Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content

Alan Lundgard and Arvind Satyanarayan

1 INTRODUCTION

The proliferation of visualizations during the COVID-19 pandemic has underscored their double-edged potential: efficiently communicating critical public health information — as with the immediately-canonical “Flatten the Curve” chart (Fig. 1) — while simultaneously excluding people with disabilities [11,28]. To promote accessible visualization via natural language descriptions (B, C), we introduce a four-level model of semantic content. Our model categorizes and color codes sentences according to the semantic content they convey.

Abstract — Natural language descriptions sometimes accompany visualizations to better communicate and contextualize their insights, and to improve their accessibility for readers with disabilities. However, it is difficult to evaluate the usefulness of these descriptions, and how effectively they improve access to meaningful information, because we have little understanding of the semantic content they convey, and how different readers receive this content. In response, we introduce a conceptual model for the semantic content conveyed by natural language descriptions of visualizations. Developed through a grounded theory analysis of 2,147 sentences, our model spans four levels of semantic content: enumerating visualization construction properties (e.g., marks and encodings); reporting statistical concepts and relations (e.g., extrema and correlations); identifying perceptual and cognitive phenomena (e.g., complex trends and patterns); and elucidating domain-specific insights (e.g., social and political context). To demonstrate how our model can be applied to evaluate the effectiveness of visualization descriptions, we conduct a mixed-methods evaluation with 30 blind and 90 sighted readers, and find that these reader groups differ significantly on which semantic content they rank as most useful. Together, our model and findings suggest that access to meaningful information is strongly reader-specific, and that research in automatic visualization captioning should orient toward descriptions that more richly communicate overall trends and statistics, sensitive to reader preferences. Our work further opens a space of research on natural language as a data interface coequal with visualization.

Index Terms — Visualization, natural language description, caption, semantic, model, theory, alt text, blind, disability, accessibility.

Fig. 1. Visualizations like “Flatten the Curve” (A) efficiently communicate critical public health information, while simultaneously excluding people with disabilities [11,28]. To promote accessible visualization via natural language descriptions (B, C), we introduce a four-level model of semantic content. Our model categorizes and color codes sentences according to the semantic content they convey.

A multi-line chart entitled “Lower and Delay the Peak” (B) that plots the Number of Cases by the Time Since First Case. The purpose of the chart is to communicate that there are multiple ways the current crisis can play out. Without control measures, the spread of the disease increases exponentially, making it harder to slow down and create a big load on the healthcare system. The number of cases is dramatically higher if we do not control the spread, which means we need to build more healthcare capacity to handle all of the cases over a longer period of time, and this will save lives.

The purpose of the chart is not to provide exact numbers, but to communicate to the public that there are multiple ways the current crisis can play out. Without control measures, the spread of the disease increases exponentially, making it harder to slow down and create a big load on the healthcare system. The number of cases is dramatically higher if we do not control the spread, which means we need to build more healthcare capacity to handle all of the cases over a longer period of time, and this will save lives.

© 2021 IEEE. This is the author’s version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: 10.1109/TVCG.2021.3114770. An accessible HTML version of the article is available at: http://vis.csail.mit.edu/pubs/vis-text-model/.

Published 30 September 2021; date of current version 8 October 2021. For reprints@ieee.org. Digital Object Identifier: 10.1109/TVCG.2021.3114770.

Manuscript received 31 March 2021; accepted 15 July 2021. Date of Publication 30 September 2021; date of current version 8 October 2021. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: 10.1109/TVCG.2021.3114770.
resultant 3,600 ranked descriptions, we find significant differences in the content favored by these reader groups: while both groups generally prefer mid-level semantic content, they sharply diverge in their rankings of both the lowest and highest levels of our model.

These findings, contextualized by readers’ open-ended feedback, suggest that access to meaningful information is strongly reader-specific, and that captions for blind readers should aim to convey a chart’s trends and statistics, rather than solely detailing its low-level design elements or high-level insights. Our model of semantic content is not only descriptive (categorizing what is conveyed by visualizations) and evaluative (helping us to study what should be conveyed to whom) but also generative [78], pointing toward novel multimodal and accessible data representations (§ 6.1). Our work further opens a space of research on natural language as a data interface coequal with the language of graphics [12], calling back to the original linguistic and semiotic motivations at the heart of visualization theory and design (§ 6.2).

2 Related Work

Multiple visualization-adjacent literatures have studied methods for describing charts and graphics through natural language — including accessible media research, Human-Computer Interaction (HCI), Computer Vision (CV), and Natural Language Processing (NLP). But, these various efforts have been largely siloed from one another, adopting divergent methods and terminologies (e.g., the terms “caption” and “description” are used inconsistently). Here, we survey the diverse terrain of literatures intersecting visualization and natural language.

2.1 Automatic Methods for Visualization Captioning

Automatic methods for generating visualization captions broadly fall into two categories: those using CV and NLP methods when the chart is a rasterized image (e.g., JPEG or PNGs); and those using structured specifications of the chart’s construction (e.g., grammars of graphics).

2.1.1 Computer Vision and Natural Language Processing

Analogous to the long-standing CV and NLP problem of automatically captioning photographic images [48, 58, 64], recent work on visualization captioning has aimed to automatically generate accurate and descriptive natural language sentences for charts [6, 22, 24, 59, 78, 84]. Following the encoder-decoder framework of statistical machine translation [9, 107], these approaches usually take rasterized images of visualizations as input to a CV model (the encoder), which learns the visually salient features for outputting a relevant caption via a language model (the decoder). Training data consists of (chart, caption) pairs, collected via web-scrapping and crowdsourcing [83], or created synthetically from pre-defined sentence templates [47]. While these approaches are well-intentioned, in aiming to address the engineering problem of how to automatically generate natural language captions for charts, they have largely sidestepped the complementary (and prior) question: which semantic content should be generated to begin with? Some captions may be more or less descriptive than others, and different readers may receive different semantic content as more or less useful, depending on their levels of data literacy, domain-expertise, and/or visual perceptual ability [59, 71, 72]. To help orient work on automatic visualization captioning, our four-level model of semantic content offers a means of asking and answering these more human-centric questions.

