

Visual Arrangements of Bar Charts Influence Comparisons in Viewer Takeaways

Cindy Xiong, Vidya Setlur, Benjamin Bach, Kylie Lin, Eunye Koh, and Steven Franconeri, *Member, IEEE*

Abstract—Well-designed data visualizations can lead to more powerful and intuitive processing by a viewer. To help a viewer intuitively compare values to quickly generate key takeaways, visualization designers can manipulate how data values are arranged in a chart to afford particular comparisons. Using simple bar charts as a case study, we empirically tested the comparison affordances of four common arrangements: vertically juxtaposed, horizontally juxtaposed, overlaid, and stacked. We asked participants to type out what patterns they perceived in a chart and we coded their takeaways into types of comparisons. In a second study, we asked data visualization design experts to predict which arrangement they would use to afford each type of comparison and found both alignments and mismatches with our findings. These results provide concrete guidelines for how both human designers and automatic chart recommendation systems can make visualizations that help viewers extract the “right” takeaway.

Index Terms—Comparison, perception, visual grouping, bar charts, recommendation systems, computational linguistics

1 INTRODUCTION

Well-chosen data visualizations can lead to powerful and intuitive processing by a viewer, both for visual analytics and data storytelling. When badly chosen, that visualization leaves important patterns opaque, misunderstood, or misrepresented. Designing a good visualization requires multiple forms of expertise, weeks of training, and years of practice. Even after this, designers still require ideation and several critique cycles before creating an effective visualization. Current visualization recommendation systems formalize existing design knowledge into rules that can be processed by a multiple constraint satisfaction algorithm. Tableau and similar products use such rules to decide whether data plotted over time should be shown as lines, or over discrete bins as bars. These systems are useful but rely on overly-simplistic rules that fail to generalize when additional constraints are added, like the goals of the viewer, their graphical literacy level, the patterns being sought, and the relevant patterns in the underlying data.

One fundamental problem with existing recommenders is that they can correctly specify a visualization type, but offer little or no suggestion for how to arrange the data within the visualization. For example, the same data values can be grouped differently by spatial proximity, as shown in Figure 1. These different visual arrangements can lead to different viewer percepts for the same dataset.

Through two studies, we generate a new set of design guidelines around chart visual arrangements so that future research can utilize visual arrangements in visualization recommendation systems to help ensure that a viewer sees the ‘right’ story in a dataset. To study this, we reverse engineer the process by showing people visualizations and asking them to write takeaways. This will help us create a mapping between the visualization and the type of takeaways they tend to elicit, such as the types of visual comparisons people tend to do upon seeing that visualization. This mapping can finally be incorporated in visualization recommendation systems so that we know which visualization to provide that best aligns with their goals.

Contributions: We reconcile linguistic terms used for comparisons

with particular visual arrangements through an empirical study and establish a preliminary visual comparison taxonomy that could be used to categorize comparisons tasks people do with visualizations. We compare the results of our empirical study with expert intuitions, generating design implications that could support natural-language based visualization recommendation tools.

2 RELATED WORK

Designing an effective visualization is a complicated process. Researchers have built tools such as visualization recommendation systems to increase the efficacy of this process and to help viewers more quickly obtain insights from data. However, there remains much to be explored in 1) what visualization to recommend to be considered effective, and 2) how to design the recommendation tool so the users can intuitively interact with it to quickly obtain insights.

2.1 Visualization Affordances

Designing a visualization involves multiple design choices, ranging from which chart type to use to whether to highlight certain visual components or not. These choices can influence how the data being visualized can be perceived, interpreted, and understood [90]. We do not however, have comprehensive knowledge on which design choices can trigger what type of user perception, task, and decisions.

Existing work has largely focused on how *visualization type* can influence viewer perception and decisions. For example, bar charts plot the data points as discrete objects, motivating people to compare them as two distinct units (e.g., A is larger than B), while line charts plot data as one single object, eliciting the interpretation of trends, changes over time, or relations (e.g., X fluctuates up and down as time passes) [101]. Pie charts and histograms are effective for finding extremes; scatterplots are good at promoting clustering; choropleth maps are effective at making comparisons of approximate values, while treemaps encourage identification of hierarchical structures [59]. Chart types that aggregate data points, such as bar charts, can lead viewers to more likely infer causality from data compared to charts that do not, such as scatterplots [98]. Charts that show probabilistic outcomes as discrete objects, such as a beeswarm chart, can promote better understanding of uncertainties [28, 35, 42, 88]. Showing difference benchmarks on bar charts can not only facilitate a wider range of comparison tasks [85], but also increase the speed and accuracy of the comparison [69].

More recently, researchers have referenced knowledge obtained from basic human perception research to generate design guidelines regarding graphical elements, such as the color and shape choices in visualizations. For example, in multi-class scatterplots, viewers can compare between classes of scatterplots more effectively when they are colored differently, compared to when they are plotted with different shapes [20, 31]. However, shape marks can be carefully chosen to maximize distinguishability. For example, open shapes, which are

-
- Cindy Xiong is with UMass Amherst.
E-mail: cindy.xiong@cs.umass.edu.
 - Vidya Setlur is with Tableau Research.
 - Benjamin Bach is with University of Edinburgh, United Kingdom.
 - Eunye Koh is with Adobe Research.
 - Kylie Lin and Steven Franconeri are with Northwestern University.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.
Digital Object Identifier: xx.xxx/TVCG.201x.xxxxxx

composed of line segments that are bounded across a region of space, such as an asterisk, are more differentiable than closed shapes [12,39]. Additionally, orienting the shapes so that they point to the direction of the regression can facilitate correlation estimations in scatterplots [61]. Visualizations are often presented in multiples so analysts that can explore different combinations and compare interesting patterns [72]. Hence, the spatial arrangement of the data point is a critical design choice, but is so far under-explored.

We identified four common spatial arrangements used to represent multiple views: vertical, overlaid, adjacent, and stacked, as shown in Figure 1. Not only are these four arrangements commonly used, they also spatially align different pairs of data variables.

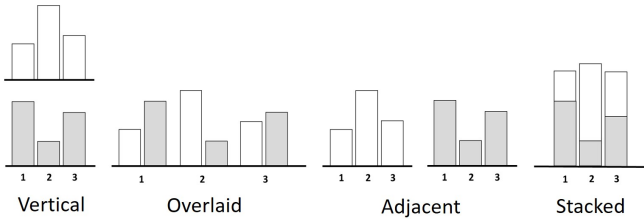


Fig. 1. Four spatial arrangements examined in the study.

2.2 Visualization Recommendation Tools

Visual analysis tools need to help people gain insights quickly by providing reasonable visualizations in order to be useful in their analytical workflows. Visualization recommendation (VisRec) systems can be classified based on whether they suggest visual encodings (i.e., encoding recommenders) or aspects of the data to visualize (i.e., data-based recommenders) [95]. The earliest VisRec systems focused on recommending visual encodings, assuming that the data attributes were already identified by the user [63, 64]. Subsequent work automatically recommended graphical encodings based on perceptual effectiveness [48, 68]. Many recent papers suggest interesting visualizations based on statistical properties of the data [18, 55, 80, 84, 92–94].

Some VisRec systems are entirely automatic [8, 94], whereas other more recent mixed-initiative systems support some user interaction to guide the recommendations [18, 84, 97]. Mixed-initiative VisRec systems combine manual specification with recommendations [38, 46, 53, 60, 91, 100]. For instance, Voyager [95–97] suggests visualizations based on user-selected fields and wildcards (e.g., a categorical wildcard will enumerate over all fields that are of categorical data type) to iterate through possible data attributes or encodings. Lee et al. [58] present a taxonomy of analytical actions used to generate and group visualization recommendations in existing systems [8, 16–18, 38, 53–55, 57, 60, 78, 80, 86, 91, 92, 94, 96, 97]. These analytical functions include distribution, correlation, enhance, generalize, pivot, and filter. Based on the analytical function, the user’s context, or data data interestingness, VisRec systems can provide a specific visualization recommendation [16–18, 55, 56]. Still, none of these systems focus on how to best provide recommendations specifically for facilitating visual comparison and offer little or no suggestions for how to arrange the data within the visualization. In this paper, we address this gap in VisRec systems by better understanding how visual arrangements affect the readers’ takeaways during their analysis and the types of comparisons that are made based on these visual arrangements.

2.3 Comparisons in Computational Linguistics

The ability to establish orderings among objects and make comparisons between them according to the amount or degree to which they possess some property is a basic component of human cognition [44]. Natural languages reflect this fact: all languages have syntactic categories (i.e., words in a language which share a common set of characteristics) that express gradable concepts, i.e., expressing explicit orderings between two objects with respect to the degree or amount to which they possess some property (e.g., “the temperatures in the Death Valley are *higher than* in Bangalore in the summer”) [77]. Research

in computational linguistics has explored the semantics of comparison based on gradable concepts [11, 15, 34, 43, 49, 79]. Bakhshandeh and Allen presented a semantic framework that describes measurement in comparative morphemes such as ‘more’, ‘less’, ‘-er’ [10].

