Seven Guiding Scenarios for Information Visualization Evaluation

Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, Sheelagh Carpendale

Technical Report
No. 2011-992-04
Department of Computer Science
University of Calgary

Seven Guiding Scenarios for Information Visualization Evaluation

Heidi Lam
Google
heidi.lam@gmail.com

Enrico Bertini
University of Konstanz
enrico.bertini@uni-konstanz.de

Petra Isenberg
INRIA
petra.isenberg@inria.fr

Catherine Plaisant
University of Maryland
plaisant@cs.umd.edu

Sheelagh Carpendale
University of Calgary
sheelagh@ucalgary.ca

Seven Guiding Scenarios for Information Visualization Evaluation

Heidi Lam Enrico Bertini Petra Isenberg Catherine Plaisant Sheelagh Carpendale

Abstract—We take a new, scenario based look at evaluation in information visualization. Our seven scenarios, evaluating visual data analysis and reasoning, evaluating user performance, evaluating user experience, evaluating environments and work practices, evaluating communication through visualization, automated evaluation of visualizations, and evaluating collaborative data analysis were derived through an extensive literature review of over 800 visualization publications. These scenarios are described through their goals, the types of questions they embody and illustrated through example studies. Through this broad survey and the distillation of these scenarios we make two contributions. One, we encapsulate the current practices in the information visualization research community and, two, we provide a different approach to reaching decisions about what might be the most effective evaluation of a given information visualization. For example, if the research goals or evaluative questions are known they can be used to map to specific scenarios, where practical existing examples can be considered for effective evaluation approaches.

Index Terms—Information visualization, evaluation

1 Introduction

Researchers and practitioners in the field of information visualization (infovis) have long identified the need to evaluate visual data representations, interaction techniques, and visualization systems. Yet, the difficulty of conducting these infovis evaluations remains a common topic. For instance, in addition to the general evaluations challenges of choosing evaluation questions, methods, and correctly executing them, the infovis focus on data and its exploratory analysis processes pose still further challenges, since both the analysis process and outputs, such as specific insights and more global growing comprehension, are difficult to capture and quantify (cf. Section 7).

While the need for facility with evaluations that are capable of addressing these challenges is much discussed [4, 11], ascertaining an approach that can improve upon our existing practices has remained elusive. For experimenters, part of the problem is the vast amount of evaluation methodologies in use. Our community draws from diverse disciplines such as psychophysics, social sciences, statistics, and computer science, using methodologies as diverse as laboratory based factorial design studies, field evaluations, statistical data analysis, and automatic image evaluation. The vastness and diversity of evaluation methodologies make it difficult for visualization researchers and practitioners to find the most appropriate approaches to achieve their evaluation goals. Another aspect of the difficulty is the lack of literature guidelines—while some guidelines are available to create and analyze visualization systems, and to evaluate visualizations, these two sets of literature are disparate as discussions on evaluation are mostly "structured as an enumeration of methods with focus on *how* to carry them out, without prescriptive advice for *when* to choose between them." ([51, p.1], author's own emphasis). We extend this by taking a different tack—we offer advice on *how* to chose between evaluation approaches.

In the paper, we take a broad community based approach, discovering from the infovis research literature common linkages between evaluation goals and evaluation approaches. We discuss these linkages via evaluation scenarios. This is based on an extensive literature analysis of over 800 papers (345 with evaluation), we systematically identified seven most commonly encountered evaluation scenarios:

1) Evaluating environments and work practices
2) Evaluating visual data analysis and reasoning
3) Evaluating communication through visualization
4) Evaluating collaborative data analysis
5) Evaluating user performance
6) Evaluating user experience
7) Automated evaluation of visualizations

For each of these scenarios, we list the most common evaluation questions and where possible illustrate them with representative published evaluation examples from the infovis community. In cases where there are gaps in our community's evaluation approaches, we suggest methods from publications from other sources.

Given the vast scope of the evaluation topic, we do

- *Heidi Lam is with Google Inc.*
  E-mail: heidi.lam@gmail.com
- *Enrico Bertini is with the University of Konstanz*
  E-mail: enrico.bertini@uni-konstanz.de
- *Petra Isenberg is with INRIA*
  E-mail: petra.isenberg@inria.fr
- *Catherine Plaisant is with the University of Maryland*
  E-mail: plaisant@cs.umd.edu
- *Sheelagh Carpendale is with the University of Calgary*
  E-mail: sheelagh@ucalgary.ca

not provide a comprehensive list of existing evaluation methods, though we do provide a wide coverage of the methodology space in our scenarios to offer a diverse set of evaluation options. We leave in depth discussions of specific methods to the ample literature that is available with these details.

The major contribution of our work is a scenario-based approach that offers advice on *how* to make evaluation choices illustrated with published examples drawn from a diverse literature. More specifically, our work aims to:

- encourage the selection of evaluation methods based on specific evaluation goals (by organizing our guide by scenarios rather than by methodologies);
- diversify evaluation methods used in our community (by providing examples from other disciplines in context of evaluation goals commonly found in our community);
- act as a first step to develop a repository of examples and scenarios as a reference.

2 The Scope of Evaluation

In this guide, we cover evaluation as part of different stages of visualization development, including:

1) **Pre-design** e. g., to understand potential users' work environment and work flow
2) **Design** e. g., to scope a visual encoding and interaction design space based on human perception and cognition
3) **Prototype** e. g., to see if a visualization has achieved its design goals, to see how a prototype compares with the current state-of-the-art systems or techniques
4) **Deployment** e. g., to see how a visualization influences workflow and work processes, to assess the visualization's effectiveness and uses in the field
5) **Re-design** e. g., to improve a current design by identifying usability problems

Note that with this broad view of evaluation, it is not restricted to the analysis of specific visual representations—it can focus on visualizations' roles on processes such as data analysis, or on specific environments to which visualizations might be applied. Our scenarios are therefore grouped based on their evaluation foci: process or visualization (Section 6). The outputs of the evaluation may be specific to a visualization such as its design decisions, or more general such as models and theories (e. g., theory formation based on grounded theory evaluation [38] and perceptual and cognitive modeling based on controlled experiments), and metrics (e. g., metrics developments based on automatic evaluation of visual quality or salience).

3 How to Use This Paper

The purpose of this paper to suggest a process that can support the design of evaluations of visualizations through the following steps:

1) **Setting a goal**: In general, before thinking about evaluation methods, we recommend starting by determining a clear evaluation goal [12, 19]. Section 2 lists a range of evaluation goals categorized by stages in visualization development; all of which are covered in our scenarios as described in Section 6. Each scenario is illustrated with a few common *evaluation questions* that can help to identify for which evaluation goals they may be most suitable.
2) **Picking suitable scenarios**: Having identified an evaluation goal, the seven scenarios can be used to help to identify close matches with the research goal and to provide descriptions of relevant information related to the goal.
3) **Considering applicable approaches**: These can be found in the *Methods and Examples* sections of Section 6. We recommend initially investigating several approaches that may suit a given evaluation need. Each scenario is illustrated with examples of published evaluations, which can be used as references for additional details.
4) **Creating evaluation design and planned analyses**: The design of an evaluation can be strengthened by considering benefits and limitations of each approach listed in the scenario selected. While we aimed to provide a diverse range of evaluation methods, the lists are not exhaustive. Also, in information visualization, research into evaluation methodologies themselves is still active and is still resulting in new methodologies and metrics. For these reasons, we encourage creativity in evaluation design starting from and also extending the work referenced here.

4 Related Work

In this section, we review related work in the areas of evaluation taxonomies, systematic reviews, and evaluation methodologies and best practices.

4.1 Evaluation Taxonomies

Others have approached the problem of guiding researchers and practitioners in visualization evaluation by providing a high-level view of available methodologies and methods as taxonomies. The metrics used for classification have been diverse, ranging from research goals, to design and development stages in which the methodologies can be applied, to methods and types of data collected, to the scope of evaluation. Table 1 summarizes existing taxonomies and their respective foci.

The diversity exhibited in Table 1 reflects the complexity and richness of existing evaluation methodologies and the difficulty in deriving an all encompassing taxonomy. For example, using research goals as a taxonomy axis is challenging because the same evaluation method may be used for different purposes. One example is laboratory-based studies measuring task completion time to compare between interfaces (also known as "head-to-head" comparisons). Such a method can be used to summarize the effectiveness of an interface ("summative") or to inform design ("formative") [2, 19]. Similar arguments apply to classifying methods based on design and development cycles—the

same method may be used differently at different stages. For example, observational technique may be first used in the pre-design stage to gather background information [38], but may also be used post-release to understand how the newly introduced technology affects user workflow. Given these difficulties, we decided on a different approach where we based our discussions on commonly encountered evaluation scenarios instead of methods. Across all the papers we examined, we explored how these scenarios relate to evaluation goals and questions (Section 5). Our goal is to encourage an approach to evaluation that is based on evaluation goals and questions instead of methods and to encourage our community to adopt and accept a more diverse range of evaluation methods.

4.2 Systematic Reviews

Our work here is closest in spirit to a subtype of systematic review known as narrative review, which is a qualitative approach and describes existing literature using narrative descriptions without performing quantitative synthesis of study results [72]. Systematic reviews is itself a type of evaluation method with the purpose to provide snapshots of existing knowledge based on published study results, where "the researcher focuses on formulating general relations among a number of variables of interest" that "hold over some relatively broad range of populations", [48, p. 158]. To the best of our knowledge, two systematic reviews on evaluation methods have been conducted, both counted the number of papers in specific corpora based on the authors' classification scheme.

