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

Alan Lundgard and Arvind Satyanarayan

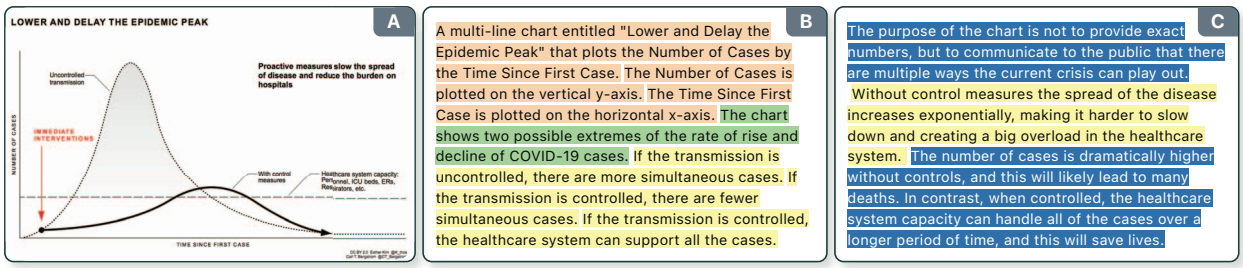


Fig. 1. Visualizations like "Flatten the Curve" (A) efficiently communicate critical public health information, while simultaneously excluding people with disabilities [11, 28]. To promote accessible visualization via natural language descriptions (B, C), we introduce a four-level model of semantic content.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

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.

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. "*For many people with various types of disabilities, graphics and the information conveyed in them is hard to read and understand,*" says software engineer Tyler Littlefield [28], who built a popular text-based COVID-19 statistics tracker after being deluged with inaccessible infographics [65, 94]. While natural language descriptions sometimes accompany visualizations in the form of chart captions or alt text (short for "alternative text"), these practices remain rare. Technology educator and researcher Chancey Fleet notes that infographics and charts usually lack meaningful and detailed descriptions, leaving disabled people with "*a feeling of uncertainty*" about the pandemic [28]. For readers with visual disabilities (approximately 8.1 million in the United States and 253 million worldwide [1]), inaccessible visualizations are, at best, demeaning and, at worst, damaging to health, if not accompanied by meaningful and up-to-date alternatives.

Predating the pandemic, publishers and education specialists have long suggested best practices for accessible visual media, including guidelines for tactile graphics [41] and for describing "complex images" in natural language [39, 99]. While valuable, visualization authors have yet to broadly adopt these practices, for lack of experience with accessible media, if not a lack of attention and resources. Contemporary visualization research has primarily attended to color vision deficiency [21, 77, 79], and has only recently begun to engage with non-visual alternatives [25, 67] and with accessibility broadly [53, 105]. Parallel to these efforts, computer science researchers have been grappling with the engineering problem of automatically generating chart captions [27, 78, 83]. While well-intentioned, these methods usually neither consult existing accessibility guidelines, nor do they evaluate their results empirically with their intended readership. As a result, it is difficult to know how useful (or not) the resultant captions are, or how effectively they improve access to meaningful information.

In this paper, we make a two-fold contribution. First, we extend existing accessibility guidelines by introducing a conceptual model for categorizing and comparing the semantic content conveyed by natural language descriptions of visualizations. Developed through a grounded theory analysis of 2,147 natural language sentences, authored by over 120 participants in an online study (§ 3), 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) (§ 4). Second, we demonstrate how this model can be applied to evaluate the effectiveness of visualization descriptions, by comparing different semantic content levels and reader groups. We conduct a mixed-methods evaluation in which a group of 30 blind and 90 sighted readers rank the usefulness

- *Alan Lundgard is with MIT CSAIL. E-mail: lundgard@mit.edu.*
- *Arvind Satyanarayan is with MIT CSAIL. E-mail: arvindsatya@mit.edu.*

Manuscript received 21 Mar. 2021; revised 13 June 2021; accepted 8 Aug. 2021. Date of publication 30 Sept. 2021; date of current version 22 Dec. 2021. Digital Object Identifier no. 10.1109/TVCG.2021.3114770

of descriptions authored at varying content levels (§ 5). Analyzing the 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* [7,8], 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., JPEGs 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 [98, 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-scraping 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 [69, 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 scenegraph 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 § 4.1.