The notion of vagueness in comparative language has been studied in the computational linguistics community [9, 25]. The semantics of comparatives can be vague as their interpretation depends on the context and the boundaries that make up the definition of the comparative. For the example, “coffee and doughnuts in the Bay Area are more expensive than in Texas,” is the statement about whether those items are more expensive *on average*, or whether both items are *each individually* more expensive? Research has focused on the conceptualization and representation of vague knowledge for these comparatives. Kessler and Kuhn present a corpus of annotated comparison sentences from English camera product reviews. Each sentence contains the comparative predicate that expresses the comparison, the type of the comparison, the two entities that are being compared, and the aspect they are compared in [45].

While linguistic vagueness has been explored for comparative expressions along with their semantic variability, little work has been done in determining how best to *visually* represent comparatives based on these variations, especially in the context of visual analysis. Our work explores the types of comparisons readers make and their inherent ambiguities when comparing bar charts in different configurations.

2.4 Natural Language Interfaces for Visual Analysis

Natural language (NL) interfaces for visualization systems [1–4] attempt to infer a user’s analytical intent and provide a reasonable visualization response. The methods of inferring intent typically rely on explicitly named data attributes, values, and chart types in the user’s input queries. DataTone [26] provides ambiguity widgets to allow a user to update the system’s default interpretation. Eviza [81] and Analyza [19] support simple pragmatics in analytical interaction through contextual inferencing. Evizeon [37] and Orko [87] extend pragmatics in analytical conversation. The Ask Data system [82] handles various analytical expressions in NL form such as grouping of attributes, aggregations, filters, and sorts. The system also handles impreciseness around numerical vague concepts such as ‘cheap’ and ‘high’ by inferring a range based on the underlying statistical properties of the data. Hearst et al. [36] explore appropriate visualization responses to singular and plural superlatives (e.g., ‘highest price’ and ‘highest prices’) and numerical graded adjectives (e.g., ‘higher’) based on the shape of the data distributions. Law et al. [52] investigated how the visual design of answers to *why* questions might influence user perceptions of a question-answering system. They found that users have a strong tendency to associate correlation with causation when systems do not provide clear explanations about the answers.

The space of analytical expression in an NL context is rich and much more nuanced than what these interfaces currently support. A study was conducted to assess NL input to visualization systems [82] where 75 participants were asked to write NL queries based on five underlying datasets (i.e., bird strikes, world indicators, superstore, mutual funds, and Olympic medals). Of the 578 NL queries, common ones included “Are there more strikes on takeoff or landing?”, “Are certain seasons more dangerous?” and “Which country has more female medalists?”. The data suggests that when participants were not restricted in the format of expression, they often chose to specify utterances with an underlying intent to either explicitly or implicitly compare values in a dataset. Current NL systems however, do not deeply explore how utterances about comparisons ought to be interpreted even though such forms of intent are prevalent. This paper explores different ways users express takeaways that compare bars in variants of visual arrangements. The implications of our work help inform NL interfaces with guidelines towards reasonable visualization responses based on the types of comparisons users specify in their utterances.

2.5 Comparisons in Visualization

Representing comparisons in data visualizations is an important aspect of supporting the user in their analytical workflow. Small multiples

make it easier to view objects side-by-side [5] or examine juxtaposed views through multi-view coordination [75]. Tufte discussed small multiples as a way to use the same graphic to display different slices of a data set for comparison [89].

Prior work surveyed a variety of visualization solutions to support comparisons. Graham and Kennedy [33] surveyed a range of visual mechanisms to compare trees, while other surveys consider methods for comparing flow fields [70]. Gleicher et al. [30] presented a broad survey with over 100 different comparative visualization tools from information visualization domains, organized by their comparative visual designs into a general taxonomy of visual designs for comparison. Designs were grouped into three categories: juxtaposition, superposition, and explicit encodings.

The perceptual and cognitive science communities have considered the problem of visual comparison for decades, including the issues around change blindness [74]. Other studies focused on comparing methods for specific tasks, such as the evaluation of scalar field comparisons [62] or brain connectivity graphs [7]. Franconeri [22, 23] discussed several limitations in the mechanisms of perception that may have direct impact on the design of comparison methods. For example, translations of an object are easy to compare, but texture, orientation, scale, space, and time may complicate comparison tasks [50, 51].

Comprehension of visual comparisons is an important aspect of determining their efficacy. Shah and Freedman [83] investigated the effect of format (line vs. bar), viewers' familiarity with variable, and their graphical literacy skills on the comprehension of multivariate (three variable) data presented in graphs. Findings show that line and bar chart features have a substantial influence on viewers' interpretations of data. The differences between people's perceptions of bar graphs and line graphs can be explained by differences in the visual chunks formed by the graphs based on Gestalt principles of proximity, similarity, and good continuity. Jardine et al. [40] conducted an empirical evaluation on two comparison tasks: identify the "biggest mean" and "biggest range" between two sets of values. Their work showed that visual comparisons of largest mean and largest range are most supported by vertically stacked chart arrangements and this pattern is substantially different across different types of tasks.

Another aspect of comprehension is exploring the interplay between visualization design and captions on readers' takeaways. Post-Grapher was a system that displayed a text and graphic pair considering the writer's goals, types and values of the variables to be presented, and the relations between variables [21]. Other work [13, 65] focused on developing and evaluating models for figure captioning. Kim et al. explored how readers integrate line charts and captions [47]. Their studies found that when readers are presented with charts and captions that describe a high-prominence feature, they treat the doubly emphasized high-prominence feature as the main takeaway from the chart. Conversely, when the caption describes a low-prominence feature, readers report a higher-prominence feature as the main takeaway.

Our work specifically focuses on how visual arrangements influence readers' takeaways when they are making comparison judgments with bar chart arrangements.

3 STUDY MOTIVATION AND OVERVIEW

In this paper, we investigate comparison affordances of four spatial arrangements of bar charts by showing crowdsourced participants bar charts and asking them to write sentences describing their most salient takeaways. We analyzed these written takeaways to create a mapping between the visualization arrangements and the takeaways they tend to elicit, such as the types of visual comparisons people tend to do upon seeing that visualization. In experiment 2, we compare our data-driven mappings with expert intuitions and generate design guidelines for visualization recommendation systems in order to recommend the spatial arrangement that best aligns with user intents and goals.

4 ELICITING VIEWER TAKEAWAYS IN NATURAL LANGUAGE

One critical challenge in investigating viewer affordances is how to elicit viewer percepts when they interact with the visualizations. A

dataset can contain many patterns to perceive [99]. For example, looking at the top panel in Figure 2, one could notice that both reviewers gave higher scores to *A* and lower scores to *B*. Alternatively, one could notice that the differences in scores given to *A* and *B* is smaller for Reviewer 2 and bigger for Reviewer 1. To communicate what patterns and relations one extracted from these visualizations, the viewer has to generate sentence descriptions of the pattern or relation, such as "*A* is greater than *B*," or "the difference between *X* and *Y* is similar to the difference between *P* and *Q*."

In order to examine affordances of different visualization spatial arrangements and create a mapping between viewer takeaways and the arrangements, we need to interpret and categorize the types of patterns and relations viewers take away from the visualizations. But, we end up facing similar challenges to that of the natural language and linguistics communities [29]. Specifically, the sentences the viewers generate to describe their percepts/takeaways in visualization can be ambiguous. There are three types of ambiguity in natural language: lexical, syntactic, and semantic [29]. Figure 2 provides an example of each type of ambiguity and how they map to different visual comparisons in the same visualization.

Lexical ambiguity represents instances when the same word is used to represent different meanings. In our study, we encountered situations where the participants used words such as "spread", which can be interpreted differently depending on their intent. As shown in Figure 2, "spread" can be interpreted as either the amount of variability in data, or the range of the data as shown in the left and right interpretations in Figure 2. These two definitions would map to separate comparison patterns. This lexical ambiguity with the word "spread" has been discovered in education setting with students as well [41].

Syntactic ambiguity occurs when there exists multiple ways to parse a sentence. For example, the takeaway "East makes more revenue from Company A and B" could be parsed as "East makes more revenue from (Company A and B)", or "East makes more revenue from Company A and (B)." The corresponding visual comparison the viewer could have done is shown in the middle panel in Figure 2. The viewer could have looked Company A and B holistically and notice that the average or combined values of the East branches is higher than that of the West branches. Alternatively, the viewers could have individually compared pairs of bars, noticing that in Company A, the East branch has a higher revenue than the West and that in Company B, the East branch has a higher revenue than the West.

Semantic ambiguity occurs when multiple meanings can still be assigned to the sentence despite being neither lexically nor syntactically ambiguous. For example, as shown in the bottom panel of Figure 2, "Bacteria 1 and Bacteria 2 are the opposite of each other" can be mapped to two comparisons. The first could be a comparison between *A* and *B* in Bacteria 1 and a comparison between *A* and *B* in Bacteria 2, where the former has a smaller than relationship, and the latter has a larger than relationship. The second could be a comparison between Bacteria 1 and 2 in *A* and another between Bacteria 1 and 2 in *B*.

