Using visual-interactive properties to support data quality visual assessment on abstract and timeless data

João Marcelo Borovina Josko[1], João Eduardo Ferreira[2]

[1] Federal University of ABC, Brazil
marcelo.josko@ufabc.edu.br
[2] São Paulo University, Brazil
jef@ime.usp.br

Abstract. Visualization systems belong to supervised tools that can make noticeable the intrinsic structures of defects on data. However, despite the significant number of these systems that assist Data Quality Assessment, few provide resources to examine these structures deeply. This situation prevents data quality appraisers from using their contextual knowledge to confirm or refute any data defect. This article explores a visualisation system's additional features and design characteristics (named *Vis4DD*) that uses visual-interactive properties to support data quality visual assessment on abstract and timeless data (e.g., Customer, Billing). Additionally, we conduct a full review and outline the state-of-art visualization systems related to data quality assessment and fit Vis4DD into this scenario.

Categories and Subject Descriptors: H.5.m [**Information Interfaces and Presentation**]: User Interface; I.3.3 [**Computer Graphics**]: Picture/Image Generation; H.2.m [**Database Management**]: Miscellaneous

Keywords: Data Quality Assessment, Information Visualization, Structured Data Defects, Visual Assessment

1. INTRODUCTION

Low data quality is an old issue that still threatens the reliability of analytical process outcomes in the Big Data era. Improving data quality requires alternatives that combine procedures, methods, and techniques. However, determining *which* the more valuable resources are and *where to* apply them implies *knowing* the current data quality state of databases. Backing this planning is the aim of the Data Quality Assessment process (DQAp).

DQAp provides valuable inputs to improve and keep data quality at levels required by analytical initiatives. Relevant computational models support such a process, especially for data defects whose detection rules are more precise (e.g., Domain Constraint Violation [Borovina Josko et al. 2016]). However, these models use quantitative [Chandola et al. 2009] or constraint-based [Maydanchik 2007] approaches that restrict the human role in interpreting their outcomes [Dasu 2013].

On the other hand, DQAp strongly depends on data context knowledge since it is impossible to confirm or refute a defect based only on data [Dasu 2013]. The context specifies the structure of meaning and relationship between data and an environment (e.g., organization departments) [Borovina Josko and Ferreira 2017b]. Hence, human supervision is essential throughout this process. Visualization systems belong to supervised approaches that combine computational capability with pattern-finding and semantic distinctions innate to human beings to permit data quality visual assessment.

The literature presents some visualization systems focused on DQAp support. However, their visual-interactive properties and other capabilities are disparate in consequence of the applied design

principles. This article uses the term "visual-interactive properties" to visualization technique elements (e.g., basis properties, dimensions, description types) and interactive techniques. A visualization technique determines how to use the visual space to expose the data of interest. An interactive technique allows modifying the visual space representation to enable data interpretation according to the task goal [Borovina Josko and Ferreira 2017b].

$Vis4DD$ is a web-based visualization system designed with proper visual-interactive properties to permit data quality visual assessment of a set of data defects that demands high human supervision (e.g., Missing Reference [Borovina Josko et al. 2016]). It can assist any enterprise in conducting the quality assessment of their structure data in relational or graph databases. This article extends our preceding $Vis4DD$ discussion [Borovina Josko and Ferreira 2017a] by detailing supplementary aspects regarding its features and design step. Furthermore, we outline the state-of-art literature chosen by a rigorous review process and fit this web-based system in this scenario.

This work has the following organization: In Section 2, we detail the requirements and principles considered to design the Vis4DD system, while in Section 3, we explain its features. Next, we discuss the Vis4DD operation and results regarding its use in a case study in Section 4. In Section 5, we sketch the state-of-art literature characteristics and fit Vis4DD into this outline. Lastly, we conclude this work in Section 6.

2. VIS4DD PROBLEM DOMAIN, REQUIREMENTS, AND DESIGN

Data quality visual assessment denotes a nonlinear analytical process of comprehending the current data quality state mediated by visualization systems. Through interactive visual representations, data quality appraisers pursue and correlate meanings (patterns and relationships) associated with a target defect structure until they integrate semantic evidence to confirm or refute it. Hence, the absence of correspondence between a visual representation and this task goal prevents data quality appraisers from accomplishing their work [Borovina Josko and Ferreira 2017b].

Visualization system design is manifold since different techniques composition may lead to an intended result, but with varying effectiveness levels [Ware 2004]. Therefore, $Vis4DD$ designs considered high-level tasks patterns to ensure proper support for the problem mentioned above. These tasks denote cognitive strategies of visual inquiry in assessing data quality according to the target defect structure.

The requirement analysis followed a two-step methodology that relied on a 6-year experienced data quality analyst. The first step associated patterns and interactive classes with each data defect, according to their structure. For instance, the Inclusion Dependency Violation defect ($d15$ on Table III) causes $R1$ tuples unrelated to $R2$ tuples. We associated it with an isolation pattern and the interactive classes named simplify, space arrangement, select. Based on these outcomes, the second step modeled and formalized high-level assessment tasks (HATS). The case study goals added other requirements, including color scales, homogeneous visual representation appearance, and data volume. For a complete requirement analysis discussion, please refer to [Borovina Josko and Ferreira 2017b].

