What's the Difference?: Evaluating Variants of Multi-Series Bar Charts for Visual Comparison Tasks

Arjun Srinivasan[1] **Matthew Brehmer**[2] **Bongshin Lee**[2] **Steven M. Drucker**[2]

[1]Georgia Institute of Technology
Atlanta, GA, USA
arjun010@gatech.edu

[2]Microsoft Research
Redmond, WA, USA
{mabrehme, bongshin, sdrucker}@microsoft.com

ABSTRACT

An increasingly common approach to data analysis involves using information dashboards to visually compare changing data. However, layout constraints coupled with varying levels of visualization literacy among dashboard users make facilitating visual comparison in dashboards a challenging task. In this paper, we evaluate variants of bar charts, one of the most prevalent class of charts used in dashboards. We report an online experiment ($N = 74$) conducted to evaluate four alternative designs: 1) grouped bar chart, 2) grouped bar chart with difference overlays, 3) bar chart with difference overlays, and 4) difference chart. Results show that charts with difference overlays facilitate a wider range of comparison tasks while performing comparably to charts without them on individual tasks. Finally, we discuss the implications of our findings, with a focus on supporting visual comparison in dashboards.

ACM Classification Keywords

H.5.2. Information Interfaces and Presentation (e.g. HCI): Graphical user interfaces (GUI).

Author Keywords

Visual comparison; visualization dashboards; task-based evaluation; online experiment.

INTRODUCTION

Comparison is integral to the analysis of data, in forming hypotheses as well as confirming or refuting them [4, 28, 33]. Accordingly, comparison has long been an active topic of research in the visualization community [18]. Researchers and practitioners have developed techniques that allow people to make comparisons, including serendipitous comparisons that would have otherwise not been made absent a visual representation of the data. Many of these techniques are specific to particular application domains and data types, and intended for experts. However, there are many other techniques that are domain-agnostic and appropriate for casual or infrequent use, intended for those with varying levels of visualization literacy. In this paper, we investigate this latter group of techniques and their viability in the context of information dashboards intended for non-experts.

CHI 2018, April 21–26, 2018, Montréal, QC, Canada

ISBN 978-1-4503-5620-6/18/04…$15.00

DOI: https://doi.org/10.1145/3173574.3173878

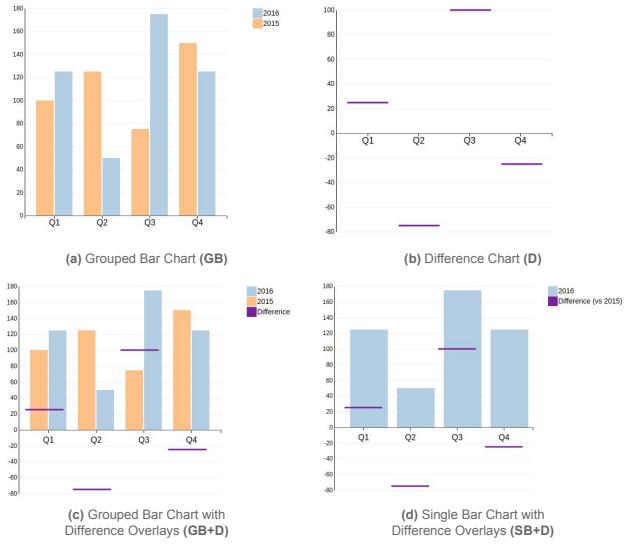

Figure 1. Four chart designs evaluated as part of our online experiment, showing quarterly sales over two years (2015 and 2016).

In describing their experience of designing dashboards with start-up companies and institutions, Froese and Tory [15] quote users of Youneeq [35], one of the many dashboard applications on the market today. These stories explicitly highlight the need to visually compare aspects of the data: one dashboard user said "*I want to compare the most important [Key Performance Indicators] for visitors who use the recommendations with those who do not.*" while another said "*I want to compare the traffic of the user during summer 2015 versus summer 2014.*" Unfortunately, it is not always possible to compare these values directly within a single chart; these values may be represented in separate charts, pages, and tabs, or they may only be visible following some interactive configuration.

Given the prevalence of bar charts in dashboards, we investigate the efficacy of variations of multi-series bar charts for comparison tasks, through an online experiment using four visualization designs (Figure 1) with 74 participants. One of these design variants, a grouped bar chart containing two series (Figure 1a), appears often in dashboards and is likely to be familiar to most readers. Meanwhile, our difference chart (Figure 1b) is a variation of the difference bar chart (Figure 2b), which also appears often in dashboards; our variant

uses mark lines instead of bars to explicitly encode the differences between corresponding values in the two series. The remaining two variants were hybrid versions of the former two: a grouped bar chart and a single bar chart superimposed with difference overlays (Figure 1c, d). Our results suggest that charts with difference overlays facilitate a wider range of comparison tasks than the charts without them, performing comparably to the grouped bar chart and difference chart on the comparison tasks for which those charts are better suited.

The primary contribution of this paper is a systematic investigation of bar chart design variants according to four different tasks and two types of multi-series data, with the results of this investigation having actionable implications for the design of information dashboards. Based on both participants' quantitative responses and subjective feedback, we discuss key observations regarding difference overlays, highlighting both their pros and cons along with design implications for their use in the context of information dashboards. In addition, we discuss potential research directions and how difference overlays can enhance both the dashboard design process and how an audience consumes a dashboard.

RELATED WORK

Our evaluation of design variants of multi-series bar charts for comparison tasks draws from existing classifications of visual comparison design choices, previous task-based evaluations, and previous perceptual studies involving bar charts.

Visualization Design for Comparison

Gleicher [17] proposes several considerations for visual comparison: comparative elements comprised of actions and targets, challenges and strategies pertaining to scalability and data complexity, and designs. In this paper, we focus on comparative elements and designs; we do not explicitly address scalability and varying levels of data complexity in our study. Gleicher's design classification for visual comparison consists of three categories: *juxtaposition*, or placing separate comparison targets adjacent to one another (e.g., juxtaposing values from multiple series in a grouped bar chart); *superposition*, or overlaying comparison targets within the same coordinate space (e.g., overlaying data points having different categorical values within the same scatterplot); and *explicit encoding*, or computing derived values between corresponding comparison targets (e.g., William Playfair's classic chart depicting the trade balance of England over time, derived from calculating the difference between imports and exports [9]).

Munzner [24] proposes a similar classification for *faceting* data into separative views and specifies two forms of juxtaposition: *multiview visualization* (same data, different encodings) and *small multiple visualization* (different data, same encoding).

Both Gleicher [17] and Munzner [24] remark that it is certainly possible to compare values by animating or navigating between comparison targets, however this places a demand on working memory; it is typically easier to compare values that are concurrently visible within the same display [24]. Consequently we do not consider designs that involve animating or navigating between multiple charts.

