Big Data and Advanced Analytics

Improving Data Quality For Big Data Using Advanced Analytics

Aytac Ozkan

Supervisor: Prof.Rachid Chelouah

Department of Engineering
Ecole internationale des sciences du traitement de l'information

This dissertation is submitted for the degree of
Master of Big Data

EISTI July 2019

Abstract

Digital data play a crucial role in the information and communication technology (ICT) society: they are managed by business and governmental applications, by all kind of applications on the Web, and are fundamental in all relationships between governments, business, and citizens.

Furthermore, quality of data is also a significant usage for operational process of business and organizations. Some disasters are due to the presence of data quality problems, among them the use of inaccurate, incomplete, out-of-data.

As a consequence, the overall quality of the information that flows between information systems may rapidly degrade over time if both process and their inputs are not themselves subject to quality control. On the other hand, the same networked information system offers new opportunities for data quality management, including possibility of selecting sources with better quality data, and of comparing sources for the purpose of error localization and correction, thus facilitating the control and improvement of data quality in the system.

Due to the described above motivations, researchers and organizations more and more need to understand and solve data quality problems, and thus answering the following questions: What is in essence, data quality? Which techniques, methodologies, and data quality issues are at a consolidated stage?

In this paper, we first review relevant works and discuss machine learning techniques, tools and statistical models. Second, we offer a creative data profiling framework based deep learning and statistical model algorithms for improving data quality.

Keywords: Deep Learning, Statistical Quality Control, Machine Learning, Data Cleaning

Table of contents

List of figures

List of tables

Chapter 1

Introduction to Data Quality

A Web search of terms "data quality" through the search engine Google, returns about three millions of pages and indicator that data quality issues are real and increasingly important (the term data quality will be shortened to the acronym DQ)

1.1 Introduction to the Concept of Data Quality

From a research perspective, data quality has been addressed in different areas, including statistics, management, and computer science. Statisticians were the first to investigate some of the problems related to data quality, by proposing a mathematical theory for considering duplicates in statistical data sets, in the late 1960's. They were followed by researchers in management, who at the beginning of the 1980's focused on how to control data manufacturing systems in order to detect and eliminate data quality problems. Only at the beginning of the 1990's computer scientists begin considering the problem of defining, measuring, and improving the quality of electronic data stored in databases, data warehouses, and legacy systems. [2]

Dr. Genichi Taguchi [11], who was a world-renowned quality engineering expert from Japan, emphasized and established the relationship between poor quality and overall loss. Dr. Taguchi (1987) used a quality loss function (QLF) to measure the loss associated with quality characteristics or parameters. The QLF describes the losses that a system suffers from an adjustable characteristic. According to the QLF, the loss increases as the characteristic y (such as thickness or strength) gets further from the target value (m). In other words, there is a loss associated if the quality characteristic diverges from the target. Taguchi regards this loss as a loss to society, and somebody must pay for this loss. The results of such losses include system breakdowns, company failures, company bankruptcies, and so forth.

Figure 1.1 shows how the loss arising from varying (on either side) from the target by Δ_0 increases and is given by $L(y)$ when y is equal to m,

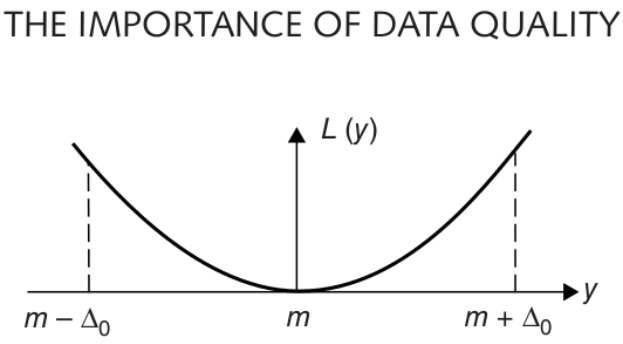

Fig. 1.1 Quality Loss Function (QLF)

the loss is zero, or at the minimum. The equation for the loss function can be expressed as follows:

$$L(y) = k(y-m)^2$$

where k is a factor that is expressed in dollars, based on direct costs, indirect costs, warranty costs, reputational costs, loss due to lost customers, and costs associated with rework and rejection. There are prescribed ways to determine the value of k. The loss function is usually not symmetrical-sometimes it is steep on one side or on both sides. Deming [6] says that the loss function need not be exact and that it is difficult to obtain the exact function. As most cost calculations are based on estimations or predictions, an approximate function is sufficient-that is, close approximation is good enough.

The concept of the loss function aptly applies in the DQ context, especially when we are measuring data quality associated with various data elements such as customer IDs, social security numbers, and account balances. Usually, the data elements are prioritized based on certain criteria, and the quality levels for data elements are measured in terms of percentages (of accuracy, completeness, etc.). The prioritized data elements are referred to as critical data elements (CDEs).

1.2 Why Data Quality is Relevant

The consequences of poor quality of data are often experienced in everyday life, but often, without making the necessary connections to their causes.

For example, the late or mistaken delivery of a letter is often blamed on a postal service, although a closer look often reveals data-related causes, typically an error in the address, orginating in the address database.

Data quality has serious consequences of far-reaching significance, for the efficiency and effectiveness of organizations and business.

1.3 Why Data Quality is Matters

Poor data quality is enemy number one to the widespread, profitable use of machine learning. While the caustic observation, “garbage-in, garbage-out” has plagued analytics and decision-making for generations, it carries a special warning for machine learning. The quality demands of machine learning are steep, and bad data can rear its ugly head twice - first in the historical data used to train the predictive model and second in the new data used by that model to make future decisions. [19]

Data quality is no less troublesome in implementation. Consider an organization seeking productivity gains with its machine learning program. While the data science team that developed the predictive model may have done a solid job cleaning the training data, it can still be compromised by bad data going forward. Again, it takes people — lots of them — to find and correct the errors. This in turns subverts the hoped-for productivity gains. Further, as machine learning technologies penetrate organizations, the output of one predictive model will feed the next, and the next, and so on, even crossing company boundaries. The risk is that a minor error at one step will cascade, causing more errors and growing ever larger across an entire process.

