Understanding and Improving Drilled-Down Information Extraction from Online Data Visualizations for Screen-Reader Users

Ather Sharif
asharif@cs.washington.edu
Paul G. Allen School of Computer Science & Engineering | DUB Group, University of Washington
Seattle, Washington, USA

Andrew M. Zhang
azhang26@cs.washington.edu
Paul G. Allen School of Computer Science & Engineering, University of Washington
Seattle, Washington, USA

Katharina Reinecke
reinecke@cs.washington.edu
Paul G. Allen School of Computer Science & Engineering | DUB Group, University of Washington
Seattle, Washington, USA

Jacob O. Wobbrock
wobbrock@uw.edu
The Information School | DUB Group, University of Washington
Seattle, Washington, USA

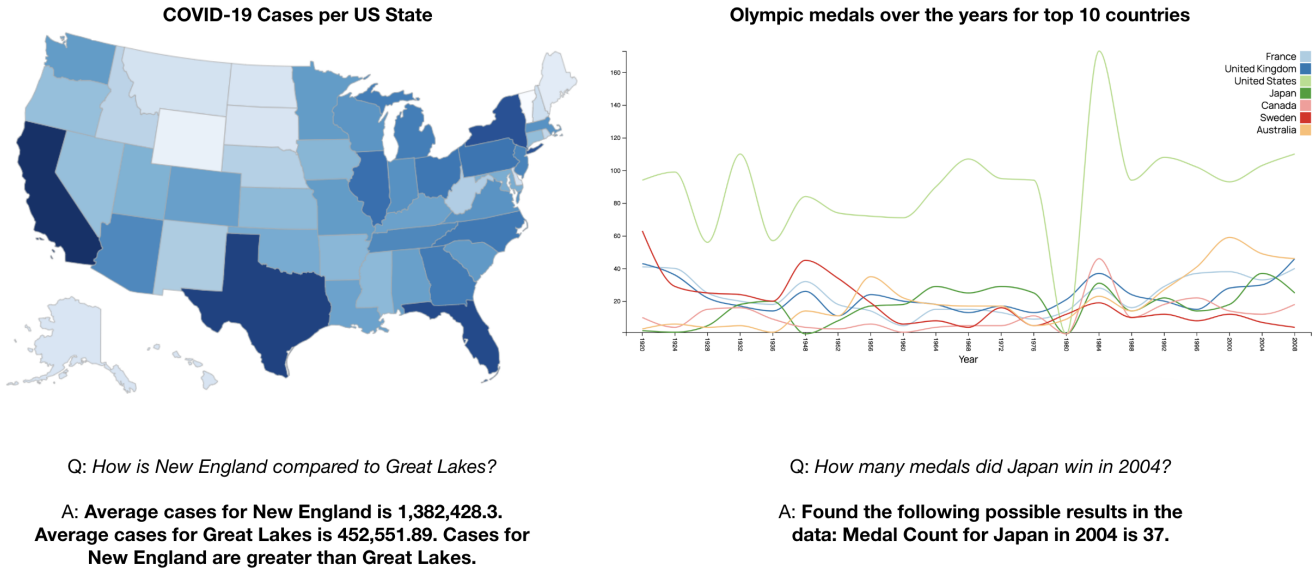


Figure 1: A screen-reader user's interaction with: (*left*) a geospatial map showing COVID-19 cases per US state, and (*right*) a multi-series line graph showing Olympic medals for the top 10 countries over multiple years. For each visualization, the user issues a question ("Q") to our system, VoxLens, which answers the user via their screen reader ("A").

ABSTRACT

Inaccessible online data visualizations can significantly disenfranchise screen-reader users from accessing critical online information. Current accessibility measures, such as adding alternative text to visualizations, only provide a high-level overview of data, limiting screen-reader users from exploring data visualizations in depth. In this work, we build on prior research to develop taxonomies of information sought by screen-reader users to interact with online data visualizations granularly through role-based and longitudinal studies with screen-reader users. Utilizing these taxonomies, we extended the functionality of VoxLens—an open-source multi-modal system that improves the accessibility of data visualizations—by supporting drilled-down information extraction. We assessed the performance of our VoxLens enhancements through task-based

W4A '23, April 30–May 01, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0748-3/23/04.
https://doi.org/10.1145/3587281.3587284

user studies with 10 screen-reader and 10 non-screen-reader users. Our enhancements "closed the gap" between the two groups by enabling screen-reader users to extract information with approximately the same accuracy as non-screen-reader users, reducing interaction time by 22% in the process.

CCS CONCEPTS

• **Human-centered computing → Information visualization**; **Accessibility systems and tools**; • **Social and professional topics → People with disabilities**.

KEYWORDS

Data visualization, accessibility, screen reader, blind, voice assistant.

ACM Reference Format:
Ather Sharif, Andrew M. Zhang, Katharina Reinecke, and Jacob O. Wobbrock. 2023. Understanding and Improving Drilled-Down Information Extraction from Online Data Visualizations for Screen-Reader Users. In *20th International Web for All Conference (W4A '23), April 30–May 01, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3587281.3587284

1 INTRODUCTION

Online data visualizations are commonly used on the web to effectively communicate large volumes of data [37]. Additionally, visualizations assist users in extracting information efficiently, helping people make informed life decisions concerning their health, finances, and activities. Recent work in politics [46, 84], health [32, 63, 85], finance [24, 76], and business analytics [6, 82] indicate the importance and wide adoption of online data visualizations.

However, the essential visual nature of data visualizations inherently disenfranchises people who cannot see [20, 47, 55, 68]. These people include people who use screen readers (over 7.6 million people in the United States) to read the contents of their computer screens. In contrast, non-screen-reader users can explore data visualizations by rapidly interpreting visual patterns [1, 28, 55]. Prior work has reported that alternative textual descriptions ("alt-text") for visualizations are often missing [68, 86]. In cases when alt-text is present, screen-reader users (SRUs) spend 211% more time and are 61% less accurate in extracting information than their non-screen-reader user counterparts [68]. Therefore, it is essential to find ways to make online data visualizations more accessible, efficient, and usable to screen-reader users (SRUs).[1]

Several prior works have attempted to improve the accessibility of online data visualizations [2, 30, 43, 44, 69, 70, 72], such as by auto-generating alt-text [44, 58, 69] and enabling verbal information extraction [70]. While these tools contribute to the accessibility of online visualizations, they either focus on simple graphs (e.g., single-series bar graphs) or the extraction of mainly high-level ("holistic") information, such as extrema and averages. Therefore, their granular ("drilled-down") interactions, such as extracting and comparing data points, especially with complex visualizations, remain unexplored.[2]

To achieve this goal, we employed a three-step process. First, we aimed to understand the granular information SRUs seek from simple and complex online data visualizations.[3] Then, we utilized these findings to develop taxonomies of the information sought by SRUs during their holistic and drilled-down explorations. Finally, using the taxonomies, we extended the functionality of VoxLens [70]—an open-source JavaScript plug-in that improves the accessibility of online data visualizations using a multi-modal approach—by supporting granular information extraction for SRUs.

To understand the granular information SRUs seek from online data visualizations, the questions they ask to extract that information, and the responses they prefer, we conducted a role-based and longitudinal user study with 12 and seven SRUs, respectively. For our role-based study, we employed a role-playing methodology [75], prompting our participants to explore data visualizations from different perspectives. Utilizing our findings, we composed taxonomies of the information sought by SRUs from visualizations.

We enhanced the capabilities of VoxLens using these taxonomies. To assess the performance of our enhancements, we conducted a task-based user study with SRUs who used VoxLens with our enhancements (N=10) and non-SRUs who did not use any tools (N=10). Our results show that using our enhancements, SRUs performed 5.6% *more* accurately than non-SRUs. (By contrast, using the original version of VoxLens, SRUs performed 15% *less* accurately than non-SRUs [70].) Our follow-up semi-structured interviews with SRUs revealed that our enhancements made VoxLens a "promising tool" (S1), helping users extract information quickly and accurately from online data visualizations.

The main contributions of this work are:

(1) We provide taxonomies of the information sought by SRUs in their holistic and drilled-down explorations of online data visualizations.
(2) We present our open-source enhancements to VoxLens [70] to support drilled-down information extraction from complex data visualizations (geospatial maps and multi-series line graphs; see Figure 1). We describe our design improvements and functional enhancements to VoxLens.
(3) We provide empirical results from a task-based user study with 10 SRUs and 10 non-SRUs to evaluate the performance of our enhancements.

2 RELATED WORK

We review research that has highlighted the need for accessible visualizations or provided recommendations and solutions to improve the interaction experiences of SRUs with online data visualizations.

