Evaluation of visual analytics environments: The road to the Visual Analytics Science and Technology challenge evaluation methodology

Jean Scholtz[1]
Catherine Plaisant[2]
Mark Whiting[1]
Georges Grinstein[3]

1Pacific Northwest National Laboratory, Richland, WA, USA
2University of Maryland, College Park, MD, USA
3University of Massachusetts Lowell, Lowell, MA, USA

Abstract
Evaluation of software can take many forms ranging from algorithm correctness and performance to evaluations that focus on the value to the end user. This paper presents a discussion of the development of an evaluation methodology for visual analytics environments. The Visual Analytics Science and Technology (VAST) Challenge was created as a community evaluation resource, that is, a resource available to researchers and developers of visual analytics environments that would allow them to test out their designs and visualizations and compare the results with the solution and the entries prepared by others. Sharing results allows the community to learn from each other and to hopefully advance more quickly. In this paper we discuss the original challenge and its evolution during the seven years since its inception. While the VAST Challenge is the focus of the paper, there are lessons for many involved in setting up a community evaluation program, including the need to understand the purpose of the evaluation, decide upon the right metrics to use and the appropriate implementation of those metrics including datasets and evaluators. For ongoing evaluations it is also necessary to track the evolution and to ensure that the evaluation methodologies are keeping pace with the science being evaluated. The discussions on the VAST Challenge on these topics should be pertinent to many interested in community evaluations.

Keywords
Community evaluations, visual analytics environments, utility evaluations

To appear in *Information Visualization* 13, 4 (2014)

Introduction

The evaluation of visual analytics environments was a topic in the R&D roadmap for visual analytics, *Illuminating the Path*, as a critical aspect of moving research into practice.[1] For a thorough understanding of the utility of the systems available, evaluation not only involves assessing the visualizations, interactions or data processing algorithms themselves, but also the complex processes that a tool is meant to support (such as exploratory data analysis and reasoning, communication through visualization, or collaborative data analysis.) [2,3] Researchers and practitioners in the field have long identified many of the challenges faced when planning, conducting, and executing an evaluation of a visualization tool or system.[4]

Evaluation is needed to verify that algorithms and software systems work correctly and that they represent improvements over the current infrastructure. Additionally to effectively transfer new software into a working environment, it is necessary to ensure that the software has utility for the end-users and that the software can be incorporated into the end-user's infrastructure and work practices. Evaluation test beds require datasets, tasks, metrics, and evaluation methodologies. As noted it is difficult and expensive for any one researcher to set up an evaluation test bed so in many cases evaluation is setup for communities of researchers or for various research projects or programs.[1] Examples of successful community evaluations can be found in such areas as message understanding, information retrieval, and facial recognition.[5, 6, 7]

As visual analytics environments are intended to facilitate the work of human analysts, one aspect of evaluation needs to focus on the utility of the software to the end-user. The end-users of visual analytics software come from such domains as intelligence analysis, law enforcement, and finance. It is difficult to gain access to these end-users due to the sensitive nature of data and tasks and the busy schedules of such analysts.

User-centered design goes beyond evaluation and starts with the user. [8, 9] Having some knowledge of the type of data, tasks, and work practices helps researchers and developers know the correct paths to pursue in their work. When access to the end-users is problematic at best and impossible at worst, user-centered design becomes difficult. Researchers are unlikely to go to work on the type of problems faced by inaccessible users. Commercial vendors have difficulties evaluating and improving their products when they cannot observe real users working with their products.

In well-established fields such as web site design or office software design, user-interface guidelines have been developed based on the results of empirical studies or the experience of experts. Guidelines can speed up the design process and replace some of the need for observation of actual users. In 2006 when the visual analytics community was initially getting organized, no such guidelines existed. Therefore, the VAST Challenge committee was faced with the problem of developing an evaluation framework for the field of visual analytics that would provide representative situations and datasets, representative tasks and utility metrics, and finally a test methodology which would include a surrogate for representative users, increase interest in conducting research in the field, and provide sufficient feedback to the researchers so that they could improve their systems.

We would like to point out the role of each author of this paper in the VAST Challenge committee.

- Jean Scholtz was on the VAST Challenge committee from 2006 to 2010 and worked primarily on the metrics used for developing the evaluation.
- Catherine Plaisant was also on the VAST Challenge committee from 2006 to 2010 and developed the Challenge Website and repository and helped with the metrics development.
- Mark Whiting has been working on the VAST Challenge since 2006 and is responsible for developing the problems and datasets.
- Georges Grinstein has also been on the VAST Challenge committee since 2006 and has worked on quantitative metrics evaluation and has also contributed to problem and dataset development.

Together the authors designed the overall format of the VAST Challenge. In addition to the authors the VAST Challenge committee has also included several analysts from PNNL who helped with problem development and evaluation. In the first several years of the Challenge several NIST employees helped with developing metrics and evaluation procedures. In later years, representatives of organizations sponsoring VAST challenges have served on the committee.