The first is Barkhuus and Rode's analysis on 24 years of publications in the proceedings of the SIGHCHI Conference on Human Factors in Computing Systems (CHI) [4]. The researchers found that while the proportion of papers with evaluations increased over time, the quality of the evaluation may not have improved, judging from the decreased median number of participants in quantitative studies, an over-reliance of students as participants, and lack of gender-balanced samples. The second is Perer and Shneiderman's analysis on three years of publications in the proceedings of the IEEE Symposium on Information Visualization (InfoVis) and one year of the IEEE Symposium on Visual Analytics Science and Technology (VAST) [54]. In these corpora, the researchers did not find an increase in proportion of papers with evaluation. Similar to Barkhuus and Rode, Perer and Shneiderman also expressed concerns over the quality of evaluation, as most evaluations conducted were controlled studies with non domain experts as test subjects. Our focus in contrast was to derive a common set of researcher questions and approaches to ground the development of our scenarios.

4.3 Evaluation Methodologies and Best Practices

There exists a large number of publications that reflect upon current practices in visualization evaluation and provide recommendations to improve our status quo. In fact, the BELIV workshop was created as a venue for researchers and practitioners to "explore novel evaluation methods, and to structure the knowledge on evaluation in information visualization around a schema, where researchers can easily identify unsolved problems and research gaps" [7]. In short, providing a complete summary of publications on evaluation probably deserves a paper of its own. In this section, we briefly outline some of the commonly discussed challenges.

In terms of study design, many papers urge researchers to think about the goals of the evaluation [12, 19]. The evaluation goal heavily influences the choice of research strategies, the types of data and methods of collection, and the methods of data analysis. For example, if the goal is to understand how a new technology affects user workflow, then realism is important. In other words, data collection should be from the field using non-intrusive collection mechanisms. Several researchers of these papers that reflect on evaluation commented on the lack of realism in the existing evaluation efforts, which are mostly laboratory based, using basic visual search tasks with non-target users. One way to ensure validity is to ensure realism in tasks, data, workflow, and participants [2, 19, 60]. An alternative is to provide an understanding of situations where some of these requirements can be released, for example, using non-domain expert participants. Other commonly discussed topics of study design include the short durations of most study periods [60], the narrowness of study measurements [60], and possibly insufficient training of participants [2]. In term of data analysis, concerns have been expressed on the narrowness of questions posed and statistical methods applied [19]. Given that most of the existing evaluation studies are one-offs, researchers have suggested doing follow-up studies to further investigate unanswered questions [19, 43].

In short, all aspects of evaluation require careful attention. This paper is therefore an effort to provide a different kind of guide for visualization researchers and practitioners through concrete scenarios illustrated with existing evaluations.

5 Methodology

Early in our project, we decided to take a *descriptive* rather than a *prescriptive* approach. In other words, our paper describes and comments on existing practices in evaluating visualizations, but we do not prescribe specific evaluation methods as we believe that the final decision on appropriate methods should be decided on a case-by-case basis. We identified seven evaluation scenarios most commonly encountered by visualization researchers which are meant to *guide* the development of appropriate evaluation strategies. The scenarios were derived from data collected through open coding [14] of publications from four information visualization publication venues (see Table 2). Our approach included the following steps to derive the scenarios:

1—Compiling an evaluation dictionary. Initially, to gather a description of existing evaluation practices in the visualization community, we compiled a dictionary

Type	Categories	Refs
Evaluation goals	Summative *(to summarize the effectiveness of an interface)*, formative *(to inform design)*	Andrews [2], Ellis and Dix [19]
Evaluation goals	Predictive *(e. g., to compare design alternatives and compute usability metrics)*, observational *(e. g., to understand user behaviour and performance)*, participative *(e. g., to understand user behaviour, performance, thoughts, and experience)*	Hilbert and Redmiles [32]
Research strategies	Axes *(generalizability, precision, realism, concreteness, obtrusiveness)* and research strategies *(field, experimental, respondent, theoretical)*	McGrath [48]
Research methods	Class *(e. g., testing, inspection)*, type *(e. g., log file analysis, guideline reviews)*, automation type *(e. g., none, capture)*, effort level *(e. g., minimal effort, model development)*	Ivory and Hearst [39]
Design stages	Nested Process Model with four stages *(domain problem characterization, data/operation abstraction, encoding/interaction technique design, algorithm design)*, each with potential threats to validity and methods of validation	Munzner [51]
Design stages	Design/development cycle stage associated with evaluation goals *("exploratory" with "before design", "predictive" with "before implementation", "formative" with "during implementation", and "summative" with "after implementation")*. Methods are further classified as inspection *(by usability specialists)* or testing *(by test users)*.	Andrews [2]
Design stages	Planning & feasibility *(e. g., competitor analysis)*, requirements *(e. g., user surveys)*, design *(e. g., heuristic evaluation)*, implementation *(e. g., style guide)*, test & measure *(e. g., diagnostic evaluation)*, and post release *(e. g., remote evaluation)*	Usability.net [87]
Design stages	Concept design, detailed design, implementation, analysis	Kulyk et al. [42]
Data and method	Data collected (qualitative, quantitative), collection method (empirical, analytical)	Barkhuus and Rode [4]
Data	Data collected (qualitative, quantitative, mixed-methods)	Creswell [14]
Evaluation scope	Work environment, system, components	Thomas and Cook [80]

TABLE 1
Taxonomies of evaluation methods and methodologies based on the type of categorization, the main categories themselves, and the corresponding references.

of terms of existing evaluation strategies and techniques and collected matching definitions and example evaluation publications. Our list was compiled based on information solicited by emails to participants of the BELIV 2008 workshop combined with our own knowledge and research (e. g., [7, 10, 36, 38, 43, 60]). The process yielded a wealth of information which required additional structure but provided us with a richer understanding of the types of evaluations commonly used and helped to provide necessary context for us to perform the open coding and tagging of the evaluation papers.

2—Open coding and tagging. From the set of terms and examples collected in the first phase we derived an initial eight tags that classified evaluations in terms of evaluation goals. These tags included topics such as data analysis, decision making, or usability and were integrated into our final result. We selected four major visualization publication venues from which to identify commonly encountered evaluations:

- Eurographics/IEEE Symposium on Visualization (EuroVis)
- IEEE Information Visualization (InfoVis)
- IEEE Visual Analytics Science and Technology (VAST)
- Palgrave's Journal of Information Visualization (IVS)

From these sources, we collected 803 papers and conducted a first coding pass that culled papers that did not mention evaluation and left 345 evaluation papers for further consideration. Publication years and number of papers involved are summarized in Table 2.

Three of us performed the open-coding [14] on parts of the dataset. For each paper, we attached one or more tags from the initial set and recorded the reported evaluation goals and methods. As we proceeded in coding selected publications, each of us independently added new tags to the initial collection, which were then shared among all coders during the tagging period. At regular intervals we discussed the definition of each tag within the group and through consensus, adopted new tags from the other coders during the process and recoded papers with the new tags.

By the end of our publication coding, we had expanded our initial tag set to seventeen tags. Details of the tags can be found in Appendix A.

3—Developing Scenarios. We engaged in one final coding pass and grouped similar tags among the 17 tags to form 7 tags, excluding tags that relate to discussing methods to develop new evaluations as we considered these to be beyond the scope of this paper. The goal was to

Venue	Year	Papers	Papers+Eval
EuroVis	2001 – 2008	292	107
InfoVis	1995 – 2008	312	145
IVS	2002 – 2008	128	67
VAST	2006 – 2008	71	26

TABLE 2
Venues included in the open-coding stage to identify commonly encountered evaluation goals, which were then distilled into scenarios.

derive scenarios which would represent the main evaluation questions encountered when evaluating visualization tools. This consolidation provides a more manageable list of elements in order to facilitate their use in practice, as described in Section 3. Scenarios, tags, and paper numbers for each are summarized in Table 3 in the Appendix.

The building of scenarios is, thus, the result of an iterative process among coders where phases of individual grouping and collective consolidation alternated. In the next section, we present the final seven scenarios we derived.

6 SCENARIOS

In this section, we present our seven evaluation scenarios. For each scenario, we define the scenario, identify the popular goals and outputs, the common evaluation questions, and the applicable evaluation methods along with concrete examples. Note that our scenarios can be roughly classified into two broad categories based on their focus. We call these two categories *process* and *visualization*. In the process group, the main goal of the evaluation is to understand the underlying process and the roles played by visualizations. While evaluators may record specific user performance and feedback, the goal is to capture a more holistic view of the user experience. Scenarios that belong to this group are: Evaluating Environment and Work Practices (EWP), Evaluating Visual Data Analysis and Reasoning (VDAR), Evaluating Communication through Visualization (CTV), and Evaluating Collaborative Data Analysis (CDA). In contrast, evaluations can focus on the visualization itself, with the goal to test design decisions, explore design space, bench-mark against existing systems, or to discover usability issues. Usually in these evaluations, a slice of the visualization system or technique is tested. Scenarios that belong to this group include Evaluating User Performance (UP), Evaluating User Experience (UE), and Automated Evaluation of Visualizations (AEV).

6.1 Evaluating Environments and Work Practices (EWP)

Evaluations in the EWP group elicit formal requirements for design. In most software development scenarios it is recommended to derive requirements from studying the people for which a tool is being designed [75] but, as noted by Munzner [51, p.7], "hardly any papers devoted solely to analysis at this level [problem characterization] have been published in venues explicitly devoted to visualization." Similarly, Plaisant [60] has argued that there is a growing need for information visualization designers to study the design context for visualization tools including tasks, work environments, and current work practices. Yet, in information visualization research studying people and their task processes is still rarely done and only few notable exceptions have published results of these analyses (e. g., [37, 83]).

