Towards Identification and Mitigation of Task-Based Challenges in Comparative Visualization Studies

Aditeya Pandey[*]
Northeastern University

Uzma Haque Syeda[†]
Northeastern University

Michelle A. Borkin[‡]
Northeastern University

ABSTRACT

The effectiveness of a visualization technique is dependent on how well it supports the tasks or goals of an end-user. To measure the effectiveness of a visualization technique, researchers often use a comparative study design. In a comparative study, two or more visualization techniques are compared over a set of tasks and commonly measure human performance in terms of task accuracy and completion time. Despite the critical role of tasks in comparative studies, the current lack of guidance in existing literature on best practices for task selection and communication of research results in evaluation studies is problematic. In this work, we systematically identify and curate the task-based challenges of comparative studies by reviewing existing visualization literature on the topic. Furthermore, for each of the presented challenges we discuss the potential threats to validity for a comparative study. The challenges discussed in this paper are further backed by evidence identified in a detailed review of comparative tree visualization studies. Finally, we recommend best practices from personal experience and the surveyed tree visualization studies to provide guidelines for other researchers to mitigate the challenges.

Index Terms: H.5 [Information Interfaces and Presentation]: User Interfaces—Evalaution/methodology; H.5 [Information Interfaces and Presentation]: User Interfaces—Theory and methods; H.5 [Information Interfaces and Presentation]: User Interfaces—Training, help, and documentation;

1 INTRODUCTION

There are numerous ways to evaluate a visualization design or tool [16]. Out of the large variety of evaluation methods, one of the most common methods employs a comparative evaluation of visualization techniques to measure user performance, e.g., in terms of accuracy and time. Within visualization literature, these studies are identified as "comparative studies" [54] or "head-to-head comparisons" [42]. In visualization literature, comparative studies are used across a wide range of application domains to evaluate a variety of visualization techniques or encodings. For instance, Bartolomeo et al. [21] designed a comparative study to evaluate the effect of different timeline shapes on a subset of timeline tasks (Fig. 1). In another example, Plaisant et al. compared a novel SpaceTree [55] to a traditional node-link tree visualization to study the effect of their novel space-efficient tree visualization layout on a set of tree visualization tasks. A critical aspect of comparative study design is the selection of analytical user tasks that are ultimately used as a proxy to the real domain goals to evaluate the effectiveness of the visualization technique [54]. Given the critical nature of tasks, it is of utmost importance that researchers choose evaluation tasks that reflect the underlying research problem [48, 65].

While tasks play a pivotal role in comparative studies, existing visualization literature argues that the methods used for task selection and communication have several shortcomings. Plaisant [54] argues that the current process of selecting tasks for designing evaluation studies remains an adhoc process. The adhoc nature of the task selection process leads to problems like gathering the wrong task for evaluation [38] or gathering an incomplete set of tasks which do not cover the goals a visualization should support [38, 54, 60]. Another major task-based challenge is associated with the selection of a task abstraction technique. Task abstraction is the process of removing domain-specific terminology from the task description to promote easy understanding and adoption of the task-based results in application domains that are not directly related to the research problem [49]. Task abstraction is limited by ambiguity in choosing the correct abstraction framework [38, 53] and the method used to communicate the abstraction in the research article [38]. The visualization community has recognized several task-based challenges in the context of evaluation studies, however these challenges have not been formalized in the context of comparative studies.

Based on our personal experience with comparative studies (e.g., [11, 12, 21, 52, 64]) and the analysis of challenges from previous research articles [27,38,54,60] we identify four **task-based challenges** (C1-C4) that can directly affect the validity of comparative studies and thus influence the overall adoption and usability of the evaluation results:

- **C1:** Insufficient Justification of Task Source
- **C2:** Missing or Incomplete Task Abstraction
- **C3:** Inconsistent Task Description Format
- **C4:** Knowledge Gap in Task-Based Evaluations

In this paper, for each challenge we discuss its effect on the validity of different stages of a comparative study. Moreover, to investigate if these challenges exist in published academic literature, we survey tree visualization comparative studies and analyze 20 studies to identify if they are threatened by these challenges. Our analysis of the surveyed papers show that many studies have insufficient task justifications and have missing or incomplete task abstractions. Overall, we also found that task descriptions in the existing studies are inconsistent and researchers focus on a limited subset of analytical tasks within the possible task design space of tree visualizations. In the paper's challenges section (Sec. 5), we provide quantitative evidence for the existing challenges in the tree visualization studies. In addition to highlighting the challenges, we also offer practical recommendations to visualization researchers on how to identify the task-based challenges in their comparative studies and mitigate them.

Through this *position* paper, we hope to draw the attention of the visualization community to a potentially problematic component of comparative studies. Through the BELIV workshop, we hope to initiate further discussion on this topic and work towards collectively identifying methods that can help our community resolve the task-based challenges in comparative evaluation studies.

[*] e-mail: pandey.ad@northeastern.edu
[†] e-mail:syeda.u@northeastern.edu
[‡] e-mail:m.borkin@northeastern.edu

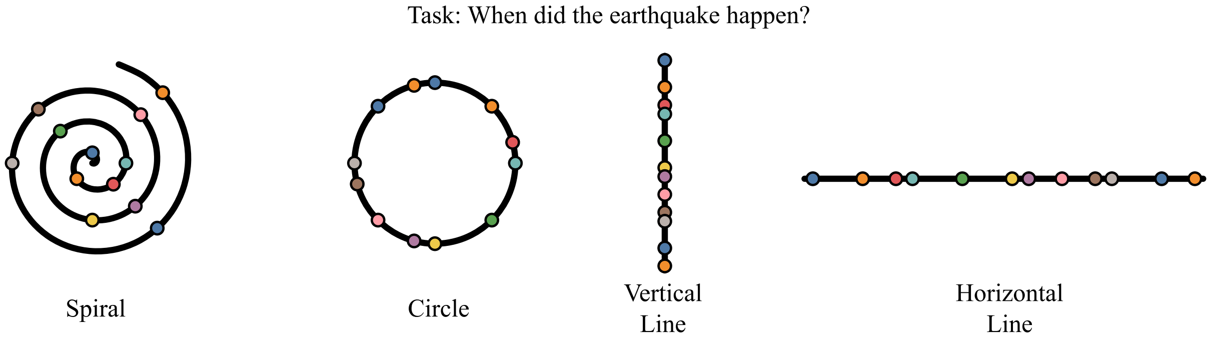


Figure 1: In a comparative study, Bartolomeo et al. [21] measured the effect of timeline shape on accuracy and time of user to find an event ("earthquake") on the timeline. The timelines shown in this figure are an abstract representation of the stimuli used in the original evaluation study.

Contributions: In this paper, we identify and present task-based challenges that exist in comparative studies. We provide evidence that existing problems manifest in published tree visualization studies by conducting a comprehensive survey of comparative studies for tree visualizations. Finally, we recommend guidelines that can assist researchers in mitigating the task-based challenges in their study.

2 Related Work

2.1 Review of Task-Based Challenges

Many researchers have raised issues and concerns within the information visualization community regarding how tasks are selected and communicated in the literature. In this section, we present an overview of the task-based challenges we found in the visualization literature.

Problems with existing task selection process: In a review of visualization evaluation challenges, Plaisant [54] highlights that the current method of task selection is ad-hoc, which can result in the collection of the wrong or incomplete tasks for an evaluation study. Saket et al. [60] also discuss the problem of limited tasks being used in existing visualization studies. In their review of existing evaluation studies, Saket et al. found that most papers only use a subset of tasks to benchmark the task-based effectiveness of basic information visualizations such as table, line chart, bar chart, scatterplot, and pie chart. They further conclude that due to the limited scope of the evaluation studies their results are hard to generalize for future work like curating task-based guidelines for visualizations. Kerracher & Kennedy [38] also discuss the challenges of task selection but in the context of task abstraction framework construction and validation. Their work identifies the threats associated with different task collection methods. For instance, if researchers use user interviews to collect tasks, they may encounter problems like identify the right people to interview or using the best interview method.

