Inferential Tasks as an Evaluation Technique for Visualization

A. Suh[1] and A. Mosca[2] and S. Robinson[1,3] and Q. Pham[1] and D. Cashman[4] and A. Ottley[5] and R. Chang[1]

[1]Tufts University, Medford, MA, USA
[2]Northeastern University, Boston, MA, USA
[3]EditShare, Watertown, MA, USA
[4]Novartis Pharmaceuticals, Cambridge, MA, USA
[5]Washington University in St. Louis, St. Louis, MO, USA

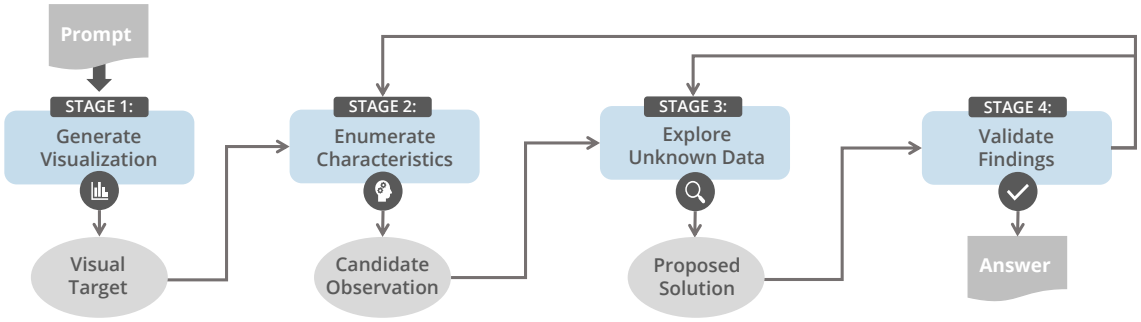


Figure 1: *The four distinct stages of evaluating interactive visualizations with an inferential task. Inferential tasks are based on the concept of inferential learning, a process requiring users to rely on their problem-solving and reasoning abilities to draw conclusions that are not explicitly prompted. In our proposed framework, users are given a prompt with the following instructions: build a specific visualization, then find a different subset of data that exhibits similar characteristics when visualized. We describe each of these stages in Section 2.*

Abstract

Designing suitable tasks for visualization evaluation remains challenging. Traditional evaluation techniques commonly rely on 'low-level' or 'open-ended' tasks to assess the efficacy of a proposed visualization, however, nontrivial trade-offs exist between the two. Low-level tasks allow for robust quantitative evaluations, but are not indicative of the complex usage of a visualization. Open-ended tasks, while excellent for insight-based evaluations, are typically unstructured and require time-consuming interviews. Bridging this gap, we propose inferential tasks: a complementary task category based on inferential learning in psychology. Inferential tasks produce quantitative evaluation data in which users are prompted to form and validate their own findings with a visualization. We demonstrate the use of inferential tasks through a validation experiment on two well-known visualization tools.

CCS Concepts

• ***Human-centered computing*** → ***Information visualization; Visualization design and evaluation methods;***

1. Introduction

When evaluating interactive visualization systems, numerous approaches have been suggested over the years to ascertain both the benefits and limitations of a proposed tool [CY00]. Central to these approaches is the use of *tasks* that users are asked to perform while using a visualization [BM13, DFP*18]. In 2005, Amar et al. presented ten low-level analytic tasks [AES05] that remain commonly used in the evaluation of visualizations today [IIC*13] (e.g., [SED18, RQ20]). However, the use of low-level tasks in visualization evaluation poses a variety of challenges, such as the task's perceived lack of complexity, as well as the task's inability to capture a visualization's insight capabilities [Kos16, PSB20].

North proposed the elimination of simple tasks altogether in experimental studies, instead suggesting "complex benchmark tasks" and insight-driven, "open-ended protocols" that more realistically assess the efficacy of visualizations [Nor06]. While open-ended tasks produce rich qualitative feedback [PAEE19, MD19], conducting and analyzing responses from think-aloud tasks through interviewing and open-coding is a time-consuming endeavor that is not always feasible or scalable for the designer [And10, TLCC17]. As a result,

the ability to evaluate a visualization's insight capabilities both realistically and quantitatively has continued to be of interest to the research community [BAB*18, BXF*20].

