Evaluating Visual Data Analysis Systems: A Discussion Report

Leilani Battle, Marco Angelini, Carsten Binnig, Tiziana Catarci, Philipp Eichmann, Jean-Daniel Fekete, Giuseppe Santucci, Michael Sedlmair, Wesley Willett

To cite this version:

Leilani Battle, Marco Angelini, Carsten Binnig, Tiziana Catarci, Philipp Eichmann, et al.. Evaluating Visual Data Analysis Systems: A Discussion Report. HILDA'18: Workshop on Human-In-the-Loop Data Analytics, Jun 2018, Houston, TX, United States. 10.1145/3209900.3209901 . hal-01786507

HAL Id: hal-01786507

https://hal.inria.fr/hal-01786507

Submitted on 5 May 2018

Evaluating Visual Data Analysis Systems: A Discussion Report

Leilani Battle[1], Marco Angelini[6], Carsten Binnig[2,3], Tiziana Catarci[6], Philipp Eichmann[3], Jean-Daniel Fekete[4], Giuseppe Santucci [6], Michael Sedlmair[7], Wesley Willett[5]

[1] University of Washington, USA [2] TU Darmstadt, Germany [3] Brown University, USA
[4] Inria, France [5] University of Calgary, Canada [6] University of Rome "La Sapienza", Italy
[7] Jacobs University Bremen, Germany

ABSTRACT

Visual data analysis is a key tool for helping people to make sense of and interact with massive data sets. However, existing evaluation methods (e.g., database benchmarks, individual user studies) fail to capture the key points that make systems for visual data analysis (or *visual data systems*) challenging to design. In November 2017, members of both the Database and Visualization communities came together in a Dagstuhl seminar to discuss the grand challenges in the intersection of data analysis and interactive visualization.

In this paper, we report on the discussions of the working group on the evaluation of visual data systems, which addressed questions centered around developing better evaluation methods, such as "How do the different communities evaluate visual data systems?" and "What we could learn from each other to develop evaluation techniques that cut across areas?". In their discussions, the group brainstormed initial steps towards new joint evaluation methods and developed a first concrete initiative — a trace repository of various real-world workloads and visual data systems — that enables researchers to derive evaluation setups (e.g., performance benchmarks, user studies) under more realistic assumptions, and enables new evaluation perspectives (e.g., broader meta analysis across analysis contexts, reproducibility and comparability across systems).

ACM Reference Format:
Leilani Battle[1], Marco Angelini[6], Carsten Binnig[2,3], Tiziana Catarci[6], Philipp Eichmann[3], Jean-Daniel Fekete[4], Giuseppe Santucci [6], Michael Sedlmair[7], Wesley Willett[5] . 2018. Evaluating Visual Data Analysis Systems: A Discussion Report. In *HILDA'18: Workshop on Human-In-the-Loop Data Analytics, June 10, 2018, Houston, TX, USA.* ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/3209900.3209901

1 INTRODUCTION

Motivation: Data visualizations are key for helping people to explore and understand data sets. To that end, it is not surprising that there exists an ever growing set of data-centric systems through which domain experts and data scientists of varying skill levels can interactively analyze and explore large data sets.

However, despite their growing popularity, it is currently unclear how to best evaluate this new class of *visual data systems*, which must balance both performance and usability. For example, existing database benchmarks are standardized and thus easy to reproduce, however they often rely on assumptions that are far from representative of real-world analytics tasks and user interactions. Unlike database benchmarks, visualization user studies can successfully capture the results of real analysis tasks performed on real-world datasets. However, these studies often rely on non-standardized workloads, and heavily restricted user populations. To that end, user studies are not easy to reproduce, limiting the comparability of their results. Finally, the reported metrics in both database benchmarks and user studies often do not cover all relevant aspects of visual data systems.

When considered individually, existing evaluation methods fail to capture the complex mix of performance considerations that make visual data systems challenging to design. A main reason for this situation is that the database and visualization communities so far have been mostly disparate, even though a lot of challenging high impact problems, and thus the best approaches to evaluating future solutions, lie at the intersection of these fields. Recently a series of workshops have been created that focus on the intersection of data management and visualization (DSIA@Vis, HILDA@SIGMOD, BigVis@EDBT, IDEA@KDD). During these workshops, first encouraging discussions have originated on how to better bridge the gap between data management systems and visualization systems to better support visual data analysis scenarios. These discussions were continued at a recent Dagstuhl Seminar in November 2017[1].

