Data Quality Management for Big Data

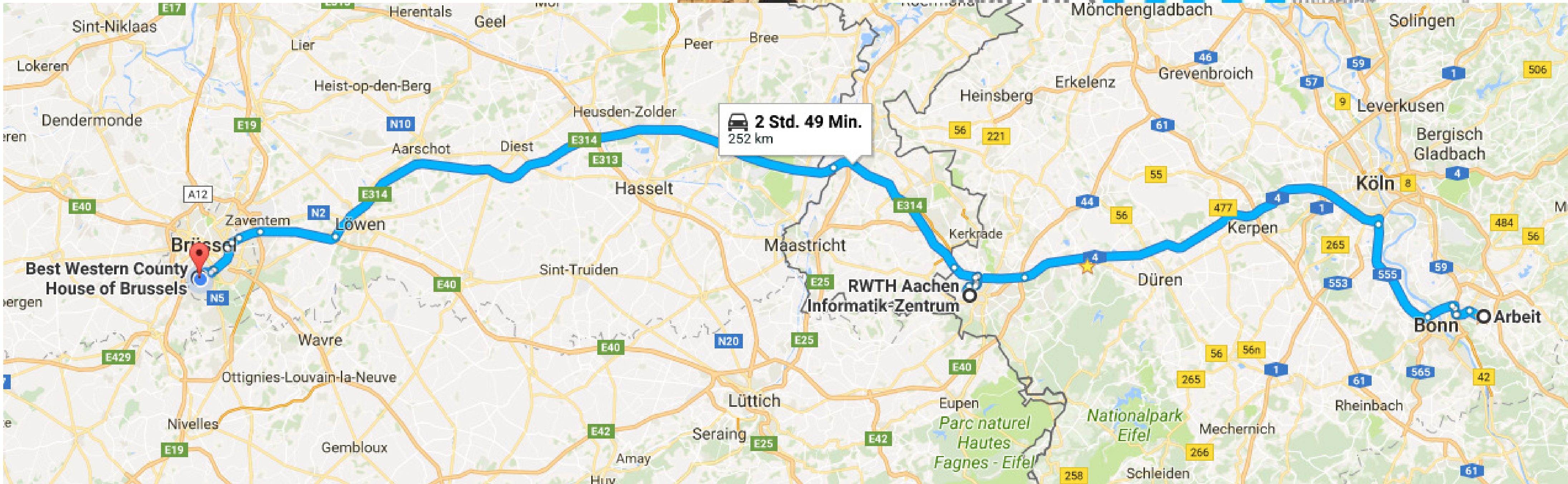
Christoph Quix

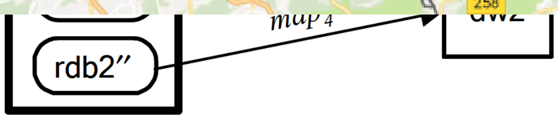
EBISS Summer School, July 4, 2017

Fraunhofer FIT
RWTH AACHEN UNIVERSITY

About Me

Diploma, PhD, Habilitation from RWTH





6. map_4 = user-defined-mapping(rdb2″, dw2
7. map_5 = Match(dw1, dw2)
8. Merge (dw1, dw2, map_5)

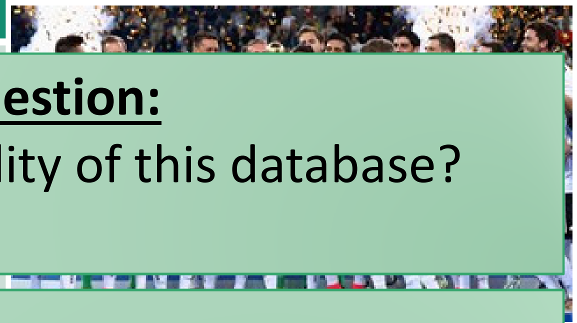
Data Quality: Motivating Example

Soccer player database for the Confed Cup 2017

Player	Bdate	SSN	Age	Club_ID	phone	City	PLZ	Experience
John Smith	12.02.90	123456	22	127	9999	Liipzig		
Peter Miller	30.13.88	123456	22	17		Dresd		
I. Vieth 26.11.92 Leipzig		073456	20	15	+49 341 808060	Germ		
John Smith	12.12.90	123456	22	127	9999	Liipzig		

ClubID	Name	Country
127	RB Leipzig	Germany
17	Śląsk Wrocław	Poland
15	FC Bayern München	Germany

CONFEDERATIONS CUP
RUSSIA 2017



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

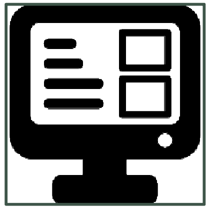
Quality
Perspectives

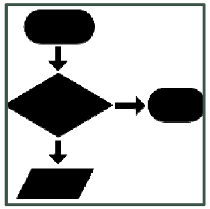
Product-oriented

- 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]

[E99] L. English: Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits. Wiley, 1999.
[ISO1] ISO/IEC 25024:2015 Measurement of Data Quality
[ISO2] ISO/TC 8000-2:2012 Data Quality – Vocabulary

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]

[Redman, 1997] T. Redman: Data Quality for the Information Age. Artech House, 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]

[ISO2] ISO/TC 8000-2:2012 Data Quality – Vocabulary

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

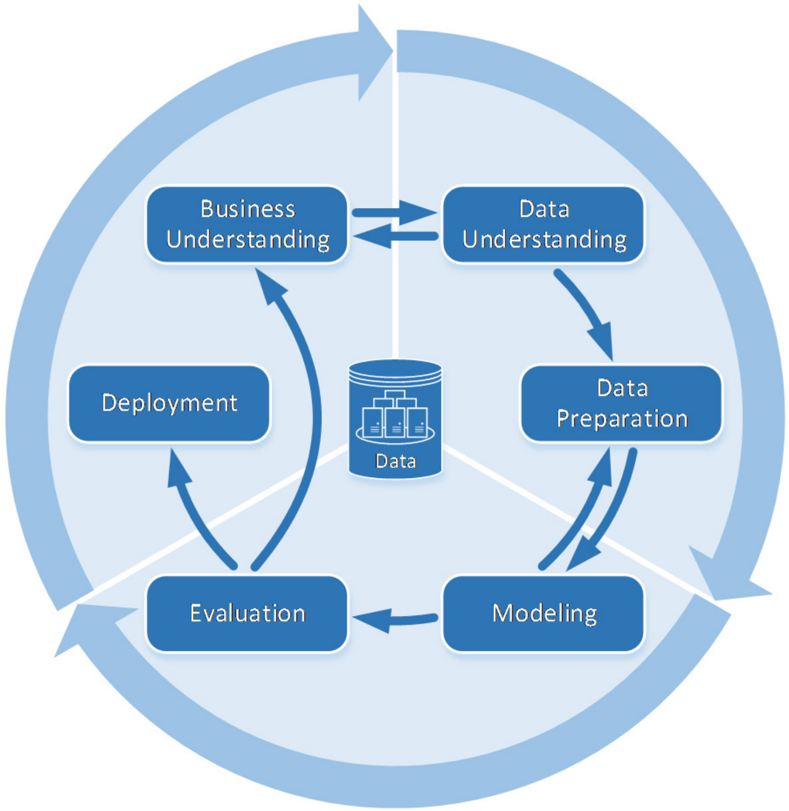
Data Quality = „fitness for use“ ➔ the use changes in data integration!

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

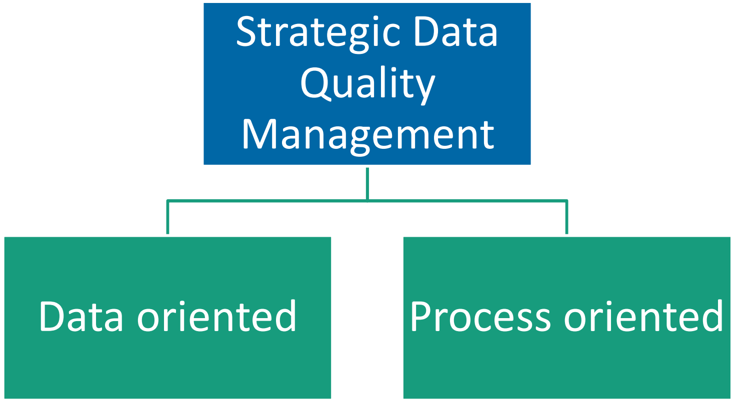
Corporate mergers or reorganizations:
Existing data is used in a new context

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

http://info.talend.com/rs/talend/images/WP_EN_DQ_Talend_10RootCauses.pdf

Data-oriented vs. Process oriented DQ 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

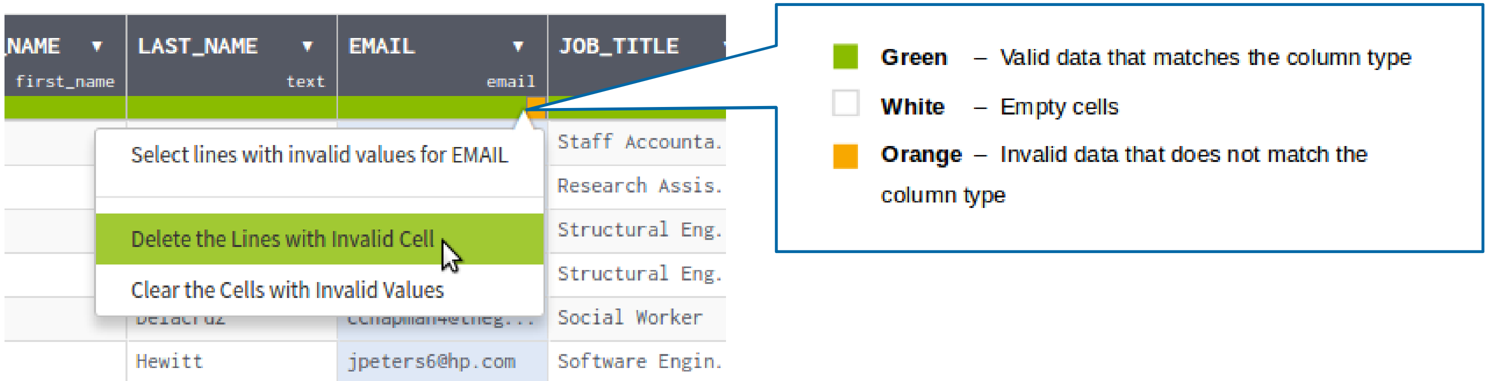
- Data validation
  - Manual and visual data validation
  - Rule based data validation
- Outlier detection
- Anomaly detection
- Automatic vs. manual procedures in data cleaning
- Entity resolution
- …

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: excecute 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

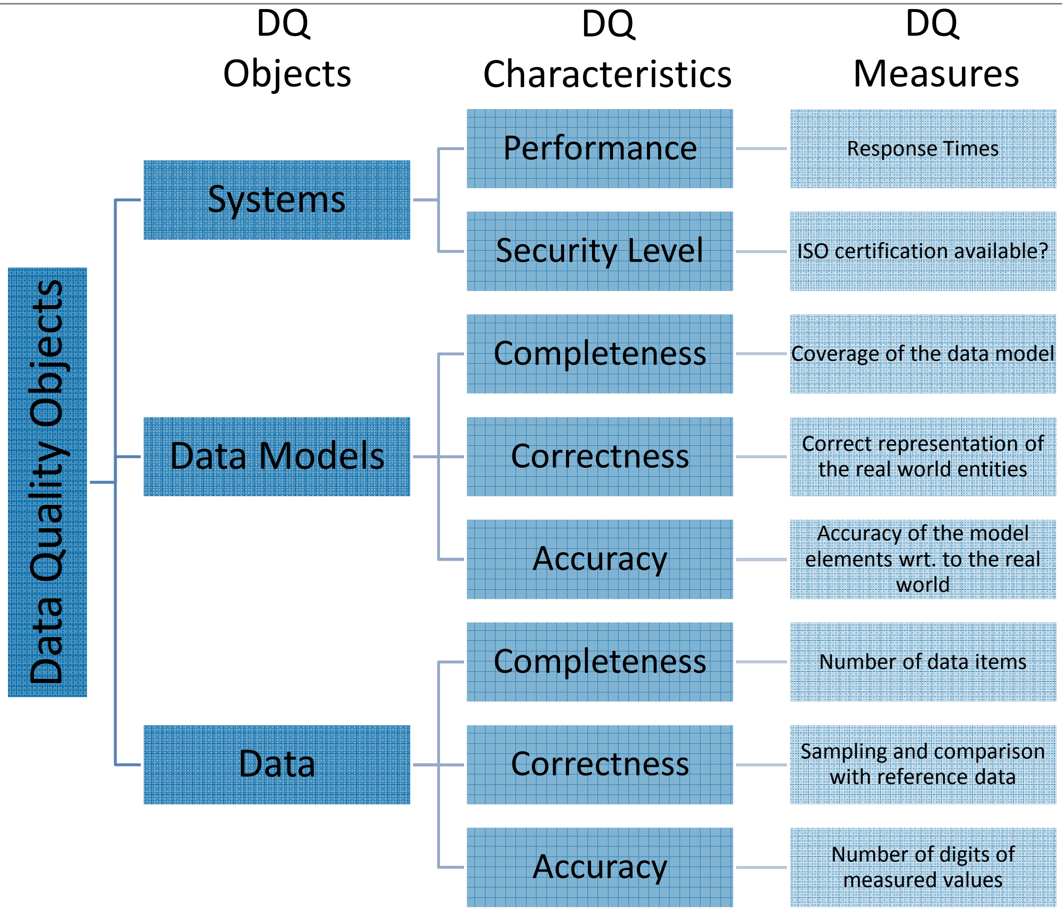
Problem: customers do not provide valid information when they register to our web shop

Current sign-up form for web-shop	Improved sign-up form for web-shop
Name	1. Step: please provide your email:
Address	Email
Birthdate	2. Step: please provide optional information
Phone	Address
Email	Birthdate
….	Phone

Data Quality Goals represent the DQ Requirements

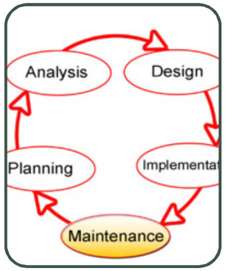
Goal WHY	Extract more data from source systems X and Y	Reduce time for transformation process	Reduce effort for adoption of ETL processes
Characteristic (Dimension) WHAT	completeness	efficiency	robustness
Measurement Method HOW	Ratio of extracted data and available data	Time needed for extraction process	Count of situations where ETL process had to be changed
Observation ➔ Result	Number of fields (or rows) in integrated DB divided by number of fields (rows) in source systems	Time per extracted row from sources X and Y	Number of versions of ETL process algorithms per week

Objects for measuring and improving Data Quality



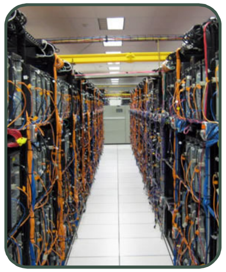
Processes relevant for Improving Data Quality

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?