2.1 Need for Accessible Data Visualizations

Prior research has identified the need for accessible online data visualizations, shedding light on the disenfranchisement caused by inaccessible visualizations for SRUs [45, 52, 55, 68]. Marriott *et al.* [55] put forward a call-to-action for inclusive visualizations, declaring the lack of access to visualizations a significant equity issue. Sharif *et al.* [68] provided empirical evidence of this inequity

[1] Following prior work [68, 70], we define "screen-reader users" as people who use screen readers (*e.g.*, JAWS [67]) and might have conditions including complete or partial blindness, low vision, learning disabilities (such as alexia), or motion sensitivity.
[2] We use the term "drilled-down" in line with its usage in the domain of *accessible* visualizations [68, 70, 86], as this term can have different meanings in different domains.
[3] We use the terms "simple visualizations" to refer to single-series bar graphs and "complex visualizations" to refer to geospatial maps and multi-series line graphs.

by conducting mixed-methods studies with 36 SRUs and 36 non-SRUs. Their results show that due to the inaccessibility of online data visualizations, SRUs extract information 61% less accurately and spend 211% more time interacting than non-SRUs.

2.2 Accessibility Recommendations

Researchers have recommended techniques to improve the accessibility of online data visualizations [17, 22, 52, 62, 68, 74]. Most recently, Sharif *et al.* [68] recommended auto-generation of alternative text ("alt-text") to represent dynamic data, and using multi-modality (e.g., tables, summaries, sonification) to enable SRUs to explore visualizations based on their preferences. Lundgard *et al.* [52] presented a set of sociotechnical considerations for accessible visualization designs, identifying participatory design and the usage of Accessible Rich Internet Application (ARIA) attributes as crucial elements in creating online data visualizations.

2.3 Solutions for Accessible Data Visualizations

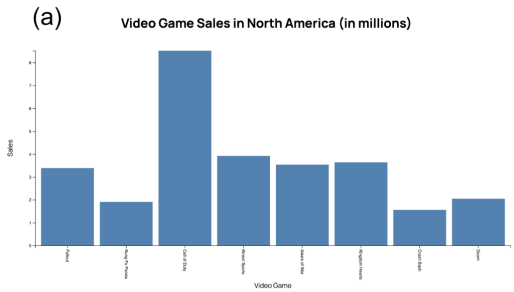
Several researchers have developed solutions to enhance the experiences of SRUs in extracting information from data visualizations, including auto-generation of alt-text [44, 53, 58, 69], sonification [2, 4, 13, 26, 36, 57, 83], data summarization [43], tables [21], and multi-modality [70]. Prior research has also developed solutions to make data visualizations accessible through other user interfaces, such as haptic graphs [78, 81] and 3-D printing [12, 40, 72]. As our work focuses on the accessibility of *online* data visualizations, we only discuss the work relevant to our exploration.

Most recently, Sharif *et al.* [70] developed VoxLens, an open-source JavaScript plug-in that enables SRUs to interact with online data visualizations using a multi-modal approach. VoxLens assists users in obtaining the data summary, listening to a sonified version of the data, and verbally interacting with visualizations. Kim *et al.* [43] generated summarization text displaying the high-level information from image-based line graphs using a multi-modal deep learning framework. Ahmetovic *et al.* [2] developed `AudioFunctions.web`, which enables blind people to explore mathematical function graphs using sonification.

These solutions focus on simple graphs (e.g., single-series two-dimensional graphs) and the extraction of holistic information (e.g., data summary or trend information). In contrast, our work: (1) presents a detailed account of screen-reader users' drilled-down explorations of simple and complex visualizations; (2) extends the findings and contributions of prior work [68, 71] by providing further insights into the holistic and drilled-down exploration of SRUs with online data visualizations; and (3) enhances the capabilities of VoxLens [70], enabling SRUs to verbally extract granular information from simple as well as complex visualizations.

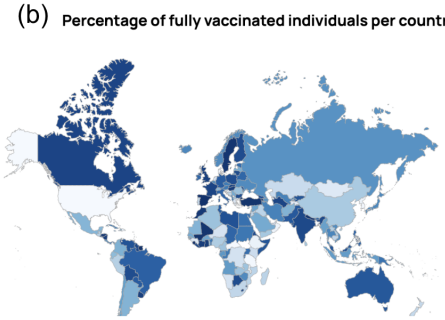
3 TAXONOMY DEVELOPMENT

We conducted role-based and longitudinal user studies [14, 48] with 12 and seven SRUs, respectively, to understand their information needs from online data visualizations. Utilizing our findings, we generated taxonomies of the information sought by screen-reader users for their holistic and drilled-down interactions. First, we present our methodology and results from the role-based and longitudinal user studies. Then, we discuss our process of taxonomy development.



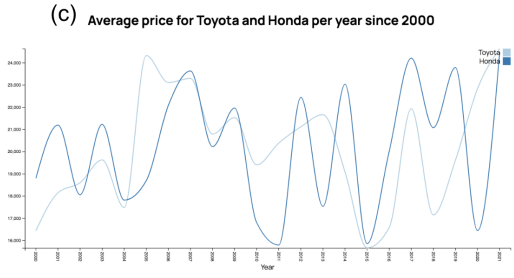
Q: *How many games are in this graph?*

A: **Data is from 8 video games. Say tell me factor levels for video games to hear all video games.**



Q: *Compare Africa to North America*

A: **Average vaccination percentage for Africa is 62.23. Average vaccination percentage for North America is 64.53. Vaccination percentage for North America is greater than Africa.**



Q: *Toyota price in 2015 versus Toyota 2020*

A: **Average price for Toyota 2015 is 15,660. Average price for Toyota 2020 is 22,832. Average Price for Toyota 2020 is greater than Toyota 2015.**

Figure 2: Sample visualizations shown during our longitudinal study. Plots (a) is a single-series bar graph, (b) is a geospatial map, and (c) is a multi-series line graph. "Q" is one of the questions participants asked during the study and "A" is the enhanced VoxLens's response.

3.1 Visualizations Selection

We selected three types of data visualizations based on their wide usage on the web: (1) *single-series bar graphs*; (2) *multi-series line graphs*; and (3) *geospatial maps*. Then, we curated a set of 30 online data visualizations (10 for each type) based on the search results for "most popular data visualizations 2021" and "most popular map visualizations 2021" on Google. Figure 2 shows three of the 30 visualizations used in our longitudinal study.

3.2 Overview of VoxLens

VoxLens is an open-source JavaScript plug-in that improves the accessibility of online data visualizations for SRUs using a multimodal approach [70]. Additionally, VoxLens requires only a single line of code for integration into existing and future online data visualizations created using D3, Google Charts, or ChartJS. There are three modes of VoxLens: (1) *Question-and-Answer mode*, where the user verbally interacts with the visualizations; (2) *Summary mode*, where VoxLens gives a summary of the visualization and the information it contains; and (3) *Sonification mode*, where VoxLens enables listeners to interpret data trends by mapping data to a musical scale. Prior to the enhancements presented in this work, VoxLens was limited to only simple visualizations created using two-dimensional single-series data, such as single-series bar graphs.

3.3 Role-Playing User Study

We conducted a Wizard-of-Oz [19, 33] role-based user study [14, 48] with 12 SRUs acting as the "wizards" and simulating responses from a hypothetical screen reader. Our goal was to elicit diverse perspectives and motivations for SRUs to perform information extraction granularly from online data visualizations. Building on and following recommendations from prior work [39, 49, 56], we identified three roles that provided in-depth perspectives: (1) *explorer*; (2) *teacher*; and (3) *news reporter*. We were unsuccessful at finding SRUs who were actual teachers or news reporters, as is often the case with recruiting disabled participants, particularly those with specialties [65]. Therefore, we used role-playing in our user study.

3.3.1 Participants. As our goal was to build on prior work [71], we recruited the same participants as in that user study. These participants were 12 SRUs (M=50.3 years, SD=13.6), recruited via word-of-mouth, snowball sampling, and advertisements through email distribution lists for disabled people. All but one participant had complete blindness; six participants had been blind since birth, and five had lost their vision gradually. The recruitment of participants ceased after reaching saturation of insights, as in prior work [68, 70, 80]. Participants received a $20 Amazon gift card for one hour of their time.

3.3.2 Procedure. We conducted hour-long user studies with our participants using Zoom videoconferencing and collected their demographic information. At least two authors took detailed notes during the sessions. We utilized Zoom's built-in features for recording and transcribing sessions. First, we presented participants with a summary of the visualization generated using the *Summary* mode from VoxLens [70] to provide them with a holistic overview of the data. Then, we had our participants explore the data in the visualization by verbally asking questions, to which we responded as "wizards," replicating the behavior of the *Question-and-Answer* mode of VoxLens. Each participant interacted with nine visualizations randomly selected from the 30 visualizations we curated (three of each type). We randomized the order of the visualizations across participants.

We randomly assigned each participant a unique role (explorer, teacher, news reporter) for each of the three visualizations in each visualization type. As an "explorer," the participants interacted with the visualizations based on their curiosity and interests; as a "teacher" and "news reporter," they had to extract information assuming they were to present it to their students and news audience, respectively. We explained the definitions of these roles to the participants before conducting the study. Hence, each participant took on each role three times. Participants did not portray any difficulty in assuming the roles. We counterbalanced the order of roles across participants. On average, our participants spent approximately five minutes interacting with each visualization.

3.4 Longitudinal Study

We conducted a longitudinal study with seven SRUs to gain more insight into their holistic and drilled-down information extraction behaviors. Our study with each participant lasted 12 days, including a tutorial session and an hour-long follow-up interview.