In this paper, we formalize a new class of complex benchmark tasks that serve to complement open-ended protocols: *Inferential tasks*. Inspired by the concept of inferential learning from psychology [See12], inferential tasks require evaluation participants to construct knowledge by inferring relations between learned concepts and new observations. Moreover, inferential tasks can be set up with clear 'correct' or 'incorrect' answers, resulting in quantitative evaluation data that can be analyzed with the same statistical methods as low-level task evaluations. An example of an inferential task for visualization evaluation is shown in Figure 2.

Our motivation for proposing tasks requiring inferential learning is their success in prior visualization literature [GJF10, ZCY*11]. Green et al. and Ziemkiewicz et al. both showed that tasks that involve inferential learning produce more nuanced and informative evaluation data than tasks that only involve *procedural learning* (i.e., traditional low-level tasks). Though the authors share these findings in their work, they do not offer guidelines for designing nor deploying inferential tasks in visualization evaluations.

We build on this previous work by defining a methodology for inferential tasks in visualization evaluation. We then demonstrate how our framework for inferential tasks can be used in practice in a validation experiment comparing two well-known exploratory visualization tools, Voyager 2 [WQM*17] and Polestar [WMA*16]. Our results indicate that the use of inferential tasks produce evaluation data that illuminates differences between the tools, while remaining straightforward to analyze quantitatively. Finally, we discuss design considerations, limitations, and future work.

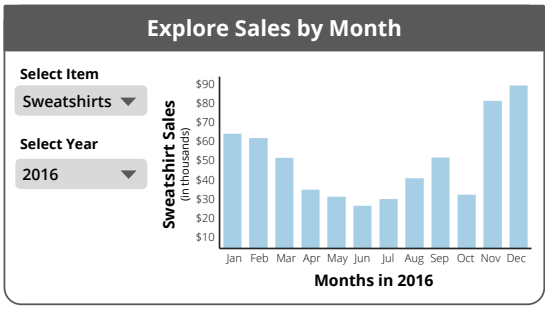


Figure 2: *Illustrative visualization tool that supports users in the exploration of monthly item sales for a specified year. This tool can be used to solve the inferential task: "Plot sweatshirt sales by month for the year 2016. Observe how monthly sweatshirt sales are affected in 2016. Find another item, besides sweatshirts, whose monthly sales exhibit a similar relationship in the year 2016."*

2. Formalization of Inferential Tasks

We formalize Green et al.'s and Ziemkiewicz et al.'s previous use of inferential tasks as a procedure consisting of four stages. Figure 1 illustrates this process. Although the authors do not describe how to construct inferential tasks for visualization evaluation, Green et al. do provide an example of the task set-up: an 'exemplar' is first shown to participants who are then asked to "*find another example that shares/does not share a variety of characteristics.*" In this section, we outline the four stages that make up an inferential task and define terminology to aid researchers in deploying inferential tasks in practice. We generalize inferential tasks for interactive visualization tools; however, researchers can modify specifics at each stage if the experiment goal is to evaluate one or more static visualization(s).

2.1. Stage 1: Generate Visualization

Visual Target: A specific visualization participants are instructed to generate (if using an interactive system) and observe.

At the start of an inferential task, participants are instructed to construct and inspect a *specific visualization* – the **visual target** for the task. For example, if using the tool in Figure 2 to solve the task "*Plot sweatshirt sales for months in 2016. Observe an interesting pattern in the chart. Find another item and year for which sales follow a similar pattern,*" participants will generate a visualization with sweatshirts selected as Item and 2016 selected as Year. The chart shown in Figure 2 is the visual target for the above task.

The purpose of having participants generate a visual target is to test the usability of a visualization tool and the design of a visual encoding. This stage is crucial when there are multiple visual encoding options for generating the visualization. When appropriate, researchers may specify the visual encoding that participants should generate for the visual target in the task prompt (e.g., build a pie chart for one task and a bar chart in another). If the participant is not working with an interactive system, the researcher can supply the visual target instead. In this case, the task begins at Stage 2.

