Reflecting on the Evaluation of Visualization Authoring Systems

Position Paper

Donghao Ren[*]
University of California, Santa Barbara

Bongshin Lee[†]
Microsoft Research

Matthew Brehmer[‡]
Microsoft Research

Nathalie Henry Riche[§]
Microsoft Research

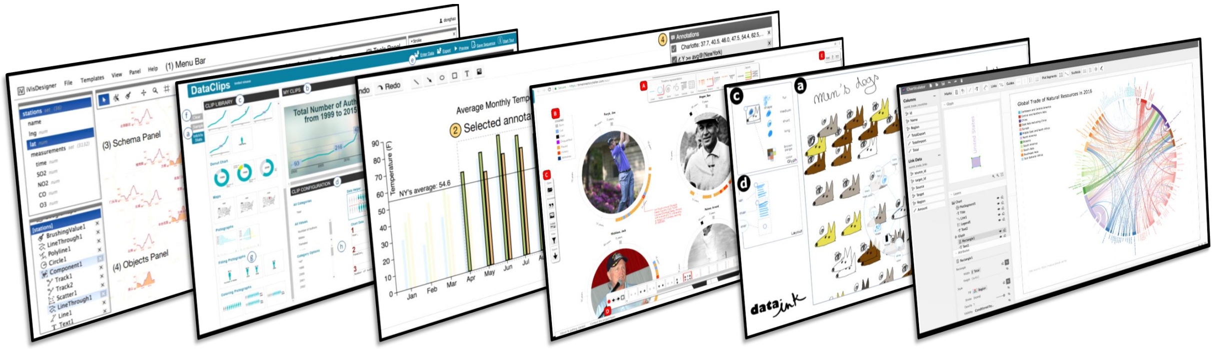

Figure 1: A selection of visualization authoring systems developed by one or more of the authors. From left to right: iVisDesigner [29], DataClips [4], ChartAccent [28], Timeline Storyteller [24], DataInk [45], and Charticulator [30].

Abstract

In this paper, we discuss the challenges one faces when evaluating authoring systems developed to help people design visualization for communication purposes. We reflect on our own experiences in evaluating the visualization authoring systems that we have developed as well as the evaluation methods used in other recent projects. We also examine alternative approaches for evaluating visualization authoring systems that we believe to be more appropriate than traditional comparative studies. We hope that our discussion is informative, not only for researchers who intend to develop novel visualization authoring systems, but also for reviewers assigned to evaluate the research contributions of these systems. Our discussion concludes with opportunities for facilitating the evaluation and adoption of deployed visualization authoring systems.

Index Terms: Human-centered computing—Visualization—Visualization design and evaluation methods

1 Introduction

With an increasing demand for data-driven storytelling, we are witnessing a proliferation of visualization authoring systems [13, 16, 21, 33]. Although these systems pursue a similar goal (i.e., enabling people to easily visualize data), it is seldom straightforward to understand and assess a novel authoring system's strengths and weaknesses relative to other systems. In this paper, we discuss more appropriate ways to evaluate visualization authoring systems.

While the difficulties and challenges in evaluating information visualization systems have been discussed at length in the visualization research community [10, 17, 26, 32], the evaluation of visualization authoring systems presents additional unique challenges, calling for alternative evaluation approaches.

Research papers describing visualization authoring systems may offer different research contributions: they may propose a new approach to visualization authoring, target a new group of prospective visualization authors, enable the creation of custom visualization designs, or any combination of these. Typical evaluation metrics employed by visualization researchers may not be suitable for evaluating these authoring systems. For instance, while efficiency and usability may be relevant in cases where the aim of the system is to improve an existing authoring workflow, these metrics may be irrelevant if the system's goal is to offer new levels of expressiveness to visualization authors.

It is challenging but critical to clearly define the specific contributions of a novel visualization authoring system, as these will guide the evaluation criteria and the approach to assessing the results of these evaluations on the part of reviewers. Being explicit about this evaluation rationale is critical for dispelling unrealistic expectations that the system should "do it all": be powerful yet easy to learn, suit novice and expert audiences alike, and outperform other authoring systems in terms of expressiveness, usability, and efficiency. In this paper, we reflect on our own experience in evaluating the visualization authoring systems that we have developed in recent years to facilitate visualization designs for communication purposes. Our intent is to provide insights to both researchers and reviewers with regards to different types of contributions, evaluation criteria, and evaluation methodologies. We conclude by discussing opportunities for the visualization community to better facilitate the evaluation of visualization authoring systems.

2 Challenges

Amini et al. [2] recently identified seven criteria relating to the evaluation of tools for authoring data-driven stories. We build upon and tailor this list of criteria to visualization authoring systems, integrating insights from our own experience. Note that this list is not exhaustive and may grow as more visualization authoring systems emerge in the future.

