Face to Face: Evaluating Visual Comparison

Brian Ondov, Nicole Jardine, Niklas Elmqvist, *Senior Member, IEEE*, Steven Franconeri

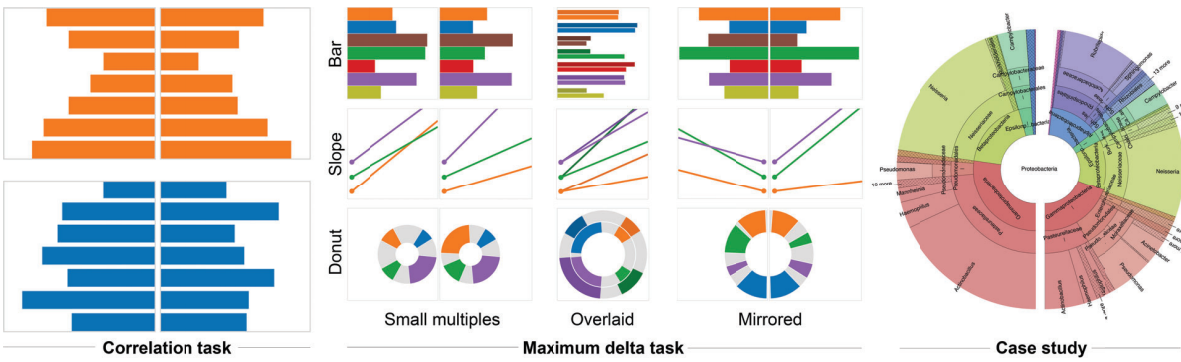


Fig. 1. Evaluation methods for visual comparison. **Left:** Participants were asked to pick the most similar pair of bar charts for a variety of arrangements and degrees of correlation. **Center:** Participants were asked to find the maximum delta, or "biggest mover," between pairs of datasets. Additional arrangements not shown are vertical small multiples and animated transitions. **Right:** Domain experts were interviewed after trying various comparative arrangements in Krona, an interactive sunburst display for biological data.

Abstract—Data are often viewed as a single set of values, but those values frequently must be compared with another set. The existing evaluations of designs that facilitate these comparisons tend to be based on intuitive reasoning, rather than quantifiable measures. We build on this work with a series of crowdsourced experiments that use low-level perceptual comparison tasks that arise frequently in comparisons within data visualizations (e.g., which value changes the most between the two sets of data?). Participants completed these tasks across a variety of layouts: overlaid, two arrangements of juxtaposed small multiples, mirror-symmetric small multiples, and animated transitions. A staircase procedure sought the difficulty level (e.g., value change delta) that led to equivalent accuracy for each layout. Confirming prior intuition, we observe high levels of performance for overlaid versus standard small multiples. However, we also find performance improvements for both mirror symmetric small multiples and animated transitions. While some results are incongruent with common wisdom in data visualization, they align with previous work in perceptual psychology, and thus have potentially strong implications for visual comparison designs.

Index Terms—Graphical perception, visual perception, visual comparison, crowdsourced evaluation

1 Introduction

While the visualization designer has myriad ways to represent information graphically, experimental evaluation has shown us that not all representations are equal [9, 19, 33]. These perceptual studies are often motivated by tasks that are typical for analyzing a single data series, e.g. averages, trends, extreme values, and outliers [13]. When comparing more than one dataset, however, the goals of the visualization can be fundamentally different [26]. For example, instead of looking for the largest or smallest data point, we may look for the largest *delta* from one set to another [52], or for an overall level of correlation [4]. While many of the perceptual lessons learned from single series no doubt extend to these tasks, introducing comparison can tax a substantially capacity-limited aspect of our visual system [25].

We present a series of graphical perception experiments designed to evaluate designs for visual comparison tasks. We choose two primitive tasks specific to the goals of comparison: (1) identification of a maximum delta (or "biggest mover") between data series, and (2) estimation of overall correlation between two series. We embed Task 1 in various stimuli (Figure 1, center): (a) length, represented as bar charts, (b) slope, represented as simple line graphs, and (c) angle, represented as donut charts. We embed Task 2 in a forced-choice between two pairs of bar charts (Figure 1, left). For each embedding, we explore performance of 5 layouts: (i) 'stacked' small multiples with a common baseline, (ii) 'adjacent' small multiples with a non-common baseline [62],[1] (iii) superposition, or 'overlaid' charts, (iv) adjacent small multiples that are mirror symmetric, and (v) animated transitions. The first three of these are commonly used and are associated with intuitive—but rarely measured—differences in efficacy [37]. The last two are less common but may leverage the visual system's sensitivity to motion [47], and in particular common motion [41], in addition to the sensitivity of the visual system to mirror symmetry of objects [65], making them valuable to evaluate.

Many of our results confirm prior expectations for common layouts (overlaid > horizontal multiples > vertical multiples). Surprisingly, however, in some cases we also find significant performance improvements when arranging small multiples in a mirror-symmetric fashion. Furthermore, counter to many prior studies, we observe animation having high performance for the biggest mover task. To validate our results in a more realistic setting, we also present a practical application of both animation as well as symmetry via mirroring in a visual comparison task for a taxonomic hierarchy browser, called

- *Brian Ondov is with the National Institutes of Health in Bethesda, MD, USA and University of Maryland in College Park, MD, USA. E-mail: ondovb@umd.edu.*
- *Nicole Jardine and Steven Franconeri are with Northwestern University in Evanston, IL, USA. E-mail: {nicole.jardine, franconeri}@northwestern.edu.*
- *Niklas Elmqvist is with University of Maryland in College Park, MD, USA. E-mail: elm@umd.edu.*

Manuscript received 31 Mar. 2018; accepted 1 Aug. 2018.
Date of publication 16 Aug. 2018; date of current version 21 Oct. 2018.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TVCG.2018.2864884

[1] We only examine a subset of (i) and (ii) for donut charts, as they have no inherent orientation. However, recent work on performance assymetries between vertical vs. horizontal display layouts [7, 44] suggests that this case merits future study.

Krona [48], based on sunburst displays [58] (Figure 1 right).

Our contributions are the following: (i) results from two graphical perception experiments measuring participant performance for (a) a maximum delta task in bar, slope, and donut charts, and (b) a correlation task for bar charts; (ii) data generation procedures designed specifically for graphical perception studies on visual comparison; and (iii) a practical application of animation and mirroring to Krona.

These results serve both to bolster existing guidelines with empirical evidence and to suggest reexamination of seldom used layouts.

2 Background and Related Work

It is not enough to make visualizations that are pleasing or engaging—empirical evaluation is a crucial part of the analytical process [63]. Cleveland & McGill informed decades of design by ranking basic visual channels by their quantitative accuracy [19]. Specific visual faculties, like the detection of outliers and salient elements, have been also been well studied [23, 36, 61, 66], and the widespread application of color theory to visualization has helped designers avoid skewed interpretations [10, 55, 70]. These types of studies typically involve relationships within a single data series, with tasks such as estimating size differences [68] or determining if points in the series are equal [2]. Often, however, real data are not so simple, requiring more complex comparisons across multiple series [27].

Expanding from a single data series to multiple constitutes a multivariate analysis, i.e. adding rows to a table in Bertin's synoptic [8]. Comparative visualization can be thought of as multivariate analysis in which a *categorical* variable is used to slice the data. For example, we may want to compare time series of the popularity of various baby names or the prices of a variety of goods in different countries. The goals of comparison are often different than those of single-series analysis and can be described as compounds of more primitive tasks [3]. Gleicher et al. provide taxonomies of tasks, as well as comprehensive reviews of techniques and best-practice guidance [26, 27]. While these reviews provide valuable intuition about the efficacy of various comparative strategies, quantitative user studies are less common in this area. Qu et al. explore the importance of consistent scale and coloring across small multiple displays, but not the efficacy of the arrangements themselves [49, 50]. Roberston et al. compare animation to a relatively high number of small multiples (8 to 80) for conveying trends in GapMinder data [1, 52]. Heer et al. compare variants of time-series representations within the context of vertical juxtaposition [34]. Javed et al. also evaluate various methods of displaying multiple time series and include both juxtaposition and superposition, but with tasks similar to those of single-view evaluations [38]. We build on these studies, taking inspiration from perceptual psychology research that is relevant to visual comparison.

