Notes for Class 12: Evaluation

Objectives
By the end of the class, you will be able to…
- Critique experimental design and research papers involving empirical studies.
- Design a strategy for evaluating visualization technique(s) or tool(s) for a given purpose.

Agenda for Class
- Discussion of paper.
- General discussion of evaluation approaches.
- (if time) Evaluation activities

Reading Assignment
- Empirical Comparison of Dynamic Sliders and Brushing Histograms, Qing Li and Chris North, InfoVis 2003, pp. 147-153.
- Ware Appendix
- Background notes on evaluation

Notes on the Empirical Comparison Paper:

What was the objective of this paper?
To empirically compare dynamic query slider interaction with brushing histogram interaction. Both of these enable you to interactively/dynamically query multidimensional data, but there was little known at the time about which one was better. The authors want to know which is best for different tasks and why.

More specifically, they want to compare these 2 techniques because they represent rather opposite approaches to interactive queries (filter vs. highlight, single range query vs. multiple range, etc as specified on page 3).

Why is it important to display the distribution of data values? What are the different ways in which this can be done?
Displaying the data distribution makes it easier for users to conduct dynamic query operations to find the data that is important to them. It does this by giving them an idea of how much data lies in a given data range. This is particularly important when the data is not normally distributed.

E.g. In the HomeFinder dynamic query application, there was no distribution data provided for any of the dimensions. So, I might choose to look for a 4 bedroom house within 5 miles of my workplace and get few results. It may be that there are many more 3 bedroom houses than 4 bedroom houses but I have no way of knowing this without trial-and-error. Showing the data distribution makes this information apparent, so that I will know ahead of time that I will get many more results if I change the bedroom level to 3.
Methods to display distributions:

- **Histogram.** Example:

  ![Histogram Example](http://www.icbl.hw.ac.uk/tdi/images/cookbook/histogram.gif)

- Bargram (Fig 4). Essentially takes all the bars in the histogram and lays them end-to-end in order from lowest to highest.

- **Dot plot** (Fig 5). Also the lines above bars in the bargram (Fig 4) are a variation on the dot plot idea.

  ![Dot Plot Example](http://www.amstat.org/education/gaise/GAISEPreK-12_files)

- **Box & whisker plot** (Fig. 5). This does not give all the data points but gives an aggregated view. It shows the median value (center line in the box), 1st and 3rd quartiles (ends of the box), and total range of the data (lines from the ends of the box, also called “whiskers”). Example:

  ![Box and Whisker Example](http://www.rhombus.be/)
• Colour bar. You could also indicate the distribution by using a coloured bar, where brightness or saturation of the colour represented the number of items with that value (like a histogram but using colour instead of height). Example:

![Colour bar example]

**Critique of the experimental design**

Positive aspects of the design:
- Carefully controlled study. The experiment appears to be technically sound in terms of its design.
- It is very nice that they used the same prototype system so that only the interface varied.
- It is also nice that they were able to use think-aloud and question-asking to help them understand the cause of participants’ confusion.
- Quite a large number of participants.
- The study does seem to address the goals identified in the paper, to a limited extent.

Questions, suggestions, and comments:
- The authors point out that they gave participants “hints” about how to use the tools. It isn’t clear exactly how often this happened or what types of hints were provided. Perhaps it might have been better to do more extensive training at the beginning of the experiment.
- Users: the study involved “technical undergraduate students” which is understandable from a practical point of view. However, based on some of the observations described, some (many?) of these people seemed rather unfamiliar with statistics concepts (e.g. histograms). Many real users investigating complex trends and relationships would have more statistical knowledge, so it makes sense to conduct another study with a more experienced group.
- Tasks: How were the tasks for this study selected? Are there other tasks that are also important? How well do these tasks generalize? Some better justification for these choices would have made a more convincing argument that these tasks are representative of real queries. (That said, these tasks do seem reasonable.) Also, how much did the actual questions change? (Only one example of each is given.)
- Some details of the design were not reported in enough detail. It would be difficult to accurately reproduce this study based on the information in the paper. E.g. What order were the tasks done in? Was this varied? What was the display size and resolution?

**Key findings of the experiment**

Dynamic query sliders were faster for the simpler tasks: single range, multiple range, and multiple criteria. The authors believe this is because the interaction is simpler and more efficient with the sliders compared to the brushing histograms.
Brushing histograms were faster for the more complex tasks: attribute relation, compare, and trend evaluation. The authors believe this is in part because the histograms themselves act as an additional visualization, providing a different view of the data that may be useful for these types of tasks. In some cases, using the histograms alone more directly answers the question. E.g. to see if two attributes are correlated, brush different sections of one histogram and see if corresponding sections of the other histogram are highlighted (without even looking at the map).

Brushing histograms were harder to learn to use than dynamic query sliders. They appear to be powerful for experienced users doing complex tasks, but difficult for novices.

**Key Concepts for Today:**
1. The paper we discussed today uses a fairly standard experimental protocol. This protocol can be adapted to compare many different types of visualization techniques.
2. Histograms or other visualizations of data distribution can provide useful data to help people effectively use dynamic queries or brushing.