Advancing User-Centered Evaluation of Visual Analytic Environments through Contests

Loura Costello, Georges Grinstein, Catherine Plaisant, and Jean Scholtz

(Draft TR version)

Abstract

In this paper the authors describe the Visual Analytics Science and Technology (VAST) Symposium contests run in 2006 and 2007 and the VAST 2008 and 2009 challenges. These contests were designed to provide researchers with a better understanding of the tasks and data that face potential end users. Access to these end users is limited due to time constraints and the classified nature of the tasks and data. In that respect, the contests serve as an intermediary, with the metrics and feedback serving as measures of utility to the end users. The authors summarize the lessons learned and the future directions for VAST Challenges.

1 Background

Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces [1]. It is a multidisciplinary field dealing with the process of analyzing data and includes analytical reasoning, visual representations of data, computer human interactions, data representations and algorithms, machine learning, and tools and techniques for communicating the results of such analyses, making it a challenge to evaluate. It is critical to involve users in the evaluation of visual analytics as analysts must be able to use the tools, interpret the visualizations and generate the analytic product (i.e. usually a briefing). However, one issue is getting access to those end users, their tasks, and data sets. A second issue is that standard evaluation methods do not deal with the scale, complexity and duration of the many tasks analysts perform [2].

Our work on user-centered evaluations began with support from the National Visualization and Analytics Center (NVAC) funded by the United States Department of Homeland Security (DHS). DHS users represent a wide variety of domains which overlap greatly with non-homeland security ones. The need for materials for user-centered evaluations is particularly important in the DHS community. First, many of the tasks performed by DHS users are critical tasks where users have little time to both learn new software and to experiment with various options. Second, it is generally difficult for researchers and developers to gain access to DHS staff because of their workload and time constraints, and in some cases, due to the classified nature of their work. As such, our initial focus was on “intelligence analysis” type of tasks but the materials we have generated in the past 3 years transcend the homeland security arena and have broader applicability (e.g., for business intelligence, epidemiology or law enforcement).

Our goal is to provide a number of data sets with ground truth, tasks, and metrics, as well as a methodology to provide user-centered assessments of visual analytics tools through contests. Our benchmark data sets are generated first by defining a problem within an area of interest (such as bioterrorism or disease spread), identifying possible scenarios in one of these areas of interest, and generating synthetic data sets around a selected scenario. Analysts participate extensively in the development of these contest data sets as well as in the review of the submitted entries, which requires both qualitative and quantitative evaluations. As a scenario is inserted in a distributed fashion into various data types, the analysis necessitates using a process to arrive at a conclusion, as opposed to finding a single “answer.” The contests help the research and development communities understand the various tasks analysts must perform and contrast and compare the different approaches available [3].

As we are primarily interested in visual analysis environments, our data sets and tasks have characteristics (fairly large, heterogeneous data) that make visual analysis more effective than manual

analysis. The ground truth inserted into the VAST contest data sets is distributed among the different types of data, making the analysis difficult for machine algorithms and suited for synthesis by analysts. In essence, the contest provides a surrogate for traditional user-centered evaluations by providing researchers with representative tasks and data sets, and by conducting evaluations that focus on the utility of the visual analytic tools to the end users, judged by end-users and experts in the field. As in traditional user-centered evaluations, contest judges are asked to rate the utility of the visualizations including interactions with them, and the efficiency, effectiveness and scalability of the process used to analyze the data, as reported in the submissions.

In all four years, we used synthetically generated data with embedded ground truth, defined tasks (so far all based on finding "who has done it?"), produced submission requirements, and refined a procedure for evaluating the entries using both quantitative and qualitative measures. The data sets have been produced by the Threat Stream Generator project at the Pacific Northwest National Laboratory (PNNL) [4]. All materials remain available after the contest (data sets, tasks and submissions from participants). Solutions are given to teams who submit their entries after the judging, and eventually made public after a delay to encourage further submissions after the end of the contest.,

In the following sections, we describe the contests run at the Visual Analytics Science and Technology (VAST) 2006 and 2007 Symposia and the VAST 2008 and 2009 Challenges that evolved from these contests. We conclude by describing our plans and some of the issues we face in future Challenges. The data sets and detailed descriptions of the contests, including tasks, judging criteria and solutions (except for the most recent challenges) are available online starting at www.cs.umd.edu/hcil/VASTchallenge09 which links to the previous years.

2 VAST 2006 Contest

The data set, "A Tale of Alderwood", consisted of about 1200 news stories from the fictional Alderwood Daily News with some images and maps, a voter registry, and a phone call log as Excel tables. Participants were asked to analyze the data set, answer *who, what,* and *where* questions concerning the embedded plot, and include reference documents from the data set as evidence. Specifically participants were asked to provide us with:

- The players who were engaging in deceptive activities
- The locations relevant to these activities
- The events that occurred that were relevant to the embedded plot
- And the time frame in which these activities occurred

Participants were also asked to generate a report (a debriefing) with screenshots summarizing their analysis and to provide a video highlighting the use of their system in performing the analysis.