3 Perceptual Factors in Comparison

We weighted three themes from the perceptual psychology literature in considering which comparison arrangements to evaluate.

3.1 Co-location

Within a single region of space, visual features such as length, orientation, and motion can rapidly convey information about stimulus deltas. Comparison between two regions is a more difficult task for the observer, because it may require an active process of storage of one region before being able to compare it with another region. "Spot the difference" games, in which observers try to detect small changes between two otherwise identical images, illustrate the difficulty of this task. Mental storage capacity, even for basic visual features like shapes and colors, is around four at maximum [12], and observer comparisons between mentally stored features and currently visible features may be subject to multiple bottlenecks [56]. Detecting a difference between two sets of data may only be possible for large change sizes, even for small datasets (e.g., 5-10 values).

3.2 Symmetry

An additional consideration for multiple displays is that the human visual system is sensitive to symmetry, and especially mirror symmetry located at the focal point [39, 65]. Specifically, the system's ability to detect visual differences is more efficient between two regions that are otherwise mirror images of each other, compared to repeated translations of each other [6, 60] and when the symmetric information is spatially close rather than far [20]. Juxtaposed datasets (e.g. small multiples) are typically translated horizontally, and with common axial directions in order to reduce the cognitive burden of understanding the different polarities of each side of the horizontal axes [26]. But mirror symmetry is occasionally used when comparing two data series that are similar, for example in population pyramids [40], suggesting that designers have an implicit awareness that this arrangement may hold benefits. We hypothesize that advantages for human symmetry detection could convey benefits for comparisons of data in mirrored arrangements.

3.3 Movement

Motion is a primitive and fundamental element of vision [47]. Estimates of velocity can originate in the retina itself [28], and at higher levels of visual processing motion can be used to extract statistics and structure from scenes [30, 41], and may be a useful cue for statistical extraction of patterns in data visualizations [59].

But motion processing is not all-powerful. In particular, when a viewer is asked to process multiple moving objects simultaneously, performance can fall drastically for more than 2-4 objects [53, 67]. When used to demonstrate processes in diagrams in teaching, its use can confuse students [64],

Evidence for the usefulness of motion in visualization is early and mixed. Animation can fill a wide variety of roles and may have similarly varied utility [18], and has shown promise in the role of maintaining context during configurational changes [5, 22, 29, 35]. Because the visual system encodes motion speed and direction as a primitive and direct feature [47] similar to orientation or length, it may be especially useful for detecting changes to values, because larger changes should co-vary with motion speed, and change direction with motion direction. Prior studies have assigned animation questionable value in similar tasks, for example when conveying correlation via oscillation [42], conveying trends in time series [52], or linking two views in a scatterplot [17]. However, these are specific instances among a wide variety of possible tasks, graphical representations, and layouts.

4 Methods

To investigate our hypotheses, we evaluated performance on two tasks (maximum delta and correlation) across multiple visualization types (bars charts, slope charts, and donut charts; see Figure 4), and arrangements (stacked, adjacent, mirrored, overlaid, and animated; see Figure 2), using a series of crowdsourced experiments. Our goal is to measure the degree of precision with which human observers can make judgments about visualized data. Because response time measures can be contaminated, or at least made noisier, by variance in a participant's level of conservatism for how certain to be before making a response, we instead relied on an accuracy-based method.

4.1 Tasks

We choose two main tasks as case studies for visual comparison: one that simulates the goal of finding single values that have changed, and another that simulates the goal of noting more global similarities between two sets of series data. They are otherwise not intended to be representative of the real-world suite of visual comparison tasks.

- MaxDelta: From one series to the next, which data point changed the most? This could be an increase or decrease, defined by absolute change, as opposed to percent change. Difficulty is increased by reducing the largest delta while increasing distractor deltas, so the maximum is less distinguishable. A bimodal distribution of absolute values decouples the largest delta from the largest or smallest absolute value in any single set (Appendix A.1).
- Correlation: Out of *two pairs* of charts, which pair exhibits the most correlation between its two series? Difficulty is adjusted

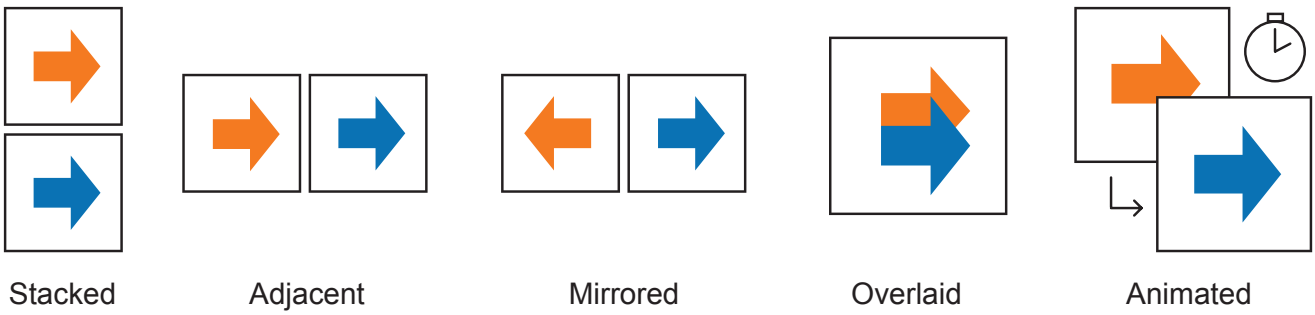


Fig. 2. Comparative arrangement methods examined. The direction of the arrows represents the orientation of the x-axis (or, in the case of donut charts, clockwise versus counterclockwise).

by varying the correlation of the target series pair, while leaving the control pair at a low, fixed correlation (Appendix A.2). Since correlation may be too esoteric of a concept for crowdsourcing, we instructed participants to choose the "most similar pair" and ensured that each chart in a pair had comparable means and standard deviations. See past studies [31, 51] for similar tasks.

4.2 Titer Staircase Method

Our goal was to quantify the magnitude of the stimulus difference required to perform the MAXDELTA task, and the magnitude of the correlation difference for the CORRELATION task, for each arrangement. To do this, instead of pre-selecting and factorially manipulating stimulus magnitudes, stimulus difficulty was titrated dynamically, using a *staircase method*. This method is commonly used in perceptual psychology because it provides a more precise measure of performance: it quantifies a *titer*, a value between 0 and 1 that scales the magnitude of the difference between stimuli to determine the stimulus threshold at which a participant can barely perform a discrimination task. This method is also practical because, by the nature of our tasks and charts, we had no a-priori hypotheses about the stimulus properties that would sufficiently span from lower to higher accuracy across arrangements.

In the MAXDELTA task, the correct answer (true biggest mover) by definition changed more than the distractors (non-biggest movers). *How much* more, however, makes a big difference for task difficulty. The titer controlled how much the correct answer stood out from the distractors in two ways: with a larger titer (easier), the biggest mover moved more, while the distractors moved less; with a lower titer (harder), the biggest mover moved less, while the distractors moved more. At top difficulty, the biggest mover barely moved more than the distractors. In the CORRELATION task, the titer controlled the difference in correlation between the baseline data (0.2) and the test data: larger titers indicate higher correlations of the test pair, and thus larger differences between baseline and test pairs. At the highest titer (easiest) the two series in the test pair were almost identical (Fig. 1, left; orange), making this pair stand out more from the baseline pair.

Titers and stimulus datasets changed trial-by-trial depending on participant performance for the previous trial (Appendix A). For each arrangement, the first trial had a titer of 0.5. Over the remainder of trials, the titer adapted to participant accuracy. An erroneous response made the next trial easier (larger titer), and a correct response made the next trial harder (smaller titer). The titer was increased three times as much following an incorrect answer as it was decreased for a correct answer. This 3:1 modulation of the signal should affect performance accordingly. By the end of the experimental trials, the titer reflects a stable magnitude of signal that allows the participant to perform with 75% accuracy for that layout. See Figure 3 for an example that shows how this method titrates the stimulus based on participant performance.

At the end of the staircase procedure, a larger titer indicates that the conditions of the given experiment made the task more difficult, requiring larger signals (and thus easier trials) to maintain accuracy. We perform within-subjects comparisons of the means of the titer values of the final 5 trials per arrangement.