3.4.1 Participants. Our participants were seven SRUs (M=48.1 years, SD=8.7), recruited similarly like our role-based user study. Five had complete blindness, four of whom had been blind since birth. We compensated people with a $230 Amazon gift card for participating in our longitudinal study of 12 days.

3.4.2 Procedure. On the first of 12 days, our participants took part in a tutorial conducted using Zoom videoconferencing, in which we asked them to share their screen and computer sound. They interacted with a sample visualization using all modes of VoxLens until they were comfortable. We used these modes because they were commonly used (Sonification and Summary modes) or a new interaction technique (Question-and-Answer mode) to make online data visualizations accessible.

On days 2–11, we asked our participants to interact daily with our curated visualizations using all three VoxLens modes and extract information from the visualizations. All participants interacted with three visualizations per day (one of each type). Therefore, each participant interacted with 30 different visualizations, spending approximately 10 minutes with each visualization. We logged their interactions, including any queries or commands they issued to VoxLens and the responses they received. Additionally, at the end of a participant's interaction with each visualization, we asked them to summarize the information they extracted from the visualization. All sessions were unsupervised. On the 12th and final day of the study, we conducted hour-long follow-up interviews. Specifically, we asked participants about their overall experiences with each type of visualization.

3.5 Data Analysis

We used a mixed-methods approach to analyze our data. Specifically, we employed both quantitative and qualitative methods to analyze the data from our role-based and longitudinal user studies. Our

primary sources of data were the interaction logs from VoxLens and our semi-structured interviews.

3.5.1 Quantitative Analysis. Our goal was to explore differences between the commands issued and the effects of different visualization types on the information sought by SRUs. Therefore, our independent variables were *Command Issued* (*CMD*)[4], representing the information sought by screen-reader users, and visualization type (*VT*). *Count* (*CNT*) was our dependent variable, calculated as the number of times each participant issued a command per chart type. We employed a mixed Poisson regression model [29] to analyze *CNT*, as is standard practice for count data. Additionally, we included $Subject_r$ as a random factor [27] to account for repeated measures on the same participants.

3.5.2 Qualitative Analysis. We used semantic thematic analysis [54, 64] to analyze the interview sessions from our role-playing and longitudinal studies. We employed Braun and Clarke's [9] "essentialist" method for our thematic analysis, combining 33 initial codes into 18 axial codes. We separated the axial codes for each exploration type (holistic and drilled-down). Each exploration type contained nine axial codes. Finally, we classified our axial codes into two broader categories within each exploration type. Additionally, we calculated inter-rater reliability, expressed as percentage agreement among raters before resolving disagreements [35]. Our percentage agreement was 90.1%, demonstrating a high level of agreement between raters [31, 35].

3.6 Quantitative Results

Command Issued (*CMD*) had a significant effect on *Count* (*CNT*) ($\chi^2(10, N=1370)=8336.1$, $p<.001$, Cramer's $V=0.93$). Specifically, the most issued command was the `Value` command, issued 28.6% of the time to extract the value of an individual data point.

We also examined the interaction between *Command Issued* and *Visualization Type* (*CMD* × *VT*), finding a significant effect ($\chi^2(20, N=1370)=5640.7$, $p<.001$, Cramer's $V=0.77$). For a *Single-Series Bar Graph*, `Factor` was the most issued command (26.7%); for *Multi-Series Line Graph* and *Geospatial Map*, it was `Average` (22.0%) and `Value` (46.4%), respectively. (The `Factor` command enables users to get information about the independent and dependent variables.)

3.7 Qualitative Results

We present our findings for each exploration type (holistic and drilled-down) and discuss our taxonomy development process.

3.7.1 Holistic Exploration. SRUs look for holistic information in data visualizations as an initial step before exploring data in detail [68]. We identified two high-level categories for SRUs' holistic explorations: (1) *Summary Statistics* and (2) *Understanding Trends*. To obtain summary statistics, our participants looked for *extrema, averages, axis ranges, factor levels, medians,* and *sums* (in that order of frequency). For example, P9 compared the two "extremes" by asking for the maximum and minimum data points:

> *I would definitely write down the minimum data point and the maximum data point. So then I can compare the two extremes.* (P9)

[4]Full list of VoxLens commands is present in Table 2 in our prior work [70]

Similarly, P4 was interested in learning about the average of COVID-19 cases in North America:

> *What's the rough average in North America?* (P4)

Our participants also sought data trend information. Specifically, they looked for *overall trend, best-fit line,* and *correlation coefficient*. P12 looked for "visual representation" of the data:

> *And then what I would do is ask for some type of sonification of that graph that'll let me get a visual representation of it.* (P12)

Overall, our findings show that SRUs look for summary statistics and data trends to explore the data holistically. In our role-playing and longitudinal user studies, *extrema* were the most commonly sought information.

3.7.2 Drilled-Down Exploration. Our user studies revealed that SRUs perform drilled-down explorations by extracting and comparing data points. These explorations were straightforward for single-series data (bar graphs, geospatial maps) with at most one independent factor. However, for multi-series data with multiple factors (i.e., multi-line graphs), our participants not only performed the extraction and comparison *within* but also *across* different factors. For example, in a graph of housing prices over the past 10 years per U.S. state, participants looked for:

- *Extraction; within factors*: The data for a given state (e.g., Texas housing price average).
- *Comparison; within factors*: The data for a given state compared to another state (e.g., Texas vs. Oregon).
- *Extraction; across factors*: The data for a given state in a given year (e.g., Texas 2017).
- *Comparison; across factors*: The data for a given state in a given year compared to another year for a different state (e.g., Texas 2017 vs. Oregon 2019).

Our participants categorized data as *regional, political, climate-related, population-related,* and *spoken-language-related* to extract information from geospatial data. For example, P1 was interested in finding the United States traffic congestion differences between eastern, western, northern, and southern regions:

> *I mean, is there a difference between the west and the east or north and south?* (P1)

Similarly, P2 categorized the data by political spectrum, whereas P10 was interested in climate categorization:

> *The federal funding... What happened when the Republicans were in power and when the Democrats were in power? That would be interesting to check out.* (P2)
>
> *Usually, there's not much information presented about that in maps, but I would like to see how states compare with climate, I mean cold, hot, whatever.* (P10)

We found that our participants performed categorization by factor levels for multi-series data. P2 wanted to categorize the alcohol consumption data for Canada for every five years:

> *I would like to be able to isolate these [data points] and say... I would ask that what is the number in Canada for the past five years? Like, by five-year increments?* (P2)

We found that SRUs ranked data based on the *top*, *bottom*, and *near-average* values. For example, P2 was interested in the average housing prices for U.S. states in 2021:

> *Well, I would like to get an order of the lowest to the highest, so I can maybe organize this in my head.* (P2)

Our findings show that SRUs categorize and rank the data to extract and compare data points. We found regional and factor-level categorization the most frequently sought information for geospatial and multi-series data, respectively.

3.7.3 Taxonomy Development Process. We developed two taxonomies of the information sought by SRUs. Our taxonomies contain three tiers: (1) *Category* (broader categories); (2) *Information Type* (axial codes); and (3) *Query* (participants questions). Appendix A shows the taxonomy for holistic exploration, organizing the categories and information types in the order of their frequency (most to least). To build the taxonomy for drilled-down exploration, we collected the information SRUs seek to extract data points. Appendix B shows the taxonomy for drilled-down exploration. Through these taxonomies, we produced generalizable knowledge that visualization creators and researchers can use to improve the accessibility of online data visualizations. To demonstrate the utility of the above discoveries and taxonomies, we extended the capabilities of VoxLens—an open-source JavaScript plug-in that makes online data visualizations accessible to SRUs using a multi-modal approach.

4 ENHANCEMENTS TO VOXLENS

Utilizing the taxonomies from our role-playing and longitudinal studies, we enhanced the functionality of VoxLens by supporting drilled-down information extraction from complex visualizations (geospatial maps and multi-series line graphs). We present the design, features, and implementation of our enhancements.

4.1 Our Additions to VoxLens

We extended the functionality of the *Question-and-Answer* mode, as the other two interaction modes only assist in holistic exploration. We added two more parameters to the existing configuration options: "chartType" and "dataModule." Five values for "chartType" are possible: (1) *bar* (single-series bar graphs); (2) *line* (single-series line graphs); (3) *scatter* (single-series scatter plots); (4) *map* (geospatial maps); and (5) *multiseries* (multi-series line graphs). VoxLens currently supports (1), (2), and (3), whereas (4) was introduced in prior work [71]. In this work, we improved upon and provided additional details for (4) (maps) and introduced (5) (multi-series line graphs) as a significant new feature for VoxLens.

We selected the most frequently sought information types from our taxonomies to implement as additional features for VoxLens. From our taxonomy of the holistic information, *extremum, average, axis ranges, factor levels,* and *overall trend* were the most frequently used. Out of these five, *extremum, average* and *overall trend* were already implemented in VoxLens. Therefore, we implemented *axis ranges* and *factor levels*. For drilled-down information extraction, the most frequently used information types were *regional categorization for geographic data, factor-level categorization for multi-series line graphs,* and *top* and *bottom* for the *ranking* category.

