Contextual Visualization

Making the Unseen Visible to Combat Bias During Visual Analysis

David Borland
University of North Carolina at Chapel Hill

Wenyuan Wang
University of North Carolina at Chapel Hill

David Gotz
University of North Carolina at Chapel Hill

Editor:
Theresa-Marie Rhyne
theresamarierhyne@gmail.com

Unseen information can lead to various "threats to validity" when analyzing complex datasets using visual tools, resulting in potentially biased findings. We enumerate sources of unseen information and argue that a new focus on contextual visualization methods is needed to inform users of these threats and to mitigate their effects.

As technology has grown more ubiquitous and affordable, the mass collection of large and complex datasets has become widespread across a range of domains. Websites gather detailed logs of user behavior. Transportation systems gather detailed logs of traffic, accidents, and other incidents. Health systems are amassing large databases of electronic medical records for patients. In these domains and others, such "big data" resources are being gathered in ever-increasing numbers with the promise of supporting precision, evidence-based decision making. This vision is tantalizing and widespread, and there is enormous enthusiasm for the value of data-driven insights across these numerous and varied application areas. The excitement reflects a core promise of big data: that by capturing data "in the wild"—with huge numbers of variables and at enormous scale—it is possible to gain more detailed, precise, and nuanced insights into complex problems.

At the same time, advances in visual analytics technologies have led to a much broader use of visualization-based exploratory analysis tools. Visual analytics systems combine interactive exploratory visualization and data analysis capabilities to put humans "in the loop," amplifying human cognition to enable improved problem solving.[1] Visual analytics techniques go well beyond the classic visualization dashboard model made popular by earlier generations of business intelligence software. As a result, visual analytics systems are being used to help users solve increasingly complex analytical tasks.

Visualization technologies designed to support these two trends—(1) increasing data complexity and (2) increasing cognitive task complexity—are key enablers for the growing use of big data analytics and data-driven decision making. Visualization is already widely viewed as a critical technology for more effective data analysis, interpretation, and communication.[2] Moreover, new advances in the field continue to extend our ability to work with more complex data and more complex cognitive tasks, increasing the value of visual analytics tools.

Published by the IEEE Computer Society
0272-1716/19/$33.00 ©2019 IEEE

Often, the challenge of increased complexity is assumed to be a problem of data volume. However, visualization can be ideally suited to communicating large volumes of data by directly summarizing huge numbers of data records with simple visual representations of aggregate measures. For instance, a bar chart showing the distribution of a single categorical variable (e.g., the proportion of men and women in a group of people) works equally well with ten records as with 1 billion records. Moreover, significant progress has been made to address other challenges of scale: (1) the computational challenges of computing aggregate statistics from large-volume data (e.g., cloud computing and the map-reduce framework[3]); and (2) progressive visualization techniques for integrating long-running computations with interactive visual representations.[4]

Data and task complexity, however, pose a more fundamental challenge to the science of visualization: *unseen information*. Well-designed visualizations are effective because they graphically communicate data in such a way that users' visual perception can detect patterns and derive insight. However, when either a dataset or a user's analytical tasks grow complex, key information may be left out of the visualization. This can be due to both (1) the inherent difficulty in visualizing and conceptualizing high-dimensional data and relationships, and (2) the fact that human cognition incorporates contextual information—such as expertise and derived insights—that is often not directly represented within the data made available to the underlying visualization system.

UNSEEN THREATS TO VISUALIZATION VALIDITY

The fact that a variety of critically important information can be omitted from a visualization threatens the fundamental validity of many visual tools that users are now employing to solve their analytical problems. The "threats to validity" posed by this unseen information can come from a variety of sources, including

- *summarization* methods that can obscure fine-grained information;
- *narrow fields of view* that focus only on portions of a dimension;
- *omitted dimensions* of data that are not visually represented within a visualization;
- *external data* not contained within the dataset under analysis;
- *cognitive challenges*, such as contextual knowledge and cognitive biases.