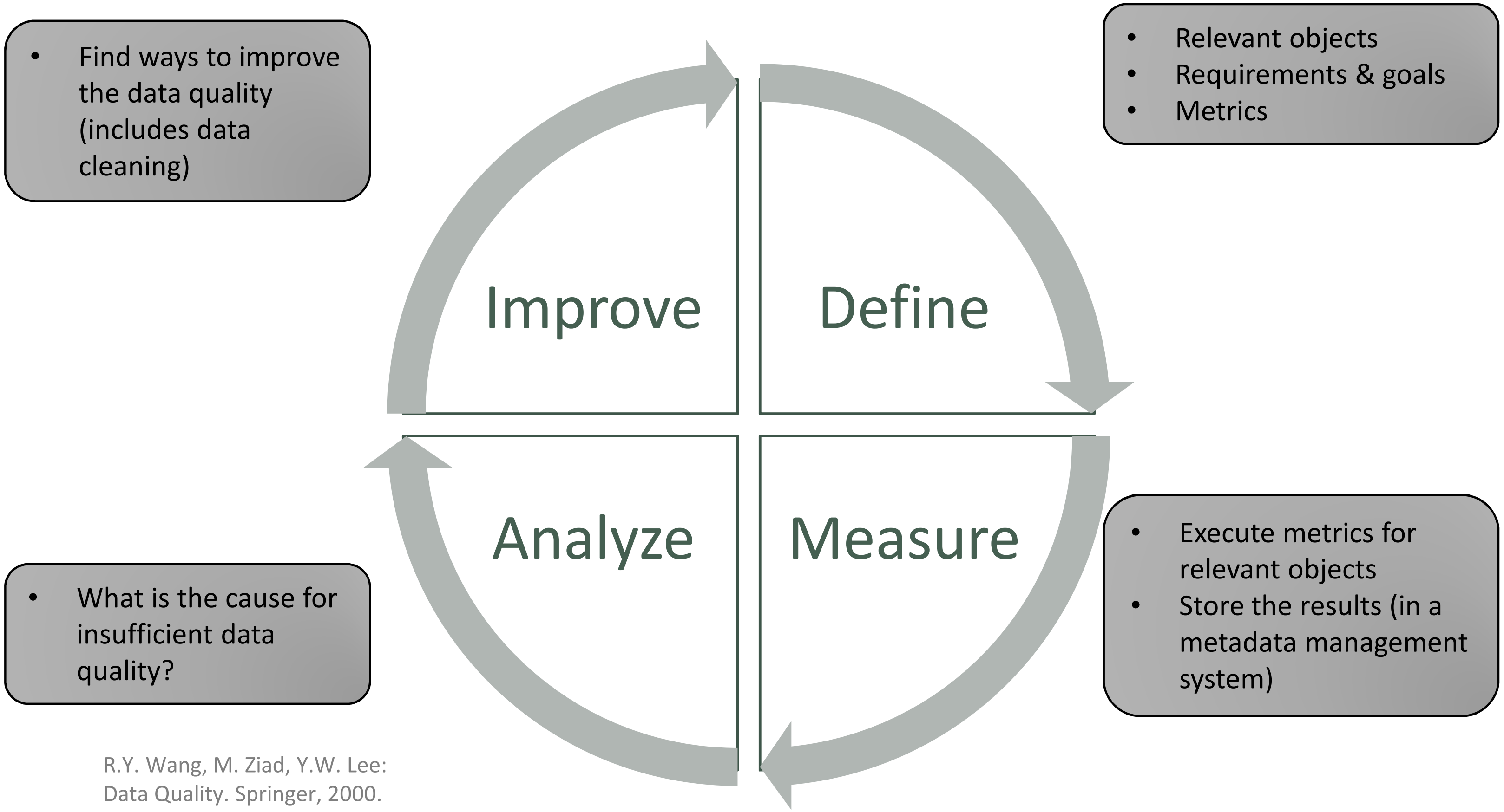
Sample Answers for the Scenario

- 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

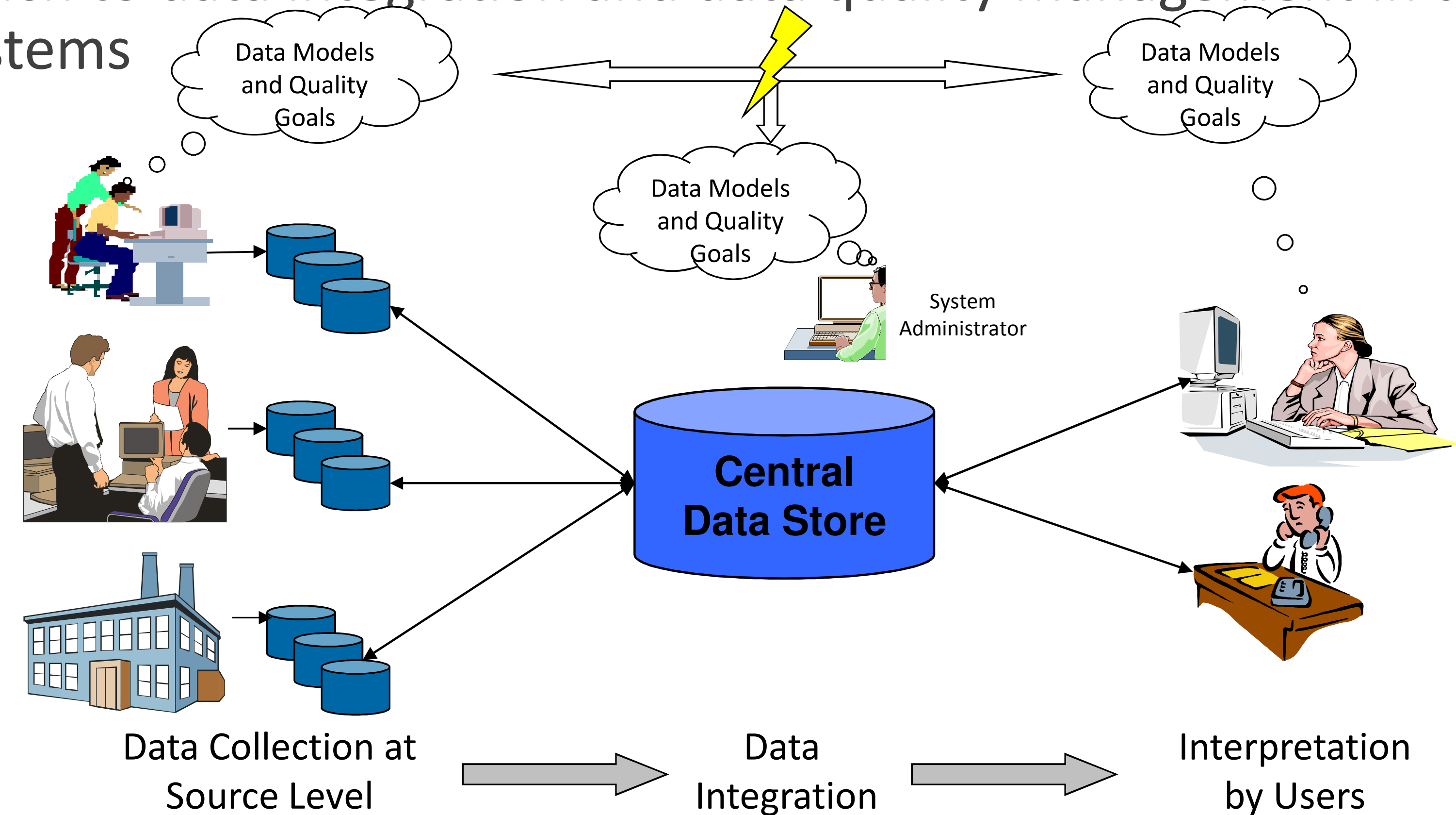
TDQM
Total Data Quality Management



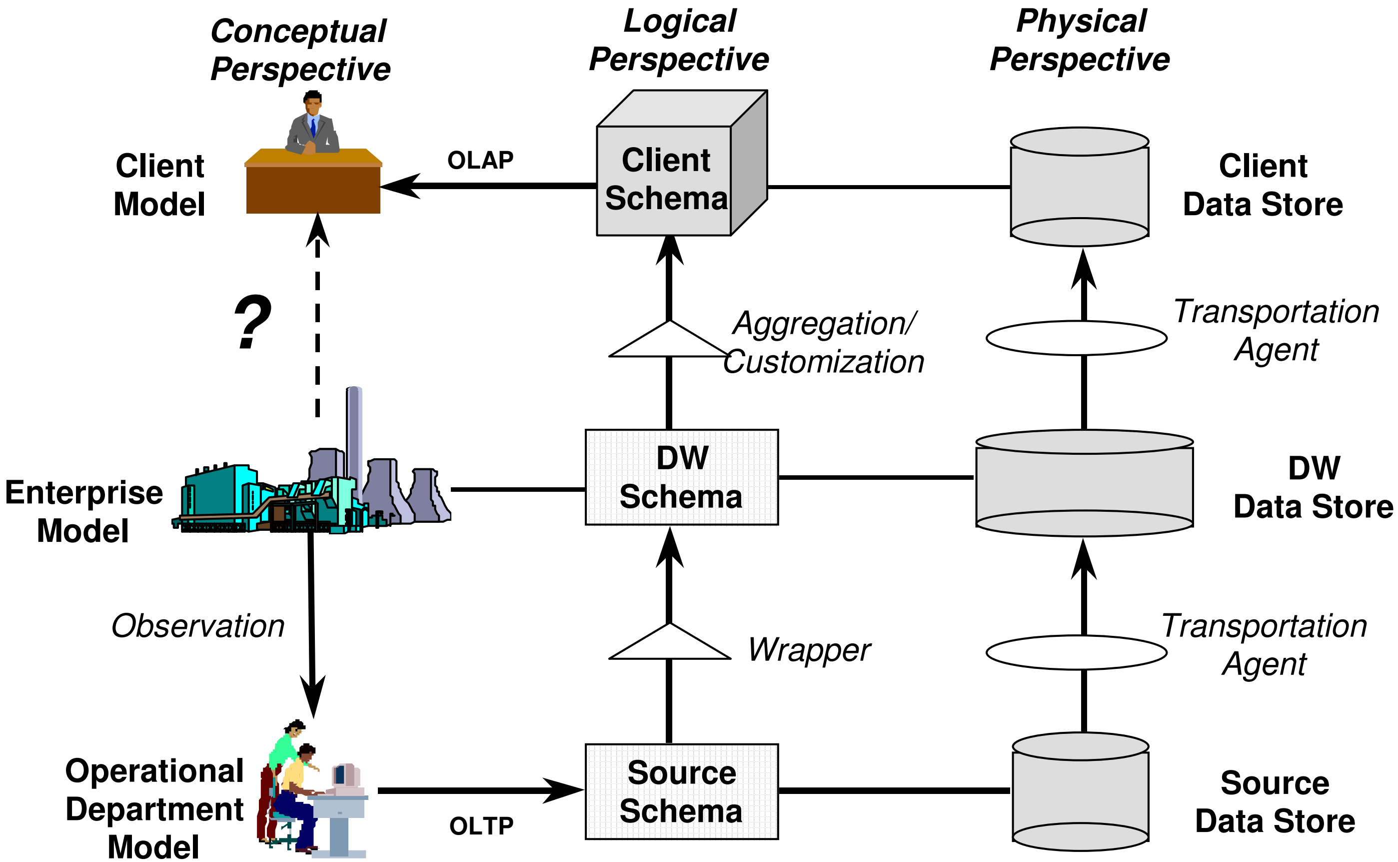
R.Y. Wang, M. Ziad, Y.W. Lee: Data Quality. Springer, 2000.

