Visual Data Encoding Revisited

Niall Williams
University of Maryland, College Park

Vishnu Dutt Sharma
University of Maryland, College Park

Noor Pratap Singh
University of Maryland, College Park

Harnaik Dhami
University of Maryland, College Park

Shilpa Roy
University of Maryland, College Park

1 Introduction/Motivation

There has been a steady increase in data generation over the past century. The use of graphical methods often is the first step in any data analysis. These methods not only help us to identify patterns in the data but at times help us to derive a hypothesis, for which we can further design our tests. They also serve the first step of evaluation of any data model. While many individuals make use of graphical elements in their study, the use of a particular element is often motivated by intuition, common sense or status quo rather than a predefined set of scientific guidelines. We want to study whether we can employ some sort of theory that can enable us to find optimal graphical representations for our analysis.

Cleveland et al. [1] laid down principles that should be employed when constructing a graphical representation. In their work, they tried to visually decode the information encoded on the graph, by breaking it down into a list of elementary perceptual tasks. They claimed that these tasks are performed by people in order to retrieve the quantifiable information from a graph. An ordering of the elementary tasks has been suggested based on their accuracy. They hypothesized that using the graphical representations that employ the higher order tasks would lead to more accurate descriptions of the underlying data compared to the graphs employing lower order tasks. We want to reevaluate some of these principles as part of our study.

Cleveland et al. proposed an ordering of the 10 elementary tasks according to the accuracy of people's ability to judge differences in each task. Their ordering, from most accurate to least accurate, is position along a common scale > positions along non-aligned scales > length = direction = angle > area > volume = curvature > shading = color saturation. To test the validity of this ordering, Cleveland et al. conducted two experiments, in which position is compared with length (first experiment) and angle (second experiment), by drawing graphs that make use of these tasks. Then, the performance between the graphs is compared in the experiments. As a part of our study, we want to replicate the second experiment - where the position and angle are compared, by using bar graphs and pie charts. While both the experiments are interesting, budget and time constraints prevent us from replicating both experiments. Due to underlying simplicity, we stick with experiment two. The graphs will be generated using the process similar to as described in the paper. However, there will be certain differences compared to the original experiment. Our process of participant recruitment would be through crowdsourcing platforms such as Amazon Mechanical Turk (MTurk). As a result, the participants would be shown images on a screen compared to on paper, as in the original paper. Due to lack of our reach and the existing atmosphere (COVID times), recruiting participants for paper based study is not feasible. The usage of screens might create its own set of problems (confounding factors, lack of quality samples etc) that need to be dealt with, which we describe in the later sections. As a part of analysis we plan to include statistical tests which are not part of paper.

Data is driving most of the decisions that our systems are making today. Thus, having a solid data analysis pipeline is the need of the hour of which good visualization is an important part. Cleveland et al. tried to codify the graph construction part, which we believe leads to more correct/robust data interpretation. Since the study is very old, rerunning the experiment on a different population would enable us to re-verify the principles and check whether perception has changed over time.

2 Methods: Data collection

The participants we plan on recruiting for this study will be crowd sourced workers from Amazon's Mechanical Turk, since the task we are distributing is relatively simple and does not rely on demographic information.

In line with our end goal of replicating Cleveland and McGill's 1984 study [1] in a modern context, we plan on reproducing the position-angle experiment detailed in Section 4.2. Specifically, in line with the original paper, we intend to first generate random data, by choosing 10 sets of 5 arbitrary

numbers that sum to 100 and generating a bar graph and pie chart for each set. Both of these graphs will highlight the largest pie segment or bar. Participants will receive these 20 graphs in random order and be asked to judge what percentage of the maximum is represented by each of the other four segments or bars. These questions are not designed to take a long time - rather, we want participants' initial thoughts after spending 10-15 seconds with the graph. In addition to these twenty questions, we also plan on adding sanity checks to our survey, comprised of questions whose answers are overly simple in order to gauge the reliability of the data.

The data we collect will be the relative percentages that our participants report for each of the questions that are not attention checks. This will be collected and recorded in a survey that we send to the workers.