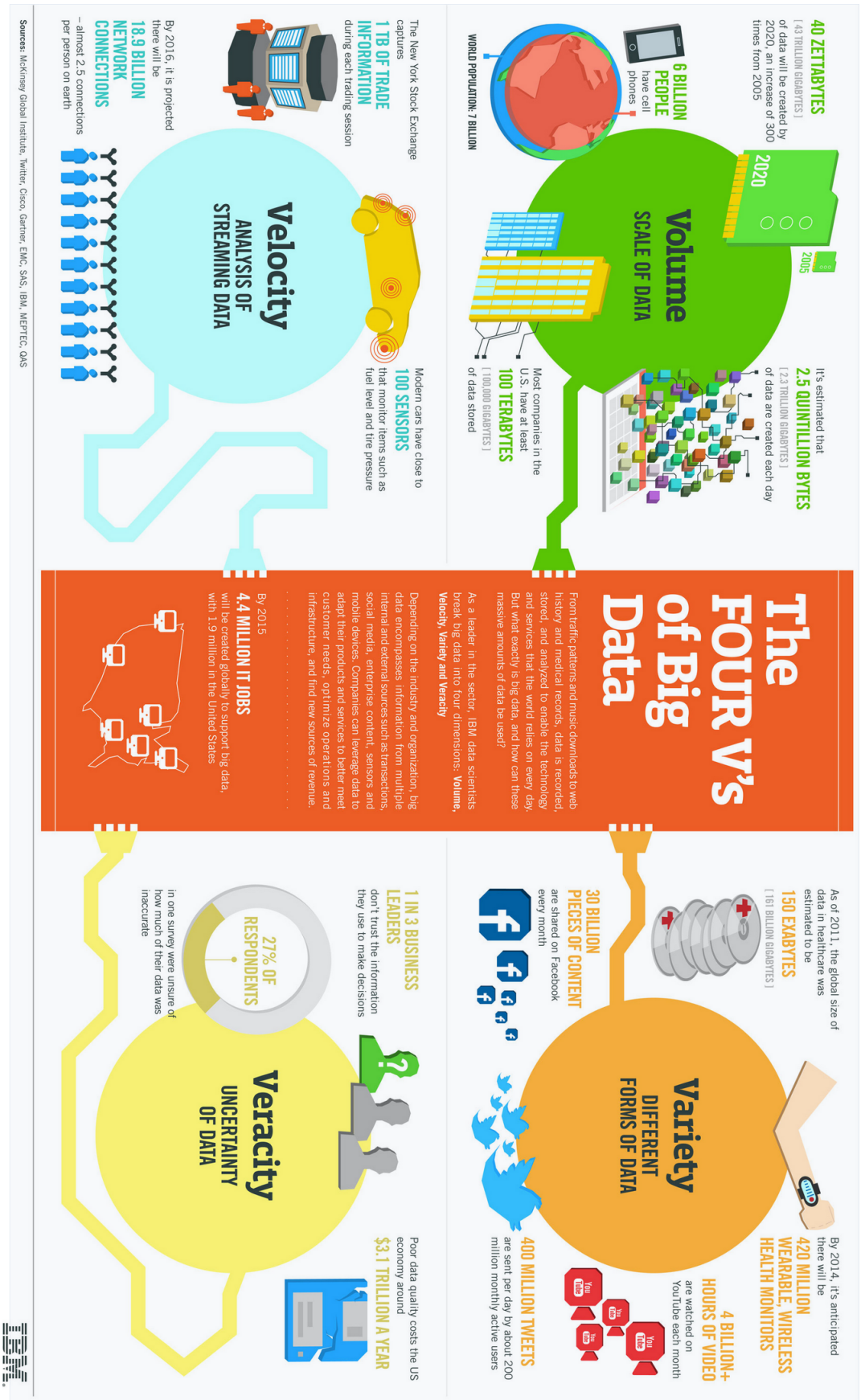

Fig. 1.2 IBM data scientists break big data into four dimensions: volume, variety, velocity and veracity. This infographic explains and gives examples of each. [18]

1.4 Data Quality and Types of Information Systems

Data are collected, stored, elaborated, retrieved, and exchanged in information systems used in organizations to provide services to business processes. Different criteria can be adopted for classifying the different types of information systems, and their corresponding architectures; they are usually related to the overall organizational model adopted by the organization or the set of the organizations that make use of the information system.

The three classifications are represented together in the classification space of Figure 1.2. Among all possible combinations, five main types of information systems are highlighted in the figure: Monolithic, Distributed, Data Warehouses,Cooperative, and Peer-to-Peer.

- In a *monolithic information system* presentation, application logic, and data management are merged into a single computational node. Many monolithic information systems are still in use. While being extremely rigid, they provide advantages to organizations, such as reduced costs due to homogenetiy of solutions and centralization of management. In monolithic systems data flows have a common format, and data quality control is facilitated by the homogenetiy and centralization of procedures and management rules.

- A *data warehouse* (DW) is a centralized set of data collected from different sources, designed to support management decision making. The most critical problem in DW design concerns the cleaning and integration of the different data sources that are loaded into the DW, in that much of the implementation budget is spent on data cleaning activities.

- A *distributed information system* relaxes the rigid centralization of monolithic systems, in that it allows the distribution of resources and applications across network of geographically distributed systems. The network can be organized in terms of several tiers, each made of one or more computational nodes. Presentation, application logic, and data management are distributed across tiers. Usually, the different tiers and nodes have a limited degree of autonomy, data design is usually performed centrally, but to a certain extent some degree of heterogenetiy can occur, due to the impossibility of establishing unified procedures. Problems of data management are more complex than in monolithic systems, due to the reduced level of centralization.

