Department: People in Practice
Editors: Daniel F. Keefe, dfk@umn.edu
Melanie Tory, mtory@tableau.com

How to Ask What to Say?: Strategies for Evaluating Natural Language Interfaces for Data Visualization

Arjun Srinivasan and John Stasko
School of Interactive Computing,
Georgia Institute of Technology

***Abstract*—In this article, we discuss challenges and strategies for evaluating natural language interfaces (NLIs) for data visualization. Through an examination of prior studies and reflecting on own experiences in evaluating visualization NLIs, we highlight benefits and considerations of three task framing strategies: Jeopardy-style facts, open-ended tasks, and target replication tasks. We hope the discussions in this article can guide future researchers working on visualization NLIs and help them avoid common challenges and pitfalls when evaluating these systems. Finally, to motivate future research, we highlight topics that call for further investigation including development of new evaluation metrics, and considering the type of natural language input (spoken versus typed), among others.**

Natural language interfaces (NLIs) for data visualization are gaining traction in both academic research[6,7,14,15,18,19] and as part of commercial tools.[1,9,20] This surge of interest has led to notable developments in terms of command interpretation and interface techniques that address challenges such as ambiguity and preserving context to support analytic conversations. That said, a persistent research

Digital Object Identifier 10.1109/MCG.2020.2986902
Date of current version 12 June 2020.

challenge in designing visualization NLIs* is evaluating the developed systems to effectively validate the systems' design and implementation while collecting actionable end-user feedback. Although visualization system evaluation has been a long standing topic of discussion (e.g., references 3, 10, 11), the evaluation of visualization NLIs presents unique challenges requiring careful consideration of evaluation strategies and procedures.

Our goal in this article is to outline benefits and considerations for different evaluation strategies to serve as a reference point for future studies. We first highlight key challenges in the evaluation of visualization NLIs. We then summarize evaluations conducted as part of prior work, wherever possible, also reflecting on our† own experiences in evaluating visualization NLIs.[13,15,16,18] Finally, we discuss topics that call for immediate consideration (e.g., defining new evaluation metrics and streamlining training procedures) to more effectively track progress and validate future visualization NLIs.

CHALLENGES

As with any visualization tool, evaluating NLIs requires addressing the general challenges of selecting appropriate evaluation metrics and methodology (e.g., comparative evaluation and qualitative study). However, besides these, experimenters of visualization NLIs constantly face two specific challenges due to the inherent nature of natural language input: *training* and *task framing*.

Training

Natural language can be used to let people perform a variety of tasks (e.g., specifying visualizations, interacting with an active visualization, performing system interface-level operations) in the context of a visualization tool.[17] As with any visualization tool, when interacting with a new visualization NLI, participants face the challenge of familiarizing themselves with the supported system functionality. Furthermore, with NLIs, participants also need to get accustomed to the type of command phrasings the system understands. However, with no prior experience of working with the tool, it is often difficult for participants to know about these features or the space of supported commands. This well-known general challenge of discoverability of NLIs (e.g., reference 21) has direct implications on the training procedure. Specifically, as experimenters, we need to think about how much information should be provided about the systems' features and interpretation capabilities during the introduction or training phase of a study.

One option to make participants familiar with the system is to tell them about the supported features and provide sample commands they can use to invoke those features. Consider the following example from our own experience: during Orko's [18] pilot studies, we initially told participants what operations (e.g., finding paths and filtering) the tool supports and gave sample commands for the same (e.g., we used the command "*Find a path between Rooney and Ronaldo*" to find the shortest path between two nodes). Although we explicitly told participants that these were only sample utterances, during the task phase, we noticed that the participants continued using the same phrasing to find paths between nodes as they thought it was the specific pattern the system understood. In other words, providing a small set of sample commands during training may give participants the false impression that the system only understands specific phrasings which may, in turn, result in participants using the same command phrasing during the task phase. This can ultimately lead to an unfair assessment of the system's capabilities as the study might never test a wider range of commands that may eventually be issued by end-users.

