How Does My Audience Read My Visualization?

Steve Rubin
UC Berkeley, Computer Science Division
CS 294-10 – Visualization Class Project
srubin@cs.berkeley.edu

ABSTRACT

Abstract abstract abstract abstract.

Author Keywords

Visualization understanding; design

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Research in information visualization and graphical perception has often focused on creating effective visualizations for low-level perception, such as "what is the value represented by this bar in the chart," and "how much bigger is this bar than that bar." This work is critical in gaining an understanding of which visual variables and encodings are most easy to understand.

A parallel line of work has looked at how higher level attributes affect the overall memorability of charts and graphs. This work aimed to answer questions like, "does chart junk make my graph easier to remember," and "what kinds of charts and graphs are most memorable."

While these two lines of work are posing interesting questions, they neglect one of the most important goals of a visualization, which is to convey a trend or a message to the viewer. The work on graphical perception can tell us how to create the visualization so it is legible, and the work on memorability can tell us how to spruce it up so it sticks in the viewer's memory, but we are interested in learning more about the how trends in a visualization strike a viewer as important or unimportant. The ultimate vision of this line of research is to take a designer's visualization, analyze it, and tell him several key facts about the visualization, such as:

- *What trends do viewers think are most important?*
- *How variable is the spread of trends that viewers think are important?*
- *How well do the viewers' thoughts about this chart match the designer's intention?*
- *What trends will the viewers retain after the visualization is taken away?*

In our methodology and system, a designer first creates a visualization. This visualization is then posted to Amazon's Mechanical Turk[1] where workers submit what they think is the most important trend displayed in the visualization. Simultaneously, the designer performs a nugget analysis on the visualization, which amounts to them writing down the smallest coherent trends that are found in the chart. The designer and his colleagues assign ratings to the nuggets to get a score for each nugget, ranging from 0 (true but unimportant) to 1 (vital importance). The designer then reads the responses from Mechanical Turk, recording the nuggets that are contained in each statement. Our system then computes scores and aggregate statistics for the statements.

Ideally our system could performing the nugget assignment and analysis in a fully automated way, but for this project we use a large amount of manual identification and classification in the pipeline in order to illustrate the potential benefits of such a system. The final output of this system is a dashboard that shows some of the above key facts to the designer of the visualization.

RELATED WORK

This work touches on several areas of prior work from information visualization as well as other subfields of computer science.

Graphical Perception

As the field of information visualization has matured over the years, a large amount of research has studied how people view charts and graphs at a perceptual level. In the foundational work on graphical perception, Cleveland and McGill [4, 5] performed experiments whose results found that some graphical encodings were significantly easier for people to read and understand than others. For example, they showed that estimating angles, as in pie charts, is more error-prone than estimating positions, as in bar or dot charts. Such guidelines are one type of deliverable in graphical perception research. The other type are tools that automate visualization creation given the guidelines we have learned from experiments. While

Woop woop woop!

[1] `http://mturk.com`

Cleveland's work gave generalized design principles for creating charts and Mackinlay provided an automated presentation tool [9], our work focuses on the perceived importance of trends in visualization rather than the viewer's ability to accurately read the data.

