

Exploring Chart Question Answering for Blind and Low Vision Users

Jiho Kim

kim999@wisc.edu

University of Wisconsin-Madison

Madison, USA

Nam Wook Kim

nam.wook.kim@bc.edu

Boston College

Chestnut Hill, USA

Arjun Srinivasan

arjunsrinivasan@tableau.com

Tableau Research

Seattle, USA

Yea-Seul Kim

yeaseul.kim@cs.wisc.edu

University of Wisconsin-Madison

Madison, USA

ABSTRACT

Data visualizations can be complex or involve numerous data points, making them impractical to navigate using screen readers alone. Question answering (QA) systems have the potential to support visualization interpretation and exploration without overwhelming blind and low vision (BLV) users. To investigate if and how QA systems can help BLV users in working with visualizations, we conducted a Wizard of Oz study with 24 BLV people where participants freely posed queries about four visualizations. We collected 979 queries and mapped them to popular analytic task taxonomies. We found that retrieving value and finding extremum were the most common tasks, participants often made complex queries and used visual references, and the data topic notably influenced the queries. We compile a list of design considerations for accessible chart QA systems and make our question corpus publicly available to guide future research and development.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in accessibility**; **Empirical studies in visualization**.

KEYWORDS

Accessibility, Visualization, Question Answering, Human-Subjects Qualitative Studies, Design Considerations

ACM Reference Format:

Jiho Kim, Arjun Srinivasan, Nam Wook Kim, and Yea-Seul Kim. 2023. Exploring Chart Question Answering for Blind and Low Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 15 pages. <https://doi.org/10.1145/3544548.3581532>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CHI '23, April 23–28, 2023, Hamburg, Germany

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9421-5/23/04...\$15.00

<https://doi.org/10.1145/3544548.3581532>

1 INTRODUCTION

Digital data visualizations are becoming increasingly common as an effective way to communicate information by allowing users to explore complex data [43]. However, their efficacy presupposes the powerful capabilities of human visual perception to process information. Visualizations can therefore disenfranchise people with *blind and low vision (BLV)*¹ unless they can be accessed through a non-visual modality [49].

Many assistive technologies have been developed to enable BLV users to access visualizations by leveraging sensory modalities beyond vision, such as sound, texture, or text (e.g., [21, 64, 89]). However, some modalities are more limited than others when it comes to their generalizability and their likelihood of being adopted outside of controlled environments. Tactile visualizations, for example, require specialized hardware such as a haptic display and embossing machine, which can be expensive or not widely available (e.g., [12, 28, 78]). Tactile perception has a steep learning curve for interpreting the signal [19, 27]. Audio channels are unable to present multiple sound sources to represent multiple data at the same time due to the limitation of auditory perception [55]. Furthermore, auditory perception in decoding data is error-prone and varies from person to person (e.g., [63, 77, 80]).

One promising alternative style of interaction that can help BLV users analyze and explore data presented in visualizations is *question answering (QA)*. QA systems allow users to formulate their queries using natural utterances without needing to interact with interface elements. Moreover, QA systems can support user agency compared to tools that try to bring accessibility with alternative text. QA systems are proposed and deployed in a variety of fields, including medical diagnosis [1], online education [82], law [17], and open-domain queries [88]. Within the visualization community too, several chart QA systems have been implemented [36, 39, 53, 64] to help users understand visualizations and the underlying data.

However, while few chart QA systems mention BLV people as possible audience, most consider accessibility only as a byproduct. To develop a comprehensive chart QA system that can address most BLV users' needs, developers must be mindful of how and why BLV people interact with such a system. The purpose of this work is to provide future developers with a concrete set of design

¹We use both *people first language* (people who are BLV) and *identity first language* (BLV people) depending on the grammar of a sentence, and in recognition that some people want their visual impairment acknowledged as an essential identifier and others do not. We also use BLV people and BLV users interchangeably.

considerations for creating QA systems intended for BLV users. More specifically, we explore the following research questions:

- **RQ1:** When, where, and why do BLV people want to use a chart QA system?
- **RQ2:** What kinds of queries do BLV people make and what factors influence their queries?
- **RQ3:** Can current QA systems for sighted people support BLV people's queries? If not, what design aspects must future accessible chart QA systems consider?

We recruited 24 BLV people and conducted a Wizard of Oz study where a researcher pretended to be a QA system. We collected a total of 979 queries and analyzed them through the lens of popular taxonomies of visualization analysis tasks. We found that 73% of the queries were data-related, whereas 27% were about the context, topic, or graphical elements of the visualization. About half of the queries focused on the tasks of *finding extrema*, *retrieving value*, or *computing derived value after filtering by attribute*. Providing a data table alongside the visualization did not influence the types of queries participants asked. To contextualize our observations, we compared the queries collected from our study to those compiled by sighted individuals from an existing study [39]. We found that participants in our study asked more complex queries. We also analyzed instances where the existing chart QA system failed to respond to queries from our study. Based on these analyses and findings, we derive design considerations for the future development of chart QA systems for BLV users.

In summary, our contributions are three-fold.

- We conduct a Wizard of Oz study with BLV people to understand if and how QA systems can help their visualization comprehension and report the observations from the study illustrating their needs and preferences of QA systems.
- We characterize the queries that BLV people ask to a chart QA system to elicit better design considerations for future systems.
- We release a collection of queries asked by BLV people to inform future research.

2 BACKGROUND & RELATED WORK

2.1 Data Visualization Accessibility

With the growing adoption of data visualizations across disciplines, addressing diverse audiences' needs has become a vital concern [42]. Among those audiences, supporting people with disabilities is a critical societal issue [49]. While the higher information processing bandwidth of vision is what makes data visualizations effective, it conversely puts a strong barrier for BLV people [49]. The inability to access information can adversely impact education [24] and employment opportunities [84] as well as people's decision-making on finances, health, and other everyday activities [70].

To confront this problem, prior work has investigated how to use non-visual modalities such as speech, sound, and texture [40]. Tactile and haptic systems provide a simultaneous and on-demand exploration of data trends but require additional motor movement, and can be difficult to perceive [27]. Sonification promptly conveys data by the dimensions of sound such as pitch and volume [80], but data details can be lost in translation. Multi-modal systems (e.g., [9]) can overcome the limitations of a single modality, but

are costly and may not be affordable for most BLV people. Due to these reasons, screen readers which use text/speech modality are the most common assistive technology, especially for browsing web-based content [41, 81].

A standard approach to making a chart accessible is to describe it with what is called an *alternative text (alt text)*. For images on the web, it is a popular method that has been empirically studied with BLV users [72]. There are in-depth guidelines on how to write effective text for communicating essential insights of a data visualization, including the overall message, visual structure, and data trends [7, 33, 48, 61]. The alt text approach has been used for decades and works reasonably well for simple charts. However, it is confined to a short text and simple conclusion [83], and faced with additional challenges due to the new advancement in data visualizations that make them more intricate and more interactive [32]. Providing long descriptions and data tables are often suggested to tackle these challenges, but they deprive data visualizations of their benefit and cannot address the users' different needs and ways in which they explore visualizations [48, 70].

Practitioners and researchers have both investigated ways to address these new challenges. The HighCharts [30] visualization library for instance, in 2017, began developing accessible chart navigation that provides a robust semantic levels of description, with the technique inspired by MathJax [13]. Visa Chart Components [74] offers a framework agnostic visualization design that grants users access to the raw data. A number of research studies have investigated the development of tools that can parse the underlying data from an image of a visualization [14, 34]. Another set of studies focuses on building systems that support interaction. Some of them can provide users with higher-level information such as the minimum and the maximum data value [26] and the average [14]. Others focus on streamlining the user experience by, for example, designing an efficient navigation strategy and having a rich natural language description [22, 23, 91].

