Data Quality Management for Big Data

Christoph Quix

EBISS Summer School, July 4, 2017
About Me

Diploma, PhD, Habilitation from RWTH
Data Quality: Motivating Example
Soccer player database for the Confed Cup 2017

<table>
<thead>
<tr>
<th>Player</th>
<th>Bdate</th>
<th>SSN</th>
<th>Age</th>
<th>Club_ID</th>
<th>phone</th>
<th>City</th>
<th>PLZ</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>12.02.90</td>
<td>123456</td>
<td>22</td>
<td>127</td>
<td>9999</td>
<td>Leipzig</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Peter Miller</td>
<td>30.13.88</td>
<td>123456</td>
<td>22</td>
<td>17</td>
<td></td>
<td>Dresden</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I. Vieth</td>
<td>26.11.92</td>
<td>073456</td>
<td>20</td>
<td>15</td>
<td>+49 341 808060</td>
<td>Germany</td>
<td></td>
<td></td>
</tr>
<tr>
<td>John Smith</td>
<td>12.12.90</td>
<td>123456</td>
<td>22</td>
<td>127</td>
<td>9999</td>
<td>Leipzig</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Question:**
What is the quality of this database?

**Queries:**
- Is there a player from a team in Poland?
- How many players are there?
- What is the phone number of Peter Miller?
- What is the meaning of PLZ?
- What is the meaning of A/B experience?
Observations

- Data quality is subjective
  - Depends on application requirements, context, user, ...

- Data quality can be measured without knowing the true values
  - Examine the intrinsic properties of the data

- Data quality is not only aspect of the data
  - Metadata and data processing systems also affect data quality

- Data quality management is more than data cleaning
  - Data cleaning is one aspect of DQM, but there is much more
Overview

- Definitions and Terminology
- Data Quality Methodologies
- Data Quality Problems in Big Data
- Empirical Explanations and Data Glitches
- Data Quality Management in Data Streams
- Data Quality Management in Data Lakes
- Data Quality Management in Data Integration Tools
- Conclusion
Quality Perspectives

- **Product-oriented**
  - Based on features of the product

- **Application-oriented**
  - Fulfills requirements of users

- **Process-oriented**
  - Compliance of production process with specifications

- **Value-oriented**
  - Price-performance ratio
Data Quality
Definitions

- Degree to which data meets user requirements [ISO2]
- As exactly as possible [E99]
- Degree to which the characteristics of data satisfy stated and implied needs when used under specified conditions [ISO1]

Data Quality
Definitions

Holistic Definition

- Ability of the information system to provide data according to the requirements of the organization
- Data as a product: fitness for use [Redman, 1997]

Data Quality Management

Definitions

Data quality management is more than just data cleaning

- Data quality problems are not caused only by the data itself, but also by the way data is processed, described, interpreted, analyzed, visualized, ...

- “Coordinated activities to direct and control an organization with regard to data quality” [ISO2]

Data Quality & Data Integration

Data quality problems are often revealed in data integration projects

Data in source systems has been collected for a specific application and in a specific context
⇒ DQ might be fine in this application context

**Data integration**
⇒ Data is used in a different application context
⇒ Data does not fulfill the requirements of the new application

Data Quality = „fitness for use“ ⇒ the use changes in data integration!
Cross Industry Standard Process for Data Mining (CRISP-DM)

Reference model for data mining and business intelligence processes

Data quality needs to be considered throughout the process, but it especially during „Data Understanding“ and „Data Preparation“ steps

Only data with good quality can lead to valuable analytics results (garbage in ➔ garbage out)
Motivation
Causes for data quality problems

**Typographical errors and non-conforming data:**
Plain errors in the data

**Information obfuscation:**
False information is given on purpose

**Renegade IT and spreadmarts:**
Data snapshots are created from central IT systems and used in subsequent business decisions

**Corporate mergers or reorganizations:**
Existing data is used in a new context

**Changing or new requirements:**
New requirements might not be satisfied by existing data

**Hidden Code or Loss of expertise:**
The interpretation or semantics of data is hidden in the code or only known to a few people
Data-oriented vs. Process oriented DQ Management

- **Strategic Data Quality Management**
  - **Data oriented**
    - Improvement by data cleaning
    - No deep analysis of the cause of DQ problems
    - Reactive
  - **Process oriented**
    - Improvement by optimizing/adapting processes
    - Considers whole data management process
    - Preventive
Data-oriented (reactive) Data Quality Management

- Data validation
  - Manual and visual data validation
  - Rule based data validation
- Outlier detection
- Anomaly detection
- Automatic vs. manual procedures in data cleaning
- Entity resolution
- ...