Since there does not exist natural language processing tools nor existing visual comparison taxonomies to aid our interpretation of chart takeaways, we could not automate the process. We had to manually read every sentence, infer the intent of the participant, and then connect the sentence to a visual pattern in the visualization. The ambiguity in these sentence descriptions can still be vague to even a human interpreter, so we also asked participants to annotate for us which chart component they compared to the best of their abilities. The human interpreter (or researcher, in our case) of these sentences could refer to these drawings and annotations to resolve ambiguities in the sentences. We decided to implement this method after a series of pilot experiments where we failed to comprehensively and accurately capture participant percepts when they viewed visualizations. We describe these failures with the hope that they can inspire future researchers to better capture viewer percepts or takeaways in visualizations.

Attempt 1: We initially thought that human interpreters of viewer-generated sentences would have little problem resolving the ambiguities in language; unlike machines, we are capable of inferring intention, understanding implicit comparisons, and correcting obvious er-

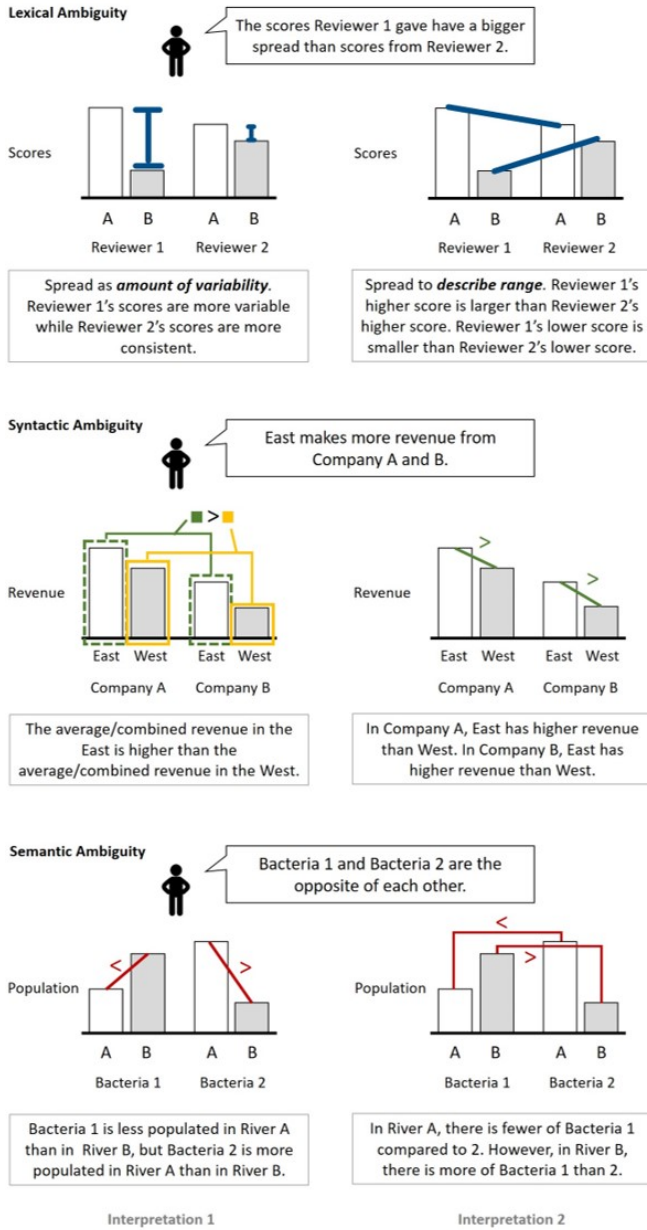


Fig. 2. Three linguistic ambiguities for various visual comparisons.

rors in text. We realized quickly that this was not the case and when a researcher read sentence descriptions as listed in Figure 2, they could not reverse engineer the visual patterns the participants extracted.

Attempt 2: We realized that we needed to ask our participants for more context than just sentence descriptions. If we knew which data values in a visualization they looked at or which pairs of data values they compared, the majority of the ambiguous cases could be resolved. After our participants generated sentence descriptions of the patterns they extracted from a visualization, we asked them to also indicate the data values they compared via a multiple choice task. Consider the chart in the bottom panel of Figure 2 as an example, the participant would be able to select a subset from the list 'Bacteria 1 A', 'Bacteria 1 B', 'Bacteria 2 A', and 'Bacteria 2 B' to indicate the ones they looked at and compared. However, most comparisons ended up containing the entire set (e.g., a comparison of A1 to B1, and then A2 to B2). In these scenarios, the multiple choice task ends up being uninformative as the participant would select all options in the entire list, because they compared every data value.

Attempt 3: A sentence typically unfolds as a comparison of two groups in which one group is the 'referent' and the other the 'target' [14, 24, 32, 76]. The target and the referent are connected by a relation. In the sentence "East makes more revenue than West in Company A," the revenue of East A is the target and the revenue of West A is the referent. The relation is 'greater than.' This process applies to both natural language and to visual comparisons across data values in a visualization [66, 67, 83]. To improve upon Failure 2, we separated the question where participants indicate which data values they compared into three questions so that they could indicate which values were the target, which were the referent, and the relation between them.

We piloted with 20 participants, including both crowd-sourced workers from Prolific.com [71] and undergraduate students enrolled in a research university (data included in the supplementary materials) and learned that while most people are able to generate sentence describing their percepts, they could not figure out which data values they compared to which other data values. They especially struggled with scenarios when their sentences contained implicit comparisons, such as "there is a decreasing trend from left to right" and "West A has the second highest revenue." Both cases could be translated into target, referent, and relation in multiple ways. For example, assuming that the participant noticed that the bars became smaller from left to right, the decreasing trend could involve a comparison of the left-most bar to the second left-most bar with the former bigger than the latter. In this case, the target is the left-most bar, the referent is the second left-most bar, and the relation is 'bigger than.' Alternatively, the participant could have compared the decreasing trend (the target) to an imagined horizontal line that is not decreasing (the referent). The training process quickly evolved to becoming more complex and its duration became less proportional to its effectiveness. We additionally collected data on participants' confidence as they translated their sentences and we observed consistent low confidence in their own translations.

Attempt 4: Inspired by the relation component in the Failure 3, we recognized that mathematical expressions such as ' $A > B$ ' contain all three elements of target, referent, and relation. Writing similar simple expressions also seems more intuitive than segmenting your own sentences into an unfamiliar new sentence structure. Mathematical expressions tend to be far less ambiguous compared to the English language. In this attempt, we asked people to write pseudo mathematical expressions to reflect the data values they compared or the pattern they've noticed. We provided examples such as ' $A = C$ ' (A is not equal to C), ' $A > B > C$ ' (decreasing from A to B to C), and ' $\max = A$ ' (A is the biggest bar) to get people started. After piloting 10 participants, who were students at a research university, we realized that this likely would not scale efficiently to crowd-sourced participants. Participants' expressions varied depending on the type of programming languages they were familiar with. Additionally, some participants wrote sentences with conjunctions with varied semantics of how they used 'and', 'or', and 'but' in their representations. Some sentences were just difficult to be intuitively represented as a mathematical expression, such as "the population is the same for both rivers, but for different bacteria types." We discuss challenges in translating natural language sentences descriptions of visualization takeaways to mathematical expressions in Section 8.

Attempt 5: This method was a success, but a temporary solution nonetheless. This is the version where we asked participants to write a sentence description and attach a digital drawing annotating the specific patterns they noticed or data values they have compared. What we ended up with was over a thousand sentences and drawings that our later-reported findings are based. We could alternatively refer to it as Failure 5, because it is more of an imperfect intermediate solution than a success – the method required dozens of hours of manual interpretation from multiple people to ensure that viewer intent is captured accurately and consistently.

5 EXPERIMENT 1 CROWDSOURCING TAKEAWAYS

In Experiment 1, we investigated the comparison affordances of four common arrangements in bar charts: vertically juxtaposed, horizontally juxtaposed, overlaid, and stacked. We asked participants to type



Fig. 3. Two datasets used to generate the bar charts, showing the overlaid arrangement as an example.

out what patterns they perceived and qualitatively coded their takeaways into types of comparisons. We then created a mapping between the visual arrangements and the comparisons they tend to afford.

5.1 Participants

We recruited 76 participants via Prolific.com [71]. They were compensated at 9 USD per hour. In order to participate in our study, the workers had to be based in the United States and be fluent in English. After excluding participants who had failed attention checks or entered illegible information for the free-response questions, we ended up with 45 participants ($M_{age} = 25.22$, $SD_{age} = 7.23$, 32 women).

5.2 Methods and Procedure

We generated two datasets for the four spatial arrangements illustrated in Figure 1, creating eight total visualizations. Each visualization depicts two groups of three data points, as shown in Figure 3. For example, the chart could be showing the sales of two ice cream flavors (Flavor A and Flavor B) in three different markets (Market 1, Market 2, and Market 3). In our analysis, we will refer to the two groups as ‘groups’ and the three data points within each group as ‘elements.’