6.1.1 EWP: Goals and Outputs

The goal of information visualization evaluations in this category is to work towards understanding the work, analysis, or information processing practices by a given group of people with or without software in use. The output of studies in this category are often design implications based on a more holistic understanding of current workflows and work practices, the conditions of the working environment itself, and potentially current tools in use. Studies that involve the assessment of people's work practices *without* a specific visualization tool typically have the goal to inform the design of a future visualization tool. Studies involving the assessment of work flow and practices *with* a specific tool in people's work environment try to assess factors that influence the adoption of a tool to find out how a tool has been appropriated and used in the intended work environments in order to elicit more specific design advice for future versions of the tool.

6.1.2 EWP: Evaluation questions

Questions in this scenario usually pertain to identifying a set of features that a potential visualization tool should have. For example:

- What is the context of use of visualizations?
- In which daily activities should the visualization tool be integrated?
- What types of analyses should the visualization tool support?
- What are the characteristics of the identified user group and work environments?
- What data is currently used and what tasks are performed on it?
- What kinds of visualizations are currently in use? How do they help to solve current tasks?
- What challenges and usage barriers can we see for a visualization tool?

6.1.3 EWP: Methods and Examples

There is a wealth of methods available for studying work environments and work practices. Most of these rely on qualitative data such as interviews or observational data, audio-visual, or written material:

Field Observation. Observational methods are the most common way to elicit information on current work practices and visualization use. We further described the goals and basics of this method in Section 6.6. In information visualization, few published examples exist of this type of

study, but more have been called for [38]. In a study of automotive engineers Sedlmair et al. [70] observed eight analysis experts at their workspace and derived information on which and how specific tools were used, why they were used, and when participants entered in collaborative analysis. The researchers then used the results of the study to derive a set of requirements for the design of data analysis systems for this domain. Both this study and another by Tory et al. [83] combined observation with interviews.

Interviews. There are several types of interviewing techniques that can be useful in this context. Contextual inquiry [34] is a user-centered design method in which people are first observed and then interviewed while engaged in their daily routines within their natural work environment. The researcher tries to interfere as little as possible. Picking the right person to interview is critical in order to gather useful results. Interviews can also be conducted within a lab context. These types of interviews are more common in information visualization: Pretorius and van Wijk interviewed domain experts about their own data to learn how they analyzed state transition graphs [62], Brewer et al. [9] interviewed geovisualization experts to learn about multi-disciplinary science collaboration and how it could be facilitated with collaborative visualization tools.

Laboratory Observation. Observational studies also sometimes occur in laboratory settings in order to allow for more control of the study situation. For example, two studies from the collaborative visualization field looked at how visualizations are used and shared by groups of people and how visualization results are integrated [37, 64]. Both studies presented rich descriptions of how people interacted with visualizations and how these activities could be supported by technology.

6.2 Evaluating Visual Data Analysis and Reasoning (VDAR)

Evaluations in the VDAR group study if and how a visualization tool supports users in generating actionable and relevant knowledge in their domain. In general, VDAR evaluation requires fairly well developed and reliable software.

6.2.1 VDAR: Goals and Outputs

Evaluations in the VDAR group involve studies that assess how an information visualization tool supports visual analysis and reasoning about data to generate information about the users' domain. Outputs are both quantifiable metrics such as the number of insights obtained during analysis (e.g., [67, 68]), or subjective feedback such as opinions on the quality of the data analysis experience (e.g., [71]).

Even though VDAR studies may collect objective participant performance measurements, studies in this category look at how an integrated visualization tool as a whole supports the analytic process, rather than studying an interactive or visual aspects of the tool in isolation. We cover the latter case in our scenario *Evaluating User Performance* in Section 6.5. Similarly, VDAR is more process oriented than just to identify usability problems in the interface to refine the prototype, which is covered in Section 6.6. Here, we first focus on the case of a single user. Collaboration is discussed in Section 6.4, *Evaluating collaborative data analysis*.

6.2.2 VDAR: Evaluation Questions

Data analysis and reasoning is a complex and ill-defined process. Our sample questions are inspired by Pirolli and Card's model of an intelligence analysis process [58], considering how a visualization tool supports:

- Data exploration? How does it support processes aimed at seeking information, searching, filtering, and reading and extracting information?
- Knowledge discovery? How does it support the schematization of information or the (re-)analysis of theories?
- Hypothesis generation? How does it support hypothesis generation and interactive examination?
- Decision making? How does it support the communication and application of analysis results?

6.2.3 VDAR: Methods and Examples

Studying how a visualization tool may support analysis and reasoning is difficult since analysis processes are typically fluid and people use a large variety of approaches [37]. In addition, the products of an analysis are difficult to standardize and quantify since both the process and its outputs are highly context-sensitive. For these reasons, evaluations in VDAR are typically field studies, mostly in the form of case studies. These studies strive to be holistic and to achieve realism by studying the tool use in its intended environment with realistic tasks and domain experts. However, we also found experiments in which parts of the analysis process were controlled in a laboratory setting.

In this section, we focus on techniques that individual researchers can use, opposed to community wide evaluation efforts such as the Visualization Contest or the Visual Analytics Challenge [13]. The Visual Analytics Challenge provides a useful collection of data sets and analysis problems that can be used in wider VDAR evaluations, and a repository of examples that demonstrate how other tools have been used to analyse the data.

Case Studies. Case studies conducted in VDAR are mostly studies on domain experts interacting with the visualization to answer questions listed in 6.2.2. For example, Trafton et al. conducted an exploratory investigation in the field to answer questions such as "How are complex visualizations used, given the large amount of data they contain?" [84, p. 16]. The researchers recruited three pairs of meteorological forecasters and asked them to prepare a written information brief for a flight. The researchers open-coded video data to capture the type of visualizations used in various stages of the analysis.

In some cases, researchers collect data over a longer period of time (from weeks to months) with participants

working on their own problems in their normal environments. Analysis activities may be captured by automated logging or by self-reporting methods such as the diary method [79]. Two examples of such long-term case studies in visualization evaluation are: Multi-dimensional In-depth Long-term Case studies (MILCs) [74] and insight-based evaluations [68].

MILC evaluations use multiple techniques such as observations, interviews, surveys, and automated logging to assess user performance, interface efficacy, and interface utility [74]. In MILC studies, researchers offer assistance to participants in learning the system, and may improve the systems based on participant feedback. MILC evaluations have been employed, for instance, to evaluate knowledge discovery tools [71] and the integration of statistics and visualization [54]. The main question Seo et al. set out to answer in their MILC case studies using the Hierarchical Clustering Explorer (HCE) was "how do HCE and the rank-by-feature framework change the way researchers explore their datasets" [71, p. 313]. To answer this data exploration question, Seo et al. used participatory observations [10, p. 36] and interviews, conducted weekly for a period of four to six weeks, during which time the participants were asked to use the tool in their everyday work.

Insight-based evaluations try to capture insight as "an individual observation about the data by the participant, a unit of discovery" [67, p. 4]. Data collection methods proposed are either the diary method or capturing video using a think-aloud protocol. For example, Saraiya et al. conducted a longitudinal study with biologists and bioinformaticans using real-life microarray data [68]. The goals of the study were to deepen understanding of the visual analytics process, to understand how existing tools were used in analysis, and to test out an evaluation methodology. Data was collected using a diary maintained by the participants to record the analysis process, the insights gained from the data, and which visualization and interaction techniques led to insights, and the successes and frustrations participants experienced with the software tools. Over the course of the study, debriefing meetings were held once every two to three weeks for the researchers to discuss data insights and participants' experience with the tools. Unlike the MILC studies, the researchers did not provide any help with the software tools or guide their participants' data analysis in any way to minimize the study's impact on the participants' normal data analysis process.

Controlled Experiment. Given the open-ended nature of exploration and the specificity of case studies, it may be beneficial to control some of the analysis process and study using laboratory experiments. For example, the Scented Widgets study measured how social navigation cues affected information foraging based on the number of revisits, unique discoveries, and user subjective preferences based on log data [92]. In some cases, experimenters may use a mixture of techniques to enrich the data collected in laboratory experiments. One example is an early insight-based evaluation [67]. The study used a think-aloud protocol and participants were asked to estimate the percentage of potential insight they would be able to obtain about the dataset with the tool every 15 minutes. In addition, Saraiya et al. coded all individual occurrences of insights from video recordings, with the characteristics of the insights coded by domain experts. Findings were expressed in five measures of insights: count, total domain value, average final amount learned, average time to first insight, and average total time spent before no more insight was felt to be gained. Both the hand-coded as well as the participant-recorded metrics helped to evaluate the most efficient of the five visualization techniques in supporting insight discovery and in influencing users' perception of data.

6.3 Evaluating Communication through Visualization (CTV)

Evaluations in the CTV group study if and how communication can be supported by visualization. Communication can pertain to aspects such as learning, teaching, and idea presentation as well as casual consumption of visual information as in ambient displays.

6.3.1 CTV: Goals and Outputs

Visualizations in this category have the goal or purpose to convey a message to one or more persons, in contrast to targeting focused data exploration or discovery. Their effectiveness is usually measured in terms of how effectively such a message is delivered and acquired. Ambient displays, for example, belong to this category as they are usually built to quickly communicate peripheral information to passers-by.

6.3.2 CTV: Evaluation questions

Studies in CTV are often interested in quantifying a tool's quality through metrics such as learning rate, information retention and accuracy, and qualitative metrics such as interaction patterns of the way people absorb information or approach the tool. Questions thus pertain to the quality with which information is acquired and the modalities with which people interact with the visualizations. Examples of questions are:

- Do people learn better and/or faster using the visualization tool?
- Is the tool helpful in explaining and communicating concepts to third parties?
- How do people interact with visualizations installed in public areas? Are they used and/or useful?
- Can useful information be extracted from a casual information visualization?