2.1.2 Structured Visualization Specifications

In contrast to rasterized images of visualizations, chart templates [96], component-based architectures [38], and grammars of graphics [87] provide not only a structured representation of the visualization’s construction, but typically render the visualization in a structured manner as well. For instance, most of these approaches either render the output visualization as Scalable Vector Graphics (SVG) or provide a scene-graph API. Unfortunately, these output representations lose many of the semantics of the structured input (e.g., which elements correspond to axes and legends, or how nesting corresponds to visual perception). As a result, most present-day visualizations are inaccessible to people who navigate the web using screen readers. For example, using Apple’s VoiceOver to read D3 charts rendered as SVG usually outputs an inscrutable mess of screen coordinates and shape rendering properties. Visualization toolkits can ameliorate this by leveraging their structured input to automatically add Accessible Rich Internet Application (ARIA) attributes to appropriate output elements, in compliance with the World Wide Web Consortium (W3C)’s Web Accessibility Initiative (WAI) guidelines [99]. Moreover, this structured input representation can also simplify automatically generating natural language captions through template-based mechanisms, as we discuss in § 6.1.

2.2 Accessible Media and Human-Computer Interaction

While automatic methods researchers often note accessibility as a worthy motivation [27, 30, 41, 78, 83, 84], evidently few have collaborated directly with disabled people [25, 71] or consulted existing accessibility guidelines [67]. Doing so is more common to HCI and accessible media literatures [73, 91], which broadly separate into two categories corresponding to the relative expertise of the description authors: those authored by experts (e.g., publishers of accessible media) and those authored by non-experts (e.g., via crowdsourcing or online platforms).

2.2.1 Descriptions Authored by Experts

Publishers have developed guidelines for describing graphics appearing in science, technology, engineering, and math (STEM) materials [19, 29]. Developed by and for authors with some expert accessibility knowledge, these guidelines provide best practices for conveying visualized content in traditional media (e.g., printed textbooks, audio books, and tactile graphics). But, many of their prescriptions — particularly those relating to the content conveyed by a chart, rather than the modality through which the chart is rendered — are also relevant to web-based visualizations. Additionally, web accessibility guidelines from W3C provide best-practices for writing descriptions of “complex images” (including canonical chart types), either in a short description alt text attribute, or as a long textual description displayed alongside the visual image [79]. While some of these guidelines have been adopted by visualization practitioners [19, 29, 32, 55, 83, 101, 102], we here bring special attention to the empirically-grounded and well-documented guidelines created by the WGBH National Center for Accessible Media [9] and by the Benetech Diagram Center [9].

2.2.2 Descriptions Authored by Non-Experts

Frequently employed in HCI and visualization research, crowdsourcing is a technique whereby remote non-experts complete tasks currently infeasible for automatic methods, with applications to online accessibility [13], as well as remote description services like Be My Eyes. For example, Morash et al. explored the efficacy of two types of non-expert tasks for authoring descriptions of visualizations: non-experts authoring free-form descriptions without expert guidance, versus those filling-in sentence templates pre-authored by experts [72]. While these approaches can yield more richly detailed and “natural”-sounding descriptions (as we discuss in § 5), and also provide training data for auto-generated captions and annotations [56, 83], it is important to be attentive to potential biases in human-authored descriptions [10].

2.3 Natural Language Hierarchies and Interfaces

Apart from the above methods for generating descriptions, prior work has adopted linguistics-inspired framings to elucidate how natural language is used to describe — as well as interact with — visualizations.

2.3.1 Using Natural Language to Describe Visualizations

Demir et al. have proposed a hierarchy of six syntactic complexity levels corresponding to a set of propositions that might be conveyed by bar charts [27]. Our model differs in that it orders semantic content — i.e., what meaning the natural language sentence conveys — rather than how it does so syntactically. Thus, our model is agnostic to a sentence’s length, whether it contains multiple clauses or conjunctions, which has also been a focus of prior work in automatic captioning [83]. Moreover, whereas Demir et al. speculatively “envision” their set of propositions to construct their hierarchy, we arrive to our model empirically through a multi-stage grounded theory process (§ 5). Perhaps
closest to our contribution are a pair of papers by Kosslyn [57] and Livingston & Brock [6]. Kosslyn draws on canonical linguistic theory, to introduce three levels for analyzing charts: the syntactic relationship between a visualization elements; the semantic meaning of these elements in what they depict or convey; and the pragmatic aspects of what these elements convey in the broader context of their reading [57].

We began gathering preliminary data by searching for descriptions of our datasets [58]. Our approach suggests opportunities for future work: might our websites of curated visualizations in journalistic publications (including the New York Times and the Financial Times), be more accessible data representations [51], while helping to clarify the semantic content that might include. The described visualizations conveyed a variety of chart types (e.g., bar charts, line charts, scatter plots) as well as dataset domains (e.g., public health, climate change, and gender equality). To complement the student-authored descriptions, from this same set of visualizations, we curated a set of 20 and wrote our (the authors’) own descriptions, attempting to be as richly descriptive as possible. Throughout, we adhered to a linguistics-inspired framing by attending to the semantic and pragmatic aspects of our writing: which content could be conveyed through the graphical sign-system alone, and which required drawing upon our individual background knowledge, experiences, and contexts.

Analyzing these preliminary data, we proceeded to the next stage in the grounded theory method: forming axial codes (i.e., open codes organized into broader abstractions, with more generalized meaning [74]) corresponding to different content. We began to distinguish between content about a visualization’s construction (e.g., its title, encodings, legends), content about trends appearing in the visualized data (e.g., correlations, clusters, extrema), and content relevant to the visualized data but not represented in the visualization itself (e.g., explanations based on current events and domain-specific knowledge). From these axial codes, different categories (i.e., groupings delineated by shared characteristics of the content) began to emerge [74], corresponding to a chart’s encoded elements, latent statistical relations, perceptual trends, and context. We refined these content categories iteratively by first writing down descriptions of new visualizations (again, as richly as possible), and then attempting to categorize each sentence appearing in that description. If we encountered a sentence that didn’t fit within any category, we either refined the specific characteristics belonging to an existing category, or we created a new category, where appropriate.

### 3.2 Gathering A Corpus

The prior inductive and empirical process resulted in a set of preliminary content categories. To test their robustness, and to further refine them, we conducted an online survey to gather a larger-scale corpus of 582 visualization descriptions comprised of 2,147 sentences.