An alternative extreme option is to provide no formal training and directly ask participants to perform specific tasks with the system. However, this is impractical particularly in lab studies evaluating prototype systems. In particular, without appropriate in-system guidance (e.g., autocomplete while typing), participants may find it challenging to use the tool freely as they have no sense of what can be done or how. Thus, it is important to consider different training protocols and choose one that gives participants a general sense of a tool's features

*In this article, we refer to any system that allows creating or interacting with visualizations using natural language (regardless of the presence of other forms of input, such as mouse or touch) as a visualization NLI.

†Usage of the term "our" throughout this article refers to prior work that the two authors were associated with and may not reflect the opinion of other coauthors of the individual papers.

Table 1. Task framing strategies used by previous studies evaluating visualization NLIs.

	Jeopardy-style facts	Open-ended tasks	Target replication tasks
Cox *et al.*[4]		✓	
DataTone[6]	✓	✓	
Eviza[14]		✓	✓
Evizeon[7]		✓	✓
Orko[13,18]	✓	✓	
FlowSense[22]		✓	
Valletto[8]		✓	
InChorus[15]	✓		✓
DataBreeze[16]		✓	

but does not bias them toward using the tool in specific ways.

Task Framing

Perhaps the biggest challenge during the evaluation of visualization NLIs is framing the tasks employed during a user study. For instance, a common practice in evaluating visualization tools is to give participants a series of questions or operations (e.g., "Which state had the highest sales in 2019?" "Highlight countries with a population under 100M") that they need to answer/perform using the tool. However, in the context of an NLI, this may lead to participants simply parroting the task or a variation of the same as a system command. This challenge is not specific to cases when tasks are phrased as questions or operations and spans, in general, to use text of any kind as part of the study task. For instance, even if a study task was presented as a chart that participants need to replicate but was provided with accompanying instruction text (e.g., "Now update the chart so it only shows Asian countries"), participants may still tend to inherit language from the instruction text. Given the uniqueness of this challenge to visualization NLIs, we make it our primary focus in this article and discuss the strategies researchers have adopted to workaround the problem.

TASK FRAMING STRATEGIES

As mentioned earlier, framing tasks to ensure unprompted interaction is a vital challenge in evaluating visualization NLIs that researchers have tackled in different ways. In the following, we discuss the three most common task framing strategies that have been used in prior work: *Jeopardy-style facts*, *open-ended tasks*, and *target replication tasks* (summarized in Table 1).

Jeopardy-Style Facts

The Jeopardy evaluation[‡] methodology, specifically devised to evaluate visualization NLIs by Gao *et al.*,[6] involves framing tasks in the form of statements or facts from a dataset. For instance, in the context of a census dataset, an example fact could be "North Dakota has the fewest number of people without jobs."[6] Participants, in turn, are expected to prove or disprove these facts using the visualization tool, typically also needing to provide a visual justification for their responses.

Benefits

Engaging for Participants. In our experience, we have noticed that this strategy evokes a sense of gamification, leading to participants being more engaged in "solving" the task.

Measuring Success is Straightforward. Since the facts are known beforehand and tasks involve true/false responses, it is straightforward to know if a task is solved correctly. Requiring participants to visually justify their response further helps validate the response, avoiding guessing.

Facts Can Mimic Realistic Analytical Findings. Depending on the dataset and the level of exploration conducted to derive the facts, having participants validate facts can emulate serious data analysis to answer realistic questions about the dataset.

Considerations Given the aforementioned benefits, we have extensively adopted Jeopardy evaluation as a primary method for evaluating our systems. However, based on our experience, there are some important considerations to keep in mind when designing tasks and using this strategy. These include the following.

[‡]Devised based on the TV show Jeopardy!, where the contestants are presented with answers and need to phrase their responses as questions.

Dataset Familiarity. As the system's designers and experimenters, we typically work extensively with the study dataset and are well versed with the different attributes and values and their implications. However, participants may not have the same level of familiarity with the dataset and, thus, may not even know which attributes to consider to validate a fact. We encountered this challenge during both Orko and InChorus's evaluations.

For instance, one of the facts in InChorus's evaluation presented in the context of a U.S. colleges dataset was "On average, schools in large cities have the lowest admission rates." During pilots, we observed that because participants would miss that "large city" is a value under the attribute "Locale," they would use the "Population" attribute (referring to the number of students at a college) and answer incorrectly. While a common technique to overcome this challenge is providing a metadata table alongside a task or as part of the system itself, it is nonetheless a factor to consider when designing tasks and test during pilot studies.