Comparing values can also be facilitated via the interactive manipulation across juxtaposed views [24], such as via brushing and linking (e.g., [6, 19]), rearrangement and alignment (e.g., [2, 10]), and linked view navigation (e.g., [13, 26]). We do not consider interactive approaches to facilitating comparison, as the forms interaction vary considerably across dashboard applications, and some usage contexts preclude interaction altogether (e.g., a dashboard shown on an ambient display or during a live presentation).

In this paper, we constrain our scope to the forms of visual comparison commonly performed within a single view. Specifically, we compare juxtaposition and explicit encoding, as well as two hybrid forms that combine each of the two approaches with superposition.

Task-Based Evaluation of Visualization Design Choices

Our evaluation draws from and continues a line of research that aims to experimentally identify effective visualization design choices, including laboratory experiments with human subjects [7] and more recent online experiments involving crowd workers [3, 20].

Methodologically, our work bears similarity to previous work that identifies a set of tasks relevant to a particular datatype, and then evaluates visualization design alternatives in terms of how well people can perform these tasks, typically using the metrics of task completion time and error rate. For instance, Albers et al. [1] evaluated several visualization design alternatives for the combination of quantitative time-series data and visual aggregation tasks, such as determining the average value within a particular span of time. Similar recent studies include an evaluation of four alternative small multiple glyph designs for time-series data and three tasks [16], and an evaluation of design variants of scatterplots for high-cardinality quantitative data and a dozen tasks [27]. Analogously, we evaluate four design variants of multi-series bar charts for the combination of categorical or ordinal count data and six comparison tasks.

Some previous work explicitly includes comparison tasks within the scope of their evaluation, albeit with datatypes and design alternatives that differ from the those considered in our current work. For example, Javed et al. [21] evaluated four design alternatives for the combination of multiple time-series data and three tasks – slope identification, value discrimination, and value comparison tasks.

Finally, the evaluation of design alternatives for visual comparison also arises in the context of visualization design studies, which are often highly specific to an application domain. One relevant recent example is a design study involving the comparison of energy consumption values from a portfolio of buildings over time [5], where multi-series bar charts and juxtaposed bar charts were among the design choices considered. However, evaluation in visualization design studies tends to differ methodologically from domain-agnostic experimental approaches such as our own, with a greater propensity toward qualitative evaluation with target users and project stakeholders that have substantial domain expertise.

Evaluating Alternative Bar Chart Designs

Skau et al. [29] recently evaluated several design alternatives of a single-series bar chart, in which the alternatives featured different illustrative embellishments typically encountered in infographics. Their evaluation measured performance with respect to two tasks: reading the value of a single bar and comparing values between bars. Though our work does not consider bar charts' illustrative embellishments, we do measure how well people can compare values between bars, albeit in design variants of the multi-series bar chart.

Most closely related to our work is an evaluation of how well people can compare the height of bars appearing in single-series and stacked bar charts [32], where Talbot et al. varied the distance of the comparison targets and the number of distractor bars (bars not involved in the specified comparison) across four separate experiments. In contrast, we examine a broader range of comparison tasks for variants of multi-series bar charts; our study conditions have a fixed number of distractor bars, a choice motivated by the data and constraints specified in the next section.

CHARACTERIZATION OF INFORMATION DASHBOARDS

To have a better understanding of the information dashboards used and needed in the real-world, we conducted informal interviews with product managers of Microsoft Power BI [22], who oversee the development of dashboards and dashboard development applications. We also surveyed 68 publicly available information dashboards listed on the Microsoft Power BI partner showcase [23] and the Tableau public gallery [30]. These repositories present curated lists of customer (individual users and organizations) created dashboards developed using popular dashboard tools including Microsoft Power BI and Tableau [31]. As selection criteria, we specifically looked for dashboards that showed data for multiple years (series) and allowed selecting multiple series (for comparison).

Data and Charts Appearing in Dashboards

One of the recurring themes that arise from the discussions with product managers was a question often posed by many dashboard users: "*What's changed?*" referring primarily to changes in data values, such as in the example described in the introduction of this paper. (Note that product managers' statements were based on both the feedback they gathered during discussions with dashboard users and the recorded usage patterns of dashboard applications.)

The product managers acknowledged that while dashboards are useful for displaying summaries of data at a single point in time, they often lack context, such as whether the values shown are better, worse, or similar relative to a previous state of the data. They further stated that charts displaying quantitative data across categorical values (e.g., regions, product types) and ordinal values (e.g., months, quarters) were the most commonly visualized form of data across dashboard users, and that this data was typically represented using bar charts.

The data used in these dashboards spanned the domains of sales, energy, and healthcare, among others. We encountered many dashboards (55/68) that featured data similar to what the product managers had described: quantitative values reported for each level of an ordinal attribute, a categorical attribute, or both, represented as bar charts, which were indeed the most common chart type used.

We observed that bar charts were commonly used to show values across months or quarters, while sorted bar charts were typically used to show categories ranked in descending or ascending order (e.g., a bar chart depicting top 10 salesmen based on the number of product units they sold in a given year, sorted in descending order). In the former subset of charts showing values across months or quarters, the same categories persist across series in the same order. However, in the latter case, each year might not contain the same categories (e.g., employees might leave or new employees might join) nor may the order persist. This addition, removal, or reordering of categories between series is a discrepancy that may not be obvious if the data is shown in separate charts or values for categories are aggregated as a single chart.

Dashboards typically included interactive toggles for displaying a series of values. When multiple series were selected, one common approach used by the dashboards was to aggregate the values for the selected series. However, this approach does not promote direct comparisons between the series since users need to repeatedly toggle the filters and mentally compute the differences. Other less common but perhaps more effective approaches for making comparisons between series involved concatenating the series into a single bar chart or using a grouped bar chart showing values for multiple years by a specific category (both forms of juxtaposition [17]), or displaying the net difference between corresponding values in two series as a difference bar chart (a form of explicit encoding [17]); examples of these approaches are shown in Figure 2. Superposition, Gleicher's third class of designs to support comparison, which in this case would correspond to overlaying values within the same chart, was not a common approach.

Constraints on Chart Design Imposed by Dashboards

From the survey of dashboards and discussions with product managers, we identified the following design constraints for comparing values across multiple series within the context of information dashboards.

Figure 2. Examples from our survey: (a) a bar chart with concatenated series, (b) a difference bar chart explicitly encoding year-over-year changes, and (c) a grouped bar chart showing values for two series.

Explicit Encoding Approaches Preclude Other Tasks
While comparison tasks are important, it is essential that the dashboards still support basic value identification tasks. In designs that solely represent derived values for the purpose of comparison (e.g., Figure 2b), it can be difficult to identify original non-derived values from individual series, particularly if derived values have been normalized to a percentage scale.

Juxtaposition Approaches are Limited by Space
Information dashboards typically contain multiple charts arranged to fill a display. This fixed layout implies that it may not always be feasible to dynamically add additional charts or increase the size of an existing chart to facilitate comparisons. Accordingly, juxtaposition-based design choices such as small multiples or concatenated views may not be feasible if they require additional space.

