Comparing Averages in Time Series Data

Michael Correll
University of Wisconsin - Madison
mcorrell@cs.wisc.edu

Danielle Albers
University of Wisconsin - Madison
dalbers@cs.wisc.edu

Steve Franconeri
Northwestern University
franconeri@northwestern.edu

Michael Gleicher
University of Wisconsin - Madison
gleicher@cs.wisc.edu

ABSTRACT
Visualizations often seek to aid viewers in assessing the big picture in the data, that is, to make judgments about aggregate properties of the data. In this paper, we present an empirical study of a representative aggregate judgment task: finding regions of maximum average in a series. We show how a theory of perceptual averaging suggests a visual design other than the typically-used line graph. We describe an experiment that assesses participants' ability to estimate averages and make judgments based on these averages. The experiment confirms that this color encoding significantly outperforms the standard practice. The experiment also provides evidence for a perceptual averaging theory.

Author Keywords
Line graphs, Information Visualization, Colorfields, Visualization Evaluation.

ACM Classification Keywords
H.5.0. Information Interfaces and Presentation: General

General Terms
Experimentation

INTRODUCTION
A common task for viewers of visualizations is finding the "needle in the haystack"- precisely identifying individual values within a dataset. This task has been well-explored using established guidelines inspired by perceptual theory. But a core advantage of visualization is giving the viewer a sense of the "bigger picture" - the higher level, aggregate properties of the data. This task is less understood - especially as it connects to underlying perceptual theory.

CHI'12, May 5–10, 2012, Austin, Texas, USA.

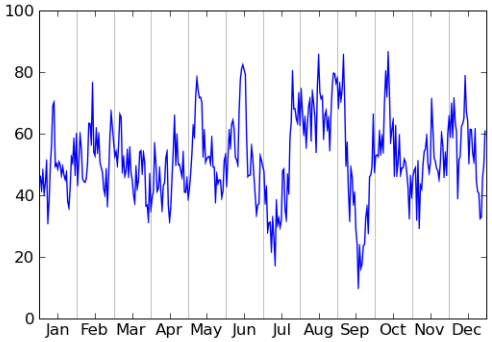


Figure 1. An aggregate judgment task: select the month with the highest *average* value. The maximum daily value (peak) may not occur in the month of maximum average.

In this paper, we explore visualization of aggregate properties, considering perceptually-motivated design and assessment to support a model summary analysis task. Our study considers a common data type: series data (an ordered sequence of numbers). Line graphs have been widely used to depict series data for over 200 years [33], and they have empirically-validated utility for identifying details within a dataset [7]. The task we consider is identifying a sub-range in a series (i.e. which month of the year) with the maximum average (see Figure 1). This task requires estimating an aggregate property over various ranges within a series and making judgments based on these estimates. Our study shows that viewers can make these judgments using a standard display (line graph), but that different visual encodings can provide improved performance especially as the task becomes more difficult. This shows potential for visual designs that better support summarization tasks.

Our approach to designing a visualization that readily presents aggregate information is to draw on efficient perceptual phenomena (i.e. preattentive and peripheral processing) that allow rapid summarization of data. Our contention was that encoding data values using color rather than the position would draw upon relevant perceptual processes to allow users to quickly summarize large regions of color. While color encodings may not be as precise as the position encodings used

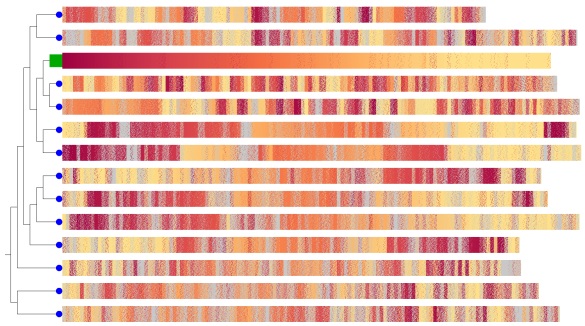


Figure 2. The Sequence Surveyor tool provides summary visualization of large genomics data sets by employing a colorfield design [1]. The present study provides an empirical grounding for this design, including its use of block-permuted displays.

by line graphs, our expectation is that the efficiency of summarization would be more valuable in this task. Our study confirms that such a colorfield design is effective for the average judgment task, providing better performance than the canonical line graph design.

Efficient summary processing by the visual system is a rapidly evolving topic in the perception community (see Background, below). There are several emerging theories for explaining our findings. Our visualization design draws on a theory of perceptual averaging , which is less popular because it cannot explain as wide a range of phenomena as other theories. For example, it does not apply to line graphs. The theory does suggest that colorfields will be even more effective if the colors are shuffled as it reduces the area over which averaging must be performed. Our experiment confirms this, providing more evidence for the theory.

In this paper, we describe an empirical examination of an aggregate judgment task for series data. We discuss how perceptual theory suggests a non-standard visual design. Data from an experiment, conducted using crowd-sourced participants, is presented and analyzed. The results confirm that viewers can adequately perform the aggregate judgment task using the standard line graph representation. The results also show that a non-standard colorfield design offers significantly better task performance, especially as the examples grow challenging. Our findings have a range of implications as they show the potential for designing visual representations that support aggregate judgment tasks, and suggest certain perceptual mechanisms that might guide this design.

This paper makes the following primary contributions:

- We show how a theory of perceptual averaging can suggest designs for information visualizations that support summarization. These designs may defy the conventional wisdom, but significantly improve task performance.
- We show empirically how a non-standard colorfield design outperforms the common positional encoding (line graph) on a model visualization task. This model task applies to a data type common in real applications.
- We provide further support for a theory of perceptual averaging by showing that permuting the data has a significant effect on performance in the colorfield case (where the theory predicts it would), but no significant effect in the line graph case (where the theory is not applicable and so has no predictions of improvement).