We assembled a team of judges consisting of human-computer interaction (HCI) and visualization experts, as well as intelligence analysts. All judges were given the results from a quantitative evaluation (identifying which parts of the plot had been uncovered by each team) and some guidelines for judging the entries. The analysts judged the report and numerically evaluated the clarity and supporting evidence. However, judging the utility of the software system and the submitted visualizations was less than systematic. The visualization and HCI experts were able to provide ratings for low-level properties of the visualizations but were less confident about utility ratings.

We received six entries for the contest with the winners in the two categories (academic and corporate) invited to an interactive session during the VAST symposium (Figure 1). During this session, the teams were able to work with an analyst and solve a similar but smaller task so that a reasonable amount of progress could be made during a two-hour session. This session was deemed valuable by the software teams who could see how the analyst used their software in an analytic process, and for the analysts who could better judge the utility of the tools.

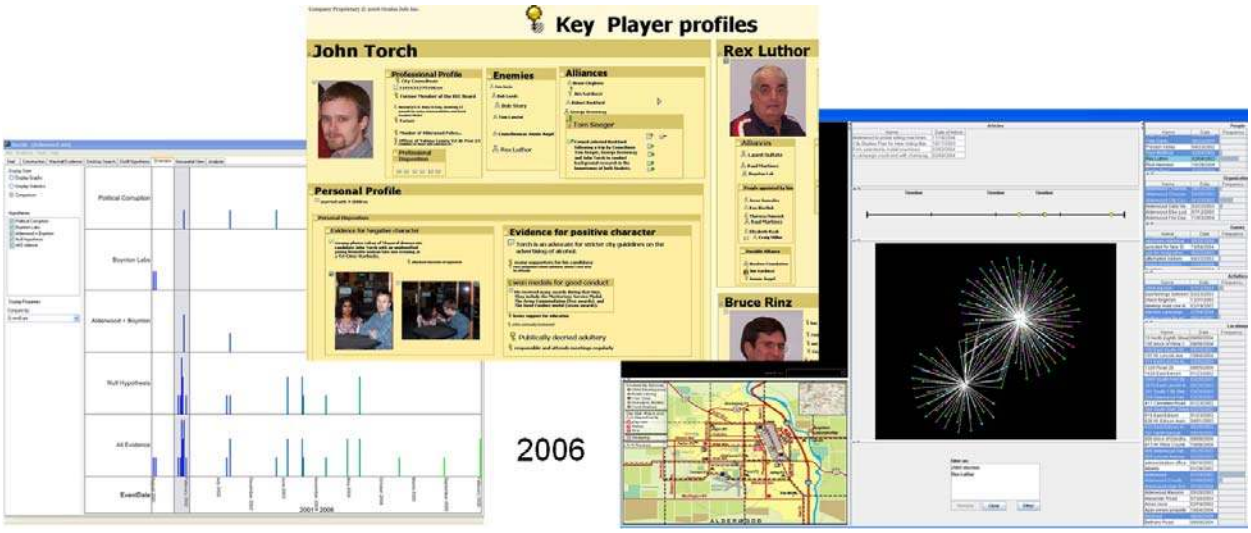

Figure 1 Examples from 2006: On the left, SSS Research Inc.'s timeline showing clusters of events. In the center, Oculus' Sandbox with annotations about key players. On the right, Georgia Tech's visual analysis environment.

Lessons Learned

The complexity of the submissions, the metrics computations, the judging, and the interactive session all had much larger than expected time demands. We were able to score the quantitative metrics systematically and consistently, even though we had failed to prescribe a simple input format for automated computation to take place. The cost was our time. The qualitative measures were less successful. Many analysts were unable to spend time reviewing the submissions prior to the meeting and therefore, it was difficult to go through all the submissions systematically. Detailed qualitative and quantitative feedback was provided to each team, which was an effort that was also time consuming.

The interactive session was well received both by the involved analysts and by the participants, but observations from these sessions were of little value in refining our qualitative ratings. A video of each team's session was made available to them, but feedback indicated that these were not used. For more information about the VAST 2006 contest, see [5].

3 VAST 2007 Contest

The VAST 2007 contest was similar in format to the VAST 2006 contest. The scenario involved both illegal animal smuggling activities and possible terrorist type activities. As in the 2007 contest, participants were asked to identify the activities, the characters involved, the locations and time frame. This time, however, the participants needed to identify whether the questionable activities were only illegal or could be a possible terrorist activity. The data set consisted of more heterogeneous data:

- about fictional 1500 news stories simulating an online news archive
- two blog excerpts with entries extracted over certain time segments
- a few pictures (in jpg format)
- a few small databases (in XLS and CVS format)
- a few pages of background information (in DOC or PDF format).

