Learning-Based Evaluation of Visual Analytic Systems

Remco Chang
UNC Charlotte
rchang@uncc.edu

Caroline Ziemkiewicz
UNC Charlotte
caziemki@uncc.edu

Roman Pyzh
UNC Charlotte
rpyzh@uncc.edu

Joseph Kielman
Dept. of Homeland Security
joseph.kielman@dhs.gov

William Ribarsky
UNC Charlotte
ribarsky@uncc.edu

ABSTRACT
Evaluation in visualization remains a difficult problem because of the unique constraints and opportunities inherent to visualization use. While many potentially useful methodologies have been proposed, there remain significant gaps in assessing the value of the open-ended exploration and complex task-solving that the visualization community holds up as an ideal. In this paper, we propose a methodology to quantitatively evaluate a visual analytics (VA) system based on measuring what is learned by its users as the users reapply the knowledge to a different problem or domain. The motivation for this methodology is based on the observation that the ultimate goal of a user of a VA system is to gain knowledge of and expertise with the dataset, task, or tool itself. We propose a framework for describing and measuring knowledge gain in the analytical process based on these three types of knowledge and discuss considerations for evaluating each. We propose that through careful design of tests that examine how well participants can reapply knowledge learned from using a VA system, the utility of the visualization can be more directly assessed.

Categories and Subject Descriptors
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

General Terms
Human Factors.

Keywords
Evaluation Methodology. Learning. Visualization.

1. INTRODUCTION
"The goal of visualization is to gain insight and knowledge." This statement has been echoed in numerous publications in various forms, dating back to the influential ViSC report in 1987 [6] to the recently published research agenda for visual analytics, *Illuminating the Path*, in 2005 [13]. Over these two decades, the goal of visualization design remains the same: to create an interactive visual form of data so that users can learn something about either the data itself or the process of solving a specific problem.

BELIV 2010, April 10-11, 2010, Atlanta, GA, USA.

While the field of visualization has grown considerably since the ViSC report, we as a visualization community unfortunately are not yet able to definitively answer the question of how our tools assist users in gaining insight or knowledge. Although many success stories exist, we are still in search of a comprehensive evaluation methodology to determine the value of visualization in terms of its goal of facilitating insight and knowledge gain. As noted by George Robertson, measuring task completion time, errors, subjective preferences, etc. have been the default practices, but these measurements have also failed to fully characterize the analytic utility of visualization [9].

In searching for a new evaluation methodology, we first note that in scientific experiments having a hypothesis is the first step in designing the experiment. In the case of evaluating visualization systems, given the goal of visualization design, the hypothesis should naturally be, "is my visualization helping the user gain insight or knowledge?" Unfortunately, in most cases, testing this hypothesis has proven difficult because knowledge and insight are inherently difficult to define [1, 2, 14]. Experts in evaluation of visualizations, therefore, turn to more repeatable and defendable measurements such as task-completion time, errors, subjective preferences, etc. [9].

In this paper, we propose that while knowledge and insight are difficult to define, it is still possible to indirectly measure them to determine the value of a visualization. For example, North et al. have suggested measuring the number of insights gained by a user when using a visualization [7, 10] by asking the user to click a button whenever a discovery (i.e., an insight) is made. Similarly, Scholtz suggested a productivity-based metric that records the number of documents (the amount of information) examined by a user during a session [11]. In both cases, the goals of the methodologies are the same in that they seek to measure how much information or knowledge is gained by the user. However, neither approach has gained widespread adoption. In insight-based evaluation, this is because there is too much variability between users on what constitutes a discovery. Similarly, in the productivity-based methodology, the metric only tells how much information is presented to the user, but not how much of that information the user absorbs.