2.2 Accessible Media and Human-Computer Interaction

While automatic methods researchers often note accessibility as a worthy motivation [27, 30, 31, 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 [9, 39]. 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 applicable 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 [99]. While some of these guidelines have been adopted by visualization practitioners [19, 29, 32, 34, 88, 101, 102], we here bring special attention to the empirically-grounded and well-documented guidelines created by the WGBH National Center for Accessible Media [39] 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 (§ 3). Perhaps

closest to our contribution are a pair of papers by Kosslyn [57] and Livingston & Brock [66]. 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 seeded our model construction with a similar linguistics-inspired framing, but also evaluated it empirically, to further decompose the semantic levels (§ 3.1). Livingston & Brock adapt Kosslyn's ideas to generate what they call "visual sentences": natural language sentences that are the result of executing a single, specific analytic task against a visualization [66]. Inspired by the Sentence Verification Technique (SVT) [85, 86], this work considers visual sentences for assessing graph comprehension, hoping to offer a more "objective" and automated alternative to existing visualization literacy assessments [35, 63]. While we adopt a more qualitative process for constructing our model, Livingston & Brock's approach suggests opportunities for future work: might our model map to similarly-hierarchical models of analytic tasks [5, 17]?

2.3.2 Using Natural Language to Interact with Visualizations

Adjacently, there is a breadth of work on Natural Language Interfaces (NLIs) for constructing and exploring visualizations [43, 50, 75, 90]. While our model primarily considers the natural language sentences that are *conveyed by* visualizations (cf., natural language as *input* for chart specification and exploration) [93], our work may yet have implications for NLIs. For example, Hearst et al. have found that many users of chatbots prefer *not* to see charts and graphics alongside text in the conversational dialogue interface [42]. By helping to decouple visual-versus-linguistic data representations, our model might be applied to offer these users a textual alternative to inline charts. Thus, we view our work as complementary to NLIs, facilitating multimodal and more accessible data representations [51], while helping to clarify the theoretical relationship between charts and captions [52, 80], and other accompanying text [2, 54, 55, 106].

3 Constructing the Model: Employing the Grounded Theory Methodology

To construct our model of semantic content we conducted a multi-stage process, following the *grounded theory* methodology. Often employed in HCI and the social sciences, grounded theory offers a rigorous method for making sense of a domain that lacks a dominant theory, and for constructing a new theory that accounts for diverse phenomena within that domain [74]. The methodology approaches theory construction *inductively* — through multiple stages of inquiry, data collection, "coding" (i.e., labeling and categorizing), and refinement — as well as *empirically*, remaining strongly based (i.e., "grounded") in the data [74]. To construct our model of semantic content, we proceeded in two stages. First, we conducted small-scale data collection and initial open coding to establish preliminary categories of semantic content. Second, we gathered a larger-scale corpus to iteratively refine those categories, and to verify their coverage over the space of natural language descriptions.

3.1 Initial Open Coding

We began gathering preliminary data by searching for descriptions accompanying visualizations in journalistic publications (including the websites of *FiveThirtyEight*, the *New York Times* and the *Financial Times*), but found that these professional sites usually provided no textual descriptions — neither as a caption alongside the chart, nor as alt text for screen readers. Indeed, often these sites were engineered so that screen readers would pass over the visualizations entirely, as if they did not appear on the page at all. Thus, to proceed with the grounded theory method, we conducted initial *open coding* (i.e., making initial, qualitative observations about our data, in an "open-minded" fashion) by studying preliminary data from two sources. We collected 330 natural language descriptions from over 100 students enrolled in a graduate-level data visualization class. As a survey-design pilot to inform future rounds of data collection (§ 3.2.1), these initial descriptions were collected with minimal prompting: students were instructed to simply "describe the visualization" without specifying what kinds of semantic content that might include. The described visualizations covered 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.

Table 1. Breakdown of the 50 curated visualizations, across the three dimensions: type, topic, and difficulty. (N.b., each column sums to 50.)