Contributions: As part of this Dagstuhl seminar, individual working groups discussed different challenges ranging from core database engine design to more user-centered questions. One of the challenges addressed how to design appropriate evaluation methods and benchmarks for visual data systems.

In this paper, as a first contribution *we report on the discussions of the Dagstuhl working group on evaluation methods.* Through these conversations, we have identified a clear need for more principled evaluation methods that are developed by both communities for evaluating visual data analysis systems in a reproducible and comparable manner, based on realistic workloads and data sets.

Furthermore, as second contribution, we propose the first concrete step towards enabling new evaluation methods through the

HILDA'18, June 10, 2018, Houston, TX, USA

ACM ISBN 978-1-4503-5827-9/18/06…$15.00
https://doi.org/10.1145/3209900.3209901

[1] https://www.dagstuhl.de/en/program/calendar/semhp/?semnr=17461

design of a new *trace repository*. Our repository design will allow researchers to share traces of real-world workloads resulting from visual data analysis systems which not only include user interactions, but also data sets that are explored, as well as important meta data about the users and their analysis tasks. The goal of this repository is to provide researchers of both communities a basis to derive and justify realistic and reproducible evaluation methods. For example, in order to make the results of user studies better comparable they would benefit from the fact that reproducible workloads and data sets that are shared in such a repository could be used without giving up the focus on users and their tasks. Furthermore, database benchmarks could be derived directly from real-world datasets, tasks and interaction logs collected through the repository, instead of using overly simplistic synthetic workloads.

Outline: The outline of the remainder of this paper is structured as follows: the paper begins with a brief overview of existing evaluation techniques for visual data systems in the different communities. We then present a vision of how new joint evaluation methods of both communities should be designed and derive requirements for a repository to share traces. Using these requirements we propose an initial design for a trace repository. We then demonstrate the potential of such a repository by describing how it can steer a recent initiative to design a new benchmark for evaluating database systems with regard to supporting users in visual and interactive data exploration scenarios. Finally, we discuss a roadmap of short- and long-term goals.

2 EXISTING EVALUATION EFFORTS

Evaluation has taken an essential role in both the database- and visualization community, but with emphasis on different evaluation goals. While the database community has mainly focused on system speed and performance, the visualization community has focused more on user-centered criteria. In this section, we briefly review the main methods and resources describing these evaluation methods, as well as recent efforts to bridge the evaluation gap between the two communities.

2.1 Efforts in the Visualization Community

Visualization systems are generally evaluated with the intention of assessing whether they support users in completing pre-defined high level analysis goals. The visualization community has partitioned the evaluation space in different ways, both hierarchically (e.g., the four-level nested model [18, 20]) and broadly as high-level analysis scenarios for evaluation (e.g., the seven guiding scenarios [16]). Evaluations range from field observations and interviews, to carefully controlled lab studies, to algorithm metrics and performance experiments.

The visualization community therefore takes a top-down approach to evaluation, where a high level analysis goal is formulated for users (e.g., to support exploratory data analysis of tabular data). A high-level goal is then broken down into tasks, and each task into sequences of smaller steps (i.e., interactions) [4], which ultimately inform the design of the analysis interface, and oftentimes the underlying system optimizations.

Plaisant et al. [22] also recommend using benchmarks and contests to perform insight-based evaluations. These benchmarks have been used in contests that were popular in the InfoVis Community in 2003–2005, and then moved to the VAST Community with the VAST Challenge[2] organized every year since 2006 [5, 23]. These contests remain a remarkable asset of the visualization community to share evaluation results and collect realistic benchmarks to test systems and engage researchers.

However, there is generally a lack of quantitative and comparative methods for evaluating the performance of large-scale visual exploration systems. In particular, user studies are designed and conducted to perform customized evaluations of individual systems, but the user interaction logs, system logs and metadata (e.g., used datasets, used interfaces) are rarely shared or published in a standardized way to enable reproducibility.

2.2 Efforts in the Database Community

The database community has developed its own evaluation methods for a long time, capitalizing on SQL as a powerful common denominator [10]. Database evaluations focus on running and comparing DBMSs using representative and tunable performance benchmarks. The most commonly used collection of benchmarks is created by the TPC consortium [21], which covers synthetic benchmarks for a variety of workloads ranging from transactional over analytical workloads to more specialized benchmarks for data integration, etc.