Varying Levels of Visualization Literacy Among Users
A product manager who leads a team that works exclusively on designing dashboards based on customer requests emphasized that dashboard consumers have varying levels of visualization literacy, and both dashboard designers and consumers are accustomed to a small set of familiar chart types. Consequently, variants of familiar chart types are preferred to introducing novel visual encodings. For example, if a single bar chart is used to represent values for one series, a variation of a bar chart such as a concatenated, grouped bar chart, or derived difference bar chart would be preferred for comparing values across multiple series, too.

ONLINE EXPERIMENT

With the design constraints mentioned above in mind, we explored variants of bar charts by conducting an online experiment to investigate how alternative bar chart designs support comparison tasks with two types of data.

Two Data Conditions

The two data conditions that we considered involved the nature of the categories: **Constant** categories, in which the 12 categories corresponded to the months of the year; and **Varying** categories, in which the 12 categories corresponded to 12 U.S. States, selected at random. In the latter condition, we also randomly selected two or three states and assigned their target-series values to NA; we then selected two to three other categories at random and assigned their source-series values to NA. This choice reflects the common scenario of missing data, of instances where cross-series comparisons are not possible. As an example, consider a sales manager reviewing the year-over-year sales revenue generated by her top-10 performing salespersons; some of these salespersons may be in the ranking two years in a row, while others may not; others still might have left or joined the company in the second year, and thus some data is unavailable in these cases.

Four Chart Design Conditions

Inspired by the classification by Gleicher [17], we considered many ways of implementing juxtaposition, explicit encoding, and superposition (overlay) for multi-series bar charts. We initially considered more than 20 alternative designs. In addition to the design variations with different category sorting choices and different derived values for explicit encoding, we explored design variations that encompassed alternative difference mark types (e.g., tick, bar, text), explicit annotations to highlight new or old categories in the varying categories condition (e.g., stroke, bolded labels), and alternative encoding channels such as hue and lightness along diverging and continuous scales. We then decided to constrain our scope in two ways. First, we fixed how the categories were sorted: in the *Constant* categories condition, the months were displayed in chronological order from left to right; in the *Varying* categories condition, the categories were ordered by their target series value. This difference in ordering between the two data conditions reflects two common types of charts encountered in our survey: a bar chart ordered by categories or one ordered by values from highest to lowest. Second, we selected only one derived value: the computed difference between corresponding values from the target series to the source series; we had initially also considered percentage change, rank change, and absolute change, among others.

Given the design constraints discussed in the previous section and informed by our review of existing dashboards, we arrived at four chart designs. Altogether, these four chart designs allowed us to study and quantify the potential benefits and drawbacks of combining juxtaposition, explicit-encoding, and superposition based approaches to comparison.

Grouped Bar Chart (GB): One of the most familiar charts that we encountered in our survey, the grouped bar chart is an instance of a *juxtaposition*-based design (Figure 2c). As shown in Figure 1a, the blue bars corresponded to values from the target series while the orange bars corresponded to values from the source series. We opted to use a grouped bar chart instead of a concatenated bar chart since comparisons are likely to be more accurate with no distractor bars in between corresponding values [32].

Difference Chart (D): The chart shown in Figure 2b is an example of a *difference bar chart*, which is an instance of an *explicit encoding*-based design that encodes derived values. A notable feature of this design is that the bars diverge from the x-axis, as the difference between corresponding values across the two series can be negative. In our implementation (Figure 1b), we represented the derived values not as bars but as purple horizontal lines, so as to be consistent with the difference overlays shown in the two overlay chart designs (Figure 1c, d). Though we did encounter difference bar charts in our survey, recall the design constraint discussed above in which *explicit encoding*-based designs such as these are undesirable due to a viewer's inability to retrieve the original absolute values from either series. We therefore included this chart design as a baseline condition that we could evaluate relative to the two overlay conditions.

Single Bar Chart with Difference Overlays (SB+D): This chart combines *explicit encoding* and *superposition*, in that we overlay differences from the source series to the target series over a single bar chart representing values from the target series (Figure 1d). This chart is a novel design that we did not encounter in our survey. Since the bars and difference

overlays appear on a common scale, both the target series values and the differences can be observed directly; the source series values require adding or subtracting a difference value from its corresponding target series value.

Grouped Bar Chart with Difference Overlays (GB+D): The final chart design used in our study combines *explicit encoding*, *juxtaposition*, and *superposition* in that we add difference overlays to values from both the target series and the source series (Figure 1c). This too is a novel design; it also has the most information encoded of all the chart design conditions, with values from both series and their derived differences being directly encoded.

Note that in the *Varying* categories data condition, the difference chart and difference overlay charts (Figures 1b, c, d) showed difference values only for the six categories that had values in both series, as shown in Figure 3. To clearly disambiguate series values from differences shown by the mark lines, we offset the mark lines a little to the left of the bars to highlight that they are changes compared to source series and not the values of the source series itself.

Data Generation

We used representative yet synthetic two-series data throughout our study. For each trial, we generated 12 pairs of quantitative values selected at random between 0 and 200, one for each category to be displayed in a bar chart; for each pair, one value corresponded to a *target series* associated with a single year, while the other value corresponded with a *source series*, the preceding year.

Task Specification

There are many visualization task classifications that include *comparison* [4], though they differ with respect to the granularity or scope of the comparison, and whether comparison is the means by which some higher-level goal is accomplished, or if comparison is an end goal in itself. According to the task typology of Brehmer and Munzner [4] and its later extension [24], comparison is an action that involves two or more targets, and comparison may occur in the context of various forms of visual search and higher-level actions, such as discovery or presentation.

Gleicher [17] recently built upon the notion of actions and targets in an effort to more precisely describe the process of visual comparison; he distinguishes between *explicit* and *implicit* comparison targets as well as six abstract comparison actions: *identifying* relationships between items, *measuring* these relationships, *dissecting* these relationships to understand their nature, *connecting* multiple relationships, *contextualizing* these relationships, and *communicating* these relationships. The tasks in our study focus primarily on *identifying* and *measuring* relationships between *explicit* and *implicit* targets; with respect to multi-series bar charts and the data described above, this translates to comparing the values within and across series: identifying extreme values, identifying and measuring maximum changes, and identifying categories that have a value in only one of the two series.