These threats to visualization validity are in many ways analogous to well-known challenges in statistics, experimental design, and observational studies.[5] For example, any approach involving statistical inference requires a set of assumptions—a statistical model—the validity of which will drastically impact the quality of any statistical analysis. Statisticians must diligently maintain awareness of these threats and mitigate their effects to ensure the validity of their analysis. Similarly, we argue, visual analysts must maintain awareness of these issues and endeavor to overcome them to produce valid and meaningful visual discoveries. Moreover, visualization tools should support this awareness by employing appropriate contextual visualization methods. Creating more robust visual analytics tools will enable more robust analysis across a wide range of disciplines.

Summarization

Widely used in visualization systems, summarization algorithms are critical tools for reducing data complexity. For instance, aggregation can be used to combine statistics for multiple variables within a single visualized measure. Consider the visualization of health data in Figure 1, showing recent estimates for the incidence rates of diabetes and heart disease in the United States. This view can provide valuable information about which of these two conditions is most prevalent, but also obscures the fact that each condition is actually a heterogeneous category representing multiple different manifestations of disease. For example, while recent estimates suggest that 9.4% of Americans were diabetic, this population was divided unevenly between Type 1 (insulin-dependent, ∼4% of diabetics) and Type 2 (adult-onset, ∼96% of diabetics) diabetes. Moreover, each of these categories can be further refined into fine-grained disease descriptors. While some summarization methods, such as aggregation and summation, hide data in predictable ways, other more sophisticated summarization methods, such as topic modeling or spatial embedding algorithms (e.g., t-SNE or MDS) can be less transparent about what information is lost. This issue is exacerbated as tasks grow in complexity,

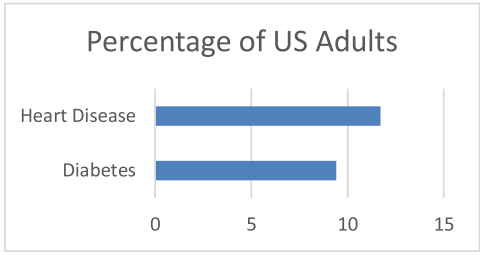

Figure 1. Disease incidence rates for adults within the United States. Based on statistics from the Centers for Disease Control and Prevention and the American Diabetes Association.

where more sophisticated (and less transparent) summarization methods may be needed, and the more nuanced information that is lost during summarization may be important to consider.

Narrow Fields of View

Visualizations will often constrain the data they display to a subset of a given dimension. For example, a map-based visualization will typically provide a view of data within a specific region rather than the entire dataset. This enables the visualization to display more fine-grained data (with less summarization), but also hides data from outside the focused region. Similar approaches are used in various visualization types, including three-dimensional volume datasets rendered within scientific visualization, times series visualizations, and even along more abstract dimensions using techniques such as zoomable hierarchical visualizations that enable users to focus on subtrees within a hierarchy.

Omitted Dimensions

When dealing with high-dimensional datasets, it is typically necessary to visualize a small subset of the dimensions at any given time. Techniques such as dimension selection and dimensionality reduction can be used to reduce the number of dimensions to be visualized, but result in a loss of information. This is true even for datasets with relatively few dimensions (e.g., 10–20 dimensions) where it can be difficult to display all attributes simultaneously. However, omitting dimensions is especially common when visualizing very high-dimensional data, such as electronic health record (EHR) data, where the number of dimensions can be in the tens of thousands or more. For example, the widely used ICD-10-CM coding system contains approximately 70 000 diagnosis codes (just one of several types of medical information in the EHR). While often necessary, omitting dimensions from a visualization can obscure shifts in variable distributions due to operations such as filtering. Yet visibility of such shifts is critical because it can signify validity problems such as selection bias or systemic data quality issues that can bias the user's visual findings.

External Data

The threats to validity described so far address challenges related to a visualization being unable to depict the entirety of a dataset. It is also critically important to recognize that even the entirety of a dataset provides an incomplete record of the real world for many tasks. Even very large and complex datasets often only capture certain aspects of a problem, with relevant variables remaining unrecorded. Moreover, even for variables where data have been captured, systemic properties of the data recording process can result in biased sampling. For example, consider a scenario related to traffic monitoring and accident prediction. A dataset may capture traffic volumes and accident locations at various times of the day. This data can be used as the basis for analysis, but if weather data—clearly a contributing a factor in some accidents and traffic delays—are omitted, it would not be possible for a visualization to provide a comprehensive representation of the factors leading to traffic delays. Similarly, if traffic data are captured predominantly from buses and taxis rather than private vehicles, any analysis of the data would be biased toward the detection of predictive factors associated with public transportation. The threat of missing external data is especially problematic given the growing complexity of tasks to

which visual analytics technologies are being applied. As analysts aim to answer questions about increasingly complex phenomena, it becomes ever more important to obtain comprehensive and representative datasets to support their analytical goals.