Still, accessing the underlying data does not warrant on-demand data operations such as filtering and sorting, and interactive features can be difficult to learn and time-consuming. Voice-based interfaces have been carefully studied and developed in the context of BLV users, especially as virtual assistants (e.g., [3, 10, 76]). However, only a few of the existing systems hint at the possibility of applying natural language interfaces to data visualizations. For instance, Murillo-Morales & Miesenberger [52], shared a prototype system where the user can ask predefined questions including asking about the mean, extremes, and the range of data. Recently, Sharif et al. [64, 65] adopted a similar approach in building a JavaScript plug-in, VoxLens, whose QA module can answer questions that contain predefined words like "maximum", "minimum", "median", and "mode". In a follow-up study [65], the authors also extended VoxLens to specifically support the querying of geospatial visualizations, including new analytic tasks and question types. We also explore the idea of QA systems to aid BLV users and consider a variety of visualizations that one might encounter as part of online news articles. In doing so, we complement prior work that largely focuses on data exploration and investigate the overlaps and unique challenges that arise when the scope of user questions expands beyond analytic functions to also include more general visualization interpretation and data understanding.

2.2 Chart Question Answering Systems

Question answering has been a long-standing topic of research in the fields of natural language processing and computer vision. Prior work has explored QA systems in the context of images (e.g., [5, 6, 47, 86]), videos (e.g., [44, 45, 87]), databases (e.g., [2, 56, 73, 79, 90]), and more recently, even data visualizations (e.g., [35–37, 39, 50, 66]). Given our focus on visualizations, we briefly expand upon prior work on chart QA systems below.

FigureQA [37] is a corpus of question-answer pairs about basic visualizations (bar charts, line charts, and pie charts). It focuses on questions that can be answered with yes/no responses and tasks such as verifying extreme values (e.g., "Is X the maximum?") and finding intersections (e.g., "Does X intersect Y?"). DVQA [35] expands the idea of visualization QA beyond yes/no questions. Questions that focus on the chart structure (e.g., "How many bars are there?"), data retrieval (e.g., "What is the value of the third bar from the left?"), and reasoning (e.g., "Which item sold the most units in any store?") are included, among others. While both FigureQA and DVQA were instrumental in promoting the development of QA models for data visualizations, they were developed using charts that show synthetically generated data. To help design models that are more applicable to real-world scenarios, Mehani et al. [50] introduced the PlotQA dataset which contains over 28.9 million questions about charts from real-world data and crowdsourced question templates. Kim et al. [39] present a system that builds upon *Sempre* [56], a table QA system, to also support questions that contain references to graphical elements (e.g., x and y axes, length, size). Besides answering questions, their system also generates a brief sentence explaining how it got the answer (e.g., "I looked up what the blue represents by looking at the legend").

While chart QA systems and datasets often advertise visualization accessibility as their potential application, it has never been their main pursuit. They do not explicitly take into account questions posed by BLV users, and often presume the interaction behavior of BLV users would be similar to that of sighted users. Nonetheless, prior work in other areas such as image QA has debunked this presumption by suggesting that the questions and phrasings used by BLV users are notably different from those posed by sighted users [18, 29]. Along these lines, we explore the task of supporting chart QA for BLV users to provide insight into how they interact with such a system. We highlight both the types of questions people ask and how they phrase their questions. Finally, we distill them into design considerations for future QA systems that can support BLV users in interpreting visualizations and querying data through visualizations.

3 WIZARD OF OZ: UNDERSTANDING IF AND HOW BLV USERS INTERACT WITH A CHART QA SYSTEM

We conducted a Wizard of Oz study to understand BLV people's expectations from a chart QA system and identify notable querying patterns. We chose the Wizard of Oz methodology to focus on defining the range of interactions of the user, while offering a realistic setting that induces natural interaction [16, 20].

3.1 Goals of the Study

The study aims to address the following questions that can guide the design of an accessible chart QA system.

- **Understanding motivations:** Understanding motivations is the first step in designing any system. We want to understand the value of QA systems as perceived by BLV people. Do they wish to use QA systems? If so, why? We also want to identify specific circumstances where they would find a QA system useful.
- **Understanding user queries:** Analyzing queries of BLV people can help us understand the target population's needs. They can then inform how we should design language parsers for chart QA systems from an accessibility standpoint. Specifically, we want to address questions such as: To what extent do BLV users ask visualization-related queries vs. data-related queries? Are the characteristics of queries different when the underlying data can be accessed via a table? What analytical tasks do users want to perform with their queries? Do any external and internal factors, such as familiarity with the topic, influence the types and the phrasing of queries?
- **Other design considerations:** We want to identify any other insights that can inform system design. For instance, how should chart QA systems for BLV people be different from current chart QA systems? In what ways can we piggyback on existing systems built for sighted users?

3.2 Participants

We recruited participants through mailing lists hosted by organizations serving BLV people (e.g., the National Federation of the Blind). Our recruitment criteria were that users must be 1) at least 18 years old, 2) legally blind, and 3) must use screen readers daily. We recruited a total of 24 participants who met the criteria. 11 participants identified themselves as female, and 13 identified themselves as male. The average age of the participants was 34 (SD=9). Participants were compensated with a \$25 gift card for their participation. Each session lasted 54 minutes on average (SD=18). Among 24 participants, 21 were blind and 3 had low vision. The detailed participants' information including education level, onset age, and assistive technology use is attached as a supplementary material.

3.3 Study Stimuli

3.3.1 User Stimuli. To emulate a real-world visualization reading scenario, we designed four study stimuli from online news articles. We included stimuli covering a variety of topics and four commonly used chart types (Figure 1). To communicate visualizations to participants, we authored an alt text for each visualization based on prior work prescribing effective alt text for BLV users [33, 48]. We ensured that each alt text conveyed the same type of information regardless of the chart type. While it is not always the case in the wild, accessibility guidelines (e.g., [60, 75]) often encourage authors to add data tables alongside visualizations to give users access to raw data. To observe potential differences in the type of queries formulated with and without data tables, we provide the corresponding data table for selected two of the stimuli, counterbalancing all stimuli presented with the tables. All participants examined all four stimuli in a randomized order.


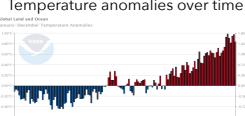
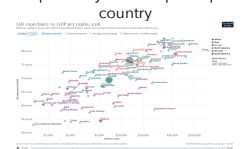
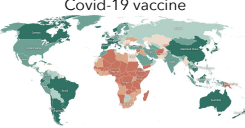
ID	Visualization	Alternative Text
V1	<p>The number of homes for sale nationally in the last 6 years</p>  <p>Line chart (2 quantitative variables)</p>	<p>A line chart depicting the number of homes for sale in the United States. The x axis represents years ranging from 2015 to 2021 in 2 years increments. The y axis represents the number of homes for sale from 250000 to 1.5 million in increments of 250000. The line is overall decreasing. There is a trend each year of the number of homes gradually increasing until the middle of a year, then decreasing at the end of the year. There is a significant decrease between 2020 to 2021. The chart shows that currently, in 2021, there are 468000 homes for sale.</p>
V2	<p>Temperature anomalies over time</p>  <p>Bar chart (2 quantitative, 1 categorical)</p>	<p>A bar chart depicting the global annual temperature anomalies. The x axis represents years from 1880 to 2021. The y axis on the right represents temperatures in Fahrenheit ranging from -1.08 Fahrenheit to 1.80 Fahrenheit. The negative temperature anomalies are represented by blue bars. They are placed on the left side of the bar chart. The overall trend is that from 1880, the temperature anomalies decrease overall to a low of around -0.75 Fahrenheit then increase until 0 Fahrenheit in 1940. The positive temperature anomalies are represented by red bars. The overall trend is that from 1940, the temperature anomalies increase from 0 Fahrenheit until 2021 up to 1.80 Fahrenheit.</p>
V3	<p>Life expectancy vs. GDP per capita by country</p>  <p>Scatterplot (2 quantitative, 1 categorical)</p>	<p>A scatterplot depicting life expectancy versus GDP per capita in 2018. The x axis represents GDP per capita from \$0 to \$100000. The increments are \$1000, \$2000, \$5000, \$10000, \$20000, \$50000, and \$100000. The y axis represents life expectancy at birth from 50 years to 80 years. The increments are by 5 years. Points that represent countries in Africa are colored in purple, Asia in green, Europe in blue, North America in orange, Oceania in brown, and South America in red. The size of a point is relative to the corresponding country's population. Points that represent Africa are in the overall lower range. Points that represent Asia are in the medium to upper range. Points that represent North America are in the medium-upper to upper range. Points that represent Europe are in the upper range. Points that represent South America are in the medium-upper range. Points that represent Oceania are in the upper range.</p>
V4	<p>Population receiving at least one dose of Covid-19 vaccine</p>  <p>Choropleth (1 quantitative, 1 categorical)</p>	<p>A global map in which countries are colored in different shades of green according to the share of the population that have been fully vaccinated with Covid-19 vaccine. 0% is the lightest shade of green, and 100% is the darkest shade of green. Most of North America is represented by the darker shades of green, likely in the 80% to 100% range. In South America, although there are some variations, the majority of the countries are darker shades of green, also in the 80% to 100% range. In Asia, there is more variation with more countries with medium to lighter shades of green in the 20% to 60% range. In Africa, there are mostly light shades of green in the 0% to 20% range. In Europe, there is variation but there are mostly darker shades of green in the 60% to 100% range. Australia is a dark shade of green in the 80% to 100% range.</p>