We created a within-subject experiment where each participant viewed all eight of the visualizations and wrote their top two main takeaways for each visualization. They were also asked to annotate their takeaways on the bar visualization by drawing circles around the bars they mentioned or using mathematical operators (e.g., $>$, $<$, $=$) to represent the patterns they saw. We examined the sentence takeaways to extract the patterns and identify the comparisons participants made upon seeing the visualizations. The takeaways and corresponding drawings can be found in the supplementary materials.

To distract the participants from noticing similarities in patterns between the charts, we added distractor tasks (e.g., demographic and visual literacy questions) between each visualization and provided each of the eight charts presented with a different context, as shown in Figure 3. We randomized the order of the charts such that the charts alternated between the two datasets and participants never saw the same spatial arrangements in back-to-back trials. Additionally, we asked at the end of the survey if the participants noticed anything unusual or have any comments regarding the visualizations shown in the survey; six out of the 48 participants mentioned that they noticed similar patterns across the visualizations seen. They mentioned that “many of the charts were the same, that’s why I gave the same answer” and “some charts were recycled as the study advanced.” We decided to keep these participants’ data since only a few noticed similarities between our stimuli, and we discuss limitations in our approach in Section 8.

5.3 Comparison Classification and Coding Approach

We took a top-down approach and identified 12 possible comparisons one could make to generate takeaways from bar charts, as shown in Figure 4. Previously, we mentioned that the charts we have shown participants all depict two groups (A,B) with three elements (1, 2, 3) in each group. A comparison could be made *across group*, which means the viewer compared something in group A to something in group B, or it could be made *within group*, which means the viewer compared something in group A to something else in group A, or compared something in group B to something else in group B. The comparison could also be classified as *within element* or *across element*. A *within element* comparison compares the same elements between two groups, such as comparing element 1 in group A to element 1 in group




B. An *across element* comparison compares different elements, such as comparing element 1 to element 2. This could be within the same group, which means element 1 in group A is compared to element 2 in group A, or across different groups, which means element 1 in group A is compared to element 2 in group B. To summarize,

- **across group** indicates that the comparison is made between group A and group B, while
- **within group** indicates that the comparison is made within the same group.
- **across element** indicates that the comparison is made among different elements, while
- **within element** indicates that the comparison is made within the same element.




A viewer could mix-and-match their comparison operations for groups and elements in four ways: *across group - within element*, *across group - across element*, *within group - across element*, and *within group - within element*. However, the last comparison operation, *within group - within element*, does not apply in most scenarios since it requires a viewer to compare the same element in the same group, such as comparing element 1 in group A to element 1 in group A. This is just a comparison of a data point to itself, therefore not interesting and we omit it from our classification system. This leaves three ways a participant could mix and match their operations comparing groups and elements, as indicated by row labels in Figure 4. To summarize,

- For an *across group - within element* comparison, the same element is identified in each group and compared to one another.
- An *across group - across element* comparison means the viewer identifies one element from one group and compares it to another element in a different group.
- For *within group - across element* comparisons, the viewer zooms in on one group, and compares different elements within that same group.

5.3.1 One-to-One Comparisons (1:1)

Viewers can compare two individual bars in their takeaways, and we refer to them as “one-to-one” comparisons, as shown in the leftmost column in Figure 4, comparison types 1, 5, and 9. For *across group - within element* operations, one-to-one comparison means that the viewer compares one element in one group to the same element in another group, such as comparing element 1 in group A (which we will refer to as A1) to element 1 in group B (which we will refer to as B1) . For *across group - across element* operations, the viewer compares one element in one group to another element in a different group, such as comparing A1 to B2 . For *within group - across element* operations, one-to-one comparison means that the viewer compares one element in one group to another element in the same group, such as comparing A1 to A3 .

5.3.2 Two-to-Two Comparisons (2:2)

Viewers can alternatively visually group together two bars and compare them as a set to another set of visually grouped two bars, which we refer to as “two-to-two” comparisons, as shown in the second column in Figure 4, comparison types 2, 6, and 10. These differ from one-to-one comparisons because the viewer is no longer comparing individual values, but rather comparing the sum/difference of two elements to the sum/difference of two other elements. For example, for an *across group - within element* two-to-two comparison, the viewer would compare element 1 to element 3 overall . Using the two ice cream flavor sales across three markets example from before, a comparison of this type will say “the overall sales in market 1 considering both flavors is lower than the overall sales in market 3.” For an *across group - across element* two-to-two comparison, the viewer will compare a set of two different elements (one from each group) to another set of two different elements (one from each group) . For a *within group - across element* two-to-two comparison, the viewer will compare a set of two different elements from the same group to another set of the same two elements from the other group .

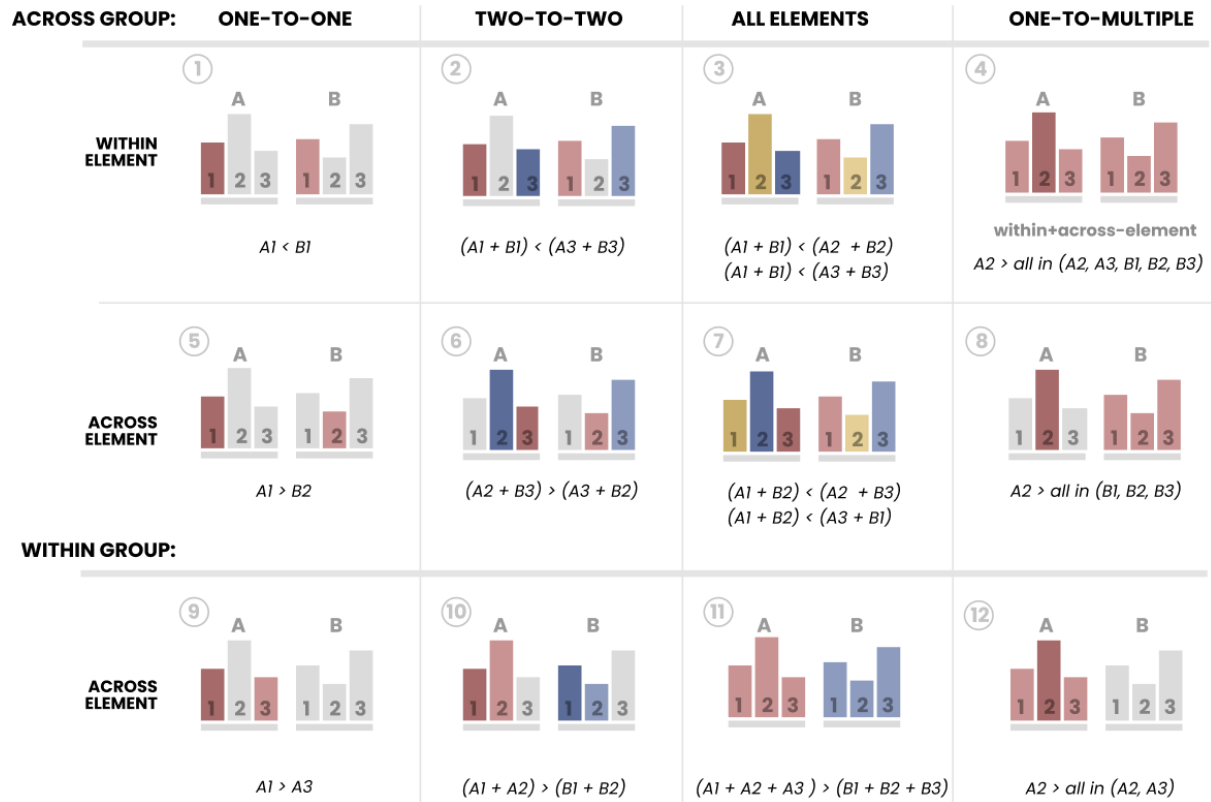


Fig. 4. Twelve categories of comparisons in two by three bar charts.

5.3.3 All Elements (All)

The previously mentioned two categories involve the viewer comparing a subset of the data in the chart. The viewer can also visually group together a set of bars and compare that set to the remaining data points, which we refer to as “all element” comparisons (third column in Figure 4), comparison types 3, 7, and 11. For this category, an *across group - within element* comparison would involve the viewer visually grouping the element 1s together, and compare them to the element 2s and 3s. Examples of this type of comparison might be the viewer identifying that the set containing element 1s is overall the smallest compared to element 2s and element 3s (e.g., considering both ice cream flavors, market 1 has the lowest amount of sales compared to market 2 and 3). An *across group - across element* comparison happens when the viewer groups together two different elements, one from each group (e.g., A1 and B2) and compares them to the other pairs of elements. Since this category requires the elements to not be matching between the two groups, it can seem arbitrary. Our data supports this point as this type of comparison is extremely rare among viewers. Finally, a *within group - across element* comparison involves the viewer to visually group together all the elements in A and compare their sum/differences to the sum/differences of all the elements in B as a whole.

