Beyond system logging: human logging for evaluating information visualization

Nathalie Henry Riche
Microsoft Research Redmond
nath@microsoft.com

ABSTRACT
In this position paper, we propose to investigate novel technologies for evaluating information visualization systems: physiological sensing. We review existing technologies and describe how advances in physiological sensing open a novel perspective for the evaluation of information visualization systems.

Author Keywords
Information Visualization, Evaluation, Human-Brain Interfaces, Physiological Sensing

ACM Classification Keywords
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION
Evaluating information visualization (infovis) systems is a difficult challenge [1, 2] for multiple reasons: the exploratory nature of the task and the difficulty to decompose it into low-level and more easily measured actions; the limited availability of expert users; the specificity of their respective data of interest; and the need to look at the same data both from multiple perspectives and over a long time. In these conditions, performing replicable evaluations with generalizable results is very difficult. How can we quantify the value of an infovis system whose role is to support exploration and provide insights to analysts who did not even know the right questions to ask? How can we compare two systems when insights cannot be ignored, nor can there be a limited, controllable and replicable number of them?

In this position paper, we briefly review existing evaluation methodologies and open a novel perspective for infovis to go beyond time and errors while also complementing qualitative methodologies: physiological sensing.

CHI 2009, April 4–9, 2009, Boston, Massachusetts, USA.

CURRENT EVALUATION METHODS
While a complete review of evaluation methodologies can be found in Carpendale [2], we provide here a brief overview of common methods found in the literature.

The simplest and probably cheapest methods to evaluate infovis systems are heuristic evaluations [3], usability studies [4] and their variants. These methods mostly aim at discovering usability issues and improve interfaces; they do not provide any empirical evidence of the system's effectiveness at supporting data exploration. More quantitative studies are commonly used to compare task performance between techniques. Controlled experiments for target acquisition [5] are probably the most used methodology in our field. However, to produce statistically significant results, these experiments necessitate controlled conditions on low-level tasks. This constraint is particularly challenging for infovis because of the difficulty to decompose a complex high-level task (e.g. find insight) into low-level ones (e.g. find a pattern in a graph), raising doubts on the ecological validity of the results.

As a result, our field gradually seeks to conduct more qualitative evaluations, performed in realistic contexts. A good example is Shneiderman et al.'s MILCS [6], inspired from ethnography and which involve a close collaboration between the infovis system designers and a small subset of experts analyzing their own data, in situ, over a long period of time. Another example is the application of grounded evaluation proposed by Isenberg et al. [7] to ensure the validity and realism of our evaluations. A very large body of qualitative methods exists and may provide us with unique and very rich insights about the use of infovis systems. However, these methods require strong time commitment for collecting and analyzing data, observers may introduce bias, and replicating and generalizing findings is often difficult if not simply impossible.

Additionally, Saraiya et al. [8] introduced the concept of insight-based evaluation. They define insights as *"an individual observation about the data by the participant, a unit of discovery*" and collect these units of discovery during the evaluation. This approach, further refined by North [9], is very promising in its attempt to capture the exploratory nature of infovis. Inspired by this concept and advocating for the use of qualitative evaluations in realistic contexts, we investigate how recent advances in physiological sensing could complement this approach.

HUMAN LOGGING FOR INFOVIS EVALUATION

In this position paper, we provide a brief overview of physiological sensing and discuss their potential use for evaluation in our field. A detailed literature review on physiological sensing can be found in Mandryk's dissertation [10].

Physiological sensing

A large variety of sensors exists to collect physiological data. They can be placed on the body such as electrodes on the skin (or more invasive technologies, placed inside the body) or close to it, such as eye trackers or cameras. These sensors may provide a solution to collect more quantitative and objective data. However, before describing relevant research in physiological sensing and how we could apply it to our field, we would like to make our readers aware that this field of research is still very recent and suffers from a large number of validity issues.

