paper:16

Data quality assessment of very large database through visualization system

João Marcelo Borovina Josko[1], **João Eduardo Ferreira**[1] **(Advisor)**

[1]Department of Computer Science
Institute of Mathematics and Statistics (IME)
University of São Paulo (USP)
São Paulo - SP - Brazil

{jmbj,jef}@ime.usp.br

Level: PhD
Enrollment in the program: February, 2011
Qualification Evaluation: September, 2013
Expected Completition: June, 2015
Steps completed: Taxonomy submitted as a paper and preliminary framework structure
Future steps: Full framework and respective case study

Abstract. *Data Quality Assessment outcomes are essential to improve data quality and are required condition to support analytical processes. There are several successful approaches to automate this support to syntactic data defects. In contrast, the dependence of semantic data defects on data context knowledge implies on human supervision. The visualization systems belong to a class of supervised tools that can turn data defects into visual items. However, there is no design support for this purpose. Hence, this paper presents a framework to assist the design of these systems fitting the visual quality assessment of semantic data defects. Such an approach is based on data defects structure, data characteristics and user-centered tasks.*

Quality in Big Data, Visual Data Quality Assessment, Semantic Data Defects, Visualization Design, Systematic Design, Information Visualization

1. Introduction

New technologies enable industry and scientific organizations to collect, store and distribute large databases to address their analytical processes. More than data processing capacity, such a knowledge-intensive processes depend on reliable data to produce useful outcomes. Improving and keeping data quality at desired level requires to reach out an alternative based on numerous methods, techniques, procedures, processes and technological approaches. However, determining which the more effective resources are and how to apply them implies knowing the current data quality state of databases; this is the aim of the Data Quality Assessment process.

Relevant computational models support this process, specially for syntactic data defects which have precise rules, like *Functional Dependency Violation* [Borovina Josko et al. 2014]. These models are based on quantitative

[Alpar and Winkelsträter 2014] or constraint [Liu et al. 2012] functionalities and share a non-interactive approach through data quality evaluation. In others words, they restrict the human role to the interpretation of their outcomes [Dasu 2013].

However, the Data Quality Assessment process strongly depends on data context knowledge since it is impossible to confirm or refute a defect based only on data [Lee et al. 2009], [Dasu 2013]. The context specifies the structure of meaning between data and an environment (e.g., a department of an organization). Hence, human supervision through this process is essential, even more to semantic data defects due to their difficult rule specification (e.g. False Tuple) [Borovina Josko et al. 2014].

Visualization systems belong to a class of supervised approach that combines computational capacities to pattern-finding and semantic distinctions innate to the human beings. There has been a huge literature in regard to design of these systems, such as the relevant works by [Bertin 1983] and [Ware 2004]. However, such literature does not provide adequate support to the design of visualization systems for the tasks of data quality assessment.

Having set the problem, this work presents an approach to assist the design of visualization system for visual data quality assessment of large databases. Such an approach is based on a systematic framework that encodes visualization system properties based on data defects structure and data characteristics. Our hypothesis is: each data defect may be connected to certain visualization system properties which enable the tasks of data quality assessment.

The work reported here is organized as follows: Section 2 introduces some basic foundations related to this work, while Section 3 reviews all related works. Section 4 describes this work's main contributions. Section 5 draws the first results, while Section 6 presents the current work and future directions.

2. Foundations

2.1. Data Quality

Inadequate data quality directly affects the outcomes and costs of different working processes, especially the knowledge-intensive ones [Redman and Blanton 1997]. Moreover, such an inadequacy also leads to enormous effort towards the required needs of relevant projects (e.g., Data Warehouse, Re-engineering) and prevents organization to provide reliable data to customers and partners.

However, the consequences aforementioned correspond to impurities that arise at any point of the data life cycle [Redman and Blanton 1997]. The life cycle is a model that exposes the various activities (data acquisition, maintenance, use, disposal) on data as well as the elements (software, hardware, working processes, people) in charge of these activities. Gathered, they determine multiple ways to affect data quality.