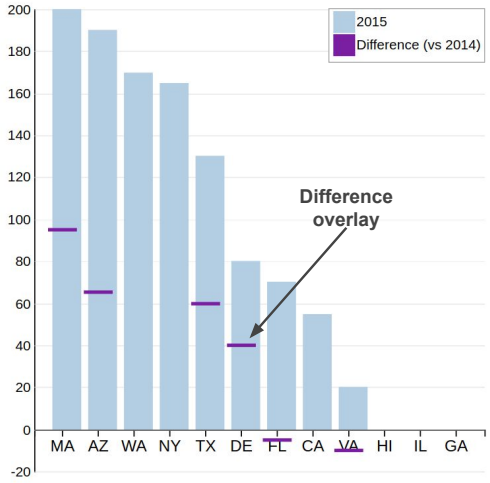

Figure 3. Example of a single bar chart with difference overlays (SB+D) displaying *Varying* categorical data. Blue bars represent values in the target series (2015), and are sorted in descending order by value from left to right. Difference overlays are shown only for categories that appear in both series. *HI*, *IL*, and *GA* have values only in 2014, and *WA*, *NY*, and *CA* have values only in 2015.

Given this task specification, we asked participants to perform multiple trials of four different comparison tasks. However, some of these tasks were either impossible or trivial with some of these combinations, and thus we did not ask participants to perform all four tasks with all (4x2) combinations of chart design condition and data condition. In Table 1, we provide a brief definition of each task, an example of the exact instruction text shown to participants, and an indication of the chart design and data conditions where the task was performed.

Participant Recruitment and Compensation

We recruited 76 participants via email from across Microsoft's mailing list containing employees who design, develop, or use visualization and dashboard applications. Data for two participants were later discarded (discussed in the following section), resulting in 74 participants between the ages of 25 and 60 (21 female, 51 male, 2 undisclosed). Out of the 36 participants who self-reported their job description, 22 were software developers, 9 were managers, and 5 were sales representatives.

All participants who completed the study received a $5 gift card for their time. To motivate participants to do their best, we rewarded the participant who achieved the best performance (considering task accuracy and completion time) for each chart design condition, with a $25 gift card. To promote study participation, we rewarded two participants, selected at random regardless of their performance, with a $100 gift card.

Procedure

The web-based study followed a within-subjects design in which each participant performed all relevant tasks with all four chart design conditions and two data conditions.

The study consisted of seven phases. Phases 2-5 were repeated four times (one for each chart design condition). For each chart design condition, phases 3-5 were repeated twice (one

			GB		GB+D		SB+D		D	
	Task	Example instruction text	C	V	C	V	C	V	C	V
T1a	Identify extremes in target series	*Click on the month with the minimum value for 2016*	✱		✱		✱			
T1b	Identify extremes in source series	*Click on the month with the maximum value for 2015*	✱		✱		✱			
T2	Identify maximum absolute change	*Click on the state with the largest absolute change in value from 2015 to 2016*	✱	✱	✱	✱	✱	✱	✱	✱
T3	Measure difference for a category	*Enter the absolute change in value for April from 2014 to 2015*	✱	✱	✱	✱	✱	✱	✱	✱
T4a	Identify categories only in target series	*Click on the states that have a value ONLY in 2016*		✱		✱		✱		
T4b	Identify categories only in source series	*Click on the states that have a value ONLY in 2015*		✱		✱		✱		

Table 1. Task overview. A ✱ indicates that participants completed one training and two testing trials for the specific combination of task (T1-T4), chart (GB, GB+D, SB+D, D), and data condition (Constant or Varying). T1 could not be performed with D since original series values are not directly shown. T1 was also not performed with V since charts in this condition were already sorted by target series value, where this task would be trivially easy. Similarly, T4 could not be performed with D since both source and target series specific categories would lack difference marks, making it impossible to differentiate the two. T4 was also not performed in C, since categories (months) maintained a fixed presence.

for each data condition). The order of the (4x2) conditions and the relevant tasks were randomized for each participant. Screenshots from an entire study session are included as part of the supplementary material.

1. *Introduction and consent*. The participant began by reading about the study goals and compensation. She then consented to study procedure, which included a disclaimer indicating that responses would be timed and that demographic data (e.g., age, gender) would be collected.
2. *Chart design introduction*. We introduced the participant to one of the four chart designs, using the example of quarterly sales over two years shown in Figure 1.
3. *Data introduction*. We then introduced the participant to one of the two data conditions. For the varying data condition, we informed the participant that no difference marks for a category (bar) in the chart indicates that the category did not have a value in one of the two years.
4. *Training phase*. We asked the participant to perform several training trials of each task relevant to the current combination of chart design and data. We informed the participant that response time for the training trials would not be recorded, and that she should take as much time as necessary to ensure that she understands how to respond correctly. We provided feedback on the participant's responses each time she submitted a response. For each trial, we gave the participant three attempts to submit a correct answer. After three failed attempts, we showed the participant the correct answer and she proceeded to the next trial.
5. *Testing phase*. We presented the participant with an indication that the training was concluded, and that performance on subsequent trials would be timed with no feedback provided. The testing phase consisted of four to eight trials (two for each task), depending on the combination of chart design and data.
   To screen out participants that appeared to be blindly guessing or responding to trials as quickly as possible, we added three trivial response tasks that appeared at random between test trials (with a maximum of one per chart design condition). These tasks required that the participant merely click on the bar corresponding to a specific category (e.g., "Click *WA*" or "Click *Jan*"). The two participants whose data we excluded failed to correctly perform these trivial tasks.
6. *Preference specification*. We asked the participant to select her most and least preferred chart design and to optionally explain her choice in a text field.
7. *Demographic information submission*. Finally, we asked the participant to provide us with demographic information including her age, gender, education level, and job role before exiting the study.

In summary, each participant performed a total of 87 trials (28 training, 56 testing, 3 guess checking) across 4 chart designs x 2 data conditions, as indicated in Table 1.

Hypotheses

Our overall hypothesis was that charts with difference overlays (GB+D) and (SB+D) facilitate more visual comparison tasks than exclusively explicit-encoding based charts (D) or exclusively juxtaposition-based charts (GB). We also hypothesized that the design featuring difference overlays without juxtaposition (SB+D) is superior to the design with juxtaposition (GB+D), since for all tasks except tasks T1b, which requires identifying source series values, SB+D will perform better than GB+D due to the lower number of distractor bars in SB+D, or bars not involved in the specified comparison.

We had the following task-specific hypotheses:

- **H1.** *Difference overlays are just as good or better than juxtaposition alone for extreme value identification tasks, at least for the target series* (T1): For T1a, GB+D and SB+D will perform comparably to GB. For T1b, GB+D and GB will outperform SB+D, since source values require adding or subtracting the difference overlay value from the target series value.
- **H2.** *Difference overlays are just as good or better than explicit encoding alone for the identification of the maximum absolute change* (T2): GB+D and SB+D will perform comparably to D.
- **H3a.** *Difference overlays are just as good or better than explicit encoding alone for difference measurement* (T3): SB+D and GB+D will perform comparably to D.
- **H3b.** *Difference measurement is difficult without explicit encoding* (T3): GB will perform poorly relative to D, SB+D, and GB+D.
- **H4.** *Difference overlays are the best way to identify missing values in either series* (T4): GB+D and SB+D will outperform GB. In GB+D and SB+D, a target series bar without a difference overlay indicates a value missing from the source series. In GB+D alone, the reverse is true, while in SB+D, this is indicated by the absence of both a bar and an overlay. Without difference overlays, a value of 'NA' could be misinterpreted as a value of zero.