DWQ
Data Warehouse Qualilty (EU Project 1997-2000)

Holistic approach to data integration and data quality management in data warehouse systems

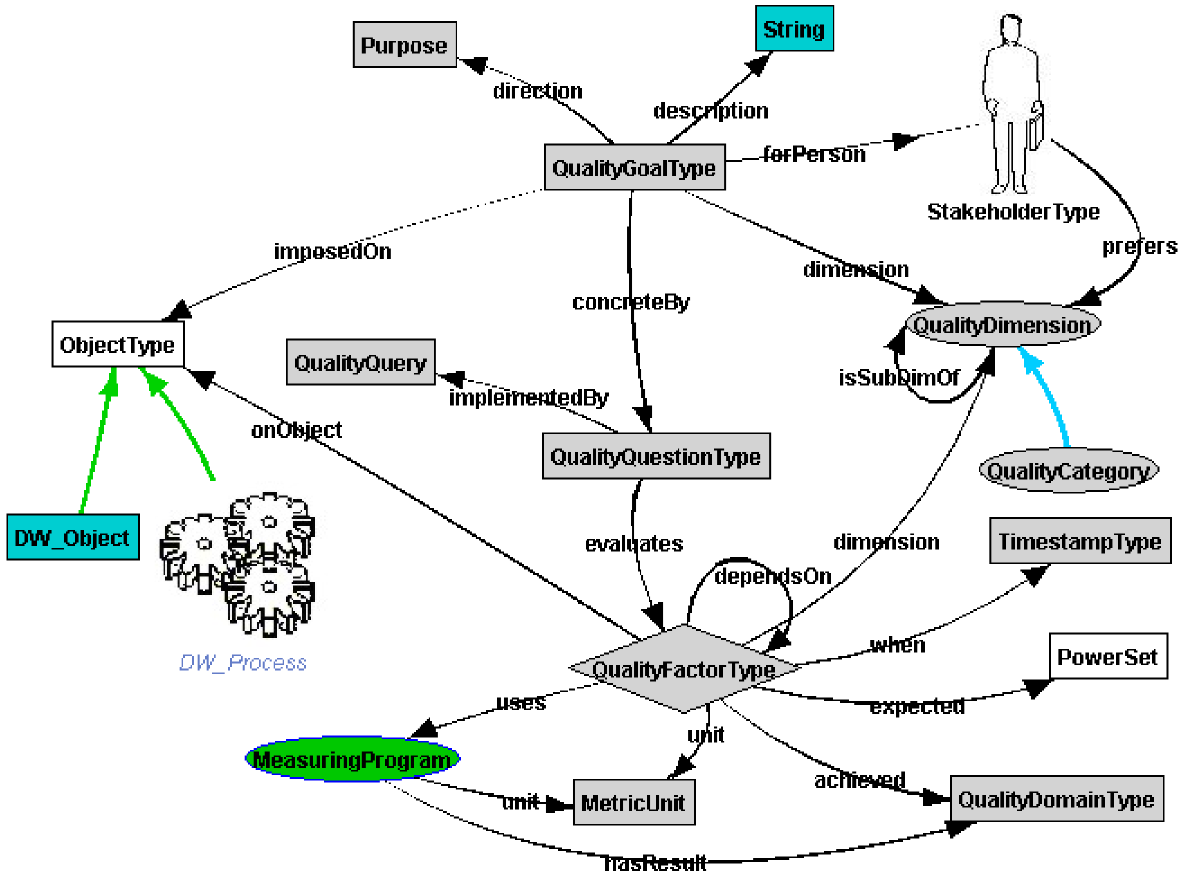


DWQ Framework for DW Metadata



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

Batini & Scannapieco: Data and Information Quality, Springer, 2016.

Step/Meth Acronym	Data Analysis	DQ Requirement Analysis	Identification of Critical Areas	Process Modeling	Measurement of Quality	Extensible to Other Dimensions and Metrics
TDQM	+		+	+	+	Fixed
DWQ	+	+	+		+	Open
TIQM	+	+	+	+	+	Fixed
AIMQ	+		+		+	Fixed
CIHI	+		+			Fixed
DQA	+		+		+	Open
IQM	+				+	Open
ISTAT	+				+	Fixed
AMEQ	+		+	+	+	Open
COLDQ	+	+	+	+	+	Fixed
DaQuinCIS	+		+	+	+	Open
QAFD	+	+	+		+	Fixed
CDQ	+	+	+	+	+	Open

Diverting definitions for DQ Characteristics (DQ Dimensions)

[Batini et al., 2009]

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

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!)

Acronym	Data Quality Dimension
TDQM	Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of manipulation, Value added, Free of error, Interpretability, Objectivity, Relevance, Reputation, Security, Timeliness, Understandability
DWQ	Correctness, Completeness, Minimality, Traceability, Interpretability, Metadata Evolution, Accessibility (System, Transactional, Security), Usefulness (Interpretability), Timeliness (Currency, Volatility), Responsiveness, Completeness, Credibility, Accuracy, Consistency, Interpretability
TIQM	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.
AIMQ	Accessibility, Appropriateness, Believability, Completeness, Concise/Consistent representation, Ease of operation, Freedom from errors, Interpretability, Objectivity, Relevancy, Reputation, Security, Timeliness, Understandability
CIHI	Dimensions: Accuracy, Timeliness Comparability, Usability, Relevance Characteristics: 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.
DQA	Accessibility, Appropriate amount of data, Believability, Completeness, Freedom from errors, Consistency, Concise Representation, Relevance, Ease of manipulation, Interpretability, Objectivity, Reputation, Security, Timeliness, Understandability, Value added.
IQM	Accessibility, Consistency, Timeliness, Conciseness, Maintainability, Currency, Applicability, Convenience, Speed, Comprehensiveness, Clarity, Accuracy, Traceability, Security, Correctness, Interactivity.
ISTAT	Accuracy, Completeness, Consistency
AMEQ	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
COLDQ	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.
DaQuinCIS	Accuracy, Completeness, Consistency, Currency, Trustworthiness
QAFD	Syntactic/Semantic accuracy, Internal/External consistency, Completeness, Currency, Uniqueness.
CDQ	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.

[Batini et al., 2009]

DQ Metrics

Several proposals for DQ metrics have been made

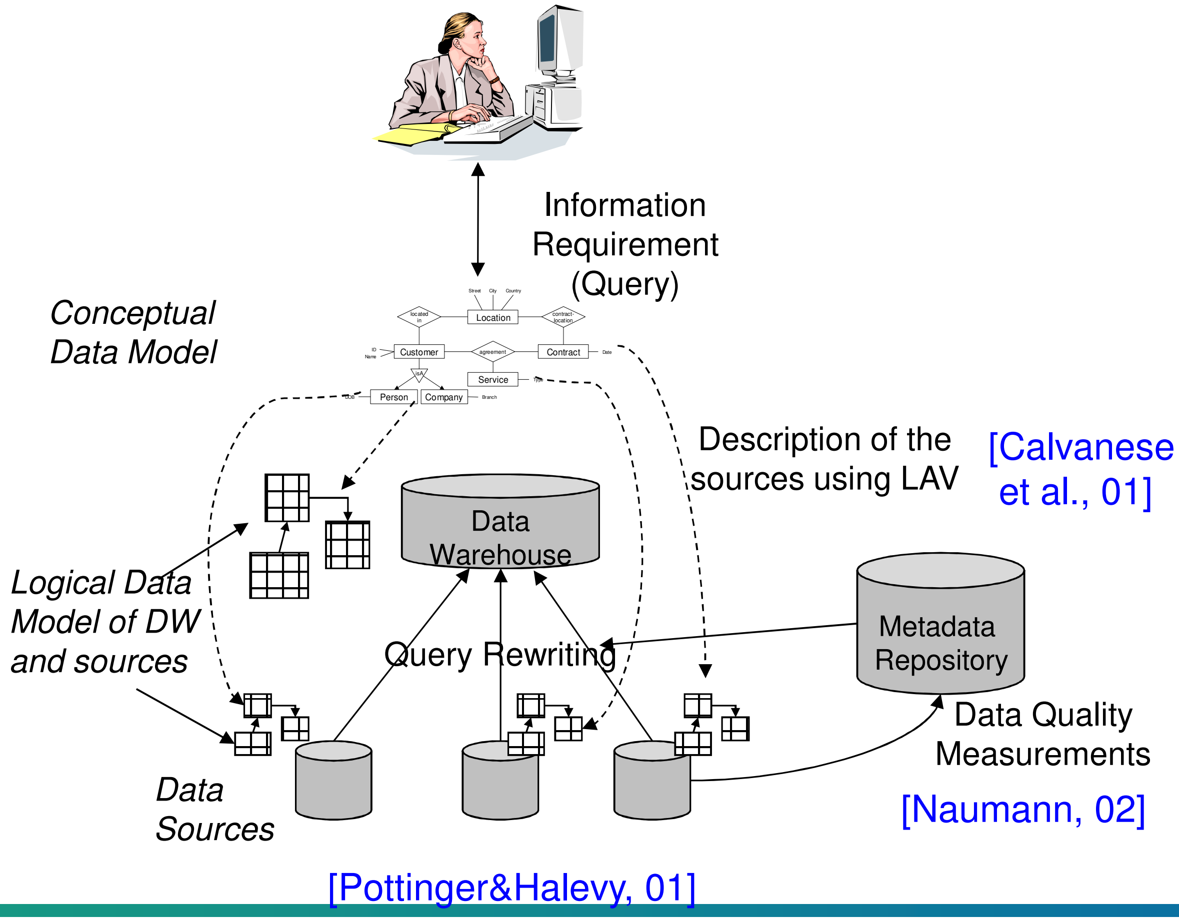
Some can be computed automatically, some require user input or knowledge of „correct“ values

➔ not scalable

Dimensions	Name	Metrics Definition
Accuracy	Acc1	Syntactic accuracy: it is measured as the distance between the value stored in the database and the correct one Syntactic Accuracy=Number of correct values/number of total values
	Acc2	Number of delivered accurate tuples
	Acc3	User Survey - Questionnaire
Completeness	Compl1	Completeness = Number of not null values/total number of values
	Compl2	Completeness = Number of tuples delivered/Expected number
	Compl3	Completeness of Web data = $(T_{max} - T_{current})^*$ $(Completeness_{Max} - Completeness_{Current})/2$
	Compl4	User Survey - Questionnaire
Consistency	Cons1	Consistency = Number of consistent values/number of total values
	Cons2	Number of tuples violating constraints, number of coding differences
	Cons3	Number of pages with style guide deviation
	Cons4	User Survey - Questionnaire
Timeliness	Time1	Timeliness = $(\max(0; 1\text{-Currency/Volatility}))^s$
	Time2	Percentage of process executions able to be performed within the required time frame
	Time3	User Survey - Questionnaire
Currency	Curr1	Currency = Time in which data are stored in the system - time in which data are updated in the real world
	Curr2	Time of last update
	Curr3	Currency = Request time- last update
	Curr4	Currency = Age + (Delivery time- Input time)
	Curr5	User Survey - Questionnaire

[Batini et al., 2009]

Quality-oriented Data Integration



Basic Assumption:
Data is redundant
➔ Multiple ways to solve

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

Teilziel	w_r	Umschreibung 1			Umschreibung 2			Umschreibung 3		
		A	V	K	A	V	K	A	V	K
PrivateCustomer	0,10	5	0,80	0,80	5	0,80	0,80	5	0,80	0,80
ID	0,05	2	1,00	1,00	2	1,00	1,00	2	0,80	0,70
DOB	0,10	5	0,70	0,60	6	0,99	0,95	6	0,99	0,95
Name	0,05	2	1,00	0,90	2	1,00	0,90	2	1,00	0,90
located-in	0,10	2	0,95	0,90	6	0,99	0,99	6	0,99	0,99
Location	0,05	2	1,00	0,70	6	1,00	0,70	6	1,00	0,70
Country	0,05	2	1,00	1,00	6	1,00	1,00	6	1,00	1,00
agreement	0,20	4	0,95	0,90	4	0,95	0,90	4	0,95	0,90
Service	0,05	4	1,00	1,00	4	1,00	1,00	2	1,00	1,00
Contract	0,05	4	0,95	0,90	4	0,95	0,90	4	0,95	0,90
Code	0,10	4	1,00	0,95	4	1,00	0,95	4	1,00	0,95
Type	0,10	4	1,00	0,95	4	1,00	0,95	2	0,80	0,80
v_{join}		5	0,90	0,81	6	0,95	0,87	6	0,90	0,83
skaliert nach Def.		1	0	0	0	1	1	0	0	0,33
alternative Skal.		0,5	0,90	0,81	0,4	0,95	0,87	0,4	0,90	0,83
Qualitätswerte		0,2 bzw. 0,784			0,8 bzw. 0,808			0,132 bzw. 0,772		

[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

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

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

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

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

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

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

Illegal value

Uniqueness for SSN violated

Violated attribute dependency

Referential integrity violation

Player	Bdate	SSN	Age	Club_ID	phone	City	Experience
John Smith	12.02.70	123456	22	127	9999	Liipzig	B
Peter Miller	30.13.70	123456		17			
J.Smith 12.02.70 Leipzig		123456	22	127	-	Germany	B
John Smith	12.12.70						

Embedded value

Duplicate record

Contradicting record

Missing values (dummy or null)

Misfielded value

Misspelling

Cryptic value

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, (NasdaqGS: GMCR)
After Hours: **95.13** -0.01 (-0.02%) 4:07PM EDT

Last Trade:	**95.14**	Day's Range:	**93.80 - 95.71**
Trade Time:	**4:00PM EDT**	52wk Range:	**25.38 - 95.71**
Change:	**1.69 (1.81%)**	Volume:	**2,384,075**
Prev Close:	**93.45**	Avg Vol (3m):	**2,512,070**
Open:	**94.01**	Market Cap:	**13.51B**
Bid:	**95.03** x 100	P/E (ttm):	**119.82**
Ask:	**95.94** x 100	EPS (ttm):	**0.79**
1y Target Est:	**92.50**	Div & ...	**N/A (N/A)**

Nasdaq

Last Sale	**$ 95.14**
Change Net / %	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	14,535,964,900
P/E Ratio	120.43
Forward P/E (1yr)	63.57
Earnings Per Share	$ 0.79
Annualized Divide...	N/A
Ex Dividend ...	N/A
Divide...	
...eld	
	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

