A Reflection on Seven Years of the VAST Challenge

Jean Scholtz
Pacific Northwest National Laboratory
PO Box 999
Richland, Washington 99352
1-503-355-2792
jean.scholtz@pnnl.gov

Mark A. Whiting
Pacific Northwest National Laboratory
PO Box 999
Richland, Washington 99352
1-509-375-2237
mark.whiting@pnnl.gov

Catherine Plaisant
Human-Computer Interface Laboratory
University of Maryland
College Park, Maryland 20742
1-301-405-2768
plaisant@cs.umd.edu

Georges Grinstein
University of Massachusetts at Lowell
Lowell, Massachusetts 01854
1-978-934-3627
grinstein@cs.uml.edu

ABSTRACT
We describe the evolution of the IEEE Visual Analytics Science and Technology (VAST) Challenge from its origin in 2006 to present (2012). The VAST Challenge has provided an opportunity for visual analytics researchers to test their innovative thoughts on approaching problems in a wide range of subject domains against realistic datasets and problem scenarios. Over time, the Challenge has changed to correspond to the needs of researchers and users. We describe those changes and the impacts they have had on topics selected, data and questions offered, submissions received, and the Challenge format.

Categories and Subject Descriptors
H.5.2 [Information Interfaces & Presentations]: User Interfaces – Evaluation/methodology

General Terms
Design, Experimentation, Human Factors, Verification.

Keywords
Visual Analytics, Software Evaluation, Contests

1. INTRODUCTION
Prior to the 2006 Visual Analytics Science and Technology (VAST) Challenge, the evaluation of visualization tools focused mostly on usability and performance (e.g. speed of algorithms, controlled experiments comparing visualization components). While this was certainly needed, there was a recognized need to educate researchers about the analytic process in the hope they would design better visual analytics tools and to provide a forum for discussing – and hopefully measuring – the utility of the proposed systems. The VAST Challenge is now seven years old, and we review how it has evolved over those seven years.

The idea of using contests or challenges to advance science is not new. The National Institute of Standards and Technology has run TREC (Text REtrieval Contest) for many years [8]. The National Science Foundation runs their Science and Engineering Visualization Contest [9] and the Defense Advanced Research Projects Agency has recently run contests for self driving automobiles [3,5] and for the use of social media [4]. Netflix recently ran a contest for recommender system algorithms [10].

BELIV 2012 Seattle, WA, USA

The first VAST Challenge was actually billed as the 2006 VAST Contest [6]. It consisted of a scenario, tasks, and data containing heterogeneous information (news stories with images and maps, a voter registry and a phone call log in Excel.) Ground truth was inserted into this synthetic data. We had six entries from both academic institutions and from corporations. Entries were judged by the six VAST contest committee members along with a small number of professional analysts. Entries were judged as to whether they found the "correct answers" and how useful their visualizations were in obtaining the answers. Three high performing teams were offered a chance to work with an analyst during the VAST symposium. In this session, a new smaller dataset was provided, and the analysts worked through the scenario with the developers while providing the team with comments about the utility of the software in analyzing the data.

In 2012, there were two Big Data challenges that focused on cyber situation awareness for over 800,000 computers across a banking enterprise [1]. One challenge included a database of health and status data and the other included encoded clues in a Firewall log and an Intrusion Detection System log. We had 40 entries from a dozen countries. We had over a hundred reviewers provide over 240 reviews of submissions. Awards were given to teams who showed particular excellence in visualizations, interaction, the analytic process and the analysis accuracy.

There has been an evolution over seven years of the VAST Challenge. The datasets have changed from single heterogeneous to multiple homogeneous; the size of data has increased; the scenarios now require more domain knowledge; the process for reviewing submissions has changed; the recognition has been modified; and the accuracy has become more qualitative and less quantitative. An early paper [2] addressed the changes from 2006 through 2008 and proposed a future path for the Challenge. In this paper, we discuss the evolution of the VAST Challenges during these past seven years and the impact of these changes in advancing the state of the art in visual analytics

2. CHALLENGE PROBLEM EVOLUTION
Benchmark data sets are also popular in the scientific community. These exist in many challenges. In particular the National Institute of Standards and Technology maintains a number of datasets to evaluate fingerprint and facial recognition [7,11]. Since visual analytics is only successful when an analyst can utilize the software tool to make her job faster, easier and/or more accurate there was a need to ensure that the software could be used in an actual task. Hence, we determined that a contest for the visual analytic community required more than just a dataset to be successful, so the VAST offering would include a scenario, a problem to be investigated, data with known ground truth, and questions to help guide and assess the analyses. We drew inspiration from many sources outside of the computer science field, including successful fiction writing and theories of

deception, in addition to our experiences in applying visual analytics tools to a broad variety of problems for our clients. We describe these founding principles in other papers [17]. We learned many lessons from our initial choices, such as the need for contestants to have or obtain domain knowledge when tackling domain-intensive problems, analytics expertise to be able to tease answers out of the data, data wrangling capabilities to be able to massage datasets into forms useable by various types of software, the enthusiasm of student groups to participate and university professors to use the data to help teach element of visual analytics, and our ability to formulate the data, scenarios, and questions to be able to enable quantitative and qualitative evaluation. Another aspect of the contest that quickly came into question was what were our real motivations for the contest? Having teams solve a challenging problem provided certain insights, but did that best support our real goal of advancing the state of the art in visual analytics?

