensburg</td>
<td>19280312</td>
<td>0461</td>
</tr>
</tbody>
</table>

**Merge**  **Save Configurations**
4.5 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?
- from the specific language edition?
- most recent?
- inserted by most trusted author?
- edited most times?
- combination of the above?

Population of Mannheim in 8 DBpedia language editions

Mannheim population

Total

"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

<table>
<thead>
<tr>
<th>ru:Mannheim:populationTotal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>nl:Mannheim:populationTotal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
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

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

* “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

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

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 DBedia
5. References

- **Provenance**

- **Data Quality**
  - Ziawasch et al.: Detecting data errors: where are we and what needs to be done? VLDB 2016.
  - Aggarwal: Managing and Mining Uncertain Data. Springer, 2010. (Focus: probabilistic models)
References

- Data Fusion

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