This scenario had missing data, requiring participants to recognize and deal with information gaps.

We handled judging in the same fashion as 2006. We extended the 2006 form for the judges to provide ratings and comments on the debriefing (i.e., the analytic product), the visualizations and interactions, and the overall utility of the visual analytic environment. We asked our judges to commit to two days of work, one for the actual meeting and one, a day prior to the meeting, to review submissions and complete rating sheets. We provided the teams with quantitative measures using a weighted scoring that penalized

guessing. For example, if we asked teams to identify all the people involved in a certain illegal activity, credit was given for those identified correctly and subtracted for identifying additional non-essential people. These scores were computed manually.

Of seven total entries, three teams (one university team and two corporate teams) were invited to the interactive session during the VAST Symposium. A detailed script and comments were sent to the individual teams. Figure 2 shows a sample of the visualizations from the VAST 2007 contest entries.

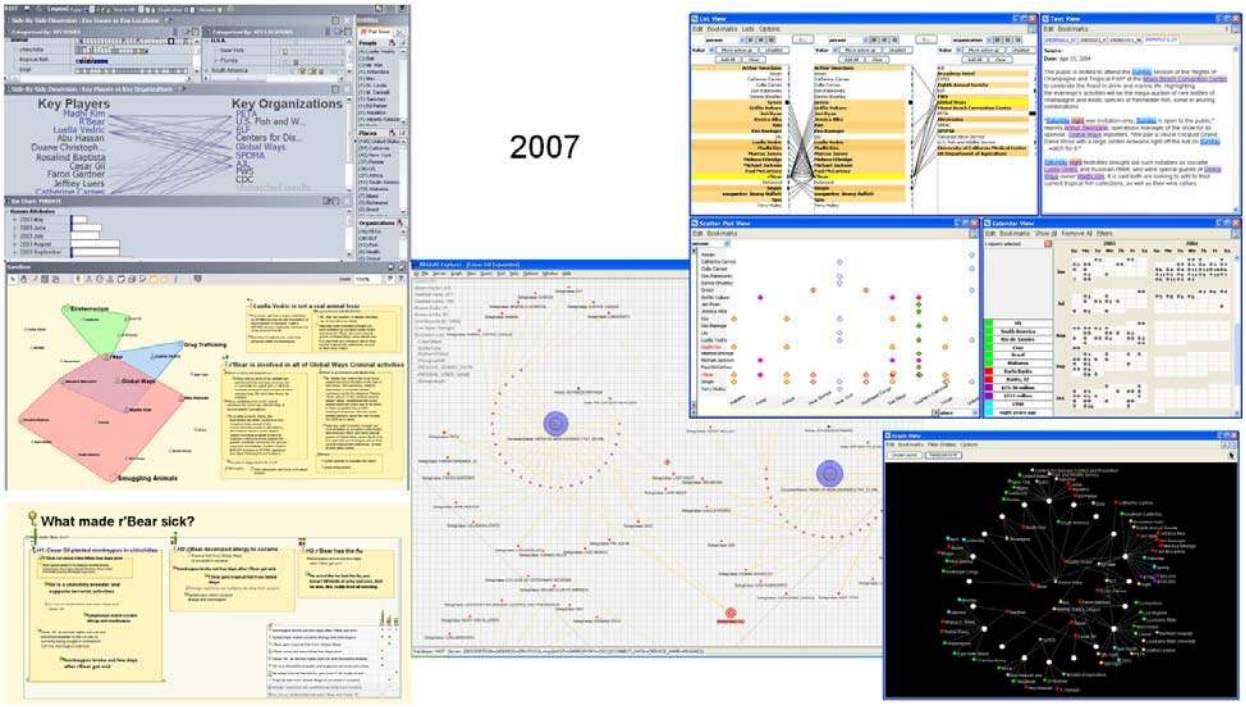

Figure 2 On the left, screens from Oculus with a sandbox for the analyst to write notes. In the center, an entry from ATS Intelligent Discovery showing relationships between characters and documents. On the right, screens from Georgia Tech, showing coordinated representations of relationships between name entities found in the documents.

Lessons Learned

In year two participants could see examples from 2006 and submitted better process descriptions and better debriefings summarizing their findings. Overall, we had made progress in organizing submissions, conducting judging sessions, and observing the interactive sessions. Our original intent had been to provide awards based primarily on visualizations and the process teams used to arrive at their conclusions, which we did, but we also found valuable contributions from entries that had done excellent analytic work or had used innovative components worth rewarding, although their overall software environment was not yet sufficiently developed. This suggested the need to broaden the type of awards given.

The refined rating sheets were an improvement for the judges but we found that analysts were much less willing to provide feedback on the quality of the displays while the HCI/Visualization experts had a difficult time providing feedback on the quality of the debrief. For more information on the VAST 2007 contest, see [6]. The main award winners were invited to coauthor with the chairs a journal paper summarizing their experience [7].

