Top 10 Unsolved Information Visualization Problems

Chaomei Chen

Drexel University

A thought-provoking panel, organized by Theresa-Marie Rhyne, at IEEE Visualization 2004 addressed the top unsolved problems of visualization.[1] Two of the invited panelists, Bill Hibbard and Chris Johnson, addressed scientific visualization problems. Steve Eick and I identified information visualization problems. The following top 10 unsolved problems list is a revised and extended version of the information visualization problems I outlined on the panel. These problems are not necessarily imposed by technical barriers; rather, they are problems that might hinder the growth of information visualization as a field. The first three problems highlight issues from a user-centered perspective. The fifth, sixth, and seventh problems are technical challenges in nature. The last three are the ones that need tackling at the disciplinary level.

In this article, I broadly define information visualization as visual representations of the semantics, or meaning, of information. In contrast to scientific visualization, information visualization typically deals with nonnumeric, nonspatial, and high-dimensional data.

1. Usability

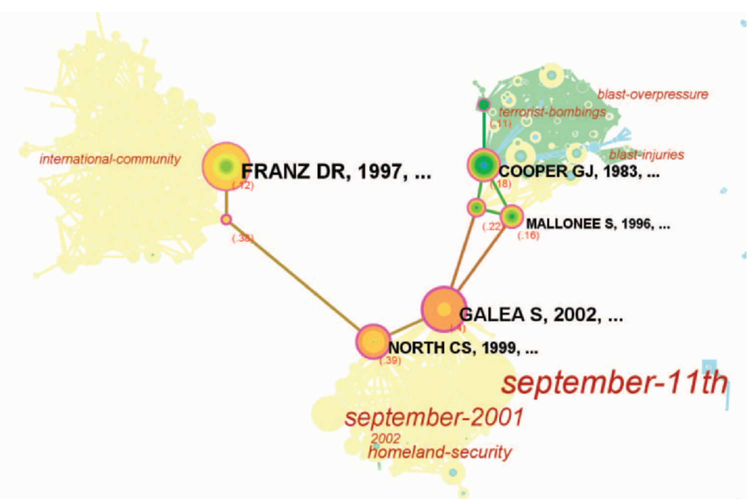
The usability issue is critical to everyone, especially in light of successful commercialization stories such as Spotfire (http://www.spotfire.com/) and Inspire (http://in-spire.pnl.gov/). Although the overall growth of information visualization is accelerating, the growth of usability studies and empirical evaluations has been relatively slow. Furthermore, usability issues still tend to be addressed in an ad hoc manner and limited to the particular systems at hand.

The complexity of the underlying analytic process involved in most information visualization systems is a major obstacle; end users cannot see how their raw data is magically turned into colorful images. The first collection of empirical studies is the 2000 special issue in *International Journal of Human–Computer Studies*.[2] Although the number of empirical studies of information visualization systems is increasing, designers and users still need to find empirical evidence that is both generic and specific enough to inform their decision-making processes.

Empirical studies tend to use open source and freely available systems.[3] A prolonged lack of low-cost, ready-to-use, and reconfigurable information visualization systems will have an adverse impact on cultivating the critical user population. A balanced portfolio of general-purpose, fully functional information visualization systems is essential from user- and learning-centered perspectives.

We need new evaluative methodologies. The majority of existing usability studies heavily relied on methodologies that predated information visualization. Such methodologies are limited because we cannot expect them to address critical details specific to information visualization needs.

There might be an even more profound reason for the shortage of usability studies. Information visualization is a visual exploration tool that enables the user to interact with the visualized content and comprehend its meaning. The comprehension process is often exploratory in nature. For example, users can interact with many possible cognitive paths in the network visualization shown in Figure 1 and interpret what they see.[4] Usability studies need to address whether users can recognize the intended patterns.



1 Prominent paths in this bibliographic network visualization highlight how different research topics are connected in research on terrorism.

 Published by the IEEE Computer Society

Because this involves interrelated perceptual–cognitive tasks, existing methodologies for empirical studies might not be readily applicable. This observation leads to the next challenging problem.