Fraunhofer FIT

RWTH AACHEN UNIVERSITY
Data Quality Management
A data cleaning example

Removing invalid Email records from a column.
- Audit data: check valid emails against predefined patterns
- Choose method: delete lines with invalid cells
- Apply method: execute method in program

Example from Talend Data Preparation

Invalid emails will still be acquired!!
Process-oriented (preventive) Data Quality Management

- What are the data quality problems?
- Where do we need to improve the data quality?
- How do you define data quality?
- What are the goals in data quality management?
- How can you measure data quality?
- How can you improve data quality?
- ...
Data Quality Management
A process-oriented example

Problem: customers do not provide valid information when they register to our web shop

Current sign-up form for web-shop
- Name
- Address
- Birthdate
- Phone
- Email
- ...

Improved sign-up form for web-shop
1. Step: please provide your email:
   - Email
2. Step: please provide optional information
   - Address
   - Birthdate
   - Phone
Data Quality Goals represent the DQ Requirements

- **Goal**: Extract more data from source systems X and Y
- **Characteristic (Dimension)**: Completeness
- **Measurement Method**: Ratio of extracted data and available data
- **Observation**: Number of fields (or rows) in integrated DB divided by number of fields (rows) in source systems
- **Result**: Time per extracted row from sources X and Y

- **Goal**: Reduce time for transformation process
- **Characteristic (Dimension)**: Efficiency
- **Measurement Method**: Time needed for extraction process
- **Observation**: Time per extracted row from sources X and Y

- **Goal**: Reduce effort for adoption of ETL processes
- **Characteristic (Dimension)**: Robustness
- **Measurement Method**: Count of situations where ETL process had to be changed
- **Observation**: Number of versions of ETL process algorithms per week

**Fraunhofer FIT**

**RWTH Aachen University**
Objects for measuring and improving Data Quality

Data Quality Objects

- Systems
- Data Models
- Data

DQ Objects

- Performance
- Security Level
- Completeness
- Correctness
- Accuracy

DQ Characteristics

- Coverage of the data model
- Correct representation of the real world entities
- Accuracy of the model elements wrt. to the real world
- Number of data items
- Number of digits of measured values

DQ Measures

- Response Times
- ISO certification available?

Fraunhofer FIT

RWTH Aachen University
Processes relevant for Improving Data Quality

Data quality can be defined in development & operation of an information system

**Development**
- Requirements Analysis
- Design
- Implementation
- Testing

**Operation**
- Data collection
- Data processing
- Data usage
- Administration and Maintenance
Scenario

You are integrating data about countries from different sources in the web (e.g., EU, UN, OECD, Wikipedia, ...). The sources contain information about the population, unemployment rates for various years.

- Which data quality problems might occur in this context?
- Define 2-3 data quality goals to address these problems.
- How can you measure the data quality in order to prove whether the DQ goals have been achieved?
- Which counteraction can be applied to improve the data quality?
Sample Answers for the Scenario

- **Inconsistent country code (DE=GER=D, B=BEL=BE, ...)**
  - Goal: All country codes should be encoded according to standard X.
  - Characteristic: Consistency, understandability
  - Metric: Number of country codes which do not conform to the standard
  - Counteraction: Transform all country codes into the required standard

- **Population or unemployment rate is not given for a year/country**
  - Goal: The information about population and unemployment should be complete
  - Characteristic: Completeness
  - Metric: Number of Null values
  - Counteraction: Integrate data from another source

- **Population/unemployment data is inconsistent between different data sources**
  - Goal: Provide consistent and correct information about population/unemployment
  - Characteristic: Consistency, trustworthiness
  - Metric: Variance between values, number of different values
  - Counteraction: Preference for a source, averaging between multiple sources
Overview

- Definitions and Terminology
- **Data Quality Methodologies**
- Data Quality Problems in Big Data
- Empirical Explanations and Data Glitches
- Data Quality Management in Data Streams
- Data Quality Management in Data Lakes
- Data Quality Management in Data Integration Tools
- Conclusion
TDQM
Total Data Quality Management

- Find ways to improve the data quality (includes data cleaning)
- What is the cause for insufficient data quality?
- Relevant objects
- Requirements & goals
- Metrics
- Execute metrics for relevant objects
- Store the results (in a metadata management system)

DWQ
Data Warehouse Quality (EU Project 1997-2000)

Holistic approach to data integration and data quality management in data warehouse systems
DWQ Framework for DW Metadata

Conceptual Perspective

Logical Perspective

Physical Perspective

Client Model

OLAP

Client Data Store

Transportation Agent

Enterprise Model

Observation

Aggregation/Customization

DW Data Store

Transportation Agent

Operational Department Model

OLTP

Source Data Store

Source Schema

Wrapper
**DWQ Data Quality Model**

Implemented as a metadata model in a metadata repository

Quality metrics could be implemented partially as queries on the (meta)data repository

Employs the Goal-Question-Metric approach from Software Quality Management
Other Data Quality Methodologies

Comprehensive survey by
Batini et al., ACM Computing Surveys, Vol. 41, No. 3, 2009