2.2. Data Quality Assessment

The Data Quality Assessment process involves inspecting and understanding prioritized data regions, according to the Data Quality Dimensions (e.g., completeness and accuracy). It is a collaborative process that provides practical inputs for choosing the most suitable alternative to solve inadequate data [Lee et al. 2009].

Besides the data context (Section 1), performing a data quality assessment requires to consider additional key issues (organizational, cultural and technical) to ensure its final cost lower than the benefits [Lee et al. 2009]. Particularly, data defects are relevant technical issues because they occur in different granularities (e.g. attribute, tuple, relationship) and share complex relationships chains that requires assessing all grains to ensure their absence. For all these reasons, automation of the Data Quality Assessment process is essential to its efficiency and efficacy [Rahm and Do 2000], [Dasu 2013].

2.3. Visualization and Visual Analytics

The visualization research area investigates the use of computational resources to synthesize visual interactive metaphors to enable interpretations of special facts within large amounts of data [Ware 2004]. Thus, its purpose is to make the corresponding behavior of data perceptible to the human mind. Historically, visualization has two interrelated disciplines: scientific visualization deals with physical references inherited from data; information visualization is concerned with visual mapping of abstract and non spatial data. This work is naturally related to the latter.

The increasing data volume prevents data analysis to be addressed only by visual or computational models. Visual Analytics represents an interdisciplinary effort (involving human-computer interaction, cognitive science, among others) to combine human, computational and visual communication capacities to enable reasoning in large data volumes [Thomas and Cook 2005]. Gathering the best of all worlds, Visual Analytics is an important support to detect inadequate quality of data.

3. Related Works

Knowledge concerning the design of visualization systems is encoded in different forms that offer distinct levels of depth and primary foci. Due to the huge literature and space restrictions, only certain papers are introduced.

Taxonomies organizes core concepts about visualization system, such as [Shneiderman 1996]. *Guidelines* describes recommendations to design visualizations systems in given conditions. While certain guidelines provide directions for particular issues [Baldonado et al. 2000], others assume a broader picture of visualization systems [Tang et al. 2004]. In contrast, *implementations* offers design examples through the description of visualization systems for data quality assessment. Such implementations reveals certain visual or systemic appeal to address syntactic data defects [Rundensteiner et al. 2007], [Kang et al. 2008], [Kandel et al. 2012].

Reference Models is the most robust support to visualization system design, which is driven by different theoretical perspectives and purpose. Certain models are based on theories as psychophysics [Csinger 1992], visual perception [Ware 2004] or cognitive psychology [Patterson et al. 2014]. In terms of purpose, certain models are concerned with automatic visualization generation [Casner 1991], [Zhu et al. 2009], rules for graphic construction [Bertin 1983], [Wilkinson et al. 2006] or activities involved with visualization design [Card et al. 1999].

Close to this study, the Spatial Data Quality (SDQ) and the Uncertainty Visualization (UV) explore the relationship between the data quality and the decision-making processes [Devillers et al. 2006]. Both research areas systematically describe models and

classifications combining data characteristics, space and time to determine the visual symbols [Potter et al. 2012] or visual variables [Thomson et al. 2005] to expose uncertainties.

Knowledge Encodings	*Main Characteristics*
Taxonomy	- Describes visualization properties. Little support on designing
Guideline	- Does not address the Data Quality Assessment process - Does not provide a systematic approach for designers - Proposals are unrelated, unstructured and susceptible to contradiction
Implementation	- Does not describe how visual data quality assessment was considered - Does not describe how data defects structures was considered - Does not provide a systematic approach for designers - Visualization is a Quality-aware media - Addresses some syntactic defects, mostly
Reference Model (All the others)	- Does not address the Data Quality Assessment process - Driven to a limited repertoire of visual attributes and techniques (Automation)
Reference Model (SDQ and UV)	- Shares a mathematical and statistical basis for data quality assessment - Visualization is a Quality-aware media (expose data uncertainties)
This study's Approach	- Driven to the Data Quality Assessment process - Visualization is media for extracting and correlating relevant information, until gathering a set of evidences to confirm or refute a data defect - Provide a systematic approach for designers - Strongly based on semantic data defects structure

Table 1. Brief Comparison between Approaches (Source: The authors)

Although such literature offers a seminal knowledge, it can not assist the design of visualization systems for assessment of the semantic data defects. Among the reasons (see Table 1), such literature does not address a key element of the Data Quality Assessment process: the *structures of the data defects*. These structures disclose the data behavior and relationships toward which reasoning and actions are directed. In other words, these are the information to be extracted from the visual stimuli for subsequent cognitive processing. Hence, the design must use this knowledge to align visualization systems properties to the demands of the tasks of visual data quality assessment [Patterson et al. 2014].