Figure 1: The study stimuli used in the Wizard of Oz session.

3.3.2 Generating Answer Sheets. Based on prior examples from chart QA systems, we expected two types of queries, namely those about *visualizations* and those about the underlying *data*. For the visualizations, we prepared information about the chart type (e.g., “what is a bar chart?”), visual elements (e.g., “What does the x-axis represent?”) and other encodings. For the underlying data, we used prior work, Calliope [68], to generate different types of data-driven facts, including data points at the extrema, the proportion of a specific data point compared to the total, the data trend, etc. We answered the participant’s query if our list of facts allowed us to. We also answered yes-no queries (e.g., “Does the chart show increasing trend?”) by referring to this list, and low-level mathematical queries (e.g., calculating the difference between two values) by performing the calculation on the spot. Scoping our responses around an existing list of facts helped us assess which queries could be answered using existing systems and identify categories of unsupported queries. For consistency in responses, we also prepared answer templates. The wizard chose one of the templates to formulate the answer. For example, to answer “What is the highest vaccination rate?” the template “The [highest/lowest] {attribute name} is {value}” is chosen to formulate the final answer “The highest vaccination rate is 89.1%.” If a query could not be answered from the information we prepared, the wizard consistently responded with “The system does not know the answer.”

3.4 Procedure

Before the session began, we distributed a survey including questions about the participants’ demographics and vision conditions. We also sent the link to each study stimulus just before the study session to prevent participants from familiarizing themselves with the visualization beforehand. Each session was led by a researcher who simultaneously performed as an experimenter and a wizard.

At the beginning of the session, we explained the overall procedure of the session, including the fact that an AI system will answer their queries, and the experimenter will read it off the system out loud. We also shared that the system may not be able to provide an answer to their queries. Then, participants were asked to examine the study stimuli in a randomly assigned order. While examining each stimulus, participants were asked to generate any queries related to the visualizations and the underlying data. Specifically, we said “Please share any questions that come up in your mind while examining the chart.” Whenever participants struggled to formulate a concrete query or showed reluctance, we encouraged them to keep posing queries, stating that their queries would be used to improve the system in the future.

After each participant finished examining the four stimuli, we revealed that the experimenter used the prepared answer sheets to answer their queries instead of dynamically interacting with a “real” system. We then asked post-tasks questions covering some aspects that we wished to probe further:

- **Process of generating queries:** We first asked participants how they generated queries. The questions we asked included: “Could you describe the process of how you came up with the questions?”, “Did your process of generating questions change based on the type of data or chart that you were given? If so, how?”, “How does having a data table affect your question generation process?”
- **Motivation for using QA systems:** We asked for what contents participants envision QA systems to be useful. The questions included: “Do you think a QA system that can answer questions like this is necessary and useful? If so (if not), why?”, “In what situations would you use this system?”
- **Other needs and preferences for QA systems:** We asked what general features participants wanted in a chart QA system. For example, we asked whether they preferred to type the queries or verbally ask them. Then, we asked for the platform where these systems could be implemented (e.g., a browser plug-in, standalone software). We also asked them to describe their idealistic version of the system, and to compare it to the experiment’s mock QA system.
- **Interest level/familiarity toward the stimuli:** Lastly, to gauge if prior knowledge or personal preferences impacted the queries posed, we asked participants to rate their interest levels toward each dataset and familiarity with each topic and dataset. We asked them to rate on a 5-point Likert scale to provide them with the option of neutral answer and to ensure high quality data [58].

The study stimuli and procedure were iteratively designed through three pilot studies.

3.5 Data Preparation

All sessions were recorded and transcribed. We created a corpus of 979 queries by collecting the queries that participants asked the wizard during the experiment.

3.5.1 Characterizing Queries. To characterize participants’ intentions and extract corresponding system design considerations, we classified the queries based on analytic tasks they focused on. Specifically, we used the low-level analysis task taxonomy from Amar et al. [4], which covers ten analytic tasks and is commonly used in the development of natural language interfaces for visualization (e.g., [25, 54, 67]). We mapped each query to one or more tasks. Queries not related to data or that did not map to one of the ten analytic tasks were tracked separately. The resulting characterization labels are as follows.

- **Retrieve Values:** This task involves returning a value of an attribute, given a key. “What is the vaccination rate in the US?”
- **Filter:** This task involves finding data entries whose attributes satisfy a given condition. “Which month did inventory go below 1 million?”
- **Compute derived values:** This task involves computing an aggregate function over a set of data entries. Common aggregate functions are *average*, *sum*, and *count*. “What is the average life expectancy for the world?”
- **Find extremum:** This task involves finding the topmost or the bottommost data entries of an attribute. “Which year had the least homes for sale?”
- **Sort:** This task involves sorting a set of data entries with respect to a metric calculated from their attributes. “Is the table organized in ascending or descending order?”
- **Determine data ranges:** This task involves determining the span of values of an attribute for a set of data entries. “What is the range of number of houses on sale in 2015?”
- **Characterize data distribution:** This task involves describing the characteristics of the distribution of an attribute in a set of data entries. “Across the world, are there more countries with a low GDP, medium GDP, or a high GDP?”
- **Find anomalies:** This task involves finding a data entry whose attribute values are extraordinary with respect to the rest of the data. “What countries have higher vaccination status despite being in the continent that does not have high vaccination status?”
- **Cluster:** This task involves finding data entries that are similar with respect to some criteria. “Which country has a similar life expectancy to Qatar?”
- **Correlate:** This task involves describing the correlation between two attributes. “Does there appear to be a relationship between GDP per capita and life expectancy?”
- **Non-data:** Queries that are unrelated to data do not fit into any of the above categories. These queries are labeled as non-data queries. “Why does the trend increase?”, “What is a scatterplot?”

Our rule of thumb was to categorize the queries from the system’s perspective (i.e., based on the task that the system needs to perform to answer them). For example, to answer “Does GDP increase with Life Expectancy?” which is a yes-no query, the system must find out whether the two variables have a positive or negative correlation. Thus we classify the query as *correlate* task. Queries that require comparisons (e.g., “How does the United States compare to Canada as far as vaccinations go?”) and more complex queries (e.g., “Was there any consistent change in negative temperature at any decades or between certain years?”) follow this rule as well.