<table>
<thead>
<tr>
<th>Step/Meth Acronym</th>
<th>Data Analysis</th>
<th>DQ Requirement Analysis</th>
<th>Identification of Critical Areas</th>
<th>Process Modeling</th>
<th>Measurement of Quality</th>
<th>Extensible to Other Dimensions and Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDQM</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Fixed</td>
</tr>
<tr>
<td>DWQ</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Open</td>
</tr>
<tr>
<td>TIQM</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Fixed</td>
</tr>
<tr>
<td>AIMQ</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Fixed</td>
</tr>
<tr>
<td>CIHI</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Fixed</td>
</tr>
<tr>
<td>DQA</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Open</td>
</tr>
<tr>
<td>IQM</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Open</td>
</tr>
<tr>
<td>ISTAT</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Fixed</td>
</tr>
<tr>
<td>AMEQ</td>
<td>+</td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Open</td>
</tr>
<tr>
<td>COLDQ</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Fixed</td>
</tr>
<tr>
<td>DaQuinCIS</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Open</td>
</tr>
<tr>
<td>QAFD</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Fixed</td>
</tr>
<tr>
<td>CDQ</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>Open</td>
</tr>
</tbody>
</table>
## Diverting definitions for DQ Characteristics (DQ Dimensions)

[Batini et al., 2009]

<table>
<thead>
<tr>
<th>Reference</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wand and Wang 1996</td>
<td>Ability of an information system to represent every meaningful state of a real world system</td>
</tr>
<tr>
<td>Wang and Wand 1996</td>
<td>Extent to which data are of sufficient breadth, depth, and scope for the task at hand</td>
</tr>
<tr>
<td>Redman 1996</td>
<td>Degree to which values are included in a data collection</td>
</tr>
<tr>
<td>Jarke et al. 1995</td>
<td>Percentage of real-world information entered in data sources and/or data warehouse</td>
</tr>
<tr>
<td>Bovee et al. 2001</td>
<td>Information having all required parts of an entity’s description</td>
</tr>
<tr>
<td>Naumann 2002</td>
<td>Ratio between the number of non-null values in a source and the size of the universal relation</td>
</tr>
<tr>
<td>Liu and Chi 2002</td>
<td>All values that are supposed to be collected as per a collection theory</td>
</tr>
</tbody>
</table>
DQ Characteristics (DQ Dimensions)

- Many different definitions for DQ Dimensions
- Hundreds of DQ Dimensions (Batini et al. enumerate ~160)
- Ignore these differences, concentrate on the definitions relevant for your context and the metrics (DQ is subjective!)

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Data Quality Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>TDQM</td>
<td>Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of manipulation, Value added, Free of error, Interpretability, Objectivity, Relevance, Reputation, Security, Timeliness, Understandability</td>
</tr>
<tr>
<td>DWQ</td>
<td>Correctness, Completeness, Minimality, Traceability, Interpretability, Metadata Evolution, Accessibility (System, Transactional, Security), Usefulness (Interpretability), Timeliness (Currency, Volatility), Responsiveness, Completeness, Credibility, Accuracy, Consistency, Interpretability</td>
</tr>
<tr>
<td>TIQM</td>
<td>Inherent dimensions: Definition conformance (consistency), Completeness, Business rules conformance, Accuracy (to surrogate source), Accuracy (to reality), Precision, Nonduplication, Equivalence of redundant data, Concurrency of redundant data, Pragmatic dimensions: accessibility, timeliness, contextual clarity, Derivation integrity, Usability, Rightness (fact completeness), cost</td>
</tr>
<tr>
<td>AIMQ</td>
<td>Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of operation, Freedom from errors, Interpretability, Objectivity, Relevancy, Reputation, Security, Timeliness, Understandability</td>
</tr>
<tr>
<td>CHII</td>
<td>Dimensions: Accuracy, Timeliness, Comparability, Usability, Relevance, Consistency: Over-coverage, Under-coverage, Simple/correlated response variance, Reliability, Collection and capture, Unit/Item non response, Edit and imputation, Processing, Estimation, Timeliness, Comprehensiveness, Integration, Standardization, Equivalence, Linkage ability, Product/Historical comparability, Accessibility, Documentation, Interpretability, Adaptability, Value</td>
</tr>
<tr>
<td>DQA</td>
<td>Accessibility, Appropriate amount of data, Believability, Completeness, Freedom from errors, Consistency, Concise Representation, Relevance, Ease of manipulation, Interpretability, Objectivity, Reputability, Security, Timeliness, Understandability, Value added</td>
</tr>
<tr>
<td>IQM</td>
<td>Accessibility, Consistency, Timeliness, Concision, Maintainability, Currency, Applicability, Convenience, Speed, Comprehensiveness, Clarity, Accuracy, Traceability, Security, Correctness, Interactivity</td>
</tr>
<tr>
<td>ISTAT</td>
<td>Accuracy, Completeness, Consistency</td>
</tr>
<tr>
<td>AIMEQ</td>
<td>Consistent representation, Interpretability, Case of understanding, Concise representation, Timeliness, Completeness Value added, Relevance, Appropriateness, Meaningfulness, Lack of confusion, Arrangement, Readable, Reasonability, Precision, Reliability, Freedom from bias, Data Deficiency, Design Deficiency, Operation, Deficiencies, Accuracy, Cost, Objectivity, Believability, Reputation, Accessibility, Correctness, Unambiguity, Consistency</td>
</tr>
<tr>
<td>COLDQ</td>
<td>Schema: Clarity of definition, Comprehensiveness, Flexibility, Robustness, Essentialness, Attribute granularity, Precision of domains, Homogeneity, Identifiability, Obtainability, Relevance, Simplicity/Complexity, Semantic consistency, Syntactic consistency, Data: Accuracy, Null Values, Completeness, Consistency, Currency, Timeliness, Agreement of Usage, Stewardship, Ubiquity, Presentation: Appropriateness, Correct Interpretation, Flexibility, Format precision, Portability, Consistency, Use of storage, Information policy: Accessibility, Metadata, Privacy, Security, Redundancy, Cost</td>
</tr>
<tr>
<td>DauQuinCIS</td>
<td>Accuracy, Completeness, Consistency, Currency, Trustworthiness</td>
</tr>
<tr>
<td>QAFD</td>
<td>Syntactic/Semantic accuracy, Internal/External consistency, Completeness, Currency, Uniqueness</td>
</tr>
<tr>
<td>CDQ</td>
<td>Schema: Correctness with respect to the model, Correctness with respect to Requirements, Completeness, Pertinence, Readability, Normalization, Data: Syntactic/Semantic Accuracy, Semantic Accuracy, Completeness, Consistency, Currency, Timeliness, Volatility, Completability, Reputation, Accessibility, Cost</td>
</tr>
</tbody>
</table>

