





A Meta-modeling Approach to Take into Account Data Domain Characteristics and Relationships in Information Visualizations

Andrea Vázquez-Ingelmo¹ (✉) , Alicia García-Holgado¹ ,
Francisco J. García-Peñalvo¹ , and Roberto Therón^{1,2} 

¹ GRIAL Research Group, Computer Sciences Department, Research Institute for Educational Sciences, University of Salamanca, Salamanca, Spain

{andreaavazquez, aliciagh, fgarcia, theron}@usal.es

² VisUSAL Research Group, University of Salamanca, Salamanca, Spain

Abstract. Visual explanations are powerful means to convey information to large audiences. However, the design of information visualizations is a complex task, because a lot of factors are involved (the audience profile, the data domain, etc.). The complexity of this task can lead to poor designs that could make users reach wrong conclusions from the visualized data. This work illustrates the process of identifying features that could make an information visualization confusing or even misleading with the goal of arranging them into a meta-model. The meta-model provides a powerful resource to automatically generate information visualizations and dashboards that take into account not only the input data, but also the audience's characteristics, the available data domain knowledge and even the data context.

Keywords: Data visualization · Information visualization · Misleading visualizations · Feature identification · Meta-modeling

1 Introduction

Visual explanations are everywhere: they convey complex information, raise attention over target topics, improves the understandability of certain domains, etc. They can take the form of infographics, simple graphs, or even elaborated information visualizations.

Visual explanations are very powerful, because they let users visually perceive information in order to generate knowledge. However, information visualizations might turn out to be a double-edged sword.

The persuasive power of visualizations [1] has its benefits (better data understanding, more attention and focus on the information, etc.) but they also can lead users to wrong conclusions. Wrong conclusions are not always predictable, because they can have its origin on a unproper data visualization design, but also be influenced by the end users' prior beliefs, biases or polarization regarding certain topics.

It is important to take all these factors into account in order to provide properly designed and honest methods of visualization, because the main goal must be focused on how the user perceives and processes the displayed data.

These factors can be taken into account through domain expertise, i.e., information visualization experts that also have knowledge regarding the visualization's data domain and can provide a well-designed product through its expertise.

However, it is very difficult that every professional that makes use of information visualizations to convey information has these levels of expertise or domain knowledge, because it might be time-consuming and visual explanations usually need to be delivered quickly (for example, for covering news stories).

For these reasons, in this paper we propose an approach based on meta-modeling in which the data domain characteristics and relationships among variables are accounted for. The goal of this work is to characterize domain expertise to include it as a part of a generative pipeline to automatically develop information visualizations and dashboards. This approach can assist novices or practitioners without a significant level of the data domain knowledge to select the best parameters for their information visualizations. To sum up, the main contribution of this paper is a new version of a meta-model for instantiating information visualizations taking into account data domain and expertise.

The rest of this paper is organized as follows. Section 2 describes the methodology employed to carry out the meta-model and the automatic generation of dashboards. Section 3 outlines the modification of a previously developed dashboard meta-model to hold information about the data domain and the data context. Section 4 presents a proof-of-concept of the visualization generation using domain knowledge. Finally, Sect. 5 discusses the results and Sect. 6 offers the conclusions derived from this work.

2 Materials and Methods

2.1 Identification of Features

Usually, the definition of an information visualization involves human-computer interaction experts specialized in visualization and also experts from the data domain. Their know-how avoids the development of misleading visualizations. The automatic development of information visualizations has to take into account the experts' know-how.

In particular, it requires the identification of the features that has an impact in the correct visualization of the data depending on the dataset and the goals related to the analysis of that data. The process to identify these features has provided the base to define the meta-modeling proposal described in this work.

The study covers a set of phases based on an experimental approach as a way to discover the features that has an impact in the automatic development of no misleading information visualizations. Figure 1 presents the three phases: testing, analysis and solution.