Following prior work on chart QA systems [39] we also classified queries along two dimensions:

- **Visual vs. Non-Visual:** Visual queries refer to graphical elements such as marks, colors, shapes, and axes (e.g., “Are there more blue bars or red bars?”, “Which country has the darkest shade of green?”). On the contrary, non-visual queries refer to the name of the data attributes instead of visual elements (e.g., “Which country has the lowest percentage of the vaccinated population?”). Therefore, interpreting non-visual queries is independent of the visualization, whereas answering visual queries require consulting the visualization. This dimension also applies to non-data queries since users may ask visual queries to clarify the visualization layout (e.g., “What is the interval on the y-axis?”).
- **Look-up vs. Compositional:** This dimension indicates the complexity of the task. Look-up queries can be answered with a simple key-value retrieval operation (e.g., “What is the vaccination status of the US?”). Compositional queries are more complex and require operations beyond a single look-up task (e.g., “Are there any countries with a lower vaccination rate on the continent with a high vaccination rate?”). Non-data queries such as “When was this data collected?” were not coded along this dimension.

Using these predefined codes, two researchers independently coded each query. The disagreements between the two were resolved after a discussion. The disagreements were mostly due to the ambiguity of the queries. 85% of the codes were initially in agreement, with Cohen’s Kappa of 0.72. The remaining 15% that were

Task Type	All Stimuli	V1 (Line/Housing)	V2 (Bar/Temperature)	V3 (Scatterplot/GDP)	V4 (Map/COVID)	With Table	Without Table
Data-related query	715 (73%)	131 (65%)	142 (68%)	240 (75%)	202 (82%)	361 (73%)	354 (73%)
Non-data query	264 (27%)	72 (35%)	68 (32%)	80 (25%)	44 (18%)	131 (27%)	133 (27%)
All queries total	979 (100%)	203 (100%)	210 (100%)	320 (100%)	246 (100%)	492 (100%)	487 (100%)

Data-related query	Find Extremum	139 (19%)	25 (19%)	27 (19%)	51 (21%)	36 (18%)	66 (18%)	73 (21%)
	Retrieve Value	137 (19%)	19 (15%)	11 (8%)	44 (18%)	63 (31%)	69 (19%)	68 (19%)
	Compute Derived Value + Filter	131 (18%)	34 (26%)	30 (21%)	39 (16%)	28 (14%)	72 (20%)	59 (17%)
	Filter	60 (8%)	4 (3%)	9 (6%)	13 (5%)	34 (17%)	39 (11%)	21 (6%)
	Filter + Find Extremum	59 (8%)	6 (5%)	12 (8%)	23 (10%)	18 (9%)	35 (10%)	24 (7%)
	Compute Derived Value	58 (8%)	14 (11%)	18 (13%)	15 (6%)	11 (5%)	21 (6%)	37 (10%)
	Correlate	53 (7%)	14 (11%)	16 (11%)	23 (10%)	—	22 (6%)	31 (9%)
	Compute Derived Value + Find Extremum	31 (4%)	11 (8%)	8 (6%)	11 (5%)	1 (0%)	11 (3%)	20 (6%)
	Cluster	10 (1%)	—	1 (1%)	6 (3%)	3 (1%)	7 (2%)	3 (1%)
	Determine Range	8 (1%)	—	4 (3%)	2 (1%)	2 (1%)	4 (1%)	4 (1%)
	Find Anomalies	7 (1%)	—	—	7 (3%)	—	3 (1%)	4 (1%)
	Characterize Distribution + Filter	7 (1%)	2 (2%)	1 (1%)	2 (1%)	2 (1%)	4 (1%)	3 (1%)
	Correlate + Filter	5 (1%)	—	4 (3%)	1 (0.4%)	—	3 (1%)	2 (1%)
	Sort	3 (0.4%)	1 (1%)	—	—	2 (1%)	2 (1%)	1 (0.3%)
	Determine Range + Filter	2 (0.3%)	1 (1%)	1 (1%)	—	—	1 (0.3%)	1 (0.3%)
	Characterize Distribution	1 (0.1%)	—	—	1 (0.4%)	—	1 (0.3%)	—
	Characterize Distribution + Determine Range	1 (0.1%)	—	—	—	1 (0.5%)	—	1 (0.3%)
	Filter + Sort	1 (0.1%)	—	—	1 (0.4%)	—	—	1 (0.3%)
	Cluster + Compute Derived Value	1 (0.1%)	—	—	—	1 (0.5%)	—	1 (0.3%)
	Find Extremum + Retrieve Value	1 (0.1%)	—	—	1 (0.4%)	—	1 (0.3%)	—
Data-related query Total	715 (100%)	131 (100%)	142 (100%)	240 (100%)	202 (100%)	361 (100%)	354 (100%)	

Table 1: Categorization of queries by stimuli, task taxonomy, and the presence/absence of a data table. A single query can map to more than one analytical task. Zero counts are noted as "-". Compound tasks are bolded and highlighted with light gray.

in disagreement were subjected to multiple sessions of discussion between the researchers and were eventually resolved.

To identify additional themes beyond the predefined ones, two researchers conducted a thematic analysis [11], generating 4 high-level themes and 10 codes.

3.5.2 Post-Task Interview. Post-task interviews were coded by a researcher. We classified the codes based on their themes. After aggregating the initial codebook, the researchers examined the transcripts again and revised the codes as needed. This process resulted in 5 high-level themes and 42 codes.

4 RESULTS

4.1 Characterizing Queries

4.1.1 Overview. We collected a total of 979 queries. On average, each participant asked 41 queries (SD=38). We analyzed the queries based on the task taxonomy proposed by Amar et al. [4] (Sec. 4.1.2), the query types suggested by Kim et al. [40] (Sec. 4.1.3), and emergent themes from the open coding process (Sec. 4.1.4). We further analyzed the queries by 1) each stimulus, 2) the presence or absence of a data table, and 3) participants' interest level and familiarity with the topic and the dataset.

4.1.2 Mapping Queries to Low-Level Analytic Tasks. Table 1 shows the result of classifying the queries according to Amar et al.'s task taxonomy [4]. Among 979 queries, 73% of the queries were relevant to the given data (715 out of 979). Non-data queries (264 out of 979, 27%) were not further classified by this taxonomy. The analysis

of the non-data queries is subsequently covered by the themes presented in Section 4.1.4.

Overall, the two most frequent tasks that the participants tried to carry out through querying were *Find Extremum* (e.g., "What is the maximum on this graph?") and *Retrieve Value* (e.g., "What is the life expectancy for Burundi?"). Approximately 40% of all queries were for these tasks. The next most frequent query type was *Compute Derived Values + Filter* (e.g., "How many countries are represented in South America on this map?"). *Filter*, *Filter + Find Extremum*, *Compute Derived Values*, and *Correlate* tasks constituted 10% of all queries. 66% of the data queries had a single type of task, while 34% contained multiple types of tasks (referred to as compound tasks by Amar et al. [4]).

We found that different stimuli induce participants to ask different types of queries. For example, *Retrieve Value* and *Filter* queries were asked more often for V4 (a map visualization) than for other stimuli. Queries involving the *Correlate* task were found for all stimuli except V4. However, providing a table did not affect the distribution of queries across task types.

4.1.3 Query Type Analysis. Table 2 shows the composition of the collected queries and Figure 2 shows how these compositions differ by stimulus and by the data table. We found that the numbers of look-up queries and compositional queries differed by visualization ($\chi^2 = 30.1, p < .01$). However, the numbers of visual and non-visual queries were not different across different stimuli ($\chi^2 = 4.6, p < .2$). We did not find reliable differences in the number of look-up vs. compositional queries ($\chi^2 = 0.1, p = .8$) or visual vs. non-visual

	Lookup	Compositional	Total
Visual	22 (3%)	72 (10%)	94 (13%)
Non-visual	113 (16%)	508 (71%)	621 (87%)
Total	135 (19%)	580 (81%)	715 (100%)

Table 2: The breakdown of number of queries by visual vs. non-visual and lookup vs. compositional.

queries ($\chi^2 = 0.1, p = .7$) by whether a data table was provided or not.