2.2. Stage 2: Enumerate Characteristics

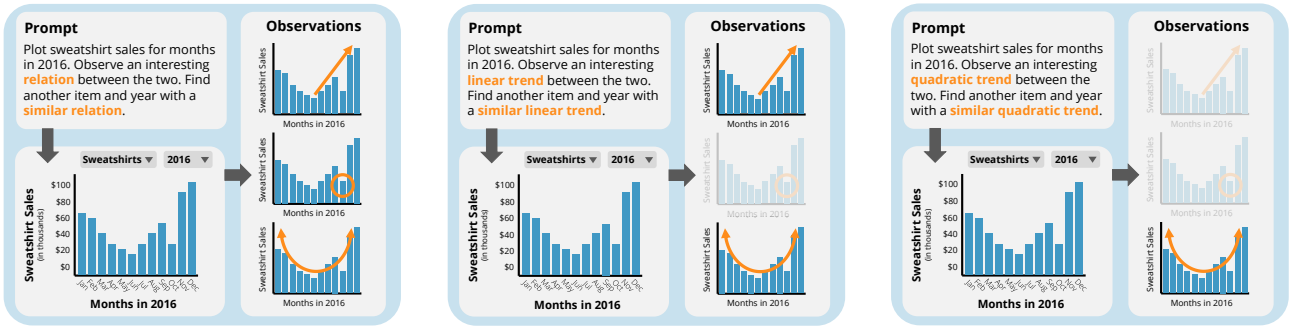
Candidate Observation: A proposed explanation or characterization of the data that is deduced from the visual target.

Once participants have constructed (or been given) the visual target, they are asked to find another example of a visualization that displays *similar or dissimilar characteristics*. This process requires participants to identify and enumerate all plausible patterns or relations shown in the visual target that could be found elsewhere in the data. We call the particular characteristic, pattern, or relation that participants discern from the visual target their **candidate observation**. To illustrate how a participant arrives at their candidate observation, consider Figure 3a. A participant may pose the following observations for relations in monthly sweatshirt sales:

- Sales spike at the end of the year (i.e., last two bars are highest)
- October is an outlier in the fall (i.e., one bar seems out of place)
- Sales decrease then increase (i.e., the bars go down then up)

Ultimately, the participant will choose a particular characterization of the visual target that they think best exemplifies a relation in monthly sweatshirt sales. During this process, participants are implicitly performing a series of analytic tasks (e.g., "correlate", "find anomalies" [AES05]) to identify their candidate observation.

Researchers can adjust the specificity of the task prompt to broaden or restrict the analytic tasks a user performs, and thus



(a) Many observations due to a vague prompt **(b)** Fewer observations due to a less vague prompt **(c)** Least observations due to a specific prompt

Figure 3: *Demonstration of tuning the complexity of an inferential task by modifying the specificity of the prompt. (a) shows a vague prompt instructing participants to observe an interesting "relation" in the chart. The open-ended nature of this task results in participants reasoning through many potential characteristics in the visualization. (b) is a less vague prompt instructing participants to observe an interesting "linear trend", supplying a more specific relation to identify in the visualization. (c) is the most explicit prompt, instructing participants to observe an interesting "quadratic trend." The specificity of this prompt gives participants the most evidence for their target of inference.*

the number of observations a participant makes with the visual target. Figure 3 demonstrates this mechanism. We note that an overly ambiguous task may result in HARKing [Ker98] (i.e., identifying a candidate observation *after* completing the task) or the multiple-comparison problem in visual analysis [ZZZK18]. To lower this risk, participants can be asked to report their candidate observation before continuing with the task (e.g., *experimental preregistration* [CGD18]), or, the researcher could provide predefined multiple choice candidate solutions for participants to choose from.

2.3. Stage 3: Explore Unknown Data 🔍

Proposed Solution: A visualization, showing different data attributes than the visual target, that the participant believes to exhibit the same characteristics as the candidate observation.

After identifying a candidate observation, participants explore the remaining data to find a different subset of data that, when visualized, exhibits those characteristics. As part of this process, participants generate and/or observe visualizations to reason about previously unknown data. Each of these visualizations is a **proposed solution** – or potential answer – to the inferential task.

For example, take again the inferential task prompt and visualization tool shown in Figure 3a. Suppose that a participant performing this task inspects the visual target and forms the observation: "Sales spike at the end of the year." The next step will be to search all possible combinations of Items and Years until finding another instance for which sales spike at the end of the year. During this search, participants will construct new visualizations with different Items and Years than those given in the prompt, i.e., proposed solutions.

