Editor: Theresa-Marie Rhyne

Toward Measuring Visualization Insight

Chris North
Virginia Polytechnic Institute and State University

> Insight: The capacity to discern the true nature of a situation; The act or outcome of grasping the inward or hidden nature of things or of perceiving in an intuitive manner. —*Merriam-Webster*

Recent visualization research literature has paid an increasing amount of attention to evaluating visualizations. For example, this trend was evident at the 2005 IEEE Visualization conference where many of the presentations included evaluation components. The conference provided some inspiring presentations that probed the philosophical boundaries of visualization and evaluation. Unfortunately, there were also a few too many Boolean usability studies that offered only two alternatives: either the users liked the visualization tool in question, or they did not. In between, there was a variety of rigorously controlled experiments. Therefore, in light of this trend, it seems an appropriate time to reopen the question about what the ultimate purpose of visualization is and how it should be evaluated.

One potential claim is: The purpose of visualization is insight. The purpose of visualization evaluation is to determine to what degree visualizations achieve this purpose.

If this claim is true, then evaluating visualizations should seek to determine how well visualizations generate insight. Measuring insight would enable the direct comparison of visualization design alternatives, or the comparison against an insight goal. But what, exactly, is insight? How can it be measured and evaluated? Do current approaches for evaluating visualizations provide measures of insight?

Defining insight

Insight has been commonly stated as the broader purpose of information visualization by many authors. The recent emphasis on visual analytics has stimulated an interest in better understanding the purpose of visualization and rigorously evaluating progress toward that purpose. Insight seems to capture the intuitive notion of visualization's purpose. However, for the most part, the definition of insight remains fairly informal, making success difficult to evaluate.

A default and implicit definition is to equate insight with user tasks, such as finding extreme values.[1] That is, the answers to questions about the data constitute insight. Yet, truly measuring the insight-generating capability of visualizations will require a more in-depth examination of insight itself. Perhaps researchers have resisted the temptation to formally define insight, believing that any formal definition would either be too restrictive to capture its essence or too general and vague to be useful. Hence, instead, it might be helpful to identify essential characteristics of insight, and then consider whether measurement methods capture those characteristics.

In the spirit of gaining understanding of insight, here I list some of its important characteristics.

- *Complex.* Insight is complex, involving all or large amounts of the given data in a synergistic way, not simply individual data values.
- *Deep.* Insight builds up over time, accumulating and building on itself to create depth. Insight often generates further questions and, hence, further insight.
- *Qualitative.* Insight is not exact, can be uncertain and subjective, and can have multiple levels of resolution.
- *Unexpected.* Insight is often unpredictable, serendipitous, and creative.
- *Relevant.* Insight is deeply embedded in the data domain, connecting the data to existing domain knowledge and giving it relevant meaning. It goes beyond dry data analysis, to relevant domain impact.

Typically, the most interesting or important insights are those that rank highly in each of the previous characteristics. For example, complexity is determined by how much data is involved in the insight. Simple insights, such as finding minimum or maximum values, involve simple answers that require only one data value. Recognizing a normal distribution of values is more complex, and might involve approximate parameters of distribution shape. Knowing that the distribution of values looks like the histogram in Figure 1 is even more complex because this understanding involves more data and thus reveals the peculiarities in its shape.

Evaluating visualizations

A variety of visualization evaluation methods exist, including empirical methods such as controlled experiments, usability testing, longitudinal studies, and analytical methods such as heuristic evaluation and cognitive walkthroughs.[2-3] Controlled experiments are increasingly common in the literature because the con-

 Published by the IEEE Computer Society

trolled nature of that method best enables researchers to rigorously measure and conclusively compare visualizations. Hence, this article examines the capability of the controlled experiment method to measure insight.

Controlled experiments on benchmark tasks

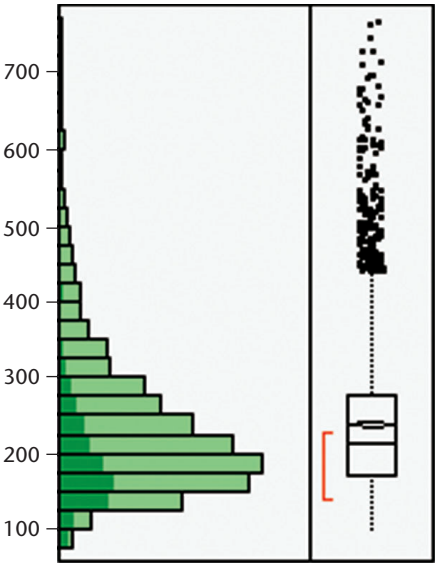
The use of controlled experiments on benchmark tasks is the primary method for rigorously evaluating visualizations.[4] In this procedure, objective metrics are captured while a sample set of human participants uses the targeted visualizations to perform a series of benchmark tasks. An example benchmark task might ask users to find the maximum value in the data set.

The two primary independent variables are the visualization design alternatives and benchmark tasks. Others common independent variables include the data set size or type, or the user class. Primary dependent measures are the user performance time to complete the task, and the accuracy of their task response (for example, the correctness of their answer). Other dependent measures include behavioral metrics such as the number of mouse-click actions. Researchers can then relatively compare the targeted visualizations based on the captured metrics to determine, for example, which visualization design resulted in the fastest user performance for a given benchmark task.