The first phase, testing, is an experimental phase. Two experiments were set up to see what happens when some features are automatic generated and their impact in achieving the goal of the developed visualization. The main difference between both

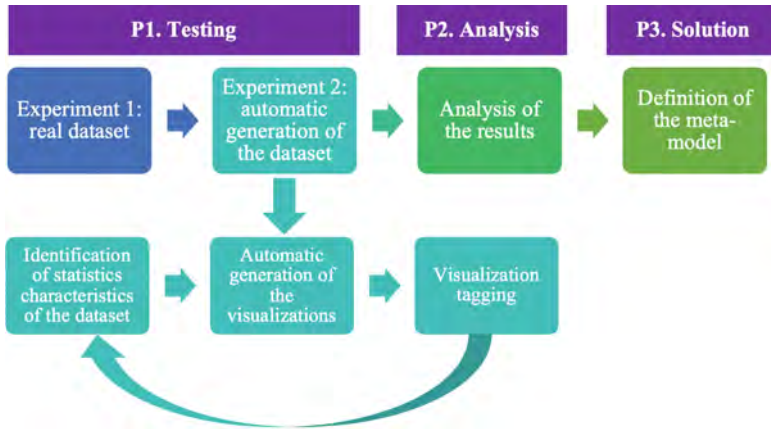


Fig. 1. Method used to identify the features that influence in the automatic generation of not misleading visualizations

experiments is the abstraction level. The first experiment is based on a real dataset from a particular domain. On the other hand, the second experiment has a high abstraction level because the dataset is random generated based on a set of statistic characteristics generated automatically.

Both experiments are focused on the automatic visualization generation. The generator use code templates based on a meta-model to define dashboards [2–4] and a Python script in which the different parameters are tuned to get a set of visualizations. The script processes the dataset changing a set of characteristics and provide a HTML and JavaScript file with the visualizations. The first experiment has the following characteristics:

- Dataset: Tri-variate dataset regarding the COVID-19 incidence rate in the 21 districts of Madrid (Spain)¹ and average income per household for each district².
- Goal: Identifying correlation among certain variables.
- Encoding channels: Two (X and Y position) and three (X and Y position and size).
- Type of visualization: Scatter plot with “circles” as visual mark.
- Scales’ domain range: We changed the minimum and maximum of the scale domain. We changed these values using these measures: the scale variable mean minus/plus two-times the standard deviation of that column in the dataset, the column’s minimum/maximum value, the column’s minimum value multiplied by 0.5 and zero and the column’s maximum value multiplied by 1.5. In the case of nominal variables, the domain holds all the existing nominal values within that column of the dataset.

¹ Official data source from the Madrid’s government open data portal: https://datos.comunidad.madrid/catalogo/dataset/covid19_tia_muni_y_distritos/resource/f22c3f43-c5d0-41a4-96dc-719214d56968.

² Official data source from the Spanish National Institute of Statistics: <https://www.ine.es/jaxiT3/Tabla.htm?t=31097>.

Regarding the second experiment, the main characteristics are:

- **Dataset:** A bivariate dataset randomly generated using a set of statistical characteristics that are changed to generate the set of visualizations. The dataset is generated to cover different types of probability distribution with different statistical dispersion. Specifically, the standard deviation and median are the modified values.
- **Goal:** Comparison between certain variables.
- **Encoding channel:** Color scale and section of the map.
- **Type of visualization:** Map.
- **Scales' domain range:** Same variations as in the first experiment.

The last part of the experiments was the labeling process. All authors worked together to tag each generated visualization as misleading or no misleading. In the first experiment, this process enabled the identification of the features that introduce misleading in a visualization for a particular goal and domain. On the other hand, we applied the same process in the second experiment but following an iterative approach (Fig. 1). The labeling process enabled the redefinition of the statistic characteristics of the dataset in order to generate the visualization. This process enabled the identification of the features that have an objective impact in the automatic visualization generation; the features of the domain itself.

Finally, we analysed the results of the experimental phase and define a solution to consider the features of the domain as an input to automatic development of information visualizations.

2.2 Meta-modeling

The model-driven development (MDD) paradigm [5, 6] enables the abstraction of the characteristics and functionalities involved in the development of information systems. The main strength of this paradigm is that it moves data and operations specifications away from technologically specific details.

Following this approach, a dashboard meta-model was developed in previous studies, obtaining a set of abstract elements and relationships to define specific products [2–4]. A fragment of the dashboard meta-model is shown in Fig. 2.

2.3 Automatic Generation

We have developed an automatic dashboard generator based on the meta-model. The code generator takes as an input a set of parameters that account for the elements and attributes of the meta-model, and the result is the source code of a dashboard according to the provided configuration.

The approach taken to automatically generate the source code is based on the software product line (SPL) paradigm [7, 8] and we developed different HTML and JavaScript code templates [9] to materialize the variability points of the product line [10].