- *Expressiveness:* The scope of possible visualization design choices enabled by the system.
- *Creativity support:* The extent to which a system aids the author in creating novel visualization designs, such as easy manipulation of graphical elements or easy specification of element layouts.

[*] e-mail: donghaoren@cs.ucsb.edu
[†] e-mail: bongshin@microsoft.com
[‡] e-mail: mabrehme@microsoft.com
[§] e-mail: nath@microsoft.com

- *Flexibility:* The number of ways an author can achieve a desired visualization design.
- *Guidance:* The extent to which an author can produce a visualization without external assistance.
- *Efficiency:* How quickly a desired visualization design can be produced using the interface, how many actions are required to produce the desired design, or how many visualization design choices can be made in a set amount of time.
- *Usability:* How easily a desired visualization can be produced using the system.
- *Learnability:* The ease of learning and recalling interactions within the system after initial guidance.
- *Integration:* The extent to which the system fits into an authors' workflow. This may include supporting specific sequences of tasks, bridging to existing tools, or supporting collaboration.

Traditional controlled experiments are useful to compare efficiency (e.g., task time, error rate). However, they are not always appropriate for evaluating other criteria, particularly in the context of visualization authoring systems, where it is rare to find an appropriate baseline to compare a new system against. Researchers often design and develop systems because existing systems are not designed to support desired capabilities. It is also unrealistic to have a study session of sufficient length such that participants learn the full capabilities of a visualization authoring system. Consider, for instance, the number of features in commercial software tools such as Adobe Illustrator, which many people use during the visualization authoring process. Therefore, it is difficult for researchers to control important factors or to select tasks without penalizing one of the systems or compromising the external validity of the study.

Suitable evaluation methods should be selected based on the evaluation criteria, which in turn should depend upon the research contribution that the researchers intend to make. For example, if the contribution is a system that allows authors to create a wider variety of visualization designs (e.g., Lyra [35]), *expressiveness* may be the primary evaluation criterion. If the system's contribution is bridging the strengths of multiple authoring tools (e.g., Hanpuku [5]), the evaluation should be centered around assessing the degree of *integration*. When an authoring system enables people to produce artifacts that may be novel or understudied, independent studies on the evaluation of the artifacts may be necessary. Note that the researchers who develop such systems may address several of these criteria, and sometimes they need to strike a balance between them. For example, if a system targets novices (e.g., ManyEyes [43]), it may provide a high level of *guidance* at the expense of *flexibility* (requiring only a limited number of simple steps to author a visualization) and *expressiveness* (providing a constrained set of visualization design choices). Finally, we must also consider the *usability* and *learnability* for the target audience. These criteria are easier to satisfy when targeting a narrow audience such as professional designers or graphic journalists, as opposed to targeting the general public.

Several evaluation methods lend themselves to a single criterion. For example, a gallery of visualization designs would be appropriate for demonstrating *expressiveness*, while a traditional usability study would be suitable for validating the usability of the authoring system. However, it is not always trivial or apparent to pick appropriate methods for certain criteria. In such cases, the researchers should justify their choice of evaluation methods.

3 Reflection on Existing Approaches

The authors of this paper have collectively designed, developed, and evaluated a set of visualization authoring systems (see Figure 1). In this section, we reflect upon the experience of evaluating these systems. We also discuss methods used to evaluate other recent visualization authoring systems whose main goal was to support "expressive" visualization design. Table 1 summarizes the evaluation methods used for each of the systems. We identify the strengths and weaknesses of each evaluation method in the hope that researchers can choose appropriate methods to evaluate their own systems.

3.1 Formative Study

Brehmer et al. [6] encourage visualization researchers and practitioners to conduct pre-design empirical studies, as they can inform system design through the characterization of work practices and associated problems. In developing Data Illustrator [18], Liu et al. held weekly meetings with three designers over the span of two years as a means to better understand how visualization could be described and produced from a graphic design perspective. Similarly, Schroeder and Keefe sought feedback from professors and students (i.e., the system's target audience) over the course of two years through short sessions of demonstrations and discussions, which helped them shape the interface design [36]. We note that these are exemplary instances of a formative or pre-design study.

Another way to establish the desired expressiveness of a novel authoring system is to review a collection of existing artifacts. For example, in designing DataClips [4], an authoring system for data videos, Amini et al. first examined 50 data videos collected from various media sources online in an effort to tease apart the design dimensions used to produce compelling narratives in film or cinematography [3]. Similarly, Ren et al. surveyed 106 existing annotated charts published by several news media outlets to generate a design space of annotations, which subsequently informed the design of ChartAccent [28]. Finally, Brehmer et al. designed Timeline Storyteller [24] based on a recently proposed design space grounded in an extensive survey [8].