[Batini et al., 2009]
DQ Metrics

Several proposals for DQ metrics have been made

Some can be computed automatically, some require user input or knowledge of „correct“ values

⇒ not scalable

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Name</th>
<th>Metrics Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Acc1</td>
<td>Syntactic accuracy: it is measured as the distance between the value stored in the database and the correct one</td>
</tr>
<tr>
<td></td>
<td>Acc2</td>
<td>Number of delivered accurate tuples</td>
</tr>
<tr>
<td></td>
<td>Acc3</td>
<td>User Survey - Questionnaire</td>
</tr>
<tr>
<td>Completeness</td>
<td>Compl1</td>
<td>Completeness = Number of not null values/total number of values</td>
</tr>
<tr>
<td></td>
<td>Compl2</td>
<td>Completeness = Number of tuples delivered/Expected number</td>
</tr>
<tr>
<td></td>
<td>Compl3</td>
<td>Completeness of Web data = (T_{max} - T_{current})^x (Completeness_{max}^y - Completeness_{current})^z</td>
</tr>
<tr>
<td></td>
<td>Compl4</td>
<td>User Survey - Questionnaire</td>
</tr>
<tr>
<td>Consistency</td>
<td>Cons1</td>
<td>Consistency = Number of consistent values/number of total values</td>
</tr>
<tr>
<td></td>
<td>Cons2</td>
<td>Number of tuples violating constraints, number of coding differences</td>
</tr>
<tr>
<td></td>
<td>Cons3</td>
<td>Number of pages with style guide deviation</td>
</tr>
<tr>
<td></td>
<td>Cons4</td>
<td>User Survey - Questionnaire</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Time1</td>
<td>Timeliness = (max (0; 1-Currency/Volatility))^x</td>
</tr>
<tr>
<td></td>
<td>Time2</td>
<td>Percentage of process executions able to be performed within the required time frame</td>
</tr>
<tr>
<td></td>
<td>Time3</td>
<td>User Survey - Questionnaire</td>
</tr>
<tr>
<td>Currency</td>
<td>Curr1</td>
<td>Currency = Time in which data are stored in the system - time in which data are updated in the real world</td>
</tr>
<tr>
<td></td>
<td>Curr2</td>
<td>Time of last update</td>
</tr>
<tr>
<td></td>
<td>Curr3</td>
<td>Currency = Request time- last update</td>
</tr>
<tr>
<td></td>
<td>Curr4</td>
<td>Currency = Age + (Delivery time- Input time)</td>
</tr>
<tr>
<td></td>
<td>Curr5</td>
<td>User Survey - Questionnaire</td>
</tr>
</tbody>
</table>

[Batini et al., 2009]
Quality-oriented Data Integration

Basic Assumption:
Data is redundant ➔ Multiple ways to solve

[Calvanese et al., 01]

[Naumann, 02]