Problems with task description and explanation: Another challenge associated with tasks in information visualization is related to their description and explanation. In a review of information visualization evaluation challenges, Carpendale [16] argues that if visualization tasks are not defined clearly, they can be hard to test empirically and fail to provide insight into the usability of the visualization technique. In another study, Ziemkiewicz & Kosara [73] find that inconsistent phrasing of concepts in task descriptions lead to varying comparative results in the evaluation studies. Yet another problem with task description can be the lack of abstraction of the visualization tasks. Due to the lack of abstraction, Plaisant [54] reports that it is difficult to compare systems even with given tasks and datasets while discussing the main findings of the InfoVis contest [35] that was established to initiate and encourage evaluation benchmarks and methods. Since, comparative studies can span across different domains and are not tied to a single one, it becomes all the more important for tasks to be expressed in an abstract form using a standard language in order to enable comparisons across different domains.

To summarize, information visualization literature recognizes that task-based research presents several challenges. In this work, we systematically organize these task-based challenges for comparative studies.

2.2 Task Abstraction and Task Space

Abstract tasks are domain and interface independent operations carried out by the user of a visualization [48]. It is rather difficult to compare different visualization tasks based on their domain-specific languages, even though the underlying abstract tasks may be exactly the same. Once the domain language is stripped away and the tasks are translated into an abstract and consistent information visualization language, the similarities between the tasks become apparent [49]. For example, even though the tasks "Which directory has more files directly inside: '/hcil/spotfire' or '/hcil/spacetree'?" [7], and "Is 'Interest on the Public Debt' more than half of the 'Department of Treasury'?" [37] may look different, the actions to accomplish both tasks are the same. If expressed in abstract task terminology, both of the tasks are the same: "**Lookup** the nodes to **compare** degree of the topology".

Presenting tasks in an abstract and consistent manner or providing task abstractions along with domain specific tasks offer various advantages to visualization evaluations and design [38]. It allows evaluators to evaluate tools in comparative studies more efficiently [3, 5, 43, 63], enables better communication between researchers [58, 63], and most importantly provides a common vocabulary for analytical task descriptions [56]. There are many task abstraction frameworks (e.g., [4, 13, 49, 66]) but the level of granularity in task abstraction varies greatly within the visualization literature [13]. For example, Amar et al. [4] provides ten low-level analytic tasks that encompass a user's activity while using visualization systems in order to understand data. Shneiderman [66] presents a task taxonomy on seven tasks based on seven different data types. In contrast, Brehmer & Munzner [13] offers a multi-level task typology that differentiates why and how a task is carried out and what are its inputs and outputs. The typology breaks down visualization tasks into abstract and interdependent high-, mid-, and low-level actions that the user performs to carry out a specific task.

Although task abstraction is necessary to facilitate fair comparisons across different domains, it can only be done once the domain tasks to be performed are selected or generated. In order to do so, it is important to consider the task space, i.e., the list of all possible tasks that can be performed using the visualization tool(s). Schulz et al. [63] defines a design space that distinguishes five aspects of visualization tasks and consolidates the vast amount of task taxonomies, classifications, and frameworks that are scattered across the visualization literature in the forms of lists, descriptions, mathematical task models, domain-specific models of tasks, and workflows derived from task procedures. The task design space defines 5 dimensions: i) why a task is performed, ii) what type of data does the task need, iii) who is performing the task, iv) when is the task performed and in what order and, v) how is the task performed in terms of actions. The authors also put their design space for task into practice by applying it in climate impact research. Amongst some of the practices of task generation are deriving from literature [71,72] or the author's own knowledge, interviews with domain experts [14,21,46,57,62] and reviewing existing systems for the tasks they support [17,26]. For a full list of the task generation techniques and the threats they pose, please refer to Schulz et al. [63].

We have personally experienced the drawbacks of the above mentioned ways of task generation in some of our previous works. Extracting tasks from literature and existing systems is hard because of the lack of task abstraction and consistency in description. Also relying on the author's knowledge greatly increases the risk of missing out on important tasks, specially in the case of comparative evaluations where multiple domains can be at play. We adopted the method of interviewing domain experts in our previous work [11,12,21,52,64] and while this is an improvement from not justifying and validating the tasks at all, it also has drawbacks. For example, the domain expert's availability might be limited [23,59]. Also, interviewing an expert from a single domain might run the risk of introducing bias in selecting representative tasks [39]. This is because while human-computer interaction (HCI) studies that involve usability testing of a tool with experts as participants indicate that three to five evaluators will suffice for usability testings depending on the particular problem [51,67], there is no agreed upon rule on how many expert interviews is sufficient for obtaining a representative task list. This is an interesting research question that needs further investigation.

2.3 Comparative Evaluations in Visualizations

Data visualizations are created based on varying goals and strategies and hence the type of evaluation in each case would be expected to vary as per the aim of the researchers. Although controlled experiments and usability studies are the cornerstones of visualization evaluation [54], there are many other diverse metrics by which evaluations have been classified. Andrews [6], Ellis et al. [22], and Hilbert et al. [30] classify evaluations based on research goals such as summarizing the efficacy of an interface (summative), providing design guidelines (formative), comparing design alternatives (predictive), and understanding user behaviour, performance, thoughts, and experiences (observational and participative). Some make the distinction based on whether the data collected is qualitative, quantitative, or mixed and whether it is collected empirically or analytically [8,19]. Researchers also differentiate evaluation methodologies based on research strategies [45], methods [34], and evaluation scope [18]. Munzner [48] breaks down evaluations based on the corresponding design stages of visualization development. Plaisant [54] surveyed approximately 50 information visualization user studies and derives four different high-level themes of evaluation: i) Comparison of design elements through controlled experiments [1,11,31], ii) usability evaluation of a visualization tool [15,69], iii) controlled experiments to compare two or more tools [55], and iv) case studies of tools in realistic settings [70]. In our work, we focus on the thematic classification by Plaisant [54], specifically on themes (i) and (iii) pertaining to comparative evaluations between visualization tools or elements of design within different visualization tools or techniques.

We narrow down our focus to comparative visualization studies because the scope of task-based challenges in visualization evaluation is vast and beyond the scope of this paper. Moreover, comparative studies are one of the most common forms of evaluation methodology in visualization literature. In Lam et al.'s [42] meta analysis of 850 papers, they found that investigating user performances (UP) and user experiences (UE) were the most common themes of evaluations. Isenberg et al. also reported UE and UP to be the dominating evaluation scenarios [33]. Comparative evaluations can encompass all of these scenarios, and are mainly used to measure UP and UE [2,11,32,47]. Finally, the tasks of comparative studies can span across different domains thus task abstraction is of paramount importance for study comparison. Therefore the focused systematic review of task-based challenges in comparative visualization studies in this paper is particularly important in order to make the comparisons fair, generalizable, and independent of domain expertise.

3 A Preliminary Life cycle of an Evaluation Project in Visualization

In this section, we present an overview of the basic steps in an evaluation project's life cycle as shown in Fig. 2. *A full literature review to develop a precise description of the life cycle of an evaluation project is beyond the scope of this paper. In this section we summarize the process in order to enable an effective discussion of the challenges (Sec. 5) in the context of the different steps a researcher goes through to conduct and publish an evaluation.*

Step 1: Formulate Evaluation Questions and Hypothesis Evaluation projects are motivated by a research question that is generally a conjecture or hypothesis and needs further evidence to make any conclusive scientific claims. An example of a research question can be: "Is a node-link based tree visualization techniques more effective in topology-based tasks than the treemap?" Therefore, in this step, researchers carefully and exhaustively determine the research question's importance. The research question's influence can be heuristically estimated through methods like expert interviews or reviews of relevant literature. The output of this phase should usually be a clear, testable hypothesis.