These benchmarks define a set of queries and data sets of different scales, and study what happens when different distributions of queries are executed over time. The emphasis of the reported metrics is on system performance (e.g., latency or throughput). Moreover, evaluations in the database community often utilize narrow, low-level methods (i.e., micro-benchmarks) that are performed on specific aspects or parts of the DBMS. The micro-benchmark results are often used to show the trade-off of design alternatives rather than the full benchmark evaluations that target a complete DBMS.

In the recent years, there has been a growing interest in the database community for many forms of data exploration [13] and in particular visualization based exploration. Supporting visualization at the database level has technical implications, such as sub-second response times to support interactivity, complex queries such as cross-filtering [24], and more recently approximate results to lower the latency of database systems on large data sets, and in particular online aggregation [12] and its more recent incarnation coined as *progressive* systems [26]. Unfortunately, the afore-mentioned existing benchmarks do not cover any of these aspects yet.

2.3 Joint Efforts of both Communities

To forge a stronger connection between the communities, a series of workshops have recently launched during the main database and visualization conferences. As a result, we see an increasing number of articles published across communities: visualization-related articles at SIGMOD and VLDB and data-management related articles at IEEE VIS, EuroVis, and ACM CHI.

Yet, the database community still relies mostly on measuring system performance, convergence speed, and statistically validated

[2] see www.vacommunity.org/About+the+VAST+Challenge for more information on the VAST challenge.

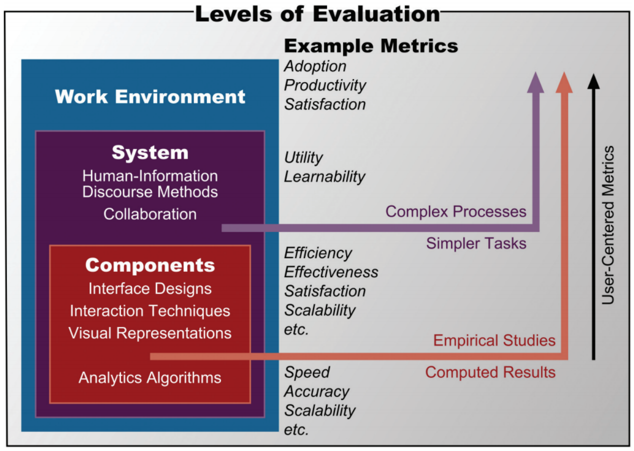

Figure 1: The three levels of evaluation [27]

measures for approximate results, whereas the visualization community still focuses on how humans can derive valid insights and make sound decisions through data visualization and interaction. Therefore, there is a need for applying human-oriented measures to database systems, performance-oriented measures to visualization systems, and to converge on accepted methods to gather these measures in a reproducible way, on collecting use-cases, scenarios, and benchmarks applicable across communities.

Recent vision papers [3, 7] discuss these issues and present initial ideas on how to design new benchmarks for visual data systems that better reflect the properties of real interactive analysis workloads and real-world data sets. They also propose measures for system efficiency and effectiveness that go beyond database query performance. While these are promising first steps, the data on which their findings are based on is limited. They would thus benefit from a trace repository that allows researchers to systematically derive characteristics for workloads and data sets that take the needs of real users and tasks into consideration.

3 DESIGNING NEW EVALUATION METHODS

In this section, we first outline the core vision for designing new evaluation methods for visual data analysis based on a shared trace repository. We then discuss the main dimensions that should be captured in a shared trace repository, and describe emerging requirements and challenges for the design of such a repository.

3.1 Vision of a Trace Repository

Visual analytics systems are complex and require evaluation methods that not only make realistic assumptions about users and workloads but are also reproducible at the same time. We therefore put forward a vision of a new shared trace repository that contains information about real workloads that can be used by researchers to derive new realistic and reproducible evaluation methods.

The main challenge is that such a new trace repository should be able to capture all the information required to enable the beforementioned goals. For example, several critical factors must be considered to fully capture how user study results were derived. The visualization interface dictates how the user will interact with the underlying data set. The use of pre-defined tasks (or lack thereof) will direct the course of the user's analysis. The user's own knowledge and expertise will influence his or her analysis strategies and behaviors. The data set itself will likely support only a specific set of analysis methods. Each of these factors represents a piece of meta data that is rarely, if ever, recorded and shared in the visualization and database communities. However, it is precisely these meta data that enable evaluation results to be comparable and reproducible. As such, we argue that *meta data should be treated as a first-class evaluation artifact*, alongside the user's interaction logs and the data set being explored.

