The Importance of Tracing Data Through the Visualization Pipeline

Aritra Dasgupta
UNC Charlotte
adasgupt@uncc.edu

Robert Kosara
UNC Charlotte
rkosara@uncc.edu

ABSTRACT

Visualization research focuses either on the transformation steps necessary to create a visualization from data, or on the perception of structures after they have been shown on the screen. We argue that an end-to-end approach is necessary that tracks the data all the way through the required steps, and provides ways of measuring the impact of any of the transformations. By feeding that information back into the pipeline, visualization systems will be able to adapt the display to the data to be shown, the parameters of the output device, and even the user.

1. INTRODUCTION

Most visualization models consider the representation on screen as the endpoint of the visualization process, and treat it as a transparent window to the data. This model is incomplete at best, and actually quite misleading. In the effort to amplify human cognition [4], the source data is transformed multiple times, in ways we know in principle, but the implications of which we are largely ignorant of: numerical instabilities, scaling, rounding effects when numbers are turned into coordinates for rendering, etc.

On the other side of the screen are the user's perceptual and cognitive systems. While we know many issues that can limit the user's ability to see and grasp the data, there are few ways of predicting when these will occur. Is the display cluttered? Is the display ambiguous? Is there data hidden behind other data? [9] Often, the user may not even be aware that data can be hidden, much less have a way of finding out whether there is hidden data or not.

So far, in visualization, the focus has primarily been on tackling *known unknowns*: check for data quality, represent uncertainty in the data, design visualizations so that users can perform certain analytical tasks, etc. We argue that it is also equally necessary to analyze what are the *unknown unknowns*: what effects the progressive data transformation in the visualization pipeline has on the artifacts produced on the screen and their subsequent perceptual implications. This necessitates a measurement framework for visualization that tracks the data through the pipeline onto the screen, and ultimately all the way into the user's mind (Figure 1). Once we have a clearer understanding of what happens in the pipeline, as well as how changes affect the results, we can create visualization tools that adapt to the data, output parameters like screen size and pixel resolution, and even the user's abilities to understand visual representations. This will enable us to reduce the probability of occurrence of unknown unknowns by increasing the number of known unknowns.

Measuring the results of the visualization process opens up further possibilities, like privacy protection. By controlling the loss of information for keeping it at a certain minimum level, data can be hidden on purpose in order to protect privacy while still allowing a high level of analytic utility. We have found the results of privacy-preservation by use of screen-space metrics to be of much higher utility than running the data through a conventional anonymization algorithm and visualizing only the results [11]. Privacy preservation is one example, but there are undoubtedly many other potential uses once the measurement framework is in place, like in cases of complex analysis scenarios like high-dimensional and/or time-varying data analysis.

2. KNOWN AND UNKNOWN UNKNOWNS

Exploring unknowns in data is a central part of visual analytics; in fact, the visual analytics mantra, *detecting the expected, discovering the unexpected* [19] is based on the philosophy of there being both known and unknown factors that analysts encounter during their exploratory data analysis process: the analysis questions are not always known by the analysts a priori, but they evolve in the course of their interactions with the data through its visual representation.

However, the visualization process itself also introduces unknowns. Despite the increase in quality and resolution of computer displays, visualization still works in a space with a limited number of discrete pixels. Our lack of understanding of how our perceptual system works also constrains our ability to design effective displays that make the most effective use of our capabilities.

Such unknown unknowns in the screen-space are shown in Figure 2. How much difference in value is represented by a single pixel in a bar chart? (Figure 2(a)) What if the