Cognitive Challenges

The mind of the analyst is itself a source for an additional class of threats to visualization validity. First, an analyst brings to any analysis task a wide range of preconceptions and background knowledge that color their analytical activities, e.g., the amount of expertise the analyst has within the domain of investigation. This background knowledge can be essential in helping an analyst successfully complete a task in many cases. For example, heuristic "short cuts" for filtering out unnecessary information are often useful in problem solving.[6] Such expert knowledge is often implicitly leveraged, without being explicitly recorded. On the other hand, recent research shows that providing background information can negatively impact analytical accuracy for novices.[7] This finding underlines the need for analytical systems that work in concert with users to help understand complex data. Other unseen forces that influence an analyst come from a variety of cognitive biases. These include anchoring and framing effects, in which an analyst's interpretation of data is influenced by what they have seen previously. Similarly, recency bias and confirmation bias suggest that users are drawn to conclusions that support previously discovered findings, especially when those findings are recent discoveries. These forces suggest that the mental order and structure with which analysts organize their findings have a significant impact on their analytical behavior even if such information is not explicitly captured within a visualization system.

SEEING THE UNSEEN WITH CONTEXTUAL VISUALIZATION

The threats to visualization validity identified above are not in and of themselves new concerns for the visualization community. Indeed, some of the challenges (especially summarization and narrow fields of view) have been studied extensively, with well-known methods designed to counter them. However, as we apply visual analytics techniques to increasingly complex datasets and analytical tasks, many of these challenges are growing in significance while simultaneously becoming more difficult to address.

In this section, we put forth the argument that this growing problem requires more attention. More specifically, we suggest that what is required is a set of *contextual visualization methods* designed to capture and display the information required to maintain awareness of these threats during visual analysis and to help analysts mitigate their effects. A variety of approaches will be required, in some cases leveraging existing approaches, while in others relying on new innovation to tackle emerging problems.

Summarization

Interactive visualization methods that follow Shneiderman's mantra—"Overview first, zoom and filter, then details-on-demand"[8]—aim to make visible this form of hidden information by providing exploratory tools that let users focus attention on more detailed representations of narrower subsets of data. For example, in Figure 1, the user might be able to click on each of the bars to view a more detailed breakdown of incidence rates within each top-level disease category. These approaches are widespread and effective and should be aggressively leveraged to help provide a richer context for high-level summaries. However, as datasets increase in size and complexity, the discoverability of interesting details can be difficult. Therefore, the development of computational methods that help guide users to discover areas of relevant or interesting detail will be important.

Narrow Fields of View

As described above, zooming is a widely used approach to providing users with a view of details that are otherwise hidden due to summarization. Focus + context techniques are designed to contextualize the narrow fields of view produced by zooming with a view of data that is beyond the area of focus. For example, mapping applications often use minimaps of large geographic regions to indicate where a

zoomed map's current narrow field of view is positioned. Other focus + context techniques, such as the volumetric visualization in Figure 2, embed the focus areas within a larger contextual object to help highlight specific areas of interest.[9]

Narrow fields of view are also used to manage nonspatial dimensions. For example, consider a scientific visualization application in which a volume rendering transfer function applies increased

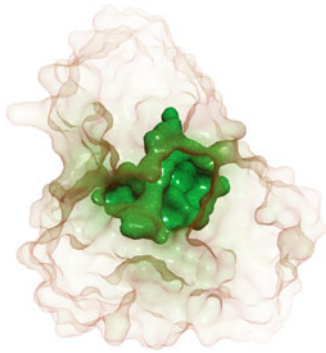

Figure 2. A focus + context view of a molecular cavity (green) within the context of the larger surface.

opacity to voxels within a given range of interest to mitigate issues with occlusion in volumetric data.[10] In this form of visual query, where users can interactively explore different ranges as part of an analysis, advanced data selection widgets could improve the visibility of context by providing visualizations of data values for the focused dimension throughout the entire range (e.g., scented widgets[11]).

