Effects of Visual Representations on Evaluating Uncertain Data

Nam Wook Kim
Stanford University
353 Serra Mall
Stanford, CA 94305
namwkim@cs.stanford.edu

ABSTRACT
Visualizations have been valuable tools in understanding data and discovering patterns in the data. Most visualization research to date has been focused on developing optimal visual encoding to help people better interpret data. However, previous research has been limited to understanding how visualization supports data analysis. As a result, little attention has been given to how visualization affects how people collect data. As an initial step toward studying effects of visualizations on data collection behaviors, this paper investigates how the saliency of visual representations affects on evaluation of uncertain data.

ACM Classification: H5.2 [Information Visualization]: User Interfaces. - Graphical user interfaces.

Keywords: Visual Salience, Uncertainty, Representation, Evaluation

INTRODUCTION
With advances in information technologies, our society is experiencing a vast amount of data. To compensate for limited human perceptual and cognitive abilities, visualizations have been valuable tools for understanding data and discovering patterns in the data. Most visualization research to date has been focused on finding optimal visual encoding to help people interpret data. Recently, a few research has recognized the role of visualization as a communication medium and studied the use of visualizations in social context.

However, previous research are limited to understanding how visualization affects how people analyze data. Considering the process of sensemaking in which data is collected, organized and analyzed [7], the value of visualization has mostly been recognized as an output device (i.e. analytic tool). Although many existing systems integrated data collection with visualization [many eyes, swivel, wikimapia, geni.com], it is still an unanswered research question how visualization influences how people gather data and refine the data.

Heer [5] observed in sense.us that people tend to discuss about unclear meaning or anomalies in data collection. This gives rise to a question about what would have happened if they are able to directly modify the underlying data. In the discussion with Heer and Viegas [6], Wattenberg also envisioned that the future of social data analysis tools is expected to see people not only engaging in data analysis but also contributing data. In the same sense, there remains interesting

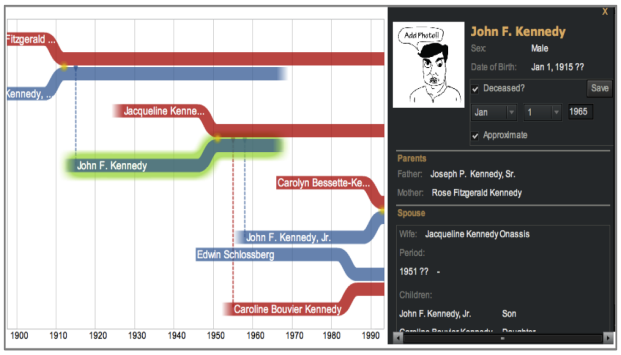

Figure 1: visual interface

questions as to how visualization play a role in data collection activity, more specifically how visualization affects the quantity and quality of entered data, or more generally how visualization can effectively support the full cycle of the sensemaking process.

As an initial step toward studying effects of visualization on information collection, this paper investigates how the saliency of visual representations affects people evaluating and correcting uncertain data.

RELATED WORK
It is the data that drives visualization. In Many Eyes [4] and Swivel [3], users contribute their own data and then use predefined visualizations to map the data into graphical forms. However, uploading the data and creating visualizations are separate processes. On the other hand, Wikimapia [2] and Geni.com [1] provide direct data manipulation through visualization. Such visualizations have proven to be easy-to-use data input interfaces. However, it is still unclear what makes visualization a good data input device.

DESIGN
We use genealogy as our study platform. It is an attractive domain for studying collaborative sensemaking practices: it engages millions in a social process of foraging for data, *evaluating uncertain data sources*, and analyzing the data. As a part of the study, We have built a prototype social genealogy web site where people can build their family trees. The web site uses a timeline-based visualization as a data input interface (Figure 1). Individuals are represented using timelines that converge and diverge to indicate marriage and

Figure 2: uncertainty representations for a birth date (a) no marker (b) subtle gradient (c) salient question mark

divorce. Parent-child relationships are depicted using a directional edge from parents to childs birth. A user can enter new data by clicking a persons lifeline, and adding new relationships to the person or editing the persons information through the data input panel.