In 2006, we introduced the fictitious town of Alderwood, Washington, with a scenario that involved political scandal on the law enforcement side and bio-terrorism on the intelligence analytics side. The scenario was stated as follows:

> In January 2003, the FBI is tipped off to possible political shenanigans in the mid-sized vacation town of Alderwood, located on the banks of the Alderwood River in south-central Washington State…routine testing at a local farm discovered bovine spongiform encephalitis (BSE, also known as 'mad cow disease") sending the remains of the local economy into a tailspin due from resulting beef import embargo imposed by the Food and Drug Administration (FDA), and several foreign nations. It's been a dismal scene. Yet, a sudden influx of young talented men and women relocating to Alderwood has caused a stir. A not-so-reliable source has stated that high-paying, high-tech employment is now "a sure thing" in Alderwood, and it's all supported by "the high-rolling big boys at City Hall". What is the situation in this scenario and what is your assessment of the situation?

Participating teams were asked to provide a solution that supplied basic information on who, what, where, when, and why. In other words, from our posting:

> For each most relevant plot (there may be only one) consider the following questions: 1. Who are the players relevant to the plot? Which of the relevant players are innocent bystanders? Which of the relevant players are deliberately engaged in deceptive activities? How are the relevant players connected? What is the time frame in which this situation unfolded? What events occurring during this time frame are relevant to the plot? What locations were relevant to the plot? What, if any, connections are there between relevant locations? What activities were going on in this time frame? Which players are involved in the different activities?

The 2007 Contest was similar to 2006, but with a theme of law enforcement and eco-terrorism.

With this style of contest, we found our participation rate did not increase, so we tried a new approach that broke a problem scenario down into individual parts called Mini-Challenges. These contain a compartmentalized problem to solve, a single (or very limited number) data type to process, and a more limited scope of questions. We also designed an overarching theme, to enable ambitious performers to solve a Grand Challenge that included additional data, and whose questions required knowledge about all Mini-Challenges. We hosted four Mini-Challenges in 2008, and participation jumped to over 70 submissions.

The VAST 2008 Challenge scenario concerned a fictitious, controversial socio-political movement. Participants were provided with an excerpt from the movement's manifesto and the following four data sets, one for each mini-challenge:

- cell phone records over a 10 day period
- a chronicle of migrant boat journeys with passenger lists, launch and landing sites and landing/interdiction status
- a catalog of wiki edits to a page discussing the movement
- geospatial data of an evacuation from a building in which a bomb exploded.

In 2008, two Mini-Challenges allowed contestants to explore geospatial temporal visualization and reasoning. Another enabled participants to analyze social media information in order to determine factions participating in the discussion and to characterize how they were using the Wikipedia pages to forward their own political opinion. The final Mini-Challenge concerning this evacuation resulted in several forms of animations of people moving over time within a building, and approaches to use visualization to determine who was responsible for the bombing.

Challenges for 2009 through 2011 were similar in format to 2008, and the participation rate remained high throughout. The themes for these years included insider threat in 2009, arms dealing and pandemic in 2010, and bioterrorism and cyber security in 2011. The individual Mini-Challenges featured a wide range of data to analyze over these years including cyber data, video, Twitter-like texts, intelligence messages, and DNA sequences.

As sponsorship of the VAST Challenge changed, the approaches and goals changed as well. Sponsors who came forth quite naturally wanted the VAST scenarios to reflect issues that were in-line with problems they were facing. Therefore, the 2012 Challenge, fully funded by a new sponsor, was entirely focused on large-scale cyber situation awareness—a topic of high interest to many organizations and one that represented a considerable challenge to the visual analytics community. Instead of a fictitious city, this year we invented both a synthetic world, "BankWorld," and a large financial institution, the venerable "Bank of Money." The Bank of Money operated a world-wide network of machines that ran 24 hours a day, 7 days a week, and operations covered 10 time zones. Large-scale cyber situation awareness in this scenario included factors such as geo-spatial elements, banking business rules, operational health, and security policies. The scenario descriptions were reasonably complex, and for this domain, Mini-Challenge 1 was designed so that any visual analytics researcher could participate, while Mini-Challenge 2 required specific cyber expertise to determine a reasonable explanation for the situation. The scenario for the 2012 VAST Challenge was as follows:

> The Bank of Money (BOM) Corporate Information Officer (CIO) has assigned you to create a situation awareness visualization of the entire enterprise. This is a considerable challenge, considering that BOM operates from BankWorld's coast to coast. In addition to observing the global situation, he would also would like to be able to detect operational changes outside of the norm.
>
> Mini-Challenge 1. You are provided with two datasets that span two days of data for BOM. One dataset contains metadata about the bank's network. The second dataset contains periodic status reports from all computing equipment in the BOM enterprise.
>
> There is also one additional smaller dataset that contains a one hour snapshot of the enterprise's activities. It has the same format as the

second dataset mentioned above, and can use the metadata contained in the first dataset.