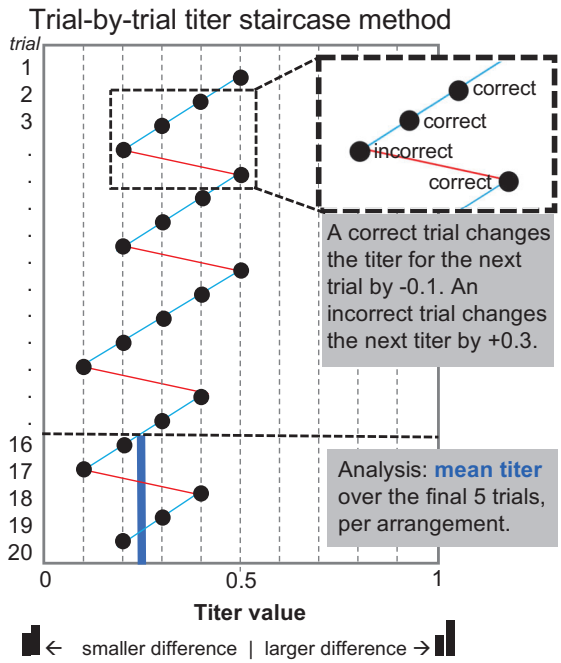


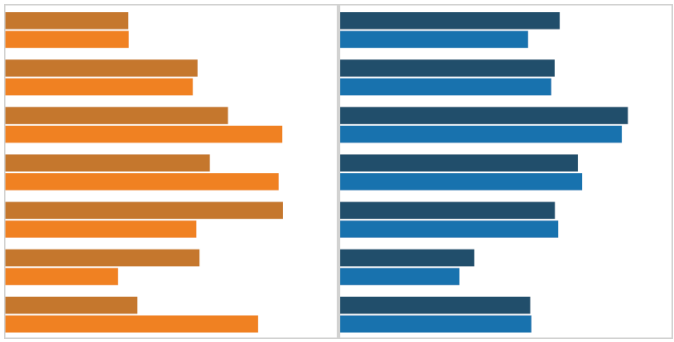
Fig. 3. During titration, the titer value (stimulus signal) increases if an erroneous response is made, and decreases if a correct response is made. Titers are calculated independently for each arrangement, and are analyzed to determine how chart arrangement affected the final staircased titer values.

4.3 Chart Types

Bar charts are versatile and intuitive, making them a natural choice for evaluating both tasks. However, since the choice of visual encoding channel could interact with the choice of arrangement, we also evaluated slope charts and donut charts for the MAXDELTA task, for a total of three chart types. To ensure each chart type provided an appropriate range of difficulty, parameters such as the number of data points had to be adjusted. These parameters were determined during internal piloting, resulting in the following configurations:

- **Bar charts**: (both tasks) Standard charts in which the length corresponds to the datum. Each series contains 7 data points.
- **Slope charts**: (MAXDELTA only) Simplified line charts with just two points in each line, (0 and a generated datum), reducing them to slopes. Each series contains 3 data points.

(a)



(b)

Fig. 4. Overlaid layout for (a) the MAXDELTA task for bar (Exp. 1A), slope (1B), and donut (1C) charts, and (b) CORRELATION task for Exp. 2.

- **Donut charts**: (MAXDELTA only) Rings in which the data are represented by angular sector. For the purposes of experimental control, they differ form standard donut charts in several ways: (i) gray distractors are used as buffers to allow adjacent data to change size while remaining in the same position, (ii) overlaid arrangements, which are non-standard for donut charts were implemented with concentric rings, aligning the centers of corresponding colors, and (iii) mirrored arrangements were implemented by limiting each chart to 180 degrees, allowing the two series to form a complete circle. Each series had 4 data points.

4.4 Arrangements

Each chart type was presented in 5 different layouts (Fig. 2):

- **Stacked**: Vertically arranged small multiples (i.e. one chart is placed above the other). Cleveland & McGill posit the aligned baselines helps judgment [19]; but this design also makes it tougher to find correspondance between paired values from each series [7, 44]. We thus include it as an expected floor to which performance of other arrangements can be compared.

- **Adjacent**: A more commonly used instance of small multiples, in which data series are placed side-by-side, allowing each pair of items to align vertically. This arrangement serves as a more realistic baseline than *stacked*.

- **Mirrored**: This "mirrored" variation of *adjacent* opposes the direction of the x-axis in each chart. For bar charts, this simply amounts to right-aligning the left chart and vice versa. For slope charts, the x-axis is reversed in the left chart, essentially negating the slope. For donut charts, we restrict each series to a semicircle. The Gestalt nature of bilateral symmetry suggests this layout could improve performance versus standard small multiples.

- **Overlaid**: A combined chart depicting both data series within the same space. Past work has claimed that overlaying values, or superposition, minimizes eye movements and memory load, and may lead to efficient comparison [27]. This technique has proven effective in a design study setting [45], but, to our knowledge, not directly confirmed empirically. We expect this condition to serve as a ceiling for performance in the context of the MAXDELTA task (Figure 4).

- **Animated**: In this "arrangement," a single chart is transitioned, or morphed, from one data series to another over time. As all marks transition for the same amount of time, the maximum velocity of a given mark becomes an emergent signal that directly encodes its delta. Movement is broadly processed as a primitive feature in the vision system, suggesting that this signal is potentially beneficial for MAXDELTA task. We used cubic interpolation to ease the transitions [21], so the maximum velocity was reached at the midpoint of the impression time.

4.5 Task and Procedure

Before each trial began, the screen contained a centrally placed fixation dot and outlines of where the charts would appear. Participants clicked a button to start the trial. After a countdown, the visualization appeared for a short, fixed time. For the MAXDELTA task, the time for both static charts and animation was 1.5 seconds. For the CORRELATION task, static charts were shown for 3 seconds, to account for the doubled number of charts, while animation remained at 1.5 seconds to preserve velocity. At the end of the impression, one of the data sets of the comparison was removed, while the other remained.[2]

Participants then clicked on the appropriate portion of the remaining chart or charts to provide a response. For the MAXDELTA task they were instructed to "Click on the bar that had the biggest difference in the two charts"; for CORRELATION to "Click on the color that had the most from similarity before to after." Participants were informed if they were correct and, if incorrect, what the correct answer was. This feedback was provided to make the task more engaging and to reinforce the goal. Between trials, the titer was adjusted based on the response (if incorrect, the titer was made larger for the next trial; if correct, the titer was made smaller). Each participant completed one experiment, each with five arrangements. There were twenty trials for each arrangement (thirty for donut charts), and arrangements were blocked. The order of the arrangement blocks was changed between participants.

4.6 Training

Before training, participants were shown examples of stimuli and the task. Before each arrangement block of trials, participants were given a time-unconstrained version of the task, which they were required to answer correctly before proceeding (once for the MAXDELTA task, 3 times for the CORRELATION task). Additionally, the first non-animated arrangement given to a participant followed untimed training with 3 timed training trials, which were identical to the real trials except that they always had the easiest (largest) titer. Data were regenerated on incorrectly answered training answers to minimize answering by elimination.

4.7 Participant Recruitment

We recruited participants through the Amazon Mechanical Turk Platform. Based on power analyses from initial pilots, we recruited 50 new participants for each experiment. Participants were asked to self-select out of the study if they had color vision deficiencies. Each participant completed all arrangements (4 for donut, 5 for others) for a single combination of stimulus type (bar, slope, donut) and task (MAXDELTA, CORRELATION). Worker IDs were used to ensure uniqueness of participants across all such combinations. All 200 workers recruited for participation in these 4 experiments were adults in the United States. 61.3% were male, 36.4% female, 1.10% other, 0.6% no response. Ages varied from 18-25 (7.7%), 26-40 (60.8%), 41-60 (28.7%), and 61-80 (2.2%).

[2]For slope charts, since the mirror metaphor is more important for interpretation, the x-axis included arrows conveying axial direction during training trials, both timed and static.