CHART TYPE		TOPIC		DIFFICULTY	
bar	18	academic	15	easy	21
line	21	business	18	medium	20
scatter	11	journalism	17	hard	9

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

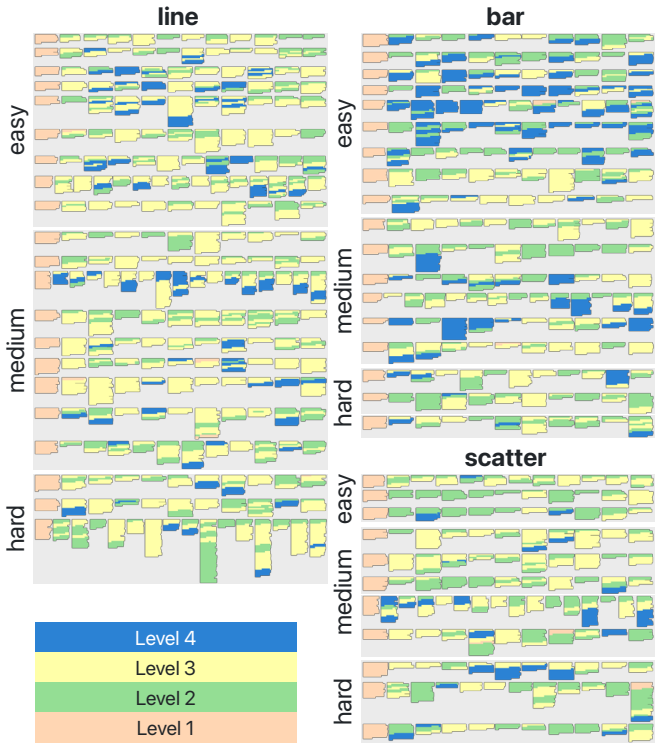


Fig. 2. A visual "fingerprint" [49] of our corpus, faceted by chart type and difficulty. Each row corresponds to a single chart. Each column shows a participant-authored description for that chart, color coded according to our model. The first column shows the provided Level 1 prompt.

basic elements and encodings (e.g., it's title, axes, colors) but to also referred to other content, trends, and insights that might be conveyed by the visualization. To make these instructions intelligible, we provided participants with a few pre-generated sentences enumerating the visualization's basic elements and encodings (e.g., the color coded sentences in Table 3 A.1, B.1, C.1), and prompted them to author semantic content *apart from* what was already conveyed by those sentences. To avoid biasing their responses, participants were *not* told that they would be read by people with visual disabilities. This prompting ensured that the survey captured a breadth of semantic content, and not only the most readily-apparent aspects of the visualization's construction.

3.2.2 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" [92, 100] through rigid or templatized 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]`").

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" [49] 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).

> 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+.

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 [12].

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. [104]. 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) [87]. 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 [47], or by extracting a visualization's structured specification from a rasterized visualization image [81].

Table 2. A four-level model of semantic content for accessible visualization. Levels are defined by the semantic content conveyed by natural language descriptions of visualizations. Additionally, we offer computational considerations for generating the semantic content at each level of the model.

#	LEVEL KEYWORDS	SEMANTIC CONTENT	COMPUTATIONAL CONSIDERATIONS
4	`contextual and domain-specific`	*domain-specific insights, current events, social and political context, explanations*	contextual knowledge and domain-specific expertise (*perceiver-dependent*)
3	`perceptual and cognitive`	*complex trends, pattern synthesis, exceptions, commonplace concepts*	reference to the rendered visualization and "common knowledge" (*perceiver-dependent*)
2	`statistical and relational`	*descriptive statistics, extrema, outliers, correlations, point-wise comparisons*	access to the visualization specification or backing dataset (*perceiver-independent*)
1	`elemental and encoded`	*chart type, encoding channels, title, axis ranges, labels, colors*	access to the visualization specification or rasterized image (*perceiver-independent*)

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 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 "lesser 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. Consider the following examples (Table 3.B.3 and 3.C.3, respectively).

> 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.

> 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.

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" of multiple graphed lines at that point in time), and do so in "natural"-sounding language, by referencing commonplace concepts (such as "fluctuate", "extreme drop", "visible gap"). N.b., "natural"-sounding articulation is necessary but insufficient for Level 3 membership, as it is also possible to articulate Level 1 or 2 content in a non-templatized fashion (§ 3.2.2).

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 *perceiver-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 *perceiver-dependent* (i.e., their content requires a perceiver of some sort; likely a human, although machine perception may increasingly suffice for generating Level 3 content). Table 2 summarizes this distinction.