Task 1.1 Create a visualization of the health and policy status of the entire Bank of Money enterprise as of 2 pm BMT (BankWorld Mean Time) on February 2. What areas of concern do you observe?

Task 1.2 Use your visualization tools to look at how the network's status changes over time. Highlight up to five potential anomalies in the network and provide a visualization of each. When did each anomaly begin and end? What might be an explanation of each anomaly?

Mini-Challenge 2. During a time period that is NOT overlapping with Mini-Challenge 1, a Region within the Bank of Money is experiencing operational difficulties. This becomes a challenge for the operations staff, particularly as they attempt to deploy their limited number of skilled administrators to address issues occurring in the enterprise. You will be provided with Firewall and IDS logs from one of the BOM networks of approximately 5000 machines

Task 2.1 Using your visual analytics tools, can you identify what noteworthy events took place for the time period covered in the firewall and IDS logs?

Task 2.2 What security trend is apparent in the firewall and IDS logs over the course of the two days included here? Illustrate the identified trend with an informative and innovative visualization.

Task 2.3 What do you suspect is (are) the root cause(s) of the events identified in Task 2.1? Understanding that you cannot shut down the corporate network or disconnect it from the internet, what actions should the network administrators take to mitigate the root cause problem(s)?

In summary, the VAST challenge evolved from a single more complex situation analysis to a simpler and more manageable model of multiple mini-challenges. The focus of the challenge is also being guided by the needs of the sponsors who see the value of the challenge in helping develop a research community around their own problems.

3. DATASET COMPLEXITY

VAST challenge datasets have been of moderate size and moderate complexity to enable a broad spectrum of users to participate. In 2006 and 2007, the diverse datasets were very closely coupled, so that all dataset elements needed to be considered to achieve an understanding of the information. The Mini-Challenges were targeted to be single data format and amenable to both large visualization packages and newly created ones by student teams. The sizes of the 2006 and 2007 datasets were relatively small. In 2006, the data set included 1800 news articles, a small Excel telephone log, a 40K record database of voter registration data, 5 photos and a few background text documents. The 2007 dataset included 1500 news stories, 150 blog entries, a small set of images, a small set of Excel worksheets, and background documents. In 2012, the Mini-Challenges were formidable in both size and structure, reflecting the nature of today's cyber challenges. The situation awareness data consisted of about 133 M rows of information resulting in 10 GB of information. The Firewall log and the Intrusion Detection System log were 22M and 230K records, respectively.

Specialized data usually attracted fewer VAST challenge contestants. In 2007, we had a blog containing both text and hand-drawn cartoons (that were digitized) that presented a stumbling block over how best to approach the mixed modes. In 2008, we had a video analytics challenge that attracted a handful of participants (although the reviewers found the few submissions to be well executed). Both 2006 and 2007 necessitated text analysis, although in 2007 we provided the tokenized text data if requested by a team. In 2010, one of the Mini Challenges contained some simplified DNA sequences for analysis, however there was considerable outreach to the biology community that kept participation for that Mini-Challenge reasonably high.

4. SUBMISSIONS

Over the years, we have requested submission materials (i.e. answers) in several different forms, mainly targeted at providing the best possible set of information about a contestant's solution for a reviewer to examine. These forms have included:

Short Answer: A short answer consists of text describing the answer and how it was arrived at. It is limited to 300 words (including captions) and a maximum of 3 screen shots.

Detailed Answer: A detailed answer allows for more explanation than a short answer. They are limited to 1000 words (including captions) with a maximum of 10 screen shots. Detailed answers should provide the answer and describe in detail the process used to arrive there. Detailed answers help reviewers judge the depth of the understanding the team has gained of the situation and the reasoning that was used.

Debrief: Debriefs were requested in the Grand Challenge. The debrief is basically the analytic product that a professional analyst would deliver after doing the analysis. A debrief is a maximum of 2000 words narrative describing the hypothesis about the situation at hand. If there are uncertainties, the debrief should include suggestions of the possible next steps to clarify those uncertainties. Making those debriefs available (especially those that won awards) has certainly helped improve the quality of the submissions in later years and helped the community understands their importance.