Memory

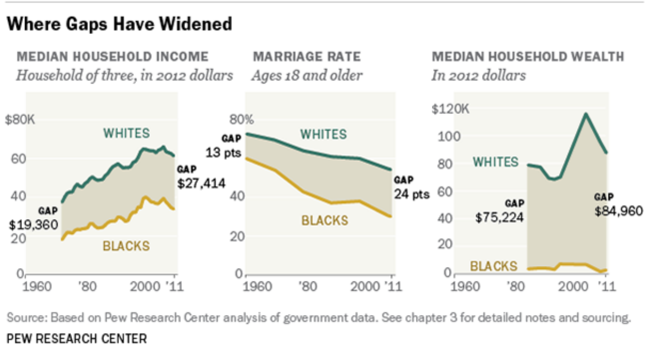
Recent work in the computer vision community has explored the question of "What makes a picture memorable?" [7]. This work posed an image memory task to workers on Amazon's Mechanical Turk and then learned predictors from image features to memorability. These tasks were focused on the overall memorability of images, and not on specific details of images. Later work took this methodology and extended it to information visualizations [3]. Because the methodology mirrored that of the earlier work on image memorability, its results focused on overall memorability and not on the memorability of specific trends in the visualization. Ultimately, when a designer is creating a visualization, he likely wants to communicate certain trends and ideas; getting the viewer to remember exactly what the visualization looked like is a secondary concern. The methodology used the measure the memorability of visualizations [3] shows, as should be expected, that charts with images (chart junk) and charts with atypical structures (i.e., not a bar chart or a line chart, etc.) are more memorable. In our work, we hope to uncover what viewers take away from the visualization rather than just whether the chart made a visual impression.

Chart Junk

The work on visualization memorability [3] showed that charts containing images tend to be more memorable. Prior work sought to quantify the ability of chart junk and visual embellishments to aid in the memorability and understanding of a chart [2]. This work suggested that, contrary to advice given by thought-leaders in information visualization [10], charts with visual embellishments gave viewers no worse accuracy in reading the data and trend, and were easier to recall later on. Several sources have pointed out numerous methodological issues with this paper [6, 1], so the question of the utility of chart junk remains open to further research. While our work in its current state does not specifically address the chart junk issue, our methods and system could ultimately be used to study the effect of visual embellishments on chart understanding.

Crowdsourcing Data Analysis

Our pipeline for learning what viewers understand about visualizations relies, obviously, on the impression of viewers. This kind of data could be collected in a lab study, but for the sake of future scalability, we decided to deploy our visualizations to Mechanical Turk. In running these tasks on Turk, we run the risk of getting noisy data. However, we avoid this by using only workers in the United States with at least a 95% "approved" rating. Prior work in crowdsourcing data analysis [11] describes additional strategies we can employ to ensure high quality responses.



Source: Based on Pew Research Center analysis of government data. See chapter 3 for detailed notes and sourcing.
PEW RESEARCH CENTER

Figure 1. A figure on racial gaps in the past 50 years, from Pew Research. We will use this visualization for our running example.

Summarization evaluation

Our analysis pipeline requires that we have way of taking a statement from a user and somehow quantifying its "goodness." This is an inherently subjective problem, so we look to prior work in summarization evaluation which shows a method for obtaining subjective "goodness" ratings for different facts about a subject by collecting the ratings from multiple trusted colleagues and combining them into final ratings [8]. This approach inspired our nugget rating and assignment approach that we use in determining the worth of each statement.

METHODS

We want to discover what trends viewers think are important about a visualization, the variability of the spread of trends that viewers think are important, and how well the viewers' thoughts about the chart reflect the designer's intention. To explain our methodology, suppose that the designer has designed the chart that shows where gaps have widened between whites and blacks in the past fifty years (Figure 1).

Our data collection and analysis pipeline contains steps that are executed by the crowd, by the designer and his colleagues, and automatically by our system. The full pipeline is shown in Figure 2, and the following subsections will walk through each phase of this pipeline.

Design

The first stage of the pipeline is for the designer to create the visualization. For this example, the designer creates the graph shown in Figure 1.

Crowdsourcing

Once the designer has created the visualization, he pushes it to Mechanical Turk. On Mechanical Turk, workers see a very simple task that first walks them through a thorough set of instructions and examples, and then asks them to write what they think is the single most important trend of the visualization that they are viewing. Figure 3 shows an example of our survey instrument.