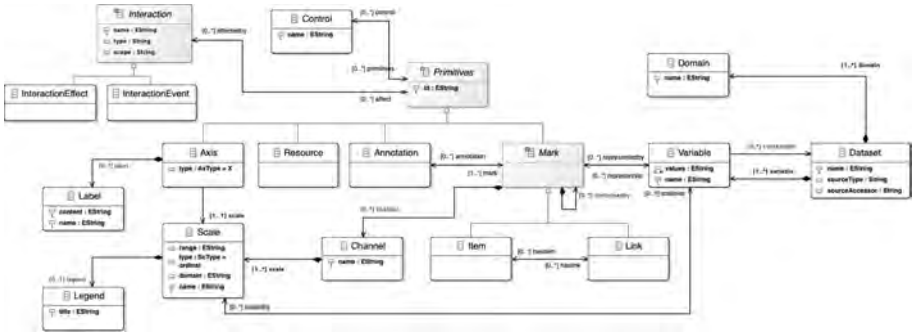


Fig. 2. Fragment of the previous version of the dashboard meta-model.

3 Meta-model Modification

3.1 Domain Characterization

It is necessary to identify relevant features to characterize the domain and to materialize those characteristics in useful visualizations. In this case, we defined the domain as a set of attributes that statistically describe the variables involved in that domain.

Specifically, we have included in the meta-model a new class named *DomainVariable*, which represent data variables that are part of the data domain. This class is associated with a domain (in the meta-model, the class *Domain*) and also with the class *Variable*, which represent a variable that belongs to a dataset to be displayed through the visualization. The *Variable* entity is seen as a sample of the *DomainVariable* entity.

The domain variable enables us to perform analytic tasks on the visualization and reach insights, because we have information about which values are normal, which values are outliers, or which tendency is being developed.

If users see a visualization with information from a domain which they don't fully understand, their conclusions might be wrong. But also, if practitioners don't fully understand the data domain of a visualization they are developing, they could end up with a misleading graphic. Another example about this concern is given. When visualizing information in X, Y coordinates it is necessary to select the domain of the scales in both axes; different scale extents might distort the whole data story being told. Figure 3 shows an example of the same data visualized through different Y-axis scales.

If we are not aware of the data domain, the first graph can be seen as misleading for not starting the Y-axis at the zero value. Starting the Y-axis at the zero value (as in the second graph), gives us the impression that the temperature change along time is very small and, indeed it is (in absolute terms) [11].

However, in this case, being aware that the tendency and average temperatures of the world over the last decades provides the context to understand that a change of 1 °C in average temperature is a huge increment in this domain, conveying a whole different story. So, although the first graph does not comply to Tufte's lie factor [12], it is more honest than the second in terms of representing data framed in this domain.

For these reasons, we included as abstract attributes of the *DomainVariable* entity the following characteristics: mean, standard deviation, median, first quartile, third quartile,

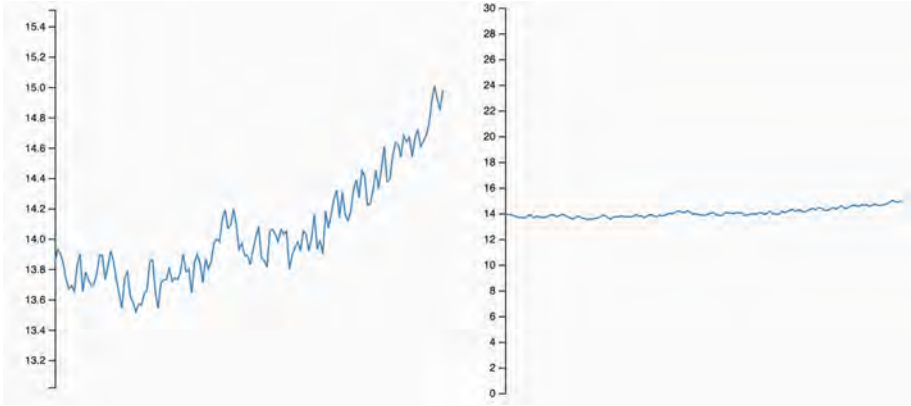


Fig. 3. Example of the effect of different scales when visualizing data.

interquartile range, maximum and minimum. These values not only describe statistically the variable, but also they help in characterizing their distribution [13], as they give notion of its dispersion, skewness and what values can be considered as outliers (Fig. 4).