The VAST Challenge as a platform for evaluation

We chose the idea of a contest so that teams could enter to see how useful their software was in analyzing data with a specific task in mind. We could locate or generate tasks and datasets that would be representative of what actual analysts might use. Additionally we needed to create a scenario for participants to analyze so we needed to embed elements of this scenario in the dataset. The contest problem solving format was motivation to the participants. As a number of the participating teams are students we get many comments that they view the challenge as fun and like the idea of trying to find an "answer" in the data.

Our key question was how to test the visual analytics systems proposed by the participants with this data and tasks. While the ideal situation would have been to install the reliable submitted software systems and recruit real analysts to conduct the task under tight time constraints, this was not a realistic solution[a].

First, we were dealing with prototype systems that might not be terribly robust and finding analysts to download, install, learn and use all those systems was not viable. Instead we could ask teams to do their own analysis of the data and submit the results to us, and we could determine if the visual analytics environment was useful (assuming the existence of ground truth). This, however, only allowed us to evaluate the end results. The process of obtaining the results is also very important to analysts so we needed the teams to explain how they reached their conclusions. Analysts could then focus on the submitted explanations and provide feedback which would cut down on the time analysts needed to spend during the evaluation process.

Having decided upon the overall format of the evaluation, our next step was to determine possible metrics. Ground truth was embedded in the dataset making it feasible to measure accuracy. Unlike some other contests where participants were asked to "find if there was something interesting in the data" or simply "communicate science, engineering and technology for education and journalistic purposes" or aim for the "wow factor", we knew the interesting information that we had embedded and could measure how well teams discovered and understood it.[10,11,12] While measuring accuracy sounds quite simple, there are issues and we discuss them in a later section on accuracy. However, we certainly wanted to include this measure in our evaluation.

Insight has been discussed by many researchers as a goal for using visual analytics systems to explore data. [13] While we agree that being able to derive insights through the use of a visual analytics tool is certainly a desired outcome, measurement of these insights is dependent on the prior knowledge of the analyst using the software. In earlier unpublished work in the Novel Intelligence from Massive Data project, an aspect of insight was used as a measure. The actions of an analyst using two different systems to investigate the same problem using the same data were recorded.[14] The analyst first used her customary tools and then used a visual analytics tool. In that effort, the experimenters used the recorded actions to determine if the analyst found new data or discovered new connections. This type of experiment was clearly not an option for the VAST Challenge as we would not be giving reviewers hands-on access to the systems and data and even if this were possible, controlling such an evaluation would have been difficult.

Usability is certainly a concern for software intended to be used by busy professionals. Again, the software we were looking at was likely to be in prototype form. While we definitely wanted to look at concepts underlying usability, we felt that the normal definition of usability (efficiency, effectiveness, and user satisfaction) could not be utilized at this particular time. One issue with usability testing is that it is most effective in testing small tasks in 1-2 hour segments. Visual analytics tools are meant to be used for solving complex problems which often take many hours, may involve many combinations of software functionality and multiple users. Currently usability testing is inadequate to measure this. [4, 15] Similarly controlled experiments measuring differences in term of speed and errors between systems are not practical.

---

[a] During 2006-2008, an "interactive session" was held at the VAST conference where a small number of selected teams were paired with professional analysts and given a new scenario, dataset, and question set to analyze on the spot. These sessions typically lasted two hours, and a wrap-up session was held where the analysts could describe their experiences with the software. These sessions were well received by both the analysts and contestants. As worthwhile as the interactive session was, it would be very expensive and logistically challenging to extend this as VAST challenge offering to all competitors.

Speed of performance and accuracy of the automated components of the systems (e.g. search or clustering) or accuracy of the data representations in the visualizations are aspects that are vital to users adopting the software, but could be tested separately and so were not the focus of the contest. Our primary concern was the overall utility of the software to the analyst: did the software help the analyst do her job better where better might be faster, more effective, with more confidence or some other aspect that contributes to success.16 Utility being multifaceted, we needed differing metrics to capture utility. A diverse set of metrics also allowed us to evaluate heterogeneous sets of systems at different levels of development, and highlight promising elements in systems that may not have performed well overall.

Metrics selected

We selected measures for accuracy, the analytics process, the visualizations and interactions as indicators of the final utility of the visual analytics software. Our goal was to be able to adequately capture those measures even without hands-on access to the software. In the sections below we discuss each of these metrics and how we have implemented them, including changes made over the course of seven years of the Challenges. Appendix A contains the URLs of the review criteria that were posted on the web sites for the years 2006 – 2011 as well as the URL that contains the VAST Challenge entries since 2006.