We collect twenty to thirty statements for the visualization, but the designer can collect as many or as few responses as he wants depending on the complexity of the visualization,

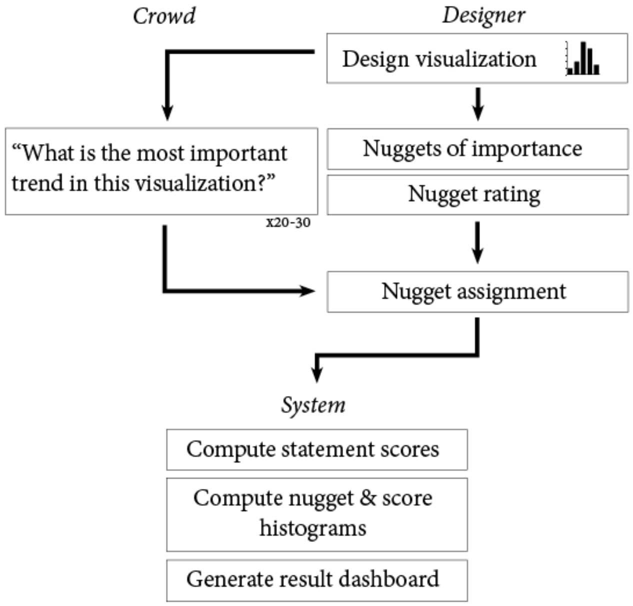


Figure 2. Our data collection and analysis pipeline.

how confident the designer wants to be of the analysis, and how much he wants to spend on crowd workers.

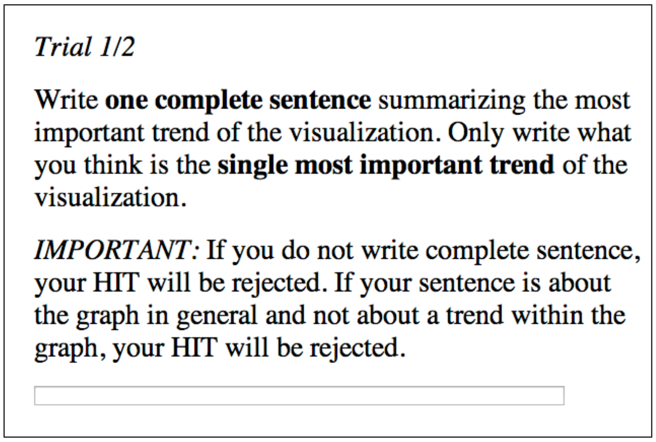
We have built an extensive system for managing tasks on Mechanical Turk that provides the designer with the following functionality:

- **Multiple visualizations can be posted for analysis at once.** This allows the designer to offer a higher price for a batch of visualizations rather than a lower price for a single visualization. While this may seem trivial, workers on Mechanical Turk (particularly the high quality 95% acceptance, US workers that we want) appear to be much quicker to accept a HIT offering fifty cents than one offering ten cents. This improves the rate in which the designer can collect data.

- **Workers will never see a visualization more than once.** This is crucial for maintaining the integrity of the collected data.

- **Multiple versions of a single visualization can be analyzed at once.** The system guarantees that no worker will see more than one version of the visualization. Again, this is crucial for maintaining integrity of the data. Our system can also accommodate the case where the designer *wants* one person to see multiple versions of a visualization.

- **HITs can be re-posted without any negative consequence.** The state logic that dispatches tasks to workers is controlled by our server rather than on Mechanical Turk itself, so canceling and re-posting a HIT will not lose any state, but it will have the positive effect of bringing the HIT back to the top of the task list. This increases the visibility of the task, which improves the rate in which the designer can collect data.



Trial 1/2

Write **one complete sentence** summarizing the most important trend of the visualization. Only write what you think is the **single most important trend** of the visualization.

IMPORTANT: If you do not write complete sentence, your HIT will be rejected. If your sentence is about the graph in general and not about a trend within the graph, your HIT will be rejected.