Based on the shortcomings of previous approaches, we propose a new "learning-based evaluation methodology" for testing the hypothesis of whether or not a user has gained insight or knowledge from using a visualization. Our approach differs from

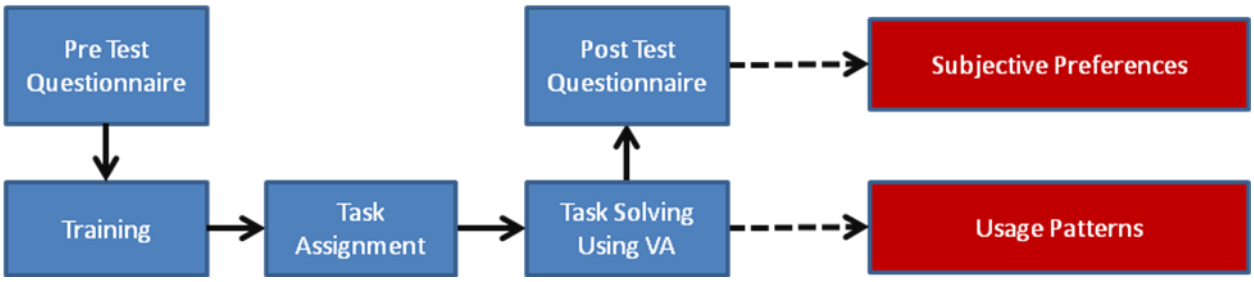

Figure 1. A pipeline for typical visualization evaluations

traditional evaluation methodologies in that we do not measure the user's performance when solving a task, but instead propose that the emphasis should be on evaluating how well a user can solve a new task after spending time using a visualization. We further categorize the types of learning that may occur while using a VA system into knowledge about how to use the system, what the particulars of the dataset are, and, more broadly, how to solve the task or other similar tasks. Furthermore, we argue that clearly distinguishing among these types of learning is important for assessing the utility of a visualization.

We note that the concept of a learning-based approach is similar in many ways to existing evaluation methodologies such as insight-based techniques [7, 8, 10] or utility and productivity-based metrics [11]. Unlike these methodologies, our proposed approach does not clearly distinguish or identify specific insights or piece of knowledge gained by the user, or when and how the user gains such insight and knowledge. Instead, the key point of our proposed approach is an emphasis on evaluating VA systems based on how well a user can reapply knowledge learned. As we discuss in this paper, many evaluation experts have designed evaluations and experiments that are similar to our proposed approach in spirit, but the emphasis has not been to explicitly demonstrate the utility of a visualization through measuring the user's reapplication of knowledge. The purpose of this work is to focus on identifying a methodology and present the challenges of designing evaluations based on such outcome.

We speculate that the results obtained using our methodology would have more applicability in real-world scenarios. For example, government agencies such as the Department of Homeland Security (DHS) do not adopt new technologies flippantly for their intelligence analysis tasks. Instead, great care and consideration are given to determining the benefits and potential dangers prior to adopting the technology. In the case of VA systems, an evaluation based on task-completion time or subjective preferences only indicates a small aspect of the value of the tool. With additional information such as how well the VA system facilitates learning and gaining knowledge, the utility of the visualization becomes more obvious and direct.

2. OVERVIEW OF LEARNING-BASED EVALUATION

The concept of a learning-based evaluation is not new. In fact, learning-based evaluations take place in all classrooms around the world every day: a student learns a subject from taking a class, and the teacher evaluates how much the student has learned by giving the student an exam. In many classrooms, students are not evaluated on the method of learning but rather, on knowledge gained as reflected by the final grade, which is based on the results of the exam. The students could have used a number of different textbooks, digital information sources, or relied solely on the lectures. The teacher's goal is primarily to make sure that the student gains knowledge on the subject matter using any available methods or resources.

To put this scenario in the context of evaluating VA systems, most existing evaluation approaches focus on how the student learns as opposed to how much the student has learned. As an analogy, most evaluation methodologies today would be similar to giving a student (a user) a book (the VA system), and testing how quickly the student can find certain passages or information (task-completion time), followed by asking the student to give a rating of the book (subjective preference) [9].

It is not difficult to see that the results of this type of evaluation do not answer how much the student has learned from using the book, but instead measures how well the book is organized (the interface of VA system). Separating the two is not a trivial task in evaluation design, but it is a task we must attempt to solve. We propose that if the goal of evaluating a VA system is to determine how much knowledge a user has gained or learned from using it, we need to adopt the evaluation methodologies used by teachers in every classroom by focusing on giving specific tests to determine what and how much the user has learned after they have had a chance to use the VA system.