4.2 Factor-Level Categorization (Multi-Series)

We used the keyword matching algorithm from VoxLens to support categorization by factor levels.[5] Specifically, we searched the user's query to find keywords matching the factor levels. For example, if the user said, "Tell me the housing price for Texas," our algorithm would identify "Texas" as the factor level and calculate the average housing prices in Texas for the past 10 years. We used "average" as the default command based on the findings from our role-playing and longitudinal studies. However, users can specify other statistical measures based on their needs (e.g., "Total housing price for Texas").

However, if the user said, "Tell me the housing price for Texas in 2017," our algorithm would identify "Texas" as the factor level for *state* and "2017" as the factor level for *year*. We employed the same strategy to perform comparisons between data points. To obtain *all* the data points, our users suggested making the data available through a table, as listening to values from large data sets can induce high cognitive overload. We discuss this further in our subsection below on additional features.

In addition to line graphs, our enhancements are generalizable to other visualizations created using multi-series data. We also note that, currently, our enhancements are limited when categorizing by factor levels if the input query contains prepositions or adverbs of time (e.g., housing price for Texas five years "ago" or housing price for Texas "between" 2017 and 2019). Future work can utilize more advanced Natural Language Processing techniques to understand such queries.

4.3 Regional Categorization (Geospatial Maps)

Our participants extracted and compared data points from geospatial maps by performing regional categorization. As all of our participants were from the United States, they categorized the data by regions within the United States (e.g., east coast); for countries of the world, they grouped the data by continents (e.g., Asia). We further expanded the *state* module based on the National Geographic Society's classification of regions in the United States [61]. Specifically, we added the following regions: *Mountain West*, *Far West*, *Northwest*, *Northeast*, *Southeast*, *Midsouth*, *New England*, *Central*, *Southcentral*, and *Great Lakes*. As the modules are open-source and engineered to be scalable, developers can easily make necessary adjustments to the VoxLens library to extend the modules to add more regions (and provinces). Figure 1 shows an interaction of an SRU with a geospatial map after our enhancements. Similar to multi-series line graphs, our enhancements for geospatial maps are generalizable to other visualizations created using geospatial data.

4.4 Improvements for Ranking

VoxLens enables users to obtain the top N and bottom N data points, where N represents the number of data points. However, the existing algorithm only recognizes specific keywords to rank the data (e.g., "top" or "bottom"). Therefore, we extended the vocabulary of VoxLens in our enhancement to recognize more keywords (e.g., "most" and "least").

[5] In a graph showing average housing prices per U.S. state for the past 10 years, *state* and *year* are factors. For *state*, the levels are the 50 U.S. states; for *year*, they are the last 10 years. Housing prices is the dependent variable.

4.5 Additional Features

We implemented additional functionality for SRUs to obtain the factor levels and range of the dependent variable and acquire the data set via a table.

4.5.1 Factor Levels. We employed a two-step interaction design for factors and their levels. The first step presents the users with the count for the factor levels and the second step enables them to obtain the list. For example, in a visualization showing the stock market prices per company, if the user asks, "How many companies are present?" the response is, "Data is from three companies. Say 'tell me factor levels for companies' to hear all companies." If the user asks for the factor levels, the response is, "Data is from three companies: Apple, Microsoft, and Google."

4.5.2 Range for the Dependent Variable. We added the functionality to obtain the range of the dependent variable, providing users with the minimum and maximum values. For example, if the user asks for the stock market price range, the response is, "Stock market price ranges from [minimum value] to [maximum value]."

4.5.3 Visually-Hidden Tables. Our participants expressed an interest in obtaining the entire data set. They noted that through the Question-and-Answer mode of VoxLens, such information could be cumbersome to process, especially for data sets with large cardinalities. They suggested presenting the data using visually-hidden tables, a strategy employed by Google Charts [21]. Therefore, we appended a table to the end of the visualization, which was visually hidden but accessible to screen readers.

5 TASK-BASED USER STUDY

To assess the performance of our VoxLens enhancements, we conducted a task-based user study with follow-up interviews.

5.1 Method

We evaluated the performance of our VoxLens enhancements through a task-based user study with 10 SRUs who used VoxLens and 10 non-SRUs who did not use any tools to aid in their interaction. We also administered the NASA-TLX questionnaire [34].

5.1.1 Participants. We recruited 10 SRUs and 10 non-SRUs for our study, advertising via word-of-mouth, snowball sampling, and email distribution lists. No participants had partaken in our role-playing or longitudinal user studies. Among SRUs, the average age was 46.1 years (SD=11.8). All participants had complete blindness; seven were blind since birth, and three had lost their vision gradually. For non-SRUs users, the average age was 45.9 (SD=8.3) years. We did not find a statistically significant difference between the ages of the two participant groups ($t(18)$=0.04, $n.s.$). We compensated SRUs with a $20 Amazon gift card and non-SRUs with a $15 Amazon gift card for 45–60 minutes and 20–35 minutes of their time, respectively.

5.1.2 Apparatus. We conducted our task-based user study online using the study platform from prior work [70], created using the JavaScript React framework [41]. We randomly selected nine data visualizations from our 30 curated visualizations (see Section 3.1; three for each visualization type). We implemented these visualizations following the WCAG 2.0 Guidelines [15]. The performance of users with VoxLens did not significantly differ between visualization libraries [70]. Therefore, we generated all visualizations using D3 [8]. All visualizations supported interactive features (e.g., hover and click), as our goal was to accurately replicate the current behavior of these visualizations on the Web. We include all the visualizations in supplementary materials for reproducibility.

We finalized the question categories for our task-based user study based on Brehmer and Munzner's task topology [11] and prior work on the type of questions users ask of graphs [10, 42, 68]. We intentionally did not base our questions on our findings from the role-playing and longitudinal studies. Overall, we identified five question categories for each visualization:

(1) *Order Statistics*: Extraction of the maximum/minimum data point, chosen randomly (e.g., "Which state has the minimum number of cases in this visualization?").
(2) *Symmetry Comparison*: Identification of the relationship between two data points (e.g., "Are the cases of state 'Oregon' greater, lesser, or equal compared to 'Michigan'?").
(3) *Value Retrieval*: Extraction of the value of an individual data point (e.g., "What is the number of cases for state 'California'?").
(4) *Ranking*: Identification of the highest/lowest ranked data points, chosen randomly (e.g., "Which of these states is not in the top three based on cases?").
(5) *Chart-Type Specific Questions*:
    - *Factor Levels*: Extraction of the total number of levels for any given independent variable (e.g., "How many countries are shown in this visualization?"). We randomly selected the independent variable for multi-series line graphs.
    - *Symmetry Comparison by Region*: Identification of the relationship between two regions (e.g., "Are the cases on the east coast greater, lesser, or equal compared to the west coast?"). We only asked this question for geographic map-based visualizations.

All questions were multiple-choice with four choices: the correct answer, two incorrect answers, and the option for "Unable to extract information." Following the study design from prior work [68, 70], we determined the incorrect answer choices programmatically, randomly choosing two data points for categorical values and two integers using a random number generator for numerical values. We randomized the order of the four choices for each question.

5.1.3 Procedure. We conducted an over-the-shoulder-style user study [77] and asked our participants to share their screens and make their screen reader's audio outputs audible using Zoom videoconferencing. The study sessions were unsupervised for non-SRUs. At the beginning of the study, we collected preliminary information from participants. In the next step, we engaged with our SRUs in an interactive tutorial session, introducing the operations and functions of our enhancements (e.g., activating the tool and issuing commands). Using a sample visualization of a single-series bar graph, we assisted our SRUs in extracting information from the visualization using the *Question-and-Answer* mode. We shared sample questions for them to ask and guided them until they expressed confidence in their familiarity with our system. We did not present a tutorial session to non-SRUs.

Table 1: Statistical test results for *Accuracy of Extracted Information* (*AEI*) from SRUs using VoxLens with our enhancements (N=10) and non-SRUs without VoxLens (N=10). *SRU* is the screen-reader factor and *VT* is the visualization type factor. Cramer's V is a measure of effect size [25].

	N	χ^2	p	Cramer's V
SRU	20	0.01	.903	.00
VT	20	9.33	< .05	.12
SRU × *VT*	20	3.92	.141	.08
Age	20	0.24	.626	.02

After the tutorial session, each participant completed five study tasks for each visualization type, totaling 5×3=15 study tasks. Each task comprised three steps. Step 1 contained the question to explore; step 2 displayed the question and visualization; step 3 presented the question with four answer choices. We randomized the order of the study tasks and the order of the multiple-choice responses. We did not interact with the participants while they performed the tasks. Finally, we asked them to fill out the NASA-TLX workload questionnaire [34]. For SRUs, they took 45–60 minutes; for non-SRUs, study sessions took from 20–35 minutes.