Managing Task Difficulty Level and Phrasing. A related point to the previous one is that of managing task difficulty. If the facts are too easy to validate, the task may seem contrived and participants may lose interest. Alternatively, if the fact is too complex and demands intricate knowledge of dataset domain, participants may find it too challenging to interpret and not fully attempt a task. Thus, as experimenters, it is critical to spend substantial time in exploring the study dataset and ideally, identifying a spectrum of facts that are incrementally difficult to verify.

Besides identifying the facts, it is also vital to try out various phrasings of the facts to ensure they are not prompting or contrived yet are easy to comprehend. For instance, for facts involving multiple attributes and values (e.g., "There is only one public school in the far west with an admission rate of under 20% requiring a minimum SAT score of 1200"), it may be useful to consider alternative phrasings, perhaps involving additional elements such as tables (e.g., "There is only one public school satisfying the following criteria" + a table listing the filtering criteria).

Engagement With Tasks May Conflict With the Evaluation Goals. While participants feel engaged when solving Jeopardy-style facts, we have also observed that the gamified nature of the task (e.g., with true/false solutions) also generates some anxiety among participants. Specifically, participants tend to be more cautious with their interactions and cognizant of time even when told that there is no hard time constraint. Particularly in multimodal systems where natural language is introduced as an additional modality, this can become challenging as in lieu of completing the task rapidly, participants may resort to more familiar modalities (e.g., mouse or touch) and refrain from trying to interact through natural language altogether. Thus, it may be beneficial to include a short free-form training phase to ensure participants feel comfortable using natural language before they attempt Jeopardy-style tasks.

Open-Ended or Scenario-Based Tasks

Another task framing strategy that has been used to evaluate visualization NLIs is asking participants to conduct open-ended data exploration (e.g., "Explore this dataset and share any insights you find interesting") or premising their analysis with high-level scenarios (e.g., "Imagine you are looking to hire a new striker for your team and your club has a budget of $400M. Which players would you bid on?"). While Jeopardy-style facts can help assess the utility of the tool in the context of targeted analysis, open-ended tasks can help mimic real-world scenarios where users are generally exploring a dataset to familiarize themselves with the data and find insights.

Benefits

Relatively Straightforward to Devise Tasks. Unlike Jeopardy-style facts that require experimenters to thoroughly explore the dataset to come up with facts, creating a high-level exploration task or scenario is relatively straightforward. Although one still needs to ensure that the phrasing does not prompt direct questions/commands, the ease of task creation and the inherently practical nature of the tasks are general advantages of this strategy.

Helps Assess Overall System Features and Usability. Especially with open-ended tasks, the nature of the task can lead to participants trying out more features, possibly following a more natural workflow and experimenting with a wide range of natural language commands and phrasings. This is useful especially when the goal is to collect feedback on the overall usage experience while also investigating the role of individual features in a more global context.

Considerations While it is relatively straightforward to devise open-ended tasks, there are some considerations to keep in mind when adopting this strategy.

Challenging to Get Feedback on Specific Features. While this framing strategy can help get feedback on the overall system, if the goal is to get feedback on specific system features (e.g., use of ambiguity widgets and support for pragmatics), this strategy may not yield desired outcomes. In other words, when performing open-ended tasks, participants may either never use certain features or forget their experience of using the feature in light of other actions amid the task.

To give an example from our own experience, during Orko's studies,[13,18] we intentionally included a relatively open-ended comparison-based task among other tasks to see how well participants use the query manipulation widgets or follow-up utterances to switch between selections and compare clusters. However, during the study, participants rarely used either of those features, repeating standard one-off commands to switch between clusters. On the other hand, participants frequently leveraged a feature where the system proactively reordered summary histograms to provide context to the active points in the node-link diagram. However, when asked about their thoughts on the feature and how it helped, most participants found it challenging to tease apart the specific feature from their other actions and provide notable feedback.