5 RESULTS

We evaluated the magnitude of differences required in the data (i.e., titers) for 200 non-expert participants to reliably identify the individual data item with the largest magnitude change (Experiments 1A-C: bar, slope, and donut charts) or identify the overall data series that was more correlated (Experiment 2, bar charts). To preview the findings, these perceptual data largely dovetail with prior recommendations arrangements (overlaid > adjacent > stacked) [27, 7, 44]. But some findings are surprising. First, animation conferred significantly improved detection of the most different data point in bar and donut charts, though not for slope charts. Second, task mattered: animation did not accrue performance benefits for participants detecting the most correlated data set. Finally, for bar charts, horizontally mirrored bar charts afforded better performance in both MAXDELTA and CORRELATION tasks.

5.1 Accuracy-based Outlier Exclusion

After data collection was completed, we assessed each participant's overall proportion of correct trials (in which the biggest-mover data point or most-correlated data series was correctly identified). A participant was excluded if their overall proportion correct was lower than two standard deviations from the mean of other workers, because this indicates that the staircase method failed to allow accurate performance. For Experiments 1A-C and 2, this procedure resulted in exclusion of 3, 4, 2, and 2 participants.

5.2 Titer Analysis

To evaluate whether arrangement affected the precision with which participants could identify the maximum delta, we computed each observer's mean titer values from the final 5 trials for each arrangement. Titers are inversely related to difficulty: smaller titers for a chart arrangement indicate that subtler, rather than larger, differences were required to elicit a mixture of correct and incorrect responses.

5.3 Exp. 1A: Bar charts (MAXDELTA task)

Figure 5 (upper) displays the mean final 5 titer values for Experiment 1A. In bar charts, two patterns in participant titer values were striking. First, the Animated bars outperformed bars that were Overlaid and all other arrangements. Second, within Small Multiples, a Mirrored arrangement is better than a Horizontal or Vertical one.

These observations were validated in a within-subjects ANOVA. Final titer values for bar charts were affected by arrangement, $F(2.98, 137.23) = 103.23$, $p < .001$, $\eta_p^2 = 0.69$, Greenhouse-Geisser corrected for violations of sphericity. Planned comparisons assessed pairwise differences between arrangement types. Titers for animated bars were significantly more precise than those that were overlaid, $t(46) = 3.42$, $p = .001$. Participants also achieved more precise titer values with horizontally mirrored small multiples compared to non-mirrored small multiples that were horizontally arranged, $t(46) = 2.73$, $p = .009$, and vertically arranged, $t(46) = 4.82$, $p < .001$.

5.4 Exp. 1A Floor Effect

For the final five trials, accuracy was low for stacked (57%), adjacent (61%), and mirrored (64%) arrangements, with large titers near the maximum titer of 0.75 (0.68, 0.64, and 0.59, respectively). By comparison, for animated arrangements, accuracy was 74.6% and the mean titer was 0.35. Participants reached the maximum titer on 28% of stacked trials and 15% of adjacent trials. By comparison, the maximum titer was reached on 6% of mirrored trials, 5% of overlaid trials, and 0% of animated trials. The histograms in Appendix C illustrate titer distributions for all trials for each arrangement. These floor effects suggest that for stacked and adjacent charts, subjects reached the artificial floor (max titer) and continued making errors without subsequent adjustments to the titer value, such that their final titer value reflects not their ability to do the task but the capped titer value. As such, Experiment 1A is not able to quantify the true floor of performance for these arrangements.

Note that this floor issue is unavoidable for many tasks. One solution for future research is a longer display time, but that could make more effective arrangements (e.g. overlaid) too easy, resulting in a ceiling effect and preventing comparison. Another solution is to conduct secondary tests of arrangements that are close in performance, using combinations of titer ranges and timings that best drive apart performance.

In summary, although this data set cannot be appropriately used to directly compare the mean titers between stacked and adjacent arrangements, it is clear that the MAXDELTA task was highly difficult in stacked and adjacent bar charts.

There was no evidence for floor effects in subsequent experiments.

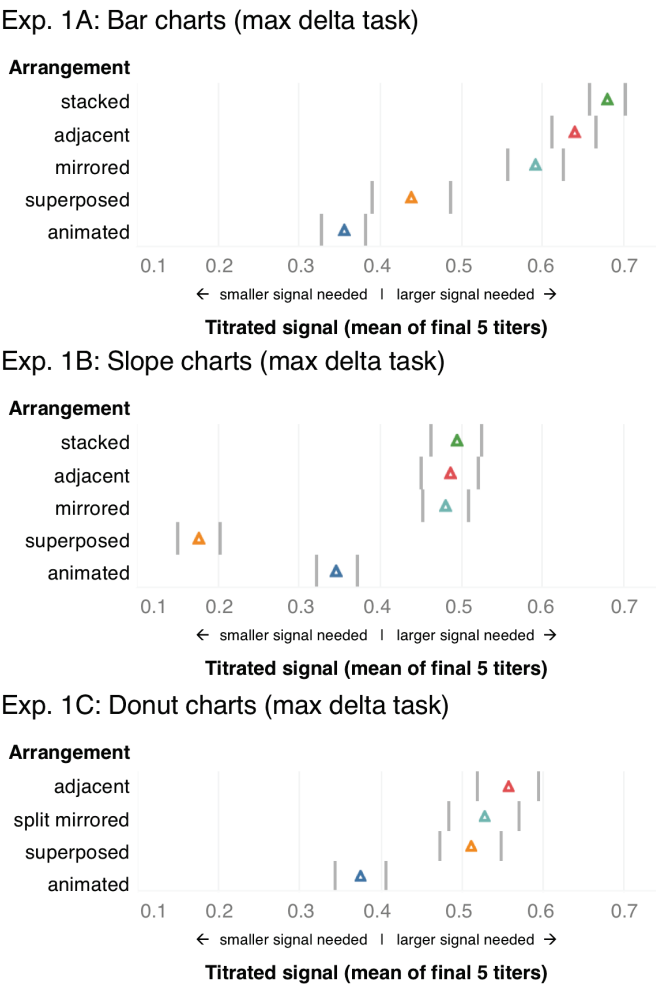


Fig. 5. Mean of final 5 titer values across participants performing the MAXDELTA task. Gray bars represent 95% confidence intervals.

5.5 Exp. 1B: Slope charts (MAXDELTA task)

In slope charts, titer values were generally more precise and there were slightly different observations as a function of arrangement. First, Overlaid slopes outperformed all other arrangements (including Animated). Second, different types of Small Multiple arrangements did not yield differing titer values (Fig. 5, center).

These observations were validated in a within-subjects ANOVA. Final titer values for slope charts were affected by arrangement, $F(4, 180) = 101.87$, $p < .001$, $\eta_p^2 = 0.69$. Titer histograms did not indicate floor effects. Planned comparisons assessed pairwise differences between arrangement types. Titers for overlaid slopes were significantly more precise than those that were animated, $t(45) = 10.13$, $p < .001$. There was no evidence that participants achieved more precise titer values with horizontally mirrored small multiples compared to non-mirrored small multiples that were horizontally arranged, $t(45) = .25$, $p = .8$, or vertically arranged, $t(45) = .77$, $p = .45$. Ac-

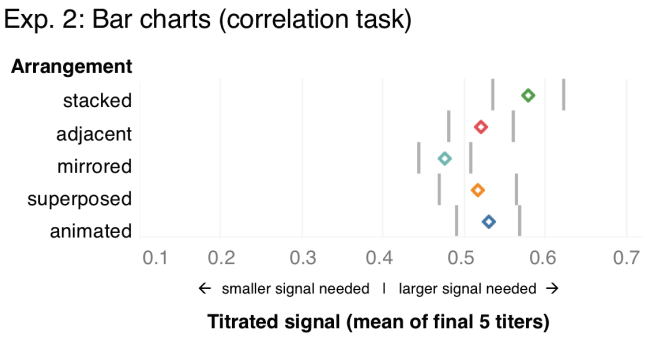


Fig. 6. Mean of final 5 titer values across participants performing the CORRELATION task with bar charts (Experiment 2). Gray bars represent 95% confidence intervals.

curacy exhibited similar patterns as titer values.