When instructing participants to "*find another example*", researchers can specify in the prompt which data should be explored for an answer to the task. Modifying the breadth of data to search in an inferential task helps evaluate a visualization's scalability. This includes the evaluation of the visualization in helping a participant navigate through large and high-dimensional spaces, as well as testing the visualization's ability to support a participant in reasoning about relationships between many attributes at a time.

2.4. Stage 4: Validate Findings ✔

Answer: The final visualization or solution, as validated by the participant, that is believed to correctly exhibit similar characteristics as the candidate observation and visual target.

At this stage of an inferential task, the participant has a proposed solution in mind that needs to be validated for correctness as a potential **answer** to the task. The validation process requires the participant to compare their proposed solution to the original visual target. This process results in three possible outcomes, illustrated as the three outgoing arrows from Stage 4 in Figure 1.

- The participant finds the proposed solution to be satisfactory. In this case, the participant has found a particular visualization that exhibits similar characteristics as the visual target.
- The participant is not satisfied with the current proposed solution, but believes their candidate observation is still correct. In this case, the participant will continue to explore the remaining data to find a visualization that better fits their candidate observation.
- The participant is not satisfied with the proposed solution, and believes that their candidate observation is incorrect. In this case, the participant will return to the second stage to identify a new candidate observation.

The process of validating a candidate observation is crucial to the evaluation of a visualization with inferential tasks, as it ensures the visualization tool (or static graphic) is capable of supporting new insights through exploratory and/or confirmatory analysis. We discuss the practicality of assessing participants' answers in Section 4.

3. Demonstration and Validation

To demonstrate how our framework for inferential tasks can be used for evaluating visualizations in practice, we present a crowdsourced study comparing the performance of two well-known and open-source visualization tools, Polestar [WMA*16] and Voyager 2 [WQM*17], using an inferential task-based evaluation.

Voyager 2 is a visual analytics tool that is designed to both manually and automatically support analysts through open-ended and

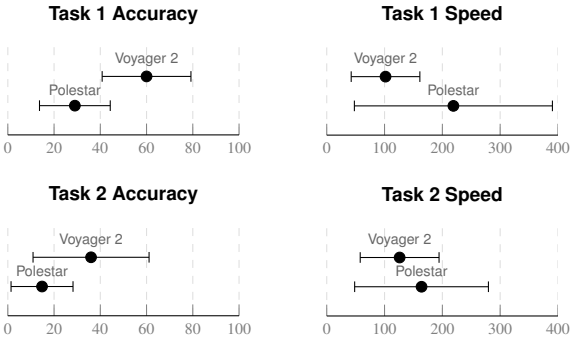


Figure 4: *Voyager 2 vs. Polestar experiment results on both tasks. Mean accuracy and 95% CIs are shown on the left. Mean speed (s) and standard deviation is shown on the right.*

focused exploration. Details are available at https://github.com/vega/voyager. In the original user study for Voyager 2 [WQM*17], the authors conducted a think-aloud protocol comparing against Polestar - a visualization tool similar to Tableau [STH02]. Unlike Voyager 2, Polestar does not provide recommendations to a user. We posit that participants will perform better on inferential tasks using Voyager 2 due to its recommendation engine.

Tasks: In our user study, each inferential task is structured with our formalism in mind: participants are asked to plot two specific data attributes (Generate Visualization), observe a relation in the chart (Enumerate Characteristics), and find one or two other attributes (Explore Unknown Data) that exhibit a similar relationship when visualized (Validate Findings). Participants completed their tasks with the *movies* dataset using either Voyager 2 or Polestar.

> **Task 1:** Plot IMDB Votes on the x-axis and IMDB Rating on the y-axis. Observe the relationship of these variables. Find another variable that shows a similar relationship with IMDB Rating.
>
> **Task 2:** Plot US Gross on the x-axis and Worldwide Gross on the y-axis. Observe the relationship of these variables. Find two different variables that show a similar relationship.

For Task 1, participants are asked to explore up to 6 combinations of attributes on the x-axis (IMDB Rating fixed on y-axis), while Task 2 asks participants to explore up to 30 combinations of attributes on the x- and y-axis. We determined ground truth by identifying data attribute(s) that display a clearly similar relationship (logarithmic for Task 1, positive linear for Task 2) to the original visualization, i.e., visual target. Task 1 had two possible correct answers, while Task 2 had only one. Examples of correct and incorrect answers are provided in the supplemental. Accuracy was recorded as a binary 'correct' or 'incorrect', and speed was recorded as the total time spent between starting the task and submitting an answer.