Accuracy

With the use of realistic analytics scenarios and datasets there is typically no simple numerical accuracy measure that can be calculated automatically. Answers are descriptions of situations that are multidimensional by nature and often best represented by a narrative. This precludes the automatic ranking of submissions which is common in other fields and allows some competitions to attract large numbers of participants by offering large monetary rewards to undisputed winners. [6, 7, 17]

Still, the initial VAST Challenge scenarios accuracy included a quantitative accuracy component. We were able to develop forms that teams were asked to fill in. Lists of names, places, times or events are examples of information we requested teams to submit in their answers. These answers were relatively easy to evaluate. Teams were given credit for correct answers and penalized for giving incorrect answers. We also asked teams to provide evidence to support their answer. Evidence was most often in the form of the name(s) of the files in which this information was found. For any given Challenge there were a number of such questions that needed to be answered to provide a picture of the entire situation. The VAST Challenge committee computed an overall weighted accuracy measure which was supplied to the reviewers.

In addition to the lists used to measure accuracy, we also asked teams to write a descriptive report or debrief (analogous to an intelligence product) of the overall situation. Initially this debrief was evaluated by the analysts on our review teams. The intelligence profession has guidelines to use in evaluating analytic debriefs.[18] Analysts used these guidelines and provided feedback in the form of comments to the teams. This feedback included comments on the clarity of the description, the level of detail covered, the distinction between the evidence found to support the conclusion and the added analytic value provided by the team.

In 2008 the Challenge transitioned from a single problem with a heterogeneous dataset to a number of smaller problems each with a separate homogeneous dataset. This allowed student teams and other teams who might be focused on one specific type of data to enter only one or two of the new mini challenges. Teams who were able to analyze all the datasets would be able to enter the Grand Challenge which required both quantitative answers and an overall debrief. Even though we had significantly more teams enter in 2008, we still had only single digit entries for the Grand Challenge. Given this small number of Grand challenge entries the analysts on the VAST Challenge committee were able to evaluate the debriefs submitted by the Grand Challenge entries.

In the VAST Challenge 2011 one of the mini challenges was quite different. Mini challenge 2 consisted of over 8 GBs of network logs and the task was to develop a situational awareness display. Evaluating the accuracy was difficult as the reviewers needed to compare the submission's findings to that supplied by the challenge committee. Basically teams were asked to locate patterns and describe them. Reviewers were then to compare those descriptions to those given by the challenge committee. Additionally, if a submission contained additional patterns the reviewers needed to judge the plausibility of those findings. Thus a reasonable degree of subjectivity is introduced.

In 2012 both mini challenges were quite complex and it was not possible to ask simplistic questions that could be easily evaluated as right or wrong. Thus the accuracy was entirely subjective. Because of the large number of entries, the accuracy was evaluated by external reviewers who were asked to compare the team debriefs with what was supplied to them by the VAST Challenge committee as ground truth.

Analytic process

In addition to the accuracy of the software analysts are concerned with the analytic process they must use. Analysts are trained in methodologies for systematic analysis and it is important for software to be flexible enough to support various analytic methodologies.[19, 20] While we do not expect non-analyst teams to use structured methodologies we are interested in the various steps that the teams take while investigating the data and responding to the given problem. The question was how we could evaluate the process that was used with a given visual analytics environment without hands-on access to the software. We accomplished this by developing the notion of a detailed answer in 2008. If a specific question required a detailed answer the team was to supply an explanation of how they used their software to arrive at the requested information. A detailed answer required a text explanation accompanied by screen shots showing the visualizations that were used.

In the early contests (2006-2007) the process was evaluated mostly by comments from reviewers. Additionally, the reviewers were asked to comment on specific aspects such as the efficiency and intuitiveness of the process. In these early contests, the reviewers consisted of the committee members and a small number of analysts who were asked to participate. With the dramatic increase in number of entries (following the change to the mini challenge format in 2008), the number of reviewers also had to increase and the Challenge committee developed review forms to normalize the comments collected from reviewers.

In some years (2011 and 2012), process has not been reviewed as such although reviewers are still asked to evaluate the visualizations and interactions (see the following sections). A case can be made that the explicit evaluation of the analytics process should be reinstated. Teams at times make faulty assumptions and arrive at incorrect answers even though they have designed software that provides an effective and efficient process. Being able to separate the evaluation of the process from the accuracy is necessary to provide this feedback. If this were to be reinstated, analytic process expertise would need to be prominently represented by the Challenge reviewers.

Visualizations and interactions

In evaluating the visualizations, we were interested in how helpful they were, how intuitive they were and if they were novel visualizations. We were also interested in whether the visualizations would scale to handle larger datasets. We asked reviewers to comment on the effectiveness and intuitiveness of the visualizations. We wanted to know if reviewers were able to easily interpret information shown in the visualizations and if that information was useful in providing evidence to support an analytic hypothesis or in helping analysts to find the next step to take in their analysis.

External reviewers are still asked to rate the visualizations provided by the team submissions. Ratings scales are provided and comments are requested to support the ratings given. On some occasion different specific questions were asked which depended on the scenario of the challenge. For example, in 2012 reviewers of the cyber challenge were asked to rate and comment on the situation awareness and the dynamic situation awareness visualizations.