Task		Factor	Completion Time (sec) Test Statistic	η_p^2	p	Error Magnitude (%) Test Statistic	p
T1a	Identify extremes in target series	Chart Design	$F_{2,146} = 2.769$	.037		$\chi^2_{2,74} = 6$	
T1b	Identify extremes in source series	Chart Design	$F_{1.44,105.09} = 85.065$	.538	***	$\chi^2_{2,74} = 30.184$	***
T2	Identify maximum absolute change	Chart Design	$F_{2.66,194.42} = 109.282$	.600	***	$F_{3,511} = 12.171$	***
		Data	$F_{1,73} = 48.197$	.398	***	$F_{1,511} = 64.801$	***
		Chart * Data	$F_{2.68,198.31} = 9.6$	.122	***	$F_{1,511} = 11.744$	***
T3	Measure difference for a category	Chart Design	$F_{2.87,209.71} = 77.603$	.515	***	$F_{3,511} = 1.704$	
		Data	$F_{1,73} = 18.722$	.204	***	$F_{3,511} = 26.069$	***
		Chart * Data	$F_{2.84,207.11} = 12.068$	.142	***	$F_{3,511} = 0.084$	
T4a	Identify categories only in target series	Chart Design	$F_{1.62,117.99} = 19.857$	.214	***	$\chi^2_{2,74} = 2.246$	
T4b	Identify categories only in source series	Chart Design	$F_{2,146} = .729$	.010		$\chi^2_{2,74} = 3.350$	

Table 2. Effects of Chart Design and Data conditions on average completion time and error magnitude for each task (* = $p < .001$).**

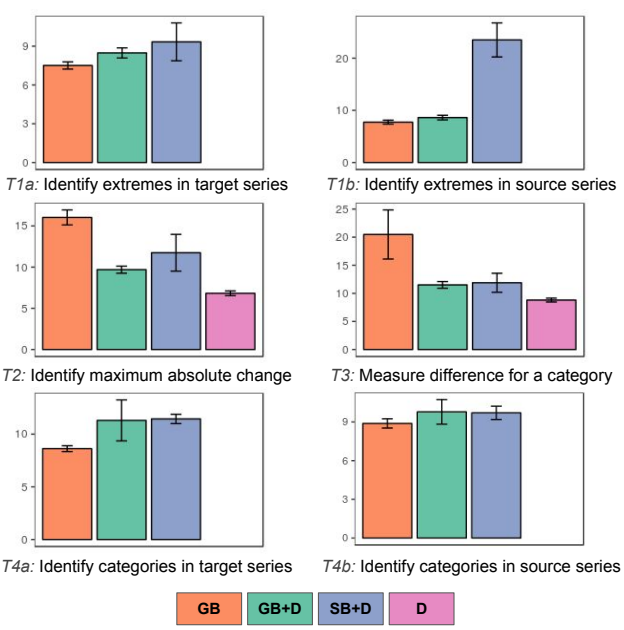

Figure 4. Means of task completion times (in seconds) for each task and chart design condition. Error bars represent standard errors.

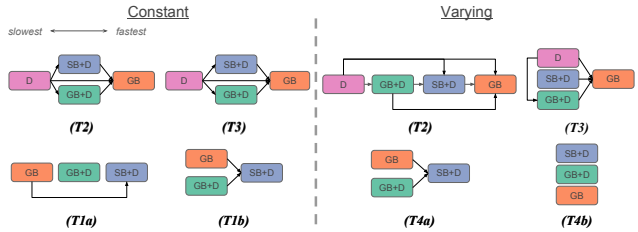

Figure 5. Pairwise relations for average completion time of individual tasks. Arrows indicate that the source is significantly faster than the destination. Designs aligned vertically are ordered top to bottom from fastest to slowest, but these differences are not significant.

RESULTS

For each combination of chart design condition, data condition, and relevant task, participants completed two trials. Accordingly, we used the average of the two repetitions in our analysis of task completion time and error.

Task Completion Time

The time to complete a trial was measured from when the chart and corresponding task were displayed until the participant clicked the "Submit" button. Task completion times as a function of chart design condition are shown in Figure 4.

We performed a one-way repeated-measures ANOVA for the analysis of completion time for T1 (identifying extremes) and T4 (identifying categories in only one series), while we performed a two-way repeated measures ANOVA for T2 (identifying the maximum change) and T3 (measuring differences); we performed a log-transformation to the completion time values in cases where they did not follow a normal distribution.

Table 2 summarizes how the factors of chart design affect completion time; for T2 and T3, we also summarize the effect of data condition and the interaction between chart design and data conditions. In cases where the sphericity assumption was violated, we report values with a Greenhouse-Geisser correction. We report partial eta-squared (η_p^2), a measure of effect size, where 0.01 is a small effect size, 0.06 is medium, and 0.14 is large [8]. Figure 5 shows a summary of Bonferroni-adjusted post-hoc comparisons.

Accuracy

The magnitude of error for a trial was computed based on the following calculation:

$$Error\% = \frac{|ParticipantResponse - CorrectResponse|}{CorrectResponse} \times 100$$

The range of responses varied across tasks due to their different response form: T1 and T2 involved identifying categories by clicking on a single region in the chart whereas T3 required

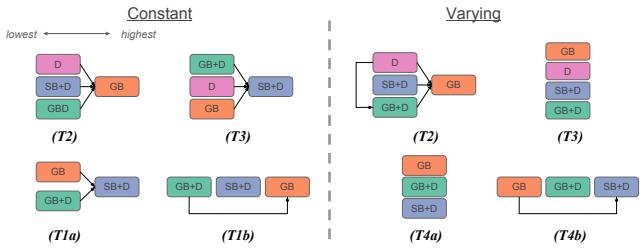

Figure 6. Pairwise relations for average magnitude of errors for individual tasks. Arrows indicate that the source had significantly lower error than the destination.