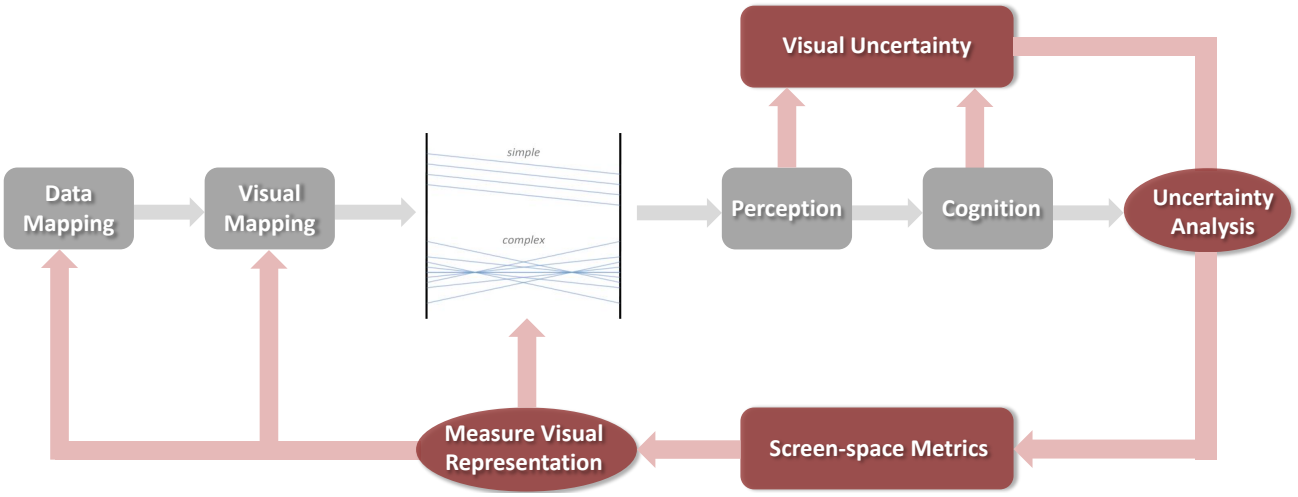


Figure 1: The conventional visualization pipeline (gray) augmented with a feedback loop (red) from the human-side to the machine-side for making visual representations better informed about visual uncertainty and reduce the number of *unknown unknowns* from a visualization designer's point-of-view.

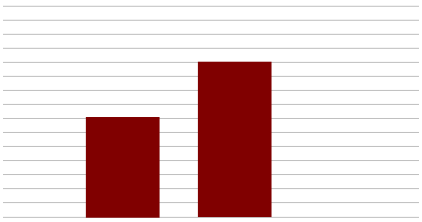
data is depicted by angle, like in a pie chart? What about color? In the case of more complex visualization techniques like parallel coordinates, are there hidden lines behind those that are visible between an axis pair? (Figure 2(b)) How many lines coincide in a particular point on an axis, or on the same line between two axes?

While the first case is an encoding problem due to the visual mapping, the second case is a decoding or perception problem on the human side. In the latter case, although the information is fully encoded, it cannot be effectively communicated to the human mind. While the encoding side is a machine-only process, decoding involves both the human and the machine. Thus there is a need to integrate the human side implications with the machine-side processes into a holistic framework.

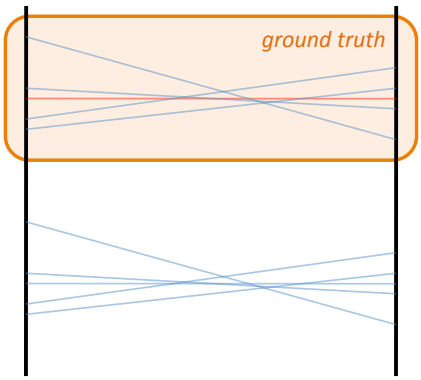
We achieve this by extending the Chi model [7] of the visualization pipeline (Figure 1) by including the perception and cognition stages: data mapping and visual mapping constitute the encoding stages on the machine-side, while perception and cognition constitute the decoding stages on the human side of the pipeline. Then we add a feedback loop that is informed about visual uncertainty [9] and can be used to measure visual representations according to he different levels of uncertainty in the screen-space.

Visual uncertainty takes into account the known and unknown unknowns that stem from the visualization process itself, rather than being present in the data. In previous work we have presented a taxonomy of the various sources, causes and effects of visual uncertainty in parallel coordinates [9].

Measures about visual uncertainty can be used to subsequently optimize the visual representation. Thus the data is followed through the pipeline, the different characteristics of end product is measured and fed back for modifying the earlier stages. This is similar to most engineering systems



(a) How much difference in value is represented by a single pixel in a bar chart?



(b) Are there hidden lines behind those that are visible between an axis pair in parallel coordinates?

Figure 2: Examples of unknown unknowns in visualization: unknown unknown can be both an encoding problem as in case of bar charts and a decoding problem as in case of parallel coordinates.

where the feedback is used to adjust the output. Our augmented visualization pipeline is similar in principle, to the one proposed by Van Wijk [20]. The important distinction is the characteristic of the feedback loop. In our case the feedback loop is used by the visualization designer to build better visual representations through controlled rendering and informed interaction design. In Van Wijk's model, the feedback loop is used by the analyst, based on his perception of the data, to build different representations of the data. Although both are manipulating visual representations, our model uses quantifiable means based on causes and effects of visual uncertainty, and therefore can be modelled and generalized as the basic principles of uncertainty are not affected by the subjectiveness of human judgement.