Guided by the requirements analysis outcomes, the design stage followed three steps. The first decomposed the system into components (Section 3.3), while the second step selected the most appropriate interactive technique for each interactive class. For example, we the interactive techniques of filter, attribute arrangement, and highlight for the interactive classes (respectively) of the data defect $d15$ above. The last step chose visualization techniques of different basis properties and description types based on [Mackinlay 1986; Bertin 2010] and the HATS needs. Basis properties determine the visual variable used to encode a target attribute for assessment purposes, including position, hue, saturation, size, connection. In turn, description types denote how the values of a target attribute are encoded by the basis property (e.g., point, line, proportionality, directed link).

3. VIS4DD SYSTEM CHARACTERISTICS

3.1 Technological Infrastructure

Our system uses the R environment due to its portability, efficiency on extensive dataset manipulation, and analytics-driven abilities tightly related to several graphical libraries or frameworks [Chambers 2008]. These features are essential requirements for proper data quality analysis support and provide an extensible property to the system, i.e., the ability to add visual representations and computational resources easily.

Vis4DD interface uses Shiny framework that provides an easy way to build interactive web solutions based on the reactive programming model [Beeley and Sukhdeve 2018]. This model permits to control of how (*reactive conductors*) the interface parameters (*reactive sources*) change elements of visual representations (*reactive endpoint*). Moreover, such a framework makes it easy to connect a Web Server product to enable Internet running mode by partitioning the graphical interface and the remaining components. Nonetheless, this procedure requires adding security features once the current version of our system runs in the local model.

The native R memory management loads datasets to RAM up to 4GB, according to hardware configuration. Our system uses special packages (named *ff* and *ffbase* [Gahlawat 2014]) to overcome such a memory constraint. In a nutshell, these packages enable extract data from different sources (e.g., CSV file, ODBC sources), loading them in chunks (in an HDD or SSD) in a particular format that enables vectorial processing. This resource allowed *Vis4DD* to support the data quality visual assessment (Section 4.2) of relations up to 10^7 tuples.

3.2 The Graphical Interface

Figure 1 shows the four visual spaces of the *Vis4DD* interface. Visual space 1 offers all visualization techniques grouped by the graphical data representation approach [Keim 2002], including parallel coordinates, radial graphs, and heat maps. Visual space 2 organizes all interaction techniques and corresponding parameters, while space 4 provides access to relation facility management discussed in the next section. Lastly, the interface's broader area (visual space 3) represents data through one visualization technique without screen scrolling.

This system provides two sets of interaction techniques: standard and specific. The first set denotes interactions available to all visualization techniques, including attribute arrangement, filtering, and storing current visual representation as a PNG file. In contrast, the specific set represents interactions associated with visualization techniques properties. For instance, jittering, trellis, data aggregation (hexagonal binning or smoothing), and opacity change are interactions available to visualizations that encode data as points (e.g., scatterplot family) or lines. In contrast, visualizations such as treemap and table plots use geometric zooming. In addition, all visual representations permit the data quality appraiser to indicate defective data regions with an "$\mathcal{X}$" sign, as illustrated by Figure 4.

Moreover, according to the visualization technique and data characteristics, *Vis4DD* chooses among two color scales based on Hue, Saturation, Lightness model. The segmented scale has twenty-two hues of maximum contrast to categorize values on dense data spaces [Green-Armytage 2010]. However, specific visualizations with high contrasting hues (e.g., Figure 4b) dispensed this scale.

In turn, the unsegmented scale has two approaches to ensure quantitative data isomorphism. The first approach uses two hues of different families (colorblind safe), increasing saturation and invariant lightness that permits analysis on visual representations with low data density. The second approach adopts a unique hue with decreasing saturation and increasing lightness to improve dense visual representations analysis [Bergman et al. 1995].

3.3 The Background Components

Figure 2 illustrates all communications between *Vis4DD* components based on the pipe-and-filter architecture style. The *Relation Facility* (1) component permits managing (e.g. loading, discarding) any relation of interest in an R workspace (R session image containing functions and datasets), as observed in Figure 3. Relations must be first extracted from source databases as a formatted file to avoid interference in their operations and provide a static data state for quality assessment. *Vis4DD* offers different separators and quotes settings (visual space 1) to load a formatted file and present its characteristics (visual space 2). This load operation keeps all original data values untouchable. However, it executes certain structural checks (e.g., each line complies with the file's header) and adjusts (e.g., convert numerical attribute into character when one value is not numerical). In future works, we intend to provide relation extraction functionalities (based on ODBC API) from different relational or graph databases.

The *Filter Engine* (2) permits select data of interest through multiple search criteria: a set of keywords (categorical attributes) or a range of values (quantitative attributes). Alternatively, a data quality appraiser can point data (individually or collectively) directly to the visual data representation (Section 3.2, visual space 3). The *Register Engine* (3) automatically logs all session interactions and their corresponding parameters, including the data marked with low data quality. It also takes a snapshot of the current visualization representation, if required by a data quality appraiser.

Finally, the *Graphic Engine* (4) produces visual representations based on visualization technique, data characteristics, and interactions parameters received from the graphical interface (e.g., spatial arranging and visual appearance). This component can handle data generated by the Filter Engine up to the limits discussed in Section 3.1.