#### 3.2.1 Survey Design

We first curated a set of 50 visualizations drawn from the MassVis dataset [15,16], Quartz’s Atlas visualization platform [81], examples from the Vega-Lite gallery [87], and the aforementioned journalistic publications. We organized these visualizations along three dimensions: the visualization type (bar charts, line charts, and scatter plots); the topic of the dataset domain (academic studies, business-related, or non-business data journalism); and their difficulty based on an assessment of their visual and conceptual complexity. We labeled visualizations as “easy” if they were basic instances of their canonical type (e.g., single-line or un-grouped bar charts), as “medium” if they were more moderate variations on canon (e.g., contained bar groupings, overlapping scatterplot clusters, visual embellishments, or simple transforms), and as “hard” if they strongly diverged from canon (e.g., contained chartjunk or complex transforms such as log scales). To ensure robustness, two authors labeled the visualizations independently, and then resolved any disagreement through discussion. Table 1 summarizes the breakdown of the 50 visualizations across these three dimensions.

In the survey interface, participants were shown a single, randomly-selected visualization at a time, and prompted to describe it in complete English sentences. In our preliminary data collection (§3.1), we found that without explicit prompting participants were likely to provide only brief and minimally informative descriptions (e.g., sometimes simply repeating the chart title and axis labels). Thus, to mitigate against this outcome, and to elicit richer semantic content, we explicitly instructed participants to author descriptions that did not only refer to the chart’s

| CHART TYPE | TOPIC | DIFFICULTY |
|------------|-------|------------|
| bar        | academic | easy      | 21 |
| line       | business | medium    | 20 |
| scatter    | journalism | hard     | 9 |
In summary, the entire grounded theory process resulted in four distinct semantic content categories, which we organize into levels in the next section. A visual “fingerprint” shows how semantic content is distributed across sentences in the corpus (Fig. 2). Level 1 (consisting of a chart’s basic elements and encodings) represents 9.1% of the sentences in the corpus. This is expected, since Level 1 sentences were pre-generated and provided as a prompt to our survey participants, as we previously discussed. The distribution of sentences across the remaining levels is as follows: Level 2 (35.1%), Level 3 (42.9%), and Level 4 (12.9%). The fairly-balanced distribution suggests that our survey prompting successfully captured natural language sentences corresponding to a breadth of visualized content.

4 A Four-Level Model of Semantic Content

Our grounded theory process yielded a four-level model of semantic content for the natural language description of visualizations. In the following subsections, we introduce the levels of the model and provide example sentences for each. Table 2 summarizes the levels, and Table 3 shows example visualizations from our corpus and corresponding descriptions, color coded according to the model’s color scale. Additionally, we offer practical computational considerations regarding the feasibility of generating sentences at each level, with reference to the present-day state-of-the-art methods described in Related Work. While we present them alongside each other for ease of explication, we emphasize that the model levels and computational considerations are theoretically decoupled: the model is indexed to the semantic content conveyed by natural language sentences, not to the computational means through which those sentences may or may not be generated.

4.1 Level 1: Elemental and Encoded Properties

At the first level, there are sentences whose semantic content refers to elemental and encoded properties of the visualization (i.e., the visual components that comprise a graphical representation’s design and construction). These include the chart type (bar chart, line graph, scatter plot, etc.), its title and legend, its encoding channels, axis labels, and the axis scales. Consider the following sentence (Table 3.A.1).

This sentence “reads off” the axis labels and scales as they appear in the bar chart, with no additional synthesizing or interpretation. Sentences such as this are placed at the lowest level in the model because they refer to content that is foundational to visualization construction—comprising the elemental properties of the “language” of graphics.

Computational Considerations. Semantic content at Level 1 is so foundational that it has long been formalized — not only theoretically, as in Bertin’s *Semiology of Graphics*, but also mathematically and programmatically, as a “grammar of graphics” that precisely defines the algorithmic rules for constructing canonical chart types. In the case of these construction grammars, Level 1 content is directly encoded in the visualization’s structured specification (i.e., mappings between data fields and visual properties). Thus, for these grammars, generating sentences at Level 1 can amount to “filling in the blank” for a pre-defined sentence template. For example, given an appropriate template, the following natural language sentence could be trivially computed using the data encoded in the visualization specification.

This is a [chart-type] entitled [chart-title]. [y-encoding] is plotted on the vertical y-axis from [y-min] to [y-max]. [x-encoding] is plotted on the horizontal x-axis from [x-min] to [x-max]."

And similarly, for other sentence templates and elemental properties encoded in the visualization’s structured specification. If the structured specification is not available, however, or if it does not follow a declarative grammar, then CV and NLP methods have also shown promise when applied to rasterized visualization images (e.g., JPEGs or PNGs). For example, recent work has shown that Level 1 semantic content can be feasibly generated provided an appropriate training dataset of pre-defined sentence templates, or by extracting a visualization’s structured specification from a rasterized visualization image.

3.2 Survey Results

We recruited 120 survey participants through the Prolific platform. In an approximately 30-minute study compensated at a rate of $10-12 per hour, we asked each participant to describe 5 visualizations (randomly selected from the set of 50), resulting in at least 10 participant-authored descriptions per visualization. For some visualizations, we collected between 10-15 responses, due to limitations of the survey logic for randomly selecting a visualization to show participants. In total, this survey resulted in 582 individual descriptions comprised of 2,147 natural language sentences. We manually cleaned each sentence to correct errors in spelling, grammar, punctuation (n.b., we did not alter the semantic content conveyed by each sentence). We then labeled each sentence according to the content categories developed through our prior grounded theory process. As before, to ensure robustness, two authors labeled each sentence independently, and then resolved any disagreement through discussion. This deliberative and iterative process helped us to further distinguish and refine our categories. For example, we were able to more precisely draw comparisons between sentences reporting computable “data facts” through rigid or templated articulation (such as “x-encoding is positively correlated with y-encoding”), with sentences conveying the same semantic content through more “natural”-sounding articulation (such as “for the most part, as x-encoding increases, so too does y-encoding”).
4.2 Level 2: Statistical Concepts and Relations

At the second level, there are sentences whose semantic content refers to abstract statistical concepts and relations that are latent in the visualization’s backing dataset. This content conveys computable descriptive statistics (such as mean, standard deviation, extrema, correlations) — what have sometimes been referred to as “data facts” because they are “objectively” present within a given dataset [92,100] (as opposed to primarily observed via visualization, which affords more opportunities for subjective interpretation). In addition to these statistics, Level 2 content includes relations between data points (such as “greater than” or “less than” comparisons). Consider the following sentences (Table 3.C.2).

For low income countries, the average life expectancy is 60 years for men and 65 years for women. For high income countries, the average life expectancy is 77 years for men and 82 years for women.