3. MEASURING VISUALIZATION

Several authors have advocated the need of measuring visual representations through quantitative metrics [2, 17]. This is a hard problem, because the representation has to be not only effective in terms of encoding data properties, but it also has to be efficient in terms of the decoding aspect, that is, communicating the nature of those properties to the end user. Freitas et al. [12] advocate metrics for general categories like presentation of the data and interaction with it. We feel the quest for useful metrics needs to be driven by more fine-grained analysis of what can be quantified and why, how much can be quantified and whether those adequately address both the purposes of encoding and decoding of the data within the visualization pipeline. In this section, we analyze some of the essential characteristics of metrics related to visual uncertainty, that have been proposed in the literature, and those that are needed to be developed to further reduce the number of unknowns in the screen-space.

Sensitivity Analysis Metrics: Visualization of large, high-dimensional data almost always involves information loss due to the disparity between number of dimensions and data points and the limited number of screen pixels. In most applications, this loss is unintended [26], as there is degradation of data fidelity and visual quality. Understanding what change in the data will have a visible effect on screen and which won't (Figure 2(a)) is critical feedback for the user to know which conclusions can be drawn from a display and which ones require additional views.

The artifacts produced due to the disparity between data and display resolution can be measured by metrics based on information theory [23] and metrics for abstraction quality [18]. Effect of dimensionality reduction, if quantified through metrics, can also convey the amount of information loss [15]. These measures are task-independent, helping the analysts build their trust in the visual information that they perceive. Besides resolution, there can also be metrics for measuring the effect of the choice of linear or non-linear scale and the effect of outliers on the visual representation. For visual variables like shape and color, metrics based on the degree of distortion in the screen-space or distinguishability can be developed. Though the current literature on visual quality measures takes some this aspects into account, we still lack the framework for a systematic definition and evaluation of the metrics. A generalization of the visual uncertainty taxonomy for different visual variables can help bridge this gap by facilitating the conceptualization, design, and evaluation of both encoding and decoding metrics.

Pixel-based Metrics: In many visualization designs, the first step to reduce the information overload for the analyst is to present a succinct overview of the data with visual cues that convey the patterns that pop out. Thus, quantifying the visual information content at the macroscopic level is also important, especially in situations where modelling the tasks is not straightforward. In that case, the best solution is to try and quantify the uncertainty at a higher level of granularity, so that the analyst is aware of the unknown unknowns, which in this case are potentially interesting patterns.

Keim at al. were among the first ones to propose pixel-based metrics [16]. Pixel-based entropy, where the pixels themselves have been considered as random variables, has been used to control the transparency [10]. Chen and Jänicke have shown how different information-theoretic concepts like entropy and mutual information can be used at different stages of the visualization pipeline [6]. Dimensions in the screen-space when considered as random variables, the probability of intersection of a record with pixel bin can be used as a basis for computing uncertainty in the screen-space based on Shannon's entropy.

Visualization leverages the pre-attentive processing capabilities of human beings. Different properties of color, like hue, luminance, saturation, etc., has been used successfully for modelling different properties of the data [21] Measuring visual distinctiveness with respect to the use of color or the semantics of color with respect to the data properties is still an open area of research.

Feature-preserving Metrics: Conveying patterns representing semantics of the data attributes is the primary goal of designing visual representations. In recent times, there have been several efforts to quantify screen-space information with respect to the visual structures. While some of them like Scagnostics [22] have been motivated solely by the structures that convey data properties, some others like Pargnostics have tried to bridge the gap between visual structures and user perception by taking Gestalt properties into account. Many of these metrics are task-independent and can be used to provide visual cues for facilitating the pre-attentive processing on the analysts' part. A problem in the current visualization literature is that most of these metrics are termed as visual quality metrics, without a proper definition of the latter. Does quality mean encoding quality? Or does it mean improving decoding capabilities of humans? Or does it take both into account. A survey of the current literature shows that most metrics have focusses mostly on encoding, while decoding is only treated implicitly.