Our results directly address the common case of 1D signal data, such as time series. Additionally, the design principles suggested by the theory apply to a wide range of applications. For example, in genomics the problem of scale requires consideration of aggregation (in that the entire genome of an individual can usually not be feasibly presented simultaneously), and the ability of users to make aggregate judgments. We have employed colorfield designs as a solution to this design problem [1][8]. The present study provides empirical validation of the colorfield approach for summarizable displays, and more generally, of the applicability of recent perceptual theory in visualization design.

BACKGROUND

There is a rich literature on the use of empirical studies to inform visual design for data presentation [7, 20], and such studies can inform a variety of aspects of the design process [27]. By experimentally validating performance characteristics at various steps of the process, designers can understand how different techniques work together to more effectively convey visual information. More recently, these empirical studies have been crowd-sourced to allow for a greater participant pool, and thus a more thorough exploration of the design space [16].

Empirical studies have also been used to illustrate how ideas from the study of perception can be applied to visual design. Notably, the work of Healey and Enns [15] use an understanding of visual attention to guide the design of visualizations, with particular focus on the use of pre-attentive phenomena to supplement visual data displays. This work shows that by exploiting "pop out" phenomena, designs can allow users to quickly find details. Other work has used a similar bottom-up understanding of perception as well as traditional best practices in visual design to develop guidelines for data presentation [21]. While these studies provide valuable groundwork for the design of data displays, they do not consider the perception of aggregates in the displays, nor do they assess making judgements based on aggregate information.

For the specific case of interpreting series data, there is a significant literature surrounding the exploration and understanding of the general properties of the line graph encoding, including categorization [30], slope, curvature [4], dimensionality [26, 22], and continuity [9]. However, these papers seldom explore the low-level mechanisms at work in line graphs and do not focus on the design space beyond conventional, single series line graphs. Recent work has considered the perception of multiple time series. Javed et al. [18] examine user performance over different encodings for multiple times series datasets, while Lam et al. [23] compare various interaction techniques. While these studies have shown the success of empirical evaluation in understanding the design of visualizations for multiple time series data, neither consider aggregate judgements, instead opting to focus on more detail-oriented tasks. These works are also more design-focused

rather than seeking to understand the perceptual mechanisms underlying the empirical results.

As discussed below, recent work in perceptual science has considered the mechanisms of how people interpret complex scenes. Some of these models are beginning to be considered in the context of interface design, for example, in Rosenholtz et al.'s work on clutter [35, 36], but are not yet commonly considered in data displays. Also, such work is generally concerned with locating details amidst visual clutter, not judging the aggregate properties of the cluttered elements. A notable exception is the work of Ramanarayanan et al [34] who consider how to simplify rendering aggregates of objects in a given scene while preserving viewer perception of the aggregate. The general case of how to design visualizations to support ready access to summary data as opposed to specific values has only recently been explored [10, 38].

Visual and Mental Averaging

Research in perception has shown that several types of peripherally processed features can support efficient operations such as visual search and outlier detection [12]. Recent research suggests that general summary information about a collection of objects can also be determined efficiently in the visual periphery [3].

One explanation for efficient summary phenomena is the theory of perceptual averaging. Perceptual averaging suggests that averages of optical parameters, such as intensity or hue, can be averaged over a spatial range efficiently, as these averages are available to the visual system. Intuitively, one could imagine doing this averaging by squinting their eyes. However, such blurring is effectively done automatically in the early visual system on many pre-attentive features (such as intensity, color, and orientation) as multiple scale encodings are believed to be used extensively. Early levels of visual processing occur in parallel across levels of spatial frequency, with representations of both high and low-pass filtered versions of the visual image [17]. By accessing these different representations, one could potentially "see" the averages directly. Such visual averaging of pooled regions of an image has also been shown to occur in the visual periphery during later phases of visual processing [12]. Observers can extract information about average values across several types of pre-attentive features of visual objects, including average size [3, 6], orientation [2, 31], or location [2], without actively attending to the objects. Several findings suggest that these averages can be computed efficiently; however, the perception of such features by sampling procedures is difficult to rule out [29].

We predict that perception of line graphs do not draw upon these same summary phenomena, and so cannot be averaged as efficiently as the colorfield encoding. The complex shape of line graphs are processed in higher-level visual areas that may not average across multiple instances in a useful way, and there is no evidence that the visual system can efficiently average across height per se. In fact, global shape perception seems to be an perceptually inefficient task [41].

There is evidence that the visual system can average across the *size* of objects [3, 6], which could act as a proxy for height. To make use of this ability it is necessary to segment out individual "objects" from the global shape of the graph. The perceptual rules which guide this segmentation are inflexible, and the observer may not be able to arbitrarily set the desired areas for aggregation [39, 11]. By artificially destroying the "shape continuity" of a line graph it may be possible to remove some of the inefficiencies of this process, but aggregate shape may not be able to "pop-out" in the same way as aggregate color.

METHODS

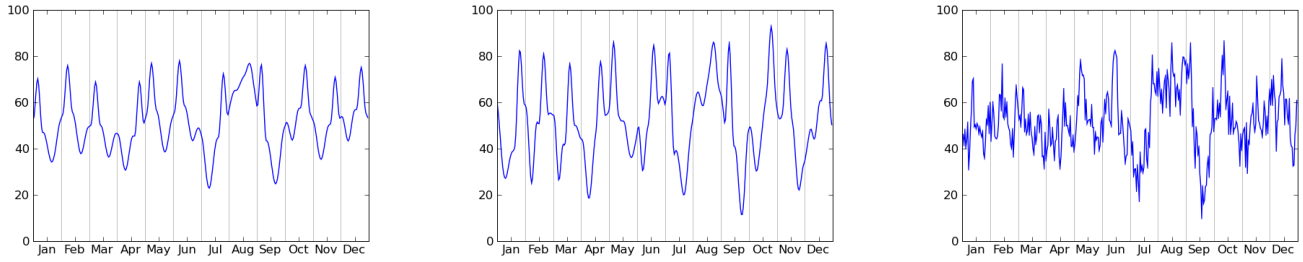
Using Amazon's Mechanical Turk infrastructure we ran a study comparing the ability of users to accurately determine aggregate information from time series data using either a colorfield or line graph encoding at various levels of task difficulty. Within the colorfield and line graph conditions, we tested different sorts of permutation to assess the validity of our underlying perceptual assumptions: a 2D permutation of data in the colorfield case (to make perceptual color averaging easier) and a 1D permutation of data in the line graph case (to break shape continuity). Our study supported our hypotheses: the colorfield outperformed line graph, and permuted colorfields outperformed non-permuted colorfields (since pre-attentive phenomena could be more easily accessed) whereas permuted line graphs did not outperform non-permuted line graphs (where our theory makes no prediction).