We were also interested in the interactions that teams provided in their software for analysts to use on refining information shown in the supplied visualizations. As a written description of interactions is not very helpful, we asked teams to provide a short video (for each given question) and external reviewers were asked to use the videos in assessing the intuitiveness and efficiency of the interactions. While we mentioned these specific qualities in early review requests, later review forms were more specific in requesting rating and comments.

In the field of human-computer interaction many guidelines have been developed for Windows based systems, Apple systems, mobile systems and web-based systems. [21, 22] There is currently no agreed upon set of guidelines in use by the visual analytics community, but several efforts have generated sets of useful guidelines. Scholtz analyzed the reviews from the 2009 VAST Challenge to identify guidelines based on comments from the reviewers.[23] She then looked at those guidelines to determine which, if any, already existed in guidelines for visual analytics or in guidelines in other domains. For example, she found that a number of reviewers commented on the lack of integration between visualizations, a guideline was created to match those comments. She found that this guideline was included in a set of guidelines for visual analytics environments developed by Carr et al.[24] Scholtz

also found guidelines from other domains that were also appropriate. For example, comments that some interactions were not available in menus but only as keyboard commands is equivalent to a Human-Computer Interaction guideline: Recognition rather than recall.[9] There were also appropriate guidelines from areas such as situation awareness, human-computer automation, and web design. Additionally there were comments not covered in other areas such as the large amount of data that analysts have to visually inspect in the visualizations. These and other issues need research to develop the appropriate guidelines for design and implementation. If an agreed upon set of guidelines for the visualizations and interactions of visual analytics systems were available, we would be able to use that and could possibly provide a heuristic review of the screens available in the submission either in static form or in the video. In the meantime for the VAST challenge we still had to rely on ratings from reviewers and the (more and useful) comments they provide as the rationale for a particular rating.

Clarity of explanation

An additional question on the review sheets has always been the clarity of the explanations. We provided this feedback to the teams as in some cases their explanations were poorly written and resulted in a wide disparity in ratings from the reviewers. An analysis of the 2009 data shows that this definitely had an impact on the submissions. [25] The submissions in 2009 that were rated higher in clarity got overall better ratings. This is presumably because the reviewers were able to better understand what the teams had done to obtain the answers.

Discussion

There are four major questions about the VAST Challenge evaluation.

- Are we using the right metrics?
- Are we using the right reviewers?
- Are we providing the reviewers with the appropriate information?
- Are we providing the appropriate feedback to participants?

The right metrics

What are the right metrics? While there is no right answer to this, it would be safe to say that the right metrics provide feedback to the participating teams that helps them refine their visual analytics software. It is also important that the metrics we are using evolve over time to adequately assess the different types of datasets and scenarios we are asking teams to analyze. Our metrics provide both quantitative and qualitative assessments. Quantitative accuracy metrics provide teams with definitive measures of whether they have found the people, places, events and/or activities that are necessary to accurately assess the situation. The fact that these can often be submitted and judged automatically, make these metrics desirable. However, these measures do not provide the rationale to tell participants why they succeeded or failed. The more qualitative measures such as evaluating the process, the analytic debrief, and the visual representations and interactions do provide judgments by the reviewers along with their reason for the assessment. However, these measures are more subjective and will often provide different perspectives on the same aspect of the software. While this may be perplexing to the teams, this is often the case in the real-world. Many of us disagree on software packages based on our preferences.

We conducted two surveys one in 2009 and another in 2012 asking participants about their experiences in the VAST Challenges.[28] These surveys were conducted over the Web and submissions were anonymous. We e-mailed out requests to past participants and provided them with the link for the survey. We have no knowledge of which responses came from academics (professors or students) or industry. While a more comprehensive survey would certainly be helpful, funding issues with the VAST Challenge make surveys and the resulting analysis difficult. Additionally the changing nature of the participants also creates a problem. While many of the same organizations participate each year, the individuals involved often are not the same. This makes year to year comparisons difficult.

The two surveys we did conduct were short and were done primarily to help us understand the usefulness of various aspects of the VAST Challenge. In 2009 we asked participants if the VAST Challenge work helped them understand the tasks of the analysts, and if it helped them improve their systems. Overall the 35 respondents agreed that the Challenge was very helpful.

We asked them if the ground truth in the data was useful and how useful the reviews were. Figure 1 shows that participants found having ground truth in the dataset very important. Figure 2 shows that some participants found the reviews useful while others did not.