Figure 3. An example of the survey instrument that is posted to Mechanical Turk. ***Note: the visualization itself is not shown in this figure to save space. In the real instrument, the visualization would appear above the text shown here.***

1. The gap in median household income, marriage rate, and median household wealth between whites and blacks have all grown.
2. There is still a large financial gap between whites and blacks
3. Trends are showing that whites income and wealth are increasing while black income and wealth remains relatively unchanged.
4. The median household wealth for whites spiked much higher than blacks in the middle of the last decade.
5. Between 1960 and 2012 the biggest gap between blacks and white has been shown in their median household wealth.

Figure 4. A sample of 5 responses ($n = 30$) from crowd workers to the prompt, "Write one complete sentence summarizing the most important trend of the visualization."

- **New studies, instructions, and prompts can easily be added to the task management system.** While this is not particularly important in light of the pipeline presented in this work, it was critical in rapidly iterating on the study in order to get high quality feedback from crowd workers.

Our task management system can be used for any kind of study on Mechanical Turk, not just the tasks described in this pipeline.

For our visualization (Figure 1), Figure 4 shows 5 of the 30 responses we collected from crowd workers. Notice that they have varying levels of "goodness;" for example, statement 5 does not make sense because there is no notion of what the "biggest" gap means because the three sub-charts of Figure 1 have incomparable units and meanings.

Nugget Analysis

In parallel with the designer collecting data on Mechanical Turk, he performs his own analysis on the visualization. The

1. Median household wealth for blacks has been relatively constant since 1960.
2. Median household income has increased for whites and blacks since 1960.
3. The gap in marriage rate between whites and blacks has grown between 1960 and 2011.
4. Median household income spiked for whites in the 90's and early 2000's but has since dropped dramatically.
5. Marriage rate has decreased for whites and blacks since 1960.
6. Median household wealth isn't much different for whites or blacks than it was in 1960.
7. The gap in household wealth between whites and blacks has grown between 1960 and 2011.
8. Marriage rates have steadily decreased since 1960.
9. The gap in median household income between whites and blacks has grown between 1960 and 2011.
10. The gap between whites and blacks in median household wealth peaked in the mid 2000's.
11. There exists a financial gap between blacks and whites.

Figure 5. Nuggets created by the designer for the chart in Figure 1.

designer must create a list of "nuggets," which are small, ideally mutually exclusive trends about the visualization. This is a manual process, but it should not be difficult for the designer because we assume that he had a good idea about *why* he made the visualization in the first place and knows what trends the visualization does and does not show. Figure 5 shows the list of nuggets that the designer could create for our visualization. Note that the list of nuggets does not necessarily need to contain every possible trend or fact about the visualization; if the designer thinks a particular trend has no importance, he can leave it off the list of nuggets.

This list of nuggets represents a set of potentially important statements that can be made about the visualization. However, some of these nuggets are likely more important than others. If the designer has a strong sense of the relative importance of these nuggets, he should assign each one a weight between 0.1 and 1.0 (here, 0 is reserved for completely unimportant or wrong statements, so none of the nuggets should have a weight of 0; more on this later).

In many cases, the designer may not be able to come up with these weights on his own. There exist multiple reasons why this might happen. For example, the designer might have a sense of what the few most important points are, but may not know how to rank the other statements. In this case, we propose a method of creating weights for nuggets that draws on an idea from the text retrieval (TREC) community [8]. The designer selects $k-1$ colleagues that he trusts will be able to properly understand the visualization. Then the designer and his colleagues independently assign a rating of either 0 or 1 to each nugget. Here, '1' means that the statement is vitally important, whereas '0' means that the statement is correct, but not as important. These ratings are given in the context of the visualization; that is, the raters are rating the nuggets on how important the nuggets appear to be in the given visualization. For example, nugget 11, "There exists a financial gap between blacks and whites" may be a very important trend in the global sense, but in the visualization, the fact that certain gaps between blacks and whites are widening seems to be more important.