Think-Aloud Protocol. One way to address the previous challenge is to employ a think-aloud protocol to collect feedback on individual features as participants work with the tool. However, given that the input modality is also natural language, the standard challenges with the think-aloud protocol such as interruption of thought processes are amplified in the context of visualization NLIs. Furthermore, in evaluating speech-based visualization NLIs, an additional challenge with using think-aloud is that the system may mistakenly try to interpret participants' comments, leading to additional errors. This was a challenge we frequently faced during pilots with DataBreeze,[16] which incorporated some implicit voice input triggering techniques (e.g., start recording voice commands when points are selected) to aid multimodal interaction. Thus, while the think-aloud protocol has been used frequently in prior work (including our own), one must try to ensure it has minimal effects on the user experience (e.g., only asking participants to think-aloud when the feel the system behaved unexpectedly).

Importance of Training. Especially if the method of open-ended tasks is the only framing strategy used during the evaluation, it is imperative that participants are well versed with the tool and different actions based on the training. If not, participants may be unsure of their actions and not use the tool freely or get frustrated if they encounter too many errors in the early stages. However, as discussed earlier, devising appropriate training procedures for visualization NLIs is a challenging task in itself. To overcome this challenge, one option may be to include an intermediate targeted analysis task set using one of the other framing strategies between the training phase and open-ended tasks. This can help participants gain more confidence in using the system and be more aware of the features and interactions before performing open-ended tasks.

Target Replication Tasks

A third strategy to frame tasks is to provide target visualization states or manipulation criteria (e.g., showing a table of attributes and values to filter by) that participants must replicate or accomplish using the given system (see Figure 1).

Benefits The inherently focused nature of these tasks provides multiple benefits, making it another commonly used task framing strategy. Some of these advantages include the following.

Minimal Risk of Phrasing Bias. This framing strategy is perhaps the least susceptible to any

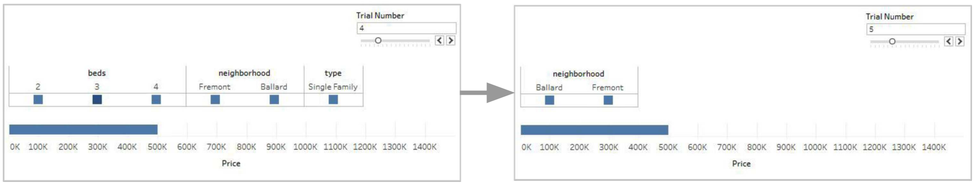

Figure 1. Examples of sequentially presented target replication tasks used during evaluations of Evizeon [7] to see how participants adjust filtering criteria.

type of phrasing bias since the instructions are provided in a graphical form. This makes it ideal to collect the most natural commands that participants would issue to create specific charts or perform given operations.

Ideal for Evaluating Low-Level Operations and Command Sequences. Since visual states can easily be broken into subsequences, this framing strategy lends itself as an ideal candidate when the goal is to evaluate a system's support for low-level operations (e.g., sorting and filtering) or follow-up commands (e.g., "*Show earthquakes in California*" > "*How about Texas?*").

Considerations While target replication is generally a low-risk and relatively low-effort strategy to implement, it also has limitations and, hence, must be cautiously applied during evaluations. The most notable drawback of this strategy is that the target state or sequences may not mimic a real-world analysis scenario. Although it helps evaluate low-level system features, the somewhat contrived nature of tasks may make it difficult for people to translate their experience of performing incremental steps during target replication tasks into actions during more complete and realistic data analysis scenarios.

Note that the aforementioned three framing strategies are not an exhaustive list, nor are the strategies mutually exclusive. For instance, although framing tasks in the form of questions without prompting participants can be challenging, it has been successfully used in prior research (e.g., reference 22). We only highlight these specific strategies due to their common adoption in prior work, allowing us to more critically reflect on their benefits and considerations.

DISCUSSION

In the following, we discuss three topics that we believe are important factors and issues to consider with respect to the evaluation of visualization NLIs going forward.

Differentiating Between Spoken and Typed Natural Language Input

In our own work,[13,15,16,18] we have largely focused on visualization NLIs where the natural language input is provided through voice. Besides interface and interaction design differences (e.g., lack of support for autocomplete in voice-based systems, additional types of errors and ambiguities caused due to speech-to-text recognition errors), we have also encountered additional challenges when conducting evaluations.