4. DATA QUALITY VISUAL ASSESSMENT THROUGH VIS4DD

4.1 Walkthrough

Vis4DD system starts working by loading the last saved R workspace and setting global parameters. In the case of an empty workspace (none relation loaded), all visualization techniques remain unavail-

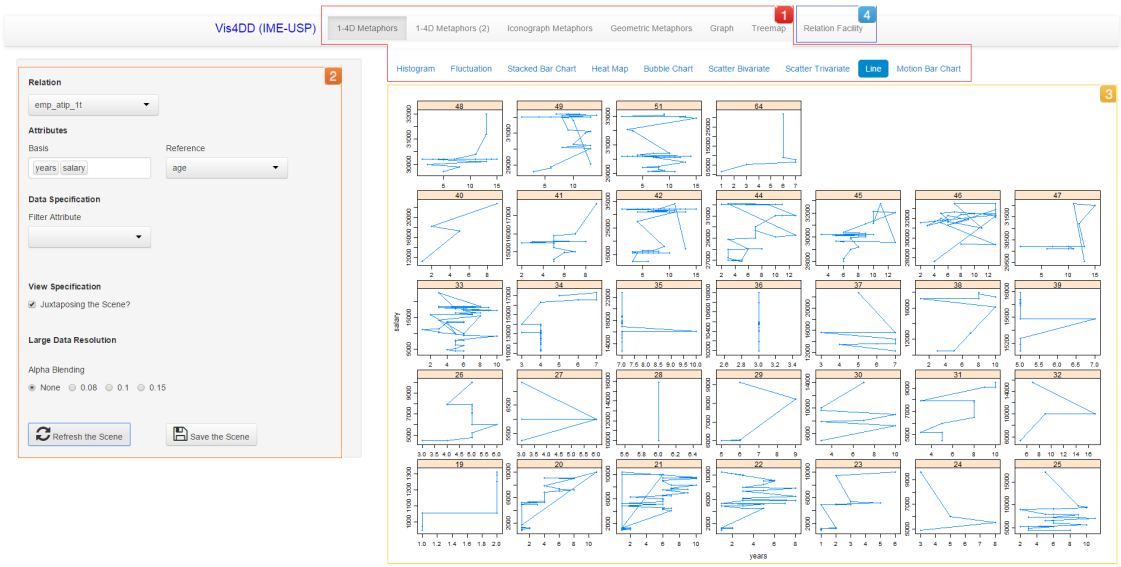


Fig. 1: The four visual spaces of *Vis4DD* interface (Source: The authors)

able. During system use, a data quality appraiser may load or discard any relation (Section 3.3).

Data quality appraisers may obtain an overall sense of all data and their patterns in the early stage of data quality assessment. Then, they select an appropriate visualization technique to expose all data of a relation of interest. They provide the corresponding target and reference attributes and may also change the default parameters of any interaction. For instance, a data quality appraiser may expose categories in different panels through a trellis (e.g., Figure 1). At the end of this setting procedure, data quality appraisers request the corresponding visual representation generation.

Interactions help data quality appraisers arrange data for comparison and correlation until they can isolate data regions potentially defective. In this stage, filtering and geometric zooming permit a comfortable and continuous refinement of data presentation to attend to the quality assessment task at hand. In suspicious solid cases, data quality appraisers can mark the defective data (e.g., "$\mathcal{X}$" in Figure 4b) and save the current visual representation. Otherwise, they can return to the overall data view (by resetting interactions parameters) and recommences quality analysis transitions until they confirm or refute the presence of a data defect. At any time, a different visualization technique may be selected, reusing the parameters already chosen.

4.2 Case Study Summary

An exploratory case study used *Vis4DD* to identify a set of relationships that exposes visual-interactive properties that permit visual assessment of different data defects. Such a case study used a Sales data model as a basis to distribute these data defects. This work outlines one of these data defects: the atypical tuple as its variants in Employee relation. For a depth discussion of this data defect and others, please refer to [Borovina Josko et al. 2016].

In a nutshell, an atypical tuple deviates from the behavior of the remaining tuples of relation for different reasons [Borovina Josko et al. 2016]. Our case study considered four atypical variants. The $1^{\underline{st}}$ and $2^{\underline{nd}}$ variants denote 0.1% and 1% of defective values in an isolated attribute, respectively. Most visual representations permitted their assessment, but position-based visualizations were outstanding. They made it easy for the data quality analyst to perceive both variants' structures, as the atypical manager and salary association indicated with "$\mathcal{X}$" (orange ellipse highlighted space) in Figure 4a.

Position-based visualizations were also the best option to assess the $3^{\underline{rd}}$ atypical value variant, although they required more interaction actions (e.g., filtering and point displacement). Such variant denotes atypical values interposed among data categories with certain superimpositions.