2. Understanding elementary perceptual–cognitive tasks

Understanding elementary and secondary perceptual–cognitive tasks is a fundamental step toward engineering information visualization systems. The general understanding of elementary perceptual–cognitive tasks must be substantially revised and updated in the context of information visualization.

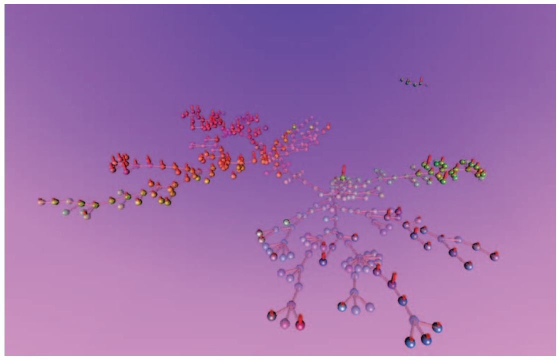
Information retrieval has had a profound impact on the evolution of information visualization as a field. Many task analysis and user studies framed interacting with information visualization as an information retrieval or an open-ended browsing problem. However, using browsing and search tasks to study users' perceptual and cognitive needs in the process of interacting with information visualization is likely to miss the target. Tasks such as browsing and searching, and even judging the relevance of information, require a level of cognitive activities higher than that of identifying and decoding visualized objects. In this sense, a mismatch exists between studying the high-level user tasks and evaluating the usefulness of visualization components.

Studies of elementary perceptual–cognitive tasks appear in the earlier psychology and statistical graphics literature, including the Cleveland–McGill study and the work of Treisman on preattentive perceptual tasks.[5,6] In the context of information visualization, while researchers have done considerable work, notably through Ware's work in characterizing motion and stereo depth perception tasks in visualization, we have a great deal more to accomplish.[7]

Above the elementary perceptual–cognitive task level, we need to collect a substantial amount of empirical evidence from the new generation of information visualization systems. The secondary level perceptual–cognitive tasks include the recognition of a cluster of dots based on their proximity, the identification of a trend based on a time series of values, or the discovery of a previously unknown connection. This would echo the moment of "aha!" when an insightful discovery is made. For example, what perceptual–cognitive tasks are in play when we see animated visualizations such as the one in Figure 2? Studies of individuals' spatial ability and fixations of eye movements are approaching this secondary level.

3. Prior knowledge

This seemingly philosophical problem has many practical implications. As a vehicle for communicating abstract information, information visualization and its users must have a common ground. This is consistent with the user-centered design tradition in human–computer interaction (HCI). A thought-provoking example of prior knowledge is the visual message carried by the Pioneer spacecraft (http://spaceprojects.arc.nasa.gov/Space_Projects/pioneer/PN10&11.html#plaque). The intended extraterrestrial audience is assumed to know



2 **3D animated visualization of the evolution of research in mad cow disease showing the importance of understanding secondary level perceptual–-cognitive tasks.**

modern physics and our solar system. The alien is also expected to figure out from the line drawings of a man and woman that the Pioneer is coming in peace from a small planet. Research in preattentive perception also studies the role of prior knowledge.

In general, users need two types of prior knowledge to understand the intended message in visualized information:

- the knowledge of how to operate the device, such as a telescope, a microscope, or, in our case, an information visualization system, and
- the domain knowledge of how to interpret the content.