6.3.3 CTV: Methods and Examples

Controlled Experiments. Quantitative studies aiming at measuring improvement in communication or learning, employ traditional controlled experiments schemes. As an example, Sedig et al. studied how students used a mathematical visualization tool aimed at teaching basic concepts in geometry [69]. A similar study was performed in the context of a basic programming class, using a tool that

visualized the role of variables in program animation [66]. This last experiment is of special interest as it highlights how measuring learning may require the study to span several weeks or months and may, thus, take longer than other traditional evaluations.

Field Observation and Interviews. Qualitative methods like direct observation and interviews are often paired up with experiments in this context. The studies mentioned above, for instance, both complement their quantitative approach with observations of tools in use to understand how information is acquired and to better investigate the process that leads to concept learning. In the context of casual visualizations, that is, visualizations that "*depict personally meaningful information in visual ways that support everyday users in both everyday work and non-work situations*" [61], direct observation and interviews are common evaluation techniques. For example, Skog et al. [76] study the use of an ambient visualization to convey real-time information of bus departure times in a public university area. The evaluation consists of interviewing people and spending enough time in the area to understand the people's interaction with the system. Viegas et al. [88], studied a visual installation in a museum's gallery. The authors observed the reactions of people to the installation and collected the people's impressions to draw conclusions on the design. In a similar context Hinrichs et al. [33] used an observational and video coding approach to analyze how visitors in an art museum approach a visualization installation and derived design considerations for information visualization in the museum context. As noted by Pousman et al. [61], this kind of observational evaluation is often needed in such context because it is necessary to capture evaluation data in a natural setting where people use the tools naturally.

6.4 Evaluating Collaborative Data Analysis (CDA)

Evaluations in the CDA group study whether a tool allows for collaboration, collaborative analysis and/or collaborative decision making processes. Collaborative data analysis differs from single-user analysis in that a group of people share the data analysis experience and often have the goal to arrive at a *joint* conclusion or discovery.

6.4.1 CDA: Goals and Outputs

Evaluations in this group study how an information visualization tool supports collaborative analysis and/or collaborative decision making processes. Collaborative systems should support both *taskwork*, the actions required to complete the task, and *teamwork*, the actions required to complete the task as a group [56]. For collaborative visualization this means that systems must not only support group work well, but also be good data analysis tools (*taskwork*). We cover the evaluation of taskwork and its questions in the other scenarios (see Section 6.2). Studies in this category have varying goals and, thus, are defined by different types of outputs. Most commonly studies in this group aim to gain a more holistic understanding of group work processes or tool use during collaboration with the goal to derive concrete design implications. It is recognized that the study of teamwork is difficult due to a number of factors including a greater number of variables to consider, the complicated logistics of evaluation, or the need to understand and judge group work processes [52]. Collaborative systems (or groupware) can be evaluated on a number of different levels such as the organization it will be embedded in, the team or group that will be using it, or the system itself. While there have been a number of papers concerned with the evaluation of groupware, only few examples of evaluations for collaborative information visualization systems exist.

6.4.2 CDA: Evaluation Questions

For the CDA evaluation of such systems any of or a combination of the following questions may be relevant to address:

- Does the tool support *effective and efficient* collaborative data analysis?
- Does the tool *satisfactorily* support or stimulate group analysis or sensemaking?
- Does the tool support group insight? [78]
- Is social exchange around and communication about the data facilitated?
- How is the collaborative visualization system used?
- How are certain system features used during collaborative work? What are patterns of system use?
- What is the process of collaborative analysis? What are users' requirements?

6.4.3 CDA: Methods and Examples

As research on collaborative visualization systems has only recently begun to receive increased research attention, there are only few examples of studies in this area. We thus draw on results from both Computer-Supported Cooperative Work (CSCW) as well as the small set of recent studies in collaborative visualization.

Within the field of CSCW a multitude of study and data collection methods have been applied to the analysis of group work [52, 55]. The *context* of group work (e. g. group configuration, work environment) has been identified as a critical factor in the evaluation and acceptance of collaborative systems (e. g. [23, 52, 86]). Yet, several research papers have outlined the practicality of early formative evaluations in less authentic environments (e. g. [57, 86]). Coupled with later more situated fieldwork a clearer picture of collaborative systems in use and their influence on groups and organizations can be won. Here we highlight a number of possible evaluation techniques.

Heuristic Evaluation. Heuristic evaluation has been previously proposed for the evaluation of visualization systems [81, 94]. Finding an appropriate set of heuristics is the main challenge for visualization systems not only to evaluate taskwork [94]. For the evaluation of teamwork a set of *heuristics* for the assessment of *effectiveness and efficiency* of collaboration has been proposed [3]. These heuristics are

based on the mechanics of collaboration [24, 57] or low-level actions and interactions that a collaborative system must support in order for group members to be able to complete a task in a shared manner. Other sets include heuristics based on the locales framework to study the influences of locales (places) on social activities [20] or awareness [16].

Log Analysis. Analysis of logs and user traces were the main sources of information analyzed in studies of distributed collaborative web-based information visualization tools [30, 89]. Both analyses resulted in descriptions and statistics of collaborative use of system features and suggestions for system improvement. Wattenberg [90] used a slightly different approach while investigating the use of his web-based *NameVoyager*. He studied off-site reviews and comments and reported on a number of examples of social interaction around the use of the tool. Studies involving the investigation of logs or comments have the advantage to be relatively easy to conduct and evaluate. Little interaction with participants is used to analyze specific system features or a tool use overall. To elicit more user-specific data these evaluation have been combined with questionnaires or interviews (e. g. [30]). On the other hand, these studies cannot clearly evaluate interaction between participants, their work or other processes that do not generate a traceable log entry.

Field or Laboratory Observation. Qualitative user studies have a long tradition within CSCW [21, 52]. Observational studies are often combined with logging of user activity, questionnaires, or interviews [47, 50]. In an analysis of group activities with an information visualization system Mark and Kobsa [47] used such a combination of techniques to analyze group coordination and analysis processes. Effectiveness and efficiency were also assessed by tracking errors and timings for group tasks. Isenberg et al. studied how effectively their collaborative social network analysis system *CoCoNutTrix* [36] supported the collaborative analysis process. They performed an observational study and post-session interview to assess how well the system supported the following factors of the collaboration: explicit communication, consequential communication, group awareness, coordination of actions, group insight, subjective work preferences, and general user reactions to the collaborative environment. Without digital systems, other more exploratory observational studies in visualization and visual analytics assessed group analysis processes [37] and collaborative information synthesis [64]. For collaborative systems studies of work processes are often seen as important prerequisites for estimating outcomes of tool use and to develop mature CSCW tools [52].

In contrast to single user systems, collaborative visual analysis systems must also consider the groups interactions and possible harmony/dis-harmony as they proceed in their joint discovery efforts. Stahl [78] defines the notion of group cognition as "computer-supported collaborative knowledge building" and recommends the study of this collaborative knowledge building through discourse analysis and observation. It would be interesting to combine this approach with insight-based methodologies (e. g. [68]) for the study of group insight.

6.5 Evaluating User Performance (UP)

Evaluations in the UP group study if and how specific features affect objectively measurable user performance.

6.5.1 UP: Goals and Outputs

User performance is predominantly measured in terms of objectively measurable metrics such as time and error rate, yet it is also possible to measure subjective performance such as work quality as long as the metrics can be objectively assessed. The most commonly used metrics are task completion time and task accuracy. Outputs are generally numerical values analyzed using descriptive statistics (such as mean, median, standard deviations, and confidence intervals) and modeled by such methods as ANalysis Of VAriance (ANOVA) to partition observed variance into components.

6.5.2 UP: Evaluation questions

Questions addressed using evaluation methods in the UP group are generally narrow and determined prior to the start of the evaluation. There are basically two types of questions:

- What are the limits of human visual perception and cognition for specific kinds of visual encoding or interaction techniques?
- How does one visualization or interaction technique compare to another as measured by human performance?

6.5.3 UP: Methods and Examples

Controlled experiments. In order to answer evaluation questions with quantitative and statistically significant results, evaluations in the UP group require high precision. The most commonly used methodologies involve an experimental design with only a small number of variables changed between experiment conditions such that the impact of such variables can be measured ([10, p. 28]; [48, p. 156]). Such methods are commonly referred to as *controlled experiments*, *quantitative evaluation*, or *factorial design experiments*. A controlled experiment often requires the abstraction of real-life tasks to simple tasks that can be performed by a large number of participants repeatedly in each study session [60]. Due to the need of a relatively large number of participants, researchers often need to recruit non-experts. As a result, study tasks have to be further abstracted to avoid the need for domain knowledge. Both types of task abstractions may sacrifice realism. One popular reason to study human perceptual and cognitive limits is to explore the design space for visualization and interaction techniques. The outcomes of these studies are usually design guidelines, and in some cases, models. For example, Tory et al. explored the design space of point displays and information landscape displays, dimensionality, and coloring method to display spatialized data [82].

Bartram et al. explored the design space of using motion as a display dimension [5]. Heer and Robertson explored the use of animated transitions in linking common statistical data graphics [29].

Another reason to study human perceptual limits is to study how people perform with specific visualization techniques under different circumstances such as data set sizes and display formats. The goal of the evaluation is to explore the scalability of particular visualization techniques. For example, Yost and North investigated the perceptual scalability of different visualizations using either a 2-megapixel display or with data scaled up using a 32 megapixel tiled display [93]. Another example is Lam et al.'s study to assess effects of image transformation such as scaling, rotation and fisheye on visual memory [44]. In some cases, these experiments can be performed outside of the laboratories. An increasingly popular approach is crowdsourcing with Amazon's Mechanical Turk web service (http://aws.amazon.com/mturk/). Interested readers are directed to a number of validation studies of the method [27, 41].