Step 2: Identify Experiment Variables In this step, researchers objectively identify the variables that will affect the hypothesis testing. In quantitative evaluations, this step will include identification of independent and dependent variables [16]. For the independent variables, researchers should decide the strategies they use to control the effect of independent variables. The dependent variables correspond to measures that will ultimately be used for analysis like task accuracy or task completion times. Further, the researchers should also determine the statistical methods they will use to analyze the results. Ideally the variables along with statistical methods should be preregistered on an open science forum such as OSF [24,28].

Step 3: Design the Study In this step of the evaluation project, researchers determine the practical aspects of the evaluation study. Here, the researchers should determine the hardware, apparatus, and experimental set-up to use (e.g., crowdsourced versus lab experiment), identify and recruit the participants of the study, design the stimuli to be used in the study, formulate the real world analytical tasks to test the hypothesis, implement the study and figure out the order of tasks and experimental conditions, and set up the methods to collect the data. After the design plan is complete, it should be submitted to IRB for approval before collecting data and analyzing results.

Step 4: Analyze the Results The data collected can then be analyzed through the application of appropriate statistics. The statistical methods should be in sync with Step 2.

Step 5: Discuss the Results In information visualization, the statistical results are often supplemented with justification or observations from the researcher [29,36]. This step aims to enable researchers to discuss the findings more informally and share their opinion about the overall experiment. Researchers who conducted the research have the first-hand experience of some exciting artifacts which are not possible to capture in the statistical results like the

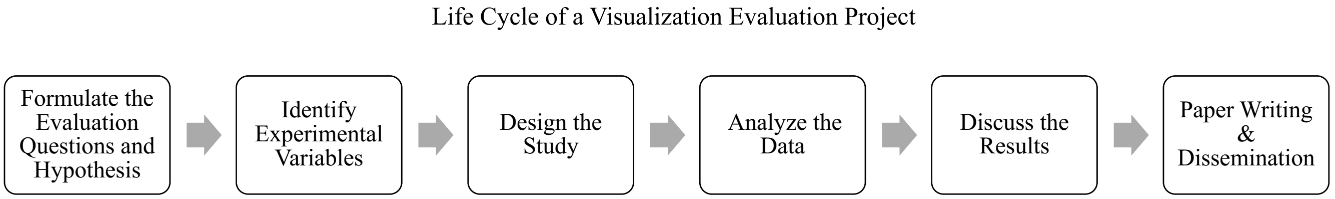

Life Cycle of a Visualization Evaluation Project
Formulate the Evaluation Questions and Hypothesis
Identify Experimental Variables
Design the Study
Analyze the Data
Discuss the Results
Paper Writing & Dissemination

Figure 2: This figure presents an overview of an evaluation project. The evaluation project begins with research question selection, progresses with evaluation design steps, and deals with the collection and analysis of data and terminates with researchers presenting the results in the form of research articles. We discuss these steps in the paper to set the background for discussion of the task-based challenges.

p-value statistics. Therefore, this step allows researchers to discuss interesting insights that might benefit the visualization community.

Step 6: Paper Writing and Dissemination Despite the outcome of the data analysis phase, the results from the experiment should be communicated to other researchers. Strong statistical evidence and other observational evidence will allow future researchers to build guidelines and theories that will benefit the entire visualization community. In some cases, it is possible that the results will not show clear trends or outcomes. Even then, it is a good practice to communicate the experiment procedure and results accurately. These results will help future researchers assess the shortcomings of the setup or prescribe future researchers to invest time in other research questions. The data and code of the experiment should be made accessible to support reproducibility of the experiment, for more information please refer to the article by Steve Haroz [28]

4 Identification of Challenges and Tree Evaluation Survey Methodology

The task-based challenges discussed in Sec. 5 were collected through a grounded theory approach [19]. The first author reviewed the visualization literature discussed in Sec. 2.1 and developed a list of task-based challenges. After this, all the authors examined the challenges and, with the backdrop of their personal experience and observations while designing comparative studies, evaluated the validity of the challenges. After a few rounds of iterations, all the authors agreed upon the four challenges discussed in Sec. 5.

As the challenges were gathered through a qualitative research method and based on the authors' experience, we found it extremely important to validate the challenges with existing comparative studies. Comparative studies are a ubiquitous evaluation method in visualization literature. Therefore an exhaustive survey of all published comparative studies was beyond the scope of this work. However, if we scope the selection of a study belonging to a particular data type, then we realized that the survey is feasible and within the scope of this paper. Since all the authors had prior research experience with tree visualizations, we chose to survey comparative studies in the tree visualization literature.

Survey Methodology: We searched for tree visualization comparative studies in the three most common digital libraries that publish the visualization research papers: IEEE Xplore, ACM Digital Library, and Eurographics Digital Library. In addition to the digital libraries, we searched for articles on Google Scholar because there is a chance that the digital libraries may not contain articles that may be relevant for the survey. On the IEEE Xplore library, we searched for the term: "Tree Visualization Evaluation". Further, we were able to filter articles only published at visualization conferences and journals, including IEEE Transactions on Visualization and Computer Graphics (TVCG), IEEE VIS, IEEE VAST, IEEE Pacific Vis, and Information Visualization(SAGE Journals). After applying the filter we found **30** papers on the IEEE Xplore library. Unlike IEEE Xplore, ACM Digital Library and Eurographics Digital Library do not have a feature to select the conference and journals. Therefore, we only selected papers with all the terms "tree", "visualization," and "evaluation" in the title. On applying these filters, we found **1** paper on ACM Digital Library and **1** paper on Eurographics Digital Library. On Google Scholar, the search term returned 770,000 results. In Scholar, the results are presented in order of relevance to the search criteria. We skimmed the first seven pages of search results (each page returns 10 papers) and determined that the first two pages had 50% relevant papers to our survey, but by the seventh page, most articles were irrelevant. However, for consistency across search phrases, with chose to include the first **100** papers (10 pages of returned search results) to ensure we captured all relevant material. From a corpus of **132** articles, we skimmed and filtered all the articles that were not comparative evaluation studies. In the end, we had a total of **20** articles that met our search criteria, i.e., they were comparative evaluation studies of tree visualizations. We provide the final list of the articles in the Supplemental Material.

5 Task-Based Challenges

Based on our analysis of the existing visualization literature and our prior research experience, we found four task-based challenges associated with comparative studies. Each challenge poses a threat to the validity and usability of visualization evaluation results. In this section, we expand on the challenges to promote awareness about these challenges to the community and offer concrete recommendations to mitigate the challenges. To ensure that we use a consistent clear method to communicate the challenges, we introduce a "challenge reporting" template:

Challenge Reporting Template:

Description: A subjective explanation of the problem.

Cause: Identification of the probable factors that contribute to the challenge. Identification of causes is essential because it directly affects how we build recommendations or guidelines to mitigate the problem.

Effect: Identification of the steps in the life cycle of an evaluation project (Sec. 3) that may be affected by this challenge and discuss them in detail. This effect discussion is important as each challenge has the potential to invalidate an evaluation project.

Evidence: Presentation of evidence from the surveyed tree visualization comparative studies to support the challenge. For an objective analysis of the problem, we quantify the evidence using measures that can generalize to future surveys.

Guidelines: A proposed set of guidelines or best practices corresponding to each challenge that we argue may help mitigate the effect of the problem. These guidelines were extrapolated from our review of related works. We divide the guidelines into two categories: "Researchers" and "Community" based on the intended target audience for the guideline.