A formative study can be used to justify an authoring system's design, such as the underlying visualization design framework or its interaction mechanisms, how the interface will integrate with existing work practices and tools, how skilled users of existing systems may transfer their skill to the new system, and whether and how multiple people collaborate during the visualization design process. For example, the partition and repetition actions in Data Illustrator [18] were attributed to findings from their formative study and observations of how designers use existing tools such as Adobe Experience Design. A formative study can also inform design trade-offs, such as between expressiveness and interface complexity; for example, ChartAccent [28] prioritizes the design of the most common forms of annotation, while some of the more esoteric forms of annotation found during the formative survey are not supported.

3.2 Reproduction Study

The focus of a reproduction study is the usability of a system and the learnability of its features. The evaluation of Lyra [35], ChartAccent [28], VisComposer [20], DataInk [45], Data Illustrator [18], and Charticulator [30] each featured a *reproduction study*, in which study participants were asked to reproduce a copy of one or more visualization designs.

A reproduction study typically employs a think-aloud approach, which helps researchers elicit the subjective impressions of participants as they use the system for visualization design tasks. Visualization design completion time can only serve as a proxy of efficiency, as participants are discovering the interface for the first time and explaining their thought process as per the think-aloud protocol; both of which can inflate the time to reproduce a design. As the study participants show different think-aloud behavior, the variance in time measurement is likely to be substantial: some participants provide more detailed explanations than others. However, researchers can still assess if participants can produce a desired visualization design, and if not, what the main barriers might be. The think-aloud protocol can also be complemented with experimenter observations, screen capture video, and post-study questionnaires.

Table 1: Evaluation methods used for each of the visualization authoring systems. 🖼: gallery of visualizations; 🎥: video showing the creation process. [1]These surveys of existing artifacts helped us extract a design space. [2]iVoLVER was evaluated against Tableau in a follow-up study [22].

	Formative Study	Reproduction Study	Free-Form Study	Comparative Study	Gallery
ChartAccent [28]	survey[1]	✓	-	-	-
Charticulator [30]	-	✓	-	interaction complexity	🖼🎥
DataClips [4]	survey[1]	-	✓	video design	-
DataInk [45]	-	✓	✓	-	-
iVisDesigner [29]	-	-	informal	-	🖼
Timeline Storyteller [24]	survey [8][1]	-	-	-	🖼
Data Illustrator [18]	design meetings	✓	-	-	🖼🎥
Ellipsis [34]	interviews	-	✓	-	🖼
Hanpuku [5]	-	-	-	-	three examples
Visualization-by-Sketching [36]	short sessions of demos and discussions	-	-	-	five examples
InfoNice [44]	-	-	✓	time	🖼
iVoLVER [23]	-	-	-	follow-up research [22][2]	🖼🎥
Lyra [35]	-	✓	-	-	🖼
VisComposer [20]	-	✓	-	time	🖼

Indirectly, a reproduction study may also shed light on whether participants understand the interaction or design framework embodied by the authoring system. For instance, Charticulator [30] incorporates a novel constraint-based chart layout design framework that separates mark and glyph construction from chart layout. By observing and listening to participants think aloud as they reproduce a set of designs with Charticulator, researchers could infer whether (and when) they understood the effects of layout constraints based on their interactions and utterances.

It should be noted that, in practice, reproduction study sessions are brief (e.g., 60–90 minutes), and thus participants may learn and use only a subset of the system's features or capabilities. Similarly, any tutorial they receive prior to using the system may only address a subset of features; ideally this subset includes those which underlie the research contributions of the system.

Reproduction studies of research prototypes are likely to be hindered by minor but common usability issues. For instance, in Data Illustrator's reproduction study [18], participants remarked upon the lack of undo/redo functionality; similarly, participants commented on the lack of z-order manipulation in ChartAccent [28]. Such usability issues are often tangential to the system's research contributions, but they can impact its usability and by extension the quality of results collected during a reproduction study. Researchers should be aware of the considerable amount of effort required to address such issues so as to conduct a successful reproduction study.

Arguably, a reproduction study will not surface all of an authoring system's usability issues, and thus additional measures of usability should be undertaken should the researchers intend to deploy a usable tool. Unfortunately, findings from additional usability studies are unlikely to form substantial visualization research contributions.

3.3 Free-Form Study

In a free-form study, participants are asked to create their own visualization designs using the system. This method was used in the evaluation of Ellipsis [34], DataClips [4], DataInk [45], and InfoNice [44]. Such studies focus on assessing if participants can create visualizations of their own imagining using the new authoring system. Thus, they may include a design phase during which participants are asked to envision a visualization design before trying to realize it using the authoring system. For example, Amini et al. [4] included an *idea generation and sketching* phase before the creation phase. During the creation process, the experimenter may aid the participants by providing reference materials, such as a cheat sheet. In addition, a free-form study can be preceded by a reproduction study (e.g., DataInk [45]), as it can serve as an additional tutorial.