These two sentences refer to a statistical property: the computed mean of the life expectancy of a population, faceted by gender and country income-level. Consider another example (Table 3.A.2).

The highest COVID-19 mortality rate is in the 80+ age range, while the lowest mortality rate is in 10-19, 20-29, 30-39, sharing the same rate.

Although this sentence is more complex, it nevertheless resides at Level 2. It refers to the extrema of the dataset (i.e., the “highest” and “lowest” mortality rates), and makes two comparisons (i.e., a comparison between the extrema, and another between age ranges sharing the lowest mortality rate). All of the above sentences above share the same characteristic, distinguishing them from those at Level 1: they refer to relations between points in the dataset, be they descriptive statistics or point-wise comparisons. Whereas Level 1 sentences “read off” the visualization’s elemental properties, Level 2 sentences “report” statistical concepts and relations within the chart’s backing dataset.

Computational Considerations. While semantic content at Level 1 requires only reference to the visualization’s specification, content at Level 2 also requires access to the backing dataset. Here, the two categories of automatic methods begin to diverge in their computational feasibility. For visualizations with a structured specification, generating sentences at Level 2 is effectively as easy as generating sentences at Level 1: it requires little more computation to calculate and report descriptive statistics when the software has access to the backing dataset (i.e., encoded as part of the visualization specification). Indeed, many visualization software systems (such as Tableau’s Summary Card, Voder [92], Quill NLG Plug-In for Power BI, and others) automatically compute summary statistics and present them in natural language captions. By contrast, for CV and NLP methods, generating Level 2 sentences from a rasterized image is considerably more difficult — although not entirely infeasible — depending on the chart type and complexity. For example, these methods can sometimes report extrema (e.g., which age ranges exhibit the highest and lowest mortality rates in [3.A.2] [26,78]). Nevertheless, precisely reporting descriptive statistics (e.g., the computed mean of points in a scatter plot) is less tractable, without direct access to the chart’s backing dataset.

4.3 Level 3: Perceptual and Cognitive Phenomena

At the third level, there are sentences whose semantic content refers to perceptual and cognitive phenomena appearing in the visual representation of the data. When compared to, and defended against, other forms of data analysis (e.g., purely mathematical or statistical methods), visualization is often argued to confer some unique benefit to human readers. That is, visualizations do not only “report” descriptive statistics of the data (as in Level 2), they also show their readers something more: they surface unforeseen trends, convey complex multi-faceted patterns, and identify noteworthy exceptions that aren’t readily apparent through non-visual methods of analysis (cf., Anscombe’s Quartet or the Datasaurus Dozen [70]). Level 3 sentences are comprised of content that refers to these perceptual and cognitive phenomena, usually articulated in “natural”-sounding (rather than templatized) language.

Computational Considerations. At Level 3, we begin to reach and exceed the limits of present-day state-of-the-art automatic methods. While there exist “off-the-shelf” statistical packages for computing basic trends and predictions in a dataset (e.g., correlations, polynomial regressions, statistical inferences), visualizations allow us to perceive and articulate complex trends for which there may exist no line of “best fit”. While automatic methods may eventually approach (or exceed) human capabilities on well-defined tasks [78], for now Level 3 semantic content is likely generated via human (rather than machine) perception and cognition [72]. Taking inspiration from the “mind-independent” versus “mind-dependent” ontological distinction [4], we define sentences at Levels 1 and 2 as perceivably-independent (i.e., their content can be generated independently of human or machine perception, without reference to the visualization), while sentences at Level 3 are perceivably-dependent (i.e., their content requires a perceivably of some sort; likely a human, although machine perception may increasingly suffice for generating Level 3 content). Table 4 summarizes this distinction.

4.4 Level 4: Contextual and Domain-Specific Insights

Finally, at the fourth level, there are sentences whose semantic content refers to contextual and domain-specific knowledge and experience. For example, sentences (Table 3.B.4 and 3.C.4).

For low income countries, the average life expectancy is 60 years for men and 65 years for women. For high income countries, the average life expectancy is 77 years for men and 82 years for women.

These sentences convey the “overall gist” of complex trends and patterns (e.g., stock prices “seem to fluctuate but nevertheless increase”), synthesize multiple trends to identify exceptions (e.g., “years 2008-2009 are exceptions as we can see an extreme drop in prices of all given corporations.

The low income countries are more scattered than the high income countries. There is a visible gap between high and low income countries, indicated by the Income-Age Divide line.

The crisis culminated with the bankruptcy of Lehman Brothers on September 15, 2008 and an international banking crisis.

People living in low-income countries tend to have a lower life expectancy than the people living in high-income countries, likely due to many societal factors including access to healthcare, food, other resources, and overall quality of life.
Table 3. Example visualizations and descriptions from our corpus. Paragraph breaks in rows A and B indicate a description authored by a unique participant from our corpus gathering survey (§5.2.1), while row C shows an curated exemplar description from our evaluation (§5.1).

**Visualization** | **Description**
--- | ---
![COVID-19 mortality rate by age](image) | [1] This is a vertical bar chart entitled “COVID-19 mortality rate by age” that plots Mortality rate by Age. Mortality rate is plotted on the vertical y-axis from 0 to 15%. Age is plotted on the horizontal x-axis in bins: 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80+. [2] The highest COVID-19 mortality rate is in the 80+ age range, while the lowest mortality rate is in 10–19, 20–29, 30–39, sharing the same rate. [3] COVID-19 mortality rate does not linearly correspond to the demographic age. [4] The mortality rate increases with age, especially around 40–49 years and upwards. [5] This relates to people’s decrease in their immunity and the increase of co-morbidity with age.

![Big Tech Stock Prices](image) | [1] This is a multi-line chart entitled “Big Tech Stock Prices” that plots price by date. The corporations include AAPL (Apple), AMZN (Amazon), GOOG (Google), IBM (IBM), and MSFT (Microsoft). The years are plotted on the horizontal x-axis from 2000 to 2010 with an increment of 2 years. The prices are plotted on the vertical y-axis from 0 to 800 with an increment of 200. [2] GOOG has the greatest price over time. MSFT has the lowest price over time. [3] Prices of particular Big Tech corporations seem to fluctuate but nevertheless increase over time. Years 2008-2009 are exceptions as we can see an extreme drop in prices of all given corporations. [4] The big drop in prices was caused by financial crisis of 2007-2008. The crisis culminated with the bankruptcy of Lehman brothers on September 15, 2008 and an international banking crisis. [5] At the beginning of 2008, every of this stock price went down, likely due to the financial crisis. [6] Then they have risen again and dropped again, more so than previously. [7] GOOG has the highest price over the years. MSFT has the lowest price over the years. [8] GOOG quickly became the richest one of the Big Tech corporations. [9] GOOG has also experienced some kind of a crisis in 2009, because their prices drop rapidly, but then rebounded.