Results & Takeaways: Our quantitative results are summarized in Figure 4, and a write-up comparing our results to those of the original study is included in the supplemental. Overall, we find that participants are more accurate with Voyager 2 than Polestar for Task 1 and faster with Voyager 2 for both Task 1 and Task 2. Our findings suggest that the recommendations of Voyager 2 are of high quality and assist participants in navigating through data efficiently.

Our results also highlight two potential limitations of the approach. First, the overall accuracy is low, particularly for Task 2. This suggests that asking participants to explore many combinations of attributes (6 in Task 1 versus 30 in Task 2) could result in poorer accuracy. Second, because of the nature of our crowdsourced study, we cannot know precisely *when* participants failed in their tasks. Fine-tuning the complexity of the experiment (e.g., by providing multiple choice answers or predefined candidate observations) and analyzing interaction logs could reduce this ambiguity, thereby providing additional context to the researcher. We discuss these trade-offs and avenues for future work further in Section 4.

4. Discussion, Limitations, and Future Work

Although prior work highlights the benefits of using inferential tasks in evaluating visualizations [GJF10, ZCY*11], they are not without limitations and shortcomings. For example, finding the balance between an open-ended versus narrow prompt can be difficult. On one hand, an open-ended prompt (e.g., Figure 3a) necessitates participants in exploring more of the data and tool being evaluated. However, as demonstrated in Section 3, participants' accuracy can suffer when searching many possible attributes for a correct answer. An open-ended prompt also requires researchers to identify all possible characteristics that a participant could possibly (and correctly) observe. A narrow prompt (e.g., Figure 3c) can be used to reduce the complexity of the task as well as limit the number of possible correct answers. Subsequently, these tasks are less indicative of the practical use a visualization tool and closer to low-level tasks [AES05]. Balancing inferential task complexity with feasibility needs to be carefully considered and studied given an experimental goal.

As such, inferential tasks are not intended to be a replacement for open-ended *nor* low-level tasks. Instead, they should be thought of as a complementary evaluation technique that can serve as a midpoint between the two. Future work is needed to better understand when an inferential task evaluation is most appropriate. In some cases, such as a strict usability study, low-level tasks are sufficient. When interested in testing how well a visualization supports new insights, the point at which an inferential task evaluation becomes as useful as an open-ended evaluation can be unclear. We leave to future work investigating the precise benefits and trade-offs of inferential tasks as an evaluation technique for visualization.

5. Conclusion

In conclusion, we formalize the use of inferential tasks as a way to build on complex benchmark tasks – creating evaluations that can be quantitatively analyzed, yet engage participants in a pattern of analysis closer to open-ended tasks for insight-based evaluations. We present a crowdsourced study to demonstrate the use of inferential tasks in practice with two interactive visualization tools. Our results suggest that our framework for inferential tasks can be successfully deployed to illuminate differences between visualization systems.

Acknowledgments

This work was supported by National Science Foundation grants IIS1452977, OAC-1940175, OAC-1939945, DGE-1855886, OAC-2118201, NRT-2021874, and DOD grants HQ0860-20-C-7137, N68335-17-C-0656. We thank the reviewers for their feedback.

References

[AES05] AMAR R., EAGAN J., STASKO J.: Low-level components of analytic activity in information visualization. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (2005), IEEE, pp. 111–117. 1, 2, 4

[And10] ANDERSON C.: Presenting and evaluating qualitative research. *American journal of pharmaceutical education 74*, 8 (2010). 1

[BAB*18] BATTLE L., ANGELINI M., BINNIG C., CATARCI T., EICHMANN P., FEKETE J.-D., SANTUCCI G., SEDLMAIR M., WILLETT W.: Evaluating visual data analysis systems: A discussion report. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics* (2018), pp. 1–6. 2

[BM13] BREHMER M., MUNZNER T.: A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2376–2385. 1

[BXF*20] BURNS A., XIONG C., FRANCONERI S., CAIRO A., MAHYAR N.: How to evaluate data visualizations across different levels of understanding. In *2020 IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization (BELIV)* (2020), IEEE, pp. 19–28. 2