There has been much literature describing data defects. Certain relevant literature has used the hierarchical model [Rahm and Do 2000] or mathematical formalization [Oliveira et al. 2005] to explain defects in a broad sense. More recently, such defects have also been analyzed using data warehouse [de Almeida et al. 2013] and temporal [Gschwandtner et al. 2012] perspectives. However, an analysis of such literature shows remarkable discrepancy in terminology, nomenclature, description, defects coverage and granularity of defects nomination.

4. Contributions

Due to the related works limits, the contributions of this work are as follow: a more representative defects taxonomy of atemporal and structure data; a framework that systematically encodes data quality assessment characteristics to core visualization systems properties. They represent the first systematic support to the design of visualization systems for visual data quality assessment.

4.1. A Data Defects Taxonomy based on a Formal Framework

The limitations of the data defect literature (see Section 3) make difficult to answer questions about data quality assessment process: *What is the problem structure behind each defect? What is the representative set of defects related to the quality dimensions of accuracy, completeness and consistency? Which defect subset requires human supervision?*

The taxonomy proposed to answer these questions resulted from applying three steps in sequence. The *review step* used a theoretical approach to the identification of data defects through re-examining topics, such as conceptual data modeling, transformation decisions between conceptual and logical models, and relational theory based on [Maier 1983], [Elsmari and Navathe 2010]. This review also determined the terminology and mathematical formalism.

		Granularity				
Nature	*Data Defect Name*	Value	Attribute	Tuple	Relation	Inter-relation
Semantic	Atypical Tuple	-	-	●	-	-
	Contradictory Attribute	-	-	-	-	●
	Contradictory Circular Reference	-	-	-	●	●
	Duplicate Tuples	-	-	-	●	●
	False Tuple	-	-	●	-	-
	Heterogeneous Granularity	-	●	-	-	-
	Heterogeneous Measurement Unit	-	●	-	-	-
	Homonymous Values	-	●	-	-	-
	Incorrect Reference	-	-	-	●	●
	Incorrect Value	●	-	-	-	-
	Missing Reference	-	-	-	●	●
	Missing Tuple	-	-	●	-	-
	Overloaded Tuple	-	-	●	-	-
	Synonymous Values	-	●	-	-	-

Table 2. Classified Data Defects (partial list) [Borovina Josko et al. 2014]

The subsequent step *classified each data defect according to the nature of the efforts involved in its detection*. Semantic nature indicates that a defect diagnosis requires data domain knowledge due to the impossibility of rule delimitation, which eliminates the use of fully computational solutions. Syntactic nature suggests that a defect has precise rules, allowing purely computational models to contribute.

The final step *classified each data defect based on its place or granularity of occurrence*, including attribute value, single attribute, single tuple, single relation or inter-relation (binary relationship), which involves one or more database instances. The resulting classes from the last two procedures are the basis to classify the data defects listed in Table 2.

4.2. A Framework based on visualization system properties for Data Quality Assessment

Only the integration of certain visualization properties can provide appropriate support to the tasks of visual data quality assessment [Ware 2004], [Casner 1991], [Bertin 1983]. However, the related works can not guide the visualization designer to such goal (see Section 3). The systematic framework proposed to address this situation resulted from applying three steps in sequence. In all of them, the interactive visualization prototypes used are based on R environment and its graphical libraries [Chambers 2008].