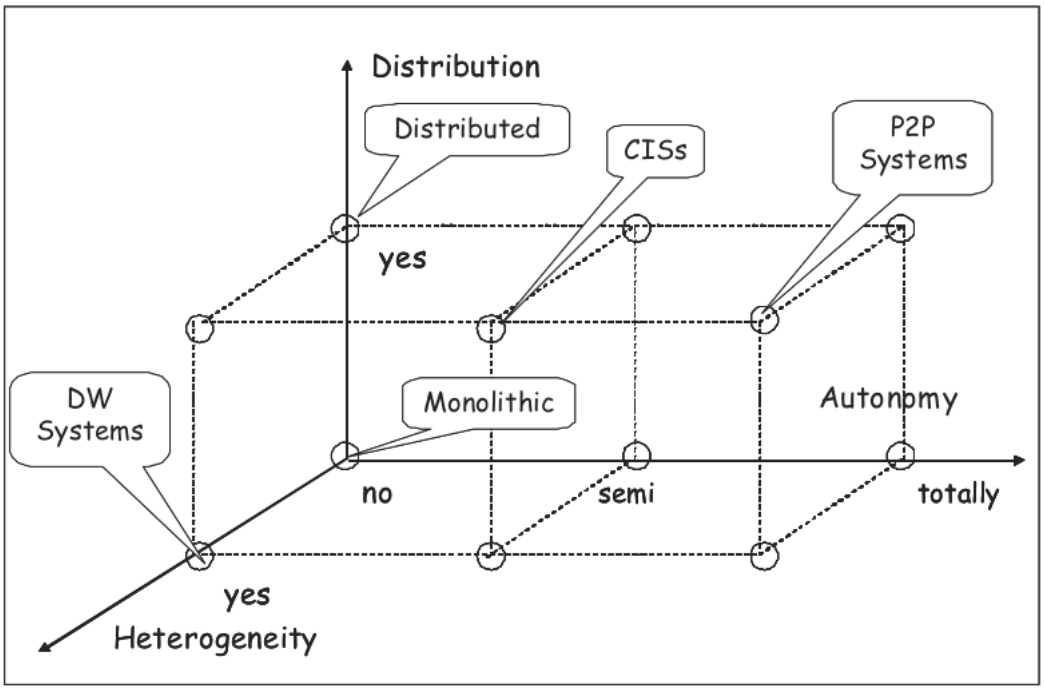

Fig. 1.3 Types of information systems

- A *coorparative information system* (CIS) can be defined as a large-scale information system that interconnects various systems of different and autonomous organizations, while sharing common objectives.
- In a *peer to peer information system* (usually abbreviated P2P), the traditional distinction between clients and servers typical of distributed systems is disappearing. A P2P system can be characterized by a number of properties: peers are highly autonomous and higly heterogeneous, they have no obligation for the quality of their services and data, no central coordination and no central database exist, no peer has a global view of the system, global behavior emerges from local intreactions.

1.5 Main Research Issues and Application Domains in Data Quality

Due to the relevance of data quality, its nature, and the variety of data types and information systems, achieving data quality is a complex, multidisciplinary area of investigation. I involves several research topics and real-life application areas

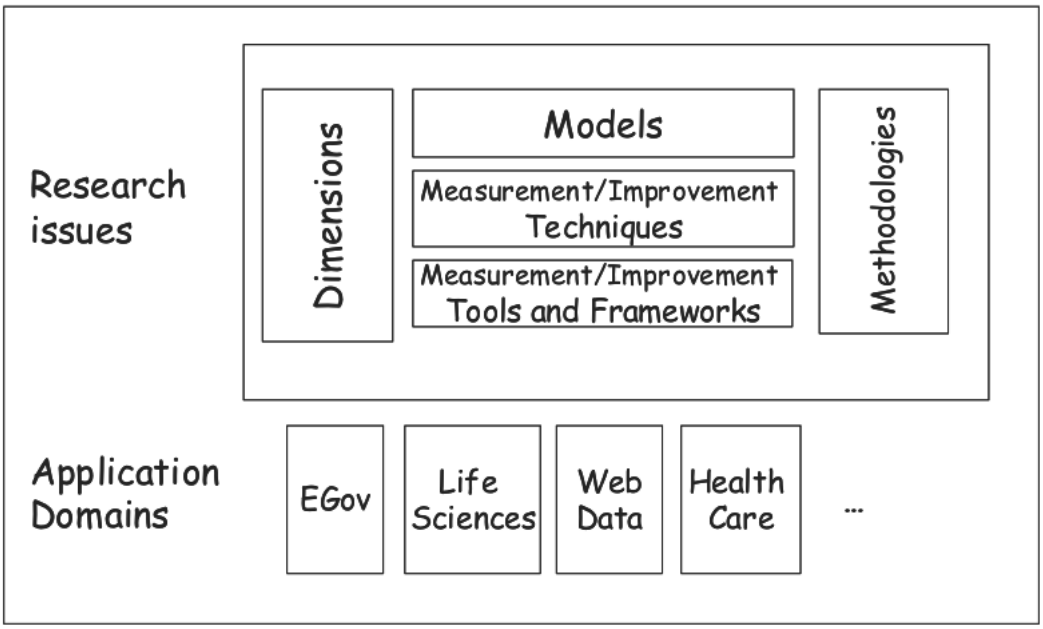

Fig. 1.4 Main issues in data quality

Chapter 2

Data Quality Dimensions

More specifically, quality dimensions can refer either to the extension of data, i.e., to data values, or to their intension, i.e., to their schema. Both data dimensions are usually defined in qualitative way, referring to general properties of data and schemas, and and related definitions do not provide any facility for assigning values to dimensions themselves.