We sent a questionnaire to the 2007 participating teams to get their input about the contest process and their rationale for participating. We also contacted researchers who did not enter to understand how we could attract more participants. The key reason for participating was the opportunity for an external

evaluation. The main reasons for not participating included the time needed to put together the entry, the notion of being either a winner or a loser, the incompatibility of the research being done and the type of data sets used in the contest. We tried to address several of those issues in 2008.

4 VAST 2008 Challenge

In order to increase participation, the VAST 2008 Challenge consisted of four mini-challenges, each containing a scenario of interest, and an overarching grand challenge using data elements from all of the mini-challenges. Teams could participate in one or more of the challenges. The reward process was also changed to emphasize outstanding performance in a variety of areas such as visualization, analysis, integration, interaction, etc. as opposed to an overall winner (hence the name change from contest to challenge). The scenario dealt with a controversial social movement involving illegal immigration and possible terrorist activities. The main task for the grand challenge was to identify the movement, its beliefs and activities. The mini-challenge had four distinct data sets and tasks:

- Wiki Edits: Characterize the factions involved and the individuals involved in each faction
- Cell Phone call logs: Describe the social network of two families involved in the movement and note how this social network changed over the time period of the data set
- Boat migrations: Analyze the patterns of landings and arrests over the three years of data given
- Evacuation traces in a building: Identify where in the building the explosion occurred, possible suspects and witnesses to the event, and which of those may have escaped during the evacuation.

The contestants were given Google Earth images of a fictitious Isla Del Sueño where the movement originated, cell phone tower locations and two artificial Wikipedia pages detailing the movement's history and views. In the mini-challenges, teams had to answer a set of questions specific to the challenge and explain how they arrived at those answers, with screenshots and an optional two-minute video. For the grand challenge teams submitted a debrief summarizing their findings, a longer process description and a longer video.

We received a record number of 73 submissions. This included 13 boat, 12 wiki, 22 phone, and 20 trace mini-challenges and 6 grand challenge entries.

The Challenge committee did a rapid preselection of the best mini-challenge entries to be reviewed and provided these to the outside judges a week prior to the face-to-face meeting. For the final judging, we supplied either analysis or visualization rating sheets to the judges, depending on their expertise. We combined the comments from the committee and the ratings and comments from the judges to provide qualitative feedback to the participants.

The quantitative feedback depended on the question asked in the mini-challenge. Questions that asked for lists of items could be easily evaluated automatically. We also experimented with an automatic scoring for the social network analysis based on precision and recall. Our ultimate goal was to allow teams to evaluate the accuracy of their answer before submitting, but we were dealing with the problem of giving away the solution after a single evaluation run (which would discourage system improvement to get to the correct solution).

We gave thirteen awards for exemplary performance with the stipulation that each team would be given only one award. Our awards included intuitive visualizations for the various mini-challenges, excellence in integrating visual analytic tools, and excellence in results, both accuracy and written reports. All awards are listed on the Challenge Website and all the submitted materials are available online.

We invited all teams who submitted an entry to a participant workshop where we presented a summary of the techniques used in the various mini-challenges, discussed the qualitative and quantitative measures used in the evaluations, and invited suggestions for future challenges. We invited three teams out of the

six grand challenge entries to participate in the interactive session, but while this session was again found useful by all participants it only reached a small portion of the participants.

Lessons Learned

By redefining the contest format into a challenge with four simpler mini-challenges, we were able to increase participation dramatically. However, logistics for handing the reviewing and judging process for this number of submissions was overwhelming. Automatic evaluation tools and submission and review management services are needed if we want this type of event to grow.
About 50 participants joined us for the interactive session, and we will definitely include it again in 2009, but we will be more careful not to duplicate the content presented during the panel summarizing the challenge during the symposium.

Figures 3 – 5 show mini-challenge visualizations that were found particularly useful by the analysts to interpret the data.