We also aim to start the process of standardization between the visualization and database communities. The implementation of a shared repository as a valuable first step in solidifying the terminology used between the communities, and for highlighting shared evaluation goals. The repository also allows us to disseminate best practices from both communities, such as creating more realistic workloads for database performance evaluations by running more user studies. Finally, the repository can help to define the right metrics for benchmark evaluations, which can directly benefit existing benchmark development efforts [3, 7].

To define the scope of our repository, we consider the seminal book by Thomas & Cook [27], which structures the evaluation methods for visual data analysis in three different levels: component, system, and work environment (see Figure 1). Each level defines a set of possible evaluation methods and metrics that should be measured. As a concrete starting point, we suggest that the visualization and database communities work together in a *bottom-up approach*, starting to derive new evaluation methods on the component level to evaluate database systems, visualizations front-ends, etc.

3.2 Requirements

The main idea for designing novel evaluation methods for visual data analysis is to leverage traces for deriving new more realistic benchmarks and reproducible user studies. A trace can be thought of as a recording of analysis sessions of a user. In the simplest version, a trace could be a simple interaction log along with the data used by the user. More sophisticated traces could also include interaction information on other levels such as screen-recordings, or even videos of users and eye-tracking but also meta data about users and tasks. A core aspect in this trace-based approach will be to define the space of possible information that should be covered in such a repository. In the following, we discuss an initial set of requirements that we believe are critical to collecting effective traces (see Section 4 for our more detailed repository proposal):

Interactions: Users reveal important information about their reasoning processes [6] and analysis goals [2] through the interactions they perform, making interaction sequences an essential component of useful traces. Interactions can range from direct to more sophisticated methods of interaction [11]. Examples for direct interactions include standard control elements such as range selection, cross filters, and binning; more sophisticated interaction methods, such as natural-language query interfaces, should also be considered.

Data: Data set characteristics like size, type, distribution, etc. also play an important role when evaluating the effectiveness of a visual data system. For example, dataset size directly affects query execution speeds (and in turn interaction speeds), motivating the development of optimizations to reduce dataset size, such as online aggregation and progressive visualization.

User: Variations in each user's background and analysis context can have a huge impact on evaluation results, but this information is often overlooked in existing evaluations. Thus, recording user characteristics will be a key component for our trace repository. Users have different levels of expertise; making visualization systems easier to use by visualization novices [9] might not necessarily work well for experts. Other important factors include the application domain and the "type" of user, such as developer, data scientist, or domain expert. Other contexts, such as collaborative data analysis [15], can also greatly influence the analysis strategies and underlying behaviors observed in trace data.

Tasks: Many evaluations, particularly user studies, require users to complete one or more tasks, to direct the user's focus in her analysis. Knowing the specific tasks that the user completed, or the broader analysis goal that the user was aiming to accomplish, provides additional insight into why the user may have performed certain interactions observed through the resulting log data. As such, capturing analysis task and goal information, and ultimately user intent, is of great importance for an effective trace repository. There is a variety of different tasks that visualization systems support. Specific tasks will depend heavily on the problem domain [18, 20]. However, task taxonomies (e.g. [4]) provide a useful starting point for generalizing tasks across domains.

Visual Representation: This dimension specifies the set of visualization encodings that were supported in the user interface, and which encodings were actually applied by users. This information is necessary to understand the user's interaction choices and analysis outcomes, since the visual interface itself has a strong effect on how users perceive and interpret the underlying data.

While this list is far from a complete set of considerations, it illustrates the diversity of information required.

3.3 Challenges

Given the vision and requirements for a new trace repository, in this section we discuss the major challenges of its implementation.

(C1) Heterogeneity of traces: The first challenge concerns the collection of the traces in large number and from a variety of real systems. Designing a common trace structure to support the heterogeneity that real systems have in terms of supported tasks, involved users, data, technologies is a critical challenge for the repository design. This includes questions like: how can the repository scale to a large number of collected traces that were captured during field studies and controlled environments, and how can these different traces be compared in a meaningful way?