5.1 C1: Insufficient Justification of Task Source

Description: Comparative studies may fail to describe the source of the tasks and provide an explicit rationale for selecting the tasks. The tasks used in a comparative study represent the type of questions or queries researchers want the user to answer with a visualization. For instance, through the task "Which directory includes a deeper hierarchy: 'Flutes' or 'Guitars'?" [40] the researchers are trying to compare four tree visualization techniques on their capability to display the depth or height of a tree dataset. The tasks used in these studies are critical to the design of a study because an incorrect selection of tasks may discard the entire evaluation. Given the essential nature of tasks, we assumed that researchers would explicitly discuss the source and rationale of the included tasks in the research articles they write to summarize the experimental procedure and results. However, for the tree papers we reviewed, we found a number of papers (discussed later in the Evidence section) that did not have an explicit rationale for the tasks included in their study. Some papers did not even have a clear description of the source of the tasks.

Cause: Existing literature effectively guides how to report experimental methods and results of an evaluation study [20, 25], but they do not focus on the specifics of how to communicate the source and rationale of tasks used in a study. The primary cause of this problem may be a lack of a clear understanding of what constitutes a source and rationale for the tasks. In the guidelines section of this challenge, we present a list of task sources and provide some role model examples of how to describe the rationale for choosing a set of tasks.

Effect: Insufficient justification of the tasks can cause a problem in the very early stages of an evaluation project. If the researcher does not provide the fundamental reason for the task source, then all the steps after **Stage 3: Design the Study** (Fig. 2) become irrelevant because other researchers will not accept or use the results. Therefore, this challenge may have the most severe consequence and should be dealt with appropriately.

Evidence: Compared to the other parts of the methodology section like the evaluation procedure and data analysis, we noticed that task source and rationale were under specified in the reviewed tree visualization papers. Out of the 20 tree visualization papers, 50% did not mention a clear source for their tasks, and a staggering 70% of these did not provide a clear rationale behind the selection of tasks. For instance, Kosba et al. [40] describe the source as *"The task selection was also informed by a very early version of the InfoVis 2003 contest tasks. In some cases, questions had to be rephrased using a more technical terminology in order to make them unambiguous."* However, the paper did not discuss aspects such as "How did they rephrase the tasks?" and "What ambiguity made them to do it?" In another example, Barlow et al. [9] mention that *"The tasks used to evaluate the compact views were based on the requirements of the user in a data mining context."* However, they do not discuss the users nor do they reveal the methodology used for collection of the tasks. These examples demonstrate how the task source is underspecified and how many tasks are included in the studies without proper rationale. We provide more details about other papers in the Supplemental Material.

Guidelines for Researchers: To mitigate this challenge, we recommend researchers mention and describe the source of the task explicitly in their published papers. Kerracher & Kennedy [38] identify seven categories of task sources: derive from literature, interview with domain experts, a survey of visualization experts, observational strategies, system review, author's knowledge, and derive from existing frameworks. All the above categories should be considered acceptable as long as researchers are clear about the motivation and reason behind choosing one source over the other. In addition to the source, researchers should also provide the rationale for selecting the set of tasks. The rationale is usually a justification that clarifies and supports the decision to choose a set of tasks. For instance, Santos et al. based their task selection on the fact that their visualization was designed to evaluate 3D information. The task selection was expected to test the tree visualization techniques that support 3D information display [61].

Guidelines for Community: Identification of tasks for evaluations can be challenging for researchers due to the lack of availability of domain or field experts to interview, or lack of understanding of the literature to correctly identify analytical tasks through survey methods [38]. To resolve this problem, Plaisant [54] proposes the creation of **task datasets**. The task dataset is an exhaustive collection of visualization tasks which can be built by a literature survey or collaborative data collection methods. Such datasets can be used as a verified resource for researchers to recognize tasks for a visualization evaluation.

5.2 C2: Missing or Incomplete Task Abstraction

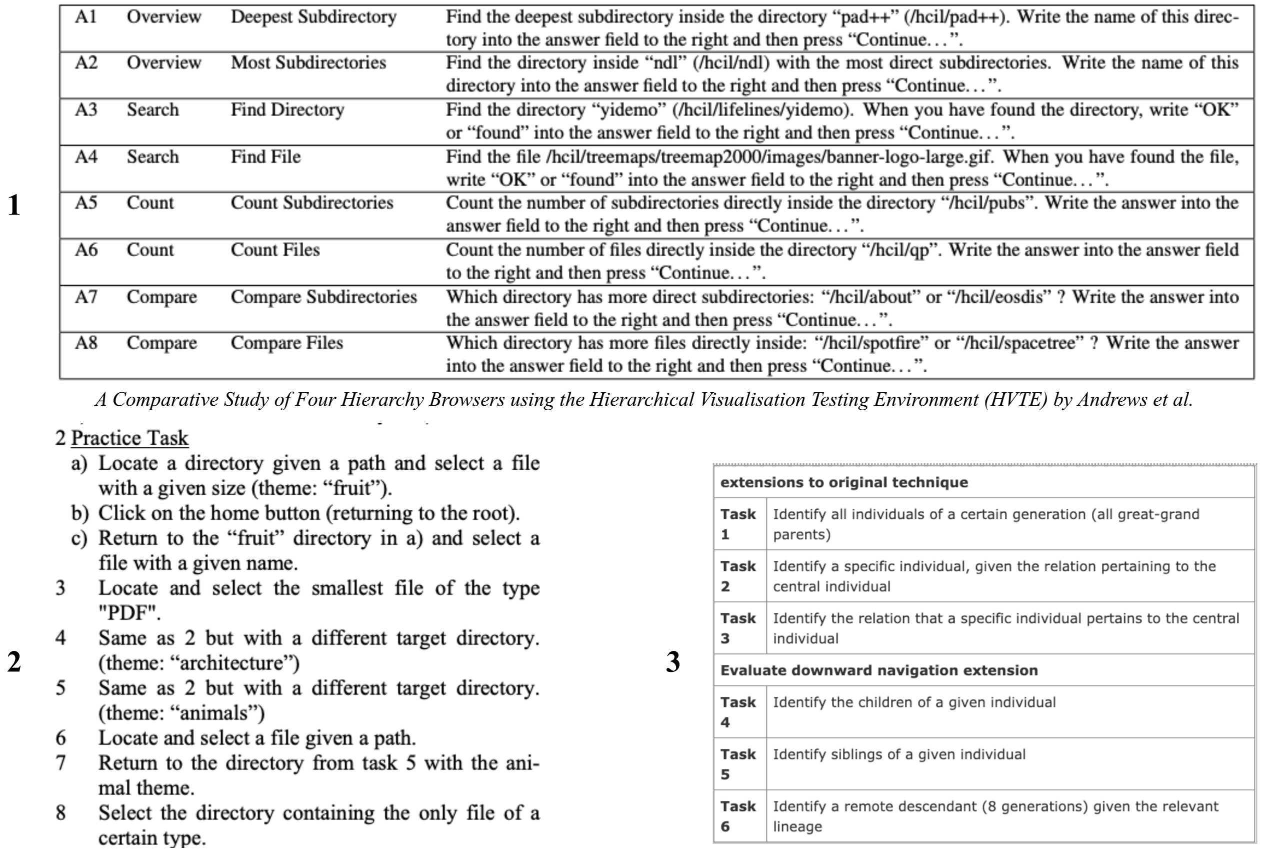
Description: Comparative studies may fail to abstractly represent a visualization task that includes domain or dataset-specific terminology. Evaluation tasks should be simple and resemble real-world problems [54]. Consequently, in the tree visualization survey we observed the tasks reflect the properties of the underlying dataset to resemble familiarity with the problem. However, the usage of dataset-specific terminology can often be misleading, and researchers may fail to notice a similarity between two tasks. Let us consider tasks **T1**: "Compare the directories '/projects/ravon/control/' and '/projects/ravon/navigator/'. Which one is the larger one?" [50] and **T2**: "Which directory has more direct sub-directories: '/hcil/about' or '/hcil/eosdis'?" [7]. T1 and T2 were used in two different studies, therefore the language used in the tasks is different. However, in both T1 and T2, researchers want the users to compare the size of two sub-directories, which is easy to miss if the tasks are not abstracted. In many research papers, we noticed that the dataset terminology influenced the evaluation study tasks, but the researchers did not present an abstract description of the dataset. We argue that this challenge reduces the transparency of a task in the evaluation study and leaves the responsibility of abstraction on the person who is reading the paper in the future.