5.1.4 Design and Analysis. The experiment was a 2 × 3 mixed-factorial design with the following factors and levels: (1) *Screen-Reader User (SRU)*, between-Ss.: {yes, no}, and (2) *Visualization Type (VT)*, within-Ss.: {single-series bar graph, multi-series line graph, geospatial map}. We used the *SRU* factor to empirically explore the gap in information access between SRUs and non-SRUs. (We did not compare our enhancements to the original VoxLens to avoid making a strawman comparison.) Hence, our SRUs interacted with visualizations using our enhancements, and non-screen-reader users interacted with the visualizations as they usually would.

We used *Accuracy of Extracted Information* (AEI) and *Interaction Time* (IT) as our dependent variables. In our analysis, *AEI* was coded as "1" for "accurate" when the user answered the question correctly and "0" for "inaccurate." Additionally, we calculated *IT* as the total task completion time. We used a mixed logistic regression model [29] and a linear mixed model [27, 51] to analyze *AEI* and *IT*, respectively. We used the above factors, their interactions, and a covariate to control for *Age* in our statistical models. We also included $Subject_r$ as a random factor to account for repeated measures [27]. Therefore, our statistical model terms were: $SRU + VT + SRU{\times}VT + Age + Subject_r$. We tested our participants over three *Visualization Type* (*VT*) conditions, resulting in a total of 3×5=15 trials per participant. With 20 participants, we had a total of 20×15=300 study trials.

5.1.5 Qualitative and Subjective Evaluation. To qualitatively assess the usability and usefulness of our enhancements, we conducted follow-up interviews with all SRUs (N=10). Specifically, we asked them about their overall experience, liked features, areas for improvement, and any issues they encountered during their interactions. To assess the subjective ratings, we administered the NASA-TLX workload questionnaire [34] with all participants (N=20).

5.2 Quantitative Results

We present the results of our task-based user study assessing the *Accuracy of Extracted Information* (*AEI*) and *Interaction Time* (*IT*) for SRUs and non-SRUs with online data visualizations. Our goal was to investigate the performance of our enhancements with users and not to assess their cognitive or intellectual abilities.

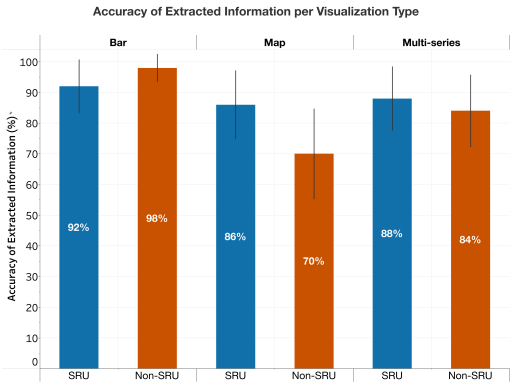


Figure 3: Accuracy of Extracted Information (*AEI*), as a percentage, for SRUs with VoxLens (N=10) and non-SRUs without VoxLens (N=10) by Visualization Type (*VT*). The percentage represents the "accurate" answers (higher is better). Error bars represent mean ±1 standard deviation.

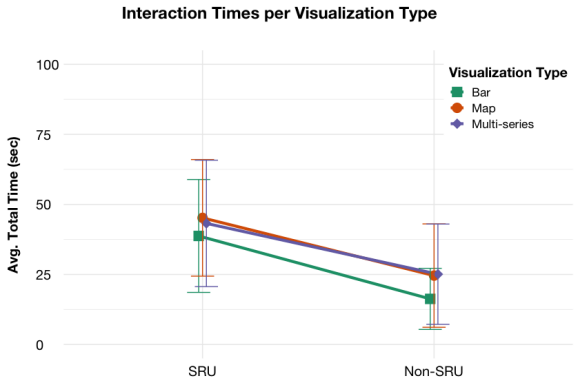


Figure 4: Interaction Times (*IT*), in seconds, for SRUs using our VoxLens enhancements (N=10) and non-SRUs without VoxLens (N=10) by Visualization Type (*VT*). Lower is better (faster). Error bars represent mean ±1 standard deviation.

5.2.1 Accuracy of Extracted Information (AEI). Our results do not show a significant effect of *Screen-Reader User* (*SRU*) on *AEI* overall ($p \approx .906$), indicating that *AEI* was not detectably different between the two *Screen-Reader User* groups. In fact, using our enhancements, SRUs extracted information 5.6% more accurately on average than non-SRUs, although this was not statistically significant. However,

Table 2: Statistical test results for *Interaction Time* (*IT*) from SRUs using VoxLens with our enhancements (*N*=10) and non-SRUs without VoxLens (*N*=10). *SRU* is the screen-reader factor and *VT* is the visualization type factor. Partial eta-squared (η_p^2) is a measure of effect size [18].

	df_n	df_d	F	p	η_p^2
SRU	1	16.98	36.94	< .001	0.69
VT	2	275.34	6.14	< .05	0.04
SRU × *VT*	2	275.15	2.46	.087	0.02
Age	1	16.95	8.21	< .05	0.33

in an evaluation prior to our enhancements [68], non-SRUs *did* outperform screen-reader users by 62%. Therefore, such a non-significant result is noteworthy because it "closed the gap."

There was a significant effect of *Visualization Type* (*VT*) on *AEI* overall (χ^2(2, *N*=300)=9.35, p<.05, Cramer's V=0.12). This result indicates that *AEI* differs significantly between different visualization types. Figure 3 show the *AEI* percentages across different *VT*. For both SRUs and non-SRUs, *Single-Series Bar Graphs* had the best performance (92% and 98% *AEI*, respectively), followed by *Multi-Series Line Graphs* (88% and 84% *AEI*, respectively) and *Geographic Maps* (86% and 70% *AEI*, respectively). Additionally, we investigated the effects of *Age* and the *SRU* × *VT* interaction, but we did not find significant effects on *AEI* (see Table 1).

5.2.2 Interaction Time (IT). Interaction times were initially conditionally lognormal, a common occurrence with temporal measures [50]. Therefore, we applied a logarithmic transformation before conducting our analysis, following standard practice for time measures [5, 38, 50]. Anderson-Darling goodness-of-fit tests of normality [3] confirmed that log-transformed interaction times were conditionally normally distributed ($p \approx .117$). For ease of understanding, we display plots of *IT* using the non-transformed values.

Screen-Reader User (*SRU*) had a significant main effect on *Interaction Time* (IT) (F(1,18)=36.94, p<.001, η_p^2=0.69). Specifically, the average *IT* for SRUs was 42.4 seconds (*SD*=21.2). For non-SRUs, it was 22.0 seconds (*SD*=16.5). The average *IT* for SRUs was 92.8% higher than non-SRUs. We also found a significant main effect of *Visualization Type (VT)* (F(2,275.3)=6.14, p<.05, η_p^2=0.04) on *IT*. These results indicate that *IT* significantly differed among visualization types (*VT*). We also examined the interaction between *SRU* × *VT*, but did not find a significant effect (see Figure 4 and Table 2). For SRUs, *Geospatial Map* had the maximum interaction time; for non-SRUs, it was the *Multi-Series Line Graph*. *Single-Series Bar Graph* had the minimum interaction time for both groups.

Age had a significant effect on *IT* (F(1,16.9)=8.21, p<.05, η_p^2 =0.33), indicating that *IT* differed significantly across the ages of our participants. Participants over 50 years old had 44.0% higher interaction times than those under the age of 50.

5.3 Qualitative Results

Through follow-up interviews with all of our SRUs (*N*=10), we assessed the usability and usefulness of our enhancements. Specifically, we asked them about the features they liked and for any areas of improvement. Additionally, we observed their interactions to identify system errors and user workarounds.

5.3.1 Liked Features. Nine out of 10 SRUs identified the "interactive dialogue" feature as one of the features they liked. Eight participants appreciated that our enhancements were "intuitive" and "easy to use." Three participants highlighted the adaptiveness of our enhancements, stating that VoxLens "adjusts to your question," and that it is "suitable for people of all ages and backgrounds." Two participants found our enhancements innovative, something they had "never seen before." One participant liked the quick responses.

5.3.2 Areas of Improvement. Our participants recognized five areas of improvement: (1) *adding a repeat command* (would enable users to re-hear the response from the previous query); (2) *building a "playground" environment* (would enable users to learn more about the tool by trying out different commands and features); (3) *making the response more succinct*; (4) *enabling responses to be explored in text form* (would append the auditory response as text on the web page, enabling them to copy it); and (5) *increasing the query input time* (would enable users to issue longer queries without feeling rushed). Based on these findings, we have started the development work needed to improve VoxLens even further.

5.3.3 System Errors. From our analysis of participants' interaction logs, we recognized system errors stemming from the limitations of the keyword matching algorithm and voice recognition—two fundamental components of VoxLens. Our system does not process the user's query when our algorithm does not find a keyword match—an unfortunate and known limitation of voice assistants [16, 59, 73]. For example, "tell me the three countries doing amazing in vaccinating people" was not recognized by our system but "tell me the top three countries by vaccination percentages" was correctly processed. Future work can utilize Natural Language Processing models trained to handle such nuanced scenarios.