The last variant ($4^{\underline{th}}$) denotes the unusual combination of values considering multiples attributes. Due to its characteristics, only multidimensional visualizations permitted partial detection of atypical cases through the intensive filter and zooming interactions. Figure 4b illustrates a $4^{\underline{th}}$ variant case

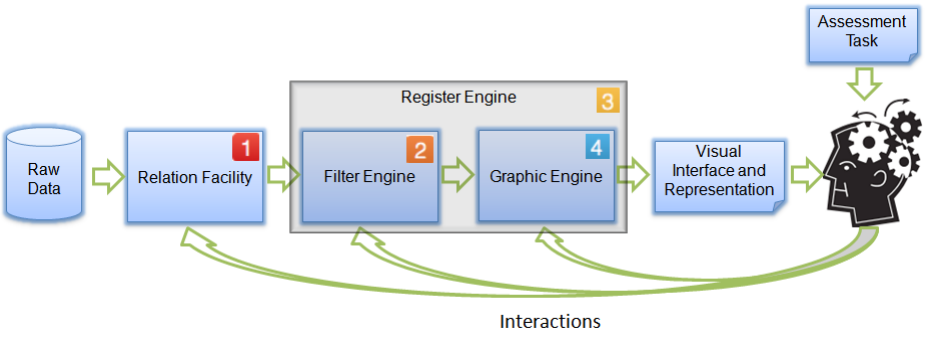


Fig. 2: Components and interface communications (Source: [Borovina Josko and Ferreira 2017a])

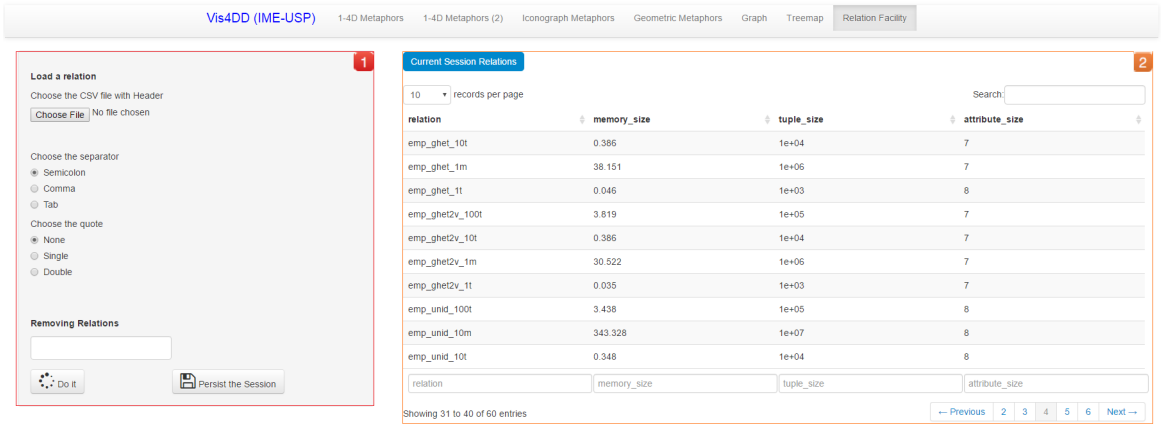
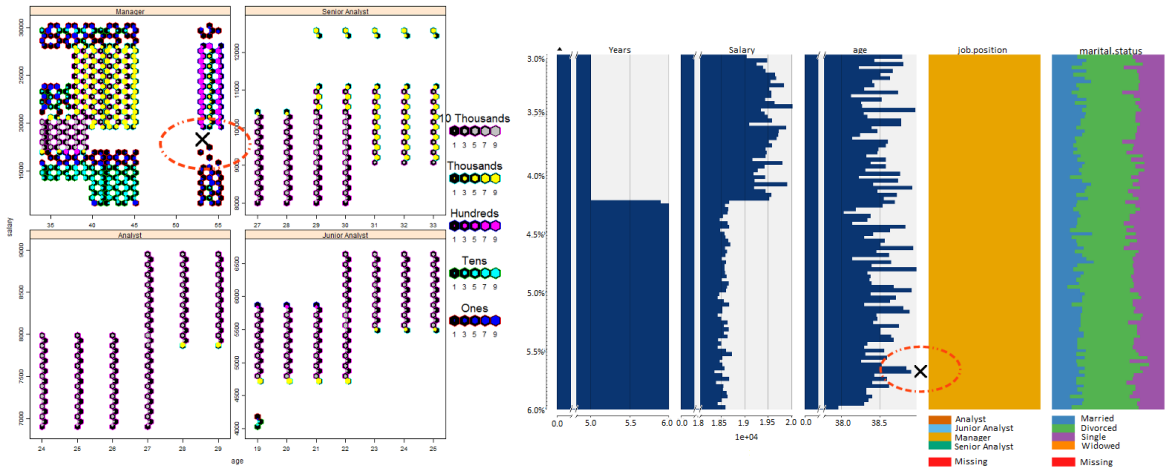


Fig. 3: The Relation Facility (Source: The authors)



(a) Atypical tuples (2^{nd} variant) detection through compacted frequency in hue in resolution of 10^7 tuples

(b) Atypical tuples (4^{th} variant) detection through size proportional to average supported by image zooming in resolution of 10^6 tuples

Fig. 4: Scenes fragments concerning Atypical Tuple assessment (Source: [Borovina Josko and Ferreira 2017a])

involving "years", "salary" and "age" attributes indicated with "$\mathcal{X}$" (orange ellipse highlighted space) by the data quality analyst. For a depth discussion of our case study outcomes, please refer to [Borovina Josko and Ferreira 2017b].