However, unlike the classroom scenario, evaluating a VA system can be more complicated. In pursuing a learning-based evaluation methodology for visualizations, one question that must be answered is what kind of learning we want to measure. Existing visualization evaluation techniques do not clearly separate knowledge learned about the data, knowledge learned about the task, and knowledge learned about the system, which can potentially confound the results of an evaluation study. In addition, these three knowledge areas represent different kinds of learning in a cognitive sense, and should be considered in this light during evaluation design.

2.1 Learning About the Data

The most basic type of learning that occurs in visualization use is learning about the dataset being visualized. While in some ways less complex than the other types of learning we consider, this is often the most important kind of learning from a user's perspective. For instance, in designing a VA system that explores a large database with billions of records (such as a database of IP logs), the ability to learn new knowledge about the data or the phenomena they describe is the primary goal.

The type of knowledge gained from learning about the data is likely to be largely declarative. That is, information about the data

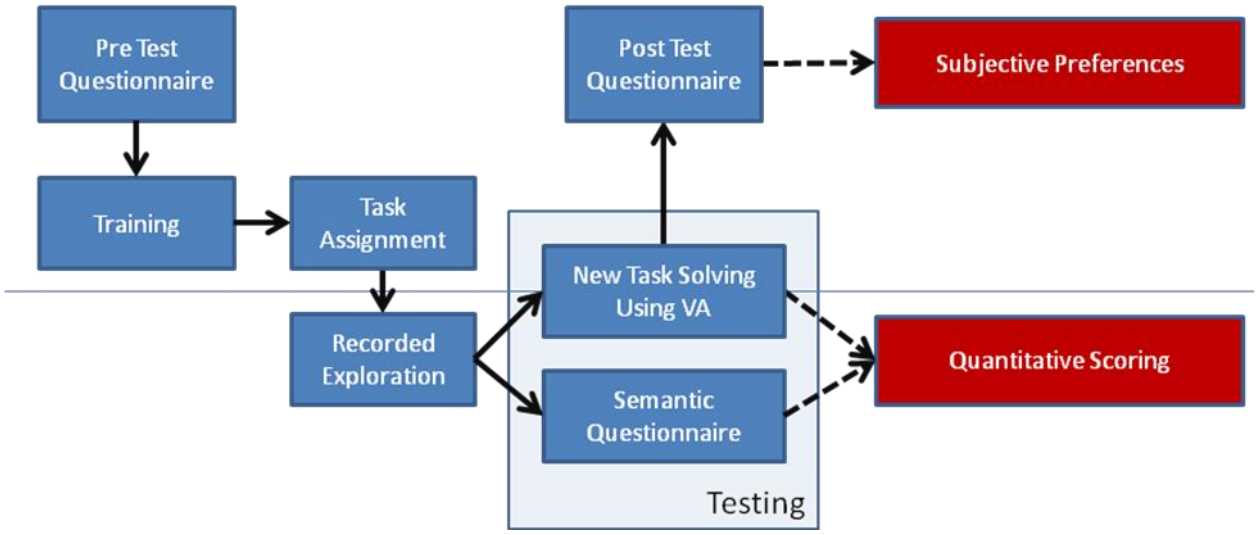

Figure 2. A pipeline for knowledge-based visualization evaluations

will be in the form of explicitly learned facts and events. This kind of knowledge can be measured through testing or by directly asking users what they have learned. Unfortunately, in the context of visual analytics, creating questions regarding a complex dataset that is not fully understood can be a challenge on its own. In some cases, domain experts can generate questions that are relevant to specific tasks and can also "grade" the responses based on the depth and breadth of the analysis results. However, in cases where the characteristics of the dataset are truly foreign to experts, we propose that the visualization be evaluated using standard datasets (such as the dataset from the VAST challenge) where there are known answers to analytical questions.

2.2 Learning About a Task

In most intelligence analyses, learning about a task implies expertise in solving a specific type of problem, and is perhaps the most important kind of knowledge to gain in certain related domains. In this framework, we wish to distinguish between learning the answer to a specific task question (which would fall under "learning about the data") and learning how to do a certain class of task. That is, we are interested in the kind of knowledge gain that leads to the learning of a skill.