The second main evaluation goal in UP is to benchmark a novel system or technique with existing counterparts. These are sometimes known as *head-to-head* comparisons as participants perform the same tasks on all study interfaces. Interface effectiveness is usually defined by objective measurements such as time and accuracy. Examples of bench-marking studies in UP include the SpaceTree study, where a novel tree browser was compared with a hyperbolic tree browser and an Explorer-type interface to display tree data in a number of node-finding and navigation tasks [59].

While most study metrics are time and accuracy, researchers are starting to look at different metrics. One example is memorability. Examples include a study on spatial location memory using Data Mountain in the short term [63], and six months later [15]. In cases where quality of work instead of objective measures are used as metrics, expert evaluators are required. One example is Hornbæk and Frokær's study on document visualization, where authors of the documents were asked to determine quality of essays produced by participants [35]. Individual differences may also play a role in user performance [11]. For example, in the evaluation of LifeLines, Alonso et al. looked at the interaction between participants' spatial visualization ability and display format (LifeLines vs. Tabular) in displaying temporal personal history information [1].

Field Logs. Systems can automatically capture logs of users interacting with a visualization. Evaluators analyze these logs to draw usage statistics or single out interesting behaviors for detailed study. This kind of evaluation, especially when performed in web-based environments, has the advantage of providing a large number of observations for evaluation. Also, participants can work in their own settings while data is collected, thus providing a good level of ecological validity. Two recent works used log-based evaluation. Mackinlay et al. [45] used computer logs to evaluate the visual effectiveness of a function inserted into Tableau to suggest users' predefined visual configurations for the data at hand. Viegas et al. [89] examined how their design decisions in ManyEyes have been received after deployment.

6.6 Evaluating User Experience (UE)

Evaluations in the UE group study people's subjective feedback and opinions in written or spoken form, both solicited and unsolicited.

6.6.1 UE: Goals and Outputs

Evaluation of user experience seeks to understand how people react to a visualization either in a short or a long time span. A visualization here may interchangeably be intended as an initial design sketch, a working prototype, as well as a finished product. The goal is to understand to what extent the visualization supports the intended tasks as seen from the participants' eyes and to probe for requirements and needs. Evaluations in UE produce subjective results in that what is observed, collected, or measured is the result of subjective user responses. Nonetheless objective user experience measurements exist, for example, recording user reactions through the use of body sensors or similar means [46]. Interestingly, several subjective measures simply mirror the measures we have in user performance, with the difference that they are recorded as they are perceived by the participant. Examples are: perceived effectiveness, perceived efficiency, perceived correctness. Other measures include satisfaction, trust, and features liked/disliked, etc. The data collected in such a study can help designers to uncover gaps in functionality and limitations in the way the interface or visualization is designed, as well as uncover promising directions to strengthen the system. In contrast to UP (Evaluating User Performance, see Section 6.5), the goal of UE is to collect user reactions to the visualization to inform design. Traditionally, studies in UP are more geared towards the production of generalizable and reproducible results whereas those in UE tend to be specific to the given design problem. While VDAR (Evaluating Visual Data Analysis and Reasoning, see Section 6.2) focuses on the output generated through the data analysis and reasoning process, UE looks more at the personal experience. EWP (Evaluating Environments and Work Practices, see Section 6.1) is similar to UE in that prolonged user observation may take place. Nonetheless, EWP focuses on studying users and their environment whereas UE focuses on a specific visualization.

6.6.2 UE: Evaluation Questions

The main question addressed by UE is: "what do my target users think of the visualization?" More specifically:

1) What features are seen as useful?
2) What features are missing?
3) How can features be reworked to improve the supported work processes?
4) Are there limitations of the current system which would hinder its adoption?
5) Is the tool understandable and can it be learned?

6.6.3 UE: Methods and Examples

Evaluations in this category can take varied forms: they can focus on understanding a small number of users' initial reactions, perhaps in depth (as in case studies) but they can also collect extensive qualitative feedback with statistical relevance, for example, in the form of questionnaires. Evaluations can be short-term to assess current or potential usage and long-term to assess the adoption of a visualization in a real usage scenario. The output consists of data recorded either during or after visualization's use. The data can be the result of indirect expert collection of user experience, as when the evaluator takes notes on observed behaviors, or of direct user feedback as when methods like structured interviews and questionnaires are used.

Informal Evaluation. An informal user feedback evaluation is performed by demoing the visualization to a group of people, often and preferably domain experts, letting them "play" with the system and/or observe typical system features as shown by representatives. The method is characterized by a very limited degree of formalism. For instance, it generally does not have a predefined list of tasks or a structured evaluation script as in usability tests. It is the simplest kind of evaluation and it is, probably for this reason, extremely common. These types of evaluations have been used to: assess "intuitiveness and functionality" [40], "probe for utility and usability" [18], "identify design flaws and users' subjective preferences" [77], "evaluate and improve [our] implementation of the ideas" [17], or "to solicit ideas for improvements and enhancements" [91].

Usability Test. A usability test is carried out by observing how users perform a set of predefined tasks. For each session, the evaluators take notes of interesting observed behaviors, remarks voiced by the user, and major problems in interaction. The set of tasks is usually defined to address a subset of features the designer deems important for the project. What differentiates this method from the other methods is the careful preparation of tasks and feedback material like questionnaires and interview scripts. Its main goal is to perfect the design by spotting major flaws and deficiencies in existing prototypes [22], nonetheless it can also serve the purpose of eliciting overlooked requirements. McGuffin et al. [49] assigned "semi-structured navigation tasks" to a genealogist to evaluate and inform the design of a visual tool used to explore large genealogical trees. Hetzler et al. [31] ran a usability test to "refine the details of the user interaction" of a visual analysis system designed to support constantly evolving text collections. The test was performed with 3 users and a series of assigned questions.

Field Observation. This method is similar to a usability test in that careful observation of users is involved. The observation however happens in a real world setting, where the system under study is used freely. The main goal of field observations is to understand how users interact with the tool in a real setting and thus to derive useful information on how it can be perfected. Often, the information extracted from this kind of study is a series of emergent patterns that can inspire new designs or improve the current one. Sometimes, this kind of study can be followed by a more formal step of questionnaires or interviews to better understand the nature of the observed patterns. An example of field observation is the study of Vizster, a visualization that explore online communities [28], where the authors observed usage in an "installation at a large party" where participants were free to use the developed tool.

Laboratory Questionnaire. The large majority of controlled experiments are followed by a subjective user experiment rating phase where participants typically fill out a questionnaire to solicit their opinions and reactions to the tested visualization. These questions may be closed ended with answers expressed in a five- or seven-point Likert Scale, or open-ended with free answers. While this phase of the evaluation is generally coupled with evaluating user experiment studies, we include it here as the method can be used alone. See Section 6.5 for examples of controlled experiments.

6.7 Automated Evaluation of Visualizations (AEV)

Evaluations in the AEV group study the aspects of visualization that can be measured automatically by a computational procedure.

6.7.1 AEV: Goals and Outputs

This class of evaluation scenarios comprises all methods that employ an automatic computer-based evaluation of visualization. The results of studies in this group usually consist of a series of numbers that represent the visualization quality or efficiency.

6.7.2 AEV: Evaluation questions

Questions in this scenario usually pertain to the visual effectiveness or computational efficiency with which data is represented. Typical questions in this domain are:

1) Is this layout algorithm faster than other state of the art techniques? Under what circumstances?
2) How does the algorithm perform under different volumes of data and number of dimensions?
3) What is the best arrangement of visual features in the visualization to optimize the detection of interesting patterns?
4) What is the extent to which the current visualization deviates from a truthful representation of underlying data?
5) What is the best ordering of visual items to speed up visual search?

6.7.3 AEV: Methods and Examples

Within this class, the evaluation of algorithmic performance plays a large role, especially when the performance of rendering algorithms is particularly relevant for the goals of the visualization. Other methods assess aspects of visual effectiveness by using some kind of computable metric. The metric can for instance represent amount of clutter, the level of optimization in the use of screen space (e. g., to compare different solutions head-to-head), the degree

of organization in a given arrangement or the degree to which the visualizations adhere to an optimal (virtual or real-world) model.

Algorithmic Performance Measurement. The analysis of algorithmic performance is so common in the whole domain of Computer Science that a full discussion of its features is beyond the scope of the paper. Information visualization, by employing algorithms to display data in clever and efficient ways, is of course also often evaluated in terms of algorithmic efficiency. The two most common and established methods in this area are algorithm complexity measures and benchmark tests. In algorithm complexity usually the goal is to demonstrate that the time complexity of the proposed algorithm is better than the state of the art or that the provided solution can be computed in linear or logarithmic time. Less often, but equally important, the complexity of the algorithm is also evaluated in terms of allocated space. Benchmarking happens in a more experimental fashion. Normally the algorithm is evaluated in terms of running time and allocated space over a predefined set of cases (usually several variations of data size or dimensionality) as in the study published by Artero et al. on an algorithm to uncover clusters in parallel coordinates, where the algorithm is tested over several variations of data size and dimensionality. Sometimes the same approach is used to compare the performance of several alternative algorithms, as in the study of Peng et al. on clutter reduction [53], where several algorithms are compared in terms of time performance over different data sizes. In many cases, standard benchmarking data sets are used for evaluation. One example is using graph data sets from the AT&T Graph Library (www.graphdrawing.org) to evaluate graph-drawing algorithms.

Quality Metrics. All studies where automatic evaluation is used share a fundamentally common model: one or more metrics must be devised to assess the quality of a given visualization against an absolute quality level or to compare alternatives. Some studies have proposed *generic* metrics that can be applied, in principle, to any kind of visualization. Edward Tufte's "data-ink ratio" is one notable example [85] that evaluates a visualization in terms efficient use of screen space. Another example is the work by Brath [8] which proposes high-level metrics for a wide set of visualizations to compare data features with visualization features as a way to find appropriate matching. More computationally intensive generic methods also exist. The clutter measure [65] permits evaluation of any generic digital image in terms of clutter. Haroz and Ma [26] proposed a method to measure the extent to which a visualization resembles the representation of a natural picture, arguing that natural visualizations are more aesthetically pleasing.

