Evaluation: How Much Evaluation is Enough?
IEEE VIS 2013 Panel

Organizer: Robert S Laramee, Swansea University, UK

Panelists:

Min Chen, Oxford University, UK
David Ebert, Purdue University, US
Brian Fisher, Simon Fraser University, Canada
Tamara Munzner, University of British Columbia, Canada

1 INTRODUCTION

Most of us agree that evaluation is a critical aspect of any visualization research paper. There are many different aspects to the topic of evaluation including: performance-based such as evaluating computational speed and memory requirements. Other angles are human-centered like user-studies, and domain expert reviews to name a few examples. In order to demonstrate that a piece of visualization work yields a contribution, it must undergo some type of evaluation.

The peer-review process itself is a type of evaluation. When we referee a research paper, we evaluate whether or not the visualization work being described has been evaluated adequately and appropriately.
In an increasing number of cases papers are rejected based on what was judged, at that time, to contain an inadequate evaluation even though the technical or design contributions are acknowledged.

However, there are differing opinions as to what constitutes an adequate or appropriate evaluation when it comes to visualization. In this panel, we discuss precisely this topic: What constitutes an adequate and appropriate evaluation of a piece of visualization work? We address questions such as:

- What is the most appropriate way to evaluate a visualization?

- To what extent should a piece of visualization work be evaluated prior to its publication?

- When presenting a new visualization approach, what proportion of a paper should be dedicated to the subject of evaluation?

- Does evaluation have to be done as the last phase of a waterfall model for visualization?

- What constitutes a convincing evaluation?

- When is a user-study an appropriate form of evaluation?

- How much responsibility lies with the reviewers to evaluate a piece of visualization work?

Why this panel at VIS 2013?
This inspiring topic touches upon the experience and sentiment of every researcher in visualization. It will form the basis of lively discussions that address these questions and perhaps more from the audience.

2 LOGISTICS

The panelists will present their positions. The introductory remarks will be made by Bob Laramee. His introduction will last for 5 minutes. Each panelist will be given 5-10 minutes, for a total of 25-45 minutes of presentations. This will allow for approximately 35-55 minutes of audience participation in the discussion. All panelists will have the opportunity to offer a summary view at the end of the panel (2 minutes each).

3 POSITION STATEMENTS

Min Chen: Less and More

In the field of visualization, the term "evaluation" may have different meanings in different contexts. In a wide context, it is a *process* for validating a piece of research. In a narrow context, it is often used as a synonym for an *empirical study*, and sometimes more specifically for a *controlled user study*. In the former, evaluation is an indispensable process for any scientific and scholarly work, whereas in the latter, it is just one of the tools in a large toolbox. Borrowing from the terminology of the UK Research Assessment Exercises, the goal of evaluation in a broad sense is to determine the level of *originality*, *rigor*, and *significance* of a piece of research. The methods for evaluation includes, but not limited to:

- Relying on peer judgement through review, publication conference presentation, and citation;

- Establishing dependence on and divergence from creditable prior work such as fundamental theories, experimental results, and to a certain degree, widely accepted models and wisdoms;

- Discovering evidence of sound technical implementation, including mathematical reasoning, algorithmic development, system engineering and deployment, experiment design, execution and analysis, and so on.

- Contemplating significance and impact through speculative discussions (short-term), usability studies (medium term), and comprehensive surveys (long term).

In terms of peer review processes, evaluation is intended as a balanced assessment of *originality*, *rigor*, and *significance* appropriate for a specific publication venue, while the priority, in my view, is reflected by the ordering of these three components. In comparison with disciplines such as computer graphics, human-computer interaction and computer vision, visualization is a relatively small community. Hence we must direct our limited resources carefully, focusing on addressing major challenges, stimulating new innovation, making new discoveries, and facilitating wide applications. In terms of evaluation, the community may benefit from:

- a bit **less** demand for controlled user studies for evaluating individual works with sufficient technical contributions, while channelling **more** energy to significant (often standalone) empirical studies with a clear aim for making new discoveries in various perceptual, cognitive and social aspects of visualization;

- a bit **less** demand for evaluation as the end in the waterfall model for design studies and application studies, while encouraging **more** iterative evaluation in the agile model, i.e., the nested model proposed by Munzner (2009);

- a bit **less** demand for evaluation as a box-ticking section in a paper, while relying **more** on multi-faceted evaluation as outlined above.