Specifically, definitions do not provide quantitative measures, and one or more metrics are to be associated with dimensions as separate, distinct properties. For each metric, one or more measurement metadata are to be provided regarding (i) where the measurement is taken, (ii) what data are included, (iii) the measurement device, and (iv) the scale on which results are reported.

According to the literature, at times we will distinguish between dimensions and metrics, while other times we will directly metrics.

Table 2.1 Data Quality Dimensions I

Dimension Name	Description
Data Governance	Do organization-wide data standards exist and are they enforced? Do clearly defined roles and responsibilities exist for data quality related activities? Does data governance strive to acquire and maintain high-quality data through proactive management?
Data Specifications	Are data standards documented in the form of a data dictionary, data models, meta data, and integrity constraints?
Data Integrity	How is data integrity maintained? How are data integrity violations detected and resolved ?
Data Consistency	If data redundancy exists, how is data consistency achieved? What methods are used to bring consistency to data that has become inconsistent? If data is geographically replicated, how is the consistency and latency managed?
Data Currency	Is the data current? Do procedures exists to keep the data current and purge stale data?
Data Duplication	Are there effective procedures in place to detect and remove duplicate data?
Data Completeness	Is the data about entities complete? How is missing data managed?
Data Provenance	Is a historical record of data and its origination maintained? If the data is acquired through multiple sources and has undergone cleaning and transformations, does the organization maintain a history of all changes to the data?
Data Heterogeneity	If multi-modality data about an entity is available, is that data captured and used?
Streaming Data	How is streaming data sampled, filtered, stored, and managed for both real-time and batch processing?
Outliers	How are outliers detected and addressed? Are there versions of datasets that are outlier-free? Does each version correspond to a different method for outlier detection and treatment?
Dimensionality	Reduction Do the datasets feature dimensionality reduced versions? How many versions are available?
Feature Selection	Do datasets have versions that exclude features that are either redundant, highly correlated, or irrelevant? How many versions are available?

Table 2.2 Data Quality Dimensions II

Dimension Name	Description
Feature Extraction	Do the datasets provide a set of derived features that are informative and non- redundant, in addition to the original set of variables/features? How many such derived feature sets are available?
Business Rules	Does a process exist to identify, refine, consolidate, and maintain business rules that pertain to data quality? Do rules exist to govern data cleaning and transformations, and integrating related data of an entity from multiple sources? What business rules govern substitutions for missing data, deleting duplicate data, and archiving historical data? Are there rules for internal data audit and regulatory compliance?
Data Accuracy	Data can be syntactically accurate and yet semantically inaccurate. For example, a customer's mailing address may meet all the syntactic patterns specified by the postal service, yet it can be inaccurate. How does the organization establish the accuracy of data?
Gender Bias	Is the data free from factors that lead to gender bias in machine learning algorithms?
Confidentiality and Privacy	Are procedures and controls implemented for data encryption, data de- identification and re-identification, and differential privacy?
Availability and Access Controls	How is high data availability achieved? What security controls are implemented to protect data from unauthorized access? How are user entitlements to data access and modifications defined and implemented?

2.1 Accuracy

Accuracy [8] is defined as the closeness between a value v and a value v' , considered as the correct representation of the real-life phenomenon that v aims to represent. As an example if the name of a person is Ayta?, the value v' = Ayta? is correct, while the value v = Ayt is incorrect. Two kinds of accuracy can be identified, namely a syntactic accuracy and a semantic accuracy.

Let us consider a relation schema **R** consisting of **K** attributes and a relational table **r** consisting of N tuples.

Let $q_{ij}(i = 1..N, j = 1..K)$ be a boolean variable defined to correspond to the cell values y_{ij}, is syntactically accurate, while otherwise it is equal to 1.

In order to identify whether or not accuracy errors affect a matching of relational table **r** with a reference table **r'** containing correct values, we introduce a further boolean variable s_i equal to 0 if the tuple t_i matches a tuple in **r'**, and otherwise equal to 1. We can introduce three metrics to distinguish the relative importance of value accuracy in context of the tuple. The first two metrics have the purpose of giving a different importance to errors on attributes that have a higher identification power, in line with the above discussion.

The first metric called *weak accuracy error,* and is defined:

$$\sum_{i=1}^{N} \frac{\beta(q_i > 0) \wedge (s_i = 0)}{N}$$

where $\beta(.)$ is a boolean variable equal to 1 if the condition in parentheses is *true*, 0 otherwise, and $q_i = \sum_{j=1}^{K} q_{ij}$. Such metric considers the case in which for a tuple t_i accuracy errors occur $(q_i > 0)$ but do not affect identification $(s_i = 0)$.

The second metric is called *strong accuracy error*, and is defined assigning

$$\sum_{i=1}^{N} \frac{\beta(q_i > 0) \wedge (s_i = 1)}{N}$$

where $\beta(.)$ and q_i have the same meaning as above. Such a metric considers the case which accuracy errors occur $(q_i > 0)$ for a tuple t_i and actually do affect identification $(s_i = 1)$.

The third metric gives the percentage of accurate tuples matched with the reference table. It is expressed by the degree of syntactic accuracy of the relational instance **r**

$$\sum_{i=1}^{N} \frac{\beta(q_i = 0) \wedge (s_i = 0)}{N}$$

by actually considering the fraction of accurate $(q_i = 0)$ matched $(s_i = 0)$ tuples.

2.2 Completeness

Completeness can be generically defined as the extent to which data are of sufficient breadth, depth, and scope for the task at hand [24] three types of completeness are identified. Schema completeness is defined as the degree to which concepts and their properties are not missing from the schema. Column completeness is defined as a measure of the missing values for a specific property or column in a table. Population completeness evaluates missing values with respect to a reference population. If focusing on a specific data model, a more precise

characterization of completeness can be given. In the following we refer to the relational model.

2.2.1 Completeness of Web Data

Data that are published in Web information systems can be characterized by evolution in time. While in the traditional paper-based media, information is published once and for all, Web information systems are characterized by information that is continuously published.