Researchers do not have a detailed understanding of physiological reactions and the factors that causes them. Signals collected by sensors are often very noisy and difficult to analyze. Thus, researchers mainly use these signals to detect variations of activity, and more particularly arousal, "*a state of heightened physiological activity*" [10]. Arousal may be caused by positive emotions (excitement, surprise or pleasure) or negative ones (stress, mental effort). It is particularly difficult to detect subtle changes of activity. Even in the case of brain activity, for which brain's functions are well mapped and sensors able to locate the source of activity (differentiating for example visual processing from sense making activity), it still requires lengthy validation in multiple context to associate a given physiological signal to a particular task or stimuli. While it seems that today physiological technologies raise more challenges than provide answers, we believe this research may bring a new light on infovis evaluation.

Eye Tracking and Pupile Dilation

Eye tracking is probably the most commonly used technique in visualization. Colin Ware [11] reflected on the importance of eye movements to understand our cognitive processes and understand visual thinking. This technique is also used in graph drawing to understand how people read graphs and what factors impact eye movements and readability performances [12]. When using an eye tracker, researchers collect eye gaze as well as additional information such as pupil dilation, blinks, saccades and eye fixations.

The analysis of eye fixations (i.e. the time spent in given regions of the screen) or the number of gaze entries in a particular area may indicate regions of interests and saliency of visual elements. Eye tracking analysis could potentially reveal visual elements that are not used in the visualization. In the ideal case, detecting eye patterns could help us investigate the strategies expert use to analyze visualizations. Research has been heading in this direction in the medical field, where medical students needs to learn how to read medical images such as X-rays from experts [13]. Successful results could help researchers teach how to read visualization and ensure participants reach a high level of expertise before any evaluation.

Heart rate, respiration and skin conductance

Variations in heat rate and blood pressure have been extensively used to detect highly stressful situations and high mental effort [14]. Mental effort and high cognitive activity are also associated with increases in respirations frequency [15]. Collecting respiratory patterns may help identify moments of high stress or surprise as these may provoke momentary cessation of respiration and irregular respiration patterns. Another solution to detect arousal is galvanic skin response (GSR). GSR sensors can measure changes in the skin conductance provoked by changes in moisture (sweat) in specific areas of the body. This sensor is used in lie detector technologies and may be one of the most commonly used sensors because of its direct relation to arousal and cognitive activity [16].

Muscle activity

Electromyography measures the muscles' activity by detecting surface voltage when muscles are contracted. When used on the face, this sensor is particularly effective to detect frustration (muscles responsible of frowning expressions on the brow) and pleasure (muscles responsible of smiling expressions on the cheeks)[17]. The measure is very sensitive; corresponding face expressions may not be visible on the subject's face. We believe that quantifying pleasure is likely to be a valuable measure for information visualization. Indeed, it could help researchers quantify aspects such as aesthetics or frustration. Aesthetical representations have been the topic of a number of works in graph drawing [18] and is commonly discussed in our field. However, the aesthetic criterion is very often overlooked in infovis because of the perceived subjectivity of its measure.

Brain Measurements

Many techniques exist to measure the brain activity, with various levels of invasiveness and precision. We only focus here on the less invasive and most practical techniques. Electroencephalograms (EEG) detect electrical activity by placing electrodes on the head. It has been proven effective to detect cognitive load when the task is well defined [19]. Recent research successfully detected the "aha moment" using EEG in specific conditions [20]. Functional near infrared (fNIR) are another type of sensors which detect the brain activity by measuring infrared light reflectivity sent in the brain. This method is particularly effective to locate where the activity occurred in the brain and is relatively non-invasive. Recent work [21] presents practical guidelines for using this technology in HCI research. From our discussions with researchers in the field, analyzing brain measurements seems to be the most challenging part of the work. It requires extensive statistical analysis or strong machine learning algorithms.

Ingenious methodologies have also been developed to assess how our brain works without actually measuring it. A good example is the implicit association test [22], measuring the response time to sort items in two sets of categories to assess if users perceive these two categories as correlated (see demo at http://implicit.harvard.edu). Such lightweight tests could provide elegant and very fast solutions to evaluate readability or aesthetics for example.

DETECTING INSIGHT: THE HOLY GRAIL?

In this position paper, we gave a brief overview of physiological sensing technologies. A number of these technologies were proven helpful to measure arousal and cognitive load in specific conditions. Measuring cognitive load required is certainly valuable for evaluating infovis systems. However, our challenge is to detect if this cognitive load is positive or negative. Indeed, when comparing two visualizations with similar task performances, a high cognitive load may either indicate a more insightful representation warranting further analysis, or a more difficult representation to interpret.