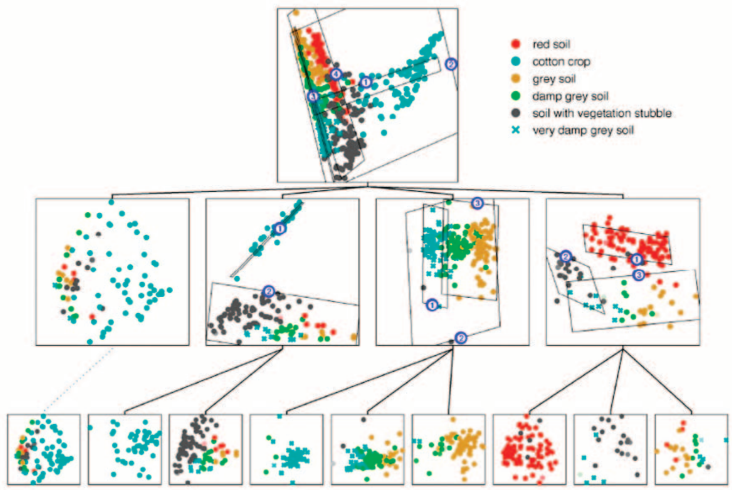
Therefore, design decisions must be made up front in terms of the level of prior knowledge necessary to understand the visualized information. The prior knowledge problem can be seen as a need for adaptive information visualization systems in response to accumulated knowledge of their users.

Solutions to the first two challenges discussed earlier can reduce the dependence on the first type of prior knowledge, but they cannot replace the need for the domain knowledge. In the Pioneer example, if the alien does not have the expected knowledge of physics and the ability to make various bold connections, then the Pioneer's message is meaningless.

4. Education and training

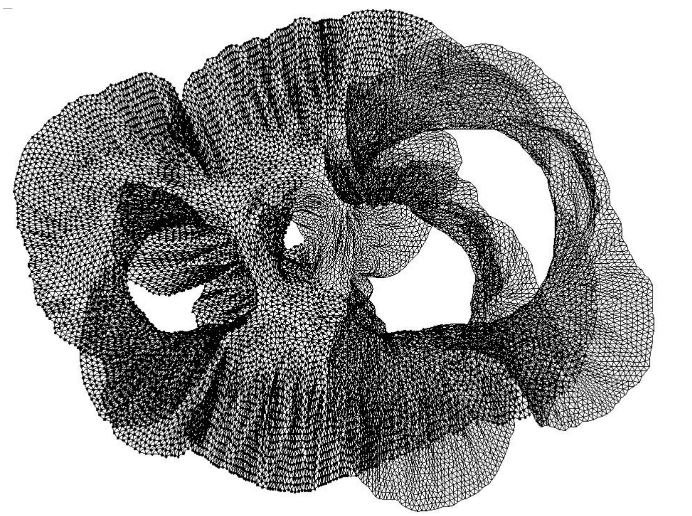
The education problem is the fourth user-centered challenge. We are facing the challenge internally and externally. The internal aspect of the challenge refers to the need for researchers and practitioners within the field of information visualization to learn and share various principles and skills of visual communication and semiotics. To reach a critical mass, the language of information visualization must become comprehensible to its potential users. Universities should connect undergraduate and graduate programs to more advanced research programs and development efforts. Regularly revising existing taxonomies in light of new systems and exemplars will consolidate the field's theoretical foundations.

The external aspect of the challenge refers to the need for potential beneficiaries outside the immediate field of information visualization to see the value of information visualization and how it might contribute to their work in an innovative way. To insiders, the value of information visualization might seem obvious. However,



3 **Hierarchical latent variable data visualization model.**

Courtesy of Christopher M. Bishop and Michael E. Tipping



4 **Large graph containing 15,606 vertices and 45,878 edges.**

Courtesy of David Harel and Yehuda Koren

researchers, practitioners, and decision makers might not see it that way. We need compelling showcase examples and widely accessible tutorials for general audiences to raise the awareness of information visualization's potential and, perhaps more importantly, the awareness of problems in other disciplines that existing or innovative approaches in information visualization could resolve.

5. Intrinsic quality measures

It's vital for the information visualization field to establish intrinsic quality metrics. Until recently, the lack of quantifiable quality measures has not been much of a concern. In part, this is because of the traditional priority of original and innovative work in this community. The lack of quantifiable measures of quality and benchmarks, however, will undermine information visualization advances, especially their evaluation and selection.