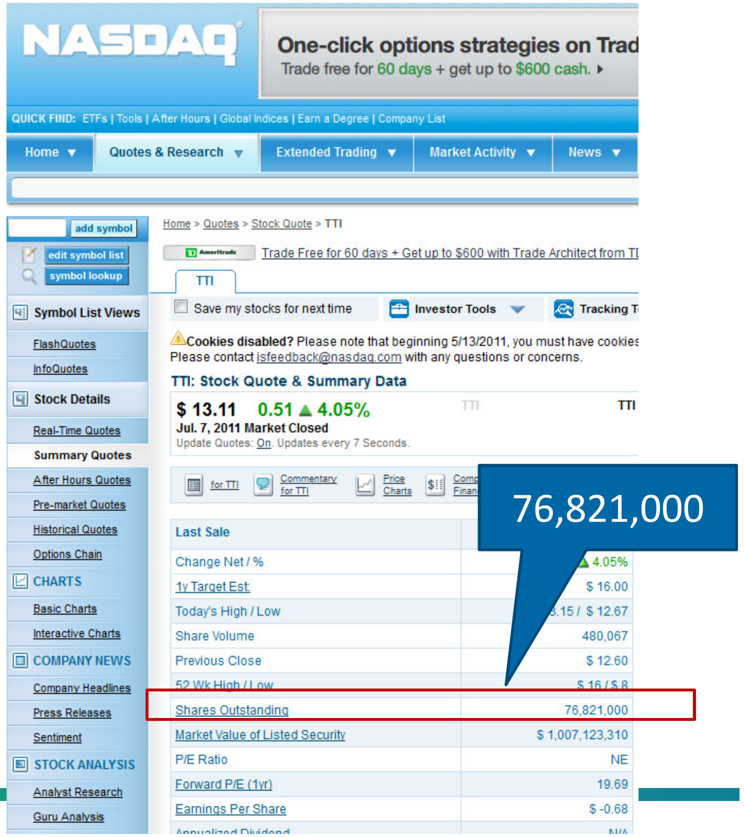
52wk Range: 25.38-95.71

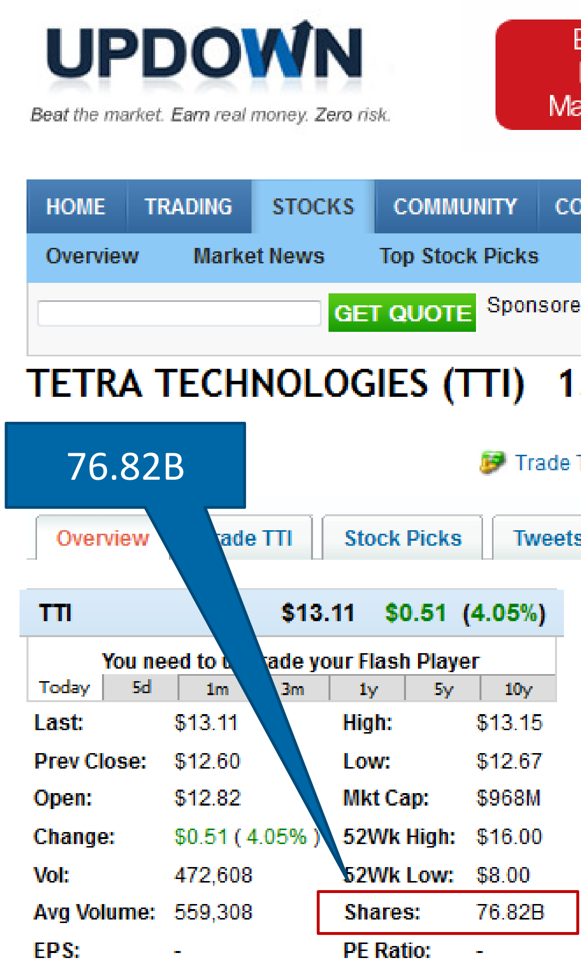
52 Wk: 25.38-93.72

Day's Range: 93.80-95.71

Typical Data Quality Problems

- Why such inconsistency?
  - Unit errors





Small Data Quality: How was It Achieved?

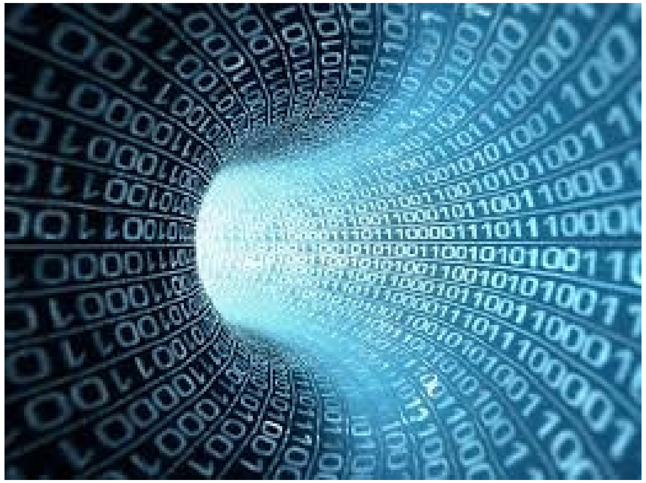
Specify all domain knowledge as **integrity constraints** on data

- **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

D. Srivastava: Data Glitches = Constraint Violations – Empirical Explanations. QDB Workshop, 2016.

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

ID	Status	Phone	Dept.	Rm.	Super_ID
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	New Hire	1AAA3608776	D2300	D284	ID_5
ID_8	New Hire	1AAA3608776	D2300	B106	ID_5

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

D. Srivastava: Data Glitches = Constraint Violations – Empirical Explanations. QDB Workshop, 2016.

Empirical Explanations

There might be many violations of the expected constraint

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

ID	Status	Phone	Dept.	Rm.	Super_ID
ID_10	Active	1AAA3605519	D8000	A132	ID_13
ID_11	Active	1AAA3605519	D8000	A132	ID_13
ID_12	Active	1AAA3605519	D8000	A132	ID_13

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

D. Srivastava: Data Glitches = Constraint Violations – Empirical Explanations. QDB Workshop, 2016.

Data Glitches

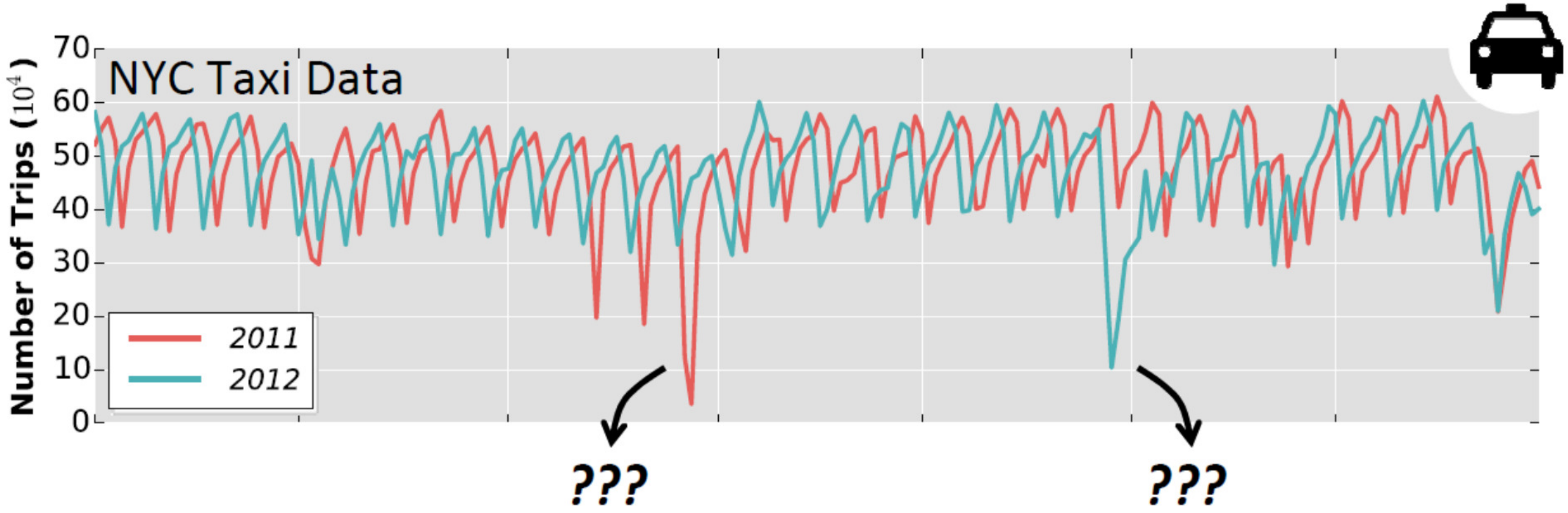
Not all violations of the expected constraint can be explained

➔ Data glitch

ID	Status	Phone	Dept.	Rm.	Super_ID
ID_1	Active	1AAA3600000	D4000	-------	ID_4
ID_2	---------	1AAA3600000	-------	-------	------
ID_3	Active	1AAA3600000	D2200	E260	ID_6

D. Srivastava: Data Glitches = Constraint Violations – Empirical Explanations. QDB Workshop, 2016.

What is an Empirical Explanation?

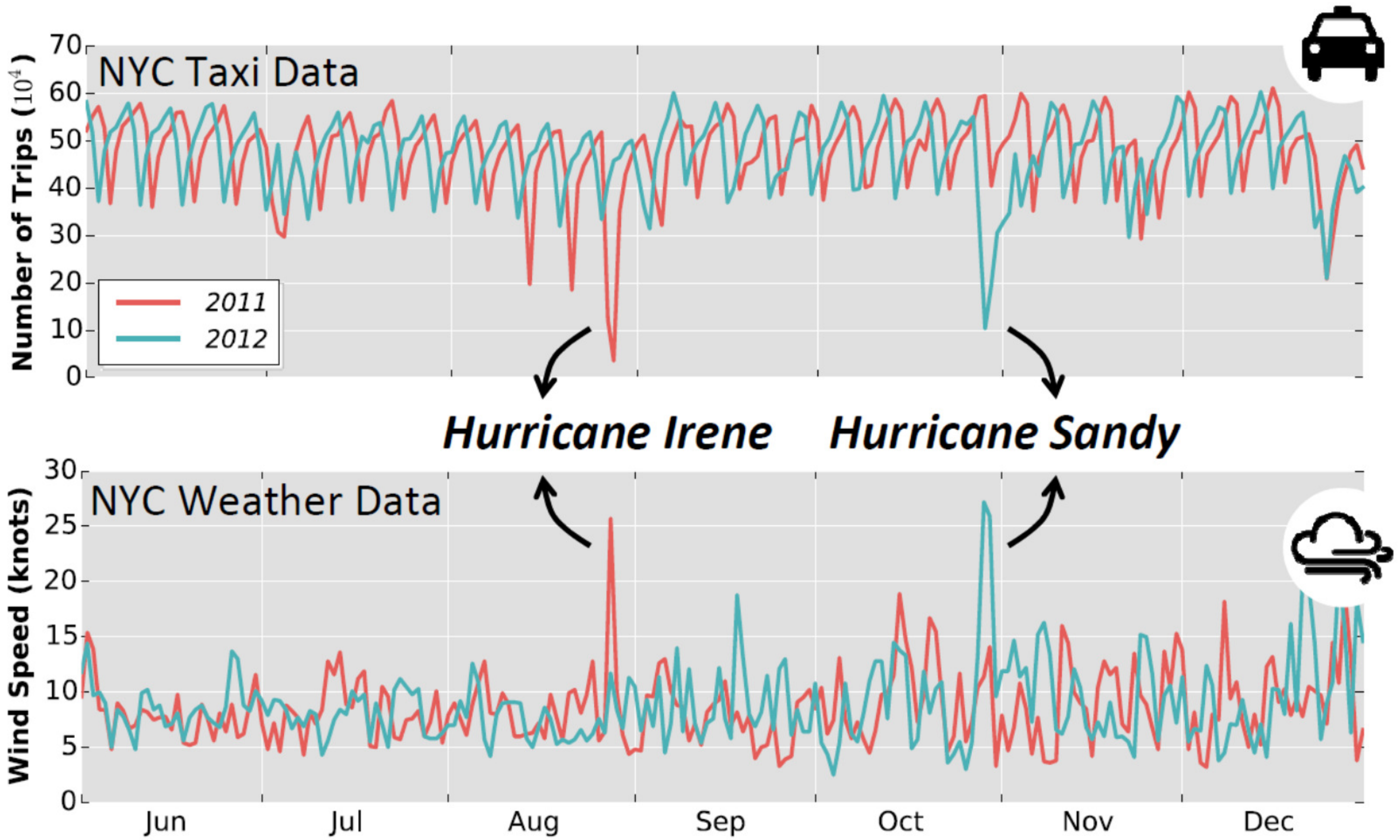


D. Srivastava: Data Glitches = Constraint Violations – Empirical Explanations. QDB Workshop, 2016.

What is an Empirical Explanation?