Video: The purpose of the video is to describe the analysis process used in arriving at the solution with the intent of highlighting the interactive functionality of the system. We allowed 4 minutes for Mini-Challenges, and 15 minutes for a Grand Challenge. To date, reviewers all agree that the video provides the best description of the challenge entries out of all submitted materials.

Specific answers: We have also requested specific answers such as a list of people involved in a plot, a number of most relevant documents used in determining an answer, or the geo-locations of events. Where possible, specific forms were given for submitting the answers so they could be automatically checked for accuracy. This is the only submission form that allows automatic evaluation.

5. JUDGING PROCESS AND CRITERIA

Evaluation of software used in analysis of complex datasets is problematic. Unlike usability evaluations of standard software, such as word processing systems, visual analytics requires a much longer time frame for evaluation and while the resulting answer is of interest, so is the process that was used to arrive at that answer [13,12]. Also, as the process requires interpreting visualizations and assessing underlying processing algorithms, a multi-disciplinary set of reviewers needs to evaluate the utility of the visual analytics software. In the VAST Contests and Challenges, we use visualization experts, human-computer interaction experts and domain analysts. As software systems are often early prototypes it is not realistic to ask our judges to download and use

the software so they must rely on the explanations including videos provided by the teams.

The VAST Contests of 2006 and 2007 were judged by the VAST Contest committee plus some additional analysts. In 2008 we did a two pass review because of the large number of submissions we received. The VAST Challenge committee did a first pass and then submitted the most promising entries to the external analysts for their consideration. The VAST Challenge committee also did a more detailed review on these submissions.

As we anticipated another year of high submissions in 2009, we recruited external reviewers from both the professional analyst community and from the visual analytics community. We devised a rating sheet that asked for ratings in the general categories of visualizations, interactions with visualizations, support for analytic process and the analytic results. Accuracy results were supplied to the reviewers where possible. If no quantitative results were feasible, the reviewers were supplied with the solution devised by the Challenge committee.

The early scenarios and corresponding ground truth were constructed so that specific answers could be given and accuracy could be, albeit with some difficulty, quantitatively specified. That is, in some early instances, we had to devise penalties for providing inaccurate information. Suppose we asked participants to give a list of all the people involved in a terrorist plot. If a team gave us the 10 actual names plus four other names, then they were penalized for providing the four names of those who were not involved.

Detailed answers and debriefs were judged not only for accuracy but were also judged as to the clarity of the explanation, the identification of the evidence, and the rationale for the overall analysis of the situation.

In the 2012 VAST Challenge, the answers necessitated an analysis of the data, rather than providing more quantitative answers. So, in fact, the answers were more like a debrief than mini challenge answers had been in previous years. Thus the accuracy was judged subjectively by the reviewers, who needed to use the Challenge team's description of what the situation was and interpret the answer the team gave in view of that. Then they were asked to rate specific aspects of the answer in addition to the visualizations and the interactions.

6. INCENTIVES

In the early Challenges of 2006 and 2007 we named winners that were determined by accuracy scores as well as subjective ratings in the areas of visualizations, interactions, support for the analytic process, and analysis reports. We differentiated between winners in the student competition and in the corporate competition. In 2008 we decided to provide awards where teams had made outstanding contributions in some area of analysis or visualization. The awards were not pre-determined but came as a result of nominations from reviewers for specific awards and from consensus of the VAST Challenge committee. Again, these awards were given in the areas of visualizations, interactions, support for the analytic process, and analysis reports. We also distinguished between the use of toolkits and the development of systems in the awards process.

In addition to receiving awards, teams participating had other opportunities for additional exposure. These have included:

- Recognition at the VAST Conference
- Two page papers in the VAST Conference proceedings
- Participation in the real-time VAST Challenge (2006, 2007, and 2008)
- Participation in a special session at the VAST Conference, ranging from a 3 hour session to a full day workshop.
- Publication of their entries on the repository web archive (www.cs.umd.edu/hcil/varepository)

Participants were also given access to the solutions soon after the submission process closed.

7. SUBMISSIONS AND TEAMS

There were six teams participating in the 2006 VAST contest, then seven in 2007. The introduction of Mini Challenges in 2008 increased participation dramatically to 63 teams. Twenty eight different organizations from thirteen countries submitted entries among which thirteen were student teams. In 2009 there were 49 entries, and teams came for 28 different organizations from 13 countries. In the following years the number and diversity of the teams remained steady.

While the number of team is still relatively limited and has never reached the high numbers of simpler contests or contests with monetary rewards, it is important to note that the number of submissions is not the true indicator of the use of datasets since the datasets are available after the contest as well (see later section on the Repository). It is also notable that we have student teams and industry teams. In particular the number of student teams has been increasing beginning in 2010.