Cause: Information visualization provides many theoretical task abstraction frameworks but lacks guidance on how to use these frameworks in practice (see Sec. 2.2). The availability of several competing frameworks with limited guidance on how to use them makes it hard for researchers to choose the right abstraction method. Moreover, the literature also lacks clear consensus on the task abstraction specificity. For instance, in Fig. 3 (1) we can notice that authors provide a group level and task level abstraction in their study, i.e., for the first "Overview" is the group-level abstraction and "Deepest Subdirectory" is the task-level abstraction. However Fig. 3 (3) provides only a group-level abstraction without "Evaluate downward navigation extension". The combined effect of multiple abstraction frameworks and lack of guidance on how and what to abstract may be the main reasons for this challenge.

Effect: Missing or incomplete task abstraction can affect an evaluation project at multiple stages. At **Stage 3: Design the Study**, without an abstraction researchers may use redundant identical tasks in the experiment. Abstraction will ensure that researchers add a variety of tasks to the evaluation. At **Stage 4: Analyze the Data** the lack of abstraction may miss the opportunity to analyze trends in the results. For instance, in tree visualizations, two common types of targets are "Topology" and "Attribute" [49]. If the researchers know the task abstraction, then they can analyze trends between topology and attribute tasks. For example, a Node-Link encoding for tree visualization is more efficient with topology tasks, but a Treemap is more efficient with attribute tasks. At **Stage 6: Paper Writing and Dissemination** without the task abstraction, other researchers may fail to analyze the results and reduce the usability of the evaluation results.

Evidence: For each tree visualization paper we surveyed, we reviewed the paper in detail and categorized it into one of three categories of abstraction:

- No Task Abstraction: The paper does not include abstract information with the tasks.



1

A1	Overview	Deepest Subdirectory	Find the deepest subdirectory inside the directory "pad++" (/hcil/pad++). Write the name of this directory into the answer field to the right and then press "Continue…".
A2	Overview	Most Subdirectories	Find the directory inside "ndl" (/hcil/ndl) with the most direct subdirectories. Write the name of this directory into the answer field to the right and then press "Continue…".
A3	Search	Find Directory	Find the directory "yidemo" (/hcil/lifelines/yidemo). When you have found the directory, write "OK" or "found" into the answer field to the right and then press "Continue…".
A4	Search	Find File	Find the file /hcil/treemaps/treemap2000/images/banner-logo-large.gif. When you have found the file, write "OK" or "found" into the answer field to the right and then press "Continue…".
A5	Count	Count Subdirectories	Count the number of subdirectories directly inside the directory "/hcil/pubs". Write the answer into the answer field to the right and then press "Continue…".
A6	Count	Count Files	Count the number of files directly inside the directory "/hcil/qp". Write the answer into the answer field to the right and then press "Continue…".
A7	Compare	Compare Subdirectories	Which directory has more direct subdirectories: "/hcil/about" or "/hcil/eosdis" ? Write the answer into the answer field to the right and then press "Continue…".
A8	Compare	Compare Files	Which directory has more files directly inside: "/hcil/spotfire" or "/hcil/spacetree" ? Write the answer into the answer field to the right and then press "Continue…".

A Comparative Study of Four Hierarchy Browsers using the Hierarchical Visualisation Testing Environment (HVTE) by Andrews et al.

2

2 Practice Task
a) Locate a directory given a path and select a file with a given size (theme: "fruit").
b) Click on the home button (returning to the root).
c) Return to the "fruit" directory in a) and select a file with a given name.
3 Locate and select the smallest file of the type "PDF".
4 Same as 2 but with a different target directory. (theme: "architecture")
5 Same as 2 but with a different target directory. (theme: "animals")
6 Locate and select a file given a path.
7 Return to the directory from task 5 with the animal theme.
8 Select the directory containing the only file of a certain type.

The effect of animated transitions on user navigation in 3D tree-maps by Bladh et al.

3

extensions to original technique	
Task 1	Identify all individuals of a certain generation (all great-grand parents)
Task 2	Identify a specific individual, given the relation pertaining to the central individual
Task 3	Identify the relation that a specific individual pertains to the central individual
Evaluate downward navigation extension	
Task 4	Identify the children of a given individual
Task 5	Identify siblings of a given individual
Task 6	Identify a remote descendant (8 generations) given the relevant lineage

Extending the H-Tree Layout Pedigree: An Evaluation by Santos et al.

Figure 3: The figure shows variation in task description format used in tree visualization comparative studies. This figure highlights that there is a lack of general technique to report the tasks used in comparative studies. The most specific description is shown in the top figure (1), where the tasks are presented with abstraction and instructions [7]. The task descriptions shown in bottom row (2 & 3) do not have instructions for the tasks [10, 61], that may hinder its replication by other researchers.

- Group-Level Task Abstraction: The paper does not abstract all tasks, but provided a group-level abstraction.
- Individual Task Abstraction: The paper contains abstractions for each task in the study.

We present the categorization results in Fig. 4 (Abstraction Distribution). The results demonstrate that 30% of the papers were missing task abstraction, while 30% only provided partial group-level task abstraction.

Guidelines for Researchers: To mitigate the challenges, we recommend researchers to adopt a suitable task abstraction framework. There is a lack of conclusive proof in the existing visualization literature that will support one abstraction framework over the other. Therefore, we argue the choice of abstraction framework should rest with the researcher. However, the researcher should justify the reason for adopting the particular task abstraction framework. Additionally, we argue that researchers should abstract each task they use in the comparative study if the task's terminology consists of domain or dataset-specific references. We believe task level abstraction will not just assist other researchers in understanding the tasks but also prove to help eliminate redundant tasks in the study design.

Guidelines for Community: This challenge also raises a broader question on the usage and evaluation of extant task classification frameworks. Kerracher & Kennedy [38] summarize that the adoption, evolution, and demise of task classifications "in the wild" may provide valuable information about their descriptive abilities, comprehensiveness, usefulness, and usability. The visualization community needs to conduct such "in the wild" studies and develop guidelines that will assist researchers in choosing the right task abstraction frameworks to use under different conditions.

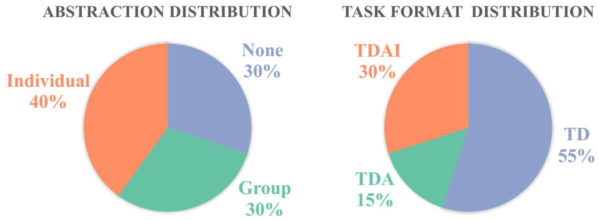


Figure 4: Abstraction Distribution corresponding to *C2: Missing or Incomplete Task Abstraction* shows the variation in task abstraction techniques. 70% of the papers had none or only group-level abstraction. Task format Distribution corresponding to *C3: Inconsistent Task Description Format* shows the variation in task description format. 55% of the papers did not present instructions and abstractions, along with the task description.

5.3 C3: Inconsistent Task Description Format

Description: Comparative studies do not have a standard method for how to frame the task used in an evaluation study which leads to inconsistent task descriptions across studies. In our survey of comparative evaluations of tree visualizations, all papers contain an explicit list of tasks in the body of the paper. We found that research papers often use more than one task to evaluate the visualization's performance in terms of task accuracy and completion times. However, we noticed that the method of task presentation was inconsistent across papers. Tasks descriptions used in three different papers are shown in Fig. 3 (1-3). In Fig. 3 (1), the authors also present the abstract categorization of the task and the instructions they presented to the participants in the study [7]. However, in Fig. 3 (2 & 3) the authors present only the task that they used [10, 61]. In a published paper the task description and experimental instructions are important to ensure that other researchers can replicate the study [25]. The lack of specificity and supplemental information about the tasks used in the comparative study may inhibit the study's accurate replication.