5.6 Exp. 1C: Donut charts (MAXDELTA task)

The mean final 5 titer values for donut charts were affected by arrangement, $F(3,141) = 22.96$, $p < .001$, $\eta_p^2 = 0.33$ (Fig. 5, lower). Titer histograms did not indicate floor effects. Animated donuts outperformed all other arrangements for the max-delta task. There was no evidence that the split mirrored hemifields arrangement outperformed the horizontal small-multiple donuts, $t(47) = 1.26$, $p = .21$. Accuracy exhibited similar patterns of titer values.

5.7 Exp. 2: Bar charts (CORRELATION task)

For this task, participants saw two pairs of bar charts (Fig. 1, left, Fig. 4(b)) and selected which pair had the highest correlation between its two series of values. One pair was fixed at a Pearson's R correlation of 0.2, and the other pair was titrated such that its correlation became closer to 1.0 (higher titer) or closer to 0.2 (lower titer; see Methods).

The mean final 5 titer values for Experiment 2 were affected by arrangement, $F(3.22, 144.95) = 6.50$, $p < .001$, $\eta_p^2 = 0.13$, with no indication of floor effects (Fig. 6).

In contrast to Experiment 1, it is apparent there was no benefit of animation over other arrangements: participants struggled to use motion to extract and compare correlations between data sets. Observer performance had resulted in staircasing of the mean correlation (Pearson's R) to 0.74 for observers to reliably choose it over the base pair correlation of 0.20.

We conducted planned comparisons to assess whether mirrored small multiples yielded more precise titers than the other small multiple arrangements. Participants achieved more precise titer values with mirrored compared to adjacent arrangements, $t(45) = 2.13$, $p = .04$. They were able to perform correlation comparison when the target correlation was 0.70 in mirrored charts, but needed a correlation of 0.75 for the same performance in adjacent chart arrangements. Adjacent bar charts outperformed stacked ones, $t(45) = 3.31, p = .002$, such that for these trials the correlation of the correct pair was 0.82 for stacked charts.

Accuracy exhibited largely similar patterns as titer values, with the exception that there was only marginally significantly higher accuracy for mirrored compared to adjacent charts, $t(45) = 1.95$, $p = .057$.

6 DISCUSSION

In the MAXDELTA task of Experiments 1A-C, in which participants identified the data series that had the most substantial change relative to other data series, animated charts consistently outperformed all small multiple arrangements. Findings were mixed for overlaid visualizations: they outperformed all other arrangements (including motion) for slope charts, were better than any arrangement of multiple bar charts, and did not seem to confer strong benefits over small multiple arrangements. Finally, mirrored small multiple arrangements marginally allowed participants to better identify the max-delta series (compared to other horizontal arrangements) only in bar charts. Although animated charts outperformed others for the goal of the MAXDELTA task, and as such is useful if an analyst's goal is to rapidly identify individual data points with the largest improvement or impairment, it might not be an optimal encoding for other goals of the observer or designer. Specifically, a maximum delta task may be a special case in which velocity information directly encodes individual data deltas but does not directly encode the visual information that observers use to inform other judgments, such as the overall correlation or mean. In Experiment 2, participants judged which of two pairs of charts had the most correlated data set. Here, animation did not lend itself to detection of correlation. Instead, mirrored bar charts outperformed all other arrangements for detection of correlated data.

6.1 Limitations

There are several limitations that should be taken into account when assessing our results. First of all, we limited our study to two data series at a time, which provided consistency across arrangements and allowed for a tractable number of combinations. In reality, however, comparative techniques often must be applied to three or more sets, adding another potential factor to the equation. This has additional repercussions for a few of the visualizations in particular: For animated charts, there is clearly an upper bound on the number of animations that a person can meaningfully perceive at the same time [53, 67]. In other words, the high performance exhibited by the animated chart may not generalize past a few charts. It is possible that animated charts did less well for the correlation task in Experiment 2, which involved two simultaneous animations, because of this limitation.

Mirroring implies only two datasets at a time, which means that this arrangement is unlikely to scale. One can imagine four data sets that are horizontal and vertical mirrors of each other. One way to scale up this method would be to test different arrangements of quadrants of data sets in biggest mover, correlation, and other tasks. Juxtaposed (adjacent/stacked) as well as overlaid arrangements could probably involve a higher number of datasets. However, as the number of simultaneous datasets increases, increasing screen distances between compared values will likely start to have an impact. Our study did not test any of these factors.

As with any graphical perception study, there are limitations how these results generalize to actual visualization tasks. First of all, the biggest mover and correlation tasks we studied here may not be representative of high-level visual comparison tasks that users of data visualization actually perform in practice. While we believe that our low-level tasks are building blocks for such higher-level tasks, we can only motivate this with intuition. Second, even if our tasks do generalize in this manner, it is not clear that measuring low-level difficulty translates to better sensemaking performance. Sometimes, spending more time to get a more accurate answer may be necessary for a specific task; sometimes it may be the inverse. Third, our study only involved three representations: bar charts, slope charts, and donut charts. While we think these are reasonably representative of many visualizations used in practice, we again do not claim to be exhaustive. Finally, our data generation process is highly controlled and may not generalize to real data. This is especially true of the MAXDELTA task, in which animated performance clearly benefits from a lack of distractions. Though CORRELATION data are more randomly distributed, we still constrain means and standard deviations to isolate similarity. In reality, a variety of statistics of interest may vary across data series.

6.2 Implications for Visualization

The present results do not suggest an easy set of guidelines that specify a best arrangement or encoding for a given task. Instead, they suggest that we are unlikely to discover such simple rules. This is not surprising if one considers that the mapping from raw data to visual comparison is mediated by a suite of visual features that the eye happens to extract and compare, and that set is determined by how the visual system evolved and develops to perceive the natural world. To produce guidelines, it will likely be necessary to empirically evaluate

multiple tasks, across several types of visualizations, and several types of arrangement designs. We suspect that some general guidelines will arise, but each will have exceptions. But if this effort is combined with attention to the visual features that could mediate [69] comparison judgments in each case (e.g., motion signals for animation; serial inspection for stacked arrangements), the inclusion of those features in a model may facilitate the generation of concrete guidelines.

We therefore consider a primary contribution of this paper to be a demonstration of a method for empirically evaluating a given arrangement design, for a given task, in a given visualization. Contrary to widely held belief, we found animation to be extremely effective for conveying salient differences between two datasets. While this is a highly controlled instance, and thus may not extend to more complex or dynamic data, our experiments at least show that animation can be a valuable addition to the designer toolbox when the emphasis of differences is a priority. However, animation is not always a feasible method of comparison, because its ephemeral nature requires constant interaction and attention. If, instead, a small multiple display is desirable, and it is only necessary to compare two datasets, our results support the intuition that center-aligning horizontal, space-filling comparisons inform the selection of axial directions to maximize preattentive detection of salient differences. Perhaps the most basic case of mirroring, in which bar charts are simply center-aligned, is already in use and has intuitive advantages. However, we provide here, for the first time, an experimental validation of the utility of this practice. Further, our results align with a hypothesized mechanism based on current understanding of the perception of symmetry.

7 Case Study: Microbiome Comparison in Krona

While highly controlled experimental conditions are crucial to empirical evaluation, they can often be at odds with the ecological validity of the results [24]. In this case, for example, our studies show that both mirror symmetry and animation can be beneficial in certain, specific contexts, but do those benefits extend to applications in the real world? While it is nearly impossible to capture both aspects in the same study, we sought to answer this question by complementing our controlled study with a more ad hoc review by expert users.

7.1 Background

For an application, we chose exploration of the human microbiome, or the communities of microorganisms that live in and on us. This domain is an extremely challenging one for visualization and an area of active development and interest. Since a community of organisms can be described at various levels of taxonomic granularity (i.e. genus, species, etc.), even single datasets are complex and challenging to represent. Various hierarchical techniques have been employed for the task, including Sunburst charts (as in Krona), Treemaps [54] (as in MetaTreeMap [32]), and Sankey/flow diagrams [15] (as in Pavian [14]). However, in each case, additional variables, such as change between datasets, are difficult to introduce. For scientific data, which often have control groups, the comparison of multiple data series is nonetheless critical to making sense of the underlying information.