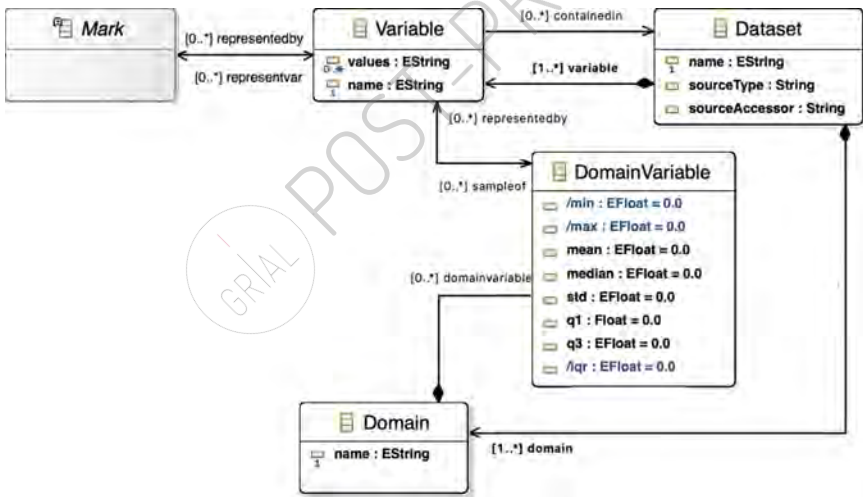


Fig. 4. Detail of the included class to characterize the domain in information visualizations.

3.2 Context Inclusion

We also included another association regarding the *DomainVariable* entity to consider the possibility of representing context in a visualization. In this case, we identify context as additional information related with a variable. For example, income household could

be related with the COVID-19 incidence rate [14], and including that information in a visualization about COVID-19 incidence rate provides context to the data to be displayed.

By including a reflexive association on the *DomainVariable* class, we enable the possibility to identify and materialize relationships among variables from a different or the same domain (for example, because they are correlated).

The inclusion of this relationship to provide the notion of context allows the accountability of potentially relevant variables to include in a visualization before selecting its technical features (Fig. 5).

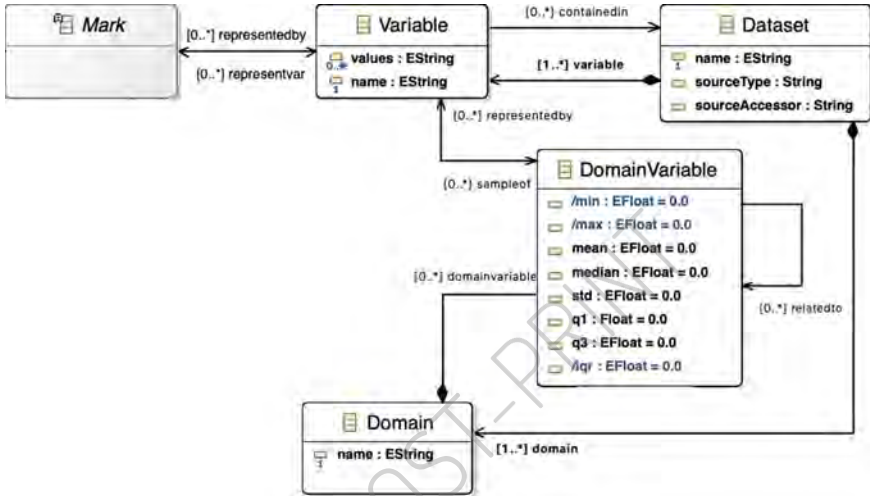


Fig. 5. Detail of the included reflexive association to represent data context in information visualizations.

4 Meta-model Instantiation Example

This section provides an instantiation example of the meta-model to illustrate the role of the included elements. Figure 6 shows the instantiation of a scatter plot for an hypothetical data domain and Fig. 7 the shows a generated visualization according to the instance.

The *DomainVariable* instance provides knowledge to select the Y-axis scale range in a way that data is not exaggerated. In this case, although the Y-axis does not start from zero (which can be seen as misleading in some domains), the domain knowledge provides us the justification: according to the domain characteristics for that variable, it is not likely to find values below 30, so it would make no sense to start the axis at zero. The visualized variable (VariableA in Fig. 6) is seen as a sample of the domain variable.