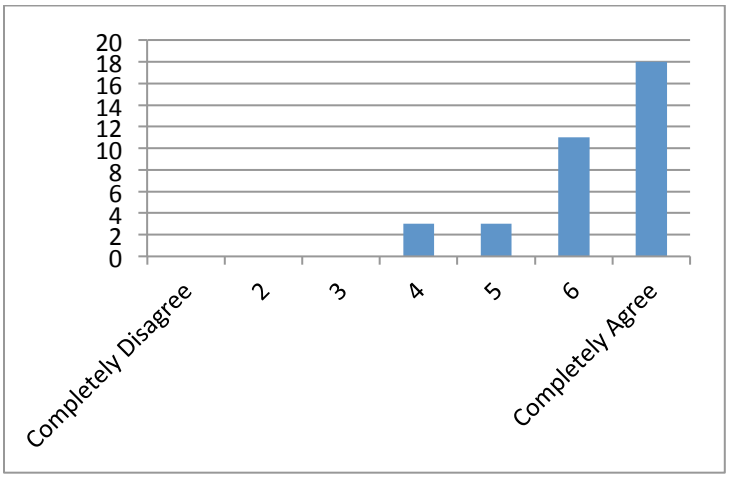

Figure 1. Ground truth in the dataset is important

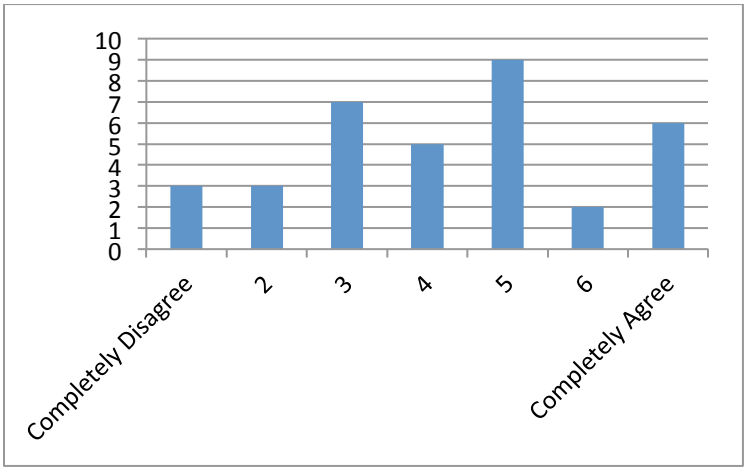

Figure 2. The reviews are useful

In our 2012 survey, 26 participants who had entered one or more VAST Challenges responded. We asked them why they entered. Table 1 shows that participants do want to see the evaluation information

Table 1: Check all reasons that you entered a VAST Challenge

Response	Number selecting this response
It looked like fun	19
We wanted to understand the analytic process more thoroughly	17
We wanted to test our software against datasets /tasks /questions with ground truth	16
We wanted to see how useful our software was	13
We could get reviews of our software by visualization experts and subject matter experts	**13**
Was required for a class I was taking	1

We asked participants to check the parts of the VAST Challenge that were most helpful to them. Table 2 provides their responses and confirms that many find the reviews helpful.

Since we know that participants found the Challenge very useful overall this suggests that for many participants working through the problem itself and seeing the ground truth at the end was more valuable than reading the reviews. Still, many participants found the reviews helpful, which is reassuring. Further analysis may look at who benefits the most from the reviews.

The accuracy metric has evolved greatly over the life of the VAST Challenge, moving from quantitative form–based and analyst evaluated debriefs to more subjective descriptions evaluated by our external reviewers. The second survey done after the 2012 VAST Challenge showed no comments about this change. We did hear from several reviewers that accuracy was difficult for them to evaluate with the solution supplied by the VAST Challenge committee. Given this, we looked at the reviews for the two mini challenges for 2012. We pulled out 5 reviews at random for each mini challenge and looked at the overall scores and comments to determine if there was a large discrepancy. This confirmed that reviewers had difficulties rating accuracy and provided inconsistent comments.

For example, in one instance, reviewers were asked to rate the noteworthy events found. The ratings that could be selected were all, most, some, very few, none. In several cases, reviewers checked most or some but gave an identical list of what they thought was specified in the report. In one instance where a team found an additional event (not listed in the supplied answer), the reviewer choose “some” to penalize the team for finding an additional

one. In some cases, reviewers declined to provide a rating and in some cases four of the five rating choices were chosen.

Table 2. Check the parts of the VAST Challenge that were most helpful to you.

Part	**No. of Responses**
Working on the software/developing the solution	23
Feedback on the accuracy of your answer	16
Receiving professional reviews	**15**
Discussions at workshop	10
Paper published in proceedings	10
Presentations at workshop	5
Poster presentations	3
Interactive sessions (only for 2006 and 2007)	2

In the future more attention needs to be given to the rating selections, criteria given to the reviewers and to the "ground truth" information provided to reviewers. We also need to ensure that these ratings are summarized in such a way that the feedback to the teams is beneficial. This is definitely an area that we will be working on in the future. As the datasets and tasks become more complex, the ground truth and "answers" could become even more subjective. In retrospect we have found that quantitative accuracy questions are more difficult to pose in cyber domains. This may improve as the domain matures but currently, the approach to this issue is elusive. When answers can only be reviewed in a subjective manner, a small number of analysts could review the subjective portion of all entries so as to consistently compare and rank the answers provided by participants.