Task Design

Our goal was to create a task that would allow us to assess a viewer's ability to make judgments about aggregate properties of data. To allow for effective experimentation, the task could not require the viewer to have domain knowledge, familiarity with displays beyond basic charts or graphs, or advanced statistical skills. The task needed to require estimating the aggregate property over subsets of the data well enough to make comparative judgments about these aggregate regions. We were not interested in the precise value of the estimate, and preferred a task with a forced choice decision.

To fulfill these requirements, we designed an experimental task of analyzing time series data over the course of a year, specifically, the task of analyzing sales data for a fictional company. Participants were asked to identify the month that had the maximum average sales. The choice of a one year time span (12 "months" of 30 days each) provided a number of data points that would be manageable for the participants but not trivial to analyze. The year length also afforded a natural subdivision for the data series, namely, trends over the course of each month.

Comparing averages across months is a task that is involved (it requires rapid analysis of hundreds of data points) but at the same time domain-invariant (only basic statistical knowledge is required). While this task is contrived in that specific data encodings could be used to precisely display the per-month average, by including all of the data and then asking for only a particular property we simulate real world data analysis tasks where users must make rapid assessments of statistical properties of large amounts of data.



(a) Low Noise (b) Medium Noise (c) High Noise

Figure 4. Difficulty for the aggregate judgment task is partly modified by adding noise at three ordinal (but non-linear) levels: low ((a)), medium ((b)), and high ((c)). Noise variation is not uniform: the different noise levels vary in both amplitude and frequency.

We needed to ensure that the task required participants to estimate the averages by precluding alternate strategies. For the maximum average task, a potential alternate strategy is to report the month containing the maximum value, a strategy otherwise known as "peak finding". We precluded this strategy by constructing our data such that the peak data point rarely occurred in the month with the maximum average (less than 0.5% of the time). The difference between maximum average judgments and peak finding was explained in the instructions to participants, who were also instructed that the maximum average may not necessarily be the peak. Participants were not told that the peak occurring in the maximum average month was rare, as this would suggest "peak avoidance" as an alternate strategy.

The more detailed design of the task can be subdivided into three areas: the design of the actual values of the time series data, the design of the visual encodings of the data, and the design for the experimental evaluation of these encodings.

Data Design

Key to the experiment was the ability to generate well-structured data signals that were both appropriate for the maximum average month task and gave sufficient control over the kinds of challenges we want to explore in the experiment. In particular, we wanted to add varying amounts of noise to the data signal and to control the sensitivity required in making judgments about the estimated averages.

To make our data generalizable across display conditions, all data values were drawn from the range $[0...100]$. To populate our 360 data points (30 days per artificial month), we selected a winning month at random with an average value u_w from the range $[50...98]$. From our initial assumptions, we attempted to have some control over the apparent difficulty of the task by selecting up to k "distracter" months, where k is in the range $[2...4]$. Distracter months were assigned average values that were influenced by a second parameter d, from the range $[1...10]$, such that the average of these distracter months was exactly $u_w - d$. The remaining months had averages randomly drawn from the range $[20...(u_w - d - 1)]$. See figure 5 for a visual description of these parameters. Once the average values for each month were determined, the daily data points were generated by using structured noise to create a smooth curve preserving the previously computed averages [32]. Additional structured noise n was added, allowing parametric control over the amount of noise (parameter n can be assigned 3 levels, Figure 4). Peaks are then added to each month to create a regular pattern of higher values and to dissuade peak-finding as a viable strategy, and smooth displacements are used to adjust the non-peak values within each month to preserve the target averages.

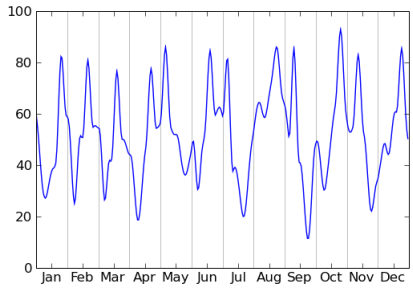
We generated one data series for each possible combination of winning month, k, d, and n, a total of 1080 distinct series. We roughly equated k and d with task difficulty, and so to fairly balance the task with respect to hardness, we evenly drew from every k and then from every d value to provide the data for each participant.

Display Design

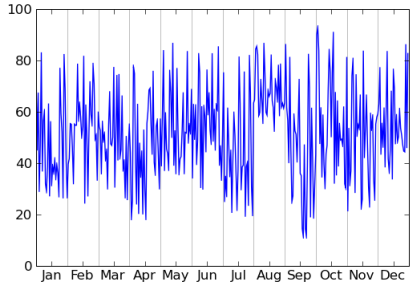
Our experiment considered different visual encodings for the series data. We used a line graph encoding as this is the standard method for displaying series data. A line graph encodes time series data by measuring time along the horizontal axis and the corresponding values along the vertical axis. The line graphs were generated using the MatPlotLib library. The vertical axis was labeled with values specified with tick marks every 20 value points, from 0 to 100. The horizontal axis was likewise labeled with tick marks at each month and with gray spans defining the boundaries between the various months.