4.1.4 Thematic Analysis. In addition to tagging participant queries based on predefined taxonomies, we also conducted an open coding to detect higher-level themes.

Theme 1: Queries due to inaccessibility.

The following set of codes summarizes queries specifically formulated due to the participants' blindness, and the lack of accessibility of the visualization.

- **Answers available to sighted individuals:** A sighted individual could directly answer some queries by referring to the visualization. The results show that 27% of queries (263 out of 979 queries) had this property. Queries about finding extremum (e.g., "What year had the greatest positive temperature anomaly?" (P1)), look-up (e.g., "What is the GDP of Qatar?" (P24)), correlation (e.g., "What is the trend of this bar chart?" (P14)), and queries about the visualization (e.g., "What is the X-axis?" (P2)) mostly make up this category.
- **Understanding trends in detail:** While it may be apparent to sighted people, understanding data trends is a non-trivial task for BLV people. We observed that participants were interested in characterizing data trends in detail, particularly when interacting with line charts. For example, P11 asked "Do we see any decreasing trend for a short period of time, or is it always increasing?" In many cases, participants tried to learn this information by themselves by asking the system to perform a series of data operations. For example, P15 first asked "What was the percentage of increase or decrease in the average number of houses on sale between 2015 and 2020?" and then followed up with a narrower time frame "What was the percentage of increase or decrease in the average number of houses on sale between 2015 and 2017?"
- **Beyond the expected visualization task:** We observed several queries that sighted individuals would not ask. BLV users' could not perceive the visual patterns, so their queries were not confined by the type of visualization. For example, a sighted user would look at a scatterplot and try to find *correlation* between two variables represented by the x and y-axes. We observed, however, nine instances where participants asked *correlation* queries with variables not mapped to the x and y-axis. For instance, some participants asked about the correlation between GDP (x-axis) and the population (size of the circle) or life expectancy (y-axis) and population (size of the circle) when they interacted with V3.
- **Ambiguous references & Out of scope:** Since participants could not see any forms of data attributes (e.g., column names in the table, labels and legends of visualizations), many references used in the queries were ambiguous. 108 out of 979 queries (11%) comprised this

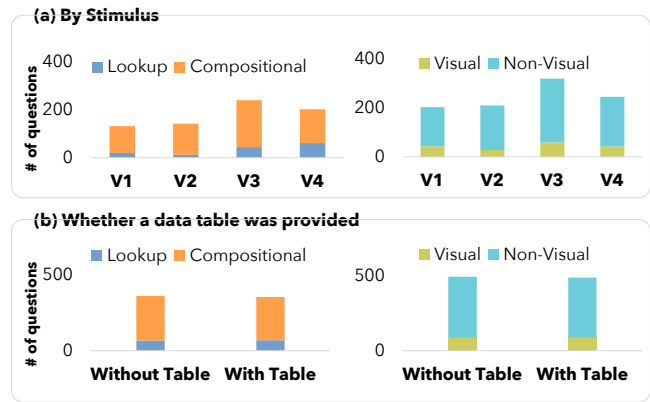


Figure 2: The breakdown of number of queries by stimulus type and data table availability.

category. For example, the column named "inventory" in V1 was referred to by the participants as P3 "homes for sale", P13 "homes on sale", and P23 "houses sold". It was also common to ask for levels of details that do not match the granularity of the actual data. For example, when V1 and V2 present monthly data (labeled yearly), participants asked queries with various levels of temporal scales, including "century" (P13), "decade" (P17), "season" (P3), "month" (P9), "week" (P6), and "day" (P12). While examining V4, where the geographical units were country and continent (indicated in the legend), participants asked about aspects of the data in "Western European" (P7) or "The Middle East" (P14), which is more granular than the actual data.

- **Misconception about the visualization:** 4% of the queries (37 out of 979) could not be answered due to the participants' misconceptions about the visualization. On such occasions, we found that they will continue to ask unanswerable queries if their misconception is not rectified immediately. For example, P12 thought V1 is about house prices instead of the number of houses, and continually asked queries that could not be answered based on the given chart's data ("What was the third most expensive house sold in 2020?", "What was the average price per home between 2015 and 2018?").

Theme 2: Non-Data queries.

The following set of codes summarizes queries that were not related to the given data.

- **Understanding visualization:** Participants frequently asked queries to understand the layout of the visualization (91 out of 979, 9.3%). These queries aimed to clarify data encodings (e.g., "What is the x-axis?", "What does red mean?") or ranges and scales (e.g., "What is the interval on the y-axis?", "Is the x-axis linear, exponential, or logarithmic?"). Participants even sought to clarify the meaning of specific visualization types (e.g., "What exactly is a scatterplot?").
- **Clarifying the alternative text and table:** Some queries regarded the alternative text or the table if provided (55 out of 979, 5.6%). Since the alternative text and the data table was the only information they could access before interacting with the system, many participants tried to ensure they understood them. For example, queries include "What is meant by medium range?" (P8), "What is a scatterplot?"

(P14), and “Is Bolivia the first column of the chart and South America the last column of the chart?” (P24). Furthermore, participants were confused when they found a discrepancy between different sources of information. They asked queries to resolve this discrepancy by asking, “The chart depicts homes for sale starting in 2015, and the graph depicts sales starting in 2014?” (P6).

- **Acquiring Contextual information:** 219 out of 979 queries (22.4%) aimed to acquire more context around data. They included queries about definitions (e.g., “What is GDP?” (P2), “What is the temperature anomaly?” (P13)), clarifications about units (e.g., “Is that Fahrenheit or Celsius?” (P1)), about the scope of data (e.g., “Is the graph showing in the US?” (P5), “What is the source of the data?” (P14)), and about the topic (e.g., “What kind of vaccine did they use?” (P11), “Is this anomaly representative of global warming?” (P12)). These queries include not only queries whose purpose is to analyze data, but also queries with epistemological or rhetorical functions, such as “Is it true that the longer a person lives the more money he makes over the lifetime?” (P12), “What is this?” (P14), “Why did you choose the color red and blue?” (P19)

Theme 3: Queries dependent on prior queries

The following set of codes summarizes queries that must be understood in relation to their preceding utterances.

- **Follow-up queries:** We observed 24 instances (2.5%) where participants formulated a follow-up query based on the previous answer. Often the queries directly referred to the previously given answer, usually with a pronoun, to formulate a new query. For example, P5 first asked, “Which country has the lowest life expectancy?” followed by “How many years do people expect to live in those countries?” P20 asked, “Which country has the lowest life expectancy?”, then followed up with: “What is its GDP?”
- **Rephrasing of previous queries:** When the system failed to answer a query, some participants rephrased it and asked again (16 out of 979, 1.6%). They assumed that reformulating the query would elicit a better answer from the system. For example, P12 asked, “what is the average vaccination of the State of Texas in the United States?” When they examined V3. However, V3 only contains country-level data, with no state information available. P12 asked again when the system failed to answer, “What is the vaccination average for Texas?” P14 asked an ambiguous query about the visualization structure, “Where does the line begin?” and asked again when the system failed to answer: “What is the first data point on the line?”

Theme 4: Phrasing variations & Testing system capabilities

This theme offers insights into how participants linguistically phrase their queries and the strategies they use to learn how the system behaves to expedite future interaction.