An intrinsic quality metric will tremendously simplify the development and evaluation of various algorithms. The intrinsic property is required so that you can still derive a quality metric in the absence of external reference sources. The provision of such quality measures will enable usability studies to evaluate the consistency between the best solution based on users' assessments and the best solution based on intrinsic measures. The stress level used by multidimensional scaling (MDS) algorithms is a good example of such metrics. The goal of MDS is to project high-dimensional data to two or three dimensions while minimizing the overall distortion. The lower the stress level is, the better the MDS solution. In general, information visualization lacks such metrics.

This is a particularly challenging problem, but it also has a potentially far-reaching reward because intrinsic quality metrics will answer key questions such as, to what extent does an information visualization design represent the underlying data faithfully and efficiently, and to what extent does it preserve intrinsic properties of the underlying phenomenon?

Integrating machine learning and information visualization is potentially fruitful. Figure 3 shows a hierarchical latent variable data visualization model.[8] Probabilistic models have received much attention in areas such as topic detection and trend tracking, adaptive information filtering, and detecting concept drifts in streaming data. An excellent overview of mixture models is available elsewhere.[9] Model-based approaches have several unique advantages in terms of probabilistic inference and model selection. Defining quantitative metrics of quality in terms of the likelihood or uncertainty that given raw data is represented by latent models is an attractive starting point.

6. Scalability

The scalability problem is a long-lasting challenge for information visualization. Figure 4 shows a large graph drawn by a fast layout algorithm.[10] The creators drew the 15,606-vertex and 45,878-edge graph within a matter of seconds. Unlike the field of scientific visualization, supercomputers have not been the primary source of data suppliers for information visualization. Parallel computing and other high-performance computing techniques have not been used in the field of information visualization as much as in scientific visualization and a few other fields. In addition to the traditional approach of developing increasingly clever ways to scale up sequential computing algorithms, the scalability issue should be studied at different levels—such as the hardware and the high-performance computing levels—as well as that of individual users.

A relatively recent research interest focuses on the visualization of data streams.[11] The challenge of visualizing data streams is due to the arrival pattern of the data stream and the urgency to understand its contents. Steve Eick identified this issue in his list of problems for the 2004 Visualization panel.

7. Aesthetics

The purpose of information visualization is the insights into data that it provides, not just pretty pictures. But what makes a picture pretty? What can we learn from making a pretty picture and enhancing the representation of insights? It's important, therefore, to understand how insights and aesthetics interact, and how these two goals could sustain insightful and visually appealing information visualization.

The graph-drawing community has done the most advanced research in relation to the aesthetics problem. However, much of the aesthetics wisdom consists more of heuristics than empirical evidence at the elementary level of perceptual–cognitive tasks. There is a lack of holistic empirical studies to characterize what visual properties make users think a graph is pretty or visually appealing. Research in this area often focuses on graph-theoretical properties and rarely involves the semantics associated with the data.

Insights should be primarily identified in the data modeling stage of the process, including feature extraction and filtering operations. Engineering aesthetics into information visualization remains a challenge.

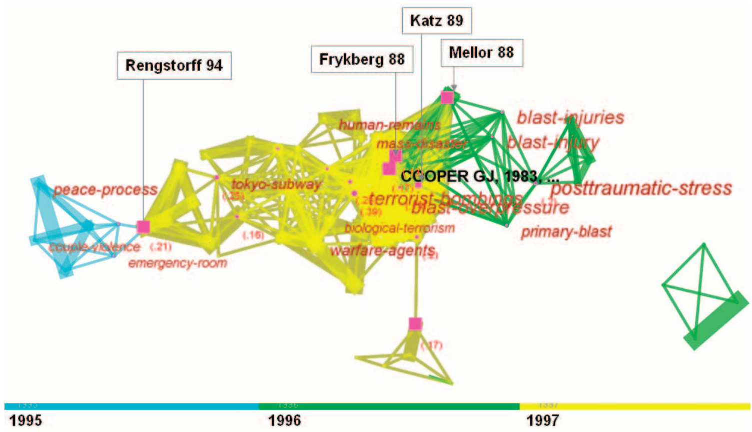
8. Paradigm shift from structures to dynamics