There are two concepts that can help bridge the encoding and decoding sides of visualization. Saliency and novelty is one of them. Capturing saliency, i.e. perceptual prominence of patterns, and novelty, i.e. uniqueness of information, with the help of information-theoretic metrics have been suggested by Chen[5] and promises to be an interesting direction. Visual uncertainty, as we have mentioned earlier, also encompasses both encoding and decoding aspects of the data transformation. For example, identity and traceability

metrics can help quantify clutter, pattern complexity metrics can describe the different visual features. Thus, metrics based on the uncertainty taxonomy can help bridge the screen-space properties and the human-side implications of visual design.

Privacy-aware Metrics: Dealing with sensitive data has so far been an open area of research. To the best of our knowledge, prior to our work [11], no previous research existed that proposes a privacy-preserving visualization technique and the only instance we are aware of uses graph-based abstraction of web data for privacy-preserving manifold visualization [25].

Visualization techniques currently have an underlying assumption that there is unrestricted access to data. In reality, access to data in many cases is restricted to protect sensitive information from being leaked. There are legal regulations like the *Health Insurance Portability and Accountability Act* (HIPAA) in the United States that regulate disclosure of private data. We have shown that adaptation of screen-space metrics based on privacy constraints enable us in achieving more usable privacy-preservation in the screen-space than when visualization is used for displaying the anonymization results after applying data-based metrics.

However, the degree of stability of the privacy-preserving visualizations with respect to attacks by malicious users through use of interactions, still needs to be investigated. There is also the bigger question of the trade-off between privacy and utility as there is significant information loss for hiding sensitive data. Thus, the treatment of unknowns has to be carefully considered based on both privacy and utility metrics. To address these issues, we are currently working on metrics for evaluating privacy-preserving visual representations based on their privacy-guarantee and loss of utility. The ultimate goal is direction is to design an optimization function that balances the privacy and utility of the application.

4. RESEARCH DIRECTIONS

Quantifying visual representations by controlling the data flow and its artifacts within the pipeline can help the visualization designer by considering different options based on the output of the metrics. In this section we outline some of the research directions, which can be facilitated by our ability to measure the impact of data transformation steps in the pipeline.

Data Fidelity Vs Perception: Perceptually beneficial designs do not always result in high data fidelity. Sometimes, certain design choices may motivated by perception, but the results can be counter-intuitive, as shown by a recent study of cluster-based variants of parallel coordinates, which shows that some of them perform worse than the ordinary line-based parallel coordinates [14]. This is mainly due to the fact that the patterns get distorted due to the preprocessing steps and therefore fidelity of the representation is affected. The fact that high data fidelity do not always benefit perception, is easily conceivable in visualization. In case of large data points, we might achieve high data fidelity by mapping all data points on screen, but would be perceptually ineffective owing to clutter. In that case, although encoding uncertainty is minimized, those due to decoding are not addressed. We believe,using visual representation metrics, these trade-offs can be quantified and the findings will be complementary to user studies, which are generally expensive and time-consuming to conduct.

Choice of visual variables: The smallest indivisible unit of all visualizations are the visual variables. While there has been some pioneering work involving the type and categorization of visual variables and their design implications [1, 8, 3], we still lack a proper understanding of the perceptual implications of the selection of visual variables for high-dimensional data analysis. While design choices and their motivations exist sporadically in different research papers, there is a dearth of a framework for their evaluation and comparison. Moreover, perception research [13] has shown that visual variables are not processed independently, but in parallel. Since encoding of high-dimensional data involves multiple visual variables, an interesting problem is how to quantify the additive affect of the visual uncertainty effected by the different variables. When the different uncertainty components are quantified, along with their additive affect, that will enable visualization designers with more informed design choices with respect to selection of visual variables.

Interaction Design: In interactive visualization, different user interaction mechanisms help maximize data fidelity. For example, zooming helps in viewing the data at multiple resolutions and dimension reordering (in case of parallel coordinates) helps in get different perspectives on the multi-dimensional relationships. On the other hand, when there is inherent loss in precision, or there is uncertainty due to traceability, when there are unknown unknowns (the existence of hidden data points), it is difficult to devise interaction techniques to recover such information. Study of interaction techniques and there effectiveness in the realm of visualization has received much less attention. While there has been recent efforts [24] to bridge the machine and human side of interaction, there is still a lack of knowledge on how visual representations and interaction complement each other for analytical tasks. One application of the metrics is to study how interaction techniques can be better informed about the causes and effects of uncertainty that can be reduced, thereby leading the development of an effective user-centric visualization optimization model.