The first step represents the *framework preliminary formulation*, illustrated at Figure 1. An user-centered task set establishes the visualization system requirements for data assessment purpose. These requirements are mapped to visualization properties (design domain), such as the analytical visual perception [Rodrigues et al. 2007]. The characteristics of these tasks are determined by the problem domain elements. The manipulated

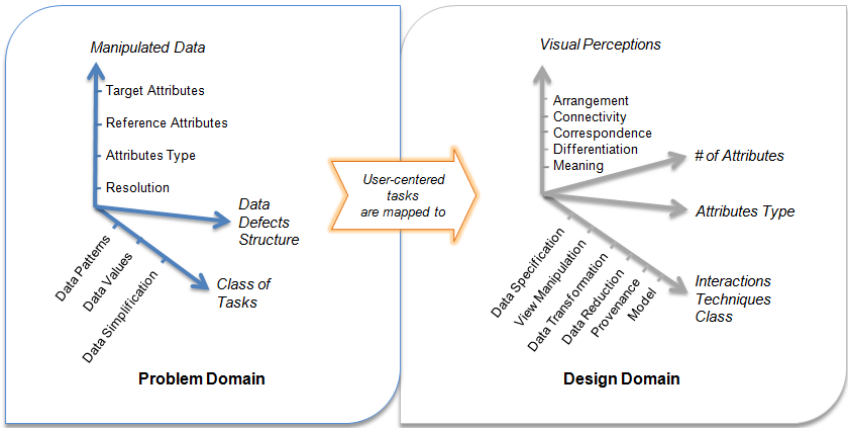

Figure 1. Framework Structure (Source: The authors)

data specify the attributes and data resolution. Data defects are essential to determine kinds of data patterns, relationships among values or granularity to look for through a task, while task class is related to a three-level analytical process: overall, characterization and assessment.

The subsequent step intends to *enhance the framework's structure* through a exploratory case study. Such method is used to comprehend which and how the properties of visualization system may favor (or not) the detection of each semantic defect, according to its structure. The analysis units of this case comprise of visualization techniques with different characteristics (planar or visual) and certain interaction techniques, such as sorting, attributes arrangement, filtering and zooming. The synthetic databases to be used will range from 1000 to 50000 tuples, which about 1% are defective.

The third step is a thoughtful exploration of the *framework in practice*. A case study will show the viability of the framework to assist the design of a interactive visualization prototype capable to support the tasks regarding to the assessment of certain semantic data defect. Such a prototype has visual resolution reduction capacities through filters because it is intended to use a synthetic database with near 50 millions tuples, which about 5% are defective.

Although the framework provides the core guiding for visualization systems design for visual data quality assessment, it does not represent exhaustive rules. To leverage the interpretation of data defects, the framework must treat the cognitive theories and the social collaborative interactions. These issues represent one of the future work directions.

5. First Results

The taxonomy (see Section 4.1) has been submitted to JDIQ (*Journal of Data Quality and Information*). It provides a greater coverage of data defects mathematically formalized and satisfies different data quality assessment needs, as guiding the basic design decisions of supervised systems. The example below is an fragment of such a taxonomy.

> *Definition (Incorrect Value):* Let Σ be the work tape alphabet of an OTM, such that $\Sigma = R(A)$. Let Γ be the oracle tape alphabet of an OTM, such that $\Gamma = \{0, 1\}$. Let A be the OTM oracle that contains all of the attributes of relation R that faithfully represent facts about objects of the interest domain. An attribute has an incorrect value *iff* $\exists a \in R(A) : O^A(a) = 0$.

An incorrect value is an unfaithful or contradictory representation of facts about objects of the interest domain.

Example 1: A customer's birth date is 03/10/1970, but it was represented as 01/01/1980.
Example 2: A customer's name is "Ridley Scott", but it was represented as "Joan Ridley".

In turn, the framework's preliminary result shows the systematic encoding of a core set of suitable visualization system properties, as illustrates the visual assessment of *atypical value defect* in Figure 2. At upper left, the *Task Set* section represents the tasks sequence in regard to such a defect. These tasks are mapped (one by one) to visualization system properties, as seen in the *Mappings* section. The *Final List* section gathers all these properties which are use to design a R interactive prototype using synthetic data (shown at right).