A common example of this approach would be time-series visualizations that contain overviews of an entire time series to contextualize a detailed visualization of a narrow time span. For instance, a stock-market visualization could help prevent misleading views of price trends over very short time-spans by showing contextually that the "short-term trend" is actually just a small fluctuation within a more significant long-term change in prices.

As datasets grow larger and more complex, the need to overcome limitations of summarization is making narrow fields-of-view an even more important tool for exposing the details of a dataset and the nuanced insights that those details inform. However, the increasing complexity also makes it more difficult for users to maintain awareness of the larger context as they shift areas of focus. We argue that this calls for a more widespread use of focus + context methods, as well as the development of more advanced capabilities for data types where the larger context is harder to represent.

Omitted Dimensions

Various techniques have been developed to facilitate navigation in high-dimensional information spaces. The most basic approach is to provide users with the ability to control which subset of variables are visualized at a given time by selecting from a list of available dimensions. To help users understand which subspaces of the data they have examined, recent work has extended the concept of scented widgets to include a visual representation of the combinations of dimensions that have already been visualized within in a given session.[12] This can help users see which dimensions are available in a dataset, as well as information about which dimensions have not yet been explored. However, these methods do not fully address the threats to validity issues outlined earlier in this paper.

Most critically, omitting dimensions from a visualization can hide issues such as selection bias introduced during filtering and zooming. Recent research has shown that the effects of selection bias in high-dimensional visualization can be communicated to users by capturing a user's exploration history and visualizing differences in variable distributions across that history (see Figure 3).[13] Future innovation in this area, including capturing more complex analyses and integrating methods to correct for bias (e.g., post-stratification[14]) is necessary to make ad hoc exploratory visual analysis more robust.

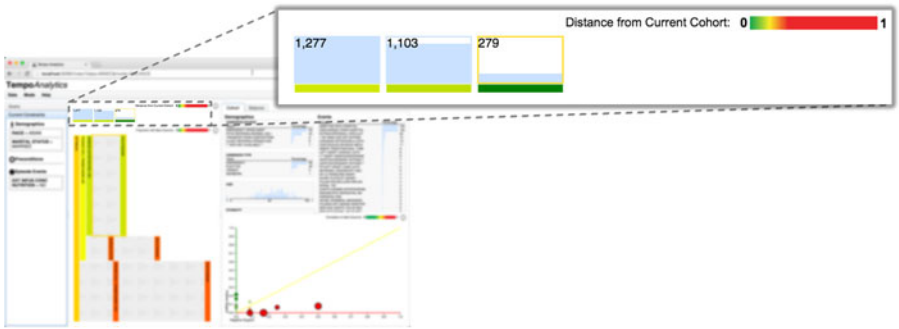

Figure 3. A visual analytics system instrumented to detect selection bias during exploratory data selection. The bread-crumb view in the callout shows increasing drift (lighter green bars) moving back from the current analytical focus (right) to the original cohort (left).

External Data

Unobserved data are perhaps the most difficult to address of the challenges enumerated in this paper, and perhaps as a consequence have not received as much attention from the visual analytics community. Any algorithmic approach must recognize that the missing variables or nonrepresentative samples at the root of this family of challenges cannot be known to the system. One potential direction to explore is the integration of structured reasoning approaches that encourage users to assess the quality and comprehensiveness of the underlying data sample as part of their interactive analysis workflow. Another possibility is formalizing the use of external "reference" data sources when such information is available. For example, an analysis of online activity that is aimed at understanding the general population could contextualize its findings using census data as a relatively complete profile of the broader population.

Cognitive Challenges

Cognitive biases can have a great impact on visual analysis quality. As with external data issues, structured reasoning techniques such as analysis of competing hypotheses have some promise as tools to help address some of these issues. However, these approaches typically require users to be diligent in following a process designed to overcome known cognitive weaknesses.