The theory of perceptual averaging suggests that a retinal measurement, such as color (intensity, hue, saturation) may be easily averaged by the visual system. Even though color encodings have been shown to be less accurate than positional encodings [7], we hypothesized that the potential facility for averaging would outweigh the reduced accuracy. Therefore, we created a color encoding of the data, where the range of values was mapped to a selected color map. Such color encodings of series data are uncommon, but not without precedent (c.f. Lasagna Plots [40]). Note that in keeping with the best-practices in color map design (c.f. [5]), we selected a color map that varied in hue, saturation, and brightness from the Color Brewer selection tool [14] (specifically a red-blue diverging scale). Color values were interpolated in CIE L*a*b* space to provide an approximately perceptually linear interpolation between color values (see Figure 3(c)).

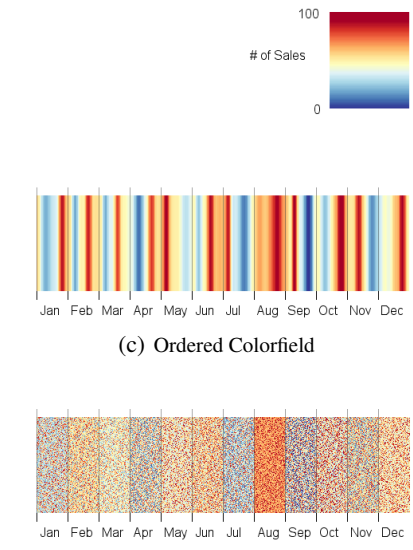
The theory of perceptual averaging would further suggest that by randomly shuffling the colors within the month, visual av-



(a) Ordered Line graph



(b) 1D Permuted Line graph



(c) Ordered Colorfield

(d) 2D Permuted Colorfield

Figure 3. An example of the four display conditions tested in the aggregate judgment task. Line graphs ((a), (b)) provide a standard method for viewing the series data, while colorfields ((c), (d)) instead use color to encode value. The effect of permutation on perceptual averaging was tested by permuting the points within each month horizontally in the line graph condition (b) and both horizontally and vertically in the colorfield condition (d).

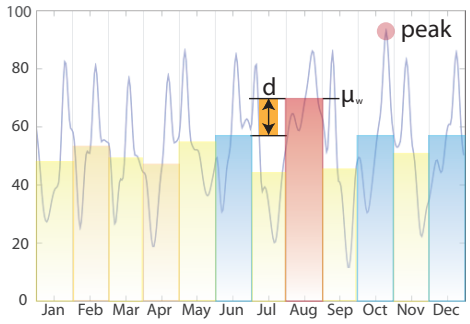


Figure 5. Each experimental stimulus has one month with the highest average value (μ_w) and k distracter months at the next highest average value (red), whose average is d below the highest months. Peaks generally did not appear in the highest month to protect against peak-finding strategies. We expect the difficulty of the aggregate judgment task to vary with both k and d.

eraging would be even easier since more local averages over smaller regions would be representative of the whole (i.e. the colors that need to be mixed would be closer together). Therefore we created a display type where within each month's block, the pixels of the block were randomly permuted (see Figure 3(d)); this technique has some precedent in the literature [1, 13]. This permutation does not preserve low-level patterns in the data, but is designed to facilitate the averaging of colors across known boundaries of aggregation.

While the line graph condition functioned as a control case based on traditional time series encodings, to further our exploration of the perceptual theories at work in aggregation we created a display condition where the line graphs were permuted. For the permuted line graph case, however, the data was only randomly shuffled in the horizontal direction, as changing vertical positions would change the encoding of the data (see Figure 3(b)). This permutation is not symmetric to the colorfield permutations (the permutations occur in one dimension rather than two), and was intended to assess the effect of intentionally breaking up continuous shapes and trends in the data.

Our experiment considered four display conditions: two different visual encodings (color and position) and a manipulation of the encodings to highlight expected perceptual results (2D permutation in the color case, to test our theory of perceptual averaging, and 1D permutation in the position case to test shape continuity theory). The horizontal spacing of the displays was made to be identical, and identical axis labels were used. For color encodings, a legend was provided immediately above every display. The height of each display type is relevant to the ease of the task: a larger height leads to increased fidelity of position in the line graph case and increased area devoted to color in the colorfield case. As a result, the display size was kept standard for each display condition. The height of the colorfield display was compressed compared to the total height available to the line graph to make room for the color legend.

Experimental Design

We considered the two basic design types (line graphs and colorfield encodings) to be sufficiently different that they required separate instructions. Therefore, we chose to use a between-participants design: each participant was randomly assigned to one display type condition. Each participant received instructions for only one type of display, and saw 30 trials of the same type of display.

In addition to this basic design type, our experiment considered several factors: whether the data was permuted within months and several properties of the data series relating to its "hardness" (the k and d levels discussed above). Noise level and permutation were considered within-participants factors. Each of the experimental factors, as well as our *a priori* k and d hardness factors, were stratified. However, the combinations and ordering of the factors were randomized. Therefore, while each participant was guaranteed to view each level of each factor an equal number of times, there were no guarantees of which combinations of factor levels they would see.

A pretest and tutorial phase used Ishihara [24] plates to detect and exclude participants with Color Vision Deficiency (CVD), as the color ramp used in our colorfield condition was not chosen with the needs of users with CVD in mind. The tutorial phase also served to familiarize the participant with the display condition, and emphasized the difference between the averaging and peak-finding task both verbally and visually.

Experimental Procedure

The experiment consisted of asking participants to answer 30 questions. Each question presented a display of a data series and asked the participant to select the month with the highest average value. The number of questions was chosen based on pilot studies that suggested that the total task would take each participant approximately 15 minutes.