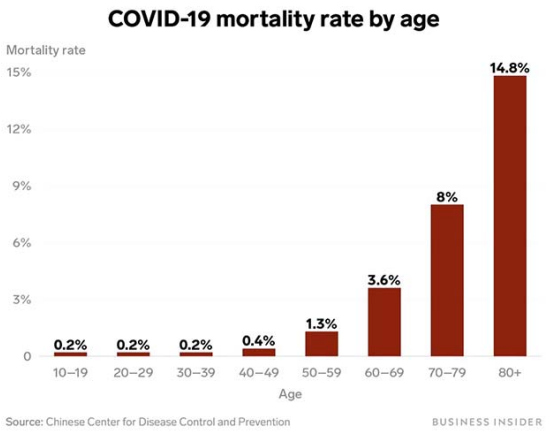
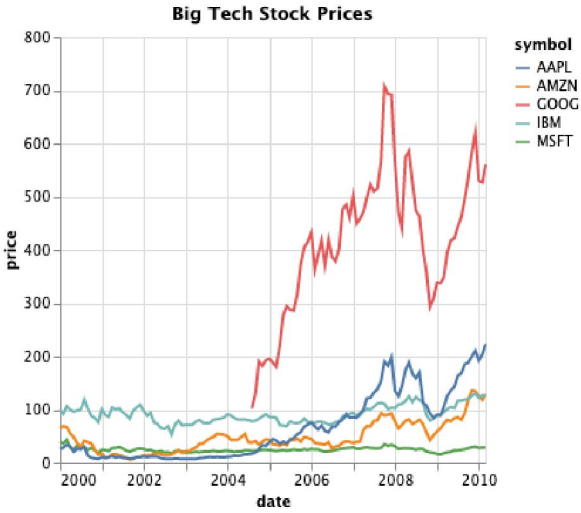
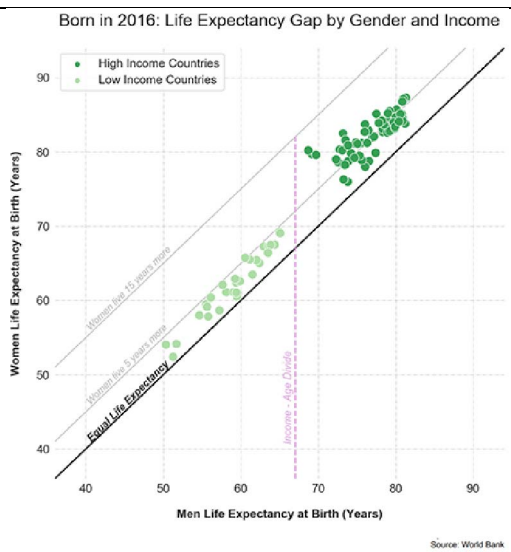
4.4 Level 4: Contextual and Domain-Specific Insights

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

> 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 (§ 3.2.1), while row C shows an curated exemplar description from our evaluation (§ 5.1).

	VISUALIZATION	DESCRIPTION
A	 [bar, easy, journalism]	[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. [6] The mortality rate increases exponentially with older people. [7] There is no difference in the mortality rate in the range between the age of 10 and 39. [8] The range of ages between 60 and 80+ are more affected by COVID-19. [9] We can observe that the mortality rate is higher starting at 50 years old due to many complications prior. [10] As we decrease the age, we also decrease the values in mortality by a lot, almost to none.
B	 [line, medium, business]	[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 had experienced some kind of a crisis in 2009, because their prices drop rapidly, but then rebounded.
C	 [scatter, hard, academic]	[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.

These sentences convey social and political explanations for an observed trend that depends on an individual reader's subjective knowledge about particular world events: the 2008 financial crisis and global socio-economic trends, respectively. This semantic content is characteristic of what is often referred to as "insight" in visualization research. Although lacking a precise and agreed-upon definition [20, 60, 61, 76, 95], an insight is often an observation about the data that is complex, deep, qualitative, unexpected, and relevant [108]. Critically, insights depend on individual perceivers, their subjective knowledge, and domain-expertise. Level 4 is where the breadth of an individual reader's knowledge and experience is brought to bear in articulating something "insightful" about the visualized data.