![Born in 2016: Life Expectancy Gap by Gender and Income](image) | [1] This is a scatter plot entitled “Born in 2016: Life Expectancy Gap by Gender and Income” that plots Women Life Expectancy at Birth (Years) by Men Life Expectancy at Birth (Years). The Women Life Expectancy at Birth is plotted on the vertical y-axis from 40 to 90 years. The Men Life Expectancy at Birth is plotted on the horizontal x-axis from 40 to 90 years. High Income Countries are plotted in dark green. Low Income Countries are plotted in light green. A 45 degree line from the origin represents Equal Life Expectancy. [2] For low income countries, the average life expectancy is 60 years for men and 65 years for women. For high income countries, the average life expectancy is 77 years for men and 82 years for women. [3] Overall, women have a slightly higher life expectancy than men. Women live around 5 to 10 years longer than men. The low income countries are more scattered than the high income countries. There is a visible gap between high and low income countries, indicated by the Income-Age Divide line. [4] People living in low-income countries tend to have a lower life expectancy than the people living in high-income countries, likely due to many societal factors, including access to healthcare, food, other resources, and overall quality of life. People who live in lower income countries are more likely to experience deprivation and poverty, which can cause related health problems.
We selected 15 visualizations for the evaluation, curated to be representative of the categories from our prior survey (§3). Specifically, we selected 5 visualizations for each of the three dimensions: type (bar, line, scatter), topic (academic, business, journalism), and difficulty (easy, medium, hard). For every visualization, participants were asked to rank the usefulness of 4 different descriptions, each corresponding to one level of semantic content, presented unlabeled and in random order. We conducted a mixed-methods evaluation in which 30 blind and 90 sighted readers first ranked the usefulness of descriptions authored at varying levels of semantic content, and then completed an open-ended questionnaire.

### 5.1 Evaluation Design

We selected 15 visualizations for the evaluation, curated to be representative of the categories from our prior survey (§3). Specifically, we selected 5 visualizations for each of the three dimensions: type (bar, line, scatter), topic (academic, business, journalism), and difficulty (easy, medium, hard). For every visualization, participants were asked to rank the usefulness of 4 different descriptions, each corresponding to one level of semantic content, presented unlabeled and in random order. We conducted a mixed-methods evaluation in which 30 blind and 90 sighted readers first ranked the usefulness of descriptions authored at varying levels of semantic content, and then completed an open-ended questionnaire.

In addition to curating a representative set of visualizations, we also curated descriptions representative of each level of semantic content. Participant-authored descriptions from our prior survey often did not contain content from all 4 levels or, if they did, this content was interleaved in a way that was not clearly-separable for the purpose of a ranking task (Fig. 2). Thus, for this evaluation, we curated and collated sentences from multiple participant-authored descriptions to create exemplar descriptions, such that each text chunk contained only content belonging to a single semantic content level. Table 3C shows one such exemplar description, whereas Table 3A and B show the original un-collated descriptions. For each ranking task, readers were presented with a brief piece of contextualizing text, such as the following.

> "Suppose that you are reading an academic paper about how life expectancy differs for people of different genders from countries with different levels of income. You encounter the following visualization. [Table 3C] Which content do you think would be most useful to include in a textual description of this visualization?"

Additionally, blind readers were presented with a brief text noting that the hypothetically-encountered visualization was inaccessible via screen reader technology. In contrast to prior work, which has evaluated chart descriptions in terms of “efficiency,” “informativeness,” and “clarity” (97, 8), we intentionally left the definition of “useful” open to the reader’s interpretation. We hypothesize that “useful” descriptions may not be necessarily efficient (i.e., they may require lengthy explanation or background context), and that both informativeness and clarity are constituents of usefulness. In short, ranking “usefulness” affords a holistic evaluation metric. Participants assigned usefulness rankings to each of the 4 descriptions by selecting corresponding radio buttons, labeled 1 (least useful) to 4 (most useful). In addition to these 4 descriptions, we included a 5th choice as an “attention check”, a sentence whose content was entirely irrelevant to the chart to ensure participants were reading each description prior to ranking them. If a participant did not rank the attention check as least useful, we filtered out their response from our final analysis. We include the evaluation interfaces and questions with the Supplemental Material.

### 5.2 Participants

Participants consisted of two reader groups: 90 sighted readers recruited through the Prolific platform, and 30 blind readers recruited through our friends in the blind community and through a call for participation sent out via Twitter (n.b., in accessibility research, it is common to compare blind and sighted readers recruited through these means [14]).

#### 5.2.1 Participant Recruitment

For sighted readers qualifications for participation included English language proficiency and no color vision deficiency, and blind readers were expected to be proficient with a screen reader, such as Job Access With Speech (JAWS), NonVisual Desktop Access (NVDA), or Apple’s VoiceOver. Sighted readers were compensated at a rate of $10-12 per hour, for an approximately 20-minute task. Blind readers were compensated at a rate of $50 per hour, for an approximately 1-hour task. This difference in task duration was for two reasons. First, participants recruited through Prolific are usually not accustomed to completing lengthy tasks — our prior surveys and pilots suggested that these participants might contribute low-quality responses on “click-through” tasks if the task duration exceeded 15–20 minutes — and thus we asked each participant to rank only 5 of the 15 visualizations at a time. Second, given the difficulty of recruiting blind readers proficient with screen readers, we asked each blind participant to rank all 15 visualizations, and compensated them at a rate commensurate with their difficult-to-find expertise [67]. In this way, we recruited sufficient numbers of readers to ensure that each of the 15 visualization ranking tasks would be completed by 30 participants from both reader groups.