Table 1: Participation in the VAST Challenge per year

Year	# of Entries	# of Organizations	# Student Teams
2006	6	6	3
2007	7	7	3
2008	73	28	9
2009	49	28	8
2010	58	31	18
2011	56	Unknown	38
2012	40	30	18

8. AWARDS

One way to determine if the state of the art is being advanced by the VAST Challenges is to look at the awards that have been given over the past years. We will ignore the contests of 2006 and 2007 as there were essentially winners declared in those. The table below lists awards by Mini Challenges and Grand Challenges for the years 2008 through 2012. In the table, a "special" award is one that was given for an unusual feature in the software or process or submission that the committee wanted to recognize but that will most likely not be repeated. For example, one year a team did user testing to see if users could arrive at the correct answer using the visualization they had developed.

If we look at the awards by domain, we see that we have text data in both 2010 and 2011, however, there was a low number of awards made for good visualizations. Data that needs to be

visualized as social networks appears in both 2008 and 2009. In 2009 there were more awards per submissions but the percentage of those for visualizations increased only slightly. However, the visualizations in 2009 were somewhat more sophisticated as they introduced visualizations of uncertainty. Situation awareness for cyber security appeared in both the 2011 and 2012 challenges but with over three times the participation in 2012. While there were no awards for visualizations in 2011, there were three in 2012 as well as one award for the analysts' choice of tool. The smaller number of participants for the firewall and log data in 2012 could probably be attributed to the large amount of data that participants had to deal with in this challenge. However, it is promising that there were 3 awards given for visualizations that had to deal with large amounts of data.

In 2009 we awarded the analysts' choice to a team in the Grand Challenge. Similar to 2011, comments from the professional analysts indicated that they would appreciate having this tool on their desktop.

Of interest is the number of awards given for good analysis and for support for the analysis process over the years. This is encouraging and leads us to believe that the VAST Challenges have provided software developers, researchers and students with a better understanding of analysts' needs.

Table 2: Number of awards and award types by year

Year	Data Type	Entries	# Awards/Type
2008	Cell phone records (social networks)	22	5 total/3 visualizations/1 toolkit/1 accuracy
	Boat landings – geospatial/time	13	2 total/1 integrated display/1 analysis
	Evacuation records	20	3 total/1 visualization/ 1 toolkit integration/1 special
	Wiki edit records	12	No awards
	Grand challenge	6	3 total/1 support for analysis/1 data integration/ 1 analytic environment
2009	Badge and network traffic	22	7 total/3 analysis/3 visualization/ 1 flexibility
	Social network	17	7 total/2 visualization of uncertainty/1 novel vis/2 debrief/1 tool/1 special
	Video	5	1 total/integration of open source tools
	Grand challenge	5	2 total/1 analyst's choice/1 integration of mini challenges
2010	Text records- arms dealings	14	4 total/1 visualization/1 integration/1 analysis/1 data ingest
	Hospital records-pandemic	22	5 total/1 visualization/2 process/1 overall/1 special
	Genetic Sequences	17	3 total/1 visualization/ 1 process/1 toolkit
	Grand challenge	5	3 total/ 1 debrief/2 special
2011	Geospatial and micro-blogging	30	4 total/2 visualization/1 toolkit/1 integration of computational and visual methods
	Cyber security/situation awareness	8	3 total/1integrated overview displays/1 tool adaption/1 scalability
	Text analysis	13	3 total/1 analytic process/1 analysis/1 special
	Grand challenge	5	1 total/1 comprehensive submission
2012	Cyber security/situation awareness	27	11 total/4 visualization/1 analysts' choice/2 visual design/1 interaction/1 comprehensive/1 support/1 special
	Firewall and IDS logs	13	5 total/ 3 visualization/2 special

There have been few awards in the area of interaction but this may be limited as the reviewers have to depend on the video demonstrations to provide an idea of the interactions the analyst has. In 2011, there was an award for potential scalability in the cyber security mini challenge.

Interestingly, the number of participants in the Grand Challenges has not increased over the years. This may be due to the amount of work and expertise required to develop software to accommodate visualizations of significantly different data types as well as the work in solving the Grand Challenge requires solving all of the mini challenges.

9. REPOSITORY

All submissions are available from the Visual Analytics Benchmark Repository web site [15,16] along with the awards, datasets, and solutions. The website is designed to make it easier to find the various benchmarks and their corresponding uses. It displays benchmark information like its description, location, solution etc. Users can view benchmarks based on their provenance and topics. The repository currently holds 30 benchmarks, 1424 uses and 49 publications. The website went live in September 2009 and had 2400 hits within 10 months. In the last two years it has received between 200 and 300 visits a month with more than 45% of visits being return visits (based on Google Analytics data).