By contrast, some metrics are designed to evaluate a *specific* technique. Hao et al. [25] use metrics to compare different design solutions in terms of "constancy of display" and "usage of display space" for a data stream monitoring tool. Constancy is measured in terms of changed pixels over time, display space in terms of used pixels in the available space. A similar study [6] compared different sorting algorithms for ordered treemaps. The configurations were compared in terms of "the average aspect ratio of a treemap layout, and the layout distance change function, which quantify the visual problems created by poor layouts."[73]).

7 Discussion

Evaluation is becoming increasingly important in the field of information visualization. The scenarios presented in this paper show the wide range of evaluation goals and questions in which the information visualization research community is currently engaged. These scenarios, and their associated evaluation goals and questions, provide an overview of the types of questions the community has asked of its tools and representations. We also provide information for each scenario about how different evaluation methodologies have been used by different researchers as they work towards discovering answers for their research questions. In this paper, we have contributed a descriptive analysis of the state of evaluation in information visualization. By analyzing several hundred papers, using tags to distinguish between approaches and categorizing and grouping the tagged evaluations into scenarios, goals, and research questions we provide a systematic overview of the diversity of evaluations and research questions that are relevant to information visualization research community.

We started this investigation by tagging evaluation papers from different information visualization venues to get a broader understanding of the types of evaluation goals present in our community. Table 3 lists the 17 tags that were used to code the papers and Appendix A provides descriptions for each tag in more detail. These tags are organized into groups according to the scenario each tag was assigned to. The scenarios are listed by acronym in the order that they are discussed in Section 6. In these scenarios we found two main categories of visualization evaluation focus: (1) the *process* of data analysis, and (2) the assessment of *visualization* use.

When analyzing this data numerically one can see a skewed distribution of papers across the different scenarios. The large majority of evaluations fall into our visualization evaluation group: *Evaluating User Performance–UP* (27%), *Evaluating User Experience–UE*(21%), and *Automated Evaluation of Visualization–AEV* (37%), together contributing to 85% of all evaluation papers. This is in sharp contrast to the 15% in the process scenarios.

The fact that the process visualization group is much less represented in the literature is somewhat surprising as the questions in these group are of high relevance to the field: how can visualization tools be integrated and used in everyday work environments (EWP), how are tasks such as reasoning, knowledge discovery, or decision making supported (VDAR), how does a visualization support communication and knowledge transfer (CTV), and how does a visualization support collaborative analysis (CDA). These questions are of high practical value beyond specific individual tools and can benefit both researchers and practitioners in all areas of visualization.

Several reasons could explain our current evaluation focus. Evaluation in the information visualization community has been following the traditions of Human-Computer Interaction (HCI) and Computer Graphics (CG), both of which also have traditionally focused on controlled experiments, usability evaluations, and algorithm evaluations [22]. Possible questions include: (1) are experiments in the process group simply not being conducted as frequently? (2) does the fact that these types of evaluations are often lengthy requiring field studies, case studies, and extensive qualitative data analysis contribute to their under representation? (3) are we as a community less welcoming to these different–often qualitative–types of evaluations? The lack of evaluations in this group raises questions about whether we as a community should take steps to encourage more evaluations in these groups to be conducted and published.

In the wider HCI community it is comparatively common that publications solely focus on evaluation, often using field and long-term evaluation approaches. In the process evaluation group we frequently used examples from venues outside of the four publications we coded (EuroVis, InfoVis, IVS, Vast) to illustrate scenarios in which these types of methodologies are more common (e. g.[37, 54, 84, 86]). As our community continues to grow we need to think critically about what types of evaluations we would like to see more of and how they can benefit our community. If the current trend continues we will likely see process and longer term qualitative evaluations published at other venues.

For this article, we have coded four main visualization venues and arrived at the codes we used through discussions and several coding passes. Yet, we encourage others to extend our coding or to re-code our results at a later point in time to see how the community has evolved in terms of what kind of evaluation papers are published. Since our coding is based on the published literature it is entirely possible that further coding can reveal new scenarios and questions which we may not have considered here. We encourage others to publish these findings and help to expand our evolving understanding of evaluation in information visualization.

8 Conclusion

Our seven evaluation scenarios encapsulate the current state of evaluation practices in our surveyed papers. From the over 800 papers we surveyed in the EuroVis, InfoVis, and VAST conferences as well as the IVS journal, we found 345 papers that included evaluations. We coded these evaluations according to seventeen tags (see Table 3) and condensed these tags into seven scenarios. The seven scenarios are:

- Evaluating environments and work practices: to derive design advice through developing a better understanding of the work, analysis, or information processing practices by a given group of people with or without software use.
- Evaluating visual data analysis and reasoning: to assess how an information visualization tool supports analysis and reasoning about data and helps to derive relevant knowledge in a given domain.
- Evaluating communication through visualization: to assess the communicative value of a visualization or visual representation in regards to goals such as teaching/learning, idea presentation, or casual use.
- Evaluating collaborative data analysis: to understand to what extent an information visualization tool supports collaborative data analysis by groups of people.
- Evaluating user performance: to objectively measure how specific features affect the performance of people with a system.
- Evaluating user experience: to elicit subjective feedback and opinions on a visualization tool.
- Automated evaluation of visualizations: to automatically capture and measure characteristics of a visualization tool or algorithm.

These scenarios can be used as a practical context-based approach to exploring evaluation options. To briefly reiterate we recommend:

1) **Setting a goal**: start by choosing an evaluation goal. In Section 6, each scenario includes a range of evaluation goals that are characterized by evaluation questions.
2) **Picking suitable scenarios**: the choice of a goal helps identify a relevant scenario.
3) **Considering applicable approaches**: from the scenarios possible methods can be investigated. Referenced examples for each method are provided.
4) **Creating evaluation design and planned analyses**: these methods and examples provide a spring board for designing evaluation methods that fit your research and your research goals. However, since evaluation is still in flux, it is important to keep abreast of new evaluation methods in your considerations.

Our scenario approach can, thus, be used as a starting point for expanding the range of evaluation studies and opens new perspectives and insights on information visualization evaluation. In contrast to other overview articles on evaluation, a major contribution of our work is that we based our evaluation categorization of evaluation questions and goals instead of on existing methods. Our intention is to encourage the information visualization community to reflect on evaluation goals and questions before choosing methods. By providing a diverse set of examples for each scenario, we hope that evaluation in our field will employ a more diverse set of evaluation methods.

Acknowledgments

We would like to thank Adam Perer and Amy Voida for early discussions on structuring evaluation methodologies. We would also like to thank the participants of Beliv 2008 who have contributed material to our early data collection on evaluation methodologies in information visualization.

Paper Tags	EuroVis	InfoVis	IVS	VAST	Scenario
Process					
1. People's workflow, work practices	2	1	0	4	EWP
2. Data analysis	0	3	3	1	VDAR
3. Decision making	0	3	2	3	VDAR
4. Knowledge management	1	1	0	3	VDAR
5. Knowledge discovery	1	1	1	1	VDAR
6. Communication, learning, teaching, publishing	0	0	4	0	CTV
7. Causal information acquisition	0	4	0	0	CTV
8. Collaboration	0	2	2	2	CDA
Visualization					
9. Visualization-analytical operation	0	3	0	0	UP
10. Perception and cognition	15	16	13	1	UP
11. Usability/effectiveness	7	57	33	9	UP&UE
12. Potential usage	1	1	4	4	UE
13. Adoption	0	0	2	0	UE
14. Algorithm performance	77	27	11	0	AEV
15. Algorithm quality	10	8	7	0	AEV
Not included in scenarios					
16. Proposed evaluation methodologies	0	3	0	2	-
17. Evaluation metric development	0	4	0	0	-

TABLE 3
Original coding tags, the number of papers classified, and the final scenario to which they were assigned.

REFERENCES

[1] D. L. Alonso, A. Rose, C. Plaisant, and K. L. Norman. Viewing personal history records: a comparison of tabular format and graphical presentation using lifelines. *Behaviour & Information Technology*, 17(5):249–262, 1998.

[2] K. Andrews. Evaluation comes in many guises. In *CHI workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, pages 7–8, 2008.

[3] K. Baker, S. Greenberg, and C. Gutwin. Heuristic evaluation of groupware based on the mechanicsof collaboration. In *Proceedings of the Conference on Engineering for Human-Computer Interaction*, volume 2254 of *LNCS*, pages 123–139, Berlin, Heidelberg, 2001. Springer Verlag.

[4] L. Barkhuus and J. A. Rode. From mice to men: 24 years of evaluation in chi. In *alt.chi: Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI)*, New York, NY, USA, 2007. ACM.

[5] L. Bartram and C. Ware. Filtering and brushing with motion. *Information Visualization*, 1(1):66–79, 2002.

[6] B. B. Bederson, B. Shneiderman, and M. Wattenberg. Ordered and quantum treemaps: Making effective use of 2d space to display hierarchies. *ACM Transactions on Graphics*, 21(4):833–854, 2002.

[7] E. Bertini, A. Perer, C. Plaisant, and G. Santucci. Beyond time and errors: novel evaluation methods for information visualization (beliv). In *Extended Abstracts of the Conference on Human Factors in Computing Systems (CHI)*, pages 3913–3916, New York, USA, 2008. ACM.

[8] R. Brath. Metrics for effective information visualization. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 108–111, Los Alamitos, USA, 1997. IEEE Computer Society.