We consider a function $C(t)$, defined as the value of completeness at the instant t, with $t \in [t\_pub, t\_max]$, where t pub is the initial instant of publication of data and t_max corresponds to the maximum time within which the series of the different scheduled updates will be completed. Starting from the function $C(t)$, we can define the completability of the published data as

$$\int_{t\_curr}^{t\_max} C(t),$$

where t_curr is the time at which completability is evaluated and $t\_curr < t\_max$.

Completability, as shown in Figure 2.1, can be graphically depicted as an area Cb of a function that represents how completeness evolves between an instant t_curr of observation and t max. Observe that the value corresponding to t_curr is indicated as c_curr; c_max is the value for completeness estimated for t max. The value c_max is a real reachable limit that can be specified for the completeness of the series of elements; if this real limit does not exist, c_max is equal to 1. In Figure 2.5, a reference area A is also shown, defined as

$$(t\_max - t\_curr) * \frac{c\_max - c\_pub}{2},$$

that, by comparison with Cb, allows us to define ranges [High, Medium, Low] for completability.

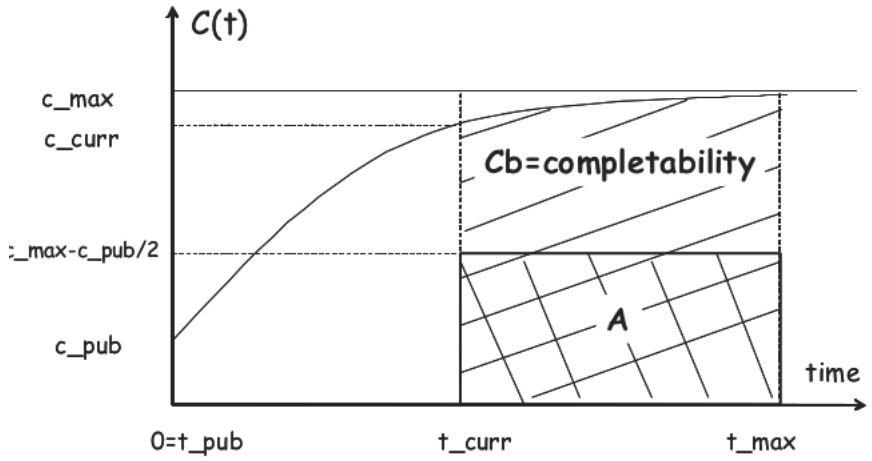

Fig. 2.1 A graphical representation of completability

With respect to the example above, considering the list of courses published on a university Web site, the completeness dimension gives information about the current degree of completeness; the completability information gives the information about how fast this degree will grow in time, i.e., how fast the list of courses will be completed. The interested reader can find further details in [17].

2.3 Consistency

The consistency dimension captures the violation of semantic rules defined over (a set of) data items, where items can be tuples of relational tables or records in a file. With reference to relational theory, integrity constraints are an instantiation of such semantic rules. In statistics, data edits are another example of semantic rules that allow for the checking of consistency.

2.3.1 Integrity Constraints

Integrity constraints are properties that must be satisfied by all instances of a database schema. Although integrity constraints are typically defined on schemas, they can at the same time be checked on a specific instance of the schema that presently represents the extension of the database. Therefore, we may define integrity constraints for schemas, describing a schema quality dimension, and for instances, representing a data dimension.

Most of the considered integrity constraints are dependencies. The following main types of dependencies can be considered:

- *Key Dependency.* This is the simplest type of dependency. Given a relation instance r, defined over a set of attributes, we say that for a subset K of the attributes a key dependency holds in r, if no two rows of r have the same K-values. For instance, an attribute like SocialSecurityNumber can serve as a key in any relation instance of a relation schema Person.

- *Inclusion Dependency.* Inclusion dependency is a very common type of constraint, and is also known as referential constraint. An inclusion de- pendency over a relational instance r states that some columns of r are contained in other columns of r or in the instances of another relational in- stance s. A foreign key constraint is an example of inclusion dependency, stating that the referring columns in one relation must be contained in the primary key columns of the referenced relation.

- *Functional Dependency.* Given a relational instance **r**, let **X** and **Y** be two nonempty sets of attributes in **r**. **r** satisfies the functional dependency $\mathbf{X} \rightarrow \mathbf{Y}$ if the following holds for every pair of tuples t_1 and t_2 in **r**:

$$\boxed{If, t_1.X = t_2.X, then, t_1.Y = t_2.Y,}$$

where the notation $t_1.X$ means the projection of the tuple t_1 onto the attributes in X.

2.4 Other Data Quality Dimensions

There are general proposals for sets of dimensions that aim to fully specify the data quality concept in a general setting Some other proposals are related to specific domains that need ad hoc dimensions in order to capture the peculiarities of the domain. For instance, specific data quality dimensions are proposed in the following domains:

1. The archival domain (see [25] and [12]) which makes use of dimensions such as condition (of a document) that refers to the physical suitability of the document for scanning.

2. The statistical domain; every National bureau of census and international organizations such as the European Union or the International Monetary Fund define several dimensions for statistical and scientific data, such as integrity, on the notion that statistical systems should be based on adherence to the principle of objectivity in the collection, compilation, and dissemination of statistics

3. The geographical and geospatial domain (see [16] [9]), where the following dimensions are proposed: (i) positional accuracy, defined as a quality parameter indicating the accuracy of geographical positions, and (ii) attribute/thematic accuracy, defined as the positional and/or value accuracy of properties such as sociodemographic attributes in thematic maps.

Chapter 3

Data Quality (DQ) Evaluation

3.1 Metrics and Measurement

Any data can have its quality measured. Using a data driven strategy, the measurements acts on the data itself to quantify the DQD (Data Quality Dimension). As mentioned before, our work is based on structured data represented in [20] a set of attributes, columns, and rows with their values. Any data quality metric should specify whether the values of data respect or not the quality attributes. The data quality measurement metrics tend to evaluate a binary results correct or incorrect or a value between 0 and 100, and use universal formulas to compute these attributes. This will apply to many quality dimensions such as accuracy, completeness, and consistency.