5.3.4 User Workarounds. Our participants employed workarounds to extract information due to the limitations mentioned above. Specifically, when our system could not process a user's query involving a comparison between data points, our participants asked for the value of each data point separately and computed the comparison mentally. Although the users successfully extracted the information, they issued multiple commands instead of a single command, consequently increasing their interaction time. We aim to address these areas of improvement and system errors to reduce the number of workarounds.

5.4 Subjective Workload Ratings

We collected subjective workload ratings for both user groups using the NASA Task Load Index questionnaire (NASA-TLX) [34]. The NASA-TLX questionnaire records users' perceived task workload on six scales: *mental demand, physical demand, temporal demand, performance, effort,* and *frustration*. For all scales, lower is better, as it corresponds to lesser perceived workload. We performed the nonparametric Aligned Rank Transform procedure [23, 79] to statistically analyze the effects of *SRU* (levels: *yes, no*). We did not find a significant effect of *SRU* on any of the six dimensions, suggesting comparable workload levels.

6 DISCUSSION

We generated taxonomies of the information sought by SRUs for their holistic and drilled-down explorations of online data visualizations using the findings from our role-playing and longitudinal user studies. Utilizing our taxonomies, we extended the capabilities of VoxLens [70], enabling granular information extraction from simple and complex data visualizations. We assessed our enhancements through a task-based user study. Our findings show that using our VoxLens enhancements, our SRUs were 5.6% *more* accurate, on average, than our non-SRUs. Furthermore, our enhancements improved their interaction time by 6.3%.

6.1 Substantial Improvement in Accuracy

Sharif *et al.* [68] reported that SRUs are 62% less accurate than non-SRUs at extracting information from online data visualizations. The original VoxLens [70] reduced this gap, resulting in SRUs extracting information only 15% less accurately than non-SRUs. With our enhancements, SRUs extracted information 6% *more* accurately than non-SRUs, constituting a 164% improvement over SRUs who did not use VoxLens, and a 19% improvement over SRUs who used the original VoxLens. These findings emphasize that tools such as VoxLens that empower SRUs in extracting information granularly and cater to their individual needs can help reduce the disparity caused by inaccessible visualizations between the two user groups.

6.2 Continuous Reduction of Interaction Time

Prior work has reported that the average interaction time for SRUs (without VoxLens) was 84.6 seconds [68]. VoxLens [70] improved these interaction times by 36%, reducing the average time to 54.1 seconds. Now with our VoxLens enhancements, the average interaction time for participants is 42.4 seconds, a 50% and 22% improvement, respectively. Although our enhancements show an improvement in interaction times compared to VoxLens, our findings show that SRUs spent 93% more time interacting with the visualizations than non-SRUs, accentuating the disparity between the two user groups' interaction times. These findings can motivate accessible data visualization solutions that use traditional methods, such as keyboard-based navigation, to incorporate the consideration of interaction times in their design. Several factors contribute to the difference in interaction times between SRUs and non-SRUs [7, 70], including the duration of the auditory responses. The findings from our follow-up interviews identify features that could further reduce these interaction times (e.g., a "repeat" command). The implications of these findings can help guide existing and future voice assistants for SRUs to improve information extraction.

6.3 A Future of Personalized Designs

A recurring observation in our studies was that each SRU exhibited a distinct way of interacting with data visualizations. Since VoxLens only identified the most common keywords, it underperformed processing variations in the queries issued by our participants, forcing them to employ workarounds. Although participants successfully extracted the information using workarounds (accuracy was not affected), their performance resulted in higher interaction times. Therefore, we recommend using personalized designs [60, 66] by identifying usage patterns and individualized preferences of users to improve overall performance.

6.4 Recommendations for Researchers

In addition to providing generalizable knowledge for visualization creators to improve the accessibility of visualizations, our taxonomies highlight the variations in SRUs' drilled-down interactions across different visualization and data types. Therefore, we recommend conducting studies with SRUs to understand their interactions with other visualization and data types, such as 3-D data visualizations. Additionally, we suggest researchers utilize our taxonomies to construct alternative textual descriptions for image-based visualizations to provide SRUs with relevant information.

7 LIMITATIONS & FUTURE WORK

Our exploration was limited to single-series bar graphs, multi-series line graphs, and geospatial maps. Future work can utilize our methodology to understand and improve SRUs' experiences with other complex data visualizations such as three-dimensional graphs and stacked bar charts. Additionally, future work can employ our methodology to investigate the interaction of SRUs with physical interfaces such as tactile maps. Furthermore, our enhancements were limited in parsing users' input commands due to restrictions from the keyword-matching algorithm and the Web Speech API's voice recognition. Future work can employ advanced Natural Language Processing and Conversational Question and Answering algorithms to handle complex and nuanced input queries.

8 CONCLUSION

In this work, we conducted role-playing and longitudinal user studies with SRUs to understand their holistic and drilled-down information extraction needs from simple and complex online data visualizations, including multi-series line graphs and geospatial maps. We used our findings to generate taxonomies of the information sought by SRUs in their interactions with online data visualizations. Then, utilizing these taxonomies, we enhanced the capabilities of VoxLens to enable them to extract information from complex data visualizations in a granular fashion. We assessed the performance of our enhancements using a mixed-methods approach through a task-based user study with SRUs and non-SRUs. Our enhancements improved the accuracy of extracted information and interaction times of SRUs compared to the original VoxLens. Additionally, using our enhancements, SRUs "closed the gap" compared to non-SRUs in the accuracy of the information they extracted from online data visualizations. Closing this gap such that the accuracy of extracted information from online data visualizations is not detectably different between SRUs and non-SRUs represents a major advancement in the accessibility of online data visualizations.

ACKNOWLEDGMENTS

This work was supported in part by the University of Washington Center for Research and Education on Accessible Technology and Experiences (CREATE). We thank and remember our recently departed team member, Zoey, for her feline support, without which the *purr*suit of this work would not have been possible. May she cross the rainbow bridge in peace and find her way to cat heaven.

REFERENCES

[1] Cengiz Acarturk and C. Habel. 2012. Eye tracking in multimodal comprehension of graphs. *CEUR Workshop Proceedings* 887 (01 2012), 11–25.

[2] Dragan Ahmetovic, Niccolò Cantù, Cristian Bernareggi, João Guerreiro, Sergio Mascetti, and Anna Capietto. 2019. Multimodal Exploration of Mathematical Function Graphs with AudioFunctions.Web. In *Proceedings of the 16th International Web for All Conference* (San Francisco, CA, USA) *(W4A '19)*. Association for Computing Machinery, New York, NY, USA, Article 8, 2 pages. https://doi.org/10.1145/3315002.3332438

[3] Theodore W. Anderson and Donald A. Darling. 1954. A test of goodness of fit. *Journal of the American statistical association* 49, 268 (1954), 765–769.

[4] Apple. n.d.. Audio Graphs | Apple Developer Documentation. https://developer.apple.com/documentation/accessibility/audio_graphs. (Accessed on 08/01/2021).

[5] Donald A Berry. 1987. Logarithmic transformations in ANOVA. *Biometrics* 43, 2 (1987), 439–456.

[6] Vittoria Biagi, Riccardo Patriarca, and Giulio Di Gravio. 2022. Business Intelligence for IT Governance of a Technology Company. *Data* 7, 1 (2022), 2.

[7] Jeffrey P. Bigham, Anna C. Cavender, Jeremy T. Brudvik, Jacob O. Wobbrock, and Richard E. Ladner. 2007. WebinSitu: A Comparative Analysis of Blind and Sighted Browsing Behavior. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility* (Tempe, Arizona, USA) *(Assets '07)*. Association for Computing Machinery, New York, NY, USA, 51–58. https://doi.org/10.1145/1296843.1296854

[8] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2301–2309.

[9] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[10] Matthew Brehmer, Bongshin Lee, Petra Isenberg, and Eun Kyoung Choe. 2018. Visualizing ranges over time on mobile phones: a task-based crowdsourced evaluation. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 619–629.

[11] Matthew Brehmer and Tamara Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2376–2385.

[12] Craig Brown and Amy Hurst. 2012. VizTouch: Automatically Generated Tactile Visualizations of Coordinate Spaces. In *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction* (Kingston, Ontario, Canada) *(TEI '12)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/2148131.2148160

[13] Lorna M. Brown, Stephen A. Brewster, Ramesh Ramloll, Mike Burton, and Beate Riedel. 2003. Design guidelines for audio presentation of graphs and tables. In *Proceedings of the 9th International Conference on Auditory Display*. Citeseer, Boston University, USA, 284–287.

[14] Marion Buchenau and Jane Fulton Suri. 2000. Experience Prototyping. In *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (New York City, New York, USA) *(DIS '00)*. Association for Computing Machinery, New York, NY, USA, 424–433. https://doi.org/10.1145/347642.347802

[15] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. 2008. Web content accessibility guidelines (WCAG) 2.0.

[16] Julia Cambre, Ying Liu, Rebecca E Taylor, and Chinmay Kulkarni. 2019. Vitro: Designing a Voice Assistant for the Scientific Lab Workplace. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. Association for Computing Machinery, New York, NY, USA, 1531–1542.