As information visualization researchers, our ultimate goal is to provide visualizations that help discover more insights about the data. If physiological sensing could help us detect when users make these discoveries, we believe this would greatly ease the evaluation of our systems. While we have mentioned a number of research pieces focusing on the discovery of the "aha moment", this research is still at an early stage and only detects strong and specific changes in physiological activity. The rich and still ill-defined concept of insights in infovis is likely to be more subtle and far more difficult to detect. However, we believe this research area opens new perspectives on the evaluation and comprehension of information visualization. We look forward to brainstorm and discuss how physiological technologies could complement our current practice.

ACKNOWLEDGMENTS

We thank Desney Tan, Ed Cutrell and Scott Counts for their enthusiasm and insights on physiological computing.

REFERENCES

1. C. Plaisant, The challenge of information visualization evaluation. In Proc. of the ACM Advanced Visual Interfaces, pp 109–116, Gallipoli, Italy, 2004.
2. S. Carpendale. Evaluating Information Visualizations. In Information Visualization: Human-Centered Issues and Perspectives, pages 19-45. Springer. 2008.
3. Nielsen, J., and Molich, R. (1990). Heuristic evaluation of user interfaces, Proc. ACM CHI'90, 249-256.
4. R.L. Mack, J. Nielsen, Executive summary. In J. Nielsen (Ed.), Usability inspection methods (pp. 1-23). New York: John Wiley and Sons, Inc. 1994.
5. P.M. Fitts, The information capacity of the human motor system in controlling amplitude of movement. Journal of Experimental Psychology, 47, 381-391, 1954
6. B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: MILCS. In Proc. of the AVI workshop BELIV'06, pp 1–7, 2006.
7. P. Isenberg, T. Zuk, C. Collins and S. Carpendale. Grounded Evaluation of Information Visualizations. In CHI workshop BELIV'08, ACM Press, pp 56-63, 2008.
8. P. Saraiya, C. North, and K. Duca, An Insight-Based Methodology for Evaluating Bioinformatics Visualizations, IEEE TVCG, 11 (4), 2005, pp. 443-456.
9. C. North. Toward measuring visualization insight. IEEE CGA, 26(3): 6–9, 2006.
10. R.L. Mandryk (2005). Modeling User Emotion in Interactive Play Environments: A Fuzzy Physiological Approach. Ph.D.dissertation, Simon Fraser University.
11. C. Ware, IEEE visualization keynote, 2009.
12. Huang, W., Hong, S.-H., Eades, P., Using eye tracking to investigate graph layout effects, Proc. APVIS 2007, IEEE Computer Society. pp. 97-100, 2007.
13. New eye-tracking device. 2004. http://www.news-medical.net/news/2004/06/21/2650.aspx
14. D.W. Rowe, J. Sibert, D. Irwin. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In Proc. SIGCHI, pp. 480-487, 2008.
15. C.J.E. Wientjes, Respiration in psychophysiology: Methods and applications. Biological Psychology, 34, 179-203, 1992.
16. W. Boucsein, Electrodermal activity. New York: Plenum Press, 1992.
17. J.T. Cacioppo, G.G. Berntson, J.T. Larsen, K.M. Poehlmann, T.A. Ito, The psychophysiology of emotion. In J. M. Haviland-Jones (Ed.), Handbook of emotions. New York: The Guilford Press. 2000.
18. H.C. Purchase, Which aesthetic has the greatest effect on human understanding?, Proc. of Graph Drawing Symposium, LNCS 1353, Springer-Verlag, 1997.
19. D. Grimes, D. Tan, S. Hudso, P. Shenoy,R. Rao. Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. Proc. Of SIGCHI, pp. 835-844, 2008.
20. E.T. Solovey, A. Girouard, K. Chauncey, L.M. Hirshfield, A. Sassaroli, F. Zheng, S. Fantini, R.J.K. Jacob. Using fNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines. ACM UIST'09.
21. S. Sandkühler, J. Bhattacharya. Deconstructing Insight: EEG Correlates of Insightful Problem Solving. PLoS ONE 3(1): e1459, 2008.
22. A.G. Greenwald, D.E. McGhee, J.L.K. Schwarz, Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464-1480, 1998