One that we alluded to earlier was that of using a think-aloud protocol. Specifically, comments provided as part of the think-aloud protocol may be interpreted as system commands if the voice recording is triggered (intentionally or otherwise) before a discourse. Another peculiar challenge involves the logistics of providing the study tasks. During pilot studies, we initially gave participants tasks on a sheet of paper. However, both with Orko [18] (running on a vertical 55" display) and InChorus [15] (running on a tablet placed on table), we noticed that participants gravitated toward using speech and would rely more on the task sheet when framing commands (often pointing their finger on the task text as they were phrasing their commands). In addition to adversely affecting the potential use of other modalities (touch in Orko or pen/touch in InChorus), such behavior may also give a false impression of the high reliance on speech.

We worked around such issues by fixing the task sheet on the screen (Orko) or using an external monitor (InChorus), but these experiences highlight considerations that arise from the differences between the nature of voice and typed input. Thus, going forward, it is important to keep such differences in mind during user studies and consider alternative approaches to overcome potential issues.

Streamlining Training Procedures

Research papers (including our prior work) often describe in-depth the system design and study details, such as tasks and results. However, the papers provide minimal details about the training procedure and the level of system detail provided during training. While this is likely unintentional, going forward, we feel it is important to streamline both the training procedure of user studies and its description in research papers. Perhaps a unique opportunity here is complement new evaluation methods, such as Jeopardy-style evaluation with new training procedures that are specific to visualization NLIs. Streamlining training procedures can ultimately ensure that findings from research are more valid (e.g., ensuring more detailed training does not bias participant behavior) and consistently derived across different systems and evaluations.

Defining Evaluation Metrics

An underlying goal of designing visualization NLIs is often to promote more fluid interaction [5] and/or improve the analytic workflow. Unfortunately, similar to subjective metrics such as engagement and enjoyment,[12] it is challenging to clearly define concepts such as "fluidity." While traditional metrics such as time and error have been used to assess the value of visualization NLIs, such metrics do not imply a direct correlation with fluidity of interaction (e.g., one may feel more engaged in a task if the system is fluid and, hence, may explore more alternatives spending more time). Thus, going forward, similar to recent developments in visualization authoring system evaluation,[2] we see the identification of concrete metrics for measuring contribution and success as an immediate area for research in the context of visualization NLIs. Eventually, these new metrics can not only help validate research progress but may also lead to the formulation of new evaluation methods and strategies that are best suited given the evaluation goals.

CONCLUSION

In this article, we highlighted key challenges in evaluating visualization NLIs including training and task framing. We describe three popular task framing strategies used while evaluating visualization NLIs—namely, *Jeopardy-style facts*, *open-ended tasks*, and *target replication tasks*. By reviewing prior studies and reflecting on our own work, we discuss the benefits and considerations to have in mind when using these different strategies. In doing so, this article aims to guide future researchers working on visualization NLIs by helping them avoid common challenges and pitfalls when evaluating these systems. Ultimately, we hope this article motivates further research not only in developing new visualization NLIs but also methods for effectively evaluating these systems and tracking research progress.

ACKNOWLEDGMENTS

The authors would like to thank Melanie Tory and Daniel F. Keefe for their feedback on this article. This work was supported in part by the National Science Foundation under Grant IIS-1717111.

REFERENCES

1. IBM Watson Analytics. 2018. [Online]. Available: http://www.ibm.com/analytics/watson-analytics/
2. F. Amini, M. Brehmer, G. Bolduan, C. Elmer, and B. Wiederkehr, "Evaluating data-driven stories and storytelling tools," in *Data-Driven Storytelling*. Natick, MA, USA: AK Peters/CRC Press, 2018, pp. 249–286.
3. S. Carpendale, "Evaluating information visualizations," in *Information Visualization*. Berlin, Germany: Springer, 2008, pp. 19–45.
4. K. Cox, R. E. Grinter, S. L. Hibino, L. J. Jagadeesan, and D. Mantilla, ,"A multi-modal natural language interface to an information visualization environment," *Int. J. Speech Technol.*, vol. 4, no. 3/4, pp. 297–314, 2001.
5. N. Elmqvist, A. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly, ,"Fluid interaction for information visualization," *Inf. Vis.*, vol. 10, no. 4, pp. 327–340, 2011.
6. T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios, "DataTone: Managing ambiguity in natural language interfaces for data visualization," in *Proc. User Interface Softw. Technol.*, ACM, 2015, pp. 489–500.