	Task	GB	GB+D	SB+D	D
T1a	Identify extremes in target series Total number of trials: 148	139 ± 5 (93.9%)	146 ± 2 (98.7%)	144 ± 3 (97.3%)	
T1b	Identify extremes in source series Total number of trials: 148	145 ± 3 (97.9%)	143 ± 4 (96.6%)	117 ± 9 (79.1%)	
T2	Identify maximum absolute change Total number of trials: 296	248 ± 12 (83.8%)	281 ± 7 (94.9%)	285 ± 6 (96.3%)	291 ± 4 (98.3%)
T3	Measure difference for a category Total number of trials: 296	279 ± 7 (94.3%)	286 ± 6 (96.6%)	279 ± 7 (94.3%)	290 ± 4 (97.9%)
T4a	Identify categories only in target series Total number of trials: 148	142 ± 4 (95.9%)	137 ± 6 (92.6%)	137 ± 6 (92.6%)	
T4b	Identify categories only in source series Total number of trials: 148	146 ± 2 (98.7%)	143 ± 4 (96.6%)	140 ± 5 (94.6%)	

Table 3. The number of correct responses ± 95% confidence intervals for each combination of task and chart design. The percentage of correct responses is shown in parentheses.

measuring values and submitting a response via numeric stepper, and T4 required clicking on multiple regions in a chart. T1 and T2 had binary errors (0 or 1); in T3, the error was continuous; and in T4, possible errors were [0, 0.5, 1] or [0, 0.33, 0.67, 1] depending on the number of correct categories selected. Table 2 summarizes the effects of chart design and data conditions on the magnitude of error while Table 3 indicates the number of correct responses.

The error results did not fit a normal distribution, so we performed a non-parametric Friedman's test to examine the effect of chart design on error rates for T1 and T4. We performed an aligned rank transform [34] to the results of T2 and T3 so as to perform the non-parametric equivalent of a two-way ANOVA.

Subjective Preference

When asked to select and justify the **chart design that they most preferred**, 63 participants (85.1%) selected the grouped bar chart with difference overlays **GB+D**. Six participants (8.1%) opted for the difference chart (D), three (4.1%) for the grouped bar chart, and two (2.7%) for bar chart with difference overlays (SB+D). 56 of the 63 participants who selected GB+D also provided a justification for their choice. GB+D "*provided the most information while being the least confusing*" (P55), and it "*decreases the time to think about the amount in the difference*" (P33). P27 stated that "*having all the information as well as the calculated visuals [difference values] allowed for faster understanding*."

Participants' choices of their **least preferred chart design** were more varied. 30 participants (40.5%) selected **D**, 26 (35.1%) selected SB+D, 10 (13.5%) selected GB, and 8 (10.8%) selected GB+D. D was deemed to be "*good if you are looking specifically only for difference, but does not provide additional context*" (P29) and "*doesn't really tell you much ... you'd need to have an accompanying chart or table to support the visual*" (P62). For SB+D, it was "*a little more difficult to imagine the previous year's value (and to remember them)*" (P15), as these values had to be calculated from the target series and difference overlay. The "*missing data points forced me to stop and do math (figure it out). The other charts were more 'honest' in what they presented. Zeros were hard to interpret.*" (P35). However, some felt that SB+D was better than GB+D but may require practice before using it competently.

Interpreting the Results

To interpret the results as a means to understand the *effectiveness* or *difficulty* associated with the chart designs used in our study, it is important to consider not only the number of errors (Table 3), but also the magnitude of these errors, task completion times, and subjective preferences. These metrics altogether provide a holistic view of relative effectiveness and difficulty across tasks and conditions.

H1: Finding Extreme Values

For the task of identifying extreme values in the target series (T1a), participants were more accurate using GB+D and SB+D relative to GB. This implies that difference overlays promote similar or even superior performance relative to an exclusively juxtaposition-based design (GB) for finding extreme values.

Participants were comparably accurate with GB and GB+D when identifying extreme values in the source series (T1b), while they had considerable difficulty performing this task with SB+D, where accuracy was almost 18% lower. They also took longer to perform the task with SB+D relative to GB and GB+D (Figure 4). Altogether, these findings confirm H1.

H2: Identifying the Maximum Change

Participants were similarly accurate with D, GB+D, and SB+D, thus confirming H2). They were less accurate using the exclusively juxtaposition-based design (GB) (< 10% compared to other designs). We attribute the lower accuracy of GB to the fact that it requires the viewer to mentally compare and compute values for each category and remember the values in order to identify the maximum change. Completion times followed a similar trend, as participants were slowest with GB. This finding highlights an advantage of difference overlays, in that they can facilitate rapid comparisons of differences.

H3: Measuring Differences

Participants were similarly accurate when measuring differences with D, GB+D, and SB+D, thus confirming H3a and suggesting that difference overlay designs perform comparably to an exclusively explicit encoding-based design for this task. The accuracy and response times were comparable between D, GB+D, and SB+D, and while participants performed the tasks more slowly with GB, they were comparably accurate using all four charts designs, so we were unable to confirm H3b. In other words, explicit encoding does not guarantee more accurate difference measurement, but it appears to reduce the time to complete the task.

H4: Identifying Missing Values

When identifying categories that appeared only in the target series (T4a), participants were fastest with GB and GB+D, and they were comparably accurate. For the converse task (T4b), participants performed the task in about the same amount of time regardless of whether they were using SB+D, GB+D, or GB, though they were most accurate using GB. Thus, we were unable to confirm H4, in that difference overlays did not assist in identifying missing values to the extent that we had expected, despite the absence of a difference mark being an indication of missing value in one of the two series.

We further analyzed the erroneous T4 trials to better understand the types of errors that participants made. In the case of identifying categories appearing only in the target series (T4a), we found that out of the 22 erroneous trials for designs with difference overlays, 18 of these responses were partially correct rather than totally incorrect, in that participants missed at least one category appearing only in the target series. In the converse task (T4b), 3 out of the 5 erroneous trials with GB+D were similarly only partially correct. Surprisingly, all 8 erroneous trials with SB+D were cases where participants selected categories exclusive to the target series rather than those appearing only in the source series, which suggests a misinterpretation of the task in the context of this chart design.

DISCUSSION

The results of our study have several implications for the design of information dashboards and the multi-series bar charts that are prevalent in this application context.

Do Difference Overlays Facilitate More Tasks?

Our overall hypothesis was that charts with difference overlays would facilitate more visual comparison tasks than exclusively explicit-encoding based charts or exclusively juxtaposition-based charts. GB+D and SB+D performed comparably well on individual tasks with the exception of T4 (identifying categories present in only one series), so this overall hypothesis was largely supported. These charts allow viewers to identify extreme values, identify large changes, and quickly assess the magnitude of changes; contrary to our expectation, they did not support the identification of missing values.

People Prefer Charts with Difference Overlays

Sixty three (85.1%) participants opted for GB+D as their most preferred chart design. Thirty (40.1%) participants selected D as their least preferred design, which was not surprising given what we had heard earlier from dashboard product managers. Given a participant pool of frequent dashboard users and their presumed familiarity with exclusively juxtaposed (GB) and explicit encoding (D) based designs, we were encouraged to learn that participants preferred designs with difference overlays (GB+D, SB+D); some stated that these designs provided a better context of what the target series values meant.

Collectively, these preferences and comments suggest designers of information dashboards add such charts to the palette of charting options or include the ability for viewers to interactively toggle difference overlays for multi-series bar charts.