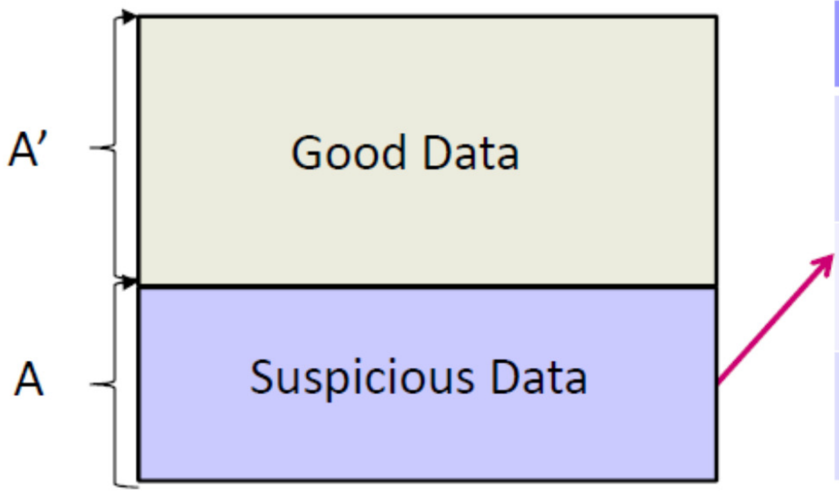
Try to find correlations in your data with other data sets

➔ Empirical explanation might involve multiple data sets



Learning Empirical Explanations with Statistical Signatures

D. Srivastava: Data Glitches = Constraint Violations – Empirical Explanations. QDB Workshop, 2016.



ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	New Hire	1AAA3608776	D2300	D284	ID_5
ID_8	New Hire	1AAA3608776	D2300	B106	ID_5

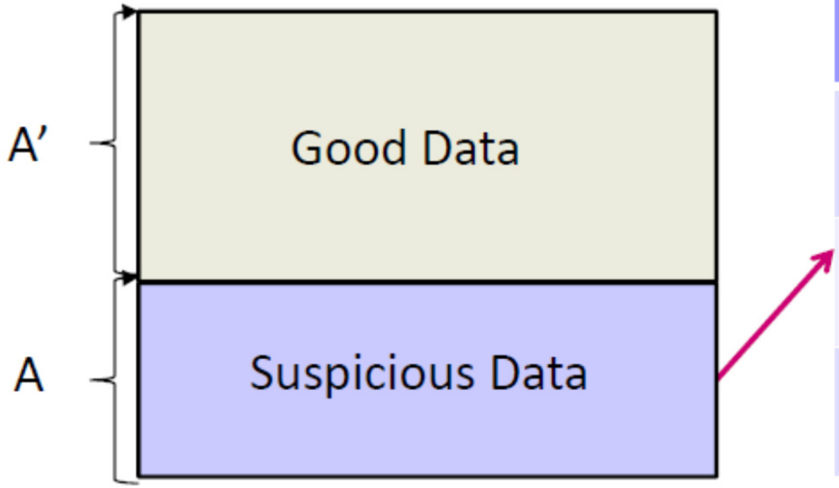
Apply constraint on D, identify violations (suspicious set) A.

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

- s_A(New Hire) = {0.67, 0.0, 0.0, 0.0, 0.0, 0.0}
- $s_{A'}$(New Hire) = {0.05, 0.0, 0.0, 0.0, 0.0, 0.0}

Learning Empirical Explanations with Statistical Signatures

D. Srivastava: Data Glitches = Constraint Violations – Empirical Explanations. QDB Workshop, 2016.



ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	New Hire	1AAA3608776	D2300	D284	ID_5
ID_8	New Hire	1AAA3608776	D2300	B106	ID_5

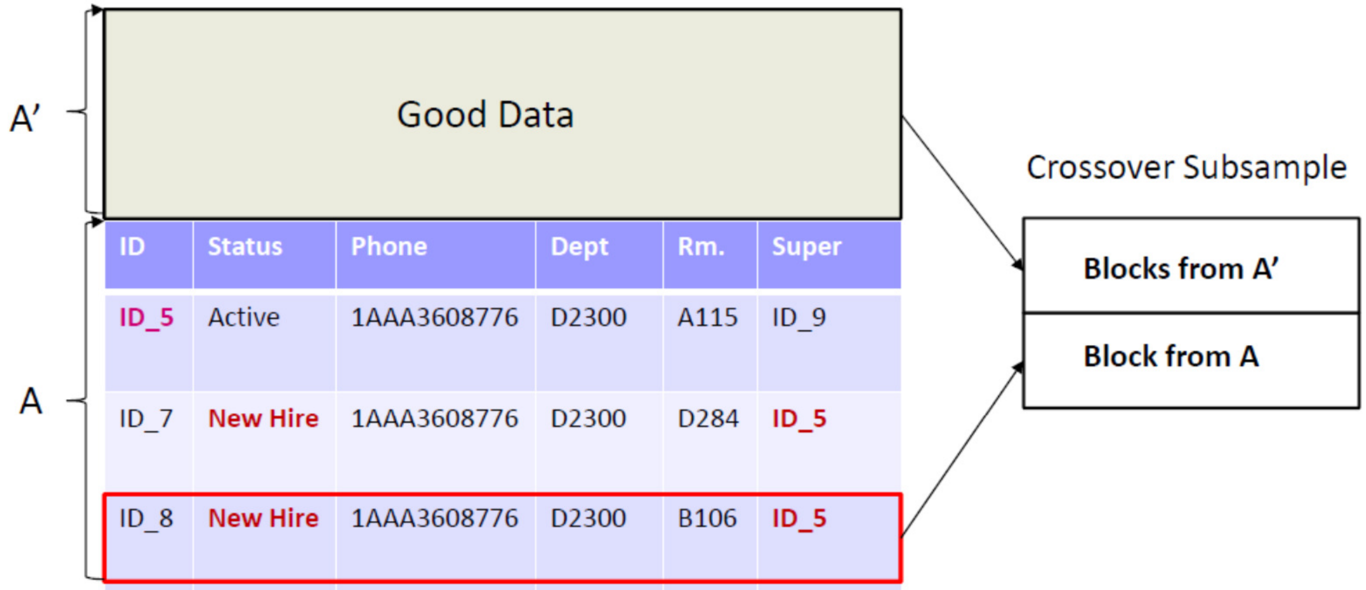
Apply constraint on D, identify violations (suspicious set) A.

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

D. Srivastava: Data Glitches = Constraint Violations – Empirical Explanations. QDB Workshop, 2016.



- 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

D. Srivastava: Data Glitches = Constraint Violations – Empirical Explanations. QDB Workshop, 2016.

ID	Status	Phone	Dept	Rm.	Super
ID_5	Active	1AAA3608776	D2300	A115	ID_9
ID_7	**New Hire**	1AAA3608776	D2300	D284	**ID_5**
ID_8	**New Hire**	1AAA3608776	D2300	B106	**ID_5**

- **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

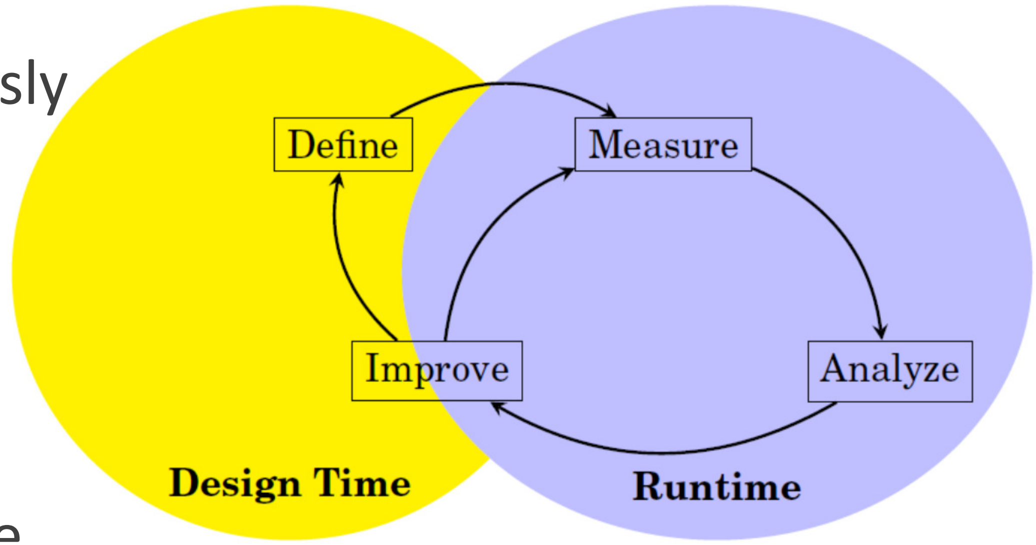
Data Streams are infinite streams of data which are processed continuously

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

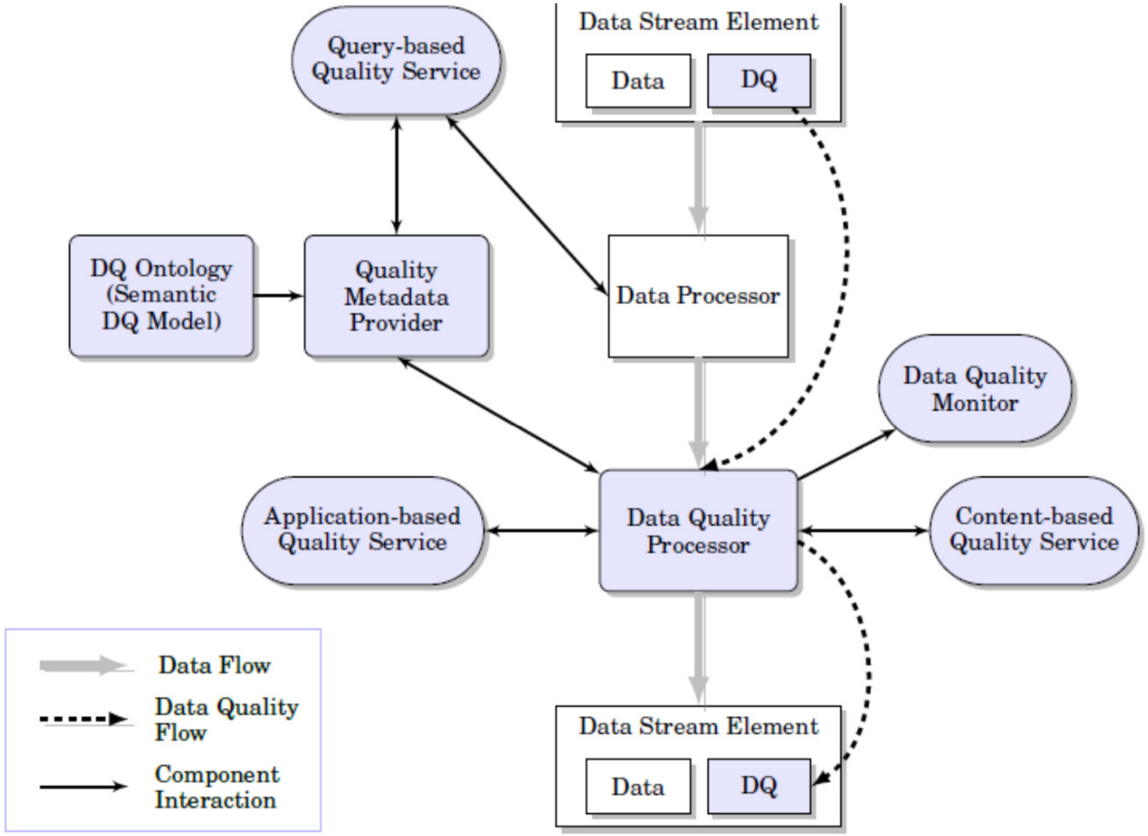


Geisler et al.: Ontology-Based Data Quality Management for Data Streams. Journal on Data and Information Quality, 2016.

A Framework for Measuring DQ in Data Streams

Different ways to mesaure 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)



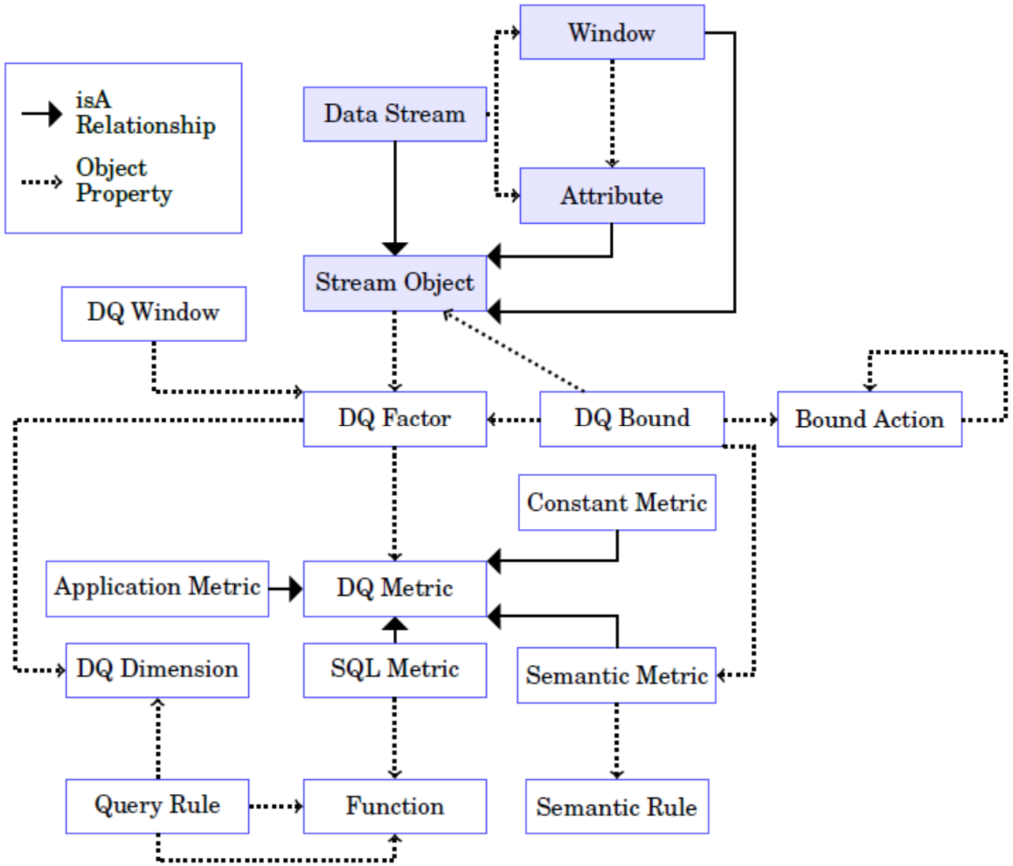
Geisler et al.: Ontology-Based Data Quality Management
for Data Streams. Journal on Data and Information Quality, 2016.