The usefulness of this method as a way to measure insight hinges entirely on whether the benchmark tasks and metrics adequately represent insight. However, benchmark tasks have four fundamental problems that are in direct opposition to the desired characteristics of insight as listed previously.

- They must be predefined by test administrators. Users must precisely follow specific instructions during the experiment, leaving little room for unexpected insight.
- They need definitive completion times. Task times must be short, typically under one minute, to examine a large number of tasks or repetitions. Stopping the timer clock at exactly the moment the user finds the answer leaves little room for deep insight.
- They must have definitive answers that measure accuracy. Multiple choice task questions enable objective mechanical (and even automated) scoring and treat answer correctness as Boolean, leaving little room for qualitative insight.
- They require simple answers that users can easily specify. Users click on the desired data object or state its unique identifier value, leaving little room for complex and relevant insight.

Because of these problems, experimenters are forced to use overly search-like tasks that don't represent insight well. Two examples of common benchmark tasks in evaluations in the information visualization literature are: find the data record that meets the following attribute criteria and find which record has the maximum value of attribute X. While such tasks might indicate that users can quickly locate or discriminate individual values, it seems far too simplistic and constrained to provide a useful indication of the kind of insight that visualization is trying to achieve. In many cases, a simple query, sorting function, or summary statistics might solve the benchmark task faster.



1 Histogram showing a data distribution of median monthly rent costs (in dollars) in all US counties. Horizontal bars indicate the relative number of counties in each rent value range. Dark portions of bars indicate counties in the southeastern US. (Generated with SAS JMP.)

Furthermore, predefined benchmark tasks are troublesome for several reasons. They force users into a line of thought that they might not otherwise take. They place an undo burden on evaluation designers, who are susceptible to bias and have little structured guidance to overcome it. The choice of tasks and the phrasing of task questions can introduce bias toward one of the visualization designs. Benchmark tasks lack completeness. A visualization might perform well at certain tasks, but at what cost? What other tasks will suffer? Finally, because the tasks must be predefined, the experiment's results are limited to only the tasks that evaluators chose. To generalize the results beyond simple benchmark tasks, researchers make the implicit claim that complex tasks will be built upon simple tasks, like a hierarchical task decomposition. Hence, if a visualization can efficiently support a variety of such simple tasks, then complex tasks will also be efficient.

The counterargument to this claim is twofold. First, the efficiency of the simple benchmark tasks is often due to specific visualization interface features that don't generalize to more complex tasks. For example, one visualization might use larger text labels for reading detailed data values, or perhaps includes a dynamic query slider that is a particularly good match for the criteria-finding tasks. Second, such a clear task decomposition does not exist as yet, and so it's unclear which simple tasks should be tested to support more complex tasks. The chasm between simple single-data-value tasks and complex synergistic tasks seems large. For example, treemaps make it easy to find the largest rectangle, but probably distorts the recognition of rectangle size distribution. Is a data distribution recognized by performing a large number of value-reading tasks?

A further confusion in the interpretation of benchmark experiment results is the tradeoff between performance and accuracy. What does it mean if a visualization has better performance time but lesser

accuracy than another visualization? Was the visualization really faster, or were participants giving up and guessing? Attempting to equalize accuracy by forcing users to continue until correctly completing the task, leads to a trial-and-error approach by users. Alternatively, filtering incorrect responses in the analysis ignores important information. None of these options corresponds well to insight.

Controlled experiments on benchmark tasks can provide a rigorous method for examining specific perceptual effects and specific tasks from a usability specification. However, it does not provide a satisfying representation of insight capability.

We need new evaluation methods that attempt to measure insight more directly. We also need to preserve the positive aspects of the controlled experiment methodology to enable rigorous comparison of visualizations. How can we modify or tweak the controlled methodology to better capture the desired characteristics of insight?

Toward insight: more complex benchmark tasks

An initial step is to include benchmark tasks of greater complexity in the experimental protocol. For example, we can ask users to characterize the distribution of data values, and include a multiple choice set of answers such as "normal," "uniform," "linearly increasing," and so on. We can carefully craft similar questions for correlations or other types of patterns. Another possibility is estimation tasks, in which users estimate various metrics, such as coverage or cluster density, without actually counting. In this case, the correctness measure can be a real value instead of Boolean.

While these benchmark tasks still suffer from some of the problems identified previously, they at least begin to test more synergistic, complex tasks that involve some uncertainty. These types of tasks generally support visualization overviews rather than detail views.

Forcing users to interpret the visualization into a textual answer ensures that they have developed their mental model of the data. However, forcing users to articulate their answers, by not providing them with multiple-choice answers, might be difficult to score. Alternatively, providing multiple-choice answers can lead users into a process of elimination, creating a challenge for the evaluation designer to provide careful wording of the possible answers. Phrases with visual metaphors, such as "densely clustered" or "higher," might bias toward a particular visualization and should be avoided.