		Data Quality Dimensions Related		
	Data Quality Issues	**Accuracy**	**Completeness**	**Consistency**
Instance Level	*Missing Data*	X	X	
	Incorrect data, Data entry errors,	X		
	Irrelevant data			X
	Outdated data	X		
	Misfiled and Contradictory values	X	X	X
Schema Level	*Uniqueness constrains, Functional dependency violation*	X		
	Wrong data type, poor schema design			X
	Lack of integrity constraints	X	X	X

Table 3.1 Data Quality Issues vs DQD

The DQDs (Data Quality Dimensions) must be relevant to the DQ problems as identified In Table 3.1 Therefore DQ Metrics are designed for each DQD to measure if the attributes respect the previously defined DQD. These measures are done for each attribute given it type, data ranges values, and if it is collected from data profiling.

For example a metric that calculates the accuracy of a data attribute is defined as follows:

- The data type of an attribute and its values.
- For numerical attributes a range or sets of acceptable values (Textual also) are defined. Any other values are incorrect.
- The accuracy of an attribute is calculated based on the number of correct values divided by number of observations or rows.
- For another data types/formats like images, videos, audio files, another type of metrics must be defined to evaluate accuracy or any other quality dimensions. The authors of [5] describe usefulness as an aspect of data quality for images. For this kind of data, feature extraction functions are defined on the data and extracted for each data item. These features have constraints that characterize the goodness or badness of data values. Some of quality metrics functions are designed based on the extracted features such as, usefulness, accuracy , completeness and any other data quality dimensions judged by domain experts to be candidate for such data type.

3.2 DQ Issues and Big Data Characteristics

Data characteristics commonly named V's are initially, Volume, Velocity, Variety, and Veracity. Since the Big Data inception; we reached now 7 V's and probably we will keep going [13]. The veracity tends more to express and describe trust and certainty of data that can be expressed mostly as quality of the data. The DQD accuracy is often related to precision, reliability and veracity [23].

A mapping tentative between these characteristics, data and data quality is complied in [10] [5] [Cai and Zhu]. The authors attempted to link the V's to the quality dimensions.

DQ Dimensions	Metric functions
Accuracy	Acc = (Ncv / N)
Completeness	Comp = (Nmv / N)
Consistency	Cons = (Nvrc / N)
Ncv	Number of correct values
Nmv	Number of missing values
Nvrc	Number of values that respects the constraints
N	Total number of values (rows) of the sample Dataset

Table 3.2 DQD metric functions

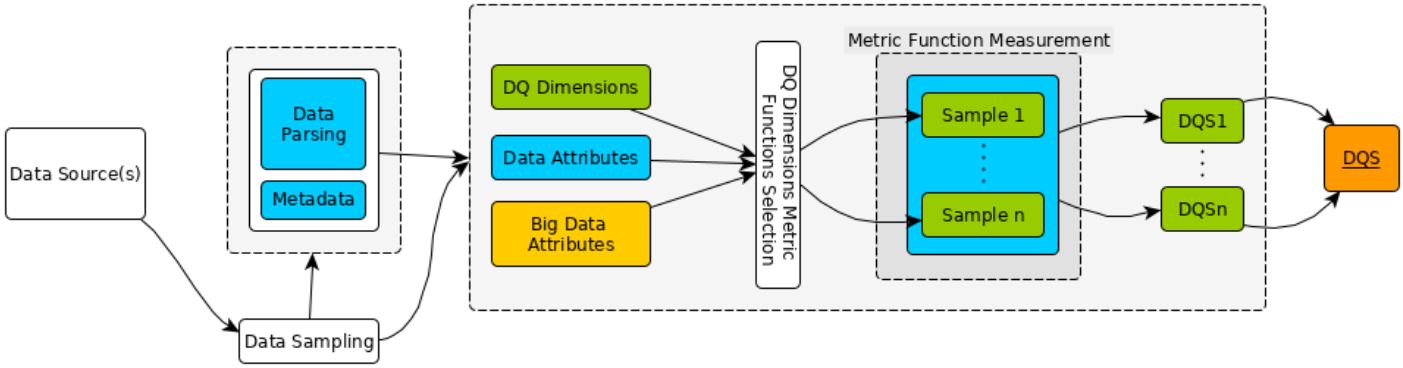

Fig. 3.1 Big Data Quality Evaluation Scheme

3.3 Big Data Quality Evaluation Scheme

The purpose of Big Data Quality Evaluation (BDQ) Scheme is to address the data quality before starting data analytics. This is done by estimating the quality of data attributes or features by applying a DQD metric to measure the quality characterized by its accuracy, completeness or/and consistency. The expected result is data quality assessment suggestions indicating the quality constraints that will increase or decrease the data quality.

The BDQ Evaluation scheme is illustrated in Figure 3.1 where the data goes through many module to estimate its quality. The key modules of our scheme consist of: (a) data sampling; and data profiling, (b) DQD vs attributes selection, (c) data quality Metric selection,

(d) samples data quality evaluation. In the following sections, we describe each module, its input(s), output(s), and the main functions.

3.3.1 Big Data Sampling

A sample is representative of a whole population. Based on a sample, we make several decision about a population. A sample is also called a subgroup. The number of observations or units in a sample is called sample size. The number of times a sample is collected is usually referred to as the sampling frequency. In designing a control chart, we must specify both of these parameters. [11]

There are several sampling strategies that can be applied on Big Data as expressed in [22] [4]. They evaluated the effect of sampling methods on Big Data and believed that sampling large datasets reduces run time and computational footprint of link prediction algorithms though maintaining sufficient prediction performance. In statistic, Bootstrap sampling technique evaluates the sampling distribution of an estimator by sampling with replacement from the original sample. In the context of Big Data, Bootstrap sampling has been addressed in many works [7] [21]. In our data evaluation scheme will used the Bag of Little Bootstrap (BLB) [A. Kleiner and Jordan], which combines the results of bootstrapping multiple small subsets of a Big data dataset. THe BLB algorithm use an original Big dataset used generate small samples without replacements. For each generated sample another set of samples are created by resampling with replacement.