The right reviewers

Who should review the submissions? When the original VAST Contest committee was formed in 2006, it consisted of visualization experts, human-computer interaction experts and practicing intelligence analysts. As we had a small number of submissions, all of us read and discussed them. Comments from the individuals were summarized and given to teams as feedback. This same procedure was followed in 2007.

In 2008 we had just a few days to come up with a way to evaluate some 80 submissions instead of the 6 submissions we had in 2006 and the 7 we had in 2007. Four members of the committee did an initial pass with each looking at the submissions in a specific mini challenge. We triaged these and then the entire committee looked at the filtered subset. In 2009 we anticipated a large number and started an external review process. We recruited both visualization researchers and intelligence analysts and assigned three reviewers to each submission; 2 visualization researchers and 1 intelligence analyst. However, the Challenge committee still reviewed the small number of Grand Challenge entries. We have continued with this external reviewer process. In 2012 we were able to assign 5 reviewers to each submission. As the domains have changed over time we have recruited reviewers from specialists in the featured domains. For example, in 2011 and 2012, we recruited reviewers from the cyber analysis domain.

Scholtz analyzed the reviews from the 2009 mini challenges to determine if there were discrepancies based on the expertise of the reviewers.[25] Existing discrepancies were not based on the expertise of the reviewer. The differences were most often between reviewers with the same expertise (analytic or visualization). However, we do feel that having a small set of reviewers focus on reviewing all of the more subjective portions of the submissions will provide more consistent feedback to the teams.

The funding model for the VAST Challenge has recently changed and companies and agencies are stepping up to fund mini challenges that are representative of problems they face. As this becomes more prevalent we anticipate that more of their analysts will become reviewers. As we have more domain knowledge experts review we may encounter more differences between their reviews and those of non-domain experts. This will be something we need to track, possibly considering asking reviewers with different backgrounds to review different aspects of the submissions. At the minimum, we will likely need a meta-reviewer to consolidate the reviews as is the case in many conferences and journals.

Currently we have limited knowledge of the background of our reviewers. We do for the most part recruit from people we know or from their references. We also cull out reviewers who do not provide reasonable reviews. In the early years, we focused on actually obtaining enough reviewers. In the past two years we have finally been able to obtain more than the bare minimum of reviewers. In the future, we may ask reviewers to provide more demographic information in order to understand their comments in this context. However, as we often use reviewers from the intelligence analysis field, they may not be willing to provide much information. Again, while this information would be useful, we would also need the funding to analyze this information and to analyze the reviewer comments accordingly.

The right information

What information is requested from the team in their submission influences our ability to evaluate the utility of the systems. Based on our own experience and the feedback of reviewers we feel that the debrief is the most valuable to judge how well a team has been able to analyze the data. On the other hand the video is clearly the most helpful to judge interaction and even the visualizations because they are typically interactive and can be explained more clearly with audio commentaries. Accuracy scores, while obviously useful to judge the submission, seem to strongly influence some reviewers who do not give good ratings on other aspects of the system or process when accuracy is low. Finally, the clarity of the explanations is an important factor in the ability of the reviewers to even start evaluating a system. A system of potentially high utility can still have poor evaluation results if the explanations are not clear, which is a handicap for newcomers and teams with low proficiency in English.

The right feedback to the participants

Are we providing our participants with feedback that is useful to them? In the previous sections we have defined the feedback primarily as the reviewers' comments that are returned to the participants. Note that our surveys did find that overall reviews were useful. The VAST Challenge entries from previous years are all made available in a repository (See Appendix A). In addition to teams receiving reviews, the VAST Challenge committee makes awards each year. These are not fixed awards but are awards based on aspects of submissions that are deemed outstanding in some respect. Awards can recognize novel visualizations, outstanding processes, interactions, good explanations, and other special awards. While the reviews are not made public, the awards given to various submissions are. So teams are able to see the various aspects of submissions that were selected for specific awards.

Additionally, the VAST Challenge runs a day long workshop during VIS Week in which participants come together to give short talks and demonstrate their software, with ample time for discussion between teams. Participants have found this interaction to be extremely helpful and have gotten feedback from attendees that is helpful. Prior to 2012 the workshop was open only to VAST Challenge participants. In 2012, the workshop was open to other VIS Week attendees as well. We have not yet been able to determine the impact of public attendance.

Conclusions

Since the VAST Challenge was created Munzner proposed a nested evaluation approach for visualization design and evaluation.[26] She looks at the problem and data characterization, the operation and data type abstraction, the visual encoding and interaction design, and the algorithm design levels. The problems and datasets provided in the VAST Challenges over the past seven years are excellent resources for use by researchers and developers using this nested approach. Comments from VAST Challenge participants suggest that they have implicitly been validating the first

three levels while preparing their entries for the VAST Challenge. The reviews provided to the participants in the VAST Challenge do not explicitly refer to these levels but they certainly contain comments on the abstractions, encodings, and interactions. The VAST Challenge has supplied a representative problem so teams can revise their software at this level. As many entries are still in the prototype form, the algorithm evaluation is not included in the Challenge but participating teams may undertake that while creating their submissions or later in their development activities.