Table 3: Early downloads per year

Year	No. of downloads as of Aug. 2010	Elapsed Time
2010	517	~ 7 months
2009	731	~ 1.5 y
2008	690	~ 2.5 y
2007	189	~ 3.5 y
2006	463	~ 4.5 y

The number of dataset downloads was fairly accurately tracked until 2010, after which spamming of the download site made it difficult to estimate legitimate download counts.

Using the limited 2010 data we could compare the email ids of all the users who downloaded, and see returning users from 2006 through 2009. There are 65 users who used 2006 and 2007

datasets and similarly there are 69 users who used 2007 and 2008 datasets. Among these users there are 39 of them who download all three (2006, 2007 and 2008) datasets. Among the five hundred odd users of the 2009 dataset very few are users from previous years. These counts are not entirely reliable since they are based on the email ids that users enter while registering to download a datasets. A probable explanation for the few users returning in 2009 is that the topic changes radically (from text to tabular data). We could also recognize people who participated in the contest for all four years but changed their email ids during that time.

Current counts of downloads of datasets varies from two to four thousands per year. Even with an approximate 30-50% spam this suggests a larger user base. We also know that users of the datasets often have their colleagues and students download the datasets, which were then shared locally among colleagues and friends, again making it difficult to estimate real usage.

The repository also invites people to submit additional datasets or to share descriptions of their usage of the datasets after the end of the corresponding challenge review process (either new entries in the challenge format, or papers describing their use of the datasets). Practically no one ever spontaneously contacted us to do this. The very few such additions were triggers by specific requests from the Challenge committee members. Incentives and reward structures are not in place to support the posting of datasets and the citation of datasets in visual analytics scientific papers.

10. SURVEYS

Two surveys were conducted in 2009 and 2012. The first survey was conducted to get user feedback from the participants of VAST Challenge 2009 with regards to the usefulness of the Challenge and the datasets. The survey was administered via Survey Monkey, and an email was sent to all 2009 teams in February 2010, i.e. 4 months after the symposium (so they had time to reflect on the long term benefit of the experience).

The survey had 35 responders. The results show that around 80% of the users agree that the VAST Challenge allowed them to better understand the tasks that would be performed by analysts (Figure 1).

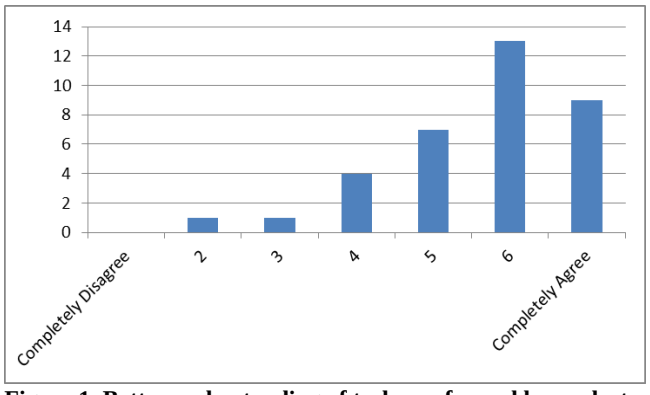

Figure 1. Better understanding of tasks performed by analysts.

Most (around 80%) of the users agree that the Challenge helped them in improving their tools and sometimes even build new tools and approaches (Figure 2). One user even commented that they developed a novel system which would not exist if not for the VAST challenge. The majority also thought having ground truth in the dataset was important (Figure 3) and one of the user commented that it increases the sense of puzzle solving. Another noteworthy result was that the users were quite divided on the usefulness of the reviews given by the reviewers and analysts. Overall the survey results are satisfactory and support the effort undertaken by the VAST Challenge committee.

We also collected feedback from users through email, which confirmed that datasets are often used to teach classes [18] and various other purposes.

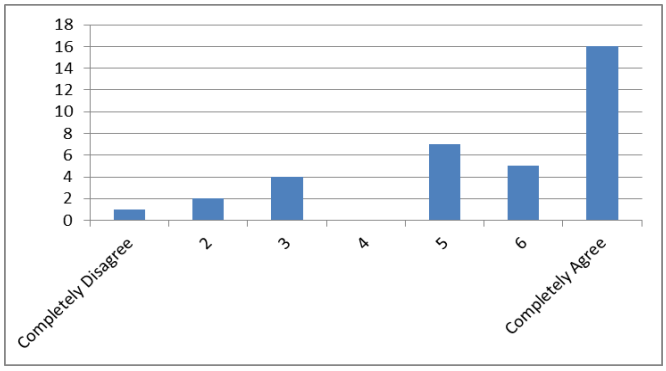

Figure 2. Helped in improving the tools.