5. CONCLUSION

Visualization has so far mostly been concerned with the forward direction through the pipeline: from the data through the transformation stages to the screen. In order to better control the visualization output and use, and control the unknowns on the screen, we need to close the loop by providing information about the result of the process back to the pipeline. This has to include not only the computational side, though, but also the human side. After all, the best image is still useless if it is not a depiction of the data a person can read and understand.

6. REFERENCES

[1] J. Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[2] R. Brath. Metrics for effective information visualization. *In Proceesings, IEEE Symposium on*

Information Visualization, 1997.

[3] S. Card and J. Mackinlay. The structure of the information visualization design space. In *Information Visualization, 1997. Proceedings., IEEE Symposium on*, pages 92–99. IEEE, 1997.

[4] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., 1999.

[5] C. Chen. An information-theoretic view of visual analytics. *Computer Graphics and Applications, IEEE*, 28(1):18–23, 2008.

[6] M. Chen and H. Jänicke. An information-theoretic framework for visualization. *Transactions on Visualization and Computer Graphics*, 16(6):1206–1215, 2010.

[7] E. H. Chi. A taxonomy of visualization techniques using the data state reference model. In *Proceedings Information Visualization*, pages 69–75. IEEE CS Press, 2000.

[8] W. Cleveland and R. McGill. Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society.*, 150(3):192–229, 1987.

[9] A. Dasgupta, M. Chen, and R. Kosara. Conceptualizing visual uncertainty in parallel coordinates. *Computer Graphics Forum*, 31(3pt2):1015–1024, 2012.

[10] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *Transactions on Visualization and Computer Graphics*, 16(6):1017–26, 2010.

[11] A. Dasgupta and R. Kosara. Adaptive privacy-preservation using parallel coordinates. *Transactions on Visualization and Computer Graphics*, 17(12):2241–2248, 2011.

[12] C. Freitas, P. Luzzardi, R. Cava, M. Winckler, M. Pimenta, and L. Nedel. On evaluating information visualization techniques. In *Proceedings of the working conference on Advanced Visual Interfaces*, pages 373–374. ACM, 2002.

[13] M. Green. Towards a perpectual science of multidimensional data visualization: Bertin and beoyond. http://www.ergogero.com/dataviz/dviz0.html, 1996.

[14] D. Holten and J. Van Wijk. Evaluation of cluster identification performance for different pcp variants. *Computer Graphics Forum*, 29(3):793–802, 2010.

[15] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics*, 15:993–1000, 2009.

[16] D. Keim. Designing pixel-oriented visualization techniques: theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6:59–78, 2000.

[17] N. Miller, B. Hetzler, G. Nakamura, and P. Whitney. The need for metrics in visual information analysis. In *NPIV '97: Proceedings of the 1997 workshop on New paradigms in information visualization and manipulation*, pages 24–28. ACM, 1997.

[18] E. A. Rundensteiner, M. O. Ward, Z. Xie, Q. Cui, C. V. Wad, D. Yang, and S. Huang. Xmdvtool: Quality-aware interactive data exploration. In *SIGMOD Conference*, pages 1109–1112, 2007.

[19] J. Thomas and K. Cook. A visual analytics agenda. *Computer Graphics and Applications, IEEE*, 26(1):10–13, 2006.

[20] J. Van Wijk. The value of visualization. In *IEEE Visualization*, pages 79–86, 2005.

[21] C. Ware. *Visual thinking for design*. Morgan Kaufmann Pub, 2008.

[22] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings Information Visualization*, pages 157–164. IEEE CS Press, 2005.

[23] J. Yang-Peláez and W. C. Flowers. Information content measures of visual displays. In *Proceedings of the IEEE Symposium on Information Vizualization 2000*, pages 99–103. IEEE Computer Society, 2000.

[24] J. Yi, Y. ah Kang, J. Stasko, and J. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1224–1231, 2007.

[25] X. Zhang, W. K. Cheung, and C. H. Li. Graph-based abstraction for privacy preserving manifold visualization. In *Proceedings of the Conference on Web Intelligence and Intelligent Agent Technology*, pages 94–97, Washington, DC, USA, 2006. IEEE Computer Society.

[26] C. Ziemkiewicz and R. Kosara. Embedding Information Visualization Within Visual Representation. *Advances in Information and Intelligent Systems*, pages 307–326, 2010.