3.3.2 Data Profiling

Data profiling is an exploratory approach to data quality analysis. Statistical approaches are used to reveal data usage patterns as well as patterns in the data [15] [14]. Several tools exist for data quality assessment using data profiling and exploratory data analysis. Such tools include Tableau and Talend Open Studio.

Data profiling module performs screening of data quality based on statistics and information summary. Since profiling is meant to discover data characteristics from data sources. It is considered as data assessment process that provides a first summary of the data quality. Such information include: data format description, different attributes, their types and values. data constraints (if any), data range, max and min. More precisely information about the data are presented in two types; technical and functional. This information can be extracted from the data itself without any additional representation using it metadata or any descriptive header file, or by parsing the data using any analysis tools. This task may become very costly in Big Data. To avoid costs generated due the data size we will use the same sampling

process BLB to reduce the data into a representative population sample, in addition to the combination of profiling results.

3.3.3 Data Quality Evaluation

The data profiling provides information about dataset,

- Data attributes (eg. type format)
- Data summary (eg. max, min)
- Big data attributes; size number of sources speed of data generation (eg. data streams)
- What DQD evaluate.

The previous information's are used to select the appropriate quality metrics functions $\boldsymbol{F}$ to evaluate a data quality dimensions d_k for an attribute a_i with a weight w_j

In the Fig 3.2 we describe how data quality evaluated using bootstrap sampling for Big data. The process follows these steps:

1. Sampling from the data set **S n** bootstrap sample size of **ss** size without replacement DS_j.
2. Each sample generated from step 1 is sampled into n' samples of size SS with replacements DS_{ij}
3. For the Each sample DS_{ij} generated in step 2, evaluate the data quality score Q_{ij}

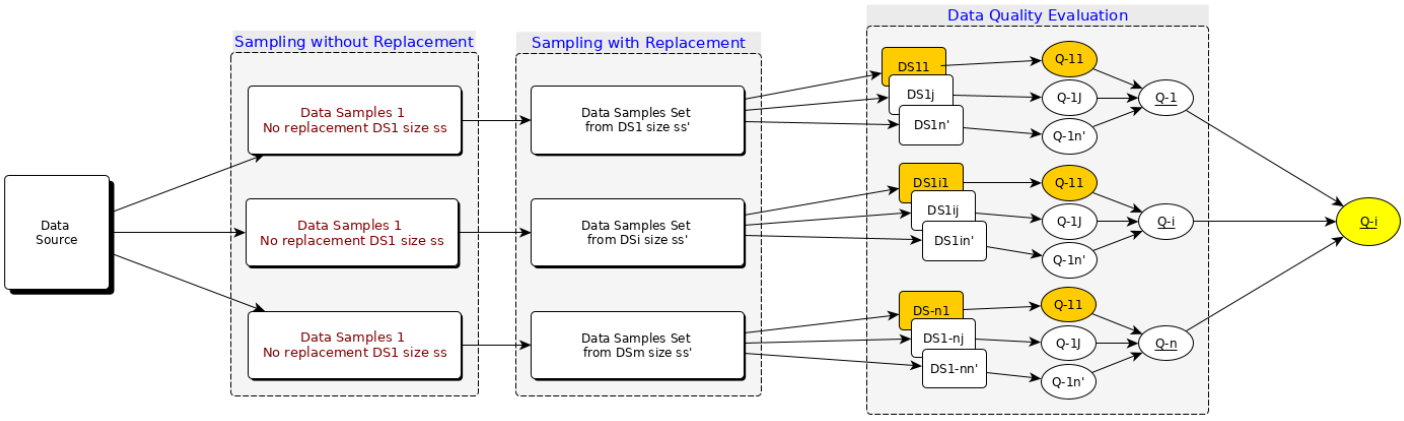

Fig. 3.2 Big Data Quality Sampling Evaluation

Table 3.3 Big Data Quality Evaluation Algorithm

Algorithm: Big Data Quality Evaluation

```
1  let ds a Original Data Set with size SS and Observation (N-SS);
2  let ss (b(SS)) the samples with ss < SS ;
3  let n samples s_i of size ss and M Observation (M-ss);
4  let D a set of DQD D = {d_0, ...d_k, ...d_q};
5  let F a metric function F (completeness, accuracy,...) ;
6  let cc ← 0 counter of correct valid attribute value(when F is true cc = cc + 1);
7  let S ={ DS_0,...DS_i,...DS_n } without replacement ;
8  for i ← 0 to n do
9      Generate sample s_i of size SS from ds;
10     for j ← 0 to n' do
11         Generate a sample _ij of size SS from sample si;
12         for k ← 0 to j do
13             MetricFunctionTuple(d_k, F)
14             for a ← 0 to j do
15                 for a_ij(x) ss values do
16                     if F(a_ij(x), value) == 1 then
17                         measure metric ;
18                         c ← cc + 1 ;
19                         Calculate the scores vector DQD(F, d_k,a_ij, DS_i)= cc/N
                           cc ←0 counter of correct valid attribute value (d_k,F)
20                     end
21                 end
22                 DQD d_k computed for all attributes for a sample ds_ij
23             end
24             DQS_ijk is the D_K scores for an attribute a_ij for sample DS_ij Q_ijk
                 sum of all d_k scores for attributes a_ij for DS_ij
25         end
26         Q_ik+ = 1/n'(Q_ijk)
27     end
28 end
29 Q_k is the mean of all Q_ik for a specif d_k Q_k+ = 1/n(Q_ik)
```