- a bit **less** demand for insight-based evaluation, while focusing **more** on tangible and realistic evaluation criteria, such as saving time. See also Chen et al. (arXiv:1305.5670).

I sometimes wonder, with today's expectation for evaluation based on usability studies, whether Shneiderman, Lorensen and Levoy would still be able to publish their seminal works on treemaps, isosurfacing and volume ray casting respectively. Surely we would not wish for a "no" answer and such a situation would not help visualization as a scientific discipline.

David Ebert: Multistep and Multi-level

The visualization community has been emphasizing/requiring evaluation in the review process and the community has responded. Unfortunately, much of the evaluation does not answer the question of whether the contribution is better than previous contributions or whether the system is effective and useful. Evaluation and feedback need to be part of the iterative development process and occur at many levels from efficiency and accuracy to usability up to effectiveness. Simple user studies with students that show statistically significant performance on toy tasks often do not lead to improved solutions in deployed applications. Better solutions are often provided when involving targeted users from the beginning, during iterative refinement and evaluation, through the development process, and into deployment without ever conducting a formal lab user study with statistical significance!

Brian Fisher: Evaluate Reasoning, Not (Just) Visualizations

Bill Lorensen pointed out in 2004 "Visualization has become a commodity". In that sense visualization has succeeded beyond our expectations. Infographics are the medium of choice for informing both decision-makers and the general public about quantitative information. Online news headlines proclaiming "8 infographics that tell you all you need to know about XXX" are increasingly common.

However, it can be argued that "Visualization (research) is (still) not having an impact in applications". I will argue that this is due to an unmet challenge: to make the case that visualization research can reliably and measurably improve the quality of human decisions. This requires evaluation not only of usability and clarity of explanation offered by visualizations but also how interaction with visual representations can predictably facilitate and direct the flow of human reasoning.

I will argue that this need to evaluate real-world impacts will require us to devise new empirical methods for assessing visually-enabled reasoning about information and the creation of policies and plans that operationalize analysis for action. These methods must bridge cognition in the lab and "cognition in the wild" and be specifically targeted to improving the design and use of visual information systems.

Tamara Munzner: Evaluation, When and How

I have struggled with the question of when and how evaluate visualization research in both private and public ways. In private, I ponder on a case-by-case basis what to do for each research project as an author and how to judge for each paper submission as a reviewer. More publicly, we have written a series of four "meta-papers" targeted at other authors and reviewers that argue for and against particular practices.

My thinking and vocabulary has evolved over time. In the first paper on pitfalls [Munzner08], I regret the earlier choice of Evaluation as the overly-broad term for a paper type that would be more accurately and narrowly called Summative User Studies. In a second paper, I present a nested model that exactly addresses the question of how to evaluate different kinds of contributions appropriately at four different levels of design [Munzner09]. In that model, I argue for interpreting "evaluation" in the broadest possible sense, as including everything from controlled laboratory studies to computational benchmarks to pre-design field studies to observe existing practices to post-deployment studies of how new visualization tools change workflows in terms of speed or capabilities. Recently, we extended that model to discuss blocks as the outcomes of the design process at a specific level, and guidelines that discuss relationships between these blocks [Meyer12]. Finally, we have proposed a nine-stage process model for design studies that directly addresses the question of how and when designers should evaluate this kind of work, as well as how reviewers might evaluate the contributions of such papers [Sedlmair12].

The field has also evolved over time: the bar for how much evaluation is enough has risen considerably. I consider this rise to be a positive sign that our field is maturing; we are still very far indeed from the point of ossification! In the early years, papers could simply propose a new technique with essentially no characterization of situations in which it might be useful, because so little of the space of possible designs had been explored. Now that many, many techniques have been proposed, it is very reasonable that there is more emphasis on comparing a new one to those that have come before. Of course, controlled user experiments are certainly not the only way to do so, so it would be ridiculous to have a litmus test that every paper should have one. I argue in the nested model [Munzner09] that they have their place for evaluating decisions at the visual encoding level, but they are particularly unsuitable for validating that system designers have addressed the correct task.