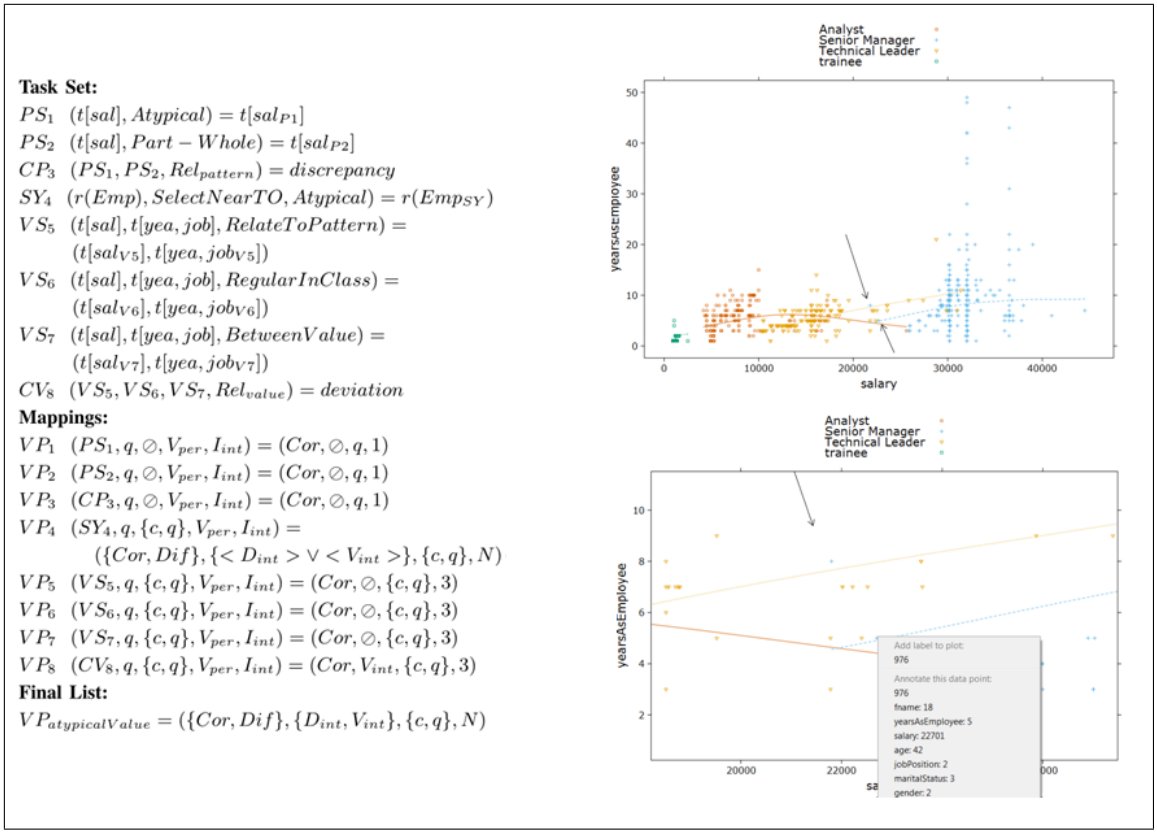

Figure 2. Framework in practice - preliminary study (Source: The authors)

It is worth to mention that this work has explored the probability sampling as a data volume reduction solution. The general idea is use the inference statistic to increase the probability to sample suspicious data in regard to defect structure. Due to the complexity and the impact on this work's scope, such a subject was left as future work.

6. Current Work and Future Directions

The current effort executes the exploratory case study to collect qualitative information about the relationship between each semantic data defect and visualization system properties. Subsequently, such an information will be basis for enhancing the framework preliminary structure. Lastly, a case study will explore the framework in practice. Additionally, a second paper will be produced to gather all these findings.

References

Alpar, P. and Winkelsträter, S. (2014). Assessment of data quality in accounting data with association rules. *Expert Systems with Applications*, 41(5):2259–2268.

Baldonado, M. Q. W., Woodruff, A., and Kuchinsky, A. (2000). Guidelines for using multiple views in information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '00, pages 110–119, New York, NY, USA. ACM.

Bertin, J. (1983). *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin press.

Borovina Josko, J. M., Oikawa, M. K., and Ferreira, J. E. (2014). A taxonomy of data defects based on a formal framework. (Submitted to ACM Journal of Data and Information Quality).