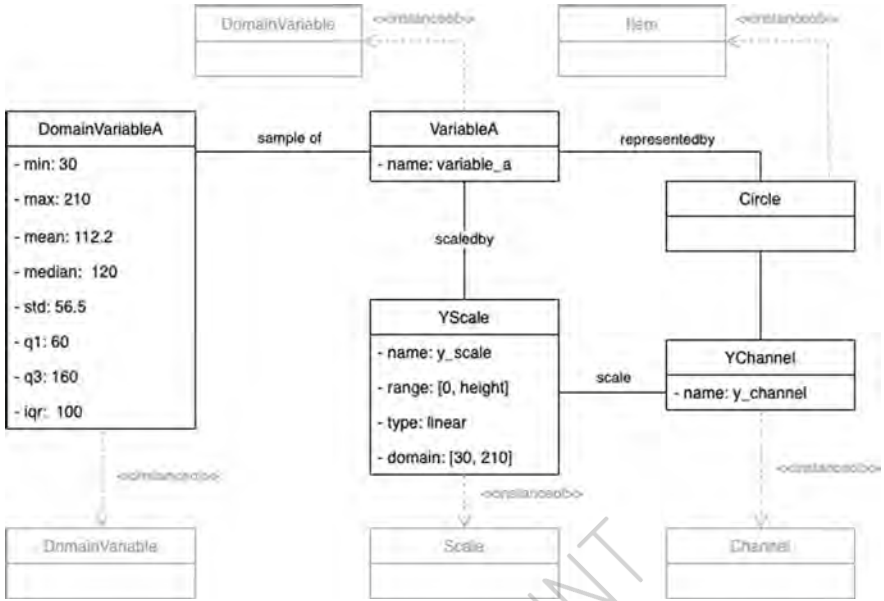


Fig. 6. Excerpt of the example visualization instantiation (Y-axis channel and scale).

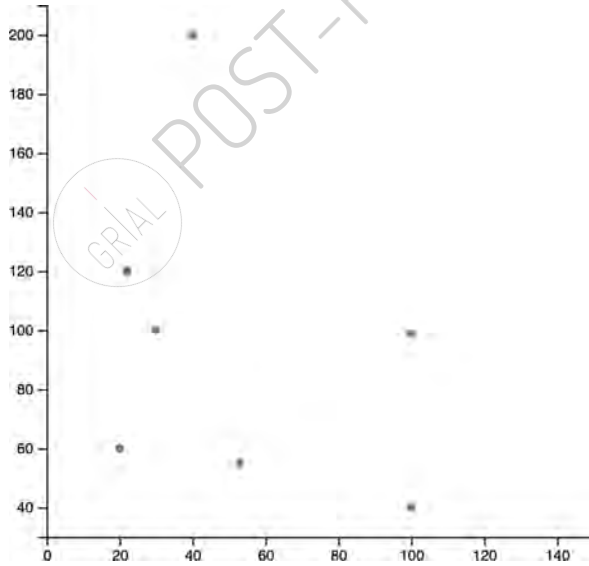


Fig. 7. Generated visualization through the meta-model's instantiation parameters.

5 Discussion

The results of this work set the foundations for characterizing conceptual concepts such as domain expertise and data context when designing information visualizations. Specifically, the testing process has yielded a new version of the meta-model that includes important factors such as the data domain characteristics. The use of a meta-model not only provides a theoretical framework to work with, but also a skeleton to instantiate real products adapted to a specific context.

We selected statistical features to define the domain variables involved in a visualization. More specifically, we included as attributes the characteristics that define box plots [13]. Although the values distribution's is better characterized by its probability density function or cumulative density function, these two functions are more difficult to incorporate to the meta-model than a set of values such as the mean or the maximum and minimum values. However, there is a limitation that will be explored in subsequent works: the inclusion of the notion of uncertainty in data (because the *DomainVariable* represents knowledge about the population), which is also a significant factor when visualizing data [15–17].

It is also important to take into account that there could exist different domain variables with different values' distributions but the same summary statistics [18]. In this case, we don't intend to use the *DomainVariable* values to compare domain variables among them, but to define visual features (such as scales) regarding that variable solely.

Moreover, the goal of the visualization must be accounted for too. In the generation example, the generated visualization would not be useful for detecting outliers, because they would fall outside the scale's domain. It is important to find balance with the visualization goal when including the domain knowledge.

Finally, this approach is limited to domain expertise: if the *DomainVariable* values are populated with wrong values, the whole process would be affected in the same way that a user can reach wrong conclusions if his or her notion of the domain differs from reality.

6 Conclusions

This work presents a meta-modeling approach to incorporate the notion of domain knowledge and data context into information visualizations. The approach is supported by a code generator that materialize the meta-model's abstract features into specific visualizations. The meta-model proposal enables not only a set of rules and guidelines to define no misleading visualizations, but also supports the automatic generation of information visualizations.