(C2) Tracing levels: Another core aspect is how to collect traces (or augment existing traces) such that they match the desired level of detail. Low-level traces such as those directly taken from systems such as Tableau may contain too many unnecessary details and can be difficult to parse. However, high-level descriptions of user interactions (e.g., from the VAST challenge), are hard to formalize. Furthermore, more abstract concepts tied to evaluation, like the analyst intent in terms of goal of analysis, are generally not explicitly present in the traces. Thus defining the right trace format to capture all the desired evaluation information is nontrivial.

(C3) Sharing traces: A trace repository must respect privacy regulations, e.g., by considering explicit consent requests for trace collection and curation and/or providing anonymization of traces in order to allow trace sharing.

(C4) Extracting workloads: Benchmarks are essentially an abstraction layer above traces. How to best transform the knowledge extracted from traces into representative workloads is a critical problem to address. Furthermore, different workloads must be derived for different evaluation goals, and the definition of these categories is an activity to conduct. Existing categorizations in visualization could be a useful starting point [16]. Determining how this abstraction process can move from manual to automated methods is an important direction for future work.

(C5) Interoperability with existing methodologies: A final challenge is to establish how the repository can support existing evaluation methodologies (e.g., the Nested Model [18, 20] and Seven Guiding Scenarios [16]).

4 A REPOSITORY PROPOSAL

In this section, we discuss a concrete first proposal of how a trace repository should be designed.

4.1 Overview

To better support a diverse range of evaluation methods, we propose a general-purpose, standardized representation for collecting, sharing, translating, integrating, and aggregating analysis logs at varying levels of abstraction. Specifically, we propose a multi-tiered representation that makes it possible to consistently share, aggregate, and analyze real-world analysis logs at a variety of different levels of abstraction (e.g., low-level systems logs but also high-level logs resulting from user studies).

Consider, for instance, the study conducted by Liu and Heer [17] on the effects of interactive latency on exploratory visual analysis. They designed a user study to investigate the effects of induced latency in an interactive data exploration scenario. To do so they asked participants to explore two datasets using operations such as brushing, panning, and zooming, on a set of linked visualizations. They recorded user traces at multiple levels, capturing lower-level events such as mouse events (including clicks and moves) and higher-level application events (such as "brush", "select", and "range select"), and used a think- aloud protocol to capture users' verbal observations . Participants were instructed to report interesting findings and the authors recorded the audio and additionally took notes. The authors then manually transcribed the audio recordings to text scripts, and segmented and coded them with seven cognitive behavior categories: observation, generalization, hypothesis, question, recall, interface, and simulation.

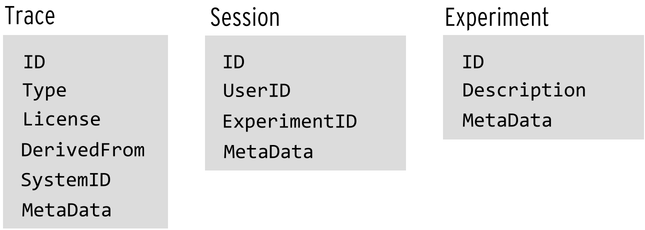

Figure 2: Elements of a trace repository

This combination of traces at a variety of levels enabled them to produce a rich and detailed analysis of analytic behaviors. For example, they were able to investigate how frequently users switched between different higher-level tasks such as brushing and zooming, which mouse movements typically precede certain cognitive behaviors, etc.

4.2 A Common Trace Format

The above example illustrates the value of collecting traces at multiple levels of abstraction for an in-depth analysis of usage and behavioral patterns. However, researchers rarely publish all recorded traces. We hope that introducing a common representation for interaction logs can enable the collection and sharing of larger repositories of realistic analyses. Establishing a shared format will also support the development of general purpose libraries for translating between representations — including automatically identifying higher-level analysis tasks or producing low-level workloads from large volumes of real-world application logs.

To address challenges *C1* and *C2* in Section 3.3, a common format should support more consistent sharing and integration of various existing types of analysis logs including:

- Contextual traces like video and audio recordings of analysis sessions as well as screen-capture and eye-tracking data
- Operating-system level interaction logs including individual movements, clicks, etc.
- Application logs with tool-level interactions and operations
- Query logs capturing individual database queries

This format would also provide a more standardized and shareable representation (*C3* and *C4*) for other kinds of task abstractions and annotations including:

- Annotations and metadata characterizing unique activities, findings, analyst productivity, etc.
- High- and low-level characterizations of analysis tasks [1, 25] for examining task-oriented subsets of analysis processes
- Abstracted workloads for more realistic evaluation of databases, machine learning platforms, and other analysis systems

Furthermore, traces might have very different characteristics — some might be sequences of events described in a semi-structured text file, while others might be entire self-contained logs (a Tableau log export, database log files, etc.) or even self-contained files such as video or audio recordings of a session. The lack of a common format to uniformly describe and identify such traces makes it difficult to publish and share this kind of data. In the following we outline a first draft of how such a common format might look like. It is intentionally designed at the meta-level of traces in order to guarantee *interoperability with existing methodologies (C5).* Conceptually we differentiate between traces, sessions and experiments (see Figure 2), each of which is described by a manifest.