Show Differences Overlays with Both Original Values

Another overall hypotheses we had was that SB+D would perform better than GB+D for all but one T1b (identifying extreme values in the source series), due to the larger number of distractor bars in GB+D. However, people were more accurate with GB+D than SB+D, and thus we were unable to confirm our hypothesis. By examining the nature of errors and participants' subjective responses, we found that participants found absence of the source series in SB+D to be confusing. When performing T3 (measuring change in value for a category) with SB+D, participants incorrectly entered the absolute value from the target series in 6 of the 17 erroneous trials, while in 6 other erroneous trials, participants responded with the target series value $\pm$ the value of the difference overlay; the correct response was simply the value of the difference overlay. In addition to the frequency and nature of errors, it is also worth pointing out the larger *magnitude* of measurement errors incurred when using SB+D to perform T3 in the *Constant* categories condition (Figure 6). With regards to difference overlays, participants said "*If you're going to use this approach, show both actual values*" (P18) and "*[it's] confusing to have one year and a difference*" (P10). Thus, it is preferable to have both target and source reference values when displaying difference overlays, even for tasks that do not directly involve these reference values.

Redundancy of Overlays May Reduce Measurement Error

GB performed considerably worse than other chart designs for T2 (identifying the category with maximum change); 48 out of 296 trials were erroneous, with 30 out of 48 appearing in the *Varying* categories data condition. In 25 of these 30 trials, participants selected categories that were unique to the target series as the category with maximum change, indicating that they misinterpreted categories that were not present in the source series as categories that had zero values. Interestingly, most of these errors (20) occurred with charts without difference overlays (GB and D). The absence of a difference overlay for a category is a way to redundantly signify that a comparison is not possible, since the category may be unique to either the target or the source series. Thus, the presence and absence of difference overlays may play a role in reducing errors when it comes to identifying and measuring differences.

Beyond Difference Overlays for Spotting Missing Values

Despite the benefit of difference overlays for reducing the frequency of measurement error, it did not appear that difference overlays were particularly good for identifying missing values (T4): the categories appearing only in one series. For example, consider a grouped bar chart with two series in which one category has only a single bar; does the absent bar signify a value of zero or a value of '*NA*?' If a difference overlay is present, one can assume the former, while if the difference overlay is absent, one can assume the latter, such as in the cases of categories *WA*, *NY*, and *CA* in Figure 3; this distinction may be too subtle. Ultimately, chart designers need to consider alternative or additional visual channels beyond difference overlays through which to indicate categories with missing values, such as by automating the annotation or highlighting of these categories.

LIMITATIONS AND FUTURE WORK

The scope of our study was constrained to several conditions and tasks. In future work, we intend to expand this scope and realize the design implications discussed above.

Beyond Bar Charts and Dual-Series Data

Visual comparison tasks and the associated high-level design choices of *juxtaposition*, *superposition*, and *explicit encoding* [17, 24] go well beyond multi-series bar charts and the type of data that they can portray. One interesting direction for future work is to consider the analog of difference overlays for other data types and chart types that are prevalent in information dashboards (e.g., scatterplots, line charts), as the horizontal lines that we use in this paper are only appropriate for superposing on bar charts.

Additionally, we only considered the common use case of comparing two series (e.g., year-over-year or quarter vs. quarter), and a relatively small but common number of categories (12). For multiple series, a single difference overlay could span from the first to last series, or multiple difference overlays could be shown between adjacent series; the latter option may result in overly dense charts that confuse viewers. Thus, follow-up studies should examine both the effect of varying the numbers of series, the number and span of difference overlays across these series, as well as the effect of varying the numbers of categories. Altogether, such studies will help to determine the scalability of difference overlays.

Combining Difference Overlays with Annotation

While we considered several chart designs before restricting our scope to the four charts, we did not formally compare difference overlays with forms of annotation or highlighting, which is a large design space to consider [25]. Prior work has shown that annotations can be a valuable tool in dashboards and are an effective means to generate context [12]. It could be interesting to compare charts that combine difference overlays and forms of annotation in different ways to charts that only leverage either one to better understand their respective advantages and disadvantages. As highlighted earlier, participants in our study struggled to identify missing values with difference overlay-based designs. Additional highlights (e.g., bolding or coloring category labels) or annotations (e.g., arrows, asterisks, and text labels) might help to signify these missing values.

Storytelling and Revealing Differences with Overlays

In their work on annotations in business intelligence (BI) dashboards, Elias et al. [12, 11] characterized four types of entities as part of a narrative prototype for BI: *information entities*, *relational entities*, *organization entities*, and *emphasis entities*. A difference overlay can be considered as an information entity, since it encodes a value directly on a chart, as well as a relational entity, since it highlights a relationship between two other entities: the values for a category between series. Another line of future work involves exploring the potential use of difference overlays in the context of storytelling with dashboards. Dashboards are often used to present data, consumed by an audience in a live presentation or asynchronously. In these contexts, the appearance of difference overlays could be staged or revealed selectively for maximum impact [14]. For example, a presenter selects the two categories in a grouped bar chart with the largest and smallest change and adds difference overlays to these categories as a means to emphasize the extent of changes.

Interactive Difference Overlays

We have yet to consider interactive techniques for facilitating visual comparison. An investigation into how interactivity might complement difference overlays is certainly an important next step. Simple tooltips and annotations could be revealed when hovering or clicking on difference overlays to promote more accurate comparisons. Beyond simple interaction, an investigation into how difference overlays might be involved in more advanced interaction techniques such as brushing and linking [6] spanning multiple charts is also certainly worthy of consideration, especially since information dashboards often contain an arrangement of multiple charts where each chart is typically tailored for a specific set of tasks and different subsets of the data. In many cases, the same categories and the same values appear in multiple charts within a dashboard. As examples, manipulating a range slider for time could trigger the temporal extent of difference overlays in charts responding to the slider. Further, brushing over a difference overlay could highlight the category and its value in both series in every other chart where it appears, whereas brushing over a single bar would only trigger a highlight of the single category value where it appears elsewhere.

CONCLUSION

We presented a study that evaluated four variants of multi-series bar charts in terms of their capacity for facilitating comparison tasks. Our choice of chart designs was motivated and also constrained by the context of information dashboards, where both bar charts and comparison tasks are particularly prevalent. We chose the four chart designs according to recent classifications of comparison appearing in the visualization literature [17, 24]. The results of our online experiment with 74 participants indicated that charts with difference overlays, or hybrid designs that combine aspects of juxtaposition and explicit encoding with superposition, are just as good or better than solely juxtaposition or explicit encoding based charts on individual tasks. Additionally, these hybrid designs have the advantage that they afford more tasks by combining elements of juxtaposition and explicit encoding-based designs. We discussed key observations regarding difference overlays and the potential implications of these in the context of information dashboards. Finally, we highlighted limitations of the current study and open areas for future research such as how difference overlays can be used to enhance the storytelling capabilities of information dashboards.

ACKNOWLEDGMENTS

The first author performed this work during an internship at Microsoft Research. We thank the product managers of Microsoft Power BI who we interviewed.