If we start with n nuggets, this procedure gives us a matrix $n \times k$ matrix N that contains 0/1 entries, with each column representing a single person's ratings. To compute the weight w_i of nugget i, we take

$$w_i = \left(\frac{\sum_{j=1}^{k} N_{ij}}{\max_{s \in \{1,2,\ldots,n\}} \sum_{j=1}^{k} N_{sj}} \right) .9 + .1$$

This guarantees that the minimum weight is at least .1, and the maximum weight is 1. The maximum weights needs to be 1 because later on, we want to consider the best possible response from a crowd worker to have score 1. Likewise, we want the worst possible response that still contains a nugget to be .1. We want completely unimportant and false statements to have a score of 0.

So, once our system has computed the weight vector $\mathbf{w} = [w_1, w_2, \ldots, w_n]$ our system can proceed in assigning scores to the crowdsourced statements. First, the designer must analyze each statement r and create an indicator vector $\delta^{\mathbf{r}}$ where

$$\delta_i^r = \begin{cases} 1 & : \text{nugget } i \text{ appears in statement } r \\ 0 & : \text{nugget } i \text{ is not in statement } r \end{cases}$$

The score s_r for statement r is the computed as the weighted average of the included nuggets:

$$s_r = \frac{\mathbf{w}^T \delta^{\mathbf{r}}}{\|\delta^{\mathbf{r}}\|_1}.$$

Once again, notice that the best statements will have a score of 1, false or completely unimportant statements will have a score of 0, and responses that contain only unimportant nuggets will have a score of .1.

Figure 6 shows the nugget assignment and score computation of the first two statements from Figure 4.

Dashboard Generation

RESULTS

DISCUSSION

FUTURE WORK

From $k = 2$ raters, we had

$$N = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}^T$$

So the computed nuggets weights were:

$$\mathbf{w} = [0.1, 0.1, 1.0, 0.1, 0.55, 0.1, 1.0, 0.55, 1.0, 0.55, 0.55]$$

Sample of statements:

1. The gap in median household income, marriage rate, and median household wealth between whites and blacks have all grown.

$$\delta^{\mathbf{1}} = [0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0]$$
$$s_1 = \frac{1+1+1}{1+1+1} = 1$$

2. There is still a large financial gap between whites and blacks

$$\delta^{\mathbf{2}} = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]$$
$$s_2 = \frac{.55}{1} = .55$$

Figure 6. Nugget assignment and score computation of a sample of crowdsourced importance statements.

REFERENCES

1. 8 Red Flags about the "Useful chartjunk" paper. `http://junkcharts.typepad.com/junk_charts/2010/05/8-red-flags-about-the-useful-chartjunk-paper.html`.

2. Bateman, S., Mandryk, R. L., Gutwin, C., Genest, A., McDine, D., and Brooks, C. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2010), 2573–2582.

3. Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., and Pfister, H. What makes a visualization memorable? *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2306–2315.

4. Cleveland, W. S. *The elements of graphing data*. Wadsworth Publ. Co., Belmont, CA, USA, 1985.

5. Cleveland, W. S., and McGill, R. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association 79*, 387 (Sept. 1984), 531.

6. Few, S. The Chartjunk Debate: A Close Examination of Recent Findings. `http://www.perceptualedge.com/articles/visual_business_intelligence/the_chartjunk_debate.pdf`.

7. Isola, P., Xiao, J., Torralba, A., and Oliva, A. What makes an image memorable? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, IEEE (2011), 145–152.

8. Lin, J., and Demner-Fushman, D. Will pyramids built of nuggets topple over? In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, Association for Computational Linguistics (2006), 383–390.

9. Mackinlay, J. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics (TOG) 5*, 2 (1986), 110–141.

10. Tufte, E. R. *The visual display of quantitative information*, vol. 2. Graphics press Cheshire, CT, 1983.

11. Willett, W., Heer, J., and Agrawala, M. Strategies for crowdsourcing social data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2012), 227–236.