Ideally, the problem-solving skills learned in visualization use should be transferable, so that they can be applied to multiple problems of the same general kind. Identifying such classes of problems will likely require the input of domain experts, much as expertise in a subject is needed to write an exam for it.

2.3 Learning About the VA System

When we say that the goal of visualization is to gain knowledge, it is usually implicit that this knowledge is something deeper than knowledge about the VA system itself. Nonetheless, studying the extent to which a user can learn how to use a system is an important part of any evaluation process. In the case of a VA system, learning how to use the system is a clear prerequisite for learning about either the data or the task. A user cannot gain knowledge from a system if they don't understand how to use it.

3. LEARNING-BASED EVALUATION METHODOLOGY

When evaluating a VA system, the researcher may be interested in different types of learning, depending on the research questions being asked. In practice, a user will likely want to learn information about the data as well as how to solve problems involved in the data domain. At the same time, it is also important to show that a user can learn how to use a specific visualization if it is to be considered useful. Therefore, when assessing whether VA systems are useful for knowledge building, we should know how to address all three of these major types of learning.

In many ways, the typical pipeline used in visualization evaluation already addresses parts of this broader model of knowledge gain. Figure 1 shows an example of an evaluation pipeline that includes some common steps in many visualization evaluations, starting with pre-test questionnaires and training. Training is typically a time period allocated to allow the user to become familiar with a particular interface or functionalities of a new tool. During this time, a user is learning about how to use the tool and the analysis process. After training, a user is normally given some specific tasks and asked to use a visualization to accomplish them. Depending on the experiment, the user's performance and usage characteristics could be captured in this stage, often in a quantitative manner (such as task-completion time). Finally, in some experiments, a post-test questionnaire is given at the end so that additional qualitative, subjective information about the user's experience could be captured.

The fact that these types of knowledge gain are to some extent already a part of the evaluation pipeline suggests that they are a good model for the kind of questions that naturally arise about a visualization's utility. We propose that an evaluation pipeline more specifically aimed at these different knowledge types would give a more concrete measure of the practical use of a visualization. Figure 2 shows the pipeline of our proposed learning-based evaluation methodology. It is loosely based on the traditional methodology, but includes two new stages: recorded exploration and testing. In the testing phase, two possible

methods can be used: semantic questionnaire and novel task solving.

3.1 Recorded Exploration

In place of using a VA system to solve specific tasks (while measuring performance or usage patterns), in this stage we will simply ask the user to explore the data. This exploration process is designed to allow the user to interact with the VA system and learn from this process. We distinguish this step from the training phase in that the tasks given to the user are real and the user is encouraged to identify a solution.

However, the user's performance during this phase is not measured in terms of task accuracy or speed. Rather, the purpose of this stage is to record a user's analysis process and findings during a natural task. While exploring data with a new tool, a user may be gaining knowledge about the system and how to solve tasks, even when not making progress towards a solution. This kind of open-ended learning is not particularly prominent in traditional visualization evaluation, but the nature of a user's unstructured experience with a system is nonetheless hugely important to his judgment of it and his ability to incorporate it into his work.

Since previous work has shown that properly designed process capturing methods can successfully record a user's findings and strategies while performing an analytical task [3], we propose a greater role for this kind of analysis in studying a user's learning about a visualization system. By capturing the exploration process, we can better understand not just what a user learns but how and why he learns it. While this proposed learning-based approach does not specifically emphasize measuring or capturing the speed or accuracy of the user's analysis, we note that there is a great deal of information within those quantitative measures. However, within the context of evaluating a VA system through reapplying learned knowledge, such statistics are less directly relevant (see section 5.2 for additional discussions on combining these statistics into our proposed learning-based approach).

3.2 The Testing Phase

The testing phase seeks to identify how much the user has learned from the recorded exploration step, and we propose that two different methods can be used: semantic questionnaires and new task solving. The analogy of these two steps to the classroom example would be the difference between an open-book (new task solving) versus a closed-book (semantic questionnaire) exam. Under new task solving, the user is asked to reapply what he has learned in the exploration step to solving a new task using the visualization. Testing with a semantic questionnaire, on the other hand, directly asks the user to answer questions that indicate how much he has learned.