Computational Considerations. As with Levels 3, we say that Level 4 semantic content is perceiver-dependent, but in a stronger sense. This is because (setting aside consideration of hypothetical future "artificial general intelligence") generating Level 4 semantic content is at-present a uniquely human endeavor. Doing so involves synthesizing background knowledge about the world (such as geographic, cultural, and political relationships between countries), contextual knowledge about current events (e.g., the fact that there was a global recession in 2008), and domain-specific knowledge (e.g., expertise in a particular field of research or scholarship). However, bespoke systems for narrowly-scoped domains (e.g., those auto-generating stock chart annotations using a corpus of human-authored news articles [45]) suggest that some Level 4 content might be feasibly generated sooner rather than later.

Lastly, we briefly note that data-driven predictions can belong to either Level 2, 3, or 4, depending on the semantic content contained therein. For example: a point-wise prediction at Level 2 (e.g., computing a stock's future expected price using the backing dataset); a prediction about future overall trends at Level 3 (e.g., observing that a steadily increasing stock price will likely continue to rise); a prediction involving contextual or domain-specific knowledge at Level 4 (e.g., the outcome of an election using a variety of poll data, social indicators, and political intuition).

5 Applying the Model: Evaluating the Effectiveness of Visualization Descriptions

The foregoing conceptual model provides a means of making structured comparisons between different levels of semantic content and reader groups. To demonstrate how it can be applied to evaluate the effectiveness of visualization descriptions (i.e., whether or not they effectively convey meaningful information, and for whom), 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 piloted this rank-choice interface with 10 sighted readers recruited via *Prolific* and 1 blind reader, a non-academic collaborator proficient with Apple's VoiceOver screen reader. Based on this pilot, we rewrote the study instructions to be more intelligible to both groups of readers, added an introductory example task to the evaluation, and improved the screen reader accessibility of our interface (e.g., by reordering nested DOM elements to be more intuitively traversed by screen reader).

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 cleanly-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 3.C shows one such exemplar description, whereas Table 3.A 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 3.C] *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" [39, 78], 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]).

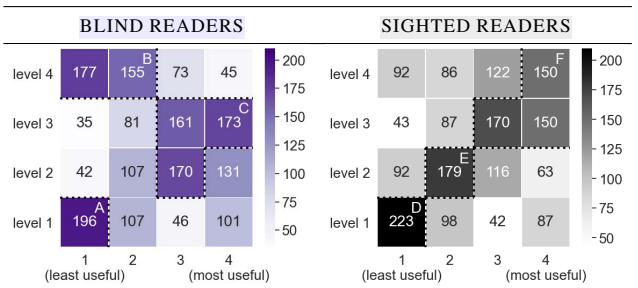
5.2.1 Participant Recruitment

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).

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 $\mu + \frac{\sigma}{2}$, each labeled with a capital letter A – F. (Lower) Shaded cells indicate significant ranking differences pair-wise between levels.



LEVELS	1 × 2	1 × 3	1 × 4	2 × 3	2 × 4	3 × 4
BLIND	$p < 0.001$	$p < 0.001$	$p < 0.321$	$p < 0.148$	$p < 0.001$	$p < 0.001$
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 participants reported 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 ($\mu + \frac{\sigma}{2}$) resulting in a value of 139 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 faceting 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" ($\mu = 6.83$, $\sigma = 0.38$).

> *"I am totally blind, and virtually all data visualizations I encounter are undescribed, and as such are unavailable. This has been acutely made clear on Twitter and in newspapers around the COVID-19 pandemic and the recent U.S. election. Often, visualizations are presented with very little introduction or coinciding text. I feel very left out of the world and left out of the ability to confidently traverse that world. The more data I am unable to access, the more vulnerable and devalued I feel."* (BR5)

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 among 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 presumptions 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, 78, 83, 84]. 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 being of much use to disabled people*" [33, 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 Beaudoin-Lafon's framework for evaluating interface models in terms of their descriptive, evaluative, and generative powers [7, 8], 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 affording readers the option to customize and combine different content levels (§ 5.4.4), depending on their individual preferences (§ 5.4.3). 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 model to develop and evaluate the 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: `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 best conveyed via language, without any reference to visual modalities [42, 82], 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, 97, 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 [3, 44, 46, 62]) 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-1900991.

REFERENCES

[1] P. Ackland, S. Resnikoff, and R. Bourne. World Blindness and Visual Impairment. *Community Eye Health*, 2017.