- *Traces* may contain log data (including system logs, audio, video, screen recordings, etc.). Each trace file includes associated metadata describing its content (such as information about the system that produced the trace). In many cases, it may be useful to group together sets of traces that share common information — such as traces from different conditions in a controlled experiment. `Experiment` and `session` manifests provide a standardized representation for connecting these kinds of related traces.
- *Sessions* provide details about individual interaction sessions, and can be used to connects multiple types of traces (query logs, interaction logs, audio and video, etc.) which describe the same set of interactions. A `session` manifest has an `ID` that uniquely identifies it, as well as an optional `UserID` that can be used to link multiple sessions that correspond to the same individual. Sessions that are part of a larger experiment can also contain an `ExperimentID` as well as additional `Metadata`, including experiment conditions.
- *Experiments* provide details about overarching experimental protocols and can be used to connect multiple sessions that were conducted as part of the same experiment. An `experiment` manifest includes a unique `ID`, along with a text `Description` of the experiment as well as optional additional `Metadata`.

While these are initial ideas as to which data should be contained in a trace format, it highlights the importance of meta-data describing traces and their context.

4.3 An Example Use Case

In a recent paper [7], the authors argue that existing database benchmarks are not suitable to evaluate the performance of interactive data exploration (IDE) systems where most queries are ad-hoc, not based on predefined reports and queries are built incrementally. The authors present a novel benchmark called IDEBench [3] [8] with the aim of providing a benchmark whose workloads are more realistic in that they mimic real users interacting with a real visual IDE frontend. The execution of these workloads can be configured in several ways: for example, to simulate the “think time” of users, the benchmark can add delays between queries triggered by consecutive user interactions. Furthermore, the metrics define a configurable threshold to measure interactiveness of a system using different latency requirements, (e.g. 500 ms) per query. Setting these parameters without knowing the users, the tasks or even the user interface is hard. The shared trace repository proposed in this paper could thus be helpful to derive and set these parameters, and motivate new workloads.

5 THE ROADMAP

In this section, we discuss immediate next steps for the implementation of our repository proposal, and new avenues for future work enabled through this project.

[3] https://idebench.github.io/

Direct Next Steps: Of the challenges discussed in Section 3.3, we believe that addressing the *Heterogeneity of Traces (C1)* and *Tracing levels (C2)* (i.e., trace granularity) to be of the greatest importance. As direct next steps to address these challenges, we plan to start collecting traces from existing commercial systems, such as Tableau, PowerBI, Spotfire. We will also begin the process of soliciting anonymized traces from existing research projects, and potentially start initiatives similar to the First Provenance Challenge [19]. Analysis of these traces will allow us to answer open questions, such as whether there are traces that are not a good fit for our repository. Based on this analysis we also plan to define a common simple abstraction that can be used to represent the traces from different systems. Furthermore, we aim to organize traces to ease the creation of structured evaluation test collections (for example, based on type of task, based on specific datasets). This is in the spirit of what is already done in other research fields like Information Retrieval [14, 28].

Long-term Opportunities: Ultimately, large realistic trace collections can support *meta-analyses of real-world analysis activities* and *realistic cross-comparisons between systems. Existing methodologies (C5)* would be useful frameworks for performing such meta-analyses across different traces and systems. For example, analyzing a large corpus of logs could allow researchers to identify where and how often people perform specific tasks during real analysis practice. Such analyses could then be used, for instance, to derive new and more realistic database benchmark workloads *(C4)* and to help reproduce user studies. Moreover, translating and unifying trace representations from multiple systems could make it possible to quantify which underlying techniques and interactions analysts and users employ, and how these vary across tools. Such analyses will highlight opportunities for optimizing both visualizations and database systems to reflect real-world use.