Cause: The inconsistency in task description may be attributed to the lack of proper information on task-reporting best practices within the information visualization community. We reviewed BELIV 2018 papers where replication and reproducibility were a common topic of discussion among the published papers [41, 44, 68]. However, in these papers, we did not find a thorough analysis or discussion of how task phrasing and description can play a role in experiment replication. This lack of acknowledgement by the community on the importance of task phrasing and reporting for study validity and replication, and lack of supporting guidelines on how to consistently report tasks, has caused this challenge.

Effect: Inconsistent task description in existing research papers can confuse researchers designing studies at **Stage 3: Design the Study**. Researchers may not find a precise method of how to phrase tasks and present it to their participants. This problem will further be reflected when the researchers write the paper (**Stage 6: Paper Writing and Dissemination**). This may inhibit future researchers from understanding the evaluation design.

Evidence: For each tree visualization paper we surveyed, we reviewed the paper content and binned them into three categories based on the task description format in the paper:

- Task Description (TD): The explicit task list only communicated the tasks used in the study e.g., Fig. 3 (2).
- Tasks with Abstraction (TDA): The papers that included abstraction along with the task description e.g., Fig. 3 (3).
- Task with Abstraction and Instructions (TDAI): The papers that included both abstraction and task-wise instructions in the task description e.g., Fig. 3 (1).

It is important to note that some papers may provide an overview of the task instructions in the paper but the information is too disconnected from the task description. As the point of this challenge is to facilitate communication of the tasks, we argue that the TDAI method is most suitable for task communication because it bundles all the important aspects of the tasks together and facilitates readers to find the information more conveniently. The tagging results of this challenge in Fig. 4 (Task Format Distribution) shows that 30% (TDAI) of the papers are already using descriptive task information. But 70% (TD and TDA) miss out on the opportunity to provide their readers detailed information about the tasks in an objective and easy to parse manner.

Guidelines For Researcher: We recommend researchers to use a task description format similar to the one shown in Fig. 3 (1) because the form is most specific and leaves little or no room for speculation about how the tasks were conducted and what was the abstract intention of the researcher behind including the task.

Guidelines for Community: Consistency problems in the task description may include sub-problems such as the consistency of the task phrasing in terms of words from conventional taxonomies, e.g., "identify" and "summarize", or identify the level of detail which tasks should be displayed to study participants and other researchers. The visualization community should recognize the low-level challenges that inhibit creating a standardized task-description format across comparative studies and suggest guidelines and methods to eliminate these challenges.

Topology Query Tasks

	Action(Search)			
Action(Query)	**Lookup**	**Locate**	**Browse**	**Explore**
Identify	18	2	2	2
Compare	7		2	1
Summarize	2		4	

Attribute Query Tasks

	Action(Search)			
Action(Query)	**Lookup**	**Locate**	**Browse**	**Explore**
Identify	13	10	9	10
Compare	8		2	6
Summarize				1

Count of Tasks

1 ▨▨▨▨▨▨▨▨▨ 18

Figure 5: Summary results of the task abstraction of 99 analytical tasks collected from the tree visualization papers. We break up the results for "Topology" and "Attribute" targets. The count in each cell of the tables corresponds to total number of tasks.

5.4 C4: Knowledge Gap in Task-Based Evaluations

Description: Existing comparative studies evaluate only a subset of the task design space. As discussed in Sec. 2.2, Schulz et al. argues that, similar to the visualization design space, tasks also have a design space, i.e., all possible combinations of analytical queries a user can perform with a visualization. However it has been observed in previous work that visualization evaluations often evaluate a subset of tasks [54, 60]. The issue of an evaluation limited in its coverage of the task-design space is common in tree visualization. The limited coverage of tasks in comparative studies creates an imbalance in the knowledge of the visualization technique. According to Plaisant [54], the effectiveness of a visualization should not be based on a task but rather depend on how well the visualization performs on all the relevant tasks. Through the case of tree visualization, we introduce a method to identify the task-design space by using a task-abstraction framework and present tools that will assist researchers in communicating exhaustiveness of their task-based comparative study.

Cause: A primary reason for this challenge is the lack of adoption of the concept of a design space for tasks [63]. The design space of tasks allows researchers to enumerate all the possible combinations of tasks that a user may want to perform with a visualization tool and may support easy identification of tasks that have previously been evaluated in studies.

Effect: This challenge directly affects **Stage 3: Design the Study** and **Stage 6: Paper Writing and Dissemination**. In Stage 3, if researchers are not adequately aware of the design space of tasks, they might fail to identify tasks that were necessary for evaluation but missed them due to an error in the task selection method. In Stage 6, due to this challenge, the authors fail to communicate the evaluated task design space and open research areas. This shortcoming in communication can inhibit researchers interested in expanding task-based knowledge about visualizations through future work. This challenge may lead to future researchers conducting redundant studies.

Evidence: We collected all the tasks used in the surveyed tree visualization papers and abstracted the tasks using the Multi-Level Task Typology (MLTT) Framework [13]. We use this framework because it allows us to specify actions and targets of the analytical task, thus providing more detailed insight into both the intention (action) of the user and the item of interest (target). After abstraction, we reviewed coverage of the task design space for the target as shown in Fig. 5. We analyze "Topology" and "Attribute" separately. From these results we observe that some aspects of the tree visualization tasks are more thoroughly evaluated than others. Studies are more inclined towards *"Identify"* and *"Lookup"* tasks. While some other tasks like *"Summarize"* are not well evaluated, researchers and practitioners have little guidance on how tree visualizations support the *"Summarize"* tasks.

Guidelines for Researchers: Researchers should explicitly communicate and justify the evaluated design space. To determine the task design space, a user should enumerate the possible combinations of analytical tasks a user can perform within an abstraction framework. For instance, in Fig. 5 we can see that a user can perform four search tasks (Lookup, Locate, Browse, Explore) and three query tasks (Identify, Compare and Summarize). Therefore, for the Multi-Level Task Typology [13], the task design space consists of 12 ($4 * 3$) analytical tasks. After enumerating the task design space, researchers can objectively communicate the tasks they have evaluated, as we have done in Fig. 5. As discussed in the evidence paragraph, the figure allows other researchers to see the tasks being assessed and open areas of research that the study did not evaluate.

Guidelines for Community: This challenge also presents the need for a task dataset. A task dataset, as discussed earlier in Challenge 1, is an exhaustive collection of visualization tasks. For example, a task dataset could be focused on a specific domain, a specific type of data, or a specific visual encoding. The task dataset can be a central resource for researchers conducting comparative studies to look up for tasks that have not been evaluated in the existing literature. Researchers can also add tasks they have evaluated in a comparative study to the benefit of other researchers.

6 Discussion and Future Work

In this paper, we present four task-based challenges of comparative evaluation studies identified through a hybrid method of literature review and personal experience. We found that common task-based problems we experienced in our own research have been discussed previously in different capacities by other researchers [27,54,60]. Furthermore, our survey and analysis of tree visualization comparative studies revealed that a large proportion of papers do not provide adequate information about the task source (C1), abstract definition of the tasks (C2), necessary information about the task procedure to support replication of the experiment (C3), and a through analysis of the open areas of the research (C4). Identification of these problems enabled us to reflect on ways researchers and visualization community can solve these problems.

Our paper provides practical guidelines to mitigate these task based challenges. Below we provide a checklist based on the **C1, C2, and C3** guidelines to assist researchers communicate tasks in their research and publications in a transparent easy to use manner:

Researcher Checklist for Publication:

Authors should,

- Explicitly mention the source of tasks, the reason for choosing the source and rationale for selecting the evaluation tasks.
- Provide task-level abstraction.
- Describe the task level procedure to ensure replication of the experiment by other researchers.