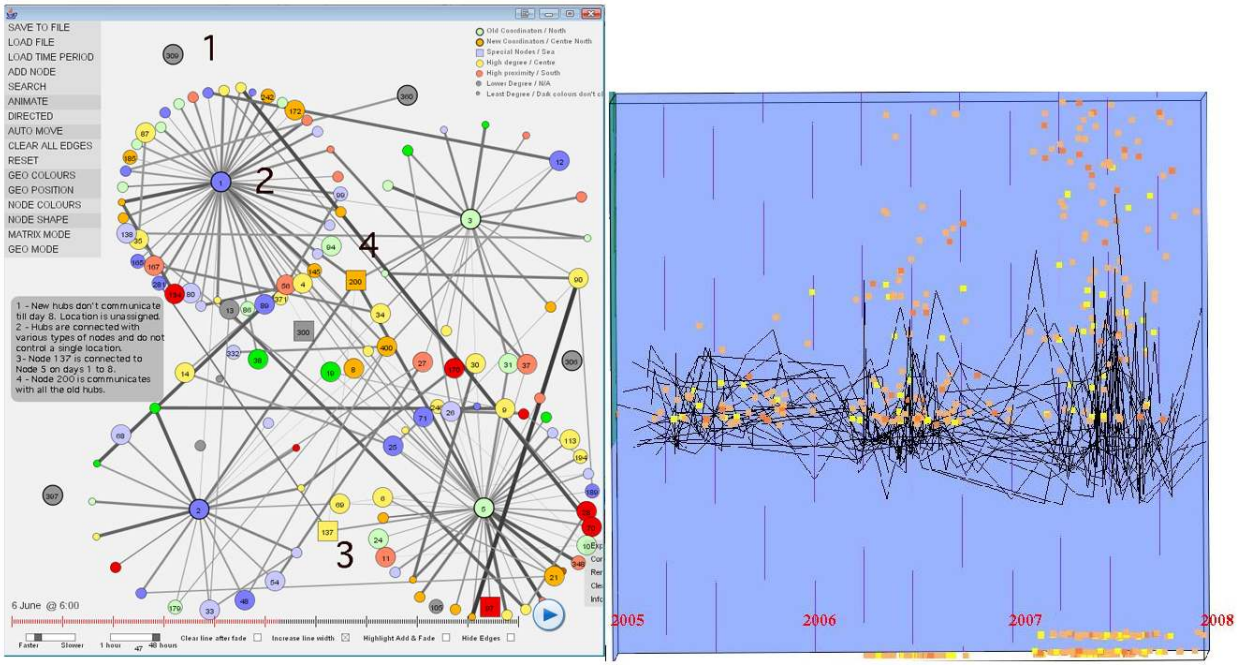

Figure 3 On the left, a dynamic social network visualization produced by University College, Dublin, Ireland. Animation shows the change over time. On the right, combining latitude locations of the Coast Guard boats and the successful landing allowed the Texas Advanced Computing Center team to recommend new areas for the Coast Guard to patrol.

In the boat migration and the evacuation mini-challenge, many teams choose to animate the data, which is the current state of the art. The problem with using only an animation for analysis is that it needs to be watched many times and this approach will not scale. But even the current state of the art was improved. For example, the visualization in Figure 4 added the concept of staining to an animation. By staining areas with different colors, the occupants who passed through an area or an exit were easily identifiable.

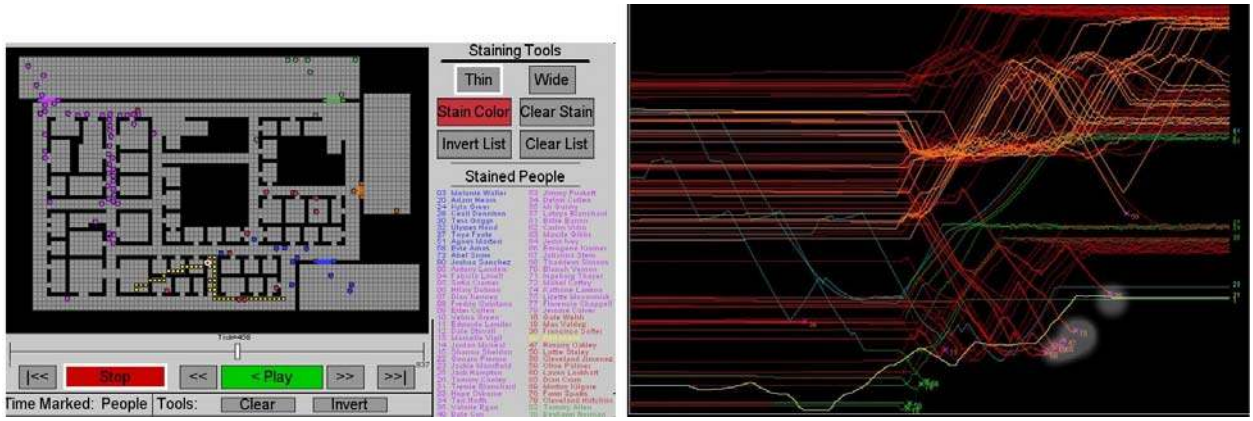

Figure 4. On the left, staining methodology developed by Southern Illinois University Edwardsville allowed users to follow groups of people. On the right, jammed exits are shown in this visualization from the University of California, Davis.

In the wiki edits mini-challenge the data could be viewed over time to determine spikes in editing behavior, and variations on social network diagrams could be used to view the different factions identified in the edit patterns. An innovative visualization was created by using a measure of disagreement over time to determine which wiki edit contributors moved into pro and con positions (Figure 5).