5.3.4 One-to-Multiple (1:M)

The last category of comparisons is “one-to-multiple” comparisons, as shown in the fourth column in Figure 4 - types 4, 8, and 12, where participants identify one data point and simultaneously compare it to multiple other bars. People typically do this type of operation when they rank the bars by value (e.g., B3 is the second highest), or when they identify extrema such as maximums or minimums. We refer to the scenario where the viewer picks out one bar and compares it to the rest of the bars as an *across group - within/across element* comparison because the comparison happened both within the same element (e.g., comparing A2 to B2) and across different elements (e.g., comparing A2 to A1, to A3, etc.). An *across group - across element*



comparison type involves the viewer identifying one element from one group, and comparing it to multiple elements in the other group. Lastly, a viewer can identify one element within one group and compare it to all of the other elements in the same group, and we refer to this as a *within group - across element* comparison.

5.4 Making Sense of Participant Takeaways

To make sense of participant takeaways, we analyzed their verbal description in conjunction with their drawings. Most takeaways involved a comparison between two chart elements, so we identified what they compared in each takeaway and what relationship described that comparison, mapping that comparison to one of our twelve categories. For example, for the takeaway “it looks like flavor B sold more than flavor A in market 2”, the two chart elements compared would be B2 (flavor B in market 2) and A2 (flavor A in market 2), and the relation is ‘greater than.’ Since this is a comparison of one individual bar to another individual bar, this would be an one-to-one comparison. Additionally, since the element is fixed (market 2 for both) and the group is changing, this would be an *across group - within element* comparison (type 1). Two authors participated in this qualitative coding process and double coded all responses. They agreed 89.5% of the time in their ratings, with a high inter-rater reliability Kappa value of 0.867 ($z = 72.3, p < 0.001$).

Some participants would make multiple accounts of the same type of comparison upon seeing one chart, such as making two instances of across group - within element one-to-one comparisons (see 1 in Figure 4). For example, one participant wrote “I noticed that in market 1, flavor B sold more than flavor A, and in market 2, flavor A sold more than flavor B.” This participant compared B1 to A1, as well as compared B2 to A2. In situations like this, we removed the duplicate and counted this participant as having done a one-to-one *across group - within element* comparison upon seeing the visualization.

We also noticed that some participants mentioned conjunction comparisons in their takeaways, touching on two comparison categories, such as saying “flavor B sold more than flavor A in market 1, but flavor

A in market 1 sold more than flavor A in market 3.” In this example, the first part of this comparison is a one-to-one *across group - within element comparison* , and the second part of this comparison is a one-to-one *within group - across element comparison* . Conjunctions like this that cover different types of comparisons were broken up and treated as separate takeaways from the same participant.

5.5 Comparison Overview

We collected a total of 584 chart takeaways with 53.09% of them being conjunction comparisons, with a total of 1100 comparisons collected from participants. Figure 5 and Table 1 summarize the number of comparisons participants made for each category and shows the frequencies of the 12 types of comparisons across the four visualization arrangements.

		%	Examples
1	Type 1	27.82%	compare one element in a group to the same element in the other group
2	Type 3	16.73%	compare one element to the other two elements considering both groups
3 (tie)	Type 11	13.36%	compare the group A as a whole group B as a whole
3 (tie)	Type 12	13.36%	compare one element in one group to the other elements in the same group
5	Type 4	7.64%	select one data point and compare it to all the other data points
6	Type 9	7.36%	compare one element in one group to another element in the same group
7	Type 2	5.00%	compare one element to one other element considering both groups

Table 1. Overview ranking the most frequently made comparisons from Experiment 1 with examples.

Participants most frequently made the *one-to-one across group - within element* comparisons (type 1 in Figure 4). Using the ice cream flavor sales across three markets example, where the two groups are flavors A and B and the three elements are markets 1, 2, and 3, participants most often wrote takeaways saying “I noticed that in Market 1, ice cream Flavor A sold less than Flavor B.”

The second most frequently made comparison was *all elements across group - within element* (type 3). For example, “the total sales in Market 1 is smaller than the total sales in Market 2, and it’s also smaller than the total sales in Market 3.”

The third and fourth (tie) most frequently made comparisons were *all elements within group - across element* (type 11), such as saying “overall, flavor A sold more than flavor B,” and *one-to-multiple within group - across element* (type 12), such as saying “for flavor A, market 2 sold more ice cream than market 1 and market 3.”

The fifth most frequently made comparison was *one-to-multiple across group - within/across element* (type 4). This type of comparison mostly involved identification of the maximum or the minimum points, or a ranking of data values. For example, “Flavor B in Market 2 is the least sold flavor, considering all the flavors and all the markets.”

The sixth most frequently made comparison was *one-to-one within group - across element* (type 9). For example, one participant said “For Flavor A, Market 1 sold more than Market 3 did.”

The seventh most frequently made comparison was the *two-to-two across group - within element* comparison (type 2). This is similar to the all element across group - within element comparison (type 3), except that the participant only compared one element to one other element. For example, “Market 1 has better business than Market 3.”

Participants very rarely did a comparison of the remaining five categories (type 5, 6, 7, 8, 10), making up less than 5% of all comparisons each. No participants made *all elements across group - across element* comparisons (type 7), which means no one visually grouped non-matching elements from the two groups and compared them. Only one participant made a *one-to-multiple across group - across element* comparison (type 8), where they compared the biggest bar in group A

to all of the bars in group B. The *one-to-one across group - across element* comparisons (type 5) that participants made may seem arbitrary, but they were always comparing two bars of similar sizes together.

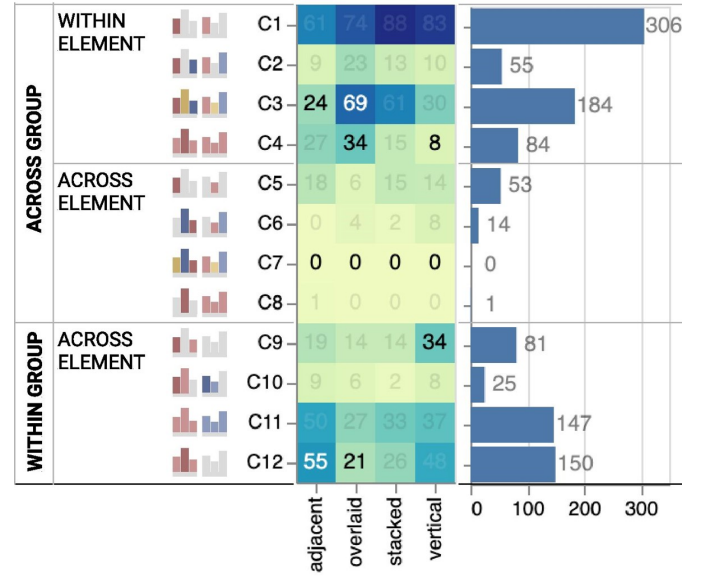


Fig. 5. Heatmap showing occurrences of categories for all techniques as well as in total across all techniques (barchart). Opaque values in the heatmap indicate significant values.

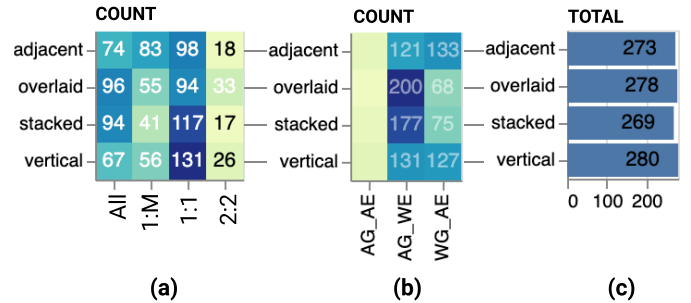


Fig. 6. Heatmaps showing (a) the total counts of techniques for each group (1:M, 1:1, 2:2), (b) per groupings (AG, WG, AE, WE)[no significant counts], and (c) the total count of observations across techniques. Opaque values in the heatmap indicate significant values.

5.6 Visual Arrangements and Number of Bars Compared

We analyzed whether different visual arrangements impacted whether a viewer would do one-to-one (1:1), two-to-two (2:2), all elements (All), or one-to-multiple (1:M) comparisons (the four columns in Figure 4). We conducted a Chi-Square analysis and, as shown in Figure 6, found that the only significant relationship is that the adjacent arrangement affords more one-to-multiple (1:M) comparisons, while the stacked arrangement affords fewer such comparisons ($\chi^2 = 38.405$, $p < 0.001$). This means visual arrangement in general does not affect how many bars people compare in a chart.

5.7 Across/Within Group and Element Comparisons

We also examined whether visual arrangements influence whether a viewer would compare *across group - within element*, *across group - across element*, or *within group - across element* (the three rows in Figure 4). We conducted a Chi-Square analysis and, as shown in Figure 6, found that participants were significantly more likely to make *across group - within element* comparisons (first row), and *within group - across element* (third row) comparisons ($\chi^2 = 65.39$,

Table 2. Summary results from Experiment 1 (crowdsourced comparisons) compared to summary results from Experiment 2 (expert intuitions).