[CGD18] COCKBURN A., GUTWIN C., DIX A.: *HARK No More: On the Preregistration of CHI Experiments*. Association for Computing Machinery, New York, NY, USA, 2018, p. 1–12. URL: https://doi.org/10.1145/3173574.3173715. 3

[CY00] CHEN C., YU Y.: Empirical studies of information visualization: a meta-analysis. *International Journal of Human-Computer Studies 53*, 5 (2000), 851–866. 1

[DFP*18] DIMARA E., FRANCONERI S., PLAISANT C., BEZERIANOS A., DRAGICEVIC P.: A task-based taxonomy of cognitive biases for information visualization. *IEEE transactions on visualization and computer graphics 26*, 2 (2018), 1413–1432. 1

[GJF10] GREEN T. M., JEONG D. H., FISHER B.: Using personality factors to predict interface learning performance. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on* (2010), IEEE, pp. 1–10. 2, 4

[IIC*13] ISENBERG T., ISENBERG P., CHEN J., SEDLMAIR M., MÖLLER T.: A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 2818–2827. 1

[Ker98] KERR N. L.: Harking: Hypothesizing after the results are known. *Personality and social psychology review 2*, 3 (1998), 196–217. 3

[Kos16] KOSARA R.: An empire built on sand: Reexamining what we think we know about visualization. In *Proceedings of the sixth workshop on beyond time and errors on novel evaluation methods for visualization* (2016), pp. 162–168. 1

[MD19] MEYER M., DYKES J.: Criteria for rigor in visualization design study. *CoRR abs/1907.08495* (2019). URL: http://arxiv.org/abs/1907.08495, arXiv:1907.08495. 1

[Nor06] NORTH C.: Toward measuring visualization insight. *IEEE computer graphics and applications 26*, 3 (2006), 6–9. 1

[PAEE19] PECK E. M., AYUSO S. E., EL-ETR O.: Data is personal: Attitudes and perceptions of data visualization in rural pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI '19, Association for Computing Machinery, p. 1–12. URL: https://doi.org/10.1145/3290605.3300474, doi:10.1145/3290605.3300474. 1

[PSB20] PANDEY A., SYEDA U. H., BORKIN M. A.: Towards identification and mitigation of task-based challenges in comparative visualization studies. In *2020 IEEE Workshop on Evaluation and Beyond-Methodological Approaches to Visualization (BELIV)* (2020), IEEE, pp. 55–64. 1

[RQ20] ROSEN P., QUADRI G. J.: Linesmooth: An analytical framework for evaluating the effectiveness of smoothing techniques on line charts. *IEEE Transactions on Visualization and Computer Graphics* (2020). 1

[SED18] SAKET B., ENDERT A., DEMIRALP Ç.: Task-based effectiveness of basic visualizations. *IEEE transactions on visualization and computer graphics 25*, 7 (2018), 2505–2512. 1

[See12] SEEL N. M.: Inferential learning and reasoning. In *Encyclopedia of the Sciences of Learning*. Springer, 2012, pp. 1550–1555. 2

[STH02] STOLTE C., TANG D., HANRAHAN P.: Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics 8*, 1 (2002), 52–65. 4

[TLCC17] THUDT A., LEE B., CHOE E. K., CARPENDALE S.: Expanding research methods for a realistic understanding of personal visualization. *IEEE computer graphics and applications 37*, 2 (2017), 12–18. 1

[WMA*16] WONGSUPHASAWAT K., MORITZ D., ANAND A., MACKINLAY J., HOWE B., HEER J.: Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)* (2016). URL: http://idl.cs.washington.edu/papers/voyager. 2, 3

[WQM*17] WONGSUPHASAWAT K., QU Z., MORITZ D., CHANG R., OUK F., ANAND A., MACKINLAY J., HOWE B., HEER J.: Voyager 2: Augmenting visual analysis with partial view specifications. In *ACM Human Factors in Computing Systems (CHI)* (2017). URL: http://idl.cs.washington.edu/papers/voyager2. 2, 3, 4

[ZCY*11] ZIEMKIEWICZ C., CROUSER R. J., YAUILLA A. R., SU S. L., RIBARSKY W., CHANG R.: How locus of control influences compatibility with visualization style. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)* (2011), IEEE, pp. 81–90. 2, 4

[ZZZK18] ZGRAGGEN E., ZHAO Z., ZELEZNIK R., KRASKA T.: Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–12. 3