C4, as discussed Sec. 5, may not directly affect the validity of results, but the task space analysis may have an impact on the exhaustiveness of the study design. Researchers can identify early on in their study design process about tasks they might have missed. The task space analysis will also benefit the broader visualization community as it can be a source to identify potential areas of future work. Therefore, we recommend researchers should include the information of the evaluated task space as supplemental material.

Our community guidelines underscore the importance of creating a task dataset. We envision the task datasets should store an exhaustive list of analytical tasks related to visualization encoding and evaluation results demonstrating the effectiveness of encoding with a set of tasks. These datasets can be valuable resources for visualization researchers designing evaluation studies or practitioners looking for guidelines to choose the right visualization encoding, given the analytical tasks.

In this paper, we present a preliminary validation of the challenges. Therefore, we want to evaluate the problems and proposed guidelines further. There can be multiple ways to assess the challenges. We are particularly interested in interviewing researchers who have worked with comparative studies in the past. In the expert interview, we could collect more information on the challenges other researchers have faced in comparative studies. We also could gather expert's opinions on the guidelines we proposed for the challenges. The opinion may further help refine and improve the guidelines for mitigating the task-based challenges in comparative visualization studies.

Although, our proposed guidelines are derived from the lens of comparative studies, we believe the guidelines may be applicable more generally to other evaluation methods discussed in Sec. 2.3 as well as design studies. For instance, design studies with usability evaluations can also follow the proposed guidelines to ensure that they are identifying the right usability tasks and also communicating the tasks more transparently.

7 Conclusion

Appropriate task selection and transparent task communication are essential to the design of comparative studies. However, the current methods of task selection and communication in comparative studies have several shortcomings. We identified four task-based challenges that can potentially affect the validity and usability of a comparative visualization study. Corresponding to each problem, we provide necessary details to enable visualization researchers and practitioners to recognize the cause of the challenge and have a precise understanding of how the challenge affects a comparative study. We also surveyed 20 tree visualization comparative studies to determine if they were affected by the proposed task-based challenges in any capacity. Our results demonstrate that several tree visualization comparative studies lacked task source, the rationale for task selection, abstract descriptions of the domain tasks, and an under specified task communication format. To ensure comparative studies in the future minimize these problems, our work proposes a checklist of guidelines to assist researchers with careful task selection and accurate task communication.

Acknowledgments

This work was supported by NSF CISE CRII award no. 1657466 and the Khoury College of Computer Sciences, Northeastern University.

References

[1] C. Ahlberg and B. Shneiderman. The alphaslider: a compact and rapid selector. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 365–371, 1994. doi: 10.1145/191666.191790

[2] B. Alper, N. Riche, G. Ramos, and M. Czerwinski. Design study of linesets, a novel set visualization technique. *IEEE transactions on visualization and computer graphics*, 17(12):2259–2267, 2011. doi: 10.1109/TVCG.2011.186

[3] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, and P. Rodgers. Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges. In R. Borgo, R. Maciejewski, and I. Viola, eds., *EuroVis - STARs*. The Eurographics Association, 2014. doi: 10.2312/eurovisstar.20141170

[4] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pp. 111–117. IEEE, 2005. doi: 10.1109/INFVIS.2005.1532136

[5] R. A. Amar and J. T. Stasko. Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):432–442, 2005. doi: 10.1109/TVCG.2005.63

[6] K. Andrews. Evaluation comes in many guises. In *AVI Workshop on BEyond time and errors (BELIV) Position Paper*, pp. 7–8, 2008.

[7] K. Andrews and J. Kasanicka. A comparative study of four hierarchy browsers using the hierarchical visualisation testing environment (hvte). In *2007 11th International Conference Information Visualization (IV '07)*, pp. 81–86, July 2007. doi: 10.1109/IV.2007.8

[8] L. Barkhuus and J. A. Rode. From mice to men-24 years of evaluation in chi. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, vol. 10, 2007.

[9] T. Barlow and P. Neville. A comparison of 2-d visualizations of hierarchies. In *IEEE Symposium on Information Visualization, 2001. INFOVIS 2001.*, pp. 131–138, Oct 2001. doi: 10.1109/INFVIS.2001.963290

[10] T. Bladh, D. A. Carr, and M. Kljun. The effect of animated transitions on user navigation in 3d tree-maps. In *Ninth International Conference on Information Visualisation (IV'05)*, pp. 297–305, July 2005. doi: 10.1109/IV.2005.122

[11] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister. Evaluation of artery visualizations for heart disease diagnosis. *IEEE transactions on visualization and computer graphics*, 17(12):2479–2488, 2011. doi: 10.1109/TVCG.2011.192

[12] M. A. Borkin, C. S. Yeh, M. Boyd, P. Macko, K. Z. Gajos, M. Seltzer, and H. Pfister. Evaluation of filesystem provenance visualization tools. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2476–2485, Dec 2013. doi: 10.1109/TVCG.2013.155

[13] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19(12):2376–2385, 2013. doi: 10.1109/TVCG.2013.124

[14] M. Brehmer, M. Sedlmair, S. Ingram, and T. Munzner. Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pp. 1–8, 2014. doi: 10.1145/2669557.2669559

[15] D. Byrd. A scrollbar-based visualization for document navigation. In *Proceedings of the fourth ACM conference on Digital libraries*, pp. 122–129, 1999. doi: 10.1145/313238.313283

[16] S. Carpendale. *Evaluating Information Visualizations*, pp. 19–45. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. doi: 10.1007/978-3-540-70956-5_2

[17] M. C. Chuah and S. F. Roth. On the semantics of interactive visualizations. In *Proceedings IEEE Symposium on Information Visualization'96*, pp. 29–36. IEEE, 1996.

[18] K. A. Cook and J. J. Thomas. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Lab.(PNNL), Richland, WA (United States), 2005.

[19] J. W. Creswell and J. D. Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2017.

[20] A. Crisan and M. Elliott. How to evaluate an evaluation study? comparing and contrasting practices in vis with those of other disciplines : Position paper. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 28–36, Oct 2018. doi: 10.1109/BELIV.2018.8634420

[21] S. Di Bartolomeo, A. Pandey, A. Leventidis, D. Saffo, U. H. Syeda, E. Carstensdottir, M. Seif El-Nasr, M. A. Borkin, and C. Dunne. Evaluating the effect of timeline shape on visualization task performance. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3313831.3376237

[22] G. Ellis and A. Dix. An explorative analysis of user evaluation studies in information visualisation. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pp. 1–7, 2006. doi: 10.1145/1168149.1168152

[23] N. Elmqvist and J. S. Yi. Patterns for visualization evaluation. *Information Visualization*, 14(3):250–269, 2015. doi: 10.1177/1473871613513228

[24] E. D. Foster and A. Deardorff. Open science framework (osf). *Journal of the Medical Library Association: JMLA*, 105(2):203, 2017. doi: 10.5195/jmla.2017.88

[25] D. Gergle and D. S. Tan. Experimental research in hci. In *Ways of Knowing in HCI*, pp. 191–227. Springer, 2014. doi: 10.1007/978-1-4939-0378-8_9

[26] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009. doi: 10.1057/ivs.2008.31

[27] G. G. Grinstein, P. E. Hoffman, and R. M. Pickett. *Benchmark Development for the Evaluation of Visualization for Data Mining*, p. 129–176. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2001.

[28] S. Haroz. Open practices in visualization research : Opinion paper. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 46–52, Oct 2018. doi: 10.1109/BELIV.2018.8634427

[29] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, p. 203–212. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753357

[30] D. M. Hilbert and D. F. Redmiles. Extracting usability information from user interface events. *ACM Computing Surveys (CSUR)*, 32(4):384–421, 2000. doi: 10.1145/371578.371593

[31] P. Irani and C. Ware. Diagramming information structures using 3d perceptual primitives. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 10(1):1–19, 2003. doi: 10.1145/606658.606659

[32] P. Isenberg, A. Bezerianos, P. Dragicevic, and J.-D. Fekete. A study on dual-scale data charts. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2469–2478, 2011. doi: 10.1109/TVCG.2011.160

[33] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013. doi: 10.1109/TVCG.2013.126

[34] M. Y. Ivory and M. A. Hearst. The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys (CSUR)*, 33(4):470–516, 2001. doi: 10.1145/503112.503114

[35] F. J.-D. and P. C. InfoVis 2003 Contest. `www.cs.umd.edu/hcil/iv03contest`, 2003.