#### 5.2.2 Participant Demographics

Among the 30 blind participants, 53% (n=16) reported their gender as male, 36% (n=11) as female, and 3 participants “preferred not to say.” The most common highest level of education attained was a Bachelor’s degree (60%, n=18), and most readers were between 20–40 years old (66%, n=20). The screen reader technology readers used to complete the study was evenly balanced: VoiceOver (n=10), JAWS (n=10), NVDA (n=9), and “other” (n=1). Among the 90 sighted participants, 69% reported their gender as male (n=62) and 31% as female (n=28). The most common highest level of education attained was a high school diploma (42%, n=38) followed by a Bachelor’s degree (40%, n=36), and most sighted readers were between 20–30 years old (64%, n=58).

---

*© 2021 IEEE. This is the author’s version of the article that has been published in IEEE Transactions on Visualization and Computer Graphics. The final version of this record is available at: [10.1109/TVCG.2021.3114770](http://dx.doi.org/10.1109/TVCG.2021.3114770).*

An accessible HTML version of the article is available at: [http://vis.csail.mit.edu/pubs/vis-text-model](http://vis.csail.mit.edu/pubs/vis-text-model)
Table 4. (Upper) Rankings [1=least useful, 4=most useful] of semantic content at each level of the model, for blind and sighted readers. The scale encodes the number of times a given level was assigned a given rank by a reader. Dotted contour lines delineate Regions with a threshold equal to $\mu + 1.96 \times \sigma$, each labeled with a capital letter A–F. (Lower) Shaded cells indicate significant ranking differences pairwise between levels.

| LEVELS | 1 × 2 | 1 × 3 | 1 × 4 | 2 × 3 | 2 × 4 | 3 × 4 |
|--------|-------|-------|-------|-------|-------|-------|
| BLIND  | $p < 0.001$ | $p < 0.001$ | $p < 0.312$ | $p < 0.148$ | $p < 0.001$ | $p < 0.001$ |
| SIGHTED | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p < 0.001$ | $p < 0.059$ |

On a 7-point Likert scale [1=strongly disagree, 7=strongly agree], blind readers participated having “a good understanding of data visualization concepts” ($\mu = 6.3$, $\sigma = 1.03$) as well as “a good understanding of statistical concepts and terminology” ($\mu = 5.90$, $\sigma = 1.01$). Sighted participants reported similar levels of understanding: ($\mu = 6.7$, $\sigma = 0.73$) and ($\mu = 5.67$, $\sigma = 1.06$), respectively. Sighted participants also considered themselves to be “proficient at reading data visualizations” ($\mu = 5.97$, $\sigma = 0.89$) and were able to “read and understand all of the visualizations presented in this study” ($\mu = 6.44$, $\sigma = 0.71$).

5.3 Quantitative Results

Quantitative results for the individual rankings (1,800 per blind and sighted reader groups) are summarized by the heatmaps in Table 4 (Upper), which aggregate the number of times a given content level was assigned a certain rank. Dotted lines in both blind and sighted heatmaps delineate regions exceeding a threshold — calculated by taking the mean plus half a standard deviation ($\mu + 0.5 \times \sigma$) resulting in a value of 1.39 and 136, respectively — and are labeled with a capital letter A–F. These results exhibit significant differences between reader groups. For both reader groups, using Friedman’s Test (a non-parametric multi-comparison test for rank-order data) the $p$-value is $p < 0.001$, so we reject the null hypothesis that the mean rank is the same for all four semantic content levels [37]. Additionally, in Table 4 (Lower), we find significant ranking differences when making pair-wise comparisons between levels, via Nemenyi’s test (a post-hoc test commonly coupled with Friedman’s to make pair-wise comparisons). There appears to be strong agreement among sighted readers that higher levels of semantic content are more useful: Levels 3 and 4 are found to be most useful (Region 4.F), while Levels 1 and 2 are least useful (Regions 4.D and 4.E). Blind readers agree with each other to a lesser extent, but strong trends are nevertheless apparent. In particular, blind readers rank content and Levels 2 and 3 as most useful (Region 4.C), and semantic content at Levels 1 and 4 as least useful (Regions 4.A and 4.B). When facetting these rankings by visualization type, topic, or difficulty we did not observe any significant differences, suggesting that both reader groups rank semantic content levels consistently, regardless of how the chart itself may vary. Noteworthy for both reader groups, the distribution of rankings for Level 1 is bimodal — the only level to exhibit this property. While a vast majority of both blind and sighted readers rank Level 1 content as least useful, this level is ranked “most useful” in 101 and 87 instances by blind and sighted readers, respectively. This suggests that both reader groups have a more complicated perspective toward descriptions of a chart’s elemental and encoded properties; a finding we explore further by analyzing qualitative data.

5.4 Qualitative Results

In a questionnaire, we asked readers to use a 7-point Likert scale [1=strongly disagree, 7=strongly agree] to rate their agreement with a set of statements about their experience with visualizations. We also asked them to offer open-ended feedback about which semantic content they found to be most useful and why. Here, we summarize the key trends that emerged from these two different forms of feedback, from both blind readers (BR) and sighted readers (SR).

5.4.1 Descriptions Are Important to Both Reader Groups

All blind readers reported encountering inaccessible visualizations: either multiple times a week (43%, n=13), everyday (20%, n=6), once or twice a month (20%, n=6), or at most once a week (17%, n=5). These readers reported primarily encountering these barriers on social media (30%, n=9), on newspaper websites (13%, n=4), and in educational materials (53%, n=16) — but, most often, barriers were encountered in all of the above contexts (53%, n=16). Blind readers overwhelmingly agreed with the statements “I often feel that important public information is inaccessible to me, because it is only available in a visual format” ($\mu = 6.1$, $\sigma = 1.49$), and “Providing textual descriptions of data visualizations is important to me” ($\mu = 6.83$, $\sigma = 0.38$).

By contrast, sighted readers neither agreed nor disagreed regarding the inaccessibility of information conveyed visually ($\mu = 4$, $\sigma = 1.57$). Similarly, they were split on whether they ever experienced barriers to reading visualizations, with 52% (n=47) reporting that they sometimes do (especially when engaging with a new topic) and 48% (n=43) reporting that they usually do not. Nevertheless, sighted readers expressed support for natural language descriptions of visualizations ($\mu = 5.60$, $\sigma = 1.27$). A possible explanation for this support is that — regardless of whether the visualization is difficult to read — descriptions can still facilitate comprehension. For instance, SR64 noted that “textual description requires far less brainpower and can break down a seemingly complex visualization into an easy to grasp overview.”