3.3 New Task Solving

In addition to the untested exploration phase, our pipeline also includes a more traditional phase in which the user solves novel tasks using the VA system. The purpose of giving the user new tasks is to evaluate how well the user can perform transfer of knowledge or skills. Specifically, how well can the user apply the knowledge learned during the exploration phase to solving the new tasks? This phase measures the skills acquired by the user by directly testing his ability to use the tool to solve problems.

This part of our pipeline is very similar to the testing phase in the traditional pipeline, although we suggest a greater focus on specifically testing analytical skills which are relevant to the user's work, rather than simply measuring the user's ability to read data values from the visualization.

3.4 Semantic Questionnaire

The other testing method of our pipeline is focused on declarative knowledge. The ability to transfer knowledge is not a definitive method for measuring knowledge; some users may gain a great deal of semantic knowledge, but might not be able to transfer or apply it. In parallel to measuring knowledge about the system and the task, we propose a questionnaire methodology for testing the amount of semantic knowledge gained. The key point to the questionnaire is that the questionnaires need to be specifically designed by domain experts to directly assess how much of what the user learned from the exploration phase is indeed new knowledge. This path assumes that the there are "known solutions" to the task and questions given to the user in the exploration phase, and through the use of these questionnaires, a VA system is evaluated based on how well it assists the user in finding and learning these solutions.

4. APPLICATION SCENARIO

We examine how this learning-based evaluation methodology could be applied to evaluating real world VA systems. The specific scenario we examine is the rating of VAST contest submissions, particularly from the perspective of the effectiveness of the use of visualizations in assisting their users in learning the analytic process[1].

The current method of scoring the contest submissions for their utility is to examine videos and text explanations of how the systems are used for analyzing the contest dataset. The scores from all reviewers are averaged to determine the final score of the system. Under this scenario, biases based on clarity of the video or the text explanations could be introduced that inadvertently affect the reviewers' scores.

Under our proposed learning-based evaluation methodology, we will seek to directly identify the utility of the systems by testing how well the systems perform under a new dataset. Specifically, we propose a two-staged process for evaluating the submissions. The first stage, which is the equivalent of the "exploration" phase in Figure 2, will consist of contestants downloading a training dataset from the VAST website. The VA system and analytic methods developed by the contestants will be based on this training dataset. The second stage is based on the "testing" phase of Figure 2, in which the contestants are given a new dataset that is similar to the training data in style, format, and content. The contestants are then asked to analyze this new dataset using their submitted VA system. All contestants will perform the second stage at the same location (e.g. at VisWeek) under the same conditions. The time and accuracy of the contestants' analysis results in the second phase will determine the sufficiency of their analytical capabilities when using their VA system.

Of the three different types of learning (learning about the VA system, data, and task), this proposed evaluation methodology specifically tests how well the contestants have learned the analysis task. Since the contestants design their own VA systems,

[1] Scoring the VAST contest submissions is not limited to effectiveness, but also includes other considerations such as novelty, analytic understanding and thoroughness, etc.

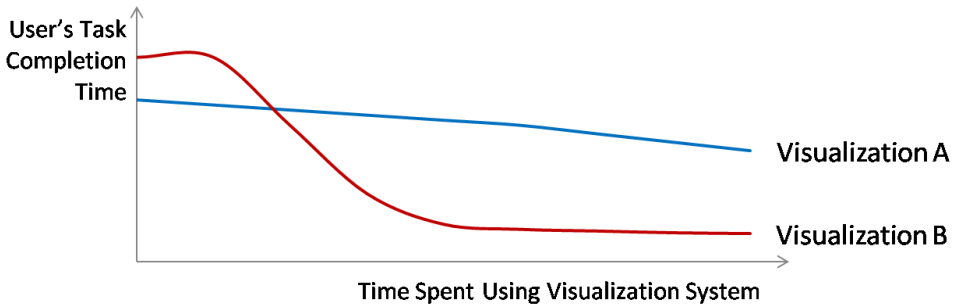

Figure 3. Usage of two visualization systems over time

the results of the testing phase would not include their familiarity with the visualizations. Similarly, since the training dataset is different from the real dataset, the contestants could not apply the specific declarative knowledge they have learned about the training data. Provided the training set and test set are designed to be sufficiently different while still representing realistic scenarios, the accuracy of the contestants' analysis result in the testing phase should be an indication of how well they have learned the task and how well they can utilize their VA system in solving it.