Three uncertainty representations are used as study stmuli: no marker, subtle gradient, and salient question mark (Figure 2, 3). In this study, we only consider uncertainties of temporal attributes including birth, death, marriage and divorce dates. Each subject is randomly allocated to one of three conditions. The subject is then given initial family data and asked to add a specific list of people to the data. During the study session, we observe how the subject reacts on the assigned uncertainty representation without explicitly asking to correct uncertain data. Considering its popularity, the Kennedy family is used as the base data. It is modified behind the scene to include uncertain attributes; for example, Jonh F. Kennedys death date, 1963, is altered to 1965. The study was conducted on Amazons MTurk [**?**] because of easy access to potential subjects and fast responses.

RESULTS

Unfortunately, we found no evidence for our hypothesis. However, even if the web site is still in an early stage of development, we are glad that almost all the participants quickly became used to the visual interface with a very brief introduction and first of all enjoyed using it. We share some of user comments in Table 1 .

DISCUSSION AND FUTURE WORK

The study results did not provide any interesting insight. We suspect two contributing factors for the failure. First, it is possible that our choice of study subjects was wrong. We have observed that most participants on Mechanical Turk tend to finish the given task as fast as they can to get paid, but they were not at all willing to do more than needed. However, in our study, it is expected that the participants not only add the list of people, but also show some reactions to uncertainty representations. Second, in worst case scenario, our study design might be wrong; that is, people do not really care about such variations of an uncertainty representation.

Based on lessons learned from the initial study, it would be reasonable to conduct the study on the genealogy community. This will provide further evidence whether the study design is wrong or not. In addition, considering the study is designed around the specific domain of genealogy, people from the genealogy community would better serve as study subjects. Another way of running the study in a desirable way is allocating a specific amount of time for each session so that each subject can have enough time to play with the visual interface, instead of leaving right after adding the required people.

Figure 3: uncertainty representations for a marriage date (a) no marker (b) subtle gradient (c) salient question mark

I really enjoyed doing so as this hit was really interesting.
It was a very pleasant and definitely a learning experience.
Thanks a lot requestor for having such a task. Really appreciate it.
This work is so interesting! I got so much fun. Thank you!
I like the visual interface, it was quick to load which is a great plus.
That was a fun and happy to learn the legacy of those great ones
I would love to do more of such task. It was really interesting.
Definitely fun! Hope you have more of these.

Table 1: User Comments.

Since data analysis and data collection are tightly coupled in the sensemaking practice, we believe that there is an avenue of investigation for the visualizations role in data collection. In this study, we varied a specific visual element and tried to see if it spurs different behaviors on evaluating uncertain attributes. Once we confirm the hypothesis, an obvious next step would be how different types of visualization affect data collection and refinement behaviors. Even larger research agenda would be to understand the relation between visualization and data collection in the context of collaborative sensemaking.

REFERENCES

1. Geni.com, June 2010. http://www.geni.com.
2. Wikimapia, June 2010. http://wikimapia.org.
3. Swivel, June 2010. http://www.swivel.com
4. Amazon Mechanical Turk, June 2010. http://www.mturk.com
5. F. B. Vigas, M. Wattenberg, F. van Ham, Jesse Kriss, Matt McKeon. Many Eyes: A Site for Visualization at Internet Scale, Infovis, 2007.
6. J. Heer, F. B. Vigas, M. Wattenberg. Voyagers and Voyeurs: Supporting Asynchronous Collaborative Information Visualization, ACM Human Factors in Computing Systems (CHI), 2007
7. J. Heer. A Conversation with Jeff Heer, Martin Wattenberg, and Fernanda Vigas, ACM Queue. March 2010
8. S.K. Card, J.D. Mackinlay, and B. Shneiderman. Readings in Information Visualization: Using Vision To Think. Morgan- Kaufmann, San Francisco. (1999)