5. RELATED WORKS

Different research areas use visualization resources for data quality assessment purposes. Therefore, we conduct a broad literature review to select significant related works to comprehensively view their features and design principles regarding data quality assessment. In Section 5.1, we describe the selection and analysis protocols, while in Section 5.2 we discuss the analysis outcomes.

5.1 Selection and Analysis Protocol

Our protocol followed four sequential steps: *i*) defining the general terms, *ii*) searching for potential works, *iii*) screening the potential works, and *iv*) analysing the selected works. In the first step, we considered different resources (e.g., [Reuters 2020], [Research and of Australasia 2020]) to identify

Table I: List of Data Quality Terms (Source: The Authors)

Data Quality Terms				
Anomaly Detection	*Data Anomaly*	*Data Assessment*	*Data Consistency*	*Data Evaluation*
Data Profiling	*Data Quality*	*Dirty Data*	*Missing Data*	*Missing Values*

the leading scientific publications (journals and conferences) and corresponding terminology strongly linked to this article aim. Having analysed these publications terminology, we defined the general and selective data quality terms shown in Table I.

In the next step, we connected the data quality terms with terms that refer to using visualization systems, which are *"visual"*, *"visualization"*, *"visualizing"*, and *"graphics"*. According to the search engine capabilities, we applied the expression below for each data quality term (Table I) to search for scientific works on the title, keywords, abstract, or entire text. On each search outcome, we analyze the abstract and keywords quickly to elect potentially related works.

{Data Quality term} AND ("visual" OR "visualization" OR "visualizing" OR "graphical")

In the third step, we disregarded duplicated works or works that refer to the same visualization system. Moreover, we examined excerpts from the remaining works' main sections to determine their compliance with the acceptance criteria below. Table II exhibits the literature review summary, while Tables III and IV exhibits all selected works.

(1) *Uses visual-interactive resources as enablers of data quality assessment*

(2) *Characterizes visualizations systems and does not describe frameworks, studies of any nature (e.g., survey, techniques evaluation), or visual-interactive techniques or approaches that data quality assessment is not the main subject*

(3) *Refers to abstract (non-spatial), structured, and timeless data*

(4) *Addresses data defects related to Completeness, Accuracy, and Consistency data quality dimensions.* The former refers to the representation degree of the relevant aspects from the objects of a Universe of Discourse (UoD). The second dimension refers to the adherence degree of data regarding the respective real-world value or reference value. The latter dimension refers to the compliance degree of data regarding all rules of a given UoD.

(5) *Publishing date starting from 1996*

Appraising a visualization system is a tricky step, as it must consider different aspects regarding the style and the perspective of evaluation [Thomas and Cook 2005]. The former refers to the degree

Table II: Literature review summary per Search Engine (Source:The Authors)

Search Engine	Search Scope	Potential Works	Selected Works
ACM Digital Library	Title, Abstract	23	1
DBLP	Title	6	1
IEEE Xplore Digital Library	Title, Abstract	28	6
Google Scholar	Title	36	6
Sage	Title, Abstract	4	-
Science Direct	Title, Abstract, Keywords	1	-
SringerLink	Full Text	2	1
Taylor and Francis Online	Full Text	2	1
Total		102	16

of formalization and duration of an evaluation (e.g., longitudinal studies). In turn, the latter sets the extent of the characteristics observed in a visualization system (e.g., isolated techniques or visual properties). Both define the results' scope, effort and cost.

In the analysis step, we examine the selected works according to the following questions: *Which principles and other inputs do guide the tool design step towards data quality assessment?*, *Which features for handling high-moderate data volume are available?*, *What data defects do they provide assessment support?*. We formulated these questions considering the above perspective and three aspects strongly related to data quality visual assessment (Section 2): data volume, interactions, and visual properties. It is worth mentioning that the present work does not intend to exhaust the subject.

5.2 Selected Works Analysis

In this section, we answer previous section questions supported by Tables III and IV. These are crosstabs that sign with "•" all data defects and the features related to each visualization system, respectively. The term "feature" refers to the *design principle* (metric communication-driven or visual diagnosis-driven), *design inputs* (informational source used to guide the visualization systems design), *basis properties* and *description types* (Section 2), *interactive technique* (Section 1), and computational methods for *handling voluminous data*.

According to [Borovina Josko et al. 2016], data defects have core structures with slight peculiarities (named variants) that occur on distinct granularities (e.g., attribute values, tuple). We used this taxonomy as the foundation to identify and discuss the selected works' coverage, as observed in Table III. Moreover, each data defect has a code (di, $i \geq 1$) to make its reference easy.

Regarding coverage, Table III reveals that 44% (12/27) of data defects have some visualization system support. However, 58% (7/12) of this total support has just one system representing. Moreover, this table exposes that most tools refer to low granularity data defects, i.e., defects in an attribute or tuple of a single relation. This situation contrasts with the significance of data relationships in any database, even more in the Big Data scenario. Only $Vis4DD$ and $VDQAM$ are exceptions to this bias by supporting some inter-relation data defects (e.g., $d15$, $d17$, $d21$ on Table III).