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

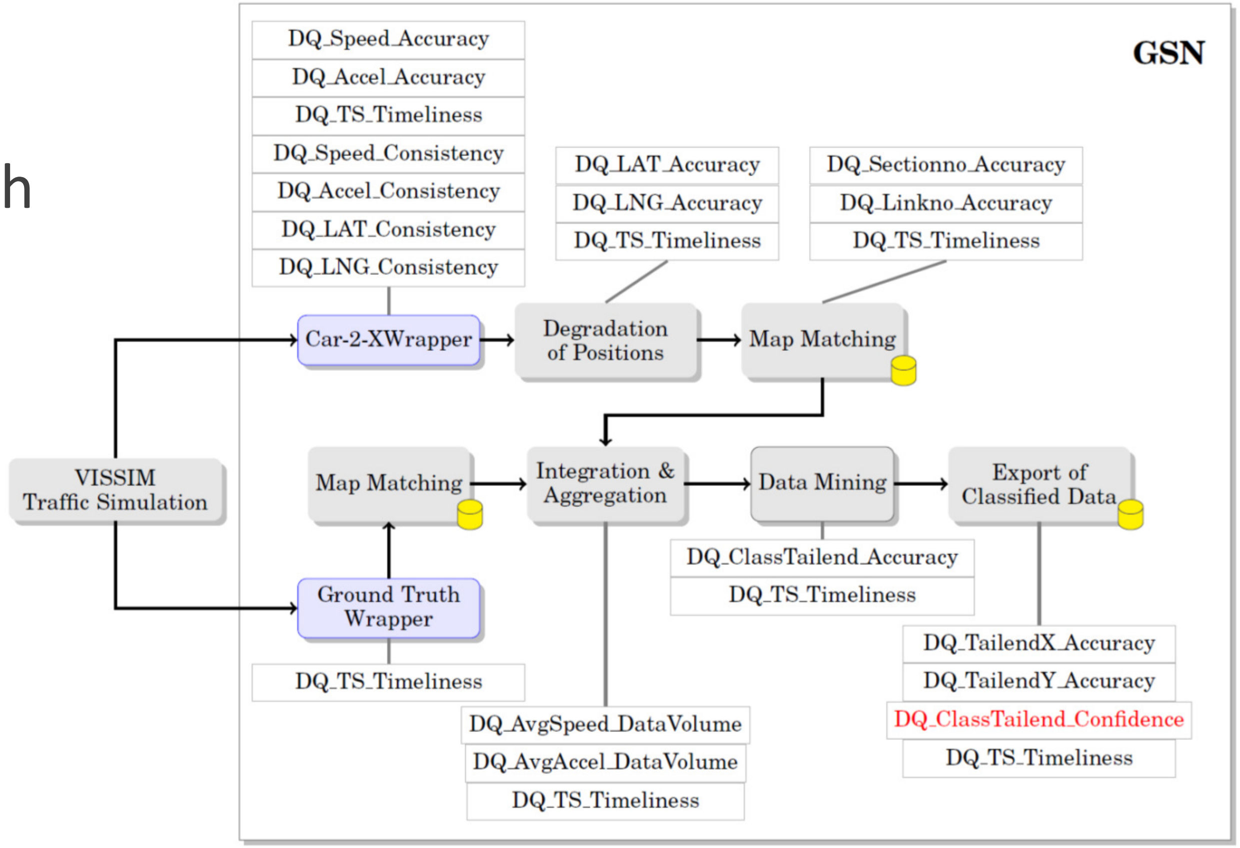
```
Q1: SELECT RoadID, AVG(Speed)
    FROM message
    GROUP BY RoadID

Q2: SELECT RoadID, AVG(Speed),
    COUNT(Speed) AS SpeedDatavolume_DQ
    FROM message
    GROUP BY RoadID
```

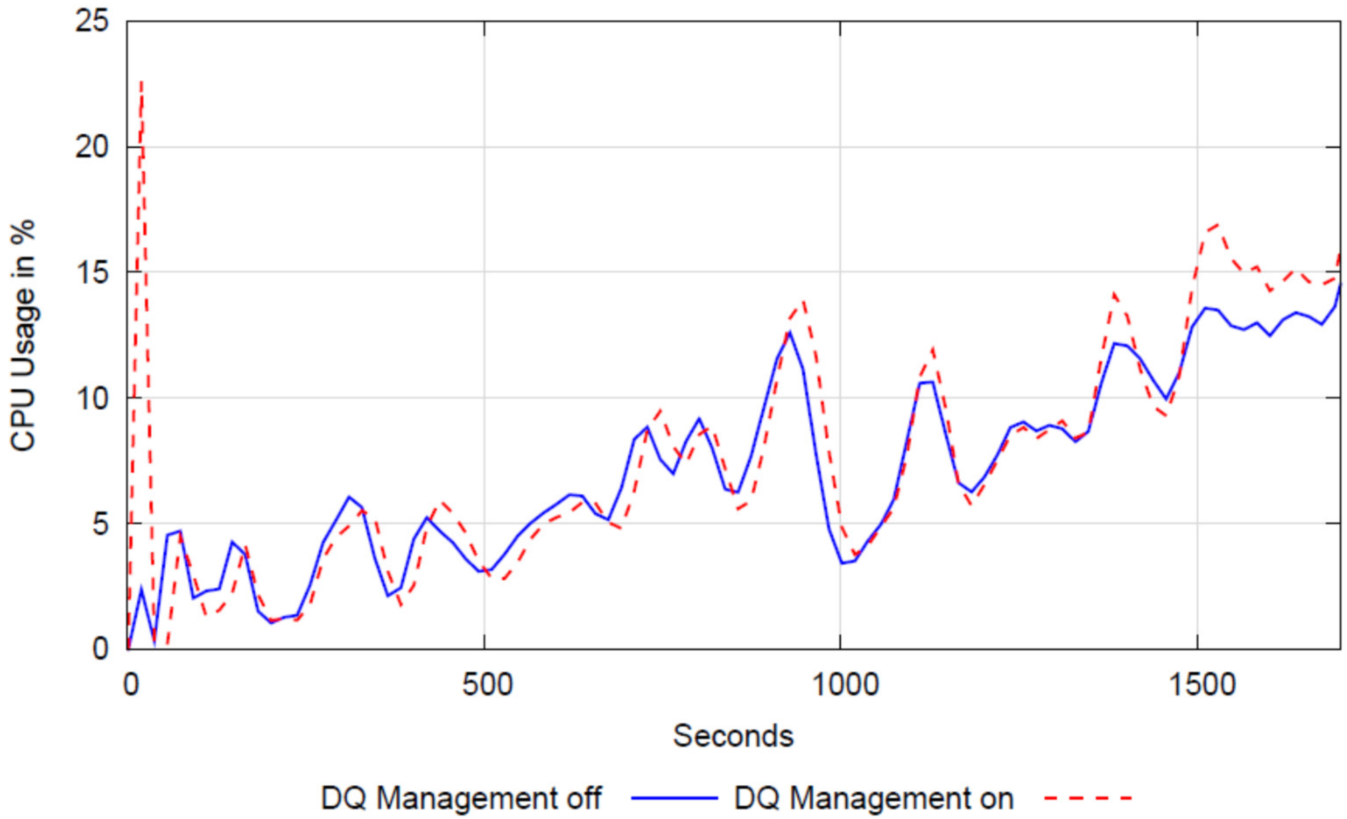
DQ Measurement along the Stream Processing

DQ values are inserted into the data stream

DQ values might depend on each other



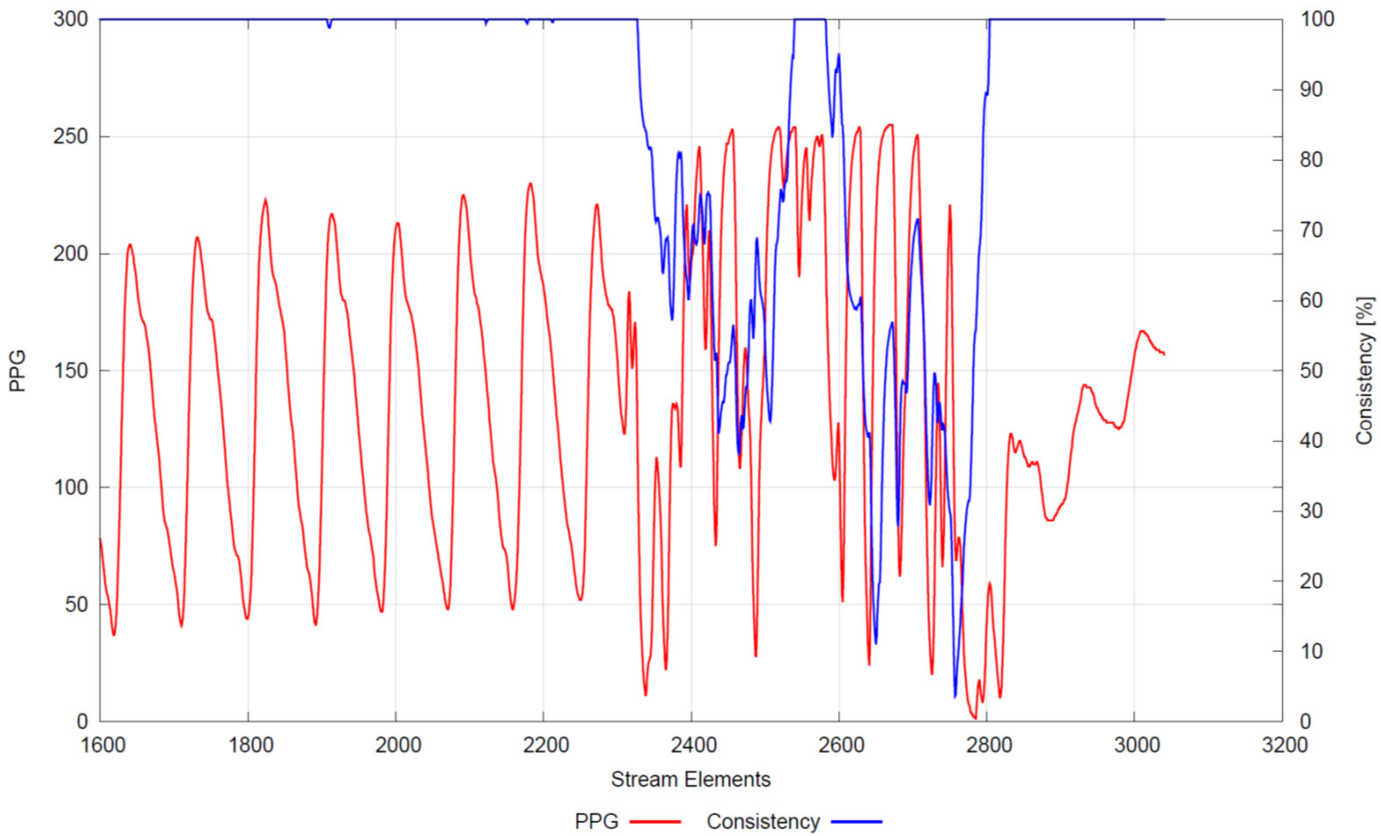
DQ Management does not create much overhead