- **Queries with binary answers:** Some queries had binary options for their answers (154 out of 979, 16%). The most common cases were queries that could be answered with yes/no responses. For example, P4 asked “Do Oceania and Europe have a similar life expectancy?” and P22 asked, “Is New Zealand included on this map?” (P22)). Other types include comparison (e.g., “What is the difference in life span between Great Britain and the US, and which

is higher?” (P7)) and clarification (e.g. “Is that a row or column?” (P14)).

- **Commands:** While it was not very common, we observed three queries articulated as a command rather than a query. All of them ordered the system to provide a natural language description. Examples include “Read me a description of this chart.” (P14), “Please describe what you mean by anomaly” (P24).
- **Testing the system:** In some instances (6 out of 979), participants tested the capabilities of the system. Some of these queries were motivated by curiosity (e.g., “What continent is green? What is Green?” (P14)), while others were to facilitate future interaction by qualifying one’s expectations about the system (e.g., “How many homes were for sale in the second week of June? How many homes were for sale on June 12, 2017?” (P13)).

4.1.5 The Effect of Interest & Familiarity with Topic. While familiarity with the topics and the datasets were similar across different stimuli (Figure 3 (b) and (c)), the interest levels toward the topics were varied (Figure 3 (a)). Participants showed the most interest in V4 COVID, followed by V3 GDP, V2 Temperature, and V1 Housing.

We constructed a mixed effect model to further understand how these factors impact the number of queries participants generated. The model shows that the self-rated interest positively impacts the number of generated queries ($t = 2.5, p < .05$). As one unit of interest increased, around three more queries were generated by a participant on average. The familiarity towards the topic ($t = 0.5, p = .63$) and the dataset ($t = -0.6, p = .58$) did not affect the number of generated queries reliably.

4.2 Phrases/Terms Analysis

To inform parsers of future natural language systems, we also collected and classified the phrases and terms participants used to refer to visual elements and tasks.

- **Visualizations:** “map”, “chart”, “graph”, “scatterplot”, “picture”
- **Marks:** “point”, “line”, “bar”
- **Axes & Range:** “x axis”, “y axis”, “coordinate”, “interval”, “increment”, “origin”, “intercept”, “baseline”, “upper range”, “lower range”
- **Color:** “shade”, “dark”, “darkest”, “bright”, “light”
- **Trend:** “ascending”, “descending”, “squiggly”, “mountainous”, “hills”, “vallies”, “linear”, “exponential”
- **Positional information:** “left”, “right”, “East(ern)”, “West(ern)”, “up(per)”, “medium”, “low(er)”, and “diagonal”
- **Tables:** “table”, “row”, “column”

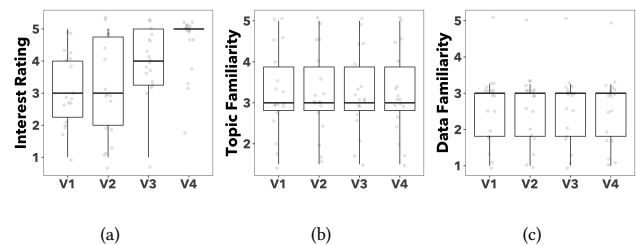


Figure 3: (a) Self-rated interest rate towards each dataset, (b) self-rated familiarity with each topic, and (c) self-rated familiarity with each dataset

- **Finding extremum task:** “most”, “least”, “highest”, “lowest”, “lightest”, “darkest”, “warmest”, “coldest”, “fewest”, “largest”, “biggest”, “greatest”
- **Correlation task:** “trend”, “relationship”, “correlation”, “trajectory”
- **Computing derived value task:** “mean”, “average”, “median”, “middle”, “variation”, “difference”, “change”, “compare”, “percentage”, “decrease”, “increase”, “increment”, “how many”
- **Determining range task:** “range”
- **Sorting task:** “sort”, “ascending”, “descending”

In many cases, the vocabulary was directly borrowed from the alt text. In V1, “homes for sale” was frequently used to refer to the name of the attribute. In V2, “degrees Fahrenheit” was commonly appended to a temperature value. In V3, “upper range” and “lower range” were frequently used to convey spatial information. In V4, “shade of green” was commonly used to refer to the brightness of the color green.

4.3 Why Could Certain Queries not be Answered?

Among all 979 queries, 60% were answered during the sessions. We analyzed why the remaining queries could not be answered by our answer sheet. First, some compound queries that require more than one data operation could not be answered, as our answer sheet only covered relatively simple tasks [68]. Second, queries about the context could not be answered as they required information beyond the given data. These queries correspond to Theme 2-contextual queries in Section 4.1.4. Lastly, ambiguously phrased queries could not be answered as their intentions could not be ascertained by the wizard.

4.4 Post-Task Queries

4.4.1 Understanding Motivation for Using Chart QA Systems. We found that participants had several motivations when they generated the queries.

- **Finding interesting data facts:** We observed several participants were particularly motivated to find some interesting data facts. For example, P22 shared that they asked queries based on “what information I think would be the most important and the most interesting.”
- **Filling the gap in knowledge:** Many participants shared that they generated queries to fill the gap in their understanding of the data. For example, P5 mentioned, “I thought about what information it (alternative text) didn’t give me.” P4 echoed that “I kind of start by thinking like what information don’t I have and what would I need to draw my own conclusions about.”
- **Conducting data analysis:** Data analysis was another common motivation to generate queries. P7 shared how they wanted to *compare* the two values: “A lot of the queries were a mathematical calculation that I would have to look at two points of data, at the same time, which I can’t do. With the screen reader, you can only look at one point of data at a time. So I was getting information about two different points of data.” P13 attempted to *retrieve value*: “If I want to know the vaccination rate of an individual country, I need to be able to ask that question. I can’t just quickly scan Africa to see, so I might want to ask about it.” Other participants wished

to *detect anomalies*: P11 “I want to know if there is any anomaly in that trend, like if there’s an increase or up and down or steady.”

4.4.2 Platform & Modality. We asked the participants in which platform or context the system should be implemented. We gave them example choices including stand-alone software, PDF reader plug-in, and Google Chrome extension. Most participants shared that a browser plug-in would be ideal for such a system due to its applicability. P8 shared “I think a web extension would be more universal.” P17 echoed: “I think a web extension that you can use on any page would be the most useful.” While P7 also wished to have the system as a web extension to parse data online, they see the value as a separate interface that they can use to examine their data: “It’d be really good as a web extension like a chrome extension. You also would probably need a website to upload your own data.” This reconfirms and further motivates the design of browser plug-in technologies such as VoxLens [61, 62, 64].

In terms of input modality, participants’ preferences were varied. About half of the participants had no preference between typing or speaking. Some participants specified circumstances that determined their preference. P8 mentioned “If I’m by myself, then verbally. If it’s a library or something, then I prefer typing.” P9 also shared that “I am okay with both. But I’m on my phone. I don’t like typing, so I like asking verbally. But if I’m on my computer, I don’t mind typing.” Some participants raised concerns about verbally asking queries due to the discursiveness of the generated queries. P4 shared “So I definitely would have rather have typed it to have questions that were not so clunky.” Some participants raised a concern with typing because of the complexity of an envisioned interaction. P13 shared, “When you’re typing with a screen reader, you have to focus on the input box. If you want to flip back to an article or something, you start jumping between things. It just takes a while to type out some questions.”

4.4.3 Perceived Usefulness & Usage Contexts. All participants expressed that the QA system would be highly helpful. P1 shared their enthusiasm for using the system: “This seems extremely useful. Actually, I’m quite excited about it.” Participants felt the system could be helpful when it allowed them to discover aspects of data not available to them. P10 shared, “Because you know, all I have to go on is the description that’s written down, so I could see myself using it for that.” P6 also mentioned that “It would be cool to have a system like that. I would want to be able to go on the Internet and find a chart or graph that needed more interpretation beyond what they provided in the alt text or even if they haven’t provided alt text and be able to plug it in and have the ability to get the data from it in a textual way, as opposed to having to ask somebody to describe it.” P6 also echoed the fact that they don’t need to ask anybody but the system: “because it is an automated machine, the machine wouldn’t judge me for asking a question that seems silly.”