5. LIMITATIONS AND EXTENSIONS

While we propose that this evaluation methodology has the potential to determine the utility of a VA system, it does have some drawbacks that require further refinement. Specifically, we note three areas of our proposed methodology that could be improved: isolating the types of learning, determining the factors for improving the visualization, and understanding the long-term effects of a system.

5.1 Isolating Learning Types

One key challenge in applying our proposed evaluation methodology is in interpreting the results. The testing phase (Figure 2) could quantitatively determine if a user has learned from the exploration process. However, in some cases, it could be difficult to isolate what type of learning contributed to the results, as well as how to isolate how much learning arose solely from the visualization and how much was driven by the user's existing knowledge. The most complicated case is in differentiating the user's increasing familiarity with the visualization (learning about the visualization system) from the user gaining knowledge in the data or the analytical task.

Isolating the differences between learning about data and learning about a task is a simpler problem in that the evaluator can use a new, but similar dataset in the testing phase (as in proposed scenario for evaluating VAST contest submissions). However, in order to separate out the user's learning the VA system from learning about the data or task requires additional understanding of the user's performance and expertise with the VA system over repeated use. There is research in the HCI community that demonstrates how a user's familiarity with a VA system could be recovered through logging and analyzing the user's interactions [4], and incorporating similar methods would help in isolating the different types of learning using our proposed evaluation methodology. Regardless, isolating the type of knowledge learned remains a major challenge of this approach, and additional research will be necessary to quantify the value of a VA system based on the combination of all three types of knowledge.

5.2 Improving the VA System

While the key strength of the proposed evaluation methodology lies in its ability to ignore the design principles or particular types of visualizations used in a system while focusing on overall utility, this aspect could also be its greatest weakness. In particular, without understanding how a system is designed, or how to relate the outcome of the evaluation to the features of a system, it could become difficult to identify features in the VA system that need to be improved.

The qualitative aspect of evaluating a visualization has been discussed in the BELIV community. Most notably, Isenberg et al. [5] proposed "grounded evaluation" in which they presented a methodology where design, implementation, and evaluation occur in a cyclic fashion. We posit that this type of evaluation methodology, which directly informs the design and implementation of a visualization, could be integrated with our proposed methodology with little conflict. Specifically, we foresee such qualitative measurements being gathered and utilized during the design and implementation phases, while our proposed evaluation methodology could be used after the completion of the VA system for determining its utility or for comparing one VA system to another. With the two evaluation approaches applied in sequence, we propose that it is possible to understand both the qualitative and quantitative values of a VA system.

In addition, the combined use of grounded evaluation and the proposed learning-based evaluation methodologies can lead to a more precise understanding of how each user uses a VA system. In evaluating VA systems, each user's innate skills in visual problem solving will directly affect the analysis outcome. Specific to the proposed learning-based evaluation, how well and how quickly each user learns and absorbs knowledge will become an equally important factor. Combining grounded evaluation and the proposed learning-based methodologies to understand the characteristics of each individual user will enable us to examine the results of each user's interaction with VA system more precisely and more holistically.

5.3 Long Term Effects of a Visualization

An important component of the value of a visualization lies in its long-term repeated use [13]. The question of how a user learns from a visualization over time, regardless of learning type, remains an open question that requires further investigation.

Figure 3 shows an example of two fictional VA systems as they are used over time. The blue line represents a VA system that is easy to learn initially and initially out-performs the red line.

However, in a longitudinal study, the VA system represented as the red line would eventually out-perform the blue line.