[9] I. Brewer, A. M. MacEachren, H. Abdo, J. Gundrum, and G. Otto. Collaborative geographic visualization: Enabling shared understanding of environmental processes. In *Proceedings of the Symposium on Information Vizualization (InfoVis)*, pages 137–141, Los Alamitos, USA, 2000. IEEE Computer Society.

[10] S. Carpendale. Evaluating information visualizations. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, editors, *Information Visualization: Human-Centered Issues and Perspectives*, pages 19–45. Springer LNCS, Berlin/Heidelberg, 2007.

[11] C. Chen and Y. Yu. Empirical studies of information visualization: A meta-analysis. *International Journal of Human-Computer Studies*, 53(5):851–866, 2000.

[12] J. Corbin and A. Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Thousand Oaks, CA, USA, 2008.

[13] L. Costello, G. Grinstein, C. Plaisant, and J. Scholtz. Advancing user-centered evaluation of visual analytic environments through contests. *Information Visualization*, 8:230–238, 2009.

[14] J. W. Creswell. *Research Design. Qualitative, Quantitative, and Mixed Methods Approaches*. Sage Publications, Inc., Thousand Oaks, CA, USA, 2nd edition, 2002.

[15] M. P. Czerwinski, M. V. Dantzich, G. Robertson, and H. Hoffman. The contribution of thumbnail image, mouse-over text and spatial location memory to web page retrieval in 3d. In *Proceedings of INTERACT*, pages 163–170. IOS Press, 1999.

[16] J. Drury and M. G. Williams. A framework for role-based specification and evaluation of awareness support in synchronous collaborative applications. In *Proceedings of the Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 12–17, Los Alamitos, USA, 2002. IEEE Computer Society.

[17] T. Dwyer and D. R. Gallagher. Visualising changes in fund manager holdings in two and a half-dimensions. *Information Visualization*, 3 (4):227–244, 2004.

[18] R. Eccles, T. Kapler, R. Harper, and W. Wright. Stories in geotime. *Information Visualization*, 7(1):3–17, 2008.

[19] G. Ellis and A. Dix. An exploratory analysis of user evaluation studies in information visualization. In *Proceedings AVI Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, 2006.

[20] S. Greenberg, G. Fitzpatrick, C. Gutwin, and S. Kaplan. Adapting the locales framework for heuristic evaluation of groupware. *Australian Journal of Information Systems (AJIS)*, 7(2):102–108, 2000.

[21] S. Greenberg. Observing collaboration: Group-centered design. In T. Erickson and D. W. McDonald, editors, *HCI Remixed: Reflections on Works That Have Influenced the HCI Community*, chapter 18, pages 111–118. MIT Press, Cambridge, Mass, 2008.

[22] S. Greenbery and B. Buxton. Usability evaluation considered

harmful (some of the time). In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 217–224, New York, USA, 2008. ACM.

[23] J. Grudin. Why cscw applications fail: Problems in the design andevaluation of organizational interfaces. In *Proceedings of Computer-Supported Cooperative Work (CSCW)*, pages 85–93, New York, USA, 1988. ACM.

[24] C. Gutwin and S. Greenberg. The mechanics of collaboration: Developing low cost usability evaluation methods for shared workspaces. In *Proceedings of the Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)*, pages 98–103. IEEE Computer Society, 2000.

[25] M. Hao, D. A. Keim, U. Dayal, D. Oelke, and C. Tremblay. Density displays for data stream monitoring. *Computer Graphics Forum*, 27 (3):895–902, 2008.

[26] S. Haroz and K.-L. Ma. Natural visualization. In *Proceedings of the European Symposium on Visualization (EuroVis)*, pages 43–50, Aire-la-Ville, Switzerland, 2006. Eurographics.

[27] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 203–212, New York, USA, 2010. ACM.

[28] J. Heer and danah boyd. Vizster: Visualizing online social networks. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 33–40, New York, USA, 2005. ACM.

[29] J. Heer and G. Robertson. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–1247, 2007.

[30] J. Heer, F. B. Viégas, and M. Wattenberg. Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 1029–1038, New York, USA, 2007. ACM.

[31] E. G. Hetzler, V. L. Crow, D. A. Payne, and A. E. Turner. Turning the bucket of text into a pipe. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 89–94, Los Alamitos, USA, 2005. IEEE Computer Society.

[32] D. M. Hilbert and D. F. Redmiles. Extracting usability information from user interface events. *ACM Computing Survey*, 32(4):384–421, 2000.

[33] U. Hinrichs, H. Schmidt, and S. Carpendale. EMDialog: Bringing information visualization into the museum. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1181–1188, 2008.

[34] K. Holtzblatt and S. Jones. *Contextual Inquiry: A Participatory Technique for Systems Design*. Lawrence Earlbaum, Hillsdale, NJ, USA, 1993.

[35] K. Hornbæk and E. Frokjær. Reading of electronic documents: The usability of linear, fisheye and overview+detail interfaces. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 293–300, New York, USA, 2001. ACM.

[36] P. Isenberg, A. Bezerianos, N. Henry, S. Carpendale, and J.-D. Fekete. CoCoNutTrix: Collaborative retrofitting for information visualization. *Computer Graphics and Applications: Special Issue on Collaborative Visualization*, 29(5):44–57, 2009.

[37] P. Isenberg, A. Tang, and S. Carpendale. An exploratory study of visual information analysis. In *Proceeding of the Conference on Human Factors in Computing Systems (CHI)*, pages 1217–1226, New York, USA, 2008. ACM.

[38] P. Isenberg, T. Zuk, C. Collins, and S. Carpendale. Grounded evaluation of information visualizations. In *Proceedings of the CHI Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, pages 56–63, New York, USA, 2008. ACM.

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

[40] F. Janoos, S. Singh, O. Irfanoglu, R. Machiraju, and R. Parent. Activity analysis using spatio-temporal trajectory volumes in surveillance applications. In *Symposium on Visual Analytics Science and Technology (VAST)*, pages 3–10, Los Alamitos, USA, 2007. IEEE.

[41] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceeding of the Conference on Human Factors in Computing Systems (CHI)*, pages 453–456, New York, USA, 2008. ACM.

[42] O. Kulyk, R. Kosara, J. Urquiza, and I. Wassin. Human-computered aspects. In G. Goos, J. Hartmanis, and J. van Leeuwen, editors, *Lecture Notes in Computer Science*, pages 13–76. Springer, 2007.

[43] H. Lam and T. Munzner. Increasing the utility of quantitative empirical studies for meta-analysis. In *Proceedings of the CHI Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, pages 21–27, New York, USA, 2008. ACM.

[44] H. Lam, R. A. Rensink, and T. Munzner. Effects of 2d geometric transformations on visual memory. In *Proceedings of the Symposium on Applied Perception in Graphics and Visualization (APGV)*, pages 119–126, New York, USA, 2006. ACM.

[45] J. D. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007.

[46] R. Mandryk. *Modeling User Emotion in Interactive Play Environments: A Fuzzy Physiological Approach*. PhD thesis, Simon Fraser University, 2005.

[47] G. Mark and A. Kobsa. The effects of collaboration and system transparency on CIVE usage: An empirical study and model. *Presence*, 14(1):60–80, 2005.

[48] J. McGrath. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000*. Morgan Kaufmann, 1994.

[49] M. J. McGuffin and R. Balakrishnan. Interactive visualization of genealogical graphs. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 17–24, Los Alamitos, USA, 2005. IEEE Computer Society.

[50] M. R. Morris, A. Paepcke, T. Winograd, and J. Stamberger. TeamTag: Exploring centralized versus replicated controls for co-located tabletop groupware. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 1273–1282, New York, USA, 2006. ACM.

[51] T. Munzner. A nested process model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[52] D. C. Neale, J. M. Carroll, and M. B. Rosson. Evaluating computer-supported cooperative work: Models and frameworks. In *Proceedings of Computer-Supported Cooperative Work (CSCW)*, pages 112–121, New York, USA, 2004. ACM.

[53] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 89–96, Los Alamitos, USA, 2004. IEEE Computer Society.

[54] A. Perer and B. Shneiderman. Integrating statistics and visualization for exploratory power: From long-term case studies to design guidelines. *IEEE Computer Graphics and Applications*, 29(3):39–51, 2009.

[55] D. Pinelle and C. Gutwin. A review of groupware evaluations. In *Proceedings 9th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE'00)*, pages 86–91, Los Alamitos, USA, 2000. IEEE Computer Society.

[56] D. Pinelle and C. Gutwin. Groupware walkthrough: Adding context to groupware usability evaluation. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 455–462, New York, USA, 2002. ACM Press.

[57] D. Pinelle, C. Gutwin, and S. Greenberg. Task analysis for groupware usability evaluation: Modeling shared-workspace tasks with the mechanics of collaboration. *ACM Transactions on Human Computer Interaction*, 10(4):281–311, 2003.

[58] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, 2005.

[59] C. Plaisant, J. Grosjean, and B. Bederson. Spacetree: Supporting exploration in large node link tree, design evolution and empirical evaluation. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 57–64, Los Alamitos, USA, 2002. IEEE Computer Society.

[60] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI)*, pages 109–116, New York, USA, 2004. ACM Press.

[61] Z. Pousman, J. Stasko, and M. Mateas. Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1145–1152, 2007.

[62] A. J. Pretorius and J. J. van Wijk. Visual inspection of multivariate graphs. *Computer Graphics Forum*, 27(3):967–974, 2008.

[63] G. Robertson, M. Czerwinski, K. Larson, D. C. Robbins, D. Thiel, and M. van Dantzich. Data mountain: Using spatial memory for

document management. In *Proceedings of the Symposium on User Interface Software and Technology*, pages 153–162, New York, USA, 1998. ACM.