For each trial, participants were shown a web page that contained a box where the display would appear. At the top of the screen, the question ("Choose the month with the highest average value") appeared. At the bottom of the screen, a radio-button list of the months was arrayed horizontally so that the month names would appear directly beneath the corresponding region of the display. A "submit button" appeared below the month buttons. The lower buttons were disabled before the display was presented.

Initially, the display box contained a button allowing the participant to click when they are ready. Once this button was clicked, an initial data series was displayed within the box. The display was shown for a maximum of 20 seconds, although the participant was free to answer the question either during this time or afterwards. After time had elapsed, the display was removed. The time limit was imposed to preclude having the participants make measurements of the screen, but selected so that it would not rush the participants.

Measures

Since our judgments were about the accuracy of our different visual designs, our measure was the number of correct answers for each condition. While this was the primary measure, to perform validation we collected additional data. Participants were tracked by IP address to ensure that the same participant did not perform the task multiple times. This identifying data was discarded upon beginning the analysis for privacy reasons.

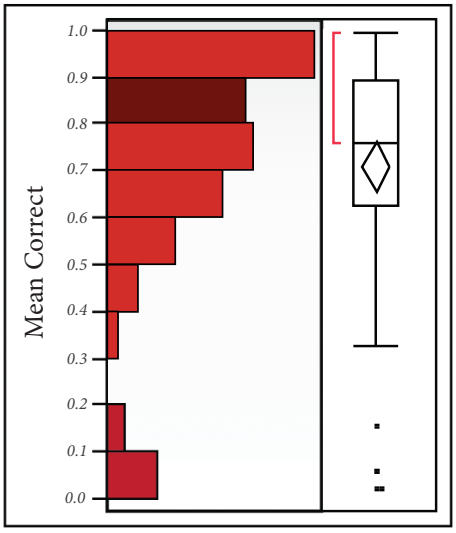


Figure 6. The accuracy distribution on the aggregate judgment task by performance. Of the 74 original participants in our study, we excluded results from eight of the participants whose performance failed the inner quartile range test.

For the primary experimental task, we recorded the data series being presented, the visual encoding used, and the various hardness parameters, as well as the identity of the winning and distracter months. After the users guessed a value we recorded their guess, whether it was correct, and the time spent on the question.

After completion of the task we asked for self-reported demographic data (age, gender, and education level) as well as asking for general comments.

Participants

Since our task could be entirely completed using the browser window, we drew our participants from Amazon's Mechanical Turk service, specifically those users registered in North America with at least a 95% "approval" rating. Mechanical Turk's large userbase and quick turnaround speed was considered an appropriate tradeoff for losing the ability to directly observe participants. In total, 74 participants completed the 30 question comparison task, 42 female and 32 male. Ages ranged from 18 to 62 ($\mu = 34.7$, $\sigma = 12.0$). The demographics data was by necessity self-reported, but does conform to the general distribution of North American Turk users [37]. We followed acknowledged best practices to improve the reliability of our experimental data, including randomizing questions, requiring mandatory questions of different input types, and preventing "click-through" behavior [19, 25].

We excluded 5 participants as significant outliers in terms of

performance (they failed the interquartile range test for k=1.5, see Figure 6). These participants were all from the line graph condition. The under-performing participants were compensated at the same rate as the other participants.

RESULTS

Given the uncertain provenance of data from Mechanical Turk, we performed validations on the dataset as a whole to eliminate possible sources of bias. We examined the selections and time taken by participants whose performance metrics were significant outliers to determine if there were any patterns indicative of adversarial participants (e.g. people clicking through as fast as possible through only a few months). There were no such red flags. We also noticed no overall effects from task learning or task fatigue. In particular, there was no statistically significant effect of question number on performance, F(29,1921) = 1.19, $p <= .211$.

We performed a two-way analysis of variance between the three levels of noise n, the ten possible settings of d (the difference between the highest and the next highest averages), the three possible settings of the number of distracters k, the two types of display (colorfield versus line graph), the two possible permutations (original ordering or permuted), and the question number $[1...30]$. All measures of statistical significance are at the 0.95 confidence level. See Table 1 for detailed results.

Our primary proposed *a priori* difficulty factor d had a statistically significant effect on performance as smaller gaps between the winning and distracter months resulted in worse performance, F(9,1921)=34.5, $p <= 0.0001$ (Figure 7). Our other tentative hardness metric k did not have a statistically significant effect, F(2,1921)=1.47, $p <= 0.231$, although there were some cross-conditional effects (Figure 8). The relative noisiness of the data n was a significant performance effect, F(2,1921)=15.2, $p <= 0.0001$ (Figure 9). Noisier data was associated with worse performance. Encoding (line graph versus colorfield) had a significant effect, F(1,1921)=143, $p <= .0001$, as line graphs significantly underperformed colorfields. Ordered versus permuted data had no statistically significant effect overall, F(2,1921) = 2.28, $p = 0.132$, but there were cross-conditional effects (Figure 11).

Based on the initial results, we performed some additional pairwise comparisons across conditions. Although permutation overall had no significant effect, there was a significant interaction between type of display in conjunction with permutation of data, F(2,1921)=16.0, $p <= 0.0001$ (Figure 11). Within the colorfield display type, participants in the permuted colorfields ($\mu = .914, \sigma = .073$) outperformed those in the ordered case ($\mu = .815, \sigma = .197$). The interaction between noise and type was not a significant effect, F(2,1921)=0.314, $p = 0.381$. With respect to our a priori hardness parameters, d had significant interactions with type, F(9,1921) = 2.08, $p <= 0.028$, but k did not, F(2,1921)=1.147, $p = 0.231$. The line graph display was more sensitive to changes in d than the colorfield display.