Type	Comparison	Comparison Goal	Crowdworkers	Experts
1) Type 1	compare the same element across two different groups	compare the sales revenue in market 3 from ice cream flavor A to the revenue in market 3 from ice cream flavor B.	all the same	vertical
2) Type 3	compare one element to other elements, considering both groups	compare the average sales revenue in market 1 to the average sales revenue in market 2, across both ice cream flavors	overlaid	stacked
3) Type 4	compare one bar to all other bars in the chart, which is often a superlative comparison	identify the lowest single sales revenue rating among all six revenues	overlaid	overlaid
4) Type 9	compare one element in one group to another element in the same group	compare the sales revenue of ice cream flavor B in market 1 to the sales revenue of ice cream flavor B in market 3	vertical	overlaid
5) Type 11	compare one group to another holistically	compare the overall ice cream sales in market A to the overall ice cream sales in market B	all the same	all the same

$p < 0.001$). Very few people made *across group - across element* comparisons (second row).

More specifically, overlaid arrangements were most likely to trigger an *across group - within element* comparison (first row, $p < 0.001$), but are least likely to trigger a *within group - across element* comparison (third row, $p < 0.001$). Adjacent arrangements were most likely to trigger a *within-group - across element* comparison (third row, $p < 0.001$), and were least likely to trigger an *across group - within element* comparison (first row, $p < 0.001$). This means that people were more likely to identify the same element and compare their values across two different groups when they view bar charts in the overlaid arrangements, and are more likely to focus on one group and compare elements within that group when they view adjacent charts.

5.8 Visual Arrangements and Comparison Categories

We examined how visual arrangements impacted the likelihood of our participants making each of the twelve types of comparison via a Chi-square analysis and found a significant effect of visual arrangements on viewer comparisons ($\chi^2 = 132.25$, $p < 0.001$). We visualize the number of comparisons participants made in each of the 12 categories for the four visual arrangements in Figure 5.

Post-hoc analysis with Bonferroni’s correction revealed that some visual arrangements particularly elicit certain comparison types. As summarized in Table 2, overlaid arrangements especially afford type 3 (all element, *across group - within element*, $p = 0.001$) and type 4 comparisons (one-to-multiple, *across group - within element*, $p = 0.037$). Vertical arrangements afford type 9 comparisons (one-to-one, *within group - across element*, $p = 0.017$). Adjacent arrangements afford type 12 comparisons (one-to-multiple, *across group - within element*, $p = 0.013$).

Some visual arrangements are also particularly bad at eliciting certain comparison types. Participants were the least likely to make type 3 comparisons with adjacent arrangements ($p = 0.002$). Vertical arrangements were the least likely to trigger type 4 comparisons ($p = 0.021$), and overlaid arrangements were the least likely to trigger type 12 comparisons ($p = 0.028$).

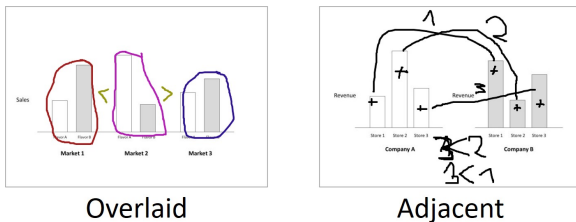


Fig. 7. Left: annotating a type 3 comparison in overlaid arrangement. Right: annotating a type 3 comparison in adjacent arrangement.

We share some participant drawings in Figure 7. You can see that the amount of effort the participant had to put into visually representing their comparisons differed between comparison type and arrangements, which also reflects the differing comparison affordances of the visual arrangement types. For overlaid arrangements, visually annotating a type 3 comparison was simple, whereas visually annotating the same comparison in an adjacent arrangement was much more complex. This corroborates with our finding that participants were more likely to make type 3 comparisons when viewing overlaid arrangements, and less likely to do so when viewing adjacent arrangements.

Surprisingly, although type 1 and type 11 comparisons were among the most frequently made comparisons (see Table 1), we did not find any difference in the likelihoods of participants making them between the four visual arrangements.

5.9 Discussion

In summary, our findings suggest that visual arrangements afford specific comparisons. Vertical arrangements and adjacent arrangements best afford comparisons that involve comparing one element in one group to another element in the same group (type 9: and type 12:). Overlaid arrangements best afford comparisons that involve comparing one bar to all other bars in the chart (which are most often superlative comparisons, type 4:), as well as comparisons that involve comparing one element to other elements, considering both groups (type 3:). Although comparisons that involve comparing one group to another holistically (type 11: and comparisons that involve comparing the same element across two different groups (type 1: are popular comparisons to make, no particular visual arrangements especially affords these comparisons.

In Experiment 2, we present the above five comparison scenarios to data visualization experts and compare their choice of the best visual arrangement to the results of our crowd-sourced experiment.

6 EXPERIMENT 2 EXPERT INTUITIONS

In Study 2, we showed data visualization experts bar charts in the four different visual arrangements and asked them which one they would choose to facilitate a specific type of comparison. We identify both alignment and mismatches between our empirical data from Experiment 1 and expert intuitions in Experiment 2.

6.1 Participant

We recruited 45 visualization expert volunteers from Adobe and Tableau ($M_{age} = 37.88$, $SD_{age} = 13.14$, 16 women) to complete an online survey through Qualtrics [73]. They reported their experience with visualizations in multiple-choice, multiple-answer questions. Among our participants, for those that chose to complete the demographic questionnaire at the end of the survey, 22 people stated that they were data analysts who used visualizations frequently or were visualization researchers, 7 people said to have taken at least one visualization design course, and 3 people indicated to be engaged with

You want to spontaneously compare the overall durability of components in Vehicle A to the overall durability of components in Vehicle B.

Please select the visualization that you think spontaneously makes the above comparison.

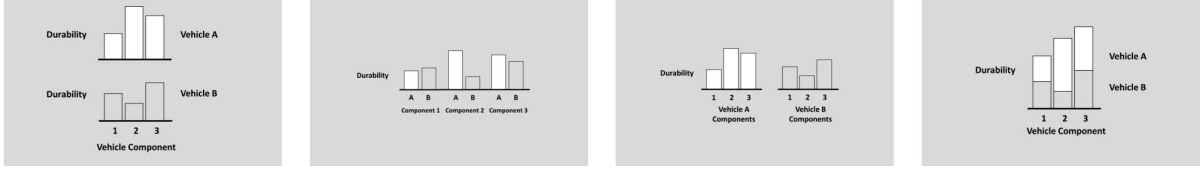


Fig. 8. Multiple choice set up for experts to indicate their preferences in Experiment 2.

data visualization design and development (e.g., engineering, graphic design, product manager). Also, 31% of our participants mentioned that they enjoyed learning about visualizations through popular media and infographics. The participants also completed a subjective graph literacy report, first introduced by [27]. They reported an average value of 4.75 out of 6 ($SD = 0.75$, 1 = not good at all, 6 = extremely good), suggesting that most participants were comfortable interpreting visualizations.

6.2 Methods and Procedures

Experts were given a comparison goal and asked to select the visualization they thought best makes that comparison from four arrangements via a multiple-choice task, as shown in Figure 8. Although we identified 12 comparison categories in Experiment 1 (see Figure 4), we shortened the list to focus on the five comparisons summarized in Section 5.9 and Table 2 to be considerate of the experts time.

Every expert-viewed five sets of data presented with five different scenarios (e.g., Figure 3), and each scenario came with one of the five listed comparison goals from Table 2. The order in which the five scenarios were presented and the mapping between datasets and scenarios follow a 5 x 5 Graeco Latin Square design, such that the order the scenarios were presented in, as well as how the datasets mapped onto the scenarios were counterbalanced.

6.3 Results

We conducted a Chi-Square analysis to investigate whether experts associate certain comparison types with certain visual arrangements. We found a significant relationship between experts' preferred visual arrangement for each comparison goal ($\chi^2 = 163.67$, $p < 0.001$). We summarize the visual arrangements the experts identified to facilitate each comparison type in Table 2 and compared their intuitions to our empirical results from Experiment 1.

Post-hoc comparisons with Bonferroni's correction suggest that experts preferred the stacked arrangement ($p < 0.001$) to make comparison type 1, in contrast to crowdworkers, who collectively suggested that all four arrangements were equally likely to elicit this comparison type. For comparison type 3, experts preferred the stacked arrangement ($p < 0.001$), while overlaid arrangement worked the most effectively with crowdworkers. For comparison type 4, experts chose the overlaid arrangement as the most effective one ($p = 0.022$), which is consistent with crowdsourced results from Experiment 1. For comparison type 9, experts preferred the overlaid arrangement ($p < 0.001$), but crowd-sourced results suggest that the vertical arrangement most affords this comparison type. For comparison type 11, crowdsourced data from Experiment 1 suggests that the four arrangements were equally likely to elicit this comparison and experts agreed ($p > 0.1$).

7 DESIGN GUIDELINES

We found that visual arrangements can afford different visual comparisons in bar charts. While experts generally have good intuitions about visualization design, there are several instances where their choices did not align with viewer intuitions. To help people design more effective visualizations, visualization researchers should continue to empirically explore the affordances of visualization design rather than solely relying on expert intuitions. We identify comparison affordances of the visual arrangements discovered in the present experiment below in Table 3. These findings provide guidelines for a variety of visual

analysis tools and applications.

Handling comparison intent in VisRec systems and NL interfaces: Insights from the study can be incorporated as rules for providing targeted visualization responses based on the type of comparison that the user may find useful or helps answer their question. For example, an NL utterance, "Are paper products doing better in the West region or East region?" is a common type of analytical inquiry. Showing a vertical arrangement of bar charts for instance, could help facilitate an effective takeaway that satisfies the user's intent.