Free-form studies are more externally valid than reproduction studies, resembling the real-world usage scenario of a visualization authoring system. They enable researchers to capture the learnability and usability of the system, as participants need to identify how to execute their own design. They may also capture additional insights relating to other criteria; for example, a free-form study with DataInk [45] enabled the researchers to capture evidence of creativity support by identifying the number of different designs that participants created, and by comparing the originality of these designs to existing ones.

Free-form studies require a relatively low degree of effort to execute and may provide useful insights that speak to evaluation criteria other than those targeted by reproduction studies, such as creativity support or expressiveness. On the other hand, they tend to be of short duration and occur in a laboratory setting. Therefore, they may not be appropriate for demonstrating the expressiveness of a system equipped with many features, which will require more time to learn and master. In these cases, other forms of evaluation are preferred, such as design galleries (Section 3.5).

3.4 Comparative Study

Several researchers have conducted a comparative study to compare their authoring system against existing commercial software tools. For example, Amini et al. [4] compared the experience of producing data videos with their DataClips prototype against the combination of Adobe Illustrator and Adobe After Effects, tools that are commonly used in conjunction to produce data videos. They compared the time that study participants took to produce each video clip as well as the number of clips they created. Similarly, Méndez et al. [22] conducted a study comparing iVoLVER [23] against Tableau to better understand the authoring process across different tools. We also note that comparison can be done without recruiting human subjects. For example, Ren et al. compared Charticulator with three existing visualization authoring systems in terms of the number of interactions required to produce an equivalent visualization design, a proxy assessment of efficiency and of their respective complexity, inspired by the keystroke-level model [9].

While a comparative study such as a controlled experiment with quantitative metrics may appear to be the most objective way to compare multiple approaches, we find it particularly difficult to yield scientifically meaningful results due to many confounding factors.

Existing visualization authoring systems differ not only in their interaction but also in their underlying design frameworks.They also differ in terms of the size and robustness of their feature sets; we have remarked above that many research prototypes prioritize the implementation of novel research features at the expense of or instead of seemingly mundane yet often-used features such as undo/redo and z-ordering. Even when researchers discover quantitative differences in completion time or interaction count in a comparative study, it is very difficult to determine which design choices or combination of features yielded the differences, and accordingly the findings are difficult to generalize. Finally, an absence of basic features or the presence of minor usability issues present in a research prototype can easily be addressed soon after the study is conducted (e.g., Data Illustrator implemented undo/redo after the completion of their study). This will often render obsolete the findings from preceding comparative studies.

3.5 Gallery

Recent visualization authoring systems such as DataInk [45], VisComposer [20], InfoNice [44], Data Illustrator [18], and Charticulator [30] are specifically intended for expressive visualization design. This expressiveness cannot be captured through a reproduction study. Even a free-form study may only capture limited insights with respect to expressiveness, as it largely depends on the participants' creativity at the time of the study. To focus on the expressive potential of the authoring system rather than of the participants, researchers will provide a collection of diverse visualization content as a gallery, either within a research paper as a gallery figure and/or as supplemental material, such as within a supplemental project website. The latter can also include video demonstrations of the authoring process for each item (an approach taken in the Data Illustrator [18] and Charticulator [30] projects), which has the additional benefit of increasing the interface's learnability and thus the adoption of the system.

Notable benefits of evaluating an authoring system via a gallery are twofold. First, it is perhaps the best way to demonstrate the expressiveness of the system. Second, these galleries can serve as a benchmark, such that developers of future authoring systems could compare the expressive range of their new system in terms of the degree of overlap relative to the content of previous galleries. For example, Charticulator's gallery [30] reproduces much of the content from Data Illustrator [18] gallery while also introducing new content that cannot be reproduced using Data illustrator.

However, it is important to consider that while the authors of an authoring system can create unique and complex visualization designs, the target users may struggle to produce such content. In other words, a gallery does not assess the usability or learnability of the system, nor does it in itself serve as creativity support for its target audience. For this reason, a gallery is often paired with at least one form of evaluation involving human subjects.

3.6 Combining Multiple Methods

Expressive visualization design is a complex and creative process, requiring a considerable amount of time and effort. To provide a comprehensive evaluation of systems to support such activity, researchers often incorporate many of the methods discussed above in conjunction, to reach a consensus regarding the overall benefits and drawbacks of a novel system, perhaps with reference to more than one of the evaluation criteria discussed above. For example, Liu et al. evaluated Data Illustrated by means of a formative study, a gallery, and a reproduction study, while Ren et al. evaluated Charticulator [30] by means of a gallery, a reproduction study, and a comparative study, Xia et al. evaluated DataInk [45] by asking participants to reproduce a visualization and then create their own custom visualization during a free-form study session, and Mei et al. combined a reproduction study and a comparative study to evaluate VisComposer [20].

When combined with other evaluation methods, a highly limited evaluation can still provide valuable insights. For example, Schroeder and Keefe deployed two versions of the system to one artist [36]. Even though this evaluation alone was not enough to validate the system, it still helped the researchers gather feedback from their target audience to improve the system.