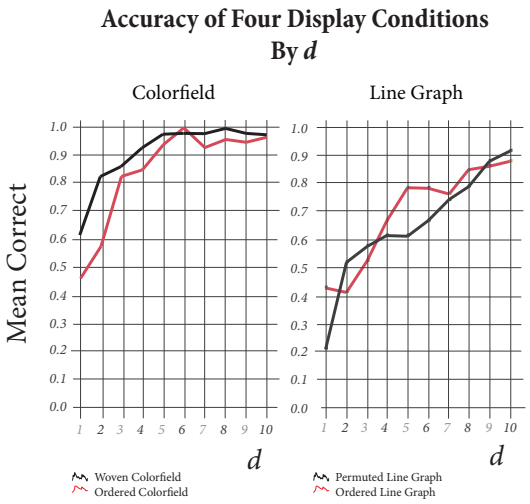


Figure 7. A comparison of the effect of d on performance in the four display conditions. Participants perform better at the aggregate judgment task when using permuted colorfields (black, left) than ordered colorfields (red, left), with a marked difference in the more difficult stimuli. Permutation has no significant effect on the overall performance of participants on the aggregate judgment task using line graphs.

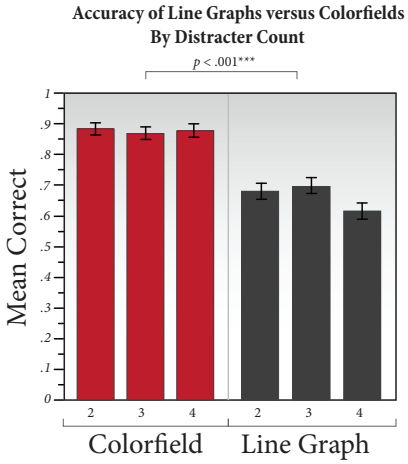


Figure 8. Performance as a factor of distracter count in the colorfield and line graph conditions. The number of distracters (k) has no significant effect on participant accuracy on either colorfields or line graphs.

Source	Number of Parameters	Degrees of Freedom	Degrees of Freedom for Dependent Measure	F	p
Display Type	1	1	1921	143.7537	< 0.0001
Permutation Type	1	1	1921	2.2769	0.1315
Noise Level (n)	2	2	1921	15.2464	< 0.0001
Distracter Difference (d)	9	9	1921	34.4739	< 0.0001
Number of Distracters (k)	2	2	1921	0.9657	0.3809
Display Type x Permutation Type	1	1	1921	15.951	< 0.0001
Display Type x Noise Level	2	2	1921	0.3144	0.7303
Display Type x Distracter Difference	9	9	1921	2.0817	0.02808
Display Type x Number of Distracters	2	2	1921	1.4669	0.2309
Question Index	29	29	1921	1.1935	0.2110

Table 1. The ANOVA Table for the aggregate judgment task.

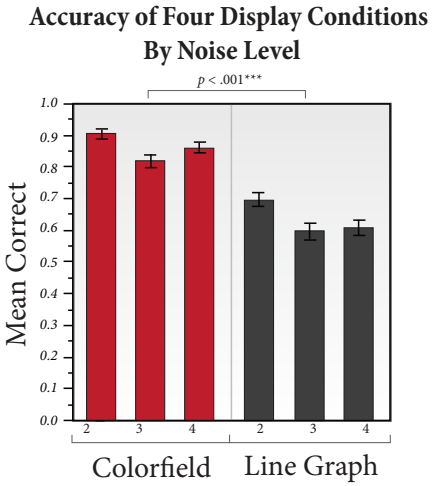


Figure 9. The effect of noise level in the colorfield and line graph conditions. While graphs with the least amount of noise are associate with higher accuracy, the ordinal but non-linear nature of the noise level prevents us from making concrete claims about a monotonic decrease in performance with respect to amount of noise.

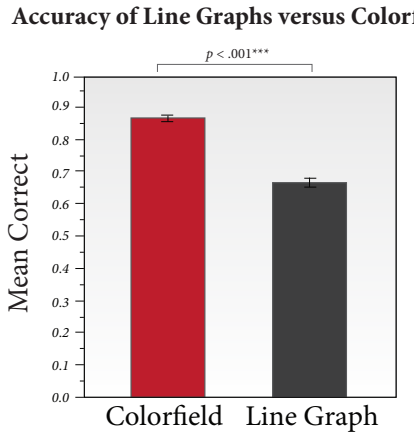


Figure 10. Accuracy in the aggregate judgment task across task type. Despite the conventional use of line graphs for series data, participants performed significantly better on the aggregate judgment task while using colorfield displays than line graphs.

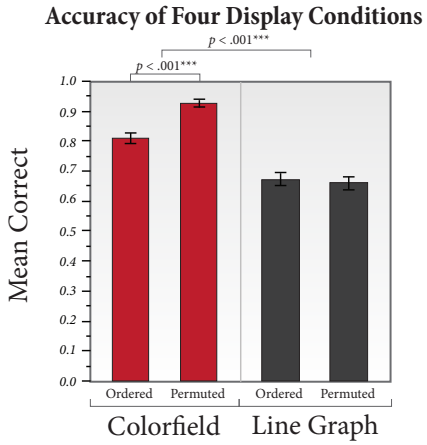


Figure 11. Accuracy in the aggregate judgment task across the four display conditions. The use of permutation to augment the aggregate judgment task yielded mixed results. Permutation in colorfields did yield statistically better performance than ordered colorfields, whereas no effect was found between ordered and permuted line graphs.

DISCUSSION

Our study confirms that people are able to retrieve high-level statistical overviews of series data efficiently. Using a standard visual encoding of the data (a line graph), participants performed significantly above chance, with degraded performance as we changed our proposed hardness parameters. We feel our data represented a large cross section of possible task difficulties.

The theory of perceptual averaging would suggest that color encodings would afford reliable determination of averages, as averaging over a range could be performed by the perceptual system instead of as a mental computation. Our experiment supports this conclusion: participants task performance was significantly better using the color encoding than the standard line graph encoding across all difficulty levels.