Smarter defaults in authoring tools: To improve the efficacy of chart-caption pairs for visual comparisons, authors could (1) design the chart with a visual arrangement that supports the comparison goal and (2) provide a caption that emphasizes the type of comparison that the arrangement affords, beyond the current practice of just describing the variables depicted in the chart. Visual analysis tools can suggest reasonable defaults and design choices to guide the author in creating such effective chart-caption comparison pairs to doubly emphasize the comparative features in the takeaways [47].

User Goal	Category	Recomm.
compare one data value to multiple other values	1:M Type 12	adjacent
compare one element to other elements considering both groups	2:2/All Type 2/3	overlaid
identify maximum and minimum points	1:M Type 4	overlaid
compare one element in one group to another element in the same group	1:1 Type 9	vertical

Table 3. Recommended visual arrangements for various user goals.

8 LIMITATION AND FUTURE DIRECTIONS

As an initial investigation of how visual arrangements affect viewer comparisons and takeaways, we identify several limitations in our study that provide promising future research directions. First, our investigation was limited to bar charts that show groups of discrete variables. Although we experimented with different datasets and scenarios, the underlying data is relatively simple. Future research can experiment with more complex bar charts and other visualization chart types such as line charts and scatterplots. This can also lead to investigations of generalizability between visual arrangements and comparison affordances across different charts and data complexity.

Second, in our approach, to address the ambiguous nature of human language, we solicited accompanying drawings that provide more detail to the sentence takeaways. As a result, we could not automate the data analysis process, and the authors had to manually read, segment, and categorize each sentence takeaway. Future research could further explore the computational linguistics aspect of this work by developing natural language processing tools that could map chart takeaways to viewer visual comparisons. This mapping can also benefit the design of the natural language interface in visualization recommendation tools by allowing a wider range of user input queries.

Third, although we found that certain visualization arrangements better afford certain visual comparisons, it is unclear whether these arrangements would also increase the accuracy of value comparisons. There might be a mismatch between what people intuitively compare

in an arrangement and how accurately they can make that comparison. Future research could investigate the effectiveness of different visual arrangements from a perception angle using psychophysical methods.

Finally, in the current investigation, we only looked at how visual arrangements could affect how people compare elements and groups. There may be other factors that could strengthen a chart's comparison affordances. For example, highlighting aspects of a visualization has been shown to help elicit takeaways [6,36]. Future work could further explore techniques to help designers choose the best arrangement that ensures that a viewer sees the 'right' story in a dataset.

REFERENCES

- [1] IBM Watson Analytics. <http://www.ibm.com/analytics/watson-analytics/>.
- [2] Microsoft Q & A. <https://powerbi.microsoft.com/en-us/documentation/powerbi-service-q-and-a/>, 2020.
- [3] Tableau's Ask Data. <https://www.tableau.com/products/new-features/ask-data>, 2020.
- [4] ThoughtSpot. <http://www.thoughtspot.com/>, 2020.
- [5] J. Ahrens, K. Heitmann, M. Petersen, J. Woodring, S. Williams, P. Fasel, C. Ahrens, C. Hsu, and B. Geveci. Verifying scientific simulations via comparative and quantitative visualization. *IEEE Computer Graphics and Applications*, 30(6):16–28, 2010.
- [6] K. Ajani, E. Lee, C. Xiong, C. N. Knaflitz, W. Kemper, and S. Franconeri. Declutter and focus: Empirically evaluating design guidelines for effective data communication. *IEEE Transactions on Visualization & Computer Graphics*, (01):1–1, 2021.
- [7] B. Alper, B. Bach, N. Henry Riche, T. Isenberg, and J.-D. Fekete. Weighted graph comparison techniques for brain connectivity analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 483â492, New York, NY, USA, 2013. Association for Computing Machinery.
- [8] A. Anand and J. Talbot. Automatic Selection of Partitioning Variables for Small Multiple Displays. 2626(c), 2015.
- [9] J. Bacsikai-Atkari. *The syntax of comparative constructions : operators, ellipsis phenomena and functional left peripheries*. 09 2014.
- [10] O. Bakhshandeh and J. Allen. Semantic framework for comparison structures in natural language. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 993–1002, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics.
- [11] M. Bierwisch. The semantics of gradation. 1989.
- [12] D. Burlinson, K. Subramanian, and P. Goolkasian. Open vs. closed shapes: New perceptual categories? *IEEE transactions on visualization and computer graphics*, 24(1):574–583, 2017.
- [13] C. Chen, R. Zhang, E. Koh, S. Kim, S. Cohen, T. Yu, R. Rossi, and R. Bunescu. Figure captioning with reasoning and sequence-level training, 06 2019.
- [14] H. H. Clark and W. G. Chase. On the process of comparing sentences against pictures. *Cognitive psychology*, 3(3):472–517, 1972.
- [15] M. CRESSWELL. The semantics of degree. In B. H. PARTEE, editor, *Montague Grammar*, pages 261–292. Academic Press, 1976.
- [16] Z. Cui, S. K. Badam, A. Yalçin, and N. Elmqvist. Datasite: Proactive visual data exploration with computation of insight-based recommendations. *CoRR*, abs/1802.08621, 2018.
- [17] T. N. Dang and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *2014 IEEE Pacific Visualization Symposium*, pages 73–80, 2014.
- [18] Ç. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati. Foresight: Rapid data exploration through guideposts. *arXiv preprint arXiv:1709.10513*, 2017.
- [19] K. Dhamdhere, K. S. McCurley, R. Nahmias, M. Sundararajan, and Q. Yan. Analyza: Exploring data with conversation. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, IUI 2017, pages 493–504, 2017.
- [20] J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological review*, 96(3):433, 1989.
- [21] M. Fasciano and G. Lapalme. Postgraphe: a system for the generation of statistical graphics and text. In *Proceedings of the 8th International Workshop on Natural Language Generation (INLG-96)*, pages 51–60, Herstmonceux, Sussex, UK, jun 1996.
- [22] S. L. Franconeri. The nature and status of visual resources. 2013.
- [23] S. L. Franconeri. Three perceptual tools for seeing and understanding visualized data. *Current Directions in Psychological Science*, in press.
- [24] S. L. Franconeri, J. M. Scimeca, J. C. Roth, S. A. Helseth, and L. E. Kahn. Flexible visual processing of spatial relationships. *Cognition*, 122(2):210–227, 2012.
- [25] C. Friedman. A general computational treatment of the comparative. In *27th Annual Meeting of the Association for Computational Linguistics*, pages 161–168, Vancouver, British Columbia, Canada, June 1989. Association for Computational Linguistics.
- [26] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visu-