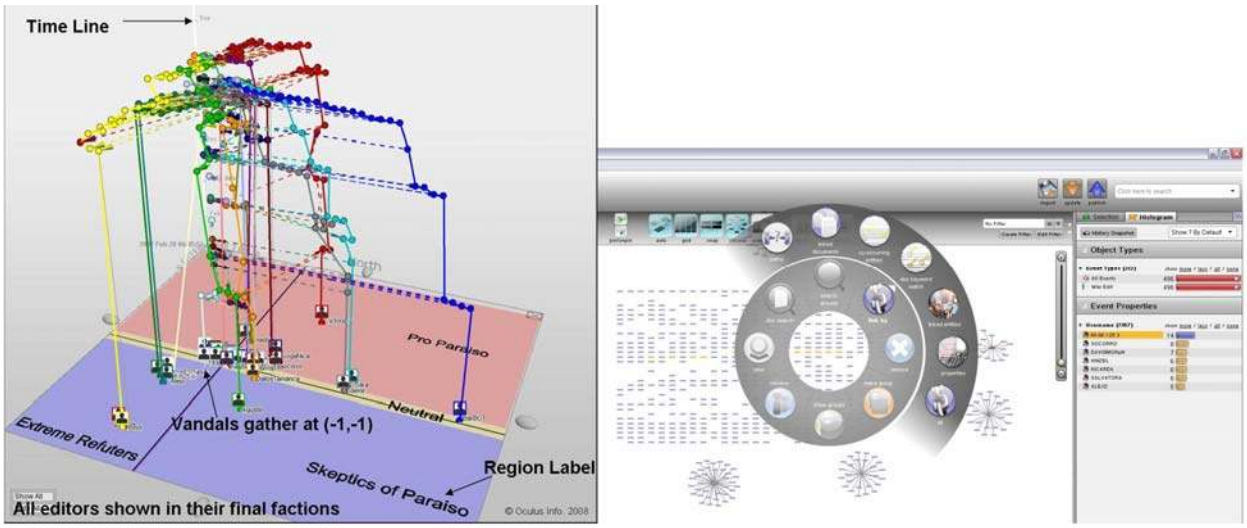

Figure 5. On the left, a visualization of faction positions over time by Oculus. On the right, interactive features of Palantir Technologies.

While no statistical analysis is possible, increased participation allowed some comparison to be done between techniques and the accuracy they obtained. For example, in the boat migration scenario, only submissions that included some mean of aggregating the data by region could identify a small island as a good destination (as animation alone did not reveal the overlapping location of the landings).

For more information on the VAST 2008 challenge, see [8].

5 Related Evaluation Efforts

The Text REtrieval Conference (TREC) [9] started as an evaluation effort for a program funded by the Defense Advanced Research Projects Agency (DARPA). It has outlived that particular program and expanded to include retrieval from many languages and media other than text. Ground truth is not inserted into TREC data sets but it is established by the National Institute of Standards and Technology

(NIST) assessors who perform the same queries that the information retrieval (IR) systems do. This assessor agreement is used as ground truth for the purpose of judging entries. The TREC tasks are straightforward (find the most relevant documents) and the TREC feedback is only quantitative - namely precision and recall, which can be calculated quickly and generate a ranking of the systems being evaluated. This differs greatly from the tasks used in the VAST challenges which are more complex and require qualitative assesment of the answers and of the process used to arive at those answers.

The tracks in TREC were our inpiration for the very successful mini-challenges incorporated into the VAST 2008 Challenge. The model of the TREC conference also inspired the Challenge workshop in 2008 which, like the TREC conference, is only opened to the participants, encouraging people to submit their entries.

The InfoVis contest started in 2003 [3] and continues today. As two of our VAST co-chairs were formerly chairs of the InfoVis contest, many similarities can be seen in our efforts. Similarities include the complex tasks and qualitative feedback for the participants. Differences include the use of real data in the InfoVis contest (for example census or publication data) while the VAST data sets are synthetic to allow ground truth scenarios to be inserted in the data. The VAST data sets have been so far smaller in size but include multiple data sets of different types with the plots distributed throughout the data sets. The embedding of ground truth in these synthetic data sets is difficult and time consuming, but we believe that the challenge of "solving a puzzle" is attracting participants. It garantees that "something interesting" is to be found in the data, which is not the case with most data sets. Finally, ground truth allows some measure of accuracy to be generated.

Many other contests have been used in the computer science community and we are preparing a survey of those contests (See www.cs.umd.edu/hcil/semvast/contestsurvey for more information).

6 Discussion

The past three contests/challenges have provided us with a number of insights into how to achieve our goal of providing researchers and developers with the necessary materials to conduct user-centered evaluations.