[17] Dustin Carroll, Suranjan Chakraborty, and Jonathan Lazar. 2013. Designing accessible visualizations: The case of designing a weather map for blind users. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 436–445.

[18] Jacob Cohen. 1973. Eta-squared and partial eta-squared in fixed factor ANOVA designs. *Educational and psychological measurement* 33, 1 (1973), 107–112.

[19] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—why and how. *Knowledge-based systems* 6, 4 (1993), 258–266.

[20] Joel J Davis. 2002. Disenfranchising the Disabled: The Inaccessibility of Internet-Based Health Information. *Journal of Health Communication* 7, 4 (2002), 355–367. https://doi.org/10.1080/10810730290001701

[21] Google Developers. 2014. Charts. https://developers.google.com/chart/

[22] Frank Elavsky, Cynthia Bennett, and Dominik Moritz. 2022. How accessible is my visualization? Evaluating visualization accessibility with Chartability. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 57–70.

[23] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 754–768. https://doi.org/10.1145/3472749.3474784

[24] Xuanhe Er and Yunqi Sun. 2021. Visualization Analysis of Stock Data and Intelligent Time Series Stock Price Prediction Based on Extreme Gradient Boosting. In *2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 272–279.

[25] Christopher J. Ferguson. 2016. An effect size primer: A guide for clinicians and researchers. In *Methodological issues and strategies in clinical research*, A.E. Kazdin (Ed.). American Psychological Association, Washington, DC, USA, 301––310.

[26] John H. Flowers, Dion C. Buhman, and Kimberly D. Turnage. 1997. Cross-modal equivalence of visual and auditory scatterplots for exploring bivariate data samples. *Human Factors* 39, 3 (1997), 341–351.

[27] Brigitte N. Frederick. 1999. Fixed-, random-, and mixed-effects ANOVA models: A user-friendly guide for increasing the generalizability of ANOVA results. In *Advances in Social Science Methodology*, B. Thompson (Ed.). JAI Press, Stamford, Connecticut, 111–122.

[28] Andreas Gegenfurtner, Erno Lehtinen, and Roger Säljö. 2011. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational psychology review* 23, 4 (2011), 523–552.

[29] Arthur Gilmour, Robert D. Anderson, and Alexander L. Rae. 1985. The analysis of binomial data by a generalized linear mixed model. *Biometrika* 72, 3 (1985), 593–599.

[30] Nicholas A. Giudice, Hari Prasath Palani, Eric Brenner, and Kevin M. Kramer. 2012. Learning Non-Visual Graphical Information Using a Touch-Based Vibro-Audio Interface. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility* (Boulder, Colorado, USA) *(ASSETS '12)*. Association for Computing Machinery, New York, NY, USA, 103–110. https://doi.org/10.1145/2384916.2384935

[31] Matthew Graham, Anthony Milanowski, and Jackson Miller. 2012. Measuring and Promoting Inter-Rater Agreement of Teacher and Principal Performance Ratings.

[32] Sylvie Grosjean, Jean-Luc Ciocca, Amélie Gauthier-Beaupré, Emely Poitras, David Grimes, and Tiago Mestre. 2022. Co-designing a digital companion with people living with Parkinson's to support self-care in a personalized way: The eCARE-PD Study. *DIGITAL HEALTH* 8 (2022), 20552076221081695.

[33] Melita Hajdinjak and France Mihelic. 2004. Conducting the Wizard-of-Oz Experiment. *Informatica (Slovenia)* 28, 4 (2004), 425–429.

[34] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, North-Holland, Netherlands, 139–183.

[35] Donald P Hartmann. 1977. Considerations in the choice of interobserver reliability estimates. *Journal of applied behavior analysis* 10, 1 (1977), 103–116.

[36] Highcharts. n.d.. Sonification | Highcharts. https://www.highcharts.com/docs/accessibility/sonification. (Accessed on 08/01/2021).

[37] Jake Holland. 2017. New York Times' Upshot editor discusses data visualization, storytelling. https://dailynorthwestern.com/2017/05/03/campus/new-york-times-upshot-editor-discusses-data-visualization-storytelling/. (Accessed on 03/05/2022).

[38] MH Hoyle. 1973. Transformations: An introduction and a bibliography. *International Statistical Review/Revue Internationale de Statistique* 41, 2 (1973), 203–223.

[39] Todd Hunt. 1982. Raising the Issue of Ethics through Use of Scenarios. *The Journalism Educator* 37, 1 (1982), 55–58.

[40] Amy Hurst. 2018. Making "Making" Accessible. In *Proceedings of the 15th International Web for All Conference* (Lyon, France) *(W4A '18)*. Association for Computing Machinery, New York, NY, USA, Article 1, 1 pages. https://doi.org/10.1145/3192714.3192715

[41] Facebook Inc. n.d.. React – A JavaScript library for building user interfaces. https://reactjs.org/. (Accessed on 08/08/2021).

[42] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering Questions about Charts and Generating Visual Explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376467

[43] Edward Kim and Kathleen F McCoy. 2018. Multimodal deep learning using images and text for information graphic classification. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 143–148.

[44] Edward Kim, Connor Onweller, and Kathleen F McCoy. 2021. Information Graphic Summarization using a Collection of Multimodal Deep Neural Networks. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 10188–10195.

[45] Mario Konecki, Charles LaPierre, and Keith Jervis. 2018. Accessible data visualization in higher education. In *2018 41st international convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 0733–0737.

[46] Michal Kubovics and Pavel Bielik. 2021. Visualization of Data And Keywords in Online Journalism. *Marketing Identity* 9, 1 (2021), 121–133.

[47] Bongshin Lee, Arjun Srinivasan, Petra Isenberg, John Stasko, et al. 2021. Post-WIMP Interaction for Information Visualization. *Foundations and Trends® in Human-Computer Interaction* 14, 1 (2021), 1–95.

[48] Michael Lewis-Beck, Alan E Bryman, and Tim Futing Liao. 2003. *The Sage encyclopedia of social science research methods*. Sage Publications, Thousand Oaks, California.

[49] Peng Liang and Onno De Graaf. 2010. Experiences of using role playing andwiki in requirements engineering course projects. In *2010 5th International Workshop on Requirements Engineering Education and Training*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 1–6.

[50] Eckhard Limpert, Werner A Stahel, and Markus Abbt. 2001. Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: that is the question. *BioScience* 51, 5 (2001), 341–352.

[51] Ramon C. Littell, Henry P. Raymond, and Clarence B. Ammerman. 1998. Statistical analysis of repeated measures data using SAS procedures. *Journal of animal science* 76, 4 (1998), 1216–1231.

[52] Alan Lundgard, Crystal Lee, and Arvind Satyanarayan. 2019. Sociotechnical considerations for accessible visualization design. In *2019 IEEE Visualization Conference (VIS)*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 16–20.

[53] Alan Lundgard and Arvind Satyanarayan. 2021. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE transactions on visualization and computer graphics* 28, 1 (2021), 1073–1083.

[54] Moira Maguire and Brid Delahunt. 2017. Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars. *All Ireland Journal of Higher Education* 9, 3 (2017), 3351–3364.

[55] Kim Marriott, Bongshin Lee, Matthew Butler, Ed Cutrell, Kirsten Ellis, Cagatay Goncu, Marti Hearst, Kathleen McCoy, and Danielle Albers Szafir. 2021. Inclusive data visualization for people with disabilities: a call to action. *Interactions* 28, 3 (2021), 47–51.

[56] Randall B Martin. 1991. The assessment of involvement in role playing. *Journal of clinical psychology* 47, 4 (1991), 587–596.

[57] David K. McGookin and Stephen A. Brewster. 2006. SoundBar: Exploiting Multiple Views in Multimodal Graph Browsing. In *Proceedings of the 4th Nordic Conference on Human-Computer Interaction: Changing Roles* (Oslo, Norway) *(NordiCHI '06)*. Association for Computing Machinery, New York, NY, USA, 145–154. https://doi.org/10.1145/1182475.1182491

[58] Silvia Mirri, Silvio Peroni, Paola Salomoni, Fabio Vitali, and Vincenzo Rubano. 2017. Towards accessible graphs in HTML-based scientific articles. In *2017 14th IEEE Annual Consumer Communications Networking Conference (CCNC)*. IEEE, Las Vegas, NV, USA, 1067–1072. https://doi.org/10.1109/CCNC.2017.7983287

[59] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for how users overcome obstacles in voice user interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–7.

[60] Esther Nathanson. 2017. Native voice, self-concept and the moral case for personalized voice technology. *Disability and rehabilitation* 39, 1 (2017), 73–81.

[61] National Geographic Society. n.d.. United States Regions | National Geographic Society. https://www.nationalgeographic.org/maps/united-states-regions/. (Accessed on 03/29/2022).

[62] Manuel M Oliveira. 2013. Towards more accessible visualizations for color-vision-deficient individuals. *Computing in Science & Engineering* 15, 5 (2013), 80–87.

[63] Rachael Rickta Patrick and Syahrul Nizam Junaini. 2021. Bibliometric Visualisation of Computer Science and COVID-19: A Review and Proposed Method. In *2021 IEEE 19th Student Conference on Research and Development (SCOReD)*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 13–18.