[2] E. Adar and E. Lee. Communicative Visualizations as a Learning Problem. In *TVCG*. IEEE, 2020.

[3] G. Aiello. Inventorizing, Situating, Transforming: Social Semiotics And Data Visualization. In M. Engebretsen and H. Kennedy, editors, *Data Visualization in Society*. Amsterdam University Press, 2020.

[4] K. M. Ali. Mind-Dependent Kinds. In *Journal of Social Ontology*, 2016.

[5] R. Amar, J. Eagan, and J. Stasko. Low-level Components Of Analytic Activity In Information Visualization. In *INFOVIS*. IEEE, 2005.

[6] A. Balaji, T. Ramanathan, and V. Sonathi. Chart-Text: A Fully Automated Chart Image Descriptor. *arXiv*, 2018.

[7] M. Beaudouin-Lafon. Instrumental Interaction: An Interaction Model For Designing Post-WIMP User Interfaces. In *CHI*. ACM, 2000.

[8] M. Beaudouin-Lafon. Designing Interaction, Not Interfaces. In *AVI*. ACM, 2004.

[9] Benetech. Making Images Accessible. http://diagramcenter.org/making-images-accessible.html/.

[10] C. L. Bennett, C. Gleason, M. K. Scheuerman, J. P. Bigham, A. Guo, and A. To. "It's Complicated": Negotiating Accessibility and (Mis)Representation in Image Descriptions of Race, Gender, and Disability. In *CHI*. ACM, 2021.

[11] C. T. Bergstrom. SARS-CoV-2 Coronavirus, 2020. http://ctbergstrom.com/covid19.html.

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

[13] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. VizWiz: Nearly Real-time Answers To Visual Questions. In *UIST*. ACM, 2010.

[14] J. P. Bigham, I. Lin, and S. Savage. The Effects of "Not Knowing What You Don't Know" on Web Accessibility for Blind Web Users. In *ASSETS*. ACM, 2017.

[15] M. A. Borkin, Z. Bylinskii, N. W. Kim, C. M. Bainbridge, C. S. Yeh, D. Borkin, H. Pfister, and A. Oliva. Beyond Memorability: Visualization Recognition and Recall. In *TVCG*. IEEE, 2016.

[16] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What Makes a Visualization Memorable? In *TVCG*. IEEE, 2013.

[17] M. Brehmer and T. Munzner. A Multi-Level Typology of Abstract Visualization Tasks. In *TVCG*. IEEE, 2013.

[18] A. Campolo. Signs and Sight: Jacques Bertin and the Visual Language of Structuralism. *Grey Room*, 2020.

[19] A. Cesal. Writing Alt Text for Data Visualization, Aug. 2020.

[20] R. Chang, C. Ziemkiewicz, T. M. Green, and W. Ribarsky. Defining Insight for Visual Analytics. In *CG&A*. IEEE, 2009.

[21] A. Chaparro and M. Chaparro. Applications of Color in Design for Color-Deficient Users. *Ergonomics in Design*, 2017.

[22] C. Chen, R. Zhang, S. Kim, S. Cohen, T. Yu, R. Rossi, and R. Bunescu. Neural Caption Generation Over Figures. In *UbiComp/ISWC '19 Adjunct*. ACM, 2019.

[23] C. Chen, R. Zhang, E. Koh, S. Kim, S. Cohen, and R. Rossi. Figure Captioning with Relation Maps for Reasoning. In *WACV*. IEEE, 2020.

[24] 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. *arXiv*, 2019.

[25] J. Choi, S. Jung, D. G. Park, J. Choo, and N. Elmqvist. Visualizing for the Non-Visual: Enabling the Visually Impaired to Use Visualization. In *CGF*. Eurographics, 2019.

[26] S. Demir, S. Carberry, and K. F. McCoy. Generating textual summaries of bar chartsGenerating Textual Summaries Of Bar Charts. In *INLG*. ACL, 2008.

[27] S. Demir, S. Carberry, and K. F. McCoy. Summarizing Information Graphics Textually. In *Computational Linguistics*. ACL, 2012.

[28] M. Ehrenkranz. Vital Coronavirus Information Is Failing the Blind and Visually Impaired. *Vice*, 2020.

[29] F. Elavsky. Chartability, 2021. https://chartability.fizz.studio/.