There could be many explanations for this. For example, the blue VA system could have a simpler, more intuitive interface. However, it does not provide the same degree of learning as the red one over time. Conversely, the red VA system might provide so much capability and information that the user initially spends a large amount of time exploring the data or the task. However, after the user becomes familiar with the analytical process of solving this problem, the user's efficiency increases. Regardless of the reason, the point of this illustration is to demonstrate the importance of evaluating the utility of a VA system over repeated use, similar to the strategies outlined by Shneiderman and Plaisant [12]. In order to understand the true utility of a VA system, our proposed evaluation methodology ought to be carried out in a repeated and controlled fashion to measure its effect over time.

6. CONCLUSION

In this paper, we propose a new evaluation methodology called learning-based evaluation. This methodology is motivated by the fact that while the visualization community has identified the value of visualization to be its ability to facilitate learning, most existing evaluation methodologies do not directly measure how much the user has learned through the use of a VA system. Our proposed methodology is loosely based on existing methodologies, but emphasizes the importance of user exploration and suggests two methods for determining the amount of knowledge gained by the user. We further distinguish three types of learning: learning about the VA system, data, and task. We propose that with careful design, this learning-based evaluation methodology could be integrated with other existing methodologies to complement each other.

7. ACKNOWLEDGMENTS

This material is based in part upon work supported by the International Program of the Department of Homeland Security under grant number 2009-ST-108-000007. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of Homeland Security.

8. REFERENCES

[1] Chen, M., Ebert, D., Hagen, H., Laramee, R. S., van Liere, R., Ma, K., Ribarsky, W., Scheuermann, G., and Silver, D. 2009. Data, Information, and Knowledge in Visualization. IEEE Comput. Graph. Appl. 29, 1 (Jan. 2009), 12-19.

[2] Chang, R., Ziemkiewicz, C., Green T. M., and Ribarsky, W. 2009. Defining insight for visual analytics. IEEE Computer Graphics and Applications, 29(2):14–17.

[3] Dou, W., Jeong, D.H., Stukes, F., Ribarsky, W., Richter Lipford, H., Chang, R. 2009. Recovering Reasoning Processes from User Interactions. IEEE Computer Graphics and Applications, vol. 29, no. 3, pp. 52-61.

[4] Hilbert, D. M. and Redmiles, D. F. 2000. Extracting usability information from user interface events. ACM Comput. Surv. 32, 4 (Dec. 2000), 384-421.

[5] Isenberg, P., Zuk, T., Collins, C., and Carpendale, S. 2008. Grounded evaluation of information visualization. In Proceedings of the 2008 Conference on Beyond Time and Errors: Novel Evaluation Methods For Information Visualization (Florence, Italy, April 05, 2008). BELIV '08. ACM, New York, NY.

[6] McCormick, B. H., DeFanti, T. A., and Brown, M. D. (eds). 1987. Visualization in Scientific Computing. ACM SIGGRAPH, New York, NY.

[7] North, C. 2006. Toward measuring visualization insight. IEEE Computer Graphics and Applications, 26(3):6–9.

[8] Perer, P., and Shneiderman, B. Integrating Statistics and Visualization: Case Studies of Gaining Clarity During Exploratory Data Analysis. ACM Conference on Human Factors in Computing Systems (CHI 2008). Florence, Italy. (2008).

[9] Robertson, G. 2008. Beyond time and errors – position statement. In Proceedings of the 2008 Conference on Beyond Time and Errors: Novel Evaluation Methods For Information Visualization (Florence, Italy, April 05, 2008). BELIV '08. ACM, New York, NY.

[10] Saraiya, P., North, C., and Duca, K. 2005. An insight-based methodology for evaluating bioinformatics visualizations. IEEE Transactions on Visualization and Computer Graphics, 11(4):443–456.

[11] Scholtz, J. 2008. Progress and Challenges in Evaluating Tools for Sensemaking. ACM Computer Human Information (CHI) conference Workshop on Sensemaking in Florence, Italy, April 6, 2008.

[12] Shneiderman, B. and Plaisant, C. 2006. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In Proceedings of the 2006 AVI Workshop on Beyond Time and Errors: Novel Evaluation Methods For Information Visualization (Venice, Italy, May 23, 2006). BELIV '06. ACM, New York, NY.

[13] Thomas, J. J. and Cook, K. A. (eds). 2005. Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE.

[14] van Wijk, J. J. 2005. The value of visualization. Visualization. VIS 05. IEEE, 79–86.