7.2 Method

Echoing our empirical studies, we adapted Krona, which already supported transitions similar to the *animated* arrangement, to implement two additional comparative strategies reviewed above:[3] *adjacent* and *mirrored* (Fig. 7). We introduced the three techniques to two scientists studying the microbiome at the National Human Genome Research Institute in Bethesda, MD, USA. Both had prior experience with the tool for exploration of single datasets.

7.3 Task

The main goal in exploring this type of data, as stated by the experts, is to find significant differences in the fractions of particular organisms, especially if they are pathogenic ones. Inherent in this task is searching for the largest delta or deltas in any given view, making it an appropriate setting for the MaxDelta task in our earlier experiments. The participants were presented with real data comparing human skin microbiomes from two time points ("M3 skin" days 0 and 1) [16]. The charts show the relative proportion of various species within each time point as well as the aggregated proportions of more general taxa. We asked the microbiologists to find significant differences between the same two time points using all three arrangements (*animated*, *adjacent*, and *mirrored*). For example, in Figure 7(a), *Gammaproteobacteria* (red wedges) decreases a large amount from day 0 to day 1 (left to right), but looking more specifically within this group, *Pseudomonas* actually increases. Rather than seeking a specific set of correct answers, however, we instead gathered more qualitative feedback about performing the task under the various conditions.

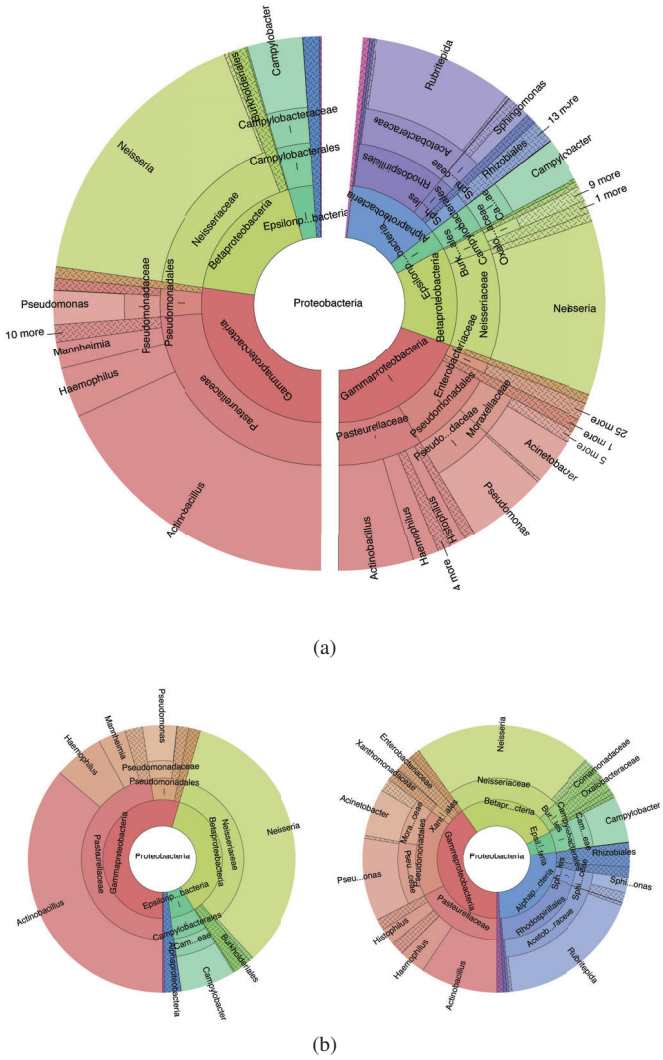


Fig. 7. Comparative modes implemented in Krona, an interactive sunburst display commonly used for biological data. Here, the skin microbiome of an individual is represented across two time points. Higher levels (i.e. innermost rings) represent more general taxonomic categories, allowing differences to be observed across various levels of granularity. In (a), a single circle is split to provide mirror symmetry, corresponding to the *mirrored* arrangement in the above experiments. In (b), the standard small multiple view of the same data is shown, corresponding to the *adjacent* arrangement.

[3]Note that we did not include overlaid as it is not clear how to implement this type of arrangement for sunburst charts, which are already nested hierarchically.

7.4 Results

Consistent with the results of our perceptual studies, both participants found that animation made differences particularly salient. However, they also noted that, if the change was large, it shifted the other wedges in a disorienting way. This is because, unlike in those studies, the positioning of the wedges could not be controlled using distractors. This is an example of a caveat of extending the lessons of the perceptual studies to a more ecological valid environment. However, it also could suggest work to be done to take advantage of the benefits of animation seen in those studies—for example, perhaps wedge ordering could be optimized to minimize offsetting during a transition, as has been done for stability in Treemaps [57].

It was also noted by the experts that animation could be engaging for an audience when highlighting a specific difference, reiterating the findings of Robertson et al. [52]. However, both participants preferred static views when performing their own exploration or investigation. One participant preferred small multiples due to its consistency with standard sunburst charts and the ability to represent more than two samples. The other, however, preferred the mirrored split view due to the better use of space and smaller eye travel distance when making direct comparisons between constituent taxa. Additionally, the case study illuminated practical considerations of implementing these arrangements. For example, the experts pointed out that small multiples may be ideal for dissemination, which is often static and must reach a wide audience that may not be familiar with the split mirrored view.

While this case study did not necessarily suggest an ideal arrangement, it did help to bridge our empirical results to a more realistic setting. Unsurprisingly, there was a consensus that each method had strengths and weaknesses, and would be more appropriate for specific contexts. If any conclusion can be made, it is perhaps that this platform, and others, should have the flexibility to support many layouts, allowing the user to switch between them to aid the task at hand.

8 Conclusion and Future Work

The present results merely scratch the surface of potential insights to be gained through empirical evaluation of combinations of visual tasks, visualization types, and comparison arrangement designs. Even that large testing space does not capture all of the important potential questions. For example, the advantages of mirror symmetry could depend on other factors, such as displacement of the values being compared, mark shape, axial distance, and the viewer's comprehension of the mirror metaphor. Further, while it was crucial for us to choose operationalized tasks to perform quantitative analyses, real users of comparative visualizations can have multiple and nuanced goals, making it important to explore more varied assays of visual efficacy.

Using very similar methods, reasonable next steps might include asking participants to perform tasks more akin to what analysts do with real data sets, such as selecting the chart with the highest overall mean value (similar to the MAXDELTA task) or the most consistent change between pairs (similar to CORRELATION). These methods could be used to match efficacies of arrangements. For example, the results of Experiment 1A suggest that in cases when animated charts are not possible (such as for handouts), the best static alternative is one that is overlaid—at least for the purposes of a max delta task. The factors we consider are also by no means the only, or even necessarily the most important, elements of vision that could impact comparative tasks. One could even test how multiple factors combine in a single arrangement, for example animation in the context of mirrored displays.

While the possibilities for additional study are so myriad as to be somewhat daunting, we hope that the initial excursion presented here can serve both as a template and motivation for discoveries to come.

Acknowledgments

We thank Adil Yalcin of Keshif, LLC for advice and assistance with the use of the Amazon Mechanical Turk platform, Catherine Plaisant for feedback on experimental design and usability, and our peer reviewers for helpful suggestions. Brian Ondov was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the respective funding agencies.

A Data Generation

Data were generated dynamically to allow real-time difficulty adjustments. Each task required its own strategy for randomization, but both were parameterized by a titer value, which represented a relative difficulty level, with higher values being easier. The STAIRCASE procedure takes a number of trials until performance stabilizes around a titer value. To determine how many trials would be suffice for titer stability across arrangements for most participants, we conducted pilot tests and analyzed the standard deviations of titer values, per arrangement, in bins of every 5 trials. Standard deviations will be highest while the titer is unstable, and decrease until reaching a stable plateau. This method suggested that 20 trials for the MAXDELTA task and 30 for CORRELATION would suffice so that the final 5 trials of reflect stable titers.