- alization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software Technology*, UIST 2015, pages 489–500, New York, NY, USA, 2015. ACM.
- [27] R. Garcia-Retamero and E. T. Cokely. Communicating health risks with visual aids. *Current Directions in Psychological Science*, 22(5):392–399, 2013.
 - [28] R. Garcia-Retamero and M. Galesic. Communicating treatment risk reduction to people with low numeracy skills: a cross-cultural comparison. *American journal of public health*, 99(12):2196–2202, 2009.
 - [29] B. Gleich, O. Creighton, and L. Kof. Ambiguity detection: Towards a tool explaining ambiguity sources. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, pages 218–232. Springer, 2010.
 - [30] 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.
 - [31] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE transactions on visualization and computer graphics*, 19(12):2316–2325, 2013.
 - [32] L. R. Gleitman, D. January, R. Nappa, and J. C. Trueswell. On the give and take between event apprehension and utterance formulation. *Journal of memory and language*, 57(4):544–569, 2007.
 - [33] M. Graham and J. Kennedy. Exploring multiple trees through dag representations. *IEEE transactions on visualization and computer graphics*, 13:1294–301, 11 2007.
 - [34] C. Hamann, I. Heim, D. Lewis, P. Seuren, and W. Sternefeld. Comparing semantic theories of comparison. *arnim von stechow*. 2005.
 - [35] S. T. Hawley, B. Zikmund-Fisher, P. Ubel, A. Jancovic, T. Lucas, and A. Fagerlin. The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient education and counseling*, 73(3):448–455, 2008.
 - [36] M. Hearst, M. Tory, and V. Setlur. Toward interface defaults for vague modifiers in natural language interfaces for visual analysis. In *2019 IEEE Visualization Conference (VIS)*, pages 21–25. IEEE, 2019.
 - [37] E. Hoque, V. Setlur, M. Tory, and I. Dykeman. Applying pragmatics principles for interaction with visual analytics. *IEEE Trans. Vis. Comput. Graph.*, 24(1):309–318, 2018.
 - [38] K. Hu, D. Orghian, and C. Hidalgo. Dive: A mixed-initiative system supporting integrated data exploration workflows. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, page 5. ACM, 2018.
 - [39] L. Huang. Space of preattentive shape features. *Journal of Vision*, 20(4):10–10, 2020.
 - [40] N. Jardine, B. D. Ondov, N. Elmqvist, and S. Franconeri. The perceptual proxies of visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1012–1021, 2020.
 - [41] J. Kaplan, D. G. Fisher, and N. T. Rogness. Lexical ambiguity in statistics: how students use and define the words: association, average, confidence, random and spread. *Journal of Statistics Education*, 18(2), 2010.
 - [42] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proc. of the 2016 CHI*, pages 5092–5103, 2016.
 - [43] C. Kennedy. Projecting the adjective: The syntax and semantics of gradability and comparison. 01 1997.
 - [44] C. Kennedy. Comparatives, semantics of. *Encyclopedia of Language Linguistics*, 08 2004.
 - [45] W. Kessler and J. Kuhn. A corpus of comparisons in product reviews. 05 2014.
 - [46] A. Key, B. Howe, D. Perry, and C. Aragon. VizDeck. *Proceedings of the 2012 international conference on Management of Data - SIGMOD '12*, page 681, 2012.
 - [47] D. H. Kim, V. Setlur, and M. Agrawala. Towards understanding how readers integrate charts and captions: A case study with line charts. *arXiv preprint arXiv:2101.08235*, 2021.
 - [48] Y. Kim, K. Wongsuphasawat, J. Hullman, and J. Heer. GraphScape: A Model for Automated Reasoning about Visualization Similarity and Sequencing. *Proc. of ACM CHI 2017*, 2017.
 - [49] E. Klein. A semantics for positive and comparative adjectives. *Linguistics and Philosophy*, 4(1):1–45, 1980.
 - [50] A. Larsen and C. Bundesen. Size scaling in visual pattern recognition. *Journal of experimental psychology. Human perception and performance*, 4 1:1–20, 1978.
 - [51] A. Larsen and C. Bundesen. Effects of spatial separation in visual pattern matching: Evidence on the role of mental translation. *Journal of experimental psychology. Human perception and performance*, 24:719–31, 07 1998.
 - [52] P. Law, L. Y.-H. Lo, A. Endert, J. Stasko, and H. Qu. Causal perception in question-answering systems. *ArXiv*, abs/2012.14477, 2020.
 - [53] P.-M. Law, R. C. Basole, and Y. Wu. Duet: Helping data analysis novices conduct pairwise comparisons by minimal specification. *IEEE transactions on visualization and computer graphics*, 25(1):427–437, 2019.
 - [54] P.-M. Law, A. Endert, and J. Stasko. Characterizing automated data insights, 2020.
 - [55] D. J.-L. Lee, H. Dev, H. Hu, H. Elmeleegy, and A. Parameswaran. Avoiding drill-down fallacies with vispilot: Assisted exploration of data subsets. In *Proc. of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 186–196. ACM, 2019.
 - [56] D. J.-L. Lee, J. Kim, R. Wang, and A. Parameswaran. Scattersearch: Visual querying of scatterplot visualizations, 2019.
 - [57] D. J.-L. Lee, J. Lee, T. Siddiqui, J. Kim, K. Karahalios, and A. Parameswaran. You can't always sketch what you want: Understanding sensemaking in visual query systems. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2019.
 - [58] D. J.-L. Lee, V. Setlur, M. Tory, K. Karahalios, and A. Parameswaran. Deconstructing categorization in visualization recommendation: A taxonomy and comparative study, 2021.
 - [59] S. Lee, S.-H. Kim, and B. C. Kwon. Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics*, 23(1):551–560, 2016.
 - [60] H. Lin, D. Moritz, and J. Heer. Dziban : Balancing Agency & Automation in Visualization Design via Anchored Recommendations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems - CHI '20*, 2020.
 - [61] T. Liu, X. Li, C. Bao, M. Correll, C. Tu, O. Deussen, and Y. Wang. Data-driven mark orientation for trend estimation in scatterplots. 2021.
 - [62] M. A. Livingston and J. W. Decker. Evaluation of trend localization with multi-variate visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2053–2062, 2011.
 - [63] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, 1986.
 - [64] J. D. Mackinlay, P. Hanrahan, and C. Stolte. Show Me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007.
 - [65] R. Metoyer, B. Lee, N. Henry Riche, and M. Czerwinski. Understanding the verbal language and structure of end-user descriptions of data visualizations. In *Proc. of the SIGCHI*, CHI '12, page 1659–1662, New York, NY, USA, 2012. Association for Computing Machinery.
 - [66] A. L. Michal and S. L. Franconeri. Visual routines are associated with specific graph interpretations. *Cognitive Research: Principles and Implications*, 2(1):1–10, 2017.
 - [67] A. L. Michal, D. Uttal, P. Shah, and S. L. Franconeri. Visual routines for extracting magnitude relations. *Psychonomic bulletin & review*, 23(6):1802–1809, 2016.
 - [68] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics*, 25(1):438–448, 2019.
 - [69] C. Nothelfer and S. Franconeri. Measures of the benefit of direct encoding of data deltas for data pair relation perception. *IEEE transactions on visualization and computer graphics*, 26(1):311–320, 2019.
 - [70] H.-G. Pagendarm and F. Post. Comparative visualization - approaches and examples. 01 1995.
 - [71] S. Palan and C. Schitter. Prolific. acâa subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.
 - [72] Z. Qu and J. Hullman. Keeping multiple views consistent: Constraints, validations, and exceptions in visualization authoring. *IEEE transactions on visualization and computer graphics*, 24(1):468–477, 2017.
 - [73] I. Qualtrics. Qualtrics. *Provo, UT, USA*, 2013.
 - [74] R. A. Rensink. Change detection. *Annual Review of Psychology*, 53(1):245–277, 2002. PMID: 11752486.
 - [75] J. C. Roberts. State of the art: Coordinated multiple views in exploratory visualization. In *CMV 2007*, pages 61–71, 2007.
 - [76] J. Roth and S. Franconeri. Asymmetric coding of categorical spatial relations in both language and vision. *Frontiers in psychology*, 3:464,

2012.

- [77] E. Sapir. Grading, a study in semantics. *Philosophy of Science*, 11:93 – 116, 1944.
- [78] A. Sarvghad, M. Tory, and N. Mahyar. Visualizing Dimension Coverage to Support Exploratory Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):21–30, 2017.
- [79] R. Schwarzschild and K. Wilkinson. Quantifiers in comparatives: A semantics of degree based on intervals. *Natural Language Semantics*, 10:1–41, 2002.
- [80] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information visualization*, 4(2):96–113, 2005.
- [81] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. 2016.
- [82] V. Setlur, M. Tory, and A. Djalali. Inferencing underspecified natural language utterances in visual analysis. 2019.
- [83] P. Shah and E. G. Freedman. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in cognitive science*, 3(3):560–578, 2011.
- [84] T. Siddiqui, A. Kim, J. Lee, K. Karahalios, and A. Parameswaran. Effortless data exploration with zenvisage: an expressive and interactive visual analytics system. *Proceedings of the VLDB Endowment*, 10(4):457–468, 2016.
- [85] A. Srinivasan, M. Brehmer, B. Lee, and S. M. Drucker. What’s the difference? evaluating variations of multi-series bar charts for visual comparison tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.
- [86] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):672 – 681, 2019.
- [87] A. Srinivasan and J. Stasko. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE transactions on visualization and computer graphics*, 24(1):511–521, 2018.
- [88] A. R. Tait, T. Voepel-Lewis, B. J. Zikmund-Fisher, and A. Fagerlin. Presenting research risks and benefits to parents: does format matter? *Anesthesia and analgesia*, 111(3):718, 2010.
- [89] E. R. Tufte. *Envisioning information*. Graphics Press, Cheshire, Conn., 1990.
- [90] B. Tversky. Visualizing thought. In *Handbook of human centric visualization*, pages 3–40. Springer, 2014.
- [91] S. van den Elzen and J. J. van Wijk. Small multiples, large singles: A new approach for visual data exploration. In *Computer Graphics Forum*, volume 32, pages 191–200. Wiley Online Library, 2013.
- [92] M. Vartak, S. Madden, and A. N. Parmeswaran. SEEDB : Supporting Visual Analytics with Data-Driven Recommendations. 2015.
- [93] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 157–164. IEEE, 2005.
- [94] G. Wills and L. Wilkinson. Autovis: automatic visualization. *Information Visualization*, 9(1):47–69, 2010.
- [95] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Towards a general-purpose query language for visualization recommendation. In *Proc. of the Workshop on Human-In-the-Loop Data Analytics*, page 4. ACM, 2016.
- [96] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, 2016.
- [97] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager 2 : Augmenting Visual Analysis with Partial View Specifications. 2017.
- [98] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri. Illusion of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):853–862, 2019.
- [99] C. Xiong, L. van Weelden, and S. Franconeri. The curse of knowledge in visual data communication. *IEEE transactions on visualization and computer graphics*, 2019.
- [100] M. A. Yalçın, N. Elmqvist, and B. B. Bederson. Keshif: Rapid and expressive tabular data exploration for novices. *IEEE transactions on visualization and computer graphics*, 24(8):2339–2352, 2018.
- [101] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory & Cognition*, 27(6):1073–1079, 1999.