4 BIOGRAPHIES

Min Chen: Min Chen is currently a professor of scientific visualization at Oxford University and a fellow of Pembroke College. Before joining Oxford, he held research and faculty positions at Swansea University. His research interests include visualization, computer graphics and human-computer interaction. His work in visualization covers a range of topics from theories to applications, and from algorithms to empirical studies. His services to the research community include papers co-chair of IEEE Visualization 2007 and 2008, co-chair of Volume Graphics 1999 and 2006, AEIC of IEEE Transactions on Visualization and Computer Graphics, and co-director of Wales Research Institute of Visual Computing. He is a fellow of the British Computer Society, European Computer Graphics Association, and Learned Society of Wales.

David Ebert: David Ebert is the Silicon Valley Professor of Electrical and Computer Engineering at Purdue University, a University Faculty Scholar, a Fellow of the IEEE, and Director of the Visual Analytics for Command Control and Interoperability Center (VACCINE), the Visualization Science team of the Department of Homeland Security's Command Control and Interoperability Center of Excellence. Dr. Ebert performs research in novel visualization techniques, visual analytics, volume rendering, information visualization, perceptually-based visualization, illustrative visualization, mobile graphics and visualization, and procedural abstraction of complex, massive data. Ebert has been very active in the visualization community, teaching courses, presenting papers, co-chairing many conference program committees, serving as Editor in Chief of IEEE Transactions on Visualization and Computer Graphics, and serving on the IEEE Computer Society Board of Governors.

Brian Fisher: Brian Fisher is Associate Professor of Interactive Arts and Technology and Cognitive Science at Simon Fraser University, and a member of the SFU Centre for Interdisciplinary Research in the Mathematical and Computational Sciences. At the University of British Columbia he is the Associate Director of the Media and Graphics Interdisciplinary Centre (MAGIC), Adjunct Professor of Computer Science, and a member of UBC Brain Research Centre and the Institute for Computing, Intelligent and Cognitive Systems. His research focuses on the cognitive science of human interaction with visual information systems, with the goal of developing new theories and methodologies for development and evaluation of technology to support human understanding, decision-making, and coordination of operations. He was a General Chair of VAST 2010 and serves on the VAST steering committee.

Tamara Munzner: Tamara Munzner is a professor at the University of British Columbia Department of Computer Science, where she has been since 2002. She was a research scientist from 2000 to 2002 at the Compaq Systems Research Center in California, earned her PhD from Stanford between 1995 and 2000, and was a technical staff member at the Geometry Center mathematical visualization research group from 1991 to 1995. She was InfoVis Co-Chair in 2003 and 2004, EuroVis Co-Chair in 2009 and 2010, and serves on the steering committees for InfoVis and BioVis. Her research interests include the development, evaluation, and characterization of information visualization systems and techniques from both user-driven and technique-driven perspectives. She has worked on visualization projects in a broad range of application domains, including evolutionary biology, microbiology, topology,
computational linguistics, large-scale system administration, web site design, and web log analysis.

References

[Meyer12] Miriah Meyer, Michael Sedlmair, and Tamara Munzner. The Four-Level Nested Model

Revisited: Blocks and Guidelines. In Proceedings of the VisWeek Workshop Beyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV). ACM Press, 2012.

[Munzner08] Tamara Munzner. Process and Pitfalls in Writing Infovis Research Papers. In A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, editors, Information Visualization: Human-Centered Issues and Perspectives, volume 4950 of LNCS, pages 133-153. Springer-Verlag, 2008.

[Munzner09] Tamara Munzner. A Nested Model for Visualization Design and Validation. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2009), 15(6):921-928, 2009.

[Sedlmair12] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis 2012), 18(12):2431-2440, 2012.