[36] N. Jardine, B. D. Ondov, N. Elmqvist, and S. Franconeri. The perceptual proxies of visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1012–1021, Jan 2020. doi: 10.1109/TVCG.2019.2934786

[37] B. S. Johnson. *Treemaps: visualizing hierarchical and categorical data*. PhD thesis, 1993. doi: 10.13016/M28G8FJ9W

[38] N. Kerracher and J. Kennedy. Constructing and evaluating visualisation task classifications: Process and considerations. *Computer Graphics Forum*, 36(3):47–59, 2017. doi: 10.1111/cgf.13167

[39] N. Kerracher, J. Kennedy, and K. Chalmers. A task taxonomy for temporal graph visualisation. *IEEE transactions on visualization and*

computer graphics, 21(10):1160–1172, 2015. doi: 10.1109/TVCG.2015.2424889

[40] A. Kobsa. User experiments with tree visualization systems. In *IEEE Symposium on Information Visualization*, pp. 9–16, Oct 2004. doi: 10.1109/INFVIS.2004.70

[41] R. Kosara and S. Haroz. Skipping the replication crisis in visualization: Threats to study validity and how to address them : Position paper. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 102–107, Oct 2018. doi: 10.1109/BELIV.2018.8634392

[42] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Seven guiding scenarios for information visualization evaluation. 2011.

[43] B. Lee, C. Plaisant, C. S. Parr, J.-D. Fekete, and N. Henry. Task taxonomy for graph visualization. In *Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization*, BELIV '06, p. 1–5. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1168149.1168168

[44] H. Lücke-Tieke, M. Beuth, P. Schader, T. May, J. Bernard, and J. Kohlhammer. Lowering the barrier for successful replication and evaluation. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 60–68, Oct 2018. doi: 10.1109/BELIV.2018.8634201

[45] J. E. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human–Computer Interaction*, pp. 152–169. Elsevier, 1995. doi: 10.1016/B978-0-08-051574-8.50019-4

[46] M. Meyer, T. Munzner, and H. Pfister. Mizbee: a multiscale synteny browser. *IEEE transactions on visualization and computer graphics*, 15(6):897–904, 2009. doi: 10.1109/TVCG.2009.167

[47] L. Micallef, P. Dragicevic, and J.-D. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE transactions on visualization and computer graphics*, 18(12):2536–2545, 2012. doi: 10.1109/TVCG.2012.199

[48] T. Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6):921–928, 2009. doi: 10.1109/TVCG.2009.111

[49] T. Munzner. *Visualization analysis and design*. CRC press, 2014. doi: 10.1201/b17511

[50] S. Muramalla, R. Al Tarawneh, S. R. Humayoun, R. Moses, S. Panis, and A. Ebert. Radial vs. rectangular: Evaluating visualization layout impact on user task performance of hierarchical data. *IADIS International Journal on Computer Science & Information Systems*, 12(2), 2017.

[51] J. Nielsen. Heuristic evaluation. In *Usability inspection methods*, pp. 25–62. John Wiley & Sons, Inc., 1994.

[52] A. Pandey, H. Shukla, G. S. Young, L. Qin, A. A. Zamani, L. Hsu, R. Huang, C. Dunne, and M. A. Borkin. Cerebrovis: Designing an abstract yet spatially contextualized cerebral artery network visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):938–948, Jan 2020. doi: 10.1109/TVCG.2019.2934402

[53] A. Pandey, Y. Zhang, J. A. Guerra-Gomez, A. G. Parker, and M. A. Borkin. Digital collaborator: Augmenting task abstraction in visualization design with artificial intelligence, 2020.

[54] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '04, p. 109–116. Association for Computing Machinery, New York, NY, USA, 2004. doi: 10.1145/989863.989880

[55] C. Plaisant, J. Grosjean, and B. B. Bederson. Spacetree: supporting exploration in large node link tree, design evolution and empirical evaluation. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002.*, pp. 57–64, Oct 2002. doi: 10.1109/INFVIS.2002.1173148

[56] A. Rind, W. Aigner, M. Wagner, S. Miksch, and T. Lammarsch. Task cube: A three-dimensional conceptual space of user tasks in visualization design and evaluation. *Information Visualization*, 15(4):288–300, 2016. doi: 10.1177/1473871615621602

[57] R. E. Roth. An empirically-derived taxonomy of interaction primitives for interactive cartography and geovisualization. *IEEE transactions on visualization and computer graphics*, 19(12):2356–2365, 2013. doi: 10.1109/TVCG.2013.130

[58] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE transactions on visualization and computer graphics*, 20(12):1604–1613, 2014. doi: 10.1109/TVCG.2014.2346481

[59] R. Sakai and J. Aerts. Card sorting techniques for domain characterization in problem-driven visualization research. In *EuroVis (Short Papers)*, pp. 121–125, 2015.

[60] B. Saket, A. Endert, and C. Demiralp. Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(7):2505–2512, July 2019. doi: 10.1109/TVCG.2018.2829750

[61] J. M. Santos, B. S. Santos, P. Dias, S. Silva, and C. Ferreira. Extending the h-tree layout pedigree: An evaluation. In *2013 17th International Conference on Information Visualisation*, pp. 422–427, July 2013. doi: 10.1109/IV.2013.56

[62] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, and R. Moorhead. A user study to compare four uncertainty visualization methods for 1d and 2d datasets. *IEEE transactions on visualization and computer graphics*, 15(6):1209–1218, 2009. doi: 10.1109/TVCG.2009.114

[63] H. Schulz, T. Nocke, M. Heitzler, and H. Schumann. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2366–2375, Dec 2013. doi: 10.1109/TVCG.2013.120

[64] M. Schwab, S. Hao, O. Vitek, J. Tompkin, J. Huang, and M. A. Borkin. Evaluating pan and zoom timelines and sliders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300786

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

[66] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*, pp. 336–343. IEEE, 1996. doi: 10.1109/VL.1996.545307

[67] J. Spool and W. Schroeder. Testing web sites: Five users is nowhere near enough. In *CHI'01 extended abstracts on Human factors in computing systems*, pp. 285–286, 2001. doi: 10.1145/634067.634236

[68] P. T. Sukumar and R. Metoyer. Towards designing unbiased replication studies in information visualization. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 93–101, Oct 2018. doi: 10.1109/BELIV.2018.8634261

[69] A. G. Sutcliffe, M. Ennis, and J. Hu. Evaluating the effectiveness of visual user interfaces for information retrieval. *International Journal of Human-Computer Studies*, 53(5):741–763, 2000. doi: 10.1006/ijhc.2000.0416

[70] J. G. Trafton, S. S. Kirschenbaum, T. L. Tsui, R. T. Miyamoto, J. A. Ballas, and P. D. Raymond. Turning pictures into numbers: extracting and generating information from complex visualizations. *International Journal of Human-Computer Studies*, 53(5):827–850, 2000. doi: 10.1006/ijhc.2000.0419

[71] S. Wehrend and C. Lewis. A problem-oriented classification of visualization techniques. In *Proceedings of the First IEEE Conference on Visualization: Visualization90*, pp. 139–143. IEEE, 1990. doi: 10.1109/VISUAL.1990.146375

[72] J. S. Yi, Y. ah Kang, J. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6):1224–1231, 2007. doi: 10.1109/TVCG.2007.70515

[73] C. Ziemkiewicz and R. Kosara. The shaping of information by visual metaphors. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1269–1276, Nov 2008. doi: 10.1109/TVCG.2008.171