The structure of abstract information was the center of attention during the first generation of information visualizations in the 1990s. The most famous exemplars dealt with structures: cone tree, treemap, and hyperbolic views, for example. The emphasis on structures is logical. An emerging trend is to shift the structure-centric paradigm to the visualization of dynamic properties of underlying phenomena. This paradigm shift is discussed in Chen.[3] Visualizing changes over time is a defining feature of the dynamics paradigm. Earlier examples of visualizing thematic trends include the work of Wong et al.[11]

There are more profound challenges, however, than visualizing change over time. Even if a trend or a sharp change is recorded in a data set, first-order visualizations of the data set's structural and dynamic properties might not feature such trends and changes prominently enough to draw users' attention. Therefore, it's necessary to distinguish two types of visualization processes: one with built-in trend detection mechanisms (see Figure 5), typically as part of the data modeling component; and one without such mechanisms. The majority of contemporary information visualization systems fall into the latter category. To address the lack of trend detection systems, I strongly recommend interdisciplinary collaborations between the data mining and artificial intelligence communities.

9. Causality, visual inference, and predictions

Visual thinking, reasoning, and analytics emphasize the role of information visualization as the powerful medium for finding causality, forming hypotheses, and



5 **Prominent terms and articles detected before and after the Oklahoma City bombing.**

assessing available evidence. Tufte showed some intriguing examples of the power of visual explanation, namely the Challenger space shuttle disaster and John Snow's map of cholera deaths.[12]

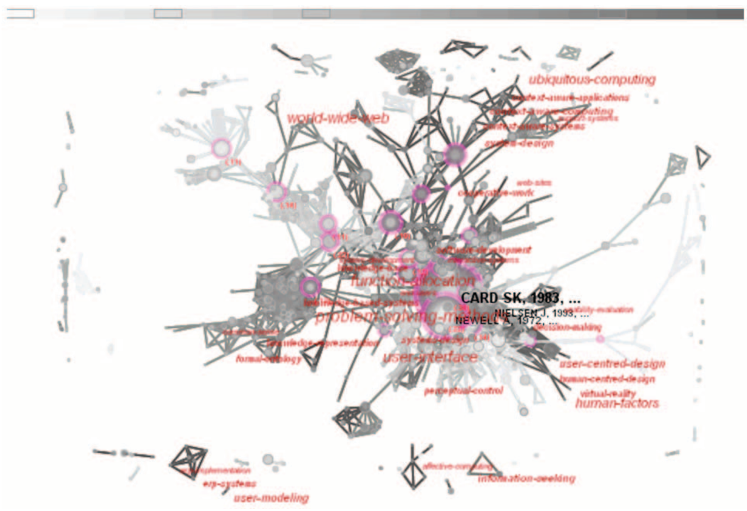
Potentially high-impact areas of applications include evidence-based medicine, technology forecasting, collaborative recommendation, intelligence analysis, and patent examination. For example, the withdrawal of the prescription drug Vioxx in 2004 has prompted researchers to re-examine the biomedical literature to establish whether available clinical evidence might have led to the withdrawal sooner.

The core challenge is to derive highly sensitive and selective algorithms that can resolve conflicting evidence and suppress background noises. Complex network analysis and link analysis are expected to continue to play an important role in this direction. Because of the exploratory and decision-making nature of such tasks, users need to freely interact with raw data as well as its visualizations to find causality. Techniques such as multiple coordinated views will enhance the discovery process. Features that facilitate users in finding what-ifs and test their hypotheses should be provided.

I grouped security and privacy issues under this category because both can be seen as requirements and constraints imposed on the amount of sufficient and necessary input to a decision-making process. Research in privacy-preserving data mining is a good source for information on this topic.