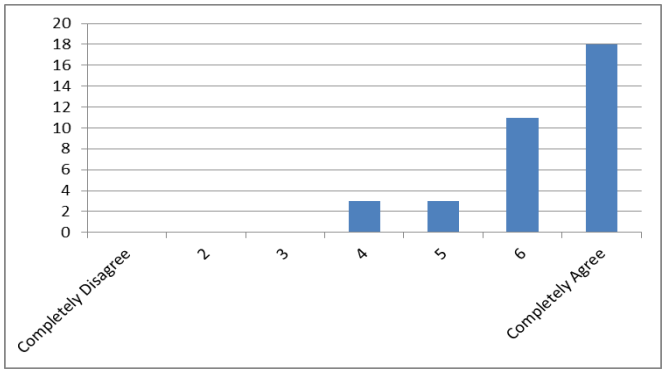

Figure 3. Ground truth in the dataset is important.

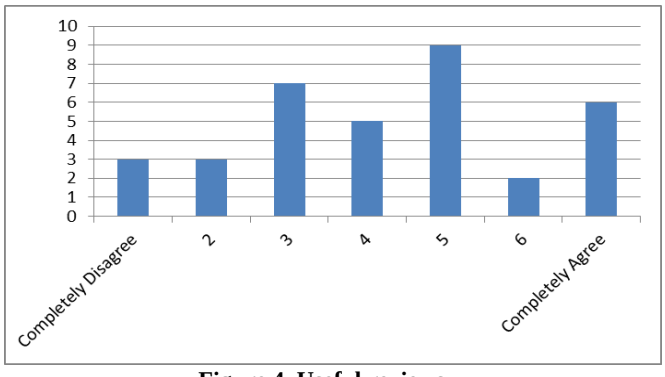

Figure 4. Useful reviews.

After the VAST 2012 Challenge judging had ended we sent a second survey to past participants in order to gauge their reactions to the changes that have been occurring in the VAST Challenges over the years. Participants had to have entered at least one of the challenges in order to answer the survey. There were 27 respondents to the survey. No one participating in this survey

entered the VAST 2006 and VAST 2007 Contests. The majority had entered the VAST 2010, 2011, and 2012 Challenges.

We asked eleven questions about the VAST Challenge and its impact on their goals for their research. .

When asked "If you didn't enter a VAST Challenge recently why was that?" Twelve responses were given to this question. The most common reasons were finding folks to work on the team and finding the time to do this. A few respondents noted that the datasets of later challenges were not appropriate for their work.

Participants were asked "Why did you enter the VAST Challenge?" and could check all that applied in a list (Table 4). An "other" option was offered as well. Of the 26 people responding to this question 19 checked "it looked like fun". Understanding the analytic process and testing their software against dataset and questions with known answers were also common reasons for entering. Comments included finding novel visualizations for the data, good to build collaborations and work as interdisciplinary teams, trying new datasets, and getting work published.

Table 4. Responses to "Why did you enter the 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

When asked "Were their specific lessons you learned in entering the VAST Challenges? Fourteen responded to this question. A number mentioned the time and effort required to pre-process the data. Others noted that code robustness is not prioritized in the academic world. One noted that focusing on one or two mini challenges was more effective than trying to solve all of them. Several noted that prototypes took a long time to build and that there was a huge gap between a prototype and one that could actually be used in an analytic process.

Twenty-four responded to the question "There were several parts to the VAST Challenge. Which were the most helpful to you?" (Table 5). The most popular response was working on the solution and developing the software. Feedback on the accuracy of the answer and receiving professional reviews were the next two most popular answers. A number commented on the importance of getting feedback on the answers and on their approach. They felt that this information helped them to know that they were on the right track. Interestingly one respondent commented that in the case of text data, the answer was a little less precise than they would have liked.

Table 5. What parts of the VAST Challenge were most useful?

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

We also asked again "Were the evaluations that you received helpful or not? Please explain" Twenty three responded to this question. Ten said the evaluations were definitely helpful. Ten said they were somewhat helpful and three said they were not at all helpful. The comments mostly said that reviews varied in usefulness with some being extremely helpful and others less so. Comments also pointed out that there were different perspectives taken in the reviews. Others said that there did not seem to be strict guidelines for the reviews. One commented that while it was nice for those entering to also do the reviews, this really necessitates a primary reviewer to mediate the reviews.

Reviewers are invited from both the visualization field and analysts in the domain field and each paper is assigned at least one domain analyst. These two types of reviewers do indeed look at things from a different perspective. Earlier work [14] showed that the clarity of the explanations submitted influenced the reviews and in some instances the accuracy of the solution submitted also impacted the reviews.

Eighteen responded to the question "Are the VAST Challenges improving each year, staying the same or becoming worse?" Six responses selected "improving", nine selected "staying the same", and three selected "becoming worse". Comments included the increasing complexity of the datasets and hence more time involved in completing the challenge; too much focus on terrorist/disaster scenarios; liked the varied coverage each year; not enough diversity in the 2012 Challenge; not all datasets are equally fun to work with. The comments reflected the diversity of participants' needs and abilities.