The metrics currently used in the VAST Challenge include measures of accuracy, analytic process, visualizations, and interactions. We have no understanding of a team's knowledge of the domain. Additionally we rely on the teams to explain their findings in any given visualization. This self-reporting and the lack of understanding of domain knowledge limit the usefulness of measures of "insights" gathered during the process. We do, however, ask them for their final "insights" on the problem.

We also realize that the process we ask teams to describe (for us to assess how they arrive at specific answers) is a process arrived at after many weeks of working on the problem. That is precisely one of our goals: to provide a problem and data that allows teams to refine their software to effectively and efficiently investigate the problem. As we said earlier, we are essentially providing resources for teams to do their own nested validation and evaluation process.

A comparative methodology was used to determine how visual analytics software (the Jigsaw system, in particular) functioned in investigative analysis. [27] The researchers also used this study to understand what metrics are important for use in practice. They concluded that useful metrics were: the percentage of important documents viewed, the time until the analyst starts making representations, the quantity of the representations, the amount of time spent reading and processing information, and the amount of time and effort needed for organization. While we agree with these measures and have used the majority of these during our own hands-on research, these metrics are impractical to use in the VAST Challenge for several reasons. First, they would have to be self-reported and would place a burden on the teams to find methods to do this. Secondly, the work in the VAST Challenge is done by teams, not a single analyst. Thirdly, the results turned into us are most likely compiled from iterative use-fix-use cycles. We applaud the researchers for undertaking this study and observing the use of the software to understand the benefits of visual analytics. It is important for the community of visual analytics researchers to note that the metrics used in our community evaluations differ from those that should be used in hands-on evaluation studies.

We track the types of organizations submitting, including whether teams are student teams or teams from organizations, along with where the teams are located. We get more submissions from students and academics than from commercial companies. The comments we have received from commercial companies indicate that they often do not have the time necessary to prepare submissions due to other responsibilities. However, some companies use the problems and datasets extensively for internal testing of their software. We track downloads, and ask about other usages, but do not have enough survey responses to make any generalizations about those other usages.

While challenges remain and the VAST Challenge evaluation methodology has some limitations, we believe that it currently remains the most economical, practical, and useful community based evaluation method. The publicly available submissions along with their reviews provide a window into the utility of the systems being compared in a particular VAST Challenge. The consistent number of participants confirms that participants view it as a useful opportunity to learn about representative tasks of analysts, improve their systems, and have fun on the way.[28] After seven years of continuous activity, reviewers continue to volunteer. Moreover, the use of datasets and problems in is expanding into the educational realm.[29] The enduring central presence of the VAST Challenge at the conference, the sustained rate of participation and dataset downloads, and the participant survey results indicate continuing interest from the researchers and practitioners and suggests that the Challenge is contributing to the overall evaluation of visual analytics systems. Evaluating accuracy remains the main challenge as we move toward more realistic scenarios and data.

Acknowledgments

The authors thank the support of Dr. Sharon Laskowski and Dr. Theresa O'Connell from the National Institute of Standards and Technology for their participation in early VAST Challenges. We also thank the many graduate students at the University of Massachusetts Lowell and the University of Maryland who worked on the VAST

Challenges. We also thank the many reviewers for their time and comments. And of course, we thank the teams who participate in the VAST Challenges. We are indebted to the late Jim Thomas from the Pacific Northwest National Laboratory who was the driving force behind the creation of the VAST Challenges. The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC05-76RL01830. Members of the committee were also supported in part by the National Science Foundation (0947343 and 0947358).