Card, S. K., Mackinlay, J. D., and Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Casner, S. M. (1991). Task-analytic approach to the automated design of graphic presentations. *ACM Transactions on Graphics (TOG)*, 10(2):111–151.

Chambers, J. M. (2008). *Software for data analysis: programming with R*. Springer.

Csinger, A. (1992). The psychology of visualization. Technical Report TR-92-28.

Dasu, T. (2013). Data glitches: Monsters in your data. In *Handbook of Data Quality*, pages 163–178. Springer.

de Almeida, W. G., de Sousa, R. T., de Deus, F. E., Daniel Amvame Nze, G., and de Mendonca, F. L. L. (2013). Taxonomy of data quality problems in multidimensional data warehouse models. In *Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on*, pages 1–7. IEEE.

Devillers, R., Jeansoulin, R., et al. (2006). *Fundamentals of spatial data quality*. ISTE London.

Elsmari, R. and Navathe, S. (2010). *Fundamentals of database systems*. Addison-Wesley, 6th. edition.

Gschwandtner, T., Gärtner, J., Aigner, W., and Miksch, S. (2012). A taxonomy of dirty time-oriented data. In *Multidisciplinary Research and Practice for Information Systems*, pages 58–72. Springer.

Kandel, S., Parikh, R., Paepcke, A., Hellerstein, J. M., and Heer, J. (2012). Profiler: integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, pages 547–554, New York, NY, USA. ACM.

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

Lee, Y. W., Pipino, L. L., Funk, J. D., and Wang, R. Y. (2009). *Journey to Data Quality*. The MIT Press.

Liu, J., Li, J., Liu, C., and Chen, Y. (2012). Discover dependencies from data-a review. *IEEE Transactions on Knowledge and Data Engineering*, 24(2):251–264.

Maier, D. (1983). *The theory of relational databases*, volume 11. Computer science press Rockville.

Oliveira, P., Rodrigues, F., and Henriques, P. (2005). A formal definition of data quality problems. In *International Conference on Information Quality*.

Patterson, R. E., Blaha, L. M., Grinstein, G. G., Liggett, K. K., Kaveney, D. E., Sheldon, K. C., Havig, P. R., and Moore, J. A. (2014). A human cognition framework for information visualization. *Computers & Graphics*, 42:42–58.

Potter, K., Rosen, P., and Johnson, C. (2012). From quantification to visualization: A taxonomy of uncertainty visualization approaches. In Dienstfrey, A. and Boisvert, R., editors, *Uncertainty Quantification in Scientific Computing*, volume 377 of *IFIP Advances in Information and Communication Technology*, pages 226–249. Springer Berlin Heidelberg.

Rahm, E. and Do, H. (2000). Data cleaning: Problems and current approaches. *IEEE Bulletin of the Technical Committee on Data Engineering*, 24:11.

Redman, T. and Blanton, A. (1997). *Data quality for the information age*. Artech House.

Rodrigues, J. F., Traina, A. J., de Oliveira, M. C. F., and Traina, C. (2007). The spatial-perceptual design space: A new comprehension for data visualization. *Information Visualization*, 6(4):261–279.

Rundensteiner, E., Ward, M., Xie, Z., Cui, Q., Wad, C., Yang, D., and Huang, S. (2007). Xmdvtool q: quality-aware interactive data exploration. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 1109–1112. ACM.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings of the 1996 IEEE Symposium on Visual Languages*, VL '96, pages 336–, Washington, DC, USA. IEEE Computer Society.

Tang, D., Stolte, C., and Bosch, R. (2004). Design choices when architecting visualizations. *Information Visualization*, 3(2):65–79.

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

Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., and Pavel, M. (2005). A typology for visualizing uncertainty. In *Electronic Imaging 2005*, pages 146–157. International Society for Optics and Photonics.

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

Wilkinson, L., Wills, D., Rope, D., Norton, A., and Dubbs, R. (2006). *The grammar of graphics*. Springer.

Zhu, Y., Suo, X., and Owen, G. S. (2009). A visual data exploration framework for complex problem solving based on extended cognitive fit theory. In *Advances in Visual Computing*, pages 869–878. Springer.