[64] Michael Quinn Patton. 1990. *Qualitative evaluation and research methods*. SAGE Publications, Inc., Thousand Oaks, CA, USA.

[65] Helen Petrie, Fraser Hamilton, Neil King, and Pete Pavan. 2006. Remote usability evaluations with disabled people. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1133–1141.

[66] Silvia Quarteroni and Suresh Manandhar. 2007. User modelling for personalized question answering. In *Congress of the Italian Association for Artificial Intelligence*. Springer, Springer Berlin Heidelberg, Berlin, Heidelberg, 386–397.

[67] Freedom Scientific. n.d.. JAWS® – Freedom Scientific. https://www.freedomscientific.com/products/software/jaws/. (Accessed on 08/08/2021).

[68] Ather Sharif, Sanjana Shivani Chintalapati, Jacob O. Wobbrock, and Katharina Reinecke. 2021. Understanding Screen-Reader Users' Experiences with Online Data Visualizations. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 14, 16 pages. https://doi.org/10.1145/3441852.3471202

[69] Ather Sharif and Babak Forouraghi. 2018. evoGraphs — A jQuery plugin to create web accessible graphs. In *2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC)*. IEEE, Las Vegas, NV, USA, 1–4. https://doi.org/10.1109/CCNC.2018.8319239

[70] Ather Sharif, Olivia H. Wang, Alida T. Muongchan, Katharina Reinecke, and Jacob O. Wobbrock. 2022. VoxLens: Making Online Data Visualizations Accessible with an Interactive JavaScript Plug-In. In *CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 478, 19 pages. https://doi.org/10.1145/3491102.3517431

[71] Ather Sharif, Andrew M Zhang, Anna Shih, Jacob O Wobbrock, and Katharina Reinecke. 2022. Understanding and Improving Information Extraction From Online Geospatial Data Visualizations for Screen-Reader Users. In *The 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, To appear pages.

[72] Lei Shi, Idan Zelzer, Catherine Feng, and Shiri Azenkot. 2016. Tickers and Talker: An Accessible Labeling Toolkit for 3D Printed Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 4896–4907. https://doi.org/10.1145/2858036.2858507

[73] Aaron Springer and Henriette Cramer. 2018. " Play PRBLMS" Identifying and Correcting Less Accessible Content in Voice Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[74] Adam Strantz. 2021. Using Web Standards to Design Accessible Data Visualizations in Professional Communication. *IEEE Transactions on Professional Communication* 64, 3 (2021), 288–301.

[75] Dag Svanaes and Gry Seland. 2004. Putting the Users Center Stage: Role Playing and Low-Fi Prototyping Enable End Users to Design Mobile Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vienna, Austria) *(CHI '04)*. Association for Computing Machinery, New York, NY, USA, 479–486. https://doi.org/10.1145/985692.985753

[76] Suppawong Tuarob, Poom Wettayakorn, Ponpat Phetchai, Siripong Traivijitkhun, Sunghoon Lim, Thanapon Noraset, and Tipajin Thaipisutikul. 2021. DAViS: a unified solution for data collection, analyzation, and visualization in real-time stock market prediction. *Financial Innovation* 7, 1 (2021), 1–32.

[77] Michael B Twidale. 2005. Over the shoulder learning: supporting brief informal learning. *Computer supported cooperative work (CSCW)* 14, 6 (2005), 505–547.

[78] Frances Van Scoy, Don McLaughlin, and Angela Fullmer. 2005. Auditory augmentation of haptic graphs: Developing a graphic tool for teaching precalculus skill to blind students.

[79] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 143–146.

[80] Susan P Wyche and Rebecca E Grinter. 2009. Extraordinary computing: religion as a lens for reconsidering the home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 749–758.

[81] Wai Yu, Rameshsharma Ramloll, and Stephen A. Brewster. 2000. Haptic Graphs for Blind Computer Users. In *Proceedings of the First International Workshop on Haptic Human-Computer Interaction*. Springer-Verlag, Berlin, Heidelberg, 41–51.

[82] Lidong Zhang, B Vinodhini, and T Maragatham. 2021. Interactive IoT Data Visualization for Decision Making in Business Intelligence. *Arabian Journal for Science and Engineering* Special Issue (2021), 1–11.

[83] Haixia Zhao, Catherine Plaisant, Ben Shneiderman, and Jonathan Lazar. 2008. Data sonification for users with visual impairment: a case study with georeferenced data. *ACM Transactions on Computer-Human Interaction (TOCHI)* 15, 1 (2008), 1–28.

[84] Luming Zhao and WeiMing Ye. 2022. Visualization as infrastructure: China's data visualization politics during COVID-19 and their implications for public health emergencies. *Convergence* 28, 1 (2022), 13548565211069872.

[85] Shuai Zheng, Jonathan R Edwards, Margaret A Dudeck, Prachi R Patel, Lauren Wattenmaker, Muzna Mirza, Sheri Chernetsky Tejedor, Kent Lemoine, Andrea L Benin, Daniel A Pollock, et al. 2021. Building an Interactive Geospatial Visualization Application for National Health Care–Associated Infection Surveillance: Development Study. *JMIR public health and surveillance* 7, 7 (2021), e23528.

[86] Jonathan Zong, Crystal Lee, Alan Lundgard, JiWoong Jang, Daniel Hajas, and Arvind Satyanarayan. 2022. Rich Screen Reader Experiences for Accessible Data Visualization. *Computer Graphics Forum* 41, 3 (2022), 15–27. https://doi.org/10.1111/cgf.14519 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14519

A TAXONOMY OF HOLISTIC INFORMATION

Table 3: Taxonomy of the holistic information screen-reader users seek when exploring online data visualizations. Information types within each category are in descending order based on their sought frequency. For each information type, the "Query" column shows some of the questions that our participants asked to extract that information. Each information type was applicable to all the visualization types used in our studies (single-series bar graphs, multi-series line graphs, and geospatial maps).

Category	Information Type	Query
Summary Statistics	Extremum	*What country has the highest number?*
		What month was the stock market doing the best?
		Which country had the lowest overall in 2020?
	Average	*What is the national average?*
		What is the rough average of data in North America?
		What is the average of the top ten countries?
	Axis Ranges	*Can you tell me what is the x-axis?*
		What is the y-axis?
		What month does it start at and what month does it end?
	Factor Levels	*What countries are included in this graph?*
		How many companies are in the graph?
		What are the states in this graph?
	Median	*What's the median?*
		Tell me the median score
		What is the exact median?
	Sum	*What is the total for each year?*
		Can you tell me the total amount for 2020?
		How about the sum of all states?
Understanding Trends	Overall Trend	*How many companies have gone down in price?*
		Could I have some type of sonification of the graph?
		Is the United States currently going up?
	Best-Fit Line	*What is the best fit line for the United States?*
		Which line of best fit has the highest slope and lowest slope?
	Correlation Coefficient	*Can you tell me the correlation coefficient?*

B TAXONOMY OF DRILLED-DOWN INFORMATION

Table 4: Taxonomy of the drilled-down information screen-reader users seek when exploring online data visualizations to extract and compare data points. Information types within each category are in descending order based on their sought frequency. For each information type, the "Query" column shows some of the questions that our participants asked to extract that information. Each information type under the *Ranking* category was applicable to all the visualization types used in our studies (single-series bar graphs, multi-series line graphs, and geospatial maps). Under the *Categorization* category, *Regional, Political, Climate-Related, Population-Related,* and *Spoken-Language-Related* were applicable to geospatial maps, whereas *Factor-Levels-Related* was only applicable to multi-series line graphs.

Category	Information Type	Query
Categorization	Regional (Geospatial)	*Is there a difference between the east and west, or the north and south?*
		Is there a continent that has a higher life expectancy?
		Tell me the values of the Southern states as opposed to the Northwest.
	Factor-Levels-Related (Multi-series)	*How is California doing now, 10 years ago, and 20 years ago?*
		How was Texas between 2010 and 2015?
		Can we filter out the data to only Apple and Walgreens?
	Political (Geospatial)	*How do the trends during the Democratic presidential campaigns compare to the trends during Republican presidential campaigns?*
		Do socialist countries have higher rates?
		When Republicans were in power, did they increase compared to when Democrats were in power?
	Climate-Related (Geospatial)	*Do warmer climate states have higher values?*
		How are colder places compared to warmer places?
	Population-Related (Geospatial)	*Can you compare these two states by population?*
		Do the states with larger population have higher traffic rates?
	Spoken-Language-Related (Geospatial)	*Can we compare Spanish speaking countries to the English speaking countries?*
Ranking	Top	*Which countries are in the top 5%?*
		What are the five top countries in Western Europe?
		I'd like to see them all in order from the highest to the lowest.
	Bottom	*What are the bottom 10 companies in the graph?*
		What about the second and third lowest?
		Can we organize from lowest to the highest?
	Surrounding Average	*Which ones are in the middle?*
		What are the three countries that are closest to the average?
		What's the distribution, the countries around the average.