4 OPPORTUNITIES

In this section, we discuss potential ways to facilitate the evaluation and adoption of visualization authoring systems.

4.1 Benchmark Repository

A benchmark is defined as "a standardized problem or test that serves as a basis for evaluation or comparison (as of computer system performance)" [1]. One way to improve and facilitate evaluation is to develop a benchmark, and thus Plaisant called for the creation of benchmark datasets and tasks for evaluating information visualization systems [26]. As a first step, Fekete and Plaisant organized the first InfoVis contest [12] to initiate the development of benchmarks and to establish a forum to promote evaluation methods. They also created the Information Visualization Benchmark Repository [25] to archive materials from this contest, which was replaced and extended by the Visual Analytics Benchmarks Repository [27] in 2006. More recently, we have seen a number of additional visualization data and technique repositories, such as vispubdata.org [14], KeyVis.org (VIS paper keywords) [15], and treevis.net [37]. Similarly, the visualization research community would benefit from such a repository with a specific focus on visualization authoring, which provides both benchmark charts and datasets: a curated collection of examples and links to existing authoring systems.

Charts and Datasets. It takes a substantial amount of effort to prepare a gallery using a variety of externally-valid data and visualizations. This process includes the curation of appropriate datasets, collecting & pre-processing them, visualizing these datasets, and documenting the results. However, it is important to respect the authorship and copyright of the source material; each research group currently needs to attain permission from the dataset owners as well as the visualization authors if a visualization design is reproduced. We envision a catalog or gallery of visualization content (which could be similar to the *Data Visualisation Catalogue* [31]), along with datasets that can be used to create this content, as well as a terms of use policy that grants permission to researchers and system developers to use the data and/or reproduce the designs. Researchers can then leverage these benchmark charts and datasets in the evaluation and comparison of visualization authoring systems.

Visualization Design Contests. The development of visualization design contests may also promote the development and use of benchmark charts and datasets. Contests are a great way to produce outstanding results by engaging the members of the visualization research and practice communities, particularly novices and students. They can also complement comparative studies, which are constrained by the recruitment of study participants who may not receive sufficient training in visualization design during a brief study session. If associated with a live event such as a conference, contests can be a exciting way to motivate attendees to subsequently try out the authoring systems.

Curated Collection of Examples. As discussed in Section 3.1, we have collected existing visualization artifacts to inform the design of several of our visualization authoring systems [4, 24, 28]. There has also been other efforts in curating collections of visualization examples, particularly in the context of data-driven storytelling [19, 39, 41]. The visualization community could benefit from such curated sets of examples, as they often include custom visualization designs that have received accolades from venues such as the Kantar Information is Beautiful Awards (https://www.informationisbeautifulawards.com), the Malofiej Infographic World Summit (http://www.malofiejgraphics.com), and various data journalism conferences.

4.2 Deployment & Adoption

With the advancement of web technology, many visualization authoring systems are developed for the web environment. This makes it easy for researchers and developers to deploy the systems online and thereby reach a large audience. However, the deployment of systems and the facilitation of adoption involves a significant amount of effort that may not necessarily lead to research contributions. On the other hand, if the evaluation criteria of interest relates to how the system integrates with existing authoring workflows, or how people other than those who participated in the design of the system express themselves by using it, the deployment of the system and a study of its adoption is highly valuable.

Barriers to Studying Adoption. Deploying a visualization authoring system and studying its adoption involves a substantial amount of effort to ensure that the system is usable and learnable, even when assessing its usability and learnability are not the main goals of the evaluation. In addition to engineering for usability and learnability, studying the adoption of a system entails tracking its usage.

Since it is useful to collect the content that authors produce using the system, researchers must consider ways to securely store and analyze this author-generated content. However, researchers should be aware of their ethical and legal responsibilities when storing author-generated content. The European Unions General Data Protection Regulation (GDPR), which came into effect in May 2018, has several implications for researchers and particularly for those employed by industrial research labs. First, content producers using the system should opt-in to sharing their content with researchers, who must clearly state how the content will be used, such as in a research paper or online gallery. Second, content producers should be able to revoke this consent and request that researchers delete their content. Finally, while the usage tracking of any system may involve the collection of personally identifiable information (PII), visualization authoring systems are especially problematic in this regard in that it may be possible to identify individuals via the content they produce, such as when authors visualize their personal data. As a consequence, the secure storage of personally identifiable information or the de-identification of personal information are both very challenging from a practical standpoint.

Researchers may also allow visualization authors to share and discuss their content within the system. Doing so may shed light on to how people use the system to collaborate with others, which was a focus of the researchers who developed the ManyEyes [11] system. However, when including such functionality, researchers should be aware of the associated ethical and legal requirements, such as moderating both the shared content and the ensuing discussion, which entails removing objectionable content and regulating access to those who abuse the discussion platform.