A.1 MaxDelta Task

A pair of datasets with controlled deltas was generated by varying points of one dataset to create another. However, simply increasing or decreasing one data point more than others—out of, say, of a normal distribution—would make it much more likely to be the largest or smallest, circumventing the task. It was thus necessary for proper evaluation of the task to devise a novel data generation algorithm. Our method creates a bimodal distribution corresponding to the two extremes of a chosen maximum delta, ensuring that these points are well masked by other data. The magnitude of this delta, and thus the difficulty of the task, is controlled parametrically by the titer value provided to the generation algorithm. In addition to changing the maximum, changing the titer also changes deltas of distractors. At the minimum (smallest difference) titer, every data point is changed a small, equal amount (note that it is, by design, impossible to do better than chance at this level, and in practice it is never reached). At the maximum (largest difference) titer, there are only two possible values for the data points—the maximum uses both, while the others stay at one and do not change at all. The data generation routine is depicted at a high level in Algorithm 1. In summary, for a given titer value t, the biggest mover will change by t times the chart's range (from minimum value to maximum value). The biggest moving distractor will change by $1-t$ of that, the next biggest moving distractor $1-t$ of the first distractor, and so on. For example, at a titer of 0.75, the delta of the biggest mover will cover $3/4$ the full range of the chart, the delta of the first (randomly placed) distractor will cover $0.75 \times 0.25 = 3/16$, and that of the next will cover $0.75 \times 0.25 \times 0.25 = 3/64$. The outputs of this algorithm were linearly transformed as appropriate for the stimuli, e.g. to add minimum width to bars. Though higher titers should always be easier, in practice, we found that difficulty increased above 0.75 due to alignment of bars. We thus capped the titer at 0.75 to prevent participants from getting stuck in a valley of (ostensibly) low difficulty. We confirmed the regularity of the data before the experiment by running multiple iterations of the data generation routine and observing the ordinal ranking of the answers among the distractors. While there do appear to be areas of bias, we deem it highly unlikely that detection of these patterns would be easier for a participant than performing the task as intended.

A.2 Correlation Task

Randomized pairs of series with given correlations were created using simulated annealing in an algorithm inspired by Matejka and Fitzmaurice [43]. Means and standard deviations were fixed within 10 percent of the range to ensure correlation was analogous to "similarity", as described in the instructions. Correlation between the series was calculated using Pearson's correlation coefficient and transformed according to the optimal formula for perceptual estimation according to Rensink & Baldridge [51]. Titers we report for this experiment thus correspond to $g(r)$ in Equation 7 of the latter study.

Algorithm 1 Max-delta data generation

```
1:  procedure MaxDelta(c, t)                    ▷ c:=cardinality, t:=titer
2:      a ← [], b ← []
3:      for i = 0 to c − 1 do
4:          r ← rand()                          ▷ r ∼ U, r ∈ ℝ, 0 ≤ r ≤ 1
5:          x ← t · sqrt(r / (c − i))
6:          y ← x + t(1 − t)^i
7:          if i%2 == 1 then
8:              x ← 1 − x
9:              y ← 1 − y
10:         if rand() < 0.5 then
11:             push a, x
12:             push b, y
13:         else
14:             push a, y
15:             push b, x
16:     return a, b
```

B Rendering

Charts were rendered in the participant's web browser in real time using the D3 [11] JavaScript library. Each individual chart (that is, for a single data series), had a square dimension of 256 pixels for all MaxDelta experiments. For Correlation experiments, renderings with four charts used a square dimension of 200 pixels, while renderings with two charts used a square dimension of 141 pixels (producing equivalent total chart area). Note that the actual number of screen elements corresponding to a "pixel" can vary with hardware configuration, due to the advent of HDPI (high dot-per-inch) displays. Bar charts would not be affected by this variable because of their orthogonal nature, and we chose sufficient line thickness to mitigate the effect for slope charts. Subsets of the Tableau 10 [46] were chosen to maximize (qualitatively) perceived uniqueness; 7 for bars, 3 for slopes, and 4 for donuts. For the overlay arrangement, the saturation and luminance of each color were slightly reduced in one dataset to distinguish adjacent elements. Other arrangements kept the original colors consistently across pairs of data sets. All charts were drawn on white backgrounds, with faint gray boundaries delimiting the chart areas. The web page automatically initiated full-screen browsing mode to avoid distraction during the study, though the persistence of this state was not enforced programmatically.

C Titer Frequencies for Experiment 1A

Frequency counts of titers (across all trials) by arrangements from Experiment 1A are shown in Figure 8, for all non-excluded participants. Participants disproportionately reached the maximum titer value (0.75) for stacked (vertical small multiple) and adjacent (horizontal small multiple) arrangements.

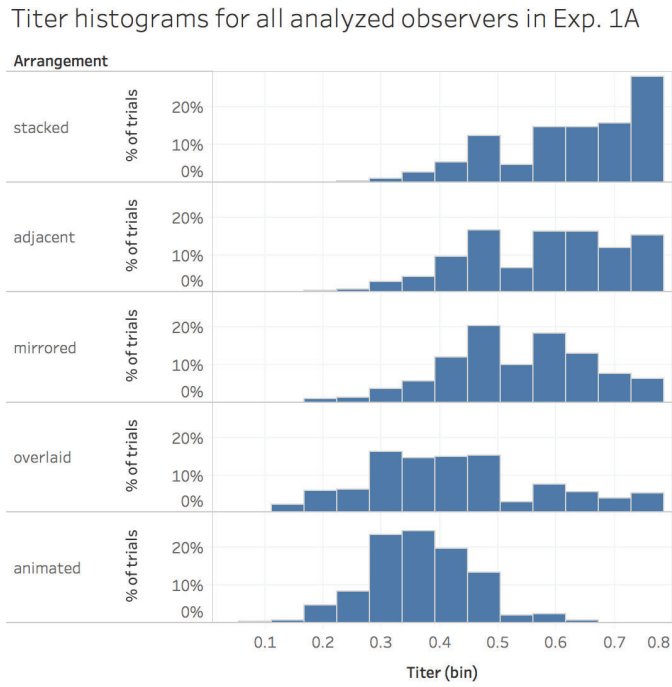


Fig. 8. Titer histograms for Experiment 1A.

References

[1] Gapminder. http://www.gapminder.org.

[2] D. Albers Szafir. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24:392–401, 2017. doi: 10.1109/TVCG.2017.2744359

[3] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *Proceedings of the IEEE Symposium on Information Visualization*, pp. 111–117. IEEE Computer Society, Washington, DC, USA, 2005. doi: 10.1109/INFOVIS.2005.24

[4] R. A. Amar and J. T. Stasko. Knowledge precepts for design and evaluation of information visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):432–442, July 2005. doi: 10.1109/TVCG.2005.63

[5] D. Archambault, H. Purchase, and B. Pinaud. Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):539–552. doi: 10.1109/TVCG.2010.78

[6] H. B. Barlow and B. C. Reeves. The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research*, 19(7):783–793, 1979. doi: 10.1016/0042-6989(79)90154-8

[7] A. Barnas and A. Greenberg. Visual field meridians modulate the reallocation of object-based attention. *Attention, Perception, & Psychophysics*, 78(7):1985–1997, 05 2016.

[8] J. Bertin. *Graphics and Graphic Information Processing*. Walter de Gruyter, Berlin, 1981.

[9] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, Madison, Wisconsin, 1983.

[10] D. Borland and R. M. Taylor II. Rainbow color map (still) considered harmful. *IEEE Computer Graphics & Applications*, 27(2):14–17, Mar. 2007. doi: 10.1109/MCG.2007.323435

[11] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, Dec. 2011. doi: 10.1109/TVCG.2011.185

[12] T. F. Brady, T. Konkle, and G. A. Alvarez. A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of vision*, 11(5):4–4, 2011.

[13] M. Brehmer, J. Ng, K. Tate, and T. Munzner. Matches, mismatches, and methods: Multiple-view workflows for energy portfolio analysis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):449–458, 2016. doi: 10.1109/TVCG.2015.2466971

[14] F. P. Breitwieser and S. L. Salzberg. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *bioRxiv*, 2016. doi: 10.1101/084715

[15] W. C. Brinton. *Graphic Presentation*. Brinton Associates, New York, New York, 1939.

[16] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, J. I. Gordon, and R. Knight. Moving pictures of the human microbiome. *Genome Biology*, 12(5):R50, May 2011. doi: 10.1186/gb-2011-12-5-r50

[17] F. Chevalier, P. Dragicevic, and S. Franconeri. The not-so-staggering effect of staggered animated transitions on visual tracking. *IEEE transactions on visualization and computer graphics*, 20(12):2241–2250, 2014.