[Pottinger & Halevy, 01]
Computation of DQ for Integrated Query

For each predicate of the query, measure the relevant quality factors

Assign weights to query predicates and quality factors

→ Integrate all values into one result

<table>
<thead>
<tr>
<th>Teilziel</th>
<th>$w_r$</th>
<th>Umschreibung 1</th>
<th>Umschreibung 2</th>
<th>Umschreibung 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>PrivateCustomer</td>
<td>0,10</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>ID</td>
<td>0,05</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>DOB</td>
<td>0,10</td>
<td>5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Name</td>
<td>0,05</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>located-in</td>
<td>0,10</td>
<td>2</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Location</td>
<td>0,05</td>
<td>2</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Country agreement</td>
<td>0,20</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Service</td>
<td>0,05</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Contract</td>
<td>0,05</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Code</td>
<td>0,10</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Type</td>
<td>0,10</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>$v_{join}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>skaliert nach Def.</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>alternative Skal.</td>
<td>0,5</td>
<td>0,90</td>
<td>0,81</td>
<td>0,87</td>
</tr>
<tr>
<td>Qualitätswerte</td>
<td>0,2</td>
<td>0,784</td>
<td>0,8</td>
<td>0,87</td>
</tr>
</tbody>
</table>

[Integrated Query]

[Quix, 2003 (sorry, in German)]
Overview

- Definitions and Terminology
- Data Quality Methodologies
- Data Quality Problems in Big Data
- Empirical Explanations and Data Glitches
- Data Quality Management in Data Streams
- Data Quality Management in Data Lakes
- Data Quality Management in Data Integration Tools
- Conclusion
### Data Cleaning of Player Profile Data

<table>
<thead>
<tr>
<th>Player</th>
<th>Bdate</th>
<th>SSN</th>
<th>Age</th>
<th>Club_ID</th>
<th>phone</th>
<th>City</th>
<th>ZIP</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>12.02.70</td>
<td>123456</td>
<td>22</td>
<td>127</td>
<td>9999</td>
<td>Liipzig</td>
<td>45257</td>
<td>B</td>
</tr>
<tr>
<td>Peter Miller</td>
<td>30.13.68</td>
<td>123456</td>
<td>22</td>
<td>17</td>
<td></td>
<td>Dresden</td>
<td>01099</td>
<td>A</td>
</tr>
<tr>
<td>I. Vieth</td>
<td></td>
<td>073456</td>
<td>20</td>
<td>15</td>
<td>+49 341 808060</td>
<td>Germany</td>
<td>04109</td>
<td>B</td>
</tr>
<tr>
<td>John Smith</td>
<td>12.12.70</td>
<td>123456</td>
<td>22</td>
<td>127</td>
<td>9999</td>
<td>Liipzig</td>
<td>45257</td>
<td>B</td>
</tr>
</tbody>
</table>

**A.** Assume that you have to work with above table containing player profile data. The table contains obvious errors. Mark and assign them to one of the following types of errors.

1. Illegal value
2. Violated attribute dependency
3. Uniqueness violation
4. Missing value
5. Misspelling
6. Kryptic value
7. Embedded value
8. Misfielded value
9. Word transposition
10. Duplicate
11. Contradicting record
12. Wrong Reference
Data Cleaning of Player Profile Data

<table>
<thead>
<tr>
<th>Player</th>
<th>Bdate</th>
<th>SSN</th>
<th>Age</th>
<th>Club_ID</th>
<th>phone</th>
<th>City</th>
<th>ZIP</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>12.02.70</td>
<td>123456</td>
<td>22</td>
<td>127</td>
<td>9999</td>
<td>Liipzig</td>
<td>45257</td>
<td>B</td>
</tr>
<tr>
<td>Peter Miller</td>
<td>30.13.68</td>
<td>123456</td>
<td>22</td>
<td>17</td>
<td></td>
<td>Dresden</td>
<td>01099</td>
<td>A</td>
</tr>
<tr>
<td>I. Vieth</td>
<td>26.11.72</td>
<td>073456</td>
<td>20</td>
<td>15</td>
<td>+49 341 808060</td>
<td>Germany</td>
<td>04109</td>
<td>B</td>
</tr>
<tr>
<td>John Smith</td>
<td>12.12.70</td>
<td>123456</td>
<td>22</td>
<td>127</td>
<td>9999</td>
<td>Liipzig</td>
<td>45257</td>
<td>B</td>
</tr>
</tbody>
</table>

B. Mark an example for an error:

- that can be found by analyzing a single attribute
- that can be found by analyzing multiple attributes
Data Cleaning of Player Profile Data

<table>
<thead>
<tr>
<th>Player</th>
<th>Bdate</th>
<th>SSN</th>
<th>Age</th>
<th>Club_ID</th>
<th>phone</th>
<th>City</th>
<th>ZIP</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>12.02.70</td>
<td>123456</td>
<td>22</td>
<td>127</td>
<td>9999</td>
<td>Liipzig</td>
<td>45257</td>
<td>B</td>
</tr>
<tr>
<td>Peter Miller</td>
<td>30.13.68</td>
<td>123456</td>
<td>22</td>
<td>17</td>
<td></td>
<td>Dresden</td>
<td>01099</td>
<td>A</td>
</tr>
<tr>
<td>I. Vieth</td>
<td>26.11.72</td>
<td>073456</td>
<td>20</td>
<td>15</td>
<td>+49 341 808060</td>
<td>Germany</td>
<td>04109</td>
<td>B</td>
</tr>
</tbody>
</table>

C. Which of the shown errors could you identify with a frequency analysis?

D. Which of the errors could be detected automatically and efficiently in a Big Data scenario?
## Data Cleaning of Player Profile Data

### Possible solution

<table>
<thead>
<tr>
<th>Player</th>
<th>Bdate</th>
<th>SSN</th>
<th>Age</th>
<th>Club ID</th>
<th>Phone</th>
<th>City</th>
<th>Experience</th>
</tr>
</thead>
<tbody>
<tr>
<td>John Smith</td>
<td>12.02.70</td>
<td>123456</td>
<td>22</td>
<td>127</td>
<td>9999</td>
<td>Liipzig</td>
<td>B</td>
</tr>
<tr>
<td>Peter Miller</td>
<td>30.13.70</td>
<td>123456</td>
<td>17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>J.Smith 12.02.70</td>
<td>123456</td>
<td>22</td>
<td></td>
<td>127</td>
<td>-</td>
<td>Germany</td>
<td>B</td>
</tr>
<tr>
<td>Leipzig</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>John Smith</td>
<td></td>
<td>12.12.70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Illegal value**: SSN for John Smith contains an illegal value that violates uniqueness for SSN.
- **Uniqueness for SSN violated**: SSN for John Smith violates uniqueness.
- **Violated attribute dependency**:_age_ for John Smith is dependent on an invalid SSN.
- **Referential integrity violation**: Phone number for Liipzig is invalid.
- **Embedded value**: Missing values (dummy or null) for John Smith.
- **Duplicate record**: J.Smith and John Smith are duplicate records.
- **Contradicting record**: J.Smith 12.02.70 Leipzig is contradicting.
- **Missing values** (dummy or null): Phone number for Liipzig is missing.
- **Misfielded value**: Club ID for J.Smith 12.02.70 Leipzig is mishandled.
- **Misspelling**: Name and date of birth for J.Smith are misspelled.
- **Cryptic value**: Experience for J.Smith 12.02.70 Leipzig is cryptic.
Overview

- Definitions and Terminology
- Data Quality Methodologies
- Data Quality Problems in Big Data
- Empirical Explanations and Data Glitches
- Data Quality Management in Data Streams
- Data Quality Management in Data Lakes
- Data Quality Management in Data Integration Tools
- Conclusion

This part to a large extent is based on B. Saha & D. Srivastava: Data Quality – The Other Face of Big Data. ICDE 2014.
A short history of Data Quality Research

- **1990s: TDQM @ MIT, DWQ @ EU, Redman, ...**
  - Data quality definitions (→ fitness for use)
  - Data quality dimensions (→ correctness, consistency, accuracy, ...)
  - Data quality methodologies (→ define, measure, analyse, improve)
  - Data cleaning in data warehouses

- **2000s: Establishing the research field**
  - Books (TDQM, DWQ, Batini/Scannapieco, ...)
  - ISO Standardization (ISO 8000, ISO 250xx, ...)
  - Conference & Workshop series: IQ, QDB, ...

- **2010s:**
  - Big Data: Volume, Velocity, Variety, **Veracity, Value**
  - Scalability of entity resolution, record linkage, similarity matching, ...
Typical Data Quality Problems

- Why such inconsistency?
- Semantic ambiguity

Yahoo! Finance

Green Mountain Coffee Roasters, Inc. (NasdaqGS: GMCR)
After Hours: 95.13 ▶ -0.01 (-0.02%) 4:07PM EDT

Day's Range: 93.80 - 95.71
52wk Range: 25.38 - 95.71

Nasdaq

| Last Sale | $ 95.14 |
| Change Nt / % | 1.69 ▶ 1.81% |
| Best Bid / Ask | $ 95.03 / $ 95.94 |
| 1y Target Est | $ 95.00 |
| Today's High / Low | $ 95.71 / $ 93.80 |
| Share Volume | 2,384,175 |
| 50 Day Avg. Daily Volume | 2,751,062 |
| Previous Close | $ 93.45 |
| 52 Wk High / Low | $ 93.72 / $ 25.38 |
| Shares Outstanding | 152,785,000 |
| Market Value of Listed Security | $ 14,535,964,000 |
| P/E Ratio | 120.43 |
| Forward P/E (yr) | 63.57 |
| Earnings Per Share | $ 0.79 |
| Annualized Dividend | N/A |
| Dividend Pay Date | N/A |
| Dividend Yield | 0.82 |

NASDAQ Official Open Price: $ 94.01
Date of NASDAQ Official Open Price: Jul 7, 2011
NASDAQ Official Close Price: $ 95.14
Date of NASDAQ Official Close Price: Jul 7, 2011

52wk Range: 25.38 - 95.71
52 Wk: 25.38 - 93.72
Typical Data Quality Problems

- Why such inconsistency?
- Unit errors
Small Data Quality: How was It Achieved?

Specify all domain knowledge as **integrity constraints** on data

- **Reject updates** that do not preserve integrity constraints
- Works well when the domain is well understood and static
Big Data Quality: A Different Approach?

Big data: integrity constraints cannot be specified a priori

- Data **diversity** → complete domain knowledge is infeasible
- Data **evolution** → domain knowledge quickly becomes obsolete
- Too much rejected data → “small” data 😊
Big Data Quality: A Different Approach?

Big data: integrity constraints cannot be specified a priori
- Data **diversity** → complete domain knowledge is infeasible
- Data **evolution** → domain knowledge quickly becomes obsolete

Solution: let the data speak for itself
- Learn **models** (semantics) from the data
- Identify **data glitches** as violations of the learned models
- Repair **data glitches and models** in a timely manner
Overview

- Definitions and Terminology
- Data Quality Methodologies
- Data Quality Problems in Big Data
- Empirical Explanations and Data Glitches
- Data Quality Management in Data Streams
- Data Quality Management in Data Lakes
- Data Quality Management in Data Integration Tools
- Conclusion
Empirical Explanations

Expectation (or constraint in small data):
Phone number is unique

Explanation for violation:
Phone numbers of new hires can be the same as the phone of their supervisor

Explanation can be learned from the data (by Data Profiling)

⇒ Empirical Explanation
Empirical Explanations

There might be many violations of the expected constraint.

Analysis and data profiling might lead to revised constraints (conditional functional dependencies).

Example: Employees in the same room can have the same phone number.

<table>
<thead>
<tr>
<th>ID</th>
<th>Status</th>
<th>Phone</th>
<th>Dept</th>
<th>Rm</th>
<th>Super_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID_10</td>
<td>Active</td>
<td>1AAA3605519</td>
<td>D8000</td>
<td>A132</td>
<td>ID_13</td>
</tr>
<tr>
<td>ID_11</td>
<td>Active</td>
<td>1AAA3605519</td>
<td>D8000</td>
<td>A132</td>
<td>ID_13</td>
</tr>
<tr>
<td>ID_12</td>
<td>Active</td>
<td>1AAA3605519</td>
<td>D8000</td>
<td>A132</td>
<td>ID_13</td>
</tr>
</tbody>
</table>
Data Glitches

Not all violations of the expected constraint can be explained

➡️ Data glitch

<table>
<thead>
<tr>
<th>ID</th>
<th>Status</th>
<th>Phone</th>
<th>Dept.</th>
<th>Rm.</th>
<th>Super_ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID_1</td>
<td>Active</td>
<td>1AAA3600000</td>
<td>D4000</td>
<td>------</td>
<td>ID_4</td>
</tr>
<tr>
<td>ID_2</td>
<td>------</td>
<td>1AAA3600000</td>
<td>------</td>
<td>------</td>
<td>--------</td>
</tr>
<tr>
<td>ID_3</td>
<td>Active</td>
<td>1AAA3600000</td>
<td>D2200</td>
<td>E260</td>
<td>ID_6</td>
</tr>
</tbody>
</table>

What is an Empirical Explanation?

NYC Taxi Data

What is an Empirical Explanation?

Try to find correlations in your data with other data sets

⇒ Empirical explanation might involve multiple data sets

Learning Empirical Explanations with Statistical Signatures

For each value $v$ in $A$, compute **propensity signatures** in $A$ and $A'$.

- $s_\text{A}(\text{New Hire}) = \{0.67, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0\}$
- $s_\text{A'}(\text{New Hire}) = \{0.05, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0\}$
Learning Empirical Explanations with Statistical Signatures


Apply constraint on D, identify violations (suspicious set) A.

For each value v in A, compute **propensity signatures** in A and A'.

- \( s_A(ID_5) = \{0.33, 0.0, 0.0, 0.0, 0.0, 0.67\} \)
- \( s_{A'}(ID_5) = \{0.02, 0.0, 0.0, 0.0, 0.0, 0.05\} \)
Step 2: Check statistical significance


Goal: Informative values that distinguish A from A'.
- Establish statistical significance using crossover subsampling.
- For an A block, sample A' blocks R times to create distribution.
Step 3: Validate by expert

<table>
<thead>
<tr>
<th>ID</th>
<th>Status</th>
<th>Phone</th>
<th>Dept</th>
<th>Rm.</th>
<th>Super</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID_5</td>
<td>Active</td>
<td>1AAA3608776</td>
<td>D2300</td>
<td>A115</td>
<td>ID_9</td>
</tr>
<tr>
<td>ID_7</td>
<td>New Hire</td>
<td>1AAA3608776</td>
<td>D2300</td>
<td>D284</td>
<td>ID_5</td>
</tr>
<tr>
<td>ID_8</td>
<td>New Hire</td>
<td>1AAA3608776</td>
<td>D2300</td>
<td>B106</td>
<td>ID_5</td>
</tr>
</tbody>
</table>

**Empirical explanation**: collection of all informative values for A.
- Learned in an **unsupervised manner**, e.g., \{ID_5, New Hire\}.
- Experts check empirical explanations, and decide on actions taken.

Overview

- Definitions and Terminology
- Data Quality Methodologies
- Data Quality Problems in Big Data
- Empirical Explanations and Data Glitches
- **Data Quality Management in Data Streams**
- Data Quality Management in Data Lakes
- Data Quality Management in Data Integration Tools
- Conclusion
Data Quality Management in Data Streams

Data Streams are infinite streams of data which are processed continuously.

Data quality improvements might need to happen at runtime.

Example: Traffic state estimation with mobile phone data.

- Data quality drops at night because insufficient number of samples is available.
- Increase sampling rate or integrate additional data source.
A Framework for Measuring DQ in Data Streams

Different ways to measure DQ

- Query-based Quality Service: Rewriting of SQL queries and inserting computations of quality values
- Content-based Quality Service: Mathematical formulas to compute data quality values
- Application-based Quality Service: Any kind of application-specific code to measure data quality (e.g., the quality of map matching)
DQ Ontology (or DQ Metadata Model)

Adapts DWQ Metadata Model to data streams

Provides three ways to measure DQ

- SQL Metric (Query-based)
- Semantic Metric (Content-based)
- Application Metric (Application-based)
Example for Query Rewriting

Q1: \textbf{SELECT} RoadID, \textbf{AVG(Speed)}
\textbf{FROM} message
\textbf{GROUP BY} RoadID

Q2: \textbf{SELECT} RoadID, \textbf{AVG(Speed)},
\textbf{COUNT(Speed)} \textbf{AS} SpeedDatavolume_DQ
\textbf{FROM} message
\textbf{GROUP BY} RoadID
DQ Measurement along the Stream Processing

DQ values are inserted into the data stream

DQ values might depend on each other
DQ Management does not create much overhead
DQ Monitoring for Health Data

DQ metric tries to measure regularity of PPG curve

Movement might introduce measurement artefacts
Overview

- Definitions and Terminology
- Data Quality Methodologies
- Data Quality Problems in Big Data
- Empirical Explanations and Data Glitches
- Data Quality Management in Data Streams
- Data Quality Management in Data Lakes
- Data Quality Management in Data Integration Tools
- Conclusion
Data Lake Architecture

Metadata and data quality management is an issue that goes across all layers

- **Ingestion:**
  - Metadata Extraction
  - Minimal requirements for ingested data
- **Storage:**
  - Metadata repository stores also DQ information
  - DQ-oriented data integration, query rewriting
- **Transformation:**
  - DQ improvement by data cleaning
- **Interaction:**
  - Show DQ information to the users
Metadata Management in DLs

Diagram showing the process of metadata management in DLs, including steps such as load, schema discovery, schema matching, schema grouping, schema summary, and human feedback, leading to users and transformation & interaction.
Metadata Types

- Structure data
- Semantic data
- Metadata properties

Metadata Model
DQ Measurements on Schemas

Homogeneity vs. Heterogeneity in Schema Summarization
Schema Summary

Concise and usable schema summary against the complex metadata

- Summary Size
- Summary Importance
- Summary Coverage
Overview

- Definitions and Terminology
- Data Quality Methodologies
- Data Quality Problems in Big Data
- Empirical Explanations and Data Glitches
- Data Quality Management in Data Streams
- Data Quality Management in Data Lakes
- Data Quality Management in Data Integration Tools
- Conclusion
Data Quality Management in Data Integration Tools

- All major data integration tools claim to support data quality management (e.g., Informatica, Talend, Pentaho, ...)

- They include methods to improve data quality (data cleaning, data transformation) as well as for measurement

- The following examples are from Talend Open Studio for Data Quality (Open Source, http://talend.com)
Define Metrics
Define expected ranges
Specific metric for checking patterns in text fields
Analyze results of the measurements
Overview

- Definitions and Terminology
- Data Quality Methodologies
- Data Quality Problems in Big Data
- Empirical Explanations and Data Glitches
- Data Quality Management in Data Streams
- Data Quality Management in Data Lakes
- Data Quality Management in Data Integration Tools
- Conclusion
Conclusion

- Data quality is subjective
  - Depends on application requirements, context, user, ...
- Data quality can be measured without knowing the true values
  - Examine the intrinsic properties of the data
- Data quality is not only aspect of the data
  - Metadata and data processing systems also affect data quality
- Data quality management is more than data cleaning
  - Data cleaning is one aspect of DQM, but there is much more
- Data quality management is closely related to data profiling