3.3.4 BDQ Evaluation Algorithm

Let ***F*** represents a set of data quality metrics, ***F*=**$\{f_0 ... f_1, ... f_m\}$ where f_1 a quality metric function that will measure and evaluate a DQD d_k for each value of an attribute a_i in the sample s_i and returns 1 if correct, 0 if not. Each f_1 function will compute if the value of the attribute reflects the d_k constraints. For example, the metric accuracy of an attribute is defined as a range of values between 0 and 100, otherwise it is correct. Similarly, it can be defined to satisfy a certain number of constraints related to the type of data such as a zip code, email, social security number, or an address. If we are evaluating the same DQD d_k for a set of attributes, if the weights are all equal, a simple mean is computed. The metric f_i will be evaluated to measure if all the attributes individually have their f_i correct. This is done for each instance (cell or row) of the sample s_i.

In Table 3.3, we describe the detail of BDQ Evaluation Algorithm. THe Q_k represents the mean quality score for a DQD d_k for measurable attributes. For data set let note. A as a set of attributes or features. THe Q_k values respectively for each attribute are represented by a set of quality scores:

$$V = \{Q_{ka_1}, ... Q_{ka_m}\}$$

where **A** is set of m attributes. With this evaluation, we have more insights, statistics and benefits about the Big data quality to ensure a well-refined analytics that targets the best precision.

3.3.5 After Evaluation Analysis

The data evaluation process done on Big data set provides data quality information and scores of quality dimensions of each attributes or features. These scores are used to identify the data that must be targeted and omitted. A set of proposals actions is generated based on many parameters, like DQD, or data quality issue. If a data attribute got a lower score than the required level (%) of accuracy or completeness the following actions are proposed:

- Discard it from the dataset.
- Tune, reformat, and normalize its values.
- Replace values, as in missing data.

Whatever the Quality evaluation results, it always contains actions to be taken on the dataset to remove any irregularities using techniques like cleaning, filtering and pre-processing based on the quality assessment.

Chapter 4

Experimentations , Results and Analysis

References

[A. Kleiner and Jordan] A. Kleiner, A. Talwalkar, P. S. and Jordan, M. The big data bootstrap.

[2] C. Batini, M. S. (2006). *Data Quality Concepts, Methodologies and Techniques*. Springer-Verlag, Berlin Heidelberg.

[Cai and Zhu] Cai, L. and Zhu, Y. The challenges of data quality and data quality assessment in the big data era. 14:2.

[4] Cormode, G. and Duffield, N. (2014). Sampling for big data: A tutorial. page 1975–1975, New York, NY, USA. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[5] D. Firmani, M. Mecella, M. S. and Batini, C. (2015). *On the Meaningfulness of 'Big Data Quality*. Springer Berlin Heidelberg.

[6] Deming, W. E. (1960). *Sample Design in Business Research*. ohn Wiley and Sons, Inc., New York.

[7] F. Liang, J. K. and Song, Q. (2016). *A Bootstrap Metropolis- Hastings Algorithm for Bayesian Analysis of Big Data*. Technometrics.

[8] Falorsi and Scannapieco (2006). *Principi Guida per la Qualità dei Dati Toponomastici nella Pubblica Amministrazione*, 12.

[9] Guptil, C. and Morrison, J. (1995). *Elements of Spatial Data Quality.* Elsevier Science Ltd, Oxford, UK.

[10] I. Caballero, M. S. and Piattini, M. (2014). A data quality in use model for big data.

[11] Jugulum, R. (2014). *Computing with High Quality Data*, page 1. John Wiley and Sons, Inc., New Jersey.

[12] Krawczyk, H. and Wiszniewski, B. (2003). Visual gqm approach to quality-driven development of electronic documents. Edinburgh UK. 2nd International Workshop on Web Document Analysis.

[13] M. Ali-ud-din Khan, M. F. U. and Gupta, N. (2014). Seven v's of big data understanding big data to extract value. page 1–5. American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the, 2014.

[14] Maydanchik, A. (2007). *Data quality assessment*. Bradley Beach, Technics Publications, New Jersey.

[15] Osborne, J. W. (2013). *Best practices in data cleaning: a complete guide to everything you need to do before and after collecting your data*. Thousand Oaks.

[16] Ostman, A. (1997). The specifications and evaluation of spatial data quality. In *18th ICA/ACI International Conference*, Stockholm, Sweden.

[17] Pernici and Scannapieco, M. (2003). Data quality in web information systems. *Journal of Data Semantics*.

[18] Platform, B. A. T. (2018). Extracting business value from the 4 v's of big data. https://www.ibmbigdatahub.com/infographic/four-vs-big-data. Accessed: 2019-07-19.

[19] Redman, T. C. (2018). If your data is bad, your machine learning tools are useless. https://hbr.org/2018/04/if-your-data-is-bad-your-machine-learning-tools-are-useless. Accessed: 2019-07-19.

[20] S., J. (2015). Overview of data quality challenges in the context of big data. pages 1–9. International Conference on Computing Communication and Security.

[21] Satyanarayana, A. (2014). Intelligent sampling for big data using bootstrap sampling and chebyshev inequality. pages 1–6. 27th Canadian Conference on Electrical and Computer Engineering (CCECE).

[22] V. Gadepally, T. Herr, L. J. L. M. M. M. and Miller, B. A. (2015). Sampling operations on big data. page 1515–1519. 49th Asilomar Conference on Signals, Systems and Computers.

[23] V. Goasdoué, S. Nugier, D. D. and Laboisse, B. (2007). An evaluation framework for data quality tools. page 280–294.

[24] Wang and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, page 12.

[25] Wiszniewski, B. and Krawczyk, H. (2003). Digital document life cycle development. In *1st International Symposium on Information and Communica- tion Technologies*, Dublin Ireland.