5.4.2 Reader Groups Disagree About Contextual Content

A majority of blind readers (63%, n=19) were emphatic that descriptions should not contain an author’s subjective interpretations, contextual information, or editorializing about the visualized data (i.e., Level 4 content). Consistent with blind readers ranking this as the least useful (Region 4.B), BR20 succinctly articulated a common sentiment: “I want the information to be simply laid out, not peppered with subjective commentary...I just prefer it to be straight facts, not suppositions or guesstimates.” BR4 also noted that an author’s “opinions” about the data “should absolutely be avoided,” and BR14 emphasized agency when interpreting data: “I want to have the time and space to interpret the numbers for myself before I read the analysis.” By contrast, many sighted readers 41% (n=37) expressed the opposite sentiment (Region 4.F) noting that, for them, the most useful descriptions often “told a story,” communicated an important conclusion, or provided deeper insights into the visualized data. As SR64 noted: “A description that simply describes the visualization and its details is hardly useful, but a description that tells a story using the data and derives a solution from it is extremely useful.” Only 4% (n=4) of sighted readers explicitly stated that a description should exclude Level 4 semantic content.

5.4.3 Some Readers Prefer Non-Statistical Content

Overall, blind readers consistently ranked both Levels 2 and 3 as the most useful (Region 4.C). But, some readers explicitly expressed preference for the latter over the former, highlighting two distinguishing characteristics of Level 3 content: that it conveys not only descriptive statistics but overall perceptible trends, and that it is articulated
in commonplace or “natural”-sounding language. For instance, BR26 remarked that a visualization description is “more useful if it contains the summary of the overall trends and distributions of the data rather than just mentioning some of the extreme values or means.” Similarly, BR21 noted that “not everyone who encounters a data visualization needs it for statistical purposes,” and further exclaimed “I want to know how a layperson sees it, not a statistician; I identify more with simpler terminology.” These preferences help to further delineate Level 3 from Levels 2 and 4. Content at Level 3 is “non-statistical” in the sense that it does only report statistical concepts and relations (as in Level 2), but neither does it do away with statistical “objectivity” entirely, so as to include subjective interpretation or speculation (as content in Level 4 might). In short, Level 3 content conveys statistically-grounded concepts in not-purely-statistical terms, a challenge that is core to visualization, and science communication more broadly.

5.4.4 Combinations of Content Levels Are Likely Most Useful

While roughly 12% readers from both blind and sighted groups indicated that a description should be as concise as possible, among blind readers, 40% (n=12) noted that the most useful descriptions would combine content from multiple levels. This finding helps to explain the bimodality in Level 1 rankings we identified in the previous section. According to BR9, Level 1 content is only useful if other information is also conveyed: “All of the descriptions provided in this survey which *only* elaborated on x/y and color-coding are almost useless.” This sentiment was echoed by BR5, who added that if Level 1 content were “combined with the [Level 2 or Level 3], that’d make for a great description.” This finding has implications for research on automatic visualization captioning: these methods should aim to generate not only the lower levels of semantic content, but to more richly communicate a chart’s overall trends and statistics, sensitive to reader preferences.

5.4.5 Some Automatic Methods Raise Ethical Concerns

Research on automatically generating visualization captions is often motivated by the goal of improving information access for people with visual disabilities [27, 28, 33, 34]. However, when deployed in real-world contexts, these methods may not confer their intended benefits, as one blind reader in our evaluation commented.

“A.I. attempting to convert these images is still in its infancy. Facebook and Apple auto-descriptions of general images are more of a timewaster than useful. As a practical matter, if I find an inaccessible chart or graph, I just move on.” (BP22)

Similarly, another participant (BR26) noted that if a description were to only describe a visualization’s encodings then “the reader wouldn’t get any insight from these texts, which not only increases the readers’ reading burden but also conveys no effective information about the data.” These sentiments reflect some of the ethical concerns surrounding the deployment of nascent CV and NLP models, which can output accurate but minimally informative content—or worse, can output erroneous content to a trusting audience [69, 78]. Facebook’s automatic image descriptions, for example, have been characterized by technology educator Chancey Fleet as “famously useless in the Blind community” while “garner[ing] a ton of glowing reviews from mainstream outlets without much of use to disabled people” [34, 40]. Such concerns might be mitigated by developing and evaluating automatic methods with disabled readers, through participatory design processes [67].

6 Discussion and Future Work

Our four-level model of semantic content—and its application to evaluating the usefulness of descriptions—has practical implications for the design of accessible data representations, and theoretical implications for the relationship between visualization and natural language.

6.1 Natural Language As An Interface Into Visualization

Divergent reader preferences for semantic content suggests that it is helpful to think of natural language—not only as an interface for constructing and exploring visualizations [36, 89, 93]—but also as an interface into visualization, for understanding the semantic content they convey. Under this framing, we can apply Beaudouin-Lafon’s framework for evaluating interface models in terms of their descriptive, evaluative, and generative powers [78], to bring further clarity to the practical design implications of our model. First, our grounded theory process yielded a model with descriptive power: it categorizes the semantic content conveyed by visualizations. Second, our study with blind and sighted readers demonstrated our model’s evaluative power: it offered a means of comparing different levels of semantic content, thus revealing divergent preferences between these different reader groups. Third, future work can now begin to study our model’s generative power: its implications for novel multimodal interfaces and accessible data representations. For instance, our evaluation suggested that descriptions primarily intending to benefit sighted readers might aim to generate higher-level semantic content (§ 5.4.2), while those intending to benefit blind readers might instead focus on affordances the option to customize and combine different content levels (§ 5.4.4), depending on their individual preferences (§ 5.4.5). This latter path might involve automatically ARIA tagging web-based charts to surface semantic content at Levels 1 & 2, with human-authors conveying Level 3 content. Or, it might involve applying our results to develop and evaluate outputs of automatic captioning systems—to probe their technological capabilities and ethical implications—in collaboration with the relevant communities (§ 5.4.5).

To facilitate this work, we have released our corpus of visualizations and labeled sentences under an open source license:

http://vis.csail.mit.edu/pubs/vis-text-model/data/

6.2 Natural Language As Coequal With Visualization

In closing, we turn to a discussion of our model’s implications for visualization theory. Not only can we think of natural language as an interface into visualization (as above), but also as an interface into data itself; coequal with and complementary to visualization. For example, some semantic content (e.g., Level 2 statistics or Level 4 explanations) may be conveyed via language, without any reference to visual modalities [22, 23], while other content (e.g., Level 3 clusters) may be uniquely suited to visual representation. This coequal framing is not a departure from orthodox visualization theory, but rather a return to its linguistic and semiotic origins. Indeed, at the start of his foundational Semiology of Graphics, Jacques Bertin introduces a similar framing to formalize an idea at the heart of visualization theory: content can be conveyed not only through speaking or writing but also through the “language” of graphics [12]. While Bertin took natural language as a point of departure for formalizing a language of graphics, we have here pursued the inverse: taking visualization as occasioning a return to language. This theoretical inversion opens avenues for future work, for which linguistic theory and semiotics are instructive [68, 77, 103].