Referring to density, Table III shows that visualization systems concentrate on the domain constraint violation ($\approx 87\%$) and a little less on atypical tuple and duplicate tuple ($\approx 43\%$ and $\approx 18\%$, respectively). Although visual-interactive resources are relevant for the last two data defects that demand human confirmation or refutation, the former is quickly assess by assertions-based methods. Such over-attention is partially explained by a significant number of statistics tools concerned with *identifying the cause* of missing values. This outline of the selected works indicates that many data defects requiring human analysis effort are still unexplored (e.g., $d22 - d25$) or under-explored (e.g., $d10 - d11$).

According to Table IV, most visualization systems ($\approx 56\%$) follow an exclusive quality-aware design principle that particularizes visualizations to highlight a specific data defect detected through some computational resource. Interestingly, all of them described no information regarding their design process, suggesting their visual-interactive properties' subjective choice. Such an approach can be sufficient to warn of data quality issues, especially for those data defects that require low human supervision (e.g., $d7$ in Table III). However, it may prevent a data quality appraiser from extracting the meanings needed to confirm or refute a data defect because of the misalignment risk between the task need and the system's visual-interactive properties [Ware 2004; Borovina Josko and Ferreira 2017b].

Several quality-aware systems show such misalignment somehow. For example, particular systems use visual variables (e.g., label) or visual arrangement of objects (e.g., DaVis) that obligate data quality appraisers to process text or consolidate information spread in multiples scrolling points, prompting them an immense cognitive load. Others (e.g., $DQVis$) communicate the presence of a

Table III: Data defect per visualization system (Source: The Authors)

Data Defect / Variant	Manet [Unwin et al. 1996]	GGobi [Cook et al. 2007]	DaVis [Sulo et al. 2005]	XMDVTOOL-Q [Xie et al. 2006]	D-DUPE [Kang et al. 2008]	VIQTOR [Führing and Naumann 2007]	Tableplot [Malik et al. 2010]	VIM [Templ and Filzmoser 2008]	Profiler [Kandel et al. 2012]	VDQAM [Teng et al. 2012]	DQVis [Wang et al. 2013]	MissingDataGui [Cheng et al. 2015]	Untitled [Sjöbergh and Tanaka 2017]	MonAT [Noselli et al. 2017]	Vis4DD [Borovina Josko and Ferreira 2017a]	TAQIH [Sánchez et al. 2019]
$d1$. Atypical Tuple																
Isolated attribute						•	•		•		•			•	•	•
Composition of attributes							•								•	•
$d2$. Cardinality Ratio Violation																
$d3$. Cond. Funct. Dependency Viol.																
$d4$. Cond. Inclusion Dependency Viol.																
$d5$. Disjoint Subdomains																
$d6$. Incorrect Temporal Reference																
$d7$. Domain Constraint Violation																
Range Constraint			•													
Enumeration Constraint																
Regular Expression Constraint									•							
Mandatory Constraint	•	•	•	•		•	•	•	•	•	•	•	•	•		•
$d8$. Duplicate Tuples																
Single Relation			•		•						•					
Multiple relations																
$d9$. False Tuple															•	
$d10$. Functional Dependency Violation											•					
$d11$. Heterogeneous Granularity															•	
$d12$. Heterogeneous Measurement Unit									•						•	
$d13$. Homonymous Values																
$d14$. Imprecise Value																
$d15$. Inclusion Dependency Violation										•						
$d16$. Incompatible Replication																
$d17$. Incorrect Reference															•	
$d18$. Incorrect Value									•					•		•
$d19$. Inference Rule Violation																
$d20$. Key Dependency Violation									•							
$d21$. Missing Reference															•	
$d22$. Missing Tuple																
$d23$. Overloaded Tuple																
$d24$. Participation Constraint Violation																
$d25$. Semantic Integrity Violation																
$d26$. Synonymous Values																
$d27$. Transition Constraint Violation																

complex data defect ($d8$ in Table III) but provide no visual-interactive features to enable data quality appraiser judgement.

The remaining visualization systems (visual diagnosis-driven) showed concern about data defect structure or the data quality appraiser's need. Still, they also lack describing the implications of this concern in their design choices. For this reason, these systems are susceptible to the misalignment risk previously mentioned. VIM and $Vis4DD$ are the only exceptions. Their design considered implica-

Table IV: Features per visualization system (Source: The Authors)