While automatically processing the answers to generate an accuracy measure is a desirable goal we are addressing, having analysts and visualization experts provide feedback to the participants is essential. This provides the researchers with input on the novelty of their visualization and adherence to good visual design principles, but also provides the analysts' input on the utility of visualizations in their work. The analysts make these judgments based not only on the tasks used in the challenge, but also on the reflection of other aspects of their work, so unlike TREC which uses precision and recall, the VAST Challenge requires tools and services to facilitate the management of a subjective assessment.

In 2008 we decided not to make the ground truth immediately available to the public (previous years are available), but we will share the solutions with anyone who submits an entry, even after the end of the event, and with instructors using the data in classes. The generation of these data sets with ground truth is extremely time consuming and costly, so we want to strongly encourage researchers to not only use the data but to participate in the challenge and contribute their examples to the repository (even though it is not a requirement). Knowing the ground truth prior to development may also contribute to "design to the solution." Ground truth is rarely black and white. Participants are asked to provide not only "answers" but evidence from the data that supports their answers. As our data sets are often too large to be reviewed manually, participants may find additional material that provides interesting data we were unaware of. Judging takes those additional insights into consideration. The NIST repository makes all the entries available to all (http://vac.nist.gov).

While the 2008 Challenge was successful (based on the number of submissions and feedback received from those who participated and other colleagues), it does not guarantee success in future years. Nevertheless, three years allows us to make some assessments as to the strengths and limitations of the VAST Challenge structure

The success of the Challenge is possibly due to the breadth of the disciplines covered and the fact that teams may work on smaller subsets of the problem. Finding the balance between realistic complexity and accessibility to newcomers is a challenge. Overly difficult problems or large data sets deter participants, while simple problems attract criticism from teams who have been working on that specific problem. The 2008 challenge provided good examples of accessible problems that could be tackled by newcomers, and still demonstrated that there is no good generic technical solution and that integrating data from multiple sources remains a challenge for everyone, even with small data sets. Because of the diversity of problems found in visual analytics applications, we need to vary the topics every year to provide more coverage. Choosing topics and finding clever scenarios will remain a yearly challenge for the organizers unless the challenge expands to repeat competitions on similar topics every year. While we aim to diversify to attract different communities (e.g., cyber security and video in 2009) some categories of problems may still elude us for a long time (e.g., tasks that require substantial domain knowledge such as studying changes in political stability or scientific capability).

Another assessment concerns the growth of the VAST Challenge. Like TREC, the VAST challenge has had a stable set of Chairs who provided continuity and instituted iterative improvements over the years. The VAST Challenge, like TREC, also benefited from the support of a funding organization to prepare the data sets, and from various sources (e.g., DHS, National Science Foundation and others providing access to judges) to help manage and run the event. Long-term success is likely tied to developing tools and services to faciliate the management of the challenges, and continuing support to prepare data sets and run the annual workshop to present and discuss Challenge results.

Sharing and discussing the results seems critical. We have seen the quality of submissions, both in interactive visualizations and in analysis, improve over the three years we have run the contests. The discussion of different approaches in our participant workshop last year was particularly encouraging. By seeing approaches that were more productive, participants should be able to devise different directions that go beyond the current state of the art. The teams used a variety of approaches, some simple, some mostly analytic, some purely visual, some combining approaches and some more innovative than others. This exchange of knowledge will help evolve and improve the visual analytics systems being developed. Conference attendees who attended the summary panel at the VAST Symposium and the general public who accesses the materials online after the event also have a chance to form their own opinion of the entries and techniques, using the awards as a guide to focus their attention on certain aspects of the entries.

Reward mechanisms vary greatly between contests. While some contests that can easily rate entries with a single metric offer monetary rewards (e.g., the Netflix Prize), the VAST Challenge offers mostly recognition in the form of awards - which can boost visibility, resumes or company profiles, and publications. Including the award summaries in the Symposium Proceedings was well received. All other summaries are published online and in the Symposium Compendium.
Still we decided to make the two page summaries optional as some participants do not value them and find the extra work unnecessary since all the other materials are available. In 2007 we offered the three award winners to co-author a journal paper. While this seemed a worthy reward, it was found to be a lot of work for “only” an invited paper which would not boost academic resumes, therefore this activity was not continued.

7 The Way Forward

The success of 2008 does not guarantee that the future challenges will continue to attract an equivalent number of participants. The VAST 2008 Challenge set a high bar with its 73 submissions. The VAST 2009 Challenge will follow the same format as the VAST 2008 Challenge – a number of mini-challenges and a grand challenge. Again, we will be drawing data sets from different domains, focusing on the type of data that researchers are using for their research within the area of cyber-terrorism. We will continue to insert ground truth into the data sets as we believe that this contributes to the fun and enjoyment that many of the teams tell us they have when entering the contest, and is of great value to the field.

We will be eliminating the interactive session and focusing on the participant workshop. While the interactive session provided informative insights about analysts' processes to the participants, the effort was too costly for the number of teams, three each year, who benefit. Instead, we will incorporate more analyst input into our participant workshop to benefit more researchers. The workshop will provide a more focused mechanism by which participants can capitalize on the various solutions submitted. This exchange of knowledge was requested several times by the participants. Not only did they want to see the published works of their competitors (with the two page papers) but they wanted to see demonstrations and have live discussions with their colleagues.