[30] S. Elzer, S. Carberry, D. Chester, S. Demir, N. Green, I. Zukerman, and K. Trnka. Exploring And Exploiting The Limited Utility Of Captions In Recognizing Intention In Information Graphics. In *ACL*. Association for Computational Linguistics, 2005.

[31] S. Elzer, E. Schwartz, S. Carberry, D. Chester, S. Demir, and P. Wu. A Browser Extension For Providing Visually Impaired Users Access To 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/1349211417744961536.

[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. DataTone: 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/ncam/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 *VIS*. IEEE, 2019.

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

[44] J. Hullman and 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. Hoque, and M. Agrawala. Answering Questions about Charts and Generating Visual Explanations. In *CHI*. ACM, Apr. 2020.

[51] D. H. Kim, E. Hoque, 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.

Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. In *IJCV*. Springer, 2017.
[59] C. Lai, Z. Lin, R. Jiang, Y. Han, C. Liu, and X. Yuan. Automatic Annotation Synchronizing with Textual Description for Visualization. In *CHI*. ACM, 2020.
[60] P.-M. Law, A. Endert, and J. Stasko. Characterizing Automated Data Insights. *arXiv*, 2020.
[61] P.-M. Law, A. Endert, and J. Stasko. What are Data Insights to Professional Visualization Users? *arXiv*, Aug. 2020.
[62] C. Lee, T. Yang, G. Inchoco, G. M. Jones, and A. Satyanarayan. Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online. In *CHI*. ACM, 2021.
[63] S. Lee, S.-H. Kim, and B. C. Kwon. Vlat: Development Of A Visualization Literacy Assessment Test. In *TVCG*. IEEE, 2016.
[64] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*. Springer, 2014.
[65] T. Littlefield. COVID-19 Statistics Tracker, 2020. https://cvstats.net.
[66] M. A. Livingston and D. Brock. Position: Visual Sentences: Definitions and Applications. In *VIS*. IEEE, 2020.
[67] A. Lundgard, C. Lee, and A. Satyanarayan. Sociotechnical Considerations for Accessible Visualization Design. In *VIS*. IEEE, Oct. 2019.
[68] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan. Visual Semiotics Uncertainty Visualization: An Empirical Study. In *TVCG*. IEEE, 2012.
[69] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell. Understanding Blind People's Experiences with Computer-Generated Captions of Social Media Images. In *CHI*. ACM, 2017.
[70] J. Matejka and G. Fitzmaurice. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In *CHI*. ACM, 2017.
[71] P. Moraes, G. Sina, K. McCoy, and S. Carberry. Evaluating The Accessibility Of Line Graphs Through Textual Summaries For Visually Impaired Users. In *ASSETS*. ACM, 2014.
[72] V. S. Morash, Y.-T. Siu, J. A. Miele, L. Hasty, and S. Landau. Guiding Novice Web Workers in Making Image Descriptions Using Templates. In *TACCESS*. ACM, 2015.
[73] M. R. Morris, J. Johnson, C. L. Bennett, and E. Cutrell. Rich Representations of Visual Content for Screen Reader Users. In *CHI*. ACM, 2018.
[74] M. Muller. Curiosity, Creativity, and Surprise as Analytic Tools: Grounded Theory Method. In J. S. Olson and W. A. Kellogg, editors, *Ways of Knowing in HCI*. Springer, 2014.
[75] A. Narechania, A. Srinivasan, and J. Stasko. NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries. In *TVCG*. IEEE, 2021.
[76] C. North. Toward Measuring Visualization Insight. In *CG&A*. IEEE, 2006.
[77] J. R. Nuñez, C. R. Anderton, and R. S. Renslow. Optimizing Colormaps With Consideration For Color Vision Deficiency To Enable Accurate Interpretation Of Scientific Data. *PLOS ONE*, 2018.
[78] J. Obeid and E. Hoque. Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. *arXiv*, 2020.
[79] M. M. Oliveira. Towards More Accessible Visualizations for Color-Vision-Deficient Individuals. In *CiSE*. IEEE, 2013.