[18] F. Chevalier, N. H. Riche, C. Plaisant, A. Chalbi, and C. Hurter. Animations 25 years later: New roles and opportunities. In *Proceedings of the ACM Conference on Advanced Visual Interfaces*, pp. 280–287. ACM, 2016.

[19] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation and application to the development of graphical methods. *Jour-*

nal of the American Statistical Association, 79(387):531–554, Sept. 1984.

[20] M. C. Corballis and C. E. Roldan. On the perception of symmetrical and repeated patterns. *Perception & Psychophysics*, 16(1):136–142, Jan 1974. doi: 10.3758/BF03203266

[21] P. Dragicevic, A. Bezerianos, W. Javed, N. Elmqvist, and J.-D. Fekete. Temporal distortion for animated transitions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 2009–2018. ACM, 2011. doi: 10.1145/1978942.1979233

[22] F. Du, N. Cao, J. Zhao, and Y.-R. Lin. Trajectory bundling for animated transitions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 289–298. ACM, New York, NY, USA, 2015. doi: 10.1145/2702123.2702476

[23] J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458, 08 1989.

[24] N. Elmqvist and J. S. Yi. Patterns for visualization evaluation. *Information Visualization*, 14(3):250–269, 2015.

[25] S. L. Franconeri. The nature and status of visual resources. *Oxford Handbook of Cognitive Psychology*, 8481:147–162, 2013.

[26] M. Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, 2017. doi: 10.1109/TVCG.2017.2744199

[27] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, Oct. 2011. doi: 10.1177/1473871611416549

[28] R. L. Gregory. *Eye and Brain*. McGraw-Hill, New York, New York, 1973.

[29] A. L. Griffin, A. M. MacEachren, F. Hardisty, E. Steiner, and B. Li. A comparison of animated maps with static small-multiple maps for visually identifying space-time clusters. *Annals of the Association of American Geographers*, 96(4):740–753, 2006. doi: 10.1111/j.1467-8306.2006.00514.x

[30] J. Haberman and D. Whitney. Ensemble perception: Summarizing the scene and broadening the limits of visual processing. *From perception to consciousness: Searching with Anne Treisman*, pp. 339–349, 2012.

[31] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using Weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014.

[32] M. Hebrard and T. D. Taylor. Metatreemap: An alternative visualization method for displaying metagenomic phylogenic trees. *PLOS ONE*, 11(6):1–6, 06 2016. doi: 10.1371/journal.pone.0158261

[33] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 203–212. ACM, New York, NY, USA, 2010. doi: 10.1145/1753326.1753357

[34] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pp. 1303–1312. ACM, New York, NY, USA, 2009. doi: 10.1145/1518701.1518897

[35] J. Heer and G. Robertson. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–1247, Nov. 2007. doi: 10.1109/TVCG.2007.70539

[36] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10):1489–1506, 2000. doi: 10.1016/S0042-6989(99)00163-7

[37] W. Javed and N. Elmqvist. Exploring the design space of composite visualization. *Proceedings of the IEEE Pacific Visualization Symposium*, pp. 1–8, 2012. doi: 10.1109/PacificVis.2012.6183556

[38] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934, 2010. doi: 10.1109/TVCG.2010.162

[39] B. Julesz. *Foundations of Cyclopean Perception*. Chicago University Press, Chicago, Illinois, 1971.

[40] S. Korenjak-Cerne, N. Kejžar, and V. Batagelj. Clustering of population pyramids. *Informatica*, 32(2), 2008.

[41] B. R. Levinthal and S. L. Franconeri. Common-fate grouping as feature selection. *Psychological science*, 22(9):1132–1137, 2011.

[42] S. Limoges, C. Ware, and W. Knight. Displaying correlations using position, motion, point size or point colour. In *Proceedings of Graphics Interface*, pp. 262–265. Canadian Man-Computer Communications Society, Toronto, Ontario, Canada, 1989.

[43] J. Matejka and G. Fitzmaurice. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pp. 1290–1294. ACM, New York, NY, USA, 2017. doi: 10.1145/3025453.3025912

[44] B. Matlen, D. Gentner, and S. Franconeri. Structure mapping in visual comparison: Embodied correspondence lines? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2014.

[45] M. Meyer, B. Wong, M. Styczynski, T. Munzner, and H. Pfister. Pathline: A tool for comparative functional genomics. In *Computer Graphics Forum*, vol. 29, pp. 1043–1052. Wiley Online Library, 2010.

[46] D. Murray. *Tableau Your Data!: Fast and Easy Visual Analysis with Tableau Software*. Wiley Publishing, 1st ed., 2013.

[47] K. Nakayama. Biological image motion processing: a review. *Vision research*, 25(5):625–660, 1985.

[48] B. D. Ondov, N. H. Bergman, and A. M. Phillippy. Interactive metagenomic visualization in a web browser. *BMC Bioinformatics*, 12(385), Sep 2011. doi: 10.1186/1471-2105-12-385

[49] Z. Qu and J. Hullman. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 24:468–477. doi: 10.1109/TVCG.2017.2744198

[50] Z. Qu and J. Hullman. Evaluating visualization sets: Trade-offs between local effectiveness and global consistency. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, pp. 44–52. ACM, New York, NY, USA, 2016. doi: 10.1145/2993901.2993910

[51] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010. doi: 10.1111/j.1467-8659.2009.01694.x

[52] G. Robertson, R. Fernandez, D. Fisher, B. Lee, and J. Stasko. Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1325–1332, Nov. 2008. doi: 10.1109/TVCG.2008.125

[53] J. M. Scimeca and S. L. Franconeri. Selecting and tracking multiple objects. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(2):109–118, 2015.

[54] B. Shneiderman. Tree visualization with tree-maps: 2-D space-filling approach. *ACM Transactions on Graphics*, 11(1):92–99, Jan. 1992. doi: 10.1145/102377.115768

[55] S. Silva, B. S. Santos, and J. Madeira. Using color in visualization: A survey. *Computers & Graphics*, 35(2):320–333, 2011. doi: 10.1016/j.cag.2010.11.015

[56] D. Simons. Current approaches to change blindness. *Visual Cognition*, 7(1–3):1–15, 01 2000.

[57] M. Sondag, B. Speckmann, and K. Verbeek. Stable treemaps via local moves. *IEEE Trans. Vis. Comput. Graph*, 24(1):729–738, 2018.

[58] J. Stasko and E. Zhang. Focus+context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. *Proceedings of the IEEE Symposium on Information Visualization*, pp. 57–65, 2000. doi: 10.1109/INFVIS.2000.885091

[59] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5):11–11, 2016.

[60] M. Treder. Behind the looking-glass: A review on human symmetry perception. *Symmetry*, 2(3):1510–1543, 09 2010.

[61] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. doi: 10.1016/0010-0285(80)90005-5

[62] E. Tufte. *Envisioning Information*. Graphics Press, Cheshire, CT, USA, 1990.

[63] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986.

[64] B. Tversky, J. B. Morrison, and M. Betrancourt. Animation: can it facilitate? *International Journal of Human-Computer Studies*, 57(4):247–262, 2002. doi: 10.1006/ijhc.2002.1017

[65] J. Wagemans. Characteristics and models of human symmetry detection. *Trends in Cognitive Sciences*, 1(9):346–352, 1997. doi: 10.1016/S1364-6613(97)01105-4

[66] J. Wolfe and T. Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1:0058, 03 2017. doi: 10.1038/s41562-017-0058

[67] Y. Xu and S. L. Franconeri. Capacity for visual features in mental rotation. *Psychological science*, 26(8):1241–1251, 2015.

[68] M. A. Yalçin, N. Elmqvist, and B. B. Bederson. Raising the bars: Evaluating treemaps vs. wrapped bars for dense visualization of sorted numeric data. In *Proceedings of the Graphics Interface Conference*, pp. 41–49. Canadian Human-Computer Communications Society / ACM, 2017.

[69] F. Yang, L. Harrison, R. A. Rensink, S. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2018. doi: 10.1109/TVCG.2018.2810918

[70] L. Zhou and C. D. Hansen. A survey of colormaps in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(8):2051–2069, 08 2016. doi: 10.1109/TVCG.2015.2489649