Several participants particularly noted that the system will be useful in an article reading scenario, further validating our choice of stimuli. P3 shared that “if I was reading an interesting article and it had the data, and I wanted to know something more about the data, I would use the system.” P20 echoed that “I like to be a consumer of news and current events, and quite often in the media there are graphs, maybe unemployment or economics, or whatever the availability of houses, I want to have access to that.”

While participants were fluent in navigating data tables and appreciated access to the raw data, they saw QA systems as a better alternative to data tables in some scenarios. For instance, P8 shared “It would be helpful, especially with long tables. You could ask the system.” P10 echoed that “I think it would be very useful, especially with charts and graphs even with tables, the last thing I want to do is scan through a really long table, so I know it would be really useful just to help people understand things in a quicker amount time.” P22 further emphasized that QA systems can support the task that takes longer when using data tables: “It would cut down on the work that I would have to do. I could ask questions and find out the answers faster. Maybe to compare things, you know different dates in different countries and stuff like that it might be faster.”

5 DISCUSSION

Our analysis characterizes the queries generated by participants to answer RQ2. We examined the queries by their constituent low-level analytical tasks and their complexity, strategies used to refer to entities, and other prominent themes. We found that a significant portion (27%) of queries were non-data queries, mostly related to the context of the visualization. Among the data-related queries, *find extremum* (19%), *retrieve value* (19%), and *compute derived value with filter* (18%) were the most prevalent analytic tasks. 13% of data-related queries contained a reference to a visual element of the visualization, and 81% were compositional queries. Thematic analysis showed that some queries are due to insufficient accessibility to the visualizations, that queries depend on prior queries, and that there are different query formats. The range of queries exceed the scope of what current accessible technology (e.g., [64]) can answer. In addition, answering RQ1, we found that BLV people have multiple motivations to use the QA system, including data analysis, discovery of interesting information, and filling the gap in their knowledge. In the following sections, we focus on RQ3 and discuss how much we can piggyback on existing chart QA systems designed for sighted individuals. We also list design considerations for future systems informed by our study.

5.1 Comparison with Sighted People

5.1.1 Composition of Queries. Table 3 shows the results from Kim et al. [40] study and contrasts them to our results. We found that BLV people tend to ask more compositional queries than sighted people ($\chi^2 = 23.4, p < .01$). In particular, the proportion of compositional queries with visual reference is more than two times higher than that of sighted people, suggesting that QA systems for BLV users need to have notably richer analytical capabilities. However, there was no difference between the two groups in the ratio of visual and non-visual queries ($\chi^2 = 0.34, p = 0.5$).

While this result is interesting, we note that it should be contextualized by the setup of each study. Specifically, participants in our study were prompted to verbally generate queries, whereas the sighted participants in the previous study were prompted to type both the query and its answer.

5.1.2 Benchmark with a QA System. We benchmarked the collected queries with a state-of-the-art system to understand how much we can piggyback on the existing functionality as well as what improvements are needed to fully support BLV people. We chose the

	Lookup		Compositional		Total	
	Sighted	BLV	Sighted	BLV	Sighted	BLV
Visual	52 (8%)	22 (3%)	24 (4%)	72 (10%)	76 (12%)	94 (13%)
Non-visual	138 (22%)	113 (16%)	415 (66%)	508 (71%)	553 (88%)	621 (87%)
Total	190 (30%)	135 (19%)	439 (70%)	580 (81%)	629 (100%)	715 (100%)

Table 3: Juxtaposition of query types observed in Kim et al.’s study with sighted individuals and queries collected during our study with people with BLV.

system based on four criteria from existing chart QA systems [31]. First, the system should be able to understand references to visual elements. Second, the system must support multiple visualization types. For instance, DVQA [35] is not considered because it only focuses on bar charts. Third, the system should demonstrate high accuracy in understanding the visualization and its underlying data so that the analysis can be about the types of queries that should be answered by the future system, instead of being about improving the system’s recognition of visual elements. Lastly, there must not be a strict restriction on the query format. For example, models for FigureQA [38] and FigureNet [57] only support yes-no queries, and are hence excluded. With these criteria, we chose the system presented by Kim et al. [39] over others suggested in the computer vision community [36, 69].

We tested the system with the queries from V1 and V2 as the original system was tested on bar and line charts. We filtered out unanswerable queries, which either had no correct answer (e.g., queries about the topic), could not be answered from the data and the visualization (e.g., queries about the metadata), or contained ambiguous words. This process yielded a total of 245 queries, which we used as input to the system. Then, we manually checked whether each generated answer is correct. Run time errors were considered incorrect answers. The system showed an accuracy of 16%. However, our goal was not to assess the performance of the existing chart QA system but instead to understand what queries can and cannot be answered by existing chart QA systems.

We found that most queries that were answered accurately were *Retrieve values* queries, queries about the layout of the visualization (e.g., “What is the X-axis?”) and *Find Extrema* queries. We also found that a few of the compound queries with more than one analysis task (e.g., “How many dates is the inventory below 500,000?”) and comparison queries (e.g., “What’s the difference in temperature of the ocean between 1995 and 2020?”) were answered correctly.

Next, we analyzed the incorrect answers and found three aspects of the system that can be improved to support BLV people’s queries. First, the *parsing of complex queries* that entail more than one task could be improved. These queries correspond to those that are tied to two task types in Table 1 as well as compositional types in Figure 2.

The system did answer a few of these queries correctly, indicating that it is the parsing of the queries into smaller tasks that can be improved to better support BLV people.

Second, the system can benefit from a *more robust disambiguation of words* in the query. It struggled when it encountered a word

or a phrase that does not match the nomenclature used in the visualization. The discrepancy is sometimes due to using a level of detail incompatible with the actual data (e.g., “20th century” (P13, V2)).

Third, the system can be improved by *supporting a wider range of answer types*. The system only provides a single data value or an attribute name for an answer. Therefore, queries that ask for a different type of answer, such as yes-no queries (e.g., “Is the temperature getting colder between 1902 and 1957?” (P12)), queries for *range* task (e.g., “What is the range of houses sold in 2015?” (P14)), and enumeration queries (“What were the years that had between 1.5 and 2.1 degrees Fahrenheit?” (P14)), were not supported. Furthermore, there was no easy way for the participants to check if the given answer is correct or not. Adding more answer types to address these situations can increase the system’s accessibility.

5.2 Design Considerations for Future Systems

We derived a set of design considerations for future systems based on our analysis. We list the considerations categorized by steps in the pipeline of chart QA systems: 1) Parsing visualizations (how to create the visualization representation to be used to answer queries), 2) Parsing queries (what types of queries the system should expect), and 3) Providing answers (how to present answers). We also offer other considerations beyond the system itself.

5.2.1 Parsing Visualizations.

- **Integrating vocabularies from alternative text and tables:** We saw that on several occasions, participants reused phrases from the alt text to frame their queries (Section 4.2). Correspondingly, QA systems should be able to understand users’ utterances that are directly borrowed from or refer to the alt text with high accuracy. Since word usage in alternative texts can prime BLV people when formulating queries, the system can leverage this when it creates the synonym dictionary. Alternative texts are supposed to use more plain language [33] (e.g., homes for sale), whereas the extracted data attribute from the visualization can use a more formal term directly from the raw data (e.g., inventory). The system can parse the alternative text to find the most semantically similar utterances and create the synonym dictionary.
- **Integrating context around the visualization:** Contextual queries were very common, taking up 22% of the total queries. Though some of these queries like “What is GDP?” (P2) and “What is a scatter-plot?” (P14) can be answered by a general-purpose QA system, other queries like “Why is the graph increase and decrease in the certain period?” (P16) and “Why does it start at 1880? What happened then?” (P21) require the system to be aware of the context to be answered properly. The system should be able to understand how the visualization is situated (e.g., inside a news article about global warming) and use this information to help answer the contextual queries. Besides incorporating information from the setting within which a chart is used, this also presents a research opportunity to explore how knowledge bases such as WolframAlpha [85] can be leveraged during chart QA.