10. Knowledge domain visualization

The knowledge domain visualization (KDViz) challenge is a holistic driving problem. In fact, this challenge contains elements from each of the previous nine challenges. The difference between knowledge and information can be seen in terms of the role of social construction. Knowledge is the result of social construction, albeit so is information, but to a substantially lesser extent. For example, the fact that an article is about mass extinctions is information, whereas the fact that the article represents the groundbreaking work in mass extinctions research is knowledge because the value, or the status, of the article is established and endorsed by



6 **Map of the human–computer interface literature. Older topics are fading out.**

a social construction process. KDViz aims to convey information structures with such added values. Figure 6 depicts the evolution of the HCI literature. The visualization combines the intrinsic properties of articles and extrinsic, socially constructed values of citations.[13]

The greatest advantage of information visualization is its ability to show the amounts of information that are beyond the capacity of textual display. Interacting with information visualization can be more than retrieving individual items of information. The entire body of domain knowledge is subject to the rendering of KDViz. The 2004 InfoVis contest on the history of InfoVis attracted an interesting set of entries addressing some relevant issues. The KDViz problem is rich in detail, large in scale, extensive in duration, and widespread in scope. KDViz makes an attractive driving problem that, if successfully solved, can be generalized to a wide range of scientific subject areas.

Conclusion

The 10 unsolved problems reflect my personal view and the list is certainly not intended to be comprehensive. For example, I leave out specific driving problems such as human genome visualization, bioinformatics, medical informatics, the World Wide Web, digital libraries, and Web-based commerce, which all have a place in the increasingly diverse field of information visualization.

When a scientific field runs out of challenging problems, it will also run out of steam. I expect this list will provoke some thinking, stimulate some debates, and most importantly, inspire a constantly revised and updated list of top unsolved problems for the field. Information visualization has advanced tremendously in the past several years, and it will evolve into a ubiquitous subject in the future with a stimulating and challenging agenda. ■

Acknowledgments

I am grateful for the insightful comments from Pak Wong and anonymous reviewers on an earlier version of this article.

References

1. T.-M. Rhyne et al., "Can We Determine the Top Unresolved Problems of Visualization?" *Proc. IEEE Visualization*, IEEE Press, 2004, pp. 563-566.
2. C. Chen and M. Czerwinski, "Empirical Evaluation of Information Visualization: An Introduction," *Int'l J. Human–Computer Studies*, vol. 53, no. 5, 2000, pp. 631-635.
3. C. Chen, *Information Visualization: Beyond the Horizon*, 2nd ed., Springer, 2004.
4. C. Chen, "Searching for Intellectual Turning Points: Progressive Knowledge Domain Visualization," *Proc. Nat'l Academy of Sciences of the United States* (PNAS), vol. 101, suppl. 1, 2004, pp. 5303-5310.
5. W.S. Cleveland and M. Robert, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *J. Am. Statistical Assoc.*, vol. 79, no. 387, 1984, pp. 531-553.
6. A. Treisman, "Preattentive Processing in Vision," *Computer Vision, Graphics and Image Processing*, vol. 31, no. 2, 1985, pp. 156-177.
7. C. Ware, *Information Visualization: Perception for Design*, 2nd ed., Morgan Kaufmann, 2004.
8. C.M. Bishop and M.E. Tipping, "A Hierarchical Latent Variable Model for Data Visualization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, 1998, pp. 281-293.
9. G. McLachlan and D. Peel, *Finite Mixture Models*, John Wiley & Sons, 2000.
10. D. Harel and Y. Koren, "A Fast Multiscale Method for Drawing Large Graphs," *Proc. 8th Int'l Symp. Graph Drawing*, 2000, pp. 183-196.
11. P.C. Wong et al., "Dynamic Visualization of Transient Data Streams," *Proc. IEEE Symp. Information Visualization*, IEEE CS Press, 2003, pp. 97-104.
12. E.R. Tufte, *Visual Explanations*, Graphics Press, 1997.
13. C. Chen et al., "Visualizing the Evolution of HCI," *Proc. Human–Computer Interaction*, Springer, 2005.

Readers may contact Chaomei Chen at chaomei.chen@cis.drexel.edu.

Readers may contact editor Theresa-Marie Rhyne at tmrhyne@ncsu.edu.