7. E. Hoque, V. Setlur, M. Tory, and I. Dykeman, "Applying pragmatics principles for interaction with visual analytics," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 309–318, Jan. 2018.
8. J.-F. Kassel and M. Rohs, "Talk to me intelligibly: Investigating an answer space to match the user's language in visual analysis," in *Proc. Des. Interactive Syst.*, ACM, 2019, pp. 1517–1529.
9. Microsoft Power BI. 2018. [Online]. Available: https://powerbi.microsoft.com/en-us/
10. C. Plaisant, "The challenge of information visualization evaluation," in *Proc. Adv. Vis. Interfaces,*. ACM, 2004, pp. 109–116.
11. D. Ren, B. Lee, M. Brehmer, and N. H. Riche, "Reflecting on the evaluation of visualization authoring systems: Position paper," in *Proc. IEEE BELIV Workshop*, VIS, 2018, pp. 86–92.
12. B. Saket, A. Endert, and J. Stasko, "Beyond usability and performance: A review of user experience-focused evaluations in visualization," in *Proc. IEEE BELIV Workshop*, VIS, 2016, pp. 133–142.
13. A. Saktheeswaran, A. Srinivasan, and J. Stasko, "Touch? Speech? or Touch and Speech? Investigating multimodal interaction for visual network exploration and analysis," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 6, pp. 2168–2179, 2020.
14. V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang, "Eviza: A natural language interface for visual analysis," in *Proc. User Interface Softw. Technol.*, ACM, 2016, pp. 365–377.
15. A. Srinivasan, B. Lee, N. H. Riche, S. M. Drucker, and K. Hinckley, "InChorus: Designing consistent multimodal interactions for data visualization on tablet devices," in *Proc. ACM CHI*, pp. 653:1–653:13, 2020.
16. A. Srinivasan, B. Lee, and J. Stasko, "Interweaving multimodal interaction with flexible unit visualizations for free-form data exploration," *IEEE Trans. Vis. Comput. Graphics*, early access, Mar. 3, 2020, doi: 10.1109/TVCG.2020.2978050.
17. A. Srinivasan and J. Stasko, "Natural language interfaces for data analysis with visualization: Considering what has and could be asked," in *Proc. EuroVis: Short Papers*. Eurographics Association, 2017, pp. 55–59.
18. A. Srinivasan and J. Stasko, "Orko: Facilitating multimodal interaction for visual exploration and analysis of networks," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 511–521, Jan. 2018.
19. Y. Sun, J. Leigh, A. Johnson, and S. Lee, "Articulate: A semi-automated model for translating natural language queries into meaningful visualizations," in *Proc. Int. Symp. Smart Graph.*, Springer, 2010, pp. 184–195.
20. Tableau Ask Data. 2018. [Online]. Available: https://www.tableau.com/about/blog/2018/10/announcing-20191-beta-96449.
21. N. Yankelovich, G.-A. Levow, and M. Marx, "Designing SpeechActs: Issues in speech user interfaces," in *Proc. Conf. Human Factors Comput. Syst.*, ACM, 1995, pp. 369–376.
22. B. Yu and C. T. Silva, "FlowSense: A natural language interface for visual data exploration within a dataflow system," *IEEE Trans. Vis. Comput. Graphics*, vol. 26, no. 1, pp. 1–11, Jan. 2020.

Arjun Srinivasan is currently working toward the Ph.D. degree in computer science with the Georgia Institute of Technology, Atlanta, GA, USA. His current research focuses on the design of intelligent and expressive visualization tools that combine multimodal input (e.g., speech and touch) and mixed-initiative interface techniques for human–data interaction. He is the corresponding author of this article. Contact him at arjun010@gatech.edu.

John Stasko is currently a Regents Professor with the School of Interactive Computing and the Director of the Information Interfaces Research Group, Georgia Institute of Technology, Atlanta, GA, USA. His research is in the areas of information visualization and visual analytics, approaching each from a human–computer interaction perspective. He was named an ACM Distinguished Scientist in 2011 and an IEEE Fellow in 2014. He received the Ph.D. degree in computer science from Brown University, Providence, RI, USA, in 1989. Contact him at stasko@cc.gatech.edu.

Contact department editor Melanie Tory at mtory@tableau.com or department editor Daniel F. Keefe at dkf@umn.edu.