5.2.2 Parsing Queries.

- **Queries that require a sequence of data operations:** As noted earlier, participants frequently asked compositional queries that

require multi-step analytical operations to generate a response. To this end, the system should be able to parse a query that asks for a complex data analysis task and break it down into a series of executable primitive data analysis tasks. Our study result alludes that BLV people may want to accomplish more complicated tasks with longer queries than sighted people. Prior work also corroborates our finding that BLV people formulate more complex queries when using speech input [3, 8].

- **Queries formulated with semantically relevant utterances:** We conjecture that the synonym dictionary for BLV people should be larger than those for sighted people. For sighted people, nomenclature for referring data and other elements are readily available visually from the title, labels, legends, etc. Therefore, they are more likely to formulate queries based on those vocabularies. For BLV people, however, there is no reference to rely on at the time of generating queries. For example, we observed a range of synonyms of “GDP per capita” including “income”, “gross domestic product”, and “money.” Because of the similar reason that the nomenclature is not readily available, participants ask many queries that are beyond the scope of data but semantically similar. For example, when the data point is available weekly, participants ask aspects of data by “day”, “month”, and “season” while sighted people’s queries may be bounded by the granularity shown in labels and ticks. Similarly, although a chart may present information at a country or state-level, users queries may involve cities. Even though we provided the tick information in the alternative text, some participants might have formed the impression that the visualization encoded temporal or geographic data at multiple levels. We envision two ways to resolve this issue. First, the system can provide a clear response to participants to adjust their mental model (e.g., “the system can only answer questions about weekly data”). Or alternatively, the system can be prepared to answer for the different levels of aggregation/granularities when they are available in the underlying dataset.
- **Rephrased queries:** The system should take its previous queries into account when parsing a query. For example, the system can identify whether the query is a follow-up query by 1) comparing the similarity between the $(N - 1)^{th}$ query and the N^{th} query and 2) examining the confidence score of the $(N - 1)^{th}$ query’s answer. When the system parses the N^{th} query, it can leverage the utterance in the $(N - 1)^{th}$ query. We envision the rephrased queries will be more prevalent among BLV people because they might utilize more diverse utterances than sighted people.
- **Follow-up queries and pronouns:** The system should be able to deal with follow-up queries involving ill-defined references and pronouns, for example, by implementing existing techniques such as pronoun disambiguation [15]. While this recommendation is not specific to chart QA systems for BLV people, the effect of having this capability will be more beneficial for BLV people than for sighted individuals. For example, since BLV users rely only on their memory to keep track of what queries they asked and what answers they received, it may reduce BLV people’s cognitive load if the system allows pronouns to refer to the previous answer or an entity in the previous query.
- **Other query formats beyond interrogative sentences:** A variety of query formats, including binary queries and commands,

should be supported by the system. We observed that participants asked many binary queries to construct a mental representation of visualizations (e.g., “Is 50 percent dark green and 25 percent light green?”), confirm the scope of the data (e.g., “Is New Zealand included on this map?”) or clarify their understanding of data-related facts (e.g., “Is Africa the lowest on this scale?”).

- **Queries with visual reference:** We found that BLV people use visual references as much as sighted people do. Therefore, the system should be able to parse utterances using graphical elements of the visualization, leveraging functionalities of systems that support sighted people (e.g., [40]). However, an important added consideration is the potential need for generating non-visual explanations for responses to visual queries such that the explanations are comprehensible by BLV users.

5.2.3 Providing Answers.

- **Adjusting & communicating uncertainty level:** The system should be able to identify and communicate whether the answer to a query is obtainable from the provided data or not. While sighted people can test the system’s accuracy by asking obvious queries, there is no way for BLV people to evaluate the accuracy of the system to tailor their accuracy perception toward the system. Also, sighted people may have more chances to adjust the misconception by visually inspecting the chart, whereas BLV people are not able to. This is evidenced by our findings on participants’ misconceptions about the visualization (e.g., “How many temperature anomalies were there in 2020?” when temperature anomaly is a measure of temperature, “Which city is the table depicting?” when the table is for the whole of the United States, and “What countries are represented by green?” when a number is represented by the brightness of green). Many of these misconceptions could have been intrinsically resolved if they were sighted. Therefore, we argue that the confidence threshold to determine whether to say “this question is unanswerable” or to provide an answer should be higher for a QA system intended for BLV people. However, further study is needed to tune the appropriate level of the threshold. Another possible approach is to provide BLV people the information about the degree of uncertainty of the systems’ answers. This may allow BLV people to better interpret the system’s answer. Again, future study is needed to understand the trade-off between the burden of interpreting uncertainty and having incorrect answers.
- **Capability to elaborate trends in detail:** The participants wanted to learn about the detailed trend and the overall trend to “visualize” the patterns of the data at a detailed level. The system should be equipped to support these tasks by expanding the vocabulary to describe the pattern. For example, “increasing trend” might not be enough to satisfy users as we observed a few participants kept asking to clarify how stiff the increasing trends are or a period where exhibited the increasing trend without a fluctuation. Possible approaches include utilizing a sonified chart [64] and generating texts that span the whole of the visualization’s semantic levels [48].

5.2.4 Other Considerations.

- **Supporting general question answering:** The system should support answering contextual queries that are out of the scope of the data. Relatively simple queries, such as asking about the definition of a word (e.g., “what is GDP?”), can be answered by a

general-purpose virtual assistant application. For BLV people, each interaction to search can cost more than for sighted people (e.g., opening a new tab to search an unfamiliar term and coming back to the original tab to examine the visualization). Thus, integrating a simple feature that allows BLV people to search easily will likely save their time and lower the distraction.

- **Providing example questions:** During the study, several participants shared that it would be useful to see possible queries before using the system. For instance, P1 said “a little bit of guidance on the type of questions using an example question would be great.” P6 also said, “Giving examples of what kinds of questions you could ask would be great, like the tutorial.” Thus, to aid the discoverability of the system’s understanding capabilities and make users aware of potentially interesting queries, QA systems should provide some example queries that can be asked. It is harder for BLV people to figure out what analysis task to perform because they cannot look at the visualization, which is typically designed to afford a specific type of data analysis (distribution for histograms, the trend for line charts, etc.). One approach to alleviate this problem could be to recommend analysis tasks and potential queries that are commonly posed on the given visualization and data type (e.g., [71]).

5.3 Factors that Influence Query Formulation

Our work also offers a deeper understanding of the factors behind participants’ query formulation. First of all, providing data tables alongside the visualization does not seemingly affect the types of queries participants generated. We expected that participants might be asking more complex queries when the data table was available, as they could conduct simple tasks using the table. However, the distribution of simple analytical tasks and the compound tasks with and without tables were very similar. This may be because the cost of making a query is lower than navigating the table, even if it is for a very simple task.

Unsurprisingly, the interest level in the topic and the dataset were positively correlated with the number of queries participants generated, validating the experimental premise. Based on the finding, we envision that a chart QA system can be highly useful when deployed alongside a thematic data-driven news article.

5.4 Limitation and Future Work

As we have mentioned, interpreting our study findings relative to prior study with sighted individuals [39] requires caution. Future work should conduct a comparative study with sighted people and BLV people to compare the two groups in an identical setup.

While our stimuli were chosen to cover various chart types and topics, our set-up does not allow us to inspect the impact of the chart type on the querying behaviors, as they are confounded. Future studies can investigate different types of visualizations with the same dataset