REFERENCES

1. Danielle Albers, Michael Correll, and Michael Gleicher. 2014. Task-driven evaluation of aggregation in time series

visualization. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 551–560. https://doi.org/10.1145/2556288.2557200.

2. Danielle Albers, Colin Dewey, and Michael Gleicher. 2011. Sequence surveyor: Leveraging overview for scalable genomic alignment visualization. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)* 17, 12 (2011), 2392–2401. DOI: http://dx.doi.org/10.1109/TVCG.2011.232
3. Rita Borgo, Bongshin Lee, Benjamin Bach, Sara Fabrikant, Radu Jianu, Andreas Kerren, Stephen Kobourov, Fintan McGee, Luana Micallef, Tatiana von Landesberger, and others. 2017. Crowdsourcing for Information Visualization: Promises and Pitfalls. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer, 96–138.
4. Matthew Brehmer and Tamara Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)* 19, 12 (2013), 2376–2385. https://doi.org/10.1109/TVCG.2013.124.
5. Matthew Brehmer, Jocelyn Ng, Kevin Tate, and Tamara Munzner. 2016. Matches, mismatches, and methods: Multiple-view workflows for energy portfolio analysis. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis 2015)* 22, 1 (2016), 449–458. https://doi.org/10.1109/TVCG.2015.2466971.
6. Andreas Buja, John Alan McDonald, John Michalak, and Werner Stuetzle. 1991. Interactive data visualization using focusing and linking. In *Proceedings of the IEEE Conference on Visualization*. 156–163. https://doi.org/10.1109/VISUAL.1991.175794.
7. William S Cleveland and Robert McGill. 1984. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. Amer. Statist. Assoc.* 79, 387 (1984), 531–554. http://dx.doi.org/10.1080/01621459.1984.10478080.
8. Jacob Cohen. 1973. Eta-squared and partial eta-squared in communication science. *Human Communication Research* 28, 473-490 (1973), 56.
9. Patricia Costigan-Eaves, Michael Macdonald-Ross, and others. 1990. William Playfair (1759-1823). *Statist. Sci.* 5, 3 (1990), 318–326. DOI: http://dx.doi.org/10.1214/ss/1177012100
10. Aaron CE Darling, Bob Mau, Frederick R Blattner, and Nicole T Perna. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* 14, 7 (2004), 1394–1403. DOI: http://dx.doi.org/10.1101/gr.2289704
11. Micheline Elias, Marie-Aude Aufaure, and Anastasia Bezerianos. 2013. Storytelling in visual analytics tools for business intelligence. In *Proceedings of the IFIP Conference on Human-Computer Interaction*. 280–297. DOI:http://dx.doi.org/10.1007/978-3-642-40477-1_18
12. Micheline Elias and Anastasia Bezerianos. 2012. Annotating BI visualization dashboards: Needs and challenges. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 1641–1650. DOI: http://dx.doi.org/10.1145/2207676.2208288
13. Niklas Elmqvist, Pierre Dragicevic, and Jean-Daniel Fekete. 2008. Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)* 14, 6 (2008), 1539–1148. DOI: http://dx.doi.org/10.1109/TVCG.2008.153
14. Hannah Fairfield. 2015. The Power of the Reveal. OpenVisConf Talk: https://youtu.be/5A0DczzXyk4.
15. Maria-Elena Froese and Melanie Tory. 2016. Lessons learned from designing visualization dashboards. *IEEE Computer Graphics and Applications (CG&A)* 36, 2 (2016), 83–89. https://doi.org/10.1109/MCG.2016.33.
16. Johannes Fuchs, Fabian Fischer, Florian Mansmann, Enrico Bertini, and Petra Isenberg. 2013. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 3237–3246. https://doi.org/10.1145/2470654.2466443.
17. Michael Gleicher. 2018. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis 2017)* 24, 1 (2018), 1–10. https://doi.org/10.1109/TVCG.2017.2744199.
18. Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D Hansen, and Jonathan C Roberts. 2011. Visual comparison for information visualization. *Information Visualization* 10, 4 (2011), 289–309. http://doi.org/10.1177/1473871611416549.
19. Helwig Hauser, Florian Ledermann, and Helmut Doleisch. 2002. Angular brushing of extended parallel coordinates. In *Proceedings of the IEEE Symposium o Information Visualization (InfoVis)*. 127–130. DOI: http://dx.doi.org/10.1109/INFVIS.2002.1173157
20. Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 203–212. https://doi.org/10.1145/1753326.1753357.
21. Waqas Javed, Bryan McDonnel, and Niklas Elmqvist. 2010. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of PacificVis)* 16, 6 (2010), 927–934. https://doi.org/10.1109/TVCG.2010.162.
22. Microsoft Power BI. https://powerbi.microsoft.com/en-us.

23. Microsoft Power BI Partner Showcase. https://powerbi.microsoft.com/en-us/partner-showcase.
24. Tamara Munzner. 2014. *Visualization Analysis and Design*. A K Peters Visualization Series, CRC press.
25. Donghao Ren, Matthew Brehmer, Bongshin Lee, Tobias Höllerer, and Eun Kyoung Choe. 2017. ChartAccent: Annotation for data-driven storytelling. In *Proceedings of IEEE Pacific Visualization Symposium (PacificVis)*. 1–10. https://doi.org/10.1109/PACIFICVIS.2017.8031599.
26. Jonathan C Roberts. 2007. State of the art: Coordinated & multiple views in exploratory visualization. In *Proceedings of the International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV)*. 61–71. https://doi.org/10.1109/CMV.2007.20.
27. Alper Sarikaya and Michael Gleicher. 2018. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis 2017)* 24, 1 (2018), 1–10. https://doi.org/10.1109/TVCG.2017.2744199.
28. Hans-Jörg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. 2013. A design space of visualization tasks. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)* 19, 12 (2013), 2366–2375. https://doi.org/10.1109/TVCG.2013.120.
29. Drew Skau, Lane Harrison, and Robert Kosara. 2015. An evaluation of the impact of visual embellishments in bar charts. *Computer Graphics Forum (Proceedings of EuroVis* 34, 3 (2015), 221–230. http://doi.org/10.1111/cgf.12634.
30. Tableau Public Gallery. https://public.tableau.com/en-us/s/gallery.
31. Tableau Software. https://www.tableau.com.
32. Justin Talbot, Vidya Setlur, and Anushka Anand. 2014. Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)* 20, 12 (2014), 2152–2160. https://doi.org/10.1109/TVCG.2014.2346320.
33. Stephen Wehrend and Clayton Lewis. 1990. A problem-oriented classification of visualization techniques. In *Proceedings of the IEEE Conference on Visualization*. 139–143. DOI: http://dx.doi.org/10.1109/VISUAL.1990.146375
34. Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only ANOVA Procedures. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 143–146. DOI: http://dx.doi.org/10.1145/1978942.1978963
35. Youneeq Inc. http://www.youneeq.ca.