Recent research has begun to explore more automated approaches that track users' interaction behavior within a visual analytics system and compute scores to assess levels of cognitive bias.[15] These sorts of automated approaches could be quite valuable in providing users a contextual knowledge of cognitive threats to their analysis results. However, it remains a grand challenge to develop meaningful metrics that can reliably distinguish between flawed analytical behavior (e.g., a false conclusion due to confirmation bias) and effective sensemaking behavior that "connects the dots" across a set of truly related observations. Steps toward improving support for handling cognitive biases in visual analytics systems may include (1) developing improved models of how cognitive biases manifest themselves within visual analytic activity, and (2) based on these models, providing recommendations for human analysis workflows or computational tools that may be able to avoid, detect, and overcome cognitive biases.

CONCLUSION

One of the great promises of interactive visualizations is to provide opportunities for supporting evidence-based decision making. However, the threats to validity due to the complexity of very high-dimensional datasets and unseen cognitive factors create an environment with a high risk of producing misleading or entirely erroneous findings. To more effectively meet this promise, the visualization community must continue efforts to uncover these hidden threats to validity and make them apparent to the user. Such contextualization methods will enable more detailed, nuanced, precise, and accurate insights into complex problems requiring sophisticated analysis. If visual analytics technologies are to reach their full potential, this topic should be a key priority for future visual analytics research.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under Grant 1704018.

REFERENCES

1. J. Thomas and K. Cook, "Illuminating the path: The research and development agenda for visual analytics," US Dept. Homeland Secur., Washington, DC, USA, 2005.
2. T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, 1st ed. Redmond, WA, USA: Microsoft Res., 2009.
3. M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets," in *Proc. 2nd USENIX Conf. Hot Topics Cloud Comput.*, Berkeley, CA, USA, 2010, pp. 10.
4. C. D. Stolper, A. Perer, and D. Gotz, "Progressive visual analytics: User-driven visual exploration of in-progress analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1653–1662, Dec. 2014.
5. W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, 2nd ed. Belmont, CA, USA: Wadsworth, 2001.
6. G. Gigerenzer, "Why heuristics work," *Perspectives Psychol. Sci.*, vol. 3, no. 1, pp. 20–29, Jan. 2008.
7. E. Dimara, A. Bezerianos, and P. Dragicevic, "Narratives in crowdsourced evaluation of visualizations: A double-edged sword," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2017, pp. 5475–5484.
8. B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Proc. IEEE Symp. Vis. Lang.*, 1996, pp. 336–343.
9. D. Borland, "Ambient occlusion opacity mapping for visualization of internal molecular structure," *J. WSCG*, vol. 19, nos. 1–3, pp. 17–24, 2011.
10. J. Kniss, G. Kindlmann, and C. Hansen, "Multidimensional transfer functions for interactive volume rendering," *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 3, pp. 270–285, Jul. 2002.
11. W. Willett, J. Heer, and M. Agrawala, "Scented widgets: Improving navigation cues with embedded visualizations," *IEEE Trans. Vis. Comput. Graph.*, vol. 13, no. 6, pp. 1129–1136, Nov. 2007.
12. A. Sarvghad, M. Tory, and N. Mahyar, "Visualizing dimension coverage to support exploratory analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 21–30, Jan. 2017.
13. D. Gotz, S. Sun, and N. Cao, "Adaptive contextualization: Combating bias during high-dimensional visualization and data selection," in *Proc. 21st Int. Conf. Intell. User Interfaces*, New York, NY, USA, 2016, pp. 85–95.
14. D. Holt and T. M. F. Smith, "Post stratification," *J. Roy. Statist. Soc.*, vol. 142, no. 1, pp. 33–46, 1979.
15. E. Wall, L. Blaha, L. Franklin, and A. Endert, "Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics," in *Proc. IEEE Vis. Anal. Sci. Technol.*, 2017.

ABOUT THE AUTHORS

David Borland is a Senior Visualization Researcher with the Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. Contact him at borland@renci.org.

Wenyuan Wang is a Ph.D. student at the School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. Contact him at vaapad@live.unc.edu.

David Gotz is an Associate Professor with the School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. Contact him at gotz@unc.edu.

Contact department editor Theresa-Marie Rhyne at theresamarierhyne@gmail.com.