[80] A. Ottley, A. Kaszowska, R. J. Crouser, and E. M. Peck. The Curious Case of Combining Text and Visualization. In *EuroVis*. Eurographics, 2019.
[81] J. Poco and J. Heer. Reverse-Engineering Visualizations: Recovering Visual Encodings from Chart Images. In *CGF*. Eurographics, 2017.
[82] V. Potluri, T. E. Grindeland, J. E. Froehlich, and J. Mankoff. Examining Visual Semantic Understanding in Blind and Low-Vision Technology Users. In *CHI*. ACM, 2021.
[83] X. Qian, E. Koh, F. Du, S. Kim, and J. Chan. A Formative Study on Designing Accurate and Natural Figure Captioning Systems. In *CHI EA*. ACM, 2020.
[84] X. Qian, E. Koh, F. Du, S. Kim, J. Chan, R. A. Rossi, S. Malik, and T. Y. Lee. Generating Accurate Caption Units for Figure Captioning. In *WWW*. ACM, 2021.
[85] J. M. Royer. Developing Reading And Listening Comprehension Tests Based On The Sentence Verification Technique (SVT). In *Journal of Adolescent & Adult Literacy*. International Literacy Association, 2001.
[86] J. M. Royer, C. N. Hastings, and C. Hook. A Sentence Verification Technique For Measuring Reading Comprehension. *Journal of Reading Behavior*, 1979.
[87] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-Lite: A Grammar of Interactive Graphics. In *TVCG*. IEEE, 2017.
[88] D. Schepers. Why Accessibility Is at the Heart of Data Visualization, 2020.
[89] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A Natural Language Interface for Visual Analysis. In *UIST*. ACM, 2016.
[90] V. Setlur, M. Tory, and A. Djalali. Inferencing Underspecified Natural Language Utterances In Visual Analysis. In *IUI*. ACM, 2019.
[91] A. Sharif, S. S. Chintalapati, J. O. Wobbrock, and K. Reinecke. Understanding Screen-Reader Users' Experiences with Online Data Visualizations. In *ASSETS*. ACM, 2021.
[92] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko. Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. In *TVCG*. IEEE, 2019.
[93] A. Srinivasan, N. Nyapathy, B. Lee, S. M. Drucker, and J. Stasko. Collecting and Characterizing Natural Language Utterances for Specifying Data Visualizations. In *CHI*. ACM, 2021.
[94] H. Sutton. Accessible Covid-19 Tracker Enables A Way For Visually Impaired To Stay Up To Date. *Disability Compliance for Higher Education*, 2020.
[95] B. Tang, S. Han, M. L. Yiu, R. Ding, and D. Zhang. Extracting Top-K Insights from Multi-dimensional Data. In *SIGMOD*. ACM, 2017.
[96] B. D. Team. *Bokeh: Python Library For Interactive Visualization*. Bokeh Development Team, 2014.
[97] P. Vickers, J. Faith, and N. Rossiter. Understanding Visualization: A Formal Approach Using Category Theory and Semiotics. In *TVCG*. IEEE, 2013.
[98] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and Tell: A Neural Image Caption Generator. In *CVPR*, 2015.
[99] W3C. WAI Web Accessibility Tutorials: Complex Images, 2019. https://www.w3.org/WAI/tutorials/images/complex/.
[100] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma, and D. Zhang. DataShot: Automatic Generation of Fact Sheets from Tabular Data. In *TVCG*. IEEE, 2020.
[101] L. Watson. Accessible SVG Line Graphs, 2017. https://tink.uk/accessible-svg-line-graphs/.
[102] L. Watson. Accessible SVG Flowcharts, 2018.
[103] W. Weber. Towards a Semiotics of Data Visualization – an Inventory of Graphic Resources. In *IV*. IEEE, 2019.
[104] L. Wilkinson. *The Grammar of Graphics*. Statistics and Computing. Springer-Verlag, 2005.
[105] K. Wu, E. Petersen, T. Ahmad, D. Burlinson, S. Tanis, and D. A. Szafir. Understanding Data Accessibility for People with Intellectual and Developmental Disabilities. In *CHI 2021*, 2021.
[106] C. Xiong, L. V. Weelden, and S. Franconeri. The Curse of Knowledge in Visual Data Communication. In *TVCG*. IEEE, 2020.
[107] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv*, 2016.
[108] J. S. Yi, Y.-a. Kang, J. T. Stasko, and J. A. Jacko. Understanding and Characterizing Insights: How Do People Gain Insights Using Information Visualization? In *BELIV*. ACM, 2008.