Within the contemporary linguistic tradition, subfields like syntax, semantics, and pragmatics suggest opportunities for further analysis at each level of our model. And, since our model focuses on English sentences and canonical chart types, extensions to other languages and bespoke charts may be warranted. Within the semiotic tradition, Christian Metz (a contemporary of Bertin’s) emphasized the pluralistic quality of graphics [18]: the semantic content conveyed by visualizations depends not only on their graphical sign-system, but also on various “social codes” such as education, class, expertise, and— we hasten to include—ability. Our evaluation with blind and sighted readers (as well as work studying how charts are deployed in particular discourse contexts [34, 40]) lends credence to Metz’s conception of graphics as pluralistic: different readers will have different ideas about what makes visualizations meaningful (Fig. 1). As a means of revealing these differences, we have here introduced a four-level model of semantic content. We leave further elucidation of the relationship between visualization and natural language to future work.

ACKNOWLEDGMENTS

For their valuable feedback, we thank Emilie Gossiaux, Chancey Fleet, Michael Correll, Frank Elavsky, Beth Semel, Stephanie Tuerk, Crystal Lee, and the MIT Visualization Group. This work was supported by National Science Foundation GRFP-1122374 and III-1909991.
The Content Of Bar Charts On The Web. In WEBIST. SciTePress, 2007.

[32] C. Fisher. Creating Accessible SVGs, 2019.

[33] C. Fleet. Things which garner a ton of glowing reviews from mainstream outlets without being of much use to disabled people. For instance, Facebook’s auto image descriptions, much loved by sighted journos but famously useless in the Blind community. Twitter, 2021. https://twitter.com/ChanceyFleet/status/134921147744961536.

[34] S. L. Fossheim. An Introduction To Accessible Data Visualizations With D3.js. 2020.

[35] M. Galesic and R. Garcia-Retamero. Graph Literacy: A Cross-cultural Comparison. In Medical Decision Making. Society for Medical Decision Making, 2011.

[36] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Data-Tone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In UIST. ACM, 2015.

[37] S. García, A. Fernández, J. Luengo, and F. Herrera. Advanced Nonparametric Tests For Multiple Comparisons In The Design Of Experiments. In Computational Intelligence And Data Mining: Experimental Analysis Of Power. Information Sciences, 2010.

[38] B. Geveci, W. Schroeder, A. Brown, and G. Wilson. VTK. The Architecture of Open Source Applications, 2012.

[39] B. Gould, T. O’Connell, and G. Freed. Effective Practices for Description of Science Content within Digital Talking Books. Technical report, The WGBH National Center for Accessible Media, 2008. https://www.wgbh.org/foundation/nacam/guidelines/effective-practices-for-description-of-science-content-within-digital-talking-books.

[40] M. Hanley, S. Barocas, K. Levy, S. Azenkot, and H. Nissenbaum. Computer Vision and Conflicting Values: Describing People with Automated Alt Text. arXiv, 2021.

[41] L. Hasty, J. Milbury, I. Miller, A. O’Day, P. Acquinas, and D. Spence. Guidelines and Standards for Tactile Graphics. Technical report, Braille Authority of North America, 2011. http://www.brailleauthority.org/tg/.

[42] M. Hearst and M. Tory. Would You Like A Chart With That? Incorporating Visualizations into Conversational Interfaces. In VIES. IEEE, 2019.

[43] M. Hearst, M. Tory, and V. Setlur. Toward Interface Defaults for Vague Modifiers in Natural Language Interfaces for Visual Analysis. In VIES. IEEE, 2019.

[44] J. Hullman, N. Diakopoulos. Visualization Rhetoric: Framing Effects in Narrative Visualization. In TVCG. IEEE, 2011.

[45] J. Hullman, N. Diakopoulos, and E. Adar. Contextifier: automatic generation of annotated stock visualizations. In CHI. ACM, 2013.

[46] J. Hullman, N. Diakopoulos, E. Momeni, and E. Adar. Content, Context, and Critique: Commenting on a Data Visualization Blog. In CSCW. ACM, 2015.

[47] S. E. Kahou, V. Michalski, A. Atkinson, A. Kadar, A. Trischler, and Y. Bengio. FigureQA: An Annotated Figure Dataset for Visual Reasoning. arXiv, 2018.

[48] A. Karpathy and L. Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In TPAMI. IEEE, Apr. 2017.

[49] D. A. Keim and D. Oelke. Literature Fingerprinting: A New Method for Visual Literary Analysis. In VAST. IEEE, 2007.

[50] D. H. Kim, E. Hooque, and M. Agrawala. Answering Questions about Charts and Generating Visual Explanations. In CHI. ACM, Apr. 2020.

[51] D. H. Kim, E. Hooque, J. Kim, and M. Agrawala. Facilitating Document Reading by Linking Text and Tables. In UIST. ACM, 2018.

[52] D. H. Kim, V. Setlur, and M. Agrawala. Towards Understanding How Readers Integrate Charts and Captions: A Case Study with Line Charts. In CHI. ACM, 2021.

[53] N. W. Kim, S. C. Joyner, A. Riegelhuth, and Y. Kim. Accessible Visualization: Design Space, Opportunities, and Challenges. In CGF. Eurographics, 2021.

[54] H.-K. Kong, Z. Liu, and K. Karahalios. Frames and Slants in Titles of Visualizations on Controversial Topics. In CHI. ACM, Apr. 2018.

[55] H.-K. Kong, Z. Liu, and K. Karahalios. Trust and Recall of Information across Varying Degrees of Title-Visualization Misalignment. In CHI. ACM, May 2019.

[56] N. Kong, M. A. Hearst, and M. Agrawala. Extracting References Between Text And Charts Via Crowdsourcing. In CHI. ACM, 2014.

[57] S. M. Kosslyn. Understanding Charts and Graphs. Applied Cognitive Psychology, 1989.

[58] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei.