Features	Manet	GGobi	DaVis	XMDVTOOL-Q	D-DUPE	VIQTOR	Tableplot	VIM	Profiler	VDQAM	DQVis	MissingDataGui	Untitled	MonAT	Vis4DD	TAQIH
Design Principle																
Quality-aware	•	•	•	•	•	•		•	•	•	•					•
Visual Diagnosis-driven							•	•				•	•	•	•	•
.....																
Design Inputs																
Data Defect Structure		•						•				•	•	•	•	
Data Quality Task															•	
Data Quality Metrics																•
Not Defined or scant mention	•		•	•	•	•	•		•	•	•					
Specialist Participation														•	•	
.....																
Basis Prop. – Description Type																
Connection - Radial node-link					•					•	•				•	
Glyphs - Hue per Object				•												
Label - Text per value					•	•								•		
Position - Density in Saturation									•						•	
Position - Points per Object	•	•	•	•				•			•	•	•	•	•	•
Position - Line per Object		•	•	•				•		•		•	•		•	•
Position - Hue per Frequency								•				•			•	•
Position - Hue per value													•			•
Position - Size per Average							•								•	
Saturation - Hue per Object								•								
Saturation - Hue per Average															•	
Recursive Hrchy - Prop. to value															•	
Size - Proportional to value			•										•			
Size - Proportional in Frequency	•							•	•	•	•			•		
.....																
Interactive Technique																
Attribute Arrangement							•	•	•			•	•	•	•	•
Brushing and Linking	•	•							•				•			
Coordinated Concurrent Views	•	•							•				•			
Data Quality Annotation				•		•									•	
Details on Demand	•		•		•				•							
Filter			•	•		•	•		•					•	•	•
Geometric Zoom		•					•						•		•	
Multiresolution Hierarchy				•												
Opacity Change															•	
Ordering					•	•	•					•			•	
Point Displacement		•						•							•	
Rotation										•						
Redimension										•						
Trellis		•		•								•		•	•	
.....																
Handling Voluminous Data																
Data Prefetching				•												
Efficient Virtual Memory use															•	
In-memory Data									•							•
Sampling		•		•												

tions from data defects structures or tasks definitions specified according to data defects structures and an experienced data quality appraiser's perspective, respectively.

Unexpectedly, several systems overlook that DQAp aims to map the current data quality, i.e.,

it associates quality metadata to the corresponding defective data. This metadata is crucial for improving quality assessment rules, crowdsourcing-based assessment, or repairing defective data (Data Cleansing). However, only three ($\approx$ 18%) of the visualization system consider (partially) such rich information produced by computational methods or data quality appraiser (*Data Quality Annotation* in Table IV). $XMDVTOOL-Q$ persists quality metrics with data, $VIQTOR$ captures user scores for each attribute value, while $Vis4DD$ marks defective data and stores the corresponding visualization.

Table IV (*Handling Voluminous Data*) exhibits that only five systems adopt resources to provide scalability with different constraints. For instance, data prefetching depends on the user access pattern to take advantage of data locality. In turn, sampling methods can separate interrelated instances, reducing their contribution to particular data defect (e.g., $d8$,$d17$,$d21$,$d24$ on Table III). Besides, specific interactive techniques (e.g., opacity change, point displacement, trellis, rotation) have a tenuous effect in highly concentrated data in small visual regions. Data volume's low relevance among most selected works is also perceptible on the extensive use of visual primitives that do not scale, including label, glyphs, connections, and position (except density in saturation encoding).

Interestingly, none of the works considered parallel processing or GPU (Graphics Processing Unit) resources to improve data or rendering operations performance, respectively. This situation is surprising as all selected works adopted a centralized architecture, i.e., they require moving all data to a particular site to perform the quality assessment procedures. However, with the sharp growth in Cloud distributed data, this architecture's moving cost tends to be prohibitive in large volumes of data. For better data quality assessment support, future visualization systems must introduce: i) a new architecture design (e.g., microservices [Yang et al. 2019]) that segments the interactive visualization from compute-intensive models and ii) GPU facilities (e.g., WebGPU [Usher and Pascucci 2020]) to enable visualizations on large data.

6. CONCLUSIONS

This article provides a supplementary discussion regarding the design approach and features of the $Vis4DD$ visualization system, whose purpose is to visually diagnose meanings (patterns and relationships) associated with data defect structures. Moreover, it also provides a snapshot of the state-of-art visualization systems concerning data quality assessment. Such a picture reveals several opportunities for extending visual analytics support in the data quality arena.

Nevertheless, $Vis4DD$ neither addresses multiple coordinated views nor offers computational approaches (e.g. data mining methods) for data defects without visual evidence. As future works, we intended to add features to capture data appraisers' assessment annotations regarding defective data regions and combine computational and visual-interactive resources to address new data defects, as indicated in this article review analysis.

REFERENCES

Beeley, C. and Sukhdeve, S. R. *Web Application Development with R Using Shiny: Build stunning graphics and interactive data visualizations to deliver cutting-edge analytics.* Packt Publishing Ltd, Birmingham, UK, 2018.

Bergman, L. D., Rogowitz, B. E., and Treinish, L. A. A rule-based tool for assisting colormap selection. In *Proceedings of the 6th conference on Visualization'95.* IEEE Computer Society, Washington DC, USA, pp. 118, 1995.

Bertin, J. *Semiology of graphics: diagrams networks maps.* Esri Press, California, US, 2010.

Borovina Josko, J. M. and Ferreira, J. E. Vis4dd: A visualization system that supports data quality visual assessment. In *Proceedings of the satellite events of 32nd Brazilian Symposium on databases.* SBC, Uberlandia, Brazil, pp. 46–51, 2017a.

Borovina Josko, J. M. and Ferreira, J. E. Visualization properties for data quality visual assessment: An exploratory case study. *Information Visualization* 16 (2): 93–112, 2017b.

Borovina Josko, J. M., Oikawa, M. K., and Ferreira, J. E. A formal taxonomy to improve data defect description. In *Database Systems for Advanced Applications: DASFAA 2016 International Workshops: BDMS, BDQM, MoI,*

and SeCoP, Dallas, TX, USA, April 16-19, 2016, Proceedings, H. Gao, J. Kim, and Y. Sakurai (Eds.). Springer International Publishing, Cham, pp. 307–320, 2016.