In order to accomplish this, the only tools we will require is the ability to generate random bar graphs and charts and the ability to create and send out a survey. Generating the graphs will be easy with the use of random number generators and an existing tool like Google Sheets. Sending out a survey will also be trivial since we have Qualtrics at our disposal.

All of our planned 50-60 subjects will receive the exact same treatment in a random order, so internal validity should be preserved in our experiment. In terms of external validity, the task that we are asking participants to perform is low-level; therefore, we can be relatively confident that crowd workers will be a reasonable approximation for the general population. Since we are measuring the percentages our participants report and comparing them with the true percentages, the construct we are employing to evaluate bar and pie charts is the accuracy of data perception. As the original paper suggests, this is not the only construct we can use to determine the effectiveness of these graphs, but since figures are designed to represent information, accuracy of perception is a reasonable construct to use.

We do not anticipate any ethical concerns to arise since our usage of MTurk implicitly grants informed consent for all of our participants. Further, the task is low-level and we do not intend to collect any sensitive information that might make our participants uncomfortable.

3 Methods: Data analysis

The initial data analysis will be a qualitative in nature. As stated in the previous section, sanity check questions will be added to the survey. The answers to these questions will be very easy and simple. We will then be able to filter out unreliable survey responses based off of these questions which will ensure the quality of the data. The authors of the original experiment conducted a similar exercise to filter their data. For the first experiment, they removed the judgements of four participants due to the unreliability of the data based on these sanity checks. The rest of the data analysis after this step will be quantitative. We will be analyzing the response accuracy similar to the original study. This is done by comparing the judgement percent (responses) with the true percent. They specifically used the following equation:

$$log_2 (| \textit{judged percent} - \textit{true percent} | + 1/8)$$

The original authors used a log scale for better visualization. They also added 1/8 because sometimes the accuracy would be really close to 0. It is too early to tell whether we need to do something similar, but the overall accuracy analysis will be the same. We will look at the absolute error.

4 Anticipated challenges and limitations

The original study design had 50 graphs for the position-length experiment and 20 graphs for the position-angle experiment. As we aim to have 50 participants similar to the original study, crowdsourcing for the first task has a budget requirement beyond the constraints. Thus, we will study only the second task in out project. The subsequent challenges include recruiting participants, setting up the experiment on MTurk and running it seamlessly. As one of the key differences between the original and the proposed setup is the mode of delivery (paper in original study, web-page in proposed study), we need to control the confounding factors originating from the sophistication in web-pages. This includes removing interactive controls, not using colors in plots as done in the original study, and instructing the participants to not zoom into the plots.

As we use web-page for delivery, we need to control how it looks on the participant's screen. Looking at charts on smartphone can affect the participant's estimation of differences in the graphs. The original study did not have this problem as it used paper for delivery. While we can put restrictions on the type of device through options provided by the hosting platform, we can't control the screen-resolution.

The original study was done 36 years ago and many news types of plots and charts have been adopted into general use. Most of these plots depend on rich usage of color (e.g. heatmaps, overlaid histograms, contour plots, etc.) for explaining the data. Although color saturation and shading are identified as elementary perceptual tasks, and can be studied similarly to the others tasks in the original study, we restrict ourselves to comparison of the point-angle task across time.

5 Proposed budget

The only thing we will need a budget for is paying participants who complete our study. The original study by Cleveland et al. recruited 55 participants. For the experiment we are replicating (the second experiment in [1]), each participant viewed 20 graphs. We are aiming to recruit about 50 participants. Based on a previous similar study that Niall conducted, we expect the experiment to take no more than 10 minutes. We plan to pay

participants no more than $2 for completing the experiment, based on the general wages paid for such tasks on MTurk. Accounting for overhead fees attached to online recruitment services, we anticipate a total cost of about $100 – $140 to run our experiment in full. We also intend to recruit participants independent of paid services, such as by recruiting friends and using snowball sampling. Thus, we hope to be able to reach our projected number of 50 participants without having to pay all 50 participants, which will hopefully bring the cost of the study closer to $100.

References

[1] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.