Conventional wisdom [28] (and prior perceptual experiments [7]) in the information visualization community would suggest that positional encodings (a line graph) support better retrieval of accurate values than color encodings, and would allow for better performance in general data analysis tasks. However, our results show that designs that allow for perceptual averaging, in particular color encodings, may be preferable when viewers must consider averages over ranges.

The utility of the theory of perceptual averaging in designing information displays inspired us to seek evidence to support it. The theory suggests that permuting the colors to be averaged would make averaging even easier, as the visual pool-

ing of color could be more local. In contrast, the theory provides no prediction about line graphs. Our experimental results support the theory: permutation significantly improved performance on the color encoding displays, but did not significantly affect performance for line graphs. Other theories of visual summarization (such as a theory of shape continuity)usually suggest that both conditions would be aided by permutation. The 2D color permutation also has other effects, such as better delineating block boundaries, which may also contribute to the performance.

Overall, our study suggests that if visualizations are meant to aid a user in getting the "big picture" (e.g. assess aggregate information over ranges), they can be designed to better support these tasks. The theory of perceptual averaging provides a foundation for such designs. The fact that the designs suggested by the theory may contradict the conventional wisdom of visual display design suggests key tradeoffs: designs best suited for finding details may not be ideal for tasks requiring aggregation, and vice versa.

Limitations

While our study does strongly suggest the possibility for displays designed for aggregate assessment, limitations in the study suggest further research may be needed. Foremost, our study explores a single kind of aggregate judgment (averages over ranges) on a single kind of data type (series). While we hypothesize that averaging is a central mechanism for exploring larger scales of data (it performs a filtering of details, in the signal processing sense), such speculation should be explored to generalize the results of this study to the assessment of other aggregate properties beyond averages, such as the assessment of variance or trends within a dataset. Similarly, while the perceptual averaging theory extends naturally to 2D displays, our study needs to be extended to other types of data, particularly 2D fields or multiple 1D series. For these data types, the use of color encoded displays is more common (e.g. heat maps and "lasagna plots"[40] respectively).

Another limitation of our study is the contrived nature of the task of asking for judgments about averages over regions. The permuted encodings make use of knowledge of the aggregation ranges, which is why we consider them more as an assessment of the perceptual theory than a practical design: if we knew the viewer wanted to assess averages over months, we could have presented those values directly. For the non-permuted colorfield encoding, there is no assumption of the aggregation duration; although further study should confirm the ranges of durations that viewers can effectively average. The extent to which the ability to precisely determine specific values is harmed by both the colorfield and especially the permuted colorfield encodings is also unknown. It is also possible that permutations might introduce patterns in the display that are not representative of patterns in the data. Generally, we view the studied designs as a having potential tradeoffs that need to be further explored empirically.

Our experimental methods also limit the power of our study. Our desire to make peak-finding a non-viable strategy resulted in series with atypical distributions of highs and lows. Our noise generation techniques were also non-uniform with respect to the frequency and amplitude of the generated noise, which prevented us from making anything other than relative judgments about noise level. The data generation process also limited our ability to control the "hardness" of the experiment: the task was too easy in some of the colorfield conditions, which made it difficult to distinguish the effects of some experimental parameters.

We believe that through careful experimental design, we have obtained reliable results from crowd-sourcing. In fact, the potential diversity of color performance of participants' display devices makes the success of the color encodings more robust: they seem to work even without calibrated monitors (modern displays exhibit little or no geometry distortion, the variety in participant displays is unlikely to be a factor for purely geometric encodings). However, the use of crowd-sourced participants precluded us from doing post-hoc interviews, for example to diagnose mistakes, understand strategies, and collect qualitative data about relative task hardness.

In the future, we hope to more fully explore the space of designs that afford aggregate judgments, and the application of the perceptual averaging theory to visualization design. This initial study did not consider many design choices, such as the color palette used. It did not consider tasks beyond maximum averaging, aggregation over variable sized regions, or datasets beyond 1D series. Our experimental methodology should extend naturally to these explorations.

CONCLUSION

Visualization offers the potential of allowing viewers to see the "big picture" and make assessments over ranges of data. In this paper, we have explored the ability of visualization to convey an aggregate over a range in an important case: series data. Our study shows that while viewers can make judgments of averages over regions using standard designs, displays can be created that improve task performance. In particular, we show how a perceptual theory suggests a colorfield design that has significantly better performance on the average judgment task than the standard line graph encoding. The study also provides further evidence for a theory of perceptual averaging. Our results have implications in suggesting how visual displays should be designed to afford effective assessments of aggregate quantities.

Acknowledgments

This project was supported in part by NSF awards IIS-0946598, CMMI-0941013, and BCS-1056730. Albers was supported in part through DoE Genomics:GTL and Sci-DAC Programs (DE-FG02-04ER25627).