Facilitating Adoption. Adoption is influenced by many factors, such as the timing of the deployment, the dispersion and quality of marketing and tutorial materials, the system developers' level of influence within the target user community, and their ability to respond to technical support requests. We realize that many of these factors are difficult to control or predict, and thus a desired level of adoption is seldom guaranteed.

Despite this uncertainty, we are aware of tangible approaches to producing tutorial content for visualization authoring systems. One approach is a guided-tour tutorial; ChartAccent [28] and Timeline Storyteller [24] both incorporate an Intro.js tutorial (https://introjs.com) that provides a step-by-step walkthrough of their interfaces. Similarly, Data Illustrator [18] provides a tutorial called "Getting Started," which mimics traditional help pages enriched with multimedia contents such as images and videos. Data Illustrator also leverages its example gallery as a way of demonstrating possible visualization designs, where each gallery image has a corresponding video documenting the visualization creation process. While such videos are great for demonstrating how to create the target visualization, it is still difficult for an individual to create one by following the video, as this process tends to involve multiple browser windows and frequent play/pause and rewind/fast-forward actions. To address this, Charticulator [30] provides a video player augmented with content segmentation and accompanying textual descriptions, a design inspired by Pluralsight (https://pluralsight.com), a popular self-learning platform. This approach is promising and versatile, and we see great potential in the design space of such interactive tutorials for visualization authoring systems.

Analysis of Author-Generated Content. Assuming that such content is collected responsibly, a corpus of visualization content generated by people using a novel authoring system provides great value to the visualization research community. More immediately, it provides an indication of a system's expressiveness, complementing other approaches such as a gallery of content produced by the system developers. The analysis of author-generated content also provides an indication of who the authors are and what data they visualize. Such was the case with ManyEyes [43], a web-based visualization authoring system deployed online between 2007 and 2015, which attracted a wide variety of users who visualized, shared, and discussed an equally varied range of datasets [42]. Researchers may report upon a meta-analysis of the visualization design choices used by authors of the system and their coverage of the system's design space. Subject to the consent of the authors, researchers may also include examples of author-generated visualization content in their research papers or on a supplemental gallery website.

Adoption Case Studies. Individual use cases can illustrate the expressiveness and efficiency of the system, as well as how the system integrates with an author's workflow. One approach involves partnering with one or more content producers and studying their process of using the system with their own data in their own environment, perhaps over the course of multiple sessions or individual authoring projects, in the spirit of the Multi-dimensional In-depth Long-term Case study (MILC) approach [40].

Our use of 'case study' is in line with Sedlmair et al.'s characterization of the term, distinguishing it from a 'usage scenario' envisioned and/or performed by the researchers [38]. In other words, case studies involve a person who did not contribute to the design and development of a visualization authoring system. Some case studies involve the solicitation of particular individuals from the target user group; in the context of visualization authoring systems, these may be journalists, designers, or educators. Researchers may engage with these individuals before, during, and after their use of the visualization authoring system via interviews and potentially directly observing their usage.

Alternatively, case studies may involve people who used the system by their own volition without direct solicitation from the researchers. For instance, consider a visualization authoring system deployed online and advertised on social media; a journalist uses the system and publishes the visualization output alongside an online news article, which in turn is shared online. This presents an opportunity for researchers to reach out to this journalist and interview them about their experience using the system, potentially demonstrating their use of the tool, either in person or via a screen-sharing interview. This approach was taken by Brehmer et al. in their interviews with journalists following their use of the Overview visualization system during the course of their journalistic investigation [7]; though Overview was a visual analysis tool, the same approach could be taken with a visualization authoring system.

5 Conclusion

We have discussed the difficulties in evaluating visualization authoring systems using traditional comparative studies, reflecting on our own experience. While comparative studies are appropriate for evaluating efficiency (e.g., task time, error), controlled experiments are not necessarily appropriate for the diverse set of evaluation criteria relevant to interactive visualization authoring systems.

Collectively, we have in recent years developed and evaluated six visualization authoring systems using the collection of existing approaches for evaluating such systems: reproduction studies, free-form studies, design galleries, comparative studies, formative studies, and various combinations thereof. Given the complex and multi-faceted nature of authoring systems for expressive visualization design, we employed more than one evaluation method in these evaluations. Even though each evaluation method may have shortcomings, their combination helped us to improve our design and demonstrate the main contribution of the authoring system, complementing one another. We thus encourage other researchers to consider combining a few evaluation methods, not following our combinations exactly but devising their own combinations.

We would like to emphasize that it is important for researchers to deliberate and justify why their selection of evaluation methods are appropriate for supporting their intended research contributions, and to clearly provide their rationale in their research papers. This will help to set the right expectations among those reviewing these papers. We believe that it would be beneficial to the visualization research community if researchers avoid conducting comparative studies of authoring systems, as these are often contrived and merely included to satisfy reviewers who expect some form of evaluation.