Chambers, J. M. *Software for data analysis: programming with R.* Springer, New York, NY, USA, 2008.

Chandola, V., Banerjee, A., and Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv.* 41 (3): 15:1–15:58, July, 2009.

Cheng, X., Cook, D., and Hofmann, H. Visually exploring missing values in multivariable data using a graphical user interface. *Journal of statistical software* 68 (1): 1–23, 2015.

Cook, D., Swayne, D. F., and Buja, A. Missing values. In *Interactive and dynamic graphics for data analysis: with R and GGobi*. Springer Science & Business Media, New York, NY, USA, pp. 47–62, 2007.

Dasu, T. Data glitches: Monsters in your data. In *Handbook of Data Quality*. Springer, Berlin, Germany, pp. 163–178, 2013.

Führing, P. and Naumann, F. Emergent data quality annotation and visualization. In *ICIQ*. MIT, Cambridge, MA, USA, pp. 424–430, 2007.

Gahlawat, A. Big data analysis using r and hadoop. *IJCEM International Journal of Computational Engineering & Management* 17 (5): 9–14, 2014.

Green-Armytage, P. A colour alphabet and the limits of colour coding. *JAIC-Journal of the International Colour Association* vol. 5, pp. 1–23, 2010.

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., and Heer, J. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. ACM, Capri Island, Italy, pp. 547–554, 2012.

Kang, H., Getoor, L., Shneiderman, B., Bilgic, M., and Licamele, L. Interactive entity resolution in relational data: A visual analytic tool and its evaluation. *IEEE Transactions on Visualization and Computer Graphics* 14 (5): 999–1014, 2008.

Keim, D. A. Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics* 8 (1): 1–8, Jan., 2002.

Mackinlay, J. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics (TOG)* 5 (2): 110–141, 1986.

Malik, W. A., Unwin, A., and Gribov, A. An interactive graphical system for visualizing data quality–tableplot graphics. In *Classification as a Tool for Research*. Springer, Berlin, Germany, pp. 331–339, 2010.

Maydanchik, A. *Data quality assessment*. Technics publications, Bradley Beach, NJ, USA, 2007.

Noselli, M., Mason, D., Mohammed, M., and Ruddle, R. Monat: a visualweb-based tool to profile health data quality. In *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)*. Vol. 5. SCITEPRESS, Porto, Portugal, pp. 26–34, 2017.

Research, C. and of Australasia, E. A. The era conference ranking exercise. http://portal.core.edu.au/conf-ranks/, 2020.

Reuters, T. Journal citation reports. https://clarivate.com/webofsciencegroup/solutions/journal-citation-reports/, 2020.

Sánchez, R. Á., Iraola, A. B., Unanue, G. E., and Carlin, P. Taqih, a tool for tabular data quality assessment and improvement in the context of health data. *Computer methods and programs in biomedicine* vol. 181, pp. 104824, 2019.

Sjöbergh, J. and Tanaka, Y. Visualizing missing values. In *2017 21st International Conference Information Visualisation (IV)*. IEEE, London, United Kingdom, pp. 242–249, 2017.

Sulo, R., Eick, S., and Grossman, R. Davis: a tool for visualizing data quality. *Posters Compendium of InfoVis* vol. 2005, pp. 45–46, 2005.

Templ, M. and Filzmoser, P. Visualization of missing values using the r-package vim. Tech. rep., Department of Statistics and Probability Therory, Vienna University of Technology, 2008.

Teng, D., Yang, H., Ma, C., and Wang, H. Vdqam: A toolkit for database quality evaluation based on visual morphology. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, Seattle, WA, USA, pp. 245–246, 2012.

Thomas, J. J. and Cook, K. A. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society Press, New York, NY, USA, 2005.

Unwin, A., Hawkins, G., Hofmann, H., and Siegl, B. Interactive graphics for data sets with missing values — manet. *Journal of Computational and Graphical Statistics* 5 (2): 113–122, 1996.

Usher, W. and Pascucci, V. Interactive visualization of terascale data in the browser: Fact or fiction? In *2020 IEEE 10th Symposium on Large Data Analysis and Visualization (LDAV)*. IEEE, Salt Lake City, Utah, USA, pp. 27–36, 2020.

Wang, K., Ma, D. T. H. Y. C., and Wang, H. Dqvis: A toolkit for visual quality analysis for relational database. In *Proceddings of 17th IEEE International Conference on Information Visualisation - Poster Session*. IEEE, Porto, Portugal, 2013.

Ware, C. *Information Visualization: Perception for Design.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

Xie, Z., Huang, S., Ward, M. O., and Rundensteiner, E. A. Exploratory visualization of multivariate data with variable quality. In *2006 IEEE Symposium on Visual Analytics Science And Technology.* IEEE, Baltimore, MD,USA, pp. 183–190, 2006.

Yang, W., Tao, Y., and Lin, H. Voxer—a platform for creating, customizing, and sharing scientific visualizations. *Journal of Visualization* 22 (6): 1161–1176, 2019.