REFERENCES

1. Albers, D., Dewey, C., and Gleicher, M. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. In *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, IEEE (December 2011).
2. Alvarez, G. A., and Oliva, A. Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences 106* (2009), 7345–7350.
3. Ariely, D. Seeing sets: representation by statistical properties. *Psychological Science 12* (2001), 157–162.
4. Best, L. A., Smith, L. D., and Stubbs, D. A. Perception of linear and nonlinear trends: using slope and curvature information to make trend discriminations. *Perceptual and motor skills 104*, 3 Pt 1 (June 2007), 707–21.
5. Brewer, C. Color use guidelines for data representation. In *Proceedings of the Section on Statistical Graphics, American Statistical Association* (1999), 55–60.
6. Chong, S. C., and Treisman, A. Representation of statistical properties. *Vision Research 43* (2003), 393–404.
7. 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), 531–554.
8. Correll, M., Ghosh, S., O'Connor, D., and Gleicher, M. Visualizing virus population variability from next generation sequencing data. In *2011 IEEE Symposium on Biological Data Visualization (BioVis)*, IEEE (oct 2011), 135–142.
9. Eaton, C., Plaisant, C., and Drizd, T. Visualizing missing data: graph interpretation user study. *Human-Computer Interaction-INTERACT 2005* (2005), 861–872.
10. Elmqvist, N., and Fekete, J.-D. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *Visualization and Computer Graphics, IEEE Transactions on 16*, 3 (may-june 2010), 439 –454.
11. Franconeri, S., Bemis, D., and Alvarez, G. Number estimation relies on a set of segmented objects. *Cognition 113*, 1 (2009), 1–13.
12. Freeman, J., and Simoncelli, E. P. Metamers of the ventral stream. *Nature Neuroscience*, 9 (Sep 2011), 1195–1201. Advance online publication: 14 Aug 2011.
13. Hagh-Shenas, H., Kim, S., Interrante, V., and Healey, C. G. Weaving versus blending: a quantitative assessment of the information carrying capacities of two alternative methods for conveying multivariate data with color. *Visualization and Computer Graphics, IEEE Transactions on 13*, 6, 1270–7.
14. Harrower, M., and Brewer, C. Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal 40*, 1 (June 2003), 27–37.
15. Healey, C. G., and Enns, J. Building perceptual textures to visualize multidimensional datasets. *Proceedings Visualization '98 (Cat. No.98CB36276)* (1998), 111–118,.
16. Heer, J., and Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, ACM (2010), 203–212.
17. Hughes, H. C., Nozawa, G., and Kitterle, F. Global precedence, spatial frequency channels, and the statistics of natural images. *J. Cognitive Neuroscience 8* (July 1996), 197–230.
18. Javed, W., McDonnel, B., and Elmqvist, N. Graphical perception of multiple time series. *Visualization and Computer Graphics, IEEE transactions on 16*, 6 (2010), 927–34.
19. Kittur, A., Chi, E. H., and Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (2008), 453.
20. Kosara, R., Healey, C. G., Interrante, V., Laidlaw, D., and Ware, C. Thoughts on user studies: Why, how, and when. *IEEE Computer Graphics and Applications 23*, 4 (2003), 20–25.
21. Kosslyn, S., and Chabris, C. Minding information graphics. *Folio: the Magazine for Magazine Management, Feb. 1, 1992, 21 2* (1992), 69–71.
22. Kumar, N., and Benbasat, I. The effect of relationship encoding, task type, and complexity on information representation: An empirical evaluation of 2D and 3D line graphs. *MIS Quarterly 28*, 2 (2004), 255–281.
23. Lam, H., Munzner, T., and Kincaid, R. Overview use in multiple visual information resolution interfaces. *Visualization and Computer Graphics, IEEE transactions on 13*, 6 (2007), 1278–85.
24. Legrand, H., Rand, G., and Rittler, C. Tests for the detection and anlysis of color-blindness i. the ishihara test: An evaluation. *Journal of the Optical Society of America 35* (1945), 268.
25. Mason, W., and Suri, S. Conducting Behavioral Research on Amazon's Mechanical Turk. *Behavior Research Methods* (2010).
26. Meserth, T., and Hollands, J. Comparing 2D and 3D displays for trend estimation: The effects of display augmentation. In *Human Factors and Ergonomics Society Annual Meeting Proceedings*, vol. 43, Human Factors and Ergonomics Society (1999), 1308–1312.
27. Munzner, T. A nested model for visualization design and validation. *Visualization and Computer Graphics, IEEE transactions on 15*, 6 (2009), 921–8.
28. Munzner, T. Visualization. In *Fundamentals of computer graphics, 3e.*, P. Shirley and S. Marschner, Eds. AK Peters, 2009, 675–707.
29. Myczek, K., and Simons, D. J. Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & Psychophysics 70* (2008), 772–788.
30. Nourbakhsh, M. R., and Ottenbacher, K. J. The statistical analysis of single-subject data: a comparative examination. *Physical therapy 74*, 8 (Aug. 1994), 768–76.
31. Parkes, L., Lund, J., Angelucci, A., Solomon, J., and Morgan, M. Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience 4* (2001), 739–744.
32. Perlin, K. Improving noise. In *ACM Transactions on Graphics (TOG)*, vol. 21, ACM (2002), 681–682.
33. Playfair, W. *The commercial and political atlas: representing, by means of stained copper-plate charts, the progress of the commerce, revenues, expenditure, and debts of England, during the whole of the eighteenth century*. Printed by T. Burton, for J. Wallis, 1801.
34. Ramanarayanan, G., Bala, K., and Ferwerda, J. A. Perception of complex aggregates. *ACM Trans. Graph. 27* (August 2008), 60:1–60:10.
35. Rosenholtz, R., Dorai, A., and Freeman, R. Do predictions of visual perception aid design? *Transactions on Applied Perception* (2011).
36. Rosenholtz, R., Li, Y., Mansfield, J., and Jin, Z. Feature congestion: a measure of display clutter. *Proceedings of the SIGCHI conference on Human factors in computing systems* (2005).
37. Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*, CHI EA '10 (2010), 2863–2872.
38. Shneiderman, B. Extreme visualization: squeezing a billion records into a million pixels. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, ACM (2008), 3–12.
39. Singh, M., and Hoffman, D. Constructing and representing visual objects. *Trends in Cognitive Sciences 1*, 3 (1997), 98–102.
40. Swihart, B. J., Caffo, B., James, B. D., Strand, M., Schwartz, B. S., and Punjabi, N. M. Lasagna plots: a saucy alternative to spaghetti plots. *Epidemiology (Cambridge, Mass.) 21*, 5 (Sept. 2010), 621–5.
41. Wolfe, J., and Bennett, S. C. Preattentive object files: Shapeless bundles of basic features. *Vision Research 37*, 1 (1997), 25–43.