While we have numerous research challenges ahead of us, we are encouraged by the growth in popularity of the Challenge and in the improved quality of the entries. Although we cannot attribute these improvements solely to the contest, several teams have told us that a majority of their time preparing entries for the Challenge is actually spent making modifications to their system. These changes occur as a result of trying to use the system in a real-world task; the tasks we are providing. We will use future Challenges to highlight particular problems that visual analytics environments can help overcome. The data sets, tasks, and metrics we use in future Challenges will be tailored to make these problems visible and to reward participants who provide solutions.

We also recognize the need for the integration of computer analysis and human analysis. Keim et al. [10] provide an excellent discussion of computer generated visualizations that are suited for human analysis. To explore this integration, we will be holding a workshop at the Knowledge Discovery and Data Mining (KDD) 2009 Conference using the VAST 2008 Challenge data set and comparing KDD solutions with the 2008 VAST Challenge entries.

Acknowledgements

This work was supported in part by the National Visualization and Analytics Center™ (NVAC™) located at the Pacific Northwest National Laboratory in Richland, WA and funded by the Department of Homeland Security. The Pacific Northwest National Laboratory is managed for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC05-76RL01830. Drs. Grinstein and Plaisant are also supported in part by the National Institute of Standards and Technology. The authors are supported in part by the National Science Foundation (NSF) under a Collaborative Research Grant to the following three institutions:
- IIS-0713198 Grinstein, Georges - University of Massachusetts, Lowell
- IIS-0713087 Plaisant, Catherine - University of Maryland, College Park
- IIS-0712770 Scholtz, Jean - Battelle Memorial Institute
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 National Science Foundation.

The authors wish to thank Heather Byrne and Adem Albayrak (our Research for Undergraduate NSF students) for all their hard work for the VAST 2008 Challenge.

References

[1] Thomas, J. and Cook, K. (Eds.) *Illuminating the Path: The Research and Development Agenda for Visual Analytics,* IEEE CS Press (2005), http://nvac.pnl.gov/agenda.stm Accessed Feb. 29, 2009.

[2] Redish, J. Expanding Usability Testing to Evaluate Complex Systems, 2007. *Journal of Usability Studies,* Vol. 2(3), 102-111.

[3] Plaisant, C., Fekete, J., Grinstein, G. Promoting Insight Based Evaluation of Visualizations: From Contest to Benchmark Repository Promoting Insight Based Evaluation of Visualizations: From Contest to Benchmark Repository, *ACM Transactions on Visualization and Computer Graphics*, 14, 1 (2008) 120-134.

[4] Whiting, M., Haack, J., and Varley, C. Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software. *BELIV '08: Proceedings of the 2008 conference on Beyond time and errors: novel evaLuation methods for Information Visualization.* ACM Press (2008).

[5] Grinstein, G., O'Connell, T., Laskowski, S., Plaisant, C., Scholtz, J., Whiting, M.,
The VAST 2006 Contest: A tale of Alderwood, *Proc. of IEEE Symposium on Visual Analytics Science and Technology* (2006) 215-216

[6] Plaisant, C., Grinstein, G., Scholtz, J., Whiting, M., O'Connell, T., Laskowski, S., Chien, L., Tat, A., Wright, W., Gorg, C., Liu, Z., Parekh, N., Singhal, K., Stasko, J. (January 2008)
Evaluating Visual Analytics: The 2007 Visual Analytics Science and Technology Symposium Contest
IEEE Computer Graphics and Applications, 28, 2, 12-21 (2008)

[7] Grinstein, G., Plaisant, C., O'Connell, T., Laskowski, S., Scholtz, J., Whiting, M., VAST 2007 Contest: Blue Iguanodon, *Proc. of IEEE Symposium on Visual Analytics Science and Technology* (2007)

[8] Grinstein, G, Plaisant, C., O'connell, T., Laskowski, S. Scholtz, J., Whiting, M. VAST 2008 Challenge: Introducing Mini-Challenges, *Proc. of IEEE Symposium on Visual Analytics Science and Technology (2008).*

[9] Voorhees, E., TREC: Continuing information retrieval's tradition of experimentation, *Communications of the ACM*, 50, 11 (2007) 51-54

[10] Keim, D., Andrienko, A., Fekete, J.-D., Görg, C., Kohlhammer, J., Melançon, G. (2008) Visual Analytics: Definition, Process, and Challenges, in Kerren, A.; Stasko, J.; Fekete, J.-D.; North, C. (Eds.), Information Visualization: Human-Centered Issues and Perspectives, Vol. 4950 of *LNCS State-of-the-Art Survey*, Springer