Finally, we see several opportunities for the future of evaluating visualization authoring systems, including the creation of benchmarks and methods for studying adoption through deployment. We hope that both researchers and reviewers will gain insights from this paper, so that they might select or recommend more appropriate evaluation criteria and methods that better demonstrate the research contributions of a novel visualization authoring system.

References

[1] Definition of *benchmark* by Merriam-Webster. `https://www.merriam-webster.com/dictionary/benchmark`.

[2] F. Amini, M. Brehmer, G. Bolduan, C. Elmer, and B. Wiederkehr. Evaluating data-driven stories & storytelling tools. In N. Henry Riche, C. Hurter, N. Diakopoulos, and S. Carpendale, eds., *Data-Driven Storytelling*. A K Peters/CRC Press, 2018.

[3] F. Amini, N. Henry Riche, B. Lee, C. Hurter, and P. Irani. Understanding data videos: Looking at narrative visualization through the cinematography lens. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 1459–1468, 2015. doi: 10.1145/2702123.2702431

[4] F. Amini, N. H. Riche, B. Lee, A. Monroy-Hernandez, and P. Irani. Authoring data-driven videos with dataclips. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 23(1):501–510, 2017. doi: 10.1109/TVCG.2016.2598647

[5] A. Bigelow, S. Drucker, D. Fisher, and M. Meyer. Iterating between tools to create and edit visualizations. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 23(1):481–490, 2017. doi: 10.1109/TVCG.2016.2598609

[6] M. Brehmer, S. Carpendale, B. Lee, and M. Tory. Pre-design empiricism for information visualization: scenarios, methods, and challenges. In *Proceedings of the ACM Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)*, pp. 147–151, 2014. doi: 10.1145/2669557.2669564

[7] M. Brehmer, S. Ingram, J. Stray, and T. Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 20(12):2271–2280, 2014. doi: 10.1109/TVCG.2014.2346431

[8] M. Brehmer, B. Lee, B. Bach, N. H. Riche, and T. Munzner. Timelines revisited: A design space and considerations for expressive storytelling. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 23(9):2151–2164, 2017. doi: 10.1109/TVCG.2016.2614803

[9] S. K. Card, T. P. Moran, and A. Newell. The keystroke-level model for user performance time with interactive systems. *Communications of the ACM*, 23(7):396–410, 1980. doi: 10.1145/358886.358895

[10] S. Carpendale. Evaluating information visualizations. In *Information visualization*, pp. 19–45. Springer, 2008. doi: 10.1007/978-3-540-70956-5_2

[11] C. M. Danis, F. B. Viegas, M. Wattenberg, and J. Kriss. Your place or mine?: Visualization as a community component. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 275–284, 2008. doi: 10.1145/1357054.1357102

[12] J.-D. Fekete and C. Plaisant. InfoVis 2003 Contest: Visualization and pair wise comparison of trees, 2003. `http://www.cs.umd.edu/hcil/iv03contest`.

[13] L. Grammel, C. Bennett, M. Tory, and M.-A. Storey. A survey of visualization construction user interfaces. In *Short Paper Proceedings of EuroVis*, 2013. doi: 10.2312/PE.EuroVisShort.EuroVisShort2013.019-023

[14] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. vispubdata.org: A metadata collection about IEEE visualization (VIS) publications. *IEEE Transactions on Visualization and Computer Graphics*, 23(9):2199–2206, Sept. 2017. doi: 10.1109/TVCG.2016.2615308

[15] P. Isenberg, T. Isenberg, M. Sedlmair, J. Chen, and T. Möller. Toward a deeper understanding of visualization through keyword analysis. Technical Report RR-8580, INRIA, France, Aug. 2014. Also published on arXiv.org (# 1408.3297).

[16] A. Kirk. The Chartmaker Directory, 2017. `http://chartmaker.visualisingdata.com`.

[17] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 18(9):1520–1536, 2012. doi: 10.1109/TVCG.2011.279

[18] Z. Liu, J. Thompson, A. Wilson, M. Dontcheva, J. Delorey, S. Grigg, B. Kerr, and J. Stasko. Data Illustrator: Augmenting vector design tools with lazy data binding for expressive visualization authoring. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 123:1–123:13, 2018. doi: 10.1145/3173574.3173697

[19] S. McKenna, N. Henry Riche, B. Lee, J. Boy, and M. Meyer. Visual narrative flow: Exploring factors shaping data visualization story reading experiences. *Computer Graphics Forum (Proceedings of EuroVis)*, 36(3):377–387, 2017. doi: 10.1111/cgf.13195

[20] H. Mei, W. Chen, Y. Ma, H. Guan, and W. Hu. VisComposer: A visual programmable composition environment for information visualization. *Visual Informatics*, 2(1):71–81, 2018. doi: j.visinf.2018.04.008

[21] H. Mei, Y. Ma, Y. Wei, and W. Chen. The design space of construction tools for information visualization: A survey. *Journal of Visual Languages & Computing*, 2017. doi: 10.1016/j.jvlc.2017.10.001

[22] G. G. Méndez, U. Hinrichs, and M. A. Nacenta. Bottom-up vs. top-down: Trade-offs in efficiency, understanding, freedom and creativity with infovis tools. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 841–852, 2017. doi: 10.1145/3025453.3025942

[23] G. G. Méndez, M. A. Nacenta, and S. Vandenheste. iVoLVER: Interactive visual language for visualization extraction and reconstruction. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 4073–4085, 2016. doi: 10.1145/2858036.2858435

[24] Microsoft. Timeline Storyteller, 2017. `https://timelinestoryteller.com`.