The study based on the automatic generation of datasets of visualizations has supported the identification process of the features, especially those related to the data domain, that influence in the development of no misleading visualizations. We introduced the results of the experiments as part of the meta-model to define dashboards.

Future research lines will involve in-depth testing of the influence of domain expertise and data context on the visual elements of a dashboard with the goal of including this knowledge into the generation pipeline.

Acknowledgements. This research work has been supported by the Spanish *Ministry of Education and Vocational Training* under an FPU fellowship (FPU17/03276). This research has been partially funded by the Spanish Government Ministry of Economy and Competitiveness throughout the DEFINES project (Ref. TIN2016-80172-R).

References

1. Pandey, A.V., Manivannan, A., Nov, O., Satterthwaite, M., Bertini, E.: The persuasive power of data visualization. *IEEE Trans. Visual Comput. Graph.* **20**(12), 2211–2220 (2014)
2. Vázquez-Ingelmo, A., García-Peñalvo, F.J., Therón, R., Conde, M.Á.: Representing data visualization goals and tasks through meta-modeling to tailor information dashboards. *Appl. Sci.* **10**(7), 2306 (2020)
3. Vázquez-Ingelmo, A., García-Holgado, A., García-Peñalvo, F.J., Therón, R.: A meta-model integration for supporting knowledge discovery in specific domains: a case study in healthcare. *Sensors* **20**(15), 4072 (2020)
4. Vázquez-Ingelmo, A., García-Peñalvo, F.J., Therón, R.: Capturing high-level requirements of information dashboards' components through meta-modeling. In: *The 7th International Conference on Technological Ecosystems for Enhancing Multiculturality (TEEM 2019)*, León, Spain (2019)
5. Pleuss, A., Wollny, S., Botterweck, G.: Model-driven development and evolution of customized user interfaces. In: *Proceedings of the 5th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, pp. 13–22. ACM (2013)
6. Kleppe, A.G., Warmer, J., Bast, W.: *MDA Explained. The Model Driven Architecture: Practice and Promise*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA (2003)
7. Clements, P., Northrop, L.: *Software product lines*. Addison-Wesley, Boston, MA, USA (2002)
8. Pohl, K., Böckle, G., Van der Linden, F.J.: *Software product line engineering: foundations, principles and techniques*. Springer-Verlag, New York Inc, New York, NY, USA (2005)
9. Vázquez-Ingelmo, A., García-Peñalvo, F.J., Therón, R.: Addressing fine-grained variability in user-centered software product lines: a case study on dashboards. In: *World Conference on Information Systems and Technologies*, pp. 855–864. Springer (2019). https://doi.org/10.1007/978-3-030-16181-1_80
10. Vázquez-Ingelmo, A., García-Peñalvo, F.J., Therón, R.: Taking advantage of the software product line paradigm to generate customized user interfaces for decision-making processes: a case study on university employability. *PeerJ Comput. Sci.* **5**, e203 (2019). <https://doi.org/10.7717/peerj-cs.203>
11. Boer, H.D.: The lie-factor/baseline paradox. Medium. <https://medium.com/@hijedeboer/the-lie-factor-baseline-paradox-ec5955393d19>. Accessed (2020)
12. Tufte, E., Graves-Morris, P.: *The visual display of quantitative information 1983*, ed. Cheshire. Graphics Press, CT, USA (2014)
13. Williamson, D.F., Parker, R.A., Kendrick, J.S.: The box plot: a simple visual method to interpret data. *Ann. Intern. Med.* **110**(11), 916–921 (1989)
14. Finch, W.H., Finch, M.E.H.: Poverty and Covid-19: rates of incidence and deaths in the United States during the first 10 weeks of the pandemic. *Front. Sociol.* **5**, 47 (2020)
15. Kale, A., Nguyen, F., Kay, M., Hullman, J.: Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE Trans. Visual Comput. Graph.* **25**(1), 892–902 (2018)
16. Hullman, J., Kay, M.: Uncertainty+Visualization, Explained <https://medium.com/multiple-views-visualization-research-explained/uncertainty-visualization-explained-67e7a73f031b>. Accessed 2019

17. Kay, M., Kola, T., Hullman, J.R., Munson, S.A.: When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 5092–5103 (2016)
18. Matejka, J., Fitzmaurice, G.: Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 1290–1294 (2017)