The difficulties with this method are longer task times, greater variability in task times and correctness, and greater difficulty in designing isomorphic tasks (different instances of the same task type for use in repeated measures). Together, these problems generally mean that researchers must test more participants to get statistically significant results.

Toward insight: eliminating benchmark tasks

A more radical step is to eliminate the pesky benchmark tasks from the protocol entirely.[5] This method's fundamental concept is to change the benchmark tasks from an independent to a dependent variable. Hence, instead of instructing users in exactly what insights to gain, researchers observe what insights users gain on their own.

This method involves the following key innovations:

- an open-ended protocol,
- a qualitative insight analysis, and
- an emphasis on domain relevance.

With an open-ended protocol, users explore the data in a way that they choose. Giving the users a chance to think of initial questions helps them get started. But soon they go beyond those initial questions in depth and unexpectedness. Users are instructed to explore the data and report their insights until they feel that they have learned all that they can from the data.

We need new evaluation methods that attempt to measure insight more directly.

Using qualitative insight analysis, users verbalize their findings in a think-aloud protocol so that evaluators can capture the users' insights. Each finding that the user reports is marked as an insight occurrence. Then, for each insight, a coding method quantifies various metrics such as insight category, complexity, time to generate, errors, and so on. For example, depth could be coded on a scale of 1 to 5, where 1 represents a simple obvious fact in the data and 5 represents a deep inference that integrates multiple data types. Insight categories can be developed by examining the entire collection of insights for common clusters. Usability and human factors experiments commonly use such coding methods. Coding converts qualitative data to quantitative and is inherently more subjective, but supports the qualitativeness of insight. Significant objectivity can be maintained through rigorous coding practices.

To emphasize domain relevance, experiment participants should be users from the target domain. Independent domain experts acting as coders provide critical metrics for the value or importance of the reported insights in the domain. Experimenters should pay special attention to cases where the user goes beyond dry data analysis, and makes domain-specific inferences and hypotheses.

The key advantage of eliminating benchmark tasks is that it reveals what insights visualization users gained. Researchers can then compare visualizations on insight-related measures such as the number of insights, the categories of insights generated, the speed at which insights were generated, and the domain value of the insights. These measures are closely related to the fundamental characteristics of insight, as identified previously. Researchers can also compare the insights that users gained with the insights that they expected users

to gain. For example, which insights did users fail to discover with the visualization?

Furthermore, since the protocol shares some similarities to that of formative usability studies, researchers can simultaneously collect a wealth of usability data and correlate it to the insight results. For example, which features of the visualization helped achieve insight, and which caused problems for the users? This directly leads to visualization refinement and improvement.

The difficulties with this method include the need for

- potentially long training and trial times depending on data and domain complexity;
- more effort by the experimenters to capture and code results;
- motivated, domain knowledgeable users who will not merely follow instructions but generate insight in a self-directed manner; and
- domain experts to assist in coding results along with visualization experts.

In general, these problems are not fundamental, in that experimenters can overcome them given sufficient resources. At the same time, however, this method provides the resource advantage of relieving the experimenter from having to design benchmark tasks, a surprisingly difficult job. As with the previous method, greater variance in the results is also a problem. Philosophically, with this method we must be willing to live with somewhat more subjective results.

Onward

In practice, both types of controlled experiments are needed. Benchmark task experiments help identify specific low-level effects. Eliminating benchmark tasks provides a much richer (more insightful) view of the broader insight capability of visualizations. If combining both approaches into a single experiment, the benchmark tasks should not precede the open-ended portion. Otherwise, user thinking will become constrained by the benchmark tasks when moving on to the open-ended portion.

Future steps should take a more rigorous and comprehensive approach to comparing these methods, perhaps running the same experiment with each method to see how they differ. Experimenters can also examine and adapt other uncontrolled evaluation methods to better gauge insight. These new methods can provide better measures of visualization insight, and ultimately determine whether visualizations are achieving their grand purpose. ■

References

1. R. Amar, J. Eagan, and J. Stasko, "Low-Level Components of Analytic Activity in Information Visualization," *Proc. IEEE Symp. Information Visualization*, IEEE Press, 2005, pp. 111-117.
2. C. Plaisant, "The Challenge of Information Visualization Evaluation," *Proc. Working Conf. Advanced Visual Interfaces* (AVI), ACM Press, 2004, pp. 109-116.
3. M. Tory and T. Möller, "Human Factors In Visualization Research," *IEEE Trans. Visualization and Computer Graphics*, vol. 10, no. 1, 2004, pp. 72-84.
4. C. Chen, and Y. Yu, "Empirical Studies of Information Visualization: A Meta-Analysis," *Int'l J. Human–Computer Studies*, vol. 53, no. 5, 2000, pp. 851-866.
5. P. Saraiya, C. North, and K. Duca, "An Insight-Based Methodology for Evaluating Bioinformatics Visualizations," *IEEE Trans. Visualizations and Computer Graphics*, vol. 11, no. 4, 2005, pp. 443-456.

Contact Chris North at north@cs.vt.edu.

Contact editor Theresa-Marie Rhyne at tmrhyne@ncsu.edu.