References

1. Thomas, J. and Cook, K. (Eds.) *Illuminating the Path: The Research and Development Agenda for Visual Analytics*, Los Alamitos, CA:IEEE CS Press, 2005.
2. Lam, H, Bertini, E , Isenberg, P, Plaisant, C and Carpendale, S. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 2012; 18, 9: 1520-1536.
3. Carpendale, S. Evaluating information visualizations. In Kerren, A Stasko, J T, Fekete, J-D and North, C (eds). *Information Visualization: Human-Centered Issues and Perspectives*, Berlin/Heidelberg : Springer LNCS, 2007, 19–45.
4. Plaisant, C. The Challenge of Information Visualization Evaluation, in *Proceedings of the working conference on Advanced Visual Interfaces* ,2004, 109-116.
5. Chinchor, N and Hirschman, L. Evaluating message understanding systems: an analysis of the third message understanding conference (MUC-3*), Computational Linguistics* 1993;19, 3:409 – 449.
6. Voorhees, E, TREC: Continuing information retrieval's tradition of experimentation, *Communications of the ACM;* 2007; 50, 11:51-54.
7. Face Recognition Grand Challenge, http://www.nist.gov/itl/iad/ig/frgc.cfm (accessed Nov 8, 2012)
8. Beyer, H., Holtzblatt, K., *Contextual Design: Defining Customer-Centered Systems*, Burlington, MA: Morgan Kaufmann ,1997.
9. Shneiderman, B, and Plaisant, C. *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* 5th ed. Boston: Addison-Wesley Publishing Co. 2009.
10. Plaisant, C, Fekete, J D, Grinstein, G. Promoting Insight Based Evaluation of Visualizations: From Contest to Benchmark Repository. *IEEE Transactions on Visualization and Computer Graphics*, 2008; 14, 1 :120-134.
11. NSF International Science & Engineering Visualization Challenge. http://www.nsf.gov/news/special_reports/scivis/challenge.jsp (accessed Nov. 2012).
12. Data Viz Challenge. 2011 http://datavizchallenge.org/about-challenge (Accessed Nov 8, 2012)
13. Saraiya, P, North, C, Lam, ., and Duca, K A. An insight-based longitudinal study of visual analytics. *Transactions on Visualization and Computer Graphics*,2006; 12(6) :1511–1522.
14. Cowley, PJ, Nowell, LT, and Scholtz, J. Glass Box: An Instrumented Infrastructure for Supporting Human Interaction with Information, In *Proc. 38th IEEE Annual Hawaii International Conference*, 2005; 296-304.
15. Redish, J. Expanding Usability Testing to Evaluate Complex Systems, *Journal of Usability Studie*s, 2007; 2, 3: 102-111.
16. Costello, L, Grinstein, G, Plaisant, C and Scholtz, J. Advancing User-Centered Evaluation of Visual Analytic Environments through Contests, *Information Visualization*, 2009; 8: 230–238.
17. Netflix Prize http://www.netflixprize.com (accessed Nov. 8, 2012)
18. ODNI Directive 203 http://www.fas.org/irp/dni/icd/icd-203.pdf (accessed March 20, 2010)
19. Heuer, R, and Pherson, R. *Structured Analytic Techniques for Intelligence Analysis*, Washington, DC: CQ Press, 2010.
20. Jones, M. D. *The Thinker's Toolkit: 14 Powerful Techniques for Problem Solving*, New York: Three Rivers Press, 1998.
21. Apple guidelines. http://developer.apple.com/library/ios/#DOCUMENTATION/UserExperience/Conceptual/MobileHIG/Introduction/Introduction.html (accessed Nov, 2012).
22. National Cancer Institute, *Research-based Web Design and Usability Guidelines*, Dept. of Health & Human Services, National Institutes of Health , 2006. http://www.usability.gov/pdfs/guidelines.html. (Accessed Nov 8 2012)

23. Scholtz, J. Developing Guidelines for Assessing Visual Analytics Environments. *Information Visualization* **July** 2011; 10 (3) : 212 -231.
24. Carr, D. Guidelines for Designing Information Visualization Applications. *Proc. of the 1999 Ericsson Conference on Usability Engineering*. ECUE'99, Stockholm, Sweden.
25. Scholtz J, Developing Qualitative Metrics for Visual Analytic Environments. In the *Proc. BELIV '10 BEyond time and errors: novel evaLuation methods for Information Visualization* , 2007: 1-7.
26. Munzner, T. A Nested Model for Visualization Design and Validation. IEEE Transactions on Visualization and Computer Graphics, Vol. 15, NO 6, November/December 2009.
27. Kang, Y., Görg, C, Stasko, J. , How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation*, IEEE Transactions on Visualzation and computer Graphics*, Vol. 17, No. 5, May 2011, pp. 570-583.
28. Scholtz, J., Whiting, M., Plaisant, C., Grinstein, G., A Reflection on Seven Years of the VAST Challenge, Proc. of BELIV 2012, *BEyond time and errors: novel evaLuation methods for Information Visualization, a workshop of the IEEE VisWeek 2012 conference*,2012: 1-8.
29. Whiting, M, North, C, Endert, A, Scholtz, J, Haack, J, Varley, C and Thomas, J. VAST Contest Dataset Use in Education, *Proc. of VAST 2009 Symposium*, 2009;115 – 122.

Appendix A

2006 Review Criteria posted on http://www.cs.umd.edu/hcil/VASTcontest06/scoring.htm

2007 Review Criteria posted on http://www.cs.umd.edu/hcil/VASTcontest07/datasetandjudging.htm

2008 Review Criteria posted on http://www.cs.umd.edu/hcil/VASTchallenge08/judging.htm

2009 Review Criteria posted on http://hcil.cs.umd.edu/localphp/hcil/vast/index.php/judging/index

2010 Review Criteria posted on http://hcil.cs.umd.edu/localphp/hcil/vast10/index.php/judging/index

2011 Review Criteria posted on http://hcil.cs.umd.edu/localphp/hcil/vast11/index.php/judging/index

2012 VAST Challenge: No judging criteria was posted on the website.

VAST Challenge Repository. Contains problems, datasets and VAST Challenge submissions.
http://hcil.cs.umd.edu/localphp/hcil/vast/archive/viewbm.php