DQ Monitoring for Health Data

DQ metric tries to measure regularity of PPG curve

Movement might introduce measurement artefacts



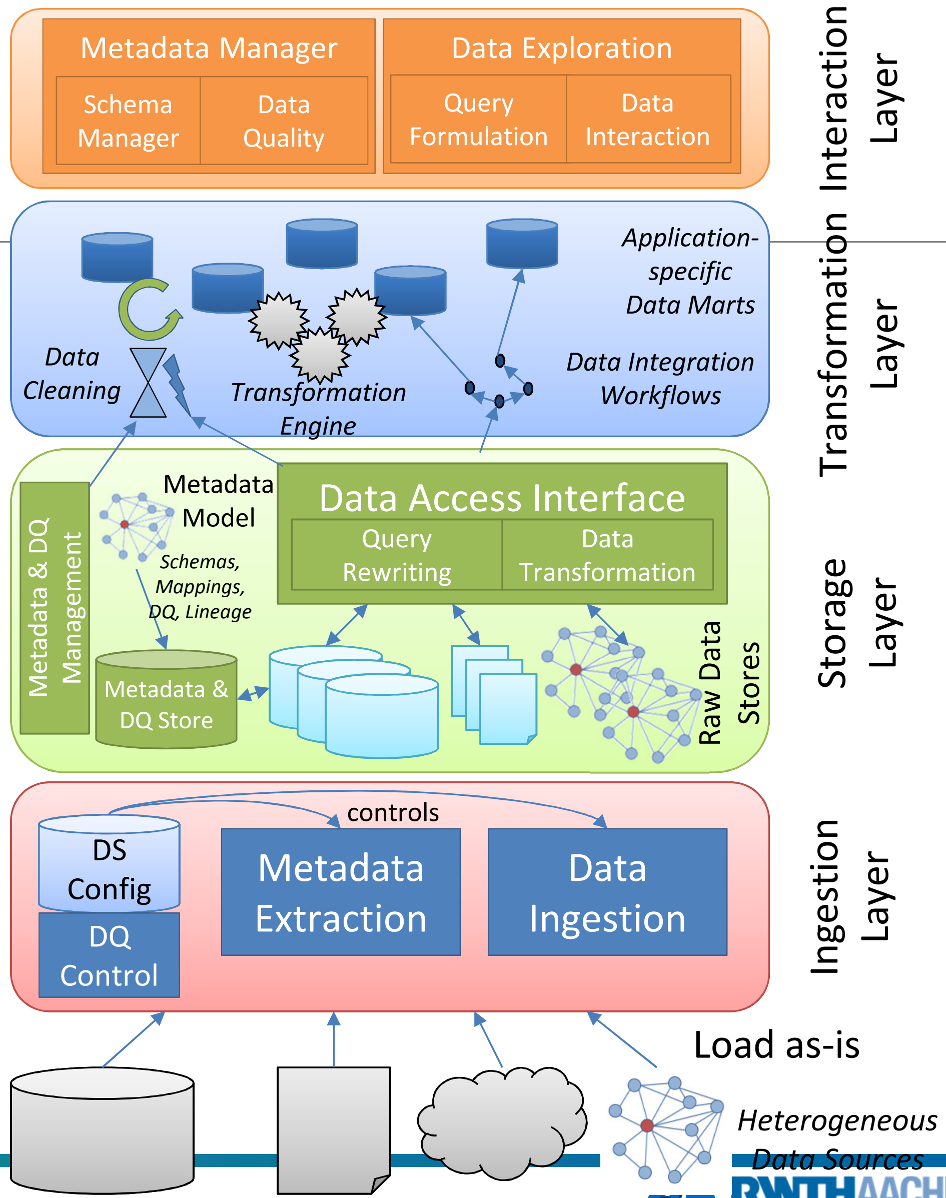
Overview

- 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

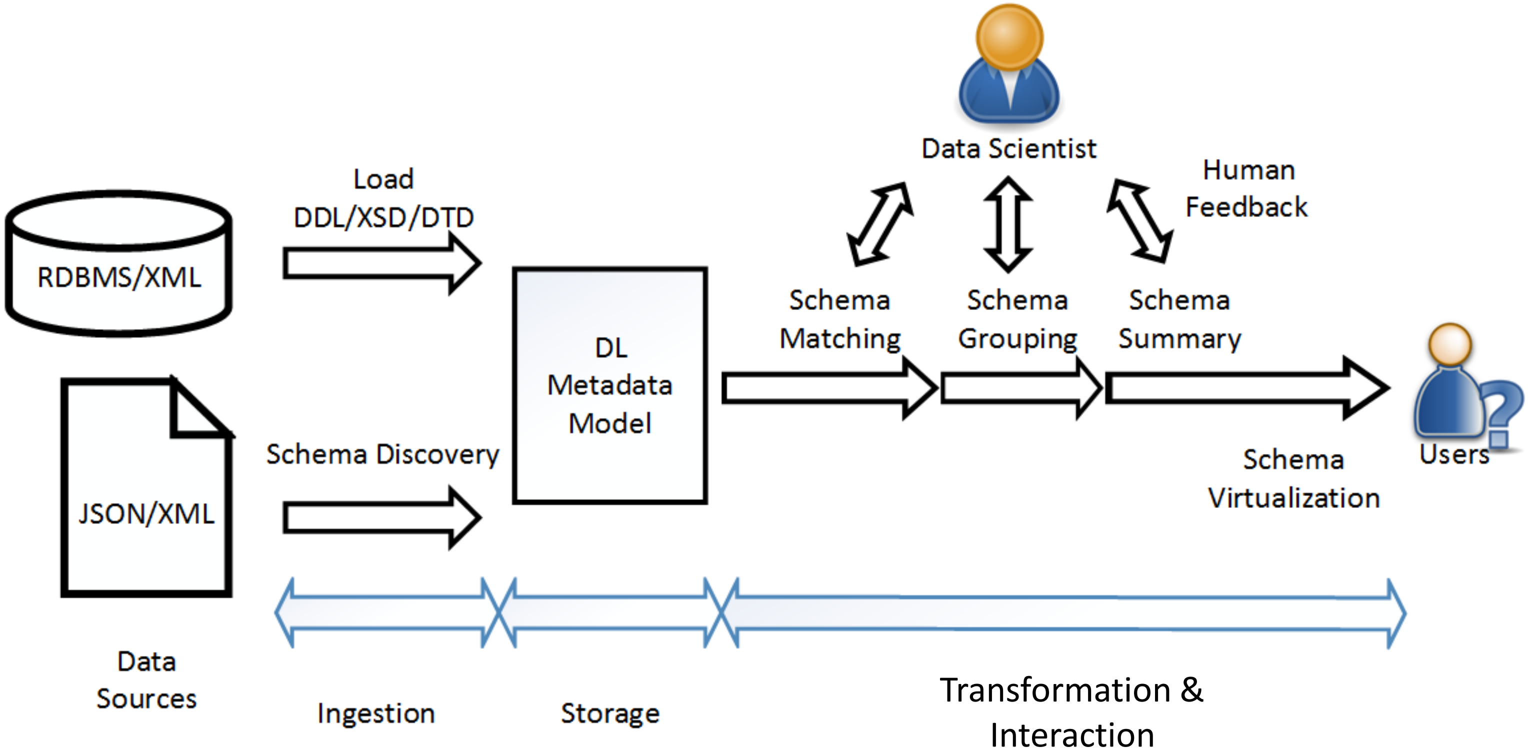
Data Lake Architecture

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



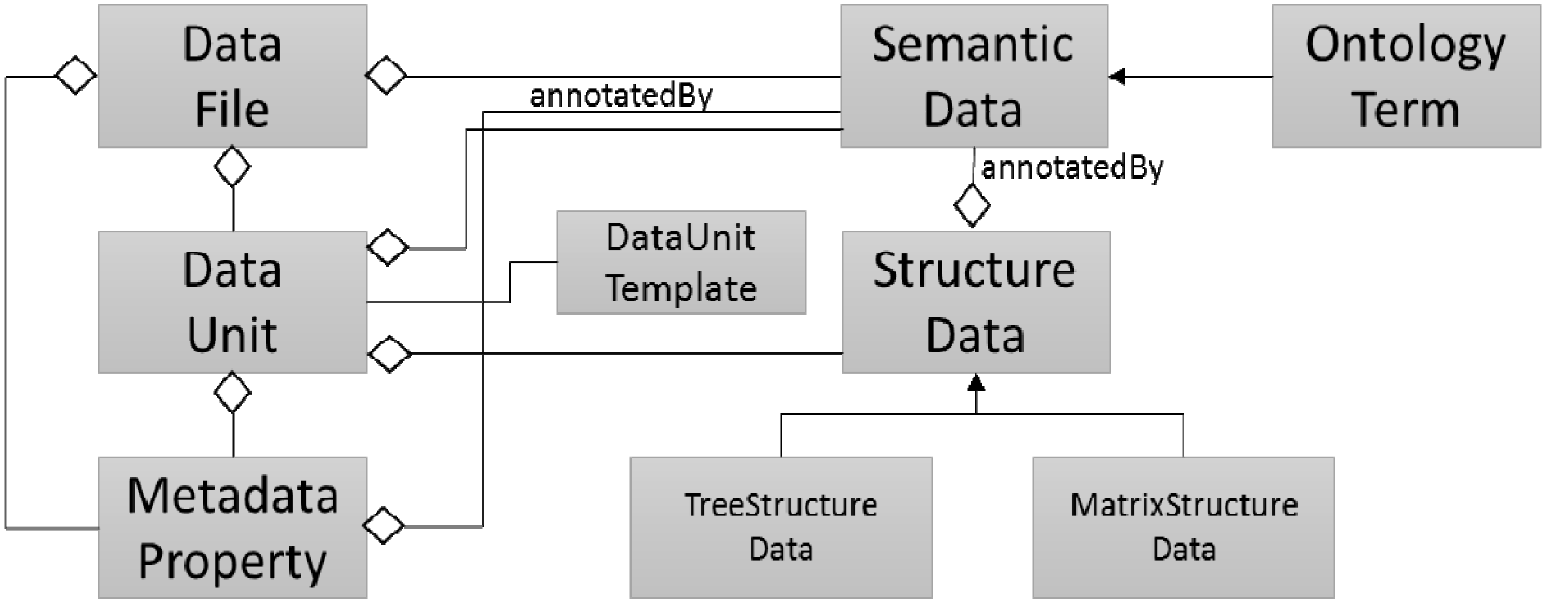
Metadata Types

- Structure data
- Semantic data
- Metadata properties

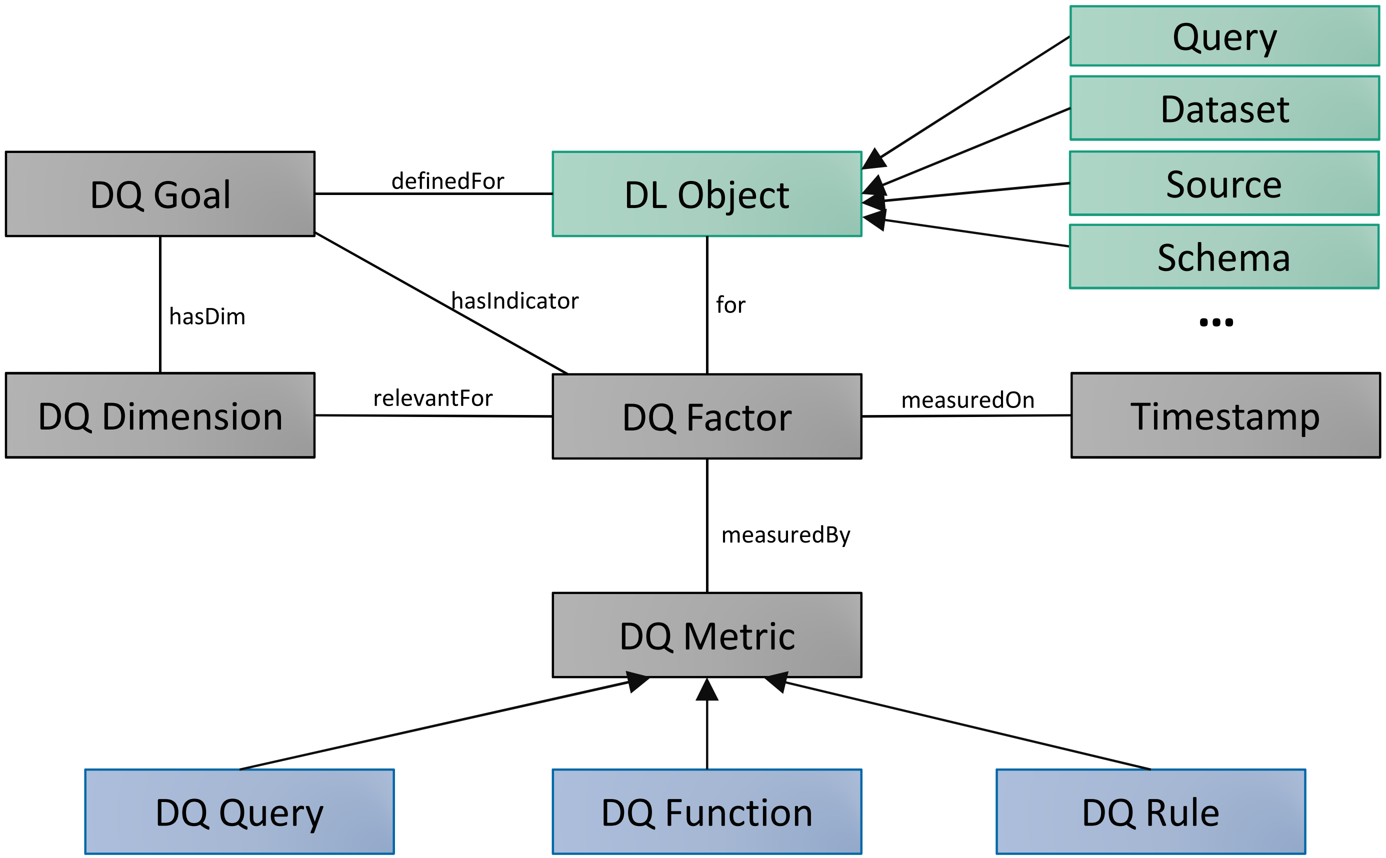
Date	09/2015					
Autor	John Doe					
Label: Label1						
Mode				Measurement from above		
Emission wavelength start				380	nm	
Emission wavelength end				600	nm	
Emissions wavelength step				2	nm	
Scan count				111		
Spectrum (Em)				280...850: 20 nm		
Spectrum (ex) (Sector 1)				230...315: 5 nm		
Spectrum (ex) (Sector 2)				316...850: 10 nm		
	Temperature: 25.5 °C					
WL	380	382	384	386	388	390
E1	966	224	162	171	206	273
E2	477	240	135	168	148	150
E3	627	235	171	174	232	263
E4	280	160	147	214	252	375
E5	657	245	164	167	157	179
E6	159	97	95	101	150	171

Reference: *GEMMS: A Generic and Extensible Metadata Management System for Data Lakes.* C. Quix, R. Hai, I. Vatov. *CAiSE'16 Forum.*