[64] A. Robinson. Collaborative synthesis of visual analytic results. In *Proceedings of the Symposium on Visual Analytics Science and Technology (VAST)*, pages 67–74, Los Alamitos, USA, 2008. IEEE Computer Society.

[65] R. Rosenholtz and J. Mansfield. Feature congestion: A measure of display clutter. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 761–770, New York, USA, 2005. ACM.

[66] J. Sajaniemi and M. Kuittinen. Visualizing roles of variables in program animation. *Information Visualization*, 3(3):137–153, 2004.

[67] P. Saraiya, C. North, and K. Duca. An evaluation of microarray visualization tools for biological insight. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 1–8, Los Alamitos, USA, 2004. IEEE Computer Society.

[68] P. Saraiya, C. North, V. Lam, and K. A. Duca. An insight-based longitudinal study of visual analytics. *Transactions on Visualization and Computer Graphics*, 12(6):1511–1522, 2006.

[69] K. Sedig, S. Rowhani, J. Morey, and H.-N. Liang. Application of information visualization techniques to the design of a mathematical mindtool: A usability study. *Information Visualization*, 2(3):142–159, 2003.

[70] M. Sedlmair, D. Baur, S. Boring, P. Isenberg, M. Jurmu, and A. Butz. Requirements for a mde system to support collaborative in-car communication diagnostics. In *CSCW Workshop on Beyond the Laboratory: Supporting Authentic Collaboration with Multiple Displays*, 2008.

[71] J. Seo and B. Shneiderman. Knowledge discovery in high-dimensional data: Case studies and a user survey for the rank-by-feature framework. *Transactions on Visualization and Computer Graphics*, 12(3):311–322, 2006.

[72] W. Shadish, T. Cook, and D. Campbell. *Experimental and Quasi-Experimental Designs*. Houghton Mifflin Company, 2002.

[73] B. Shneiderman and M. Wattenberg. Ordered treemap layouts. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 73–78, 2001.

[74] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: Multi-dimensional in-depth long-term case studies. In *Proceedings of the AVI workshop on BEyond time and errors: novel evaluation methods for information visualization (BELIV)*, New York, USA, 2006. ACM.

[75] B. Shneiderman and C. Plaisant. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Addison-Wesley, 5th edition, 2009.

[76] T. Skog, S. Ljungblad, and L. E. Holmquist. Between aesthetics and utility: Designing ambient information visualizations. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 233–240, Los Alamitos, USA, 2003. IEEE Computer Society.

[77] H. Song, E. Curran, and R. Sterritt. Multiple foci visualisation of large hierarchies with flextree. *Information Visualization*, 3(1):19–35, 2004.

[78] G. Stahl. *Group Cognition*. MIT Press, 2006.

[79] G. Symon. *Qualitative Research Diaries*, pages 94–117. Sage Publications, 1998.

[80] J. Thomas and K. A. Cook, editors. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society Press, 2005.

[81] M. Tory and T. Moller. Evaluating visualizations: Do expert reviews work? *Computer Graphics and Applications*, 25(5):8–11, 2005.

[82] M. Tory, D. W. Sprague, F. Wu, W. Y. So, and T. Munzner. Spatialization design: Comparing points and landscapes. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1262–1285, 2007.

[83] M. Tory and S. Staub-French. Qualitative analysis of visualization: A building design field study. In *Proceedings of the CHI Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, New York, USA, 2008. ACM.

[84] 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.

[85] E. Tufte. *The visual display of quantitative information*. Graphics Press Cheshire, CN, 1983.

[86] M. Twidale, D. Randall, and R. Bentley. Situated evaluation for cooperative systems. In *Proceedings of Computer-Supported Cooperative Work (CSCW)*, pages 441–452, New York, USA, 1994. ACM Press.

[87] Usability.net. Usability.net methods. Website, 2009. http://www.usabilitynet.org/tools/methods.htm.

[88] F. Viégas, E. Perry, E. Howe, and J. Donath. Artifacts of the presence era: Using information visualization to create an evocative souvenir. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 105–111, Los Alamitos, USA, 2004. IEEE Computer Society.

[89] F. Viégas, M. Wattenberg, F. van Ham, J. Kriss, and M. McKeon. Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1121–1128, 2007.

[90] M. Wattenberg. Baby names, visualization, and social data analysis. In *Proceedings of the Symposium on Information Visualization (InfoVis)*, pages 1–8, Los Alamitos, USA, 2005. IEEE Computer Society.

[91] C. Weaver, D. Fyfe, A. Robinson, D. Holdsworth, D. Peuquet, and A. M. MacEachren. Visual exploration and analysis of historic hotel visits. *Information Visualization*, 6(1):89–103, 2007.

[92] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1129–1136, 2007.

[93] B. Yost and C. North. The perceptual scalability of visualization. *IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization 2007)*, 13(6):837–844, 2007.

[94] T. Zuk, L. Schlesier, P. Neumann, M. S. Hancock, and S. Carpendale. Heuristics for information visualization evaluation. In *Proceedings of the AVI Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization (BELIV)*, pages 55–60, New York, USA, 2006. ACM.

APPENDIX A
TAGS USED IN OPEN CODING

We developed our seven scenarios based on the following 17 tags. These tags were used to open code publications from four venues. The distribution of publication by tags and venue is listed in Table 3.

1) **Data analysis:** Evaluate how visualization is used in exploratory data analysis to generate hypotheses.
2) **Decision making:** Evaluate how visualization is used to confirm or refute solutions to problems or hypotheses.
3) **Collaboration:** Evaluate how visualization supports collaboration activities.
4) **Adoption:** Observe how a visualization is adopted after deployment.
5) **Communication, learning, teaching, publishing:** Study how visualization is used in multiple forms of communication activities.
6) **Usability/Effectiveness:** Solicit usability feedback or determine effectiveness of visualization based on user performance.
7) **People's workflow, work practices:** Understand potential users' work environment and practices.
8) **Perception and cognition:** Study low-level human perception and cognition to evaluate or explore the visualization design space.
9) **Algorithm performance:** Study efficiency and performance of algorithm behind visualizations.
10) **Knowledge discovery:** Study how visualization support knowledge discovery.

11) **Potential usage:** Solicit users' opinions on how the visualization may be useful.
12) **Proposed Evaluation Methodologies:** Propose new methodologies on evaluation.
13) **Visualization-analytical operation:** Study how visualization affects users' performance of simple visual tasks.
14) **Causal information acquisition:** Study how users causally acquire information, especially in ambient displays.
15) **Evaluation metrics development:** Propose new evaluation metrics.
16) **Algorithm quality:** Evaluate algorithms when compared to accepted gold standards such as human judgments.
17) **Knowledge management:** Evaluate how effectively does the system support management of knowledge generated in the sense-making loop.

Scenario	Description	Questions	Example Methods
EWP: Evaluating Environments and Work Practices	Derive design advice through an understanding of the work, analysis, or information processing practices by a given group of people with our without software use	What is the context of use of visualizations? In which daily activities should the visualization tool be integrated? What are the characteristics of the identified user group and work environments? What data is currently used and what tasks are performed on it? What kinds of visualizations are currently in use? How do they help to solve current tasks? What challenges and usage barriers can we see for a visualization tool?	Field Observation Interviews Laboratory Observations
VDAR: Evaluating Visual Data Analysis and Reasoning	Assess a how an information visualization tool supports supports analysis and reasoning about data and helps to derive relevant knowledge in a given domain	How does a visualization or tool support…data exploration?; processes aimed at seeking information, searching, filtering, and reading and extracting information?; knowledge discovery?; the schematization of information or the (re-)analysis of theories?; hypothesis generation?; interactive hypothesis examination?; decision making?; ommunication and application of analysis results?	Case Studies Controlled Experiments.
CTV: Evaluating Communication through Visualization	Assess the communicative value of a visualization or visual representation in regards to goals such as teaching/learning, idea presentation, or casual use	Do people learn better and/or faster using the visualization tool? Is the tool helpful in explaining and communicating concepts to third parties? How do people interact with visualizations installed in public areas? Are they used and/or useful? Can useful information be extracted from a casual information visualization?	Controlled Experiments Field Observation and Interviews.
CDA: Evaluating Collaborative Data Analysis	Understand how (well) an information visualization tool supports collaborative team work and data analysis by groups of people	Does the tool support *effective and efficient* collaborative data analysis? Does the tool *satisfactorily* support or stimulate group analysis or sensemaking? Does the tool support group insight? Is social exchange around and communication about the data facilitated? How is a collaborative visualization system used? How are certain system features used during collaborative work? What are patterns of system use? What is the process of collaborative analysis? What are users' requirements?	Heuristic Evaluation Log Analysis Field or Laboratory Observation.
UP: Evaluating User Performance	Objectively measure how specific features affect the performance of people with a system	What are the limits of human visual perception and cognition for specific kinds of visual encoding or interaction techniques? How does one visualization or interaction technique compare to another as measured by human performance?	Controlled Experiments Field Logs.
UE: Evaluating User Experience	Elicit subjective feedback and opinions on a visualization tool	What features are seen as useful? What features are missing? How can features be reworked to improve the supported work processes? Are there limitations of the current system which would hinder its adoption? Is the tool understandable and can it be learned?	Informal Evaluation Usability Test Laboratory Questionnaire.
AEV: Automated Evaluation of Visualizations	Automatically capture and measure characteristics of a visualization tool or algorithm	Is this layout algorithm faster than state of the art techniques? Under what circumstances? How does the algorithm perform under different volumes of data and number of dimensions? What is the best arrangement of visual features in the visualization to optimize the detection of interesting patterns? What is the extent to which the current visualization deviates from a truthful representation of underlying data? What is the best ordering of visual items to speed up visual search?	Algorithmic Performance Measurement Quality Metrics.

TABLE 4
Summary table of the proposed scenarios.