[25] C. Plaisant. Information visualization benchmarks repository, 2003. `http://www.cs.umd.edu/hcil/InfovisRepository`.

[26] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the ACM Conference on Advanced Visual Interfaces (AVI)*, pp. 109–116, 2004. doi: 10.1145/989863.989880

[27] C. Plaisant. Visual analytics benchmark repository, 2006. `https://www.cs.umd.edu/hcil/varepository`.

[28] D. Ren, M. Brehmer, B. Lee, T. Höllerer, and E. K. Choe. ChartAccent: Annotation for data-driven storytelling. In *Proceedings of the IEEE*

Pacific Visualization Symposium (PacificVis), pp. 230–239, 2017. doi: 10.1109/PACIFICVIS.2017.8031599

[29] D. Ren, T. Höllerer, and X. Yuan. iVisDesigner: Expressive interactive design of information visualizations. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 20(12):2092–2101, 2014. doi: 10.1109/TVCG.2014.2346291

[30] D. Ren, B. Lee, and M. Brehmer. Charticulator: Interactive construction of bespoke chart layouts. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 25(1), 2019. doi: 10.1109/TVCG.2018.2865158

[31] S. Ribecca. The data visualisation catalogue. `https://datavizcatalogue.com/`.

[32] G. Robertson, M. Czerwinski, D. Fisher, and B. Lee. Selected human factors issues in information visualization. *Reviews of Human Factors and Ergonomics*, 5(1):41–81, 2009. doi: 10.1518/155723409X448017

[33] L. C. Rost. What I learned recreating one chart using 24 tools. *Source*, 2016. `https://goo.gl/uGE5dc`.

[34] A. Satyanarayan and J. Heer. Authoring narrative visualizations with Ellipsis. *Computer Graphics Forum (Proceedings of EuroVis)*, 33(3), 2014. doi: 10.1111/cgf.12392

[35] A. Satyanarayan and J. Heer. Lyra: An interactive visualization design environment. *Computer Graphics Forum (Proceedings of EuroVis)*, 33(3), 2014. doi: 10.1111/cgf.12391

[36] D. Schroeder and D. F. Keefe. Visualization-by-sketching: An artist's interface for creating multivariate time-varying data visualizations. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 22(1):877–885, 2016. doi: 10.1109/TVCG.2015.2467153

[37] H.-J. Schulz. Treevis. net: A tree visualization reference. *IEEE Computer Graphics & Applications (CG&A)*, (6):11–15, 2011. doi: 10.1109/MCG.2011.103

[38] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 18(12):2431–2440, 2012. `https://doi.org/10.1109/TVCG.2012.213`.

[39] E. Segel and J. Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 16(6):1139–1148, 2010. doi: 10.1109/TVCG.2010.179

[40] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the ACM Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)*, pp. 1–7, 2006. doi: 10.1145/1168149.1168158

[41] C. D. Stolper, B. Lee, N. Henry Riche, and J. Stasko. Data-driven storytelling techniques: Analysis of a curated collection of visual stories. In N. Henry Riche, C. Hurter, N. Diakopoulos, and S. Carpendale, eds., *Data-Driven Storytelling*. A K Peters/CRC Press, 2018.

[42] F. B. Viegas, M. Wattenberg, M. McKeon, F. Van Ham, and J. Kriss. Harry Potter and the meat-filled freezer: A case study of spontaneous usage of visualization tools. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, pp. 159–159, 2008. doi: 10.1109/HICSS.2008.188

[43] F. B. Viegas, M. Wattenberg, F. Van Ham, J. Kriss, and M. McKeon. ManyEyes: a site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 13(6), 2007. doi: 10.1109/TVCG.2007.70577

[44] Y. Wang, H. Zhang, H. Huang, X. Chen, Q. Yin, Z. Hou, D. Zhang, Q. Luo, and H. Qu. InfoNice: Easy creation of information graphics. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 335:1–335:12, 2018. doi: 10.1145/3173574.3173909

[45] H. Xia, N. Riche, F. Chevalier, B. D. Araujo, and D. Wigdor. DataInk: Enabling direct and creative data-oriented drawing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 223:1–223:13, 2018. doi: 10.1145/3173574.3173797