Collections of shared traces could also provide the basis for *cross-cutting benchmarks* that include all components of a visual analysis system as well as other activities such as model building and data cleaning. Moreover, benchmarks derived from these collections could preserve links back to the original traces, allowing researchers to examine them in order to contextualize, compare, and interpret costly or complex operations.

6 CONCLUSIONS

In this paper, we reported on the ongoing discussion between members of the data management and the visualization community regarding the evaluation of visual data systems. While the initial discussions were focused on exchanging information about how evaluations are conducted in both communities, the group has started to work on a trace repository to enable better evaluation methods. This repository will help researchers to derive more effective metrics and realistic workloads for evaluating visual data systems, as well as to inform the design of new performance benchmarks. More importantly, the repository will foster a broader understanding of how users perform visual data analysis in different contexts, and help to better characterize users, problems, and work environments in an effort to standardize the evaluation process for better reproducibility and comparability across visual data systems.

REFERENCES

[1] R. Amar et al. Low-level components of analytic activity in information visualization. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 111–117. IEEE, 2005.

[2] L. Battle et al. Dynamic Prefetching of Data Tiles for Interactive Visualization. In *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, pages 1363–1375, New York, NY, USA, 2016. ACM.

[3] L. Battle et al. Position statement: The case for a visualization performance benchmark. In *DSIA'18*, 2018.

[4] M. Brehmer et al. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[5] K. Cook et al. The vast challenge: history, scope, and outcomes: An introduction to the special issue. *Information Visualization*, 13(4):301–312, 2014.

[6] W. Dou et al. Recovering Reasoning Processes from User Interactions. *IEEE Computer Graphics and Applications*, 29(3):52–61, May 2009.

[7] P. Eichmann et al. Towards a benchmark for interactive data exploration. *IEEE Data Eng. Bull.*, 39(4):50–61, 2016.

[8] P. Eichmann et al. Idebench: A benchmark for interactive data exploration, 2018.

[9] L. Grammel et al. How Information Visualization Novices Construct Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):943–952, Nov. 2010.

[10] J. Gray. *Benchmark Handbook: For Database and Transaction Processing Systems.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1992.

[11] J. Heer et al. Interactive dynamics for visual analysis. *Commun. ACM*, 55(4):45–54, Apr. 2012.

[12] J. M. Hellerstein et al. Online aggregation. In *SIGMOD '97*, pages 171–182. ACM, 1997.

[13] T. Hey et al. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, October 2009.

[14] C. Initiative. Conference and labs of the evaluation forum, 2018.

[15] P. Isenberg et al. An exploratory study of visual information analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1217–1226, New York, NY, USA, 2008. ACM.

[16] H. Lam et al. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Trans. Vis. Comput. Graphics*, 18(9):1520–1536, 2012.

[17] Z. Liu et al. The effects of interactive latency on exploratory visual analysis. *IEEE transactions on visualization and computer graphics*, 20(12):2122–2131, 2014.

[18] M. Meyer et al. The nested blocks and guidelines model. *Information Visualization*, 14(3):234–249, 2015.

[19] L. Moreau et al. Special issue: The first provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):409–418, Apr. 2008.

[20] T. Munzner. A nested model for visualization design and validation. *IEEE Trans. Vis. Comput. Graphics*, 15(6):921–928, Nov. 2009.

[21] R. Nambiar et al. Transaction Processing Performance Council (TPC): State of the Council 2010. In *Performance Evaluation, Measurement and Characterization of Complex Systems*, Lecture Notes in Computer Science, pages 1–9. Springer, Berlin, Heidelberg, Sept. 2010.

[22] C. Plaisant et al. Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository. *IEEE Trans. Vis. Comput. Graphics*, 14(1):120–134, 2008.

[23] J. Scholtz et al. A reflection on seven years of the vast challenge. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors - Novel Evaluation Methods for Visualization*, BELIV '12, pages 13:1–13:8, New York, NY, USA, 2012. ACM.

[24] B. Shneiderman. Dynamic queries for visual information seeking. *IEEE Softw.*, 11(6):70–77, Nov. 1994.

[25] B. Shneiderman. The eyes have it: a task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

[26] C. D. Stolper et al. Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1653–1662, 2014.

[27] J. Thomas et al. *Illuminating the Path: The Research and Development Agenda for Visual Analytics.* National Visualization and Analytics Ctr, 2005.

[28] E. M. Voorhees et al. *TREC: Experiment and evaluation in information retrieval*, volume 1. MIT press Cambridge, 2005.