11. PARTICIPANTS' WORKSHOP

A participants' workshop has been held since 2006. That year it was held for several hours the first evening of the Conference. Since that time it has been increased to a full day, and in 2012 it was opened to all Conference attendees. The 2011 survey showed that participants appreciated the presentations by the award teams and the demonstrations that teams gave. The networking and seeing how others approached problems was highly valued.

12. SUMMARY

The Challenge continues to attract significant numbers of participants and feedback suggests that the event and the problems that made available are of great value to the Visual Analytics

community. How has the evolution of the Challenge impacted the state of the art of visualization? We are encouraged by the awards given, showing an understanding of the analytic process and the support needed, as well as novel visualizations in a number of different domains. We are also encouraged that participation by students has increased. We hope that this will continue even as more domain knowledge becomes necessary to solve scenarios. Much more could be done to widen the range and diversity of datasets and problems, but continuing support for this activity remains a yearly struggle. We are hopeful now that we are seeing new sponsors approaching the Challenge Committee to investigate supporting the development of datasets and the management of this multi-step interdisciplinary event. This will most certainly provide researchers with more current real-world problems to tackle. Challenge participants and reviewers struggled this year with answers that were much more subjective. It required reviewers to interpret both the answer we supplied to them and the answer given in the submission. We will need to develop a better process for reviewing the submissions if the answers continue to be more qualitative. However, as this is certainly the case in the real-world, it is probably the direction we should head. We will be interested to see if this is a deterrent to the fun of arriving at the answer.

The number of downloads from the repository continue to grow so we assume that others in the community are finding our Challenge packages useful.

The biggest issue will most likely be the domains tackled. A reliance on multiple sponsors also makes the generation of a grand challenge problem difficult. As the field grows more sophisticated it will be interesting to see if a grand challenge is necessary or if the increased complexity of the mini-challenges will suffice.

Although the evolution has not all been planned, we feel that the VAST Challenges are currently on track to grow the state of the art in the field of visual analytics. We will continue to track both the participation and the quality of the visual analytics submissions and adjust the parameters as necessary.

13. REFERENCES

[1] K. Cook, G. Grinstein, M. Whiting, M. Cooper, P. Havig, K. Liggett, B. Nebesh, and C. Paul. VAST Challenge 2012: Visual Analytics for Big Data, In *Proc. VAST 2012*, (Seattle,14-19 Oct 2012).

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

[3] The DARPA Grand Challenge 2005. http://en.wikipedia.org/wiki/DARPA_Grand_Challenge. Accessed Oct. 30th, 2012.

[4] The DARPA Red Balloon Challenge. http://archive.darpa.mil/networkchallenge/. Accessed Oct. 30, 2012.

[5] The DARPA Urban Challenge. http://archive.darpa.mil/grandchallenge/. Accessed Oct. 30th, 2012.

[6] G. Grinstein, T. O'Connell, S. Laskowski, C. Plaisant, J. Scholtz, and M. Whiting. VAST 2006 Contest: A Tale of Alderwood, In *Proc. VAST 2006*, IEEE Computer Society Press, (2006), 215-216.

[7] Image Group Fingerprint Overview, http://www.nist.gov/itl/iad/ig/fingerprint.cfm. Accessed Oct. 30th, 2012.

[8] Text REtrival Conference (TREC).. http://trec.nist.gov/. Accessed Oct. 30th, 2012.

[9] National Science Foundation Science and Engineering Visualization Contest. http://www.nsf.gov/news/special_reports/scivis/challenge.jsp . Accessed Oct. 30th , 2012.

[10] Netflix Recommender Systems Contest. http://www.netflixprize.com//index. Accessed Oct. 30th, 2012.

[11] NIST Facial Recognition, http://www.nist.gov/itl/iad/ig/face.cfm. Accessed Oct. 30th, 2012.

[12] C. Plaisant. The Challenge of Information Visualiation Evaluation, In *IEEE Proceedings of AVI*, 2004.

[13] J. Redish, Expanding Usability Evaluation to Test Complex Systems. *Journal of Usability Studies*. 2, 3 (May 2007), 102-111.

[14] J. Scholtz, Developing Qualitative Metrics for Visual Analytic Environments. In *Proceedings of BELIV '10*, 1-7.

[15] Scientific Evaluations Methods for Visual Analytics Science and Technology (SEMVAST) www.cs.umd.edu/hcil/semvast

[16] Visual Analytics Benchmark Repository. http://hcil.cs.umd.edu/localphp/hcil/vast/archive/

[17] M. Whiting, J. Haack, and C. Varley. Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software. In *Proc .BELIV '08*. ACM, (2008), 1-9.

[18] M. Whiting, C. North, A. Endert, J. Scholtz, J. Haack, C. Varley, J. Thomas. VAST contest dataset use in education. In *Proc. IEEE VAST 2009 Symposium* (2009), 115-122.