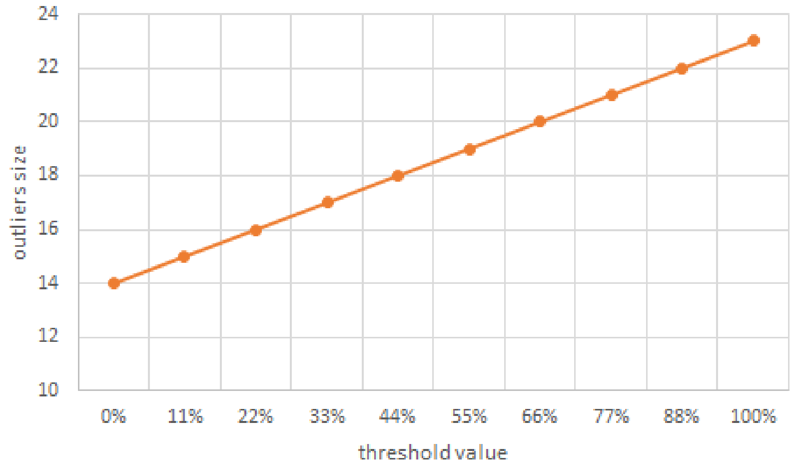
Metadata Model



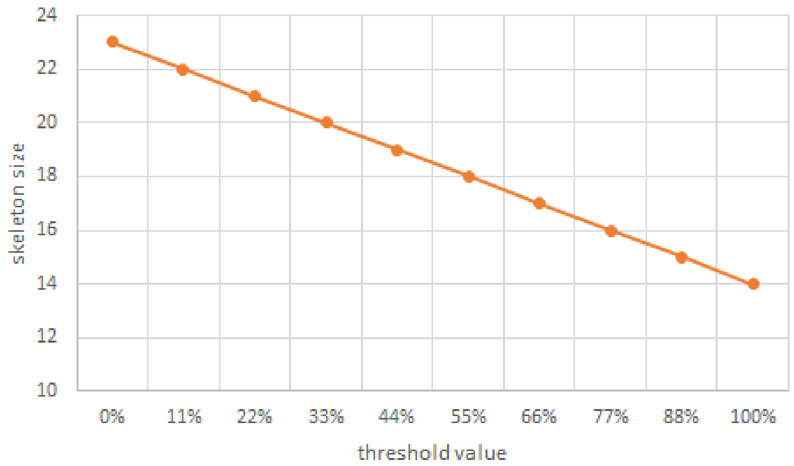
Data Quality Model for Data Lakes



DQ Measurements on Schemas



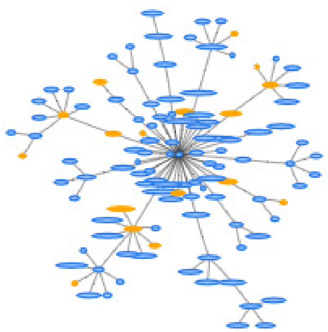
Homogeneity vs. Heterogeneity
in Schema Summarization

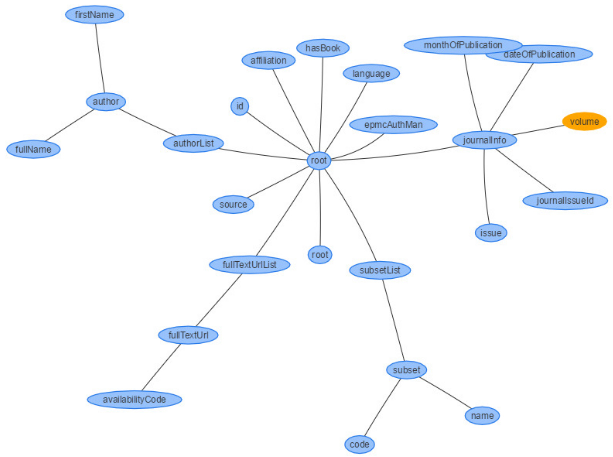


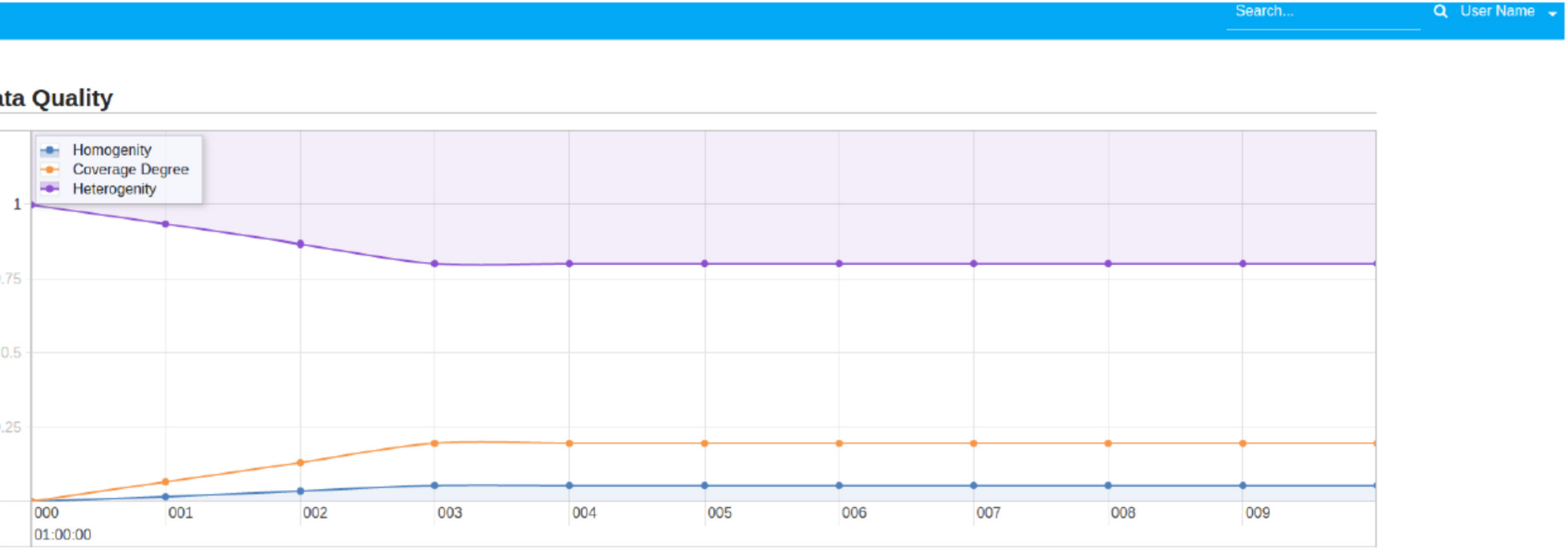
Schema Summary

Concise and usable schema summary against the complex metadata

- Summary Size
- Summary Importance
- Summary Coverage







Search...
User Name
ata Quality
Homogenity
Coverage Degree
Heterogenity
1
.75
0.5
.25
000
01:00:00
001
002
003
004
005
006
007
008
009

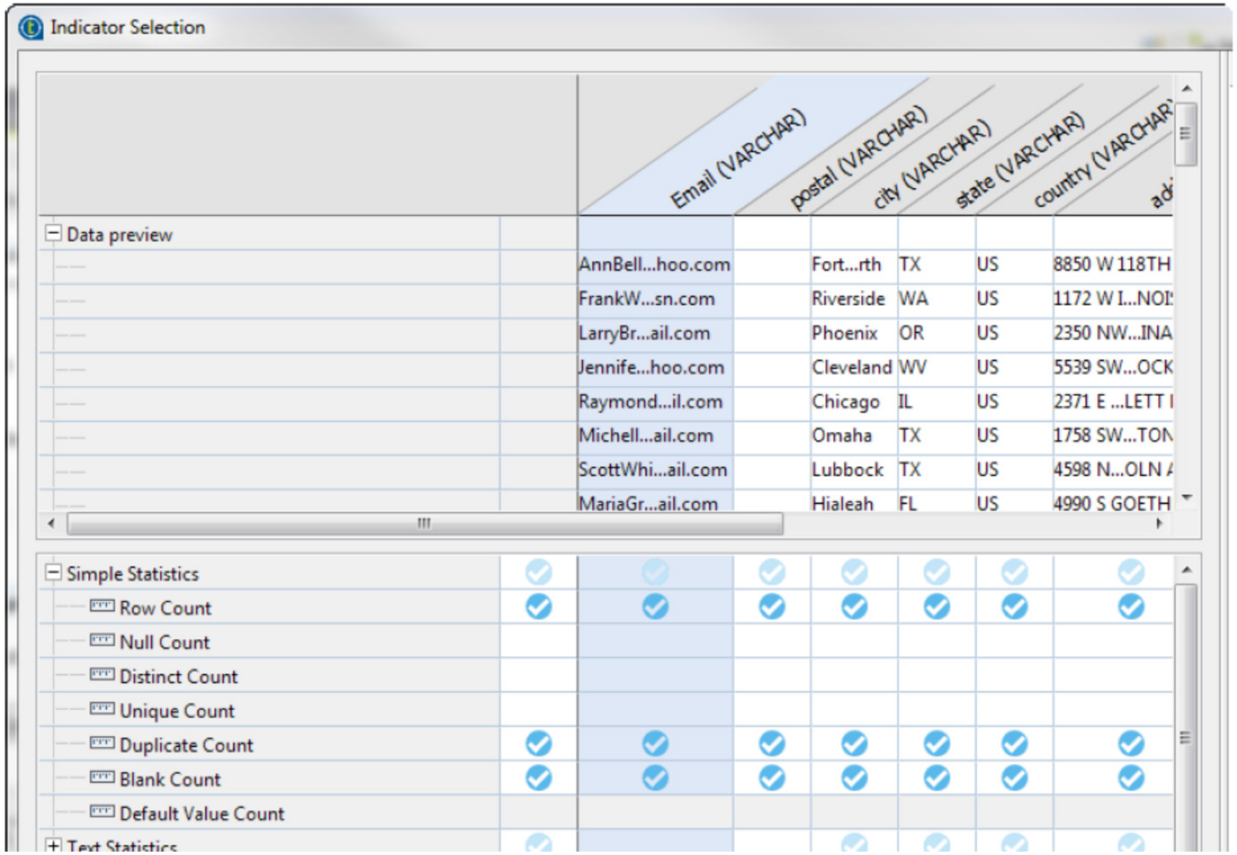
Overview

- 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

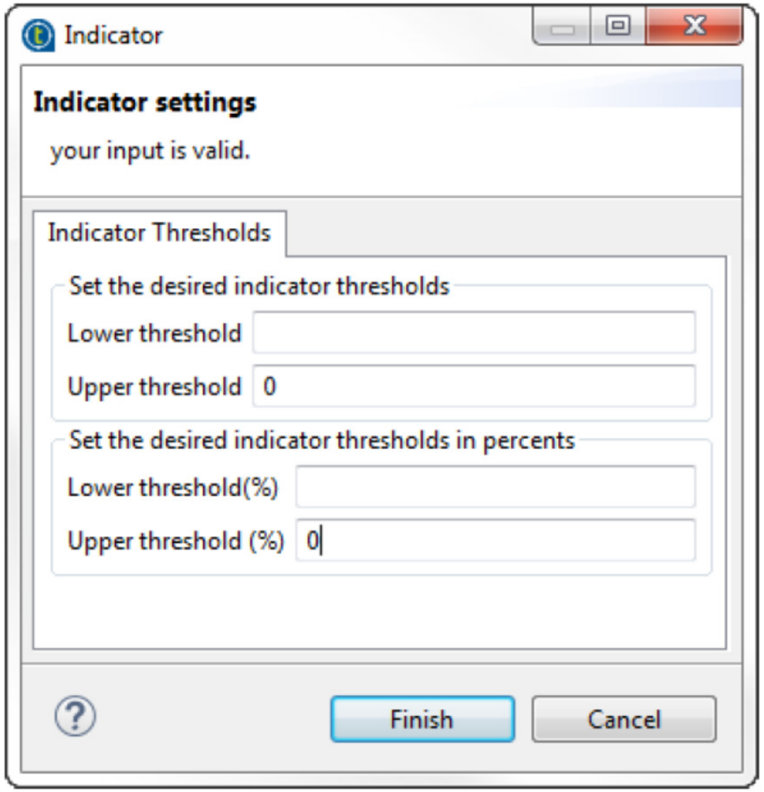
Data Quality Management in Data Integration Tools

- 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)

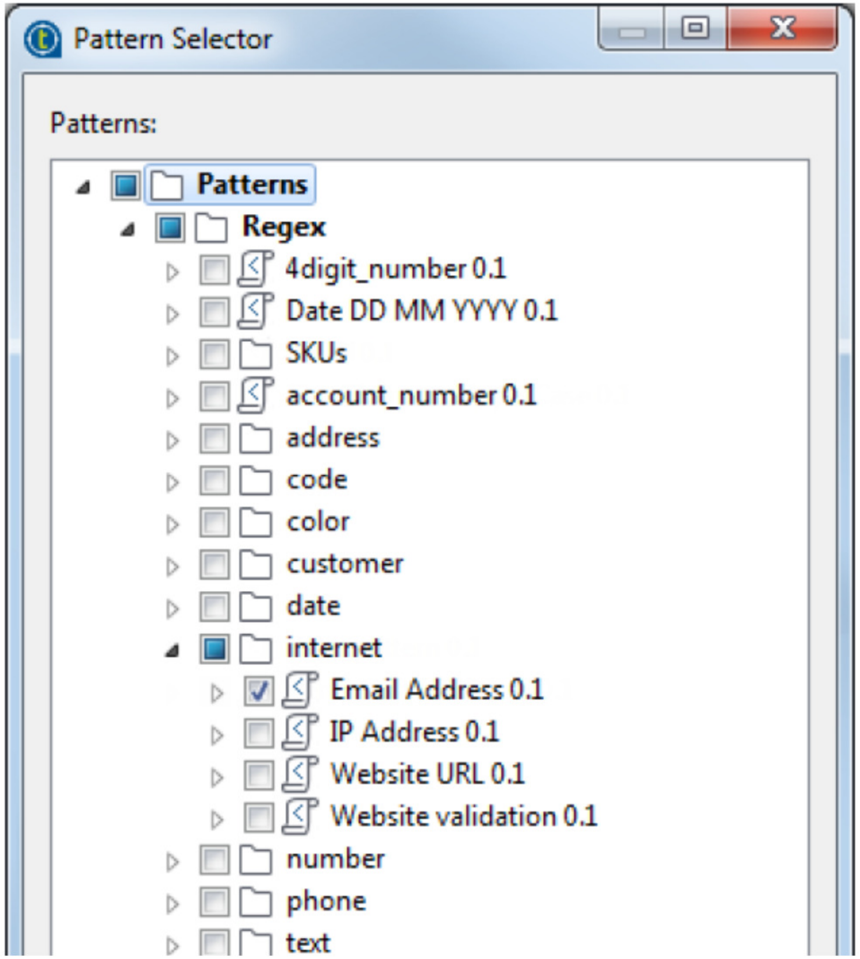
Define Metrics



Define expected ranges



Specific metric for checking patterns in text fields



Analyze results of the measurements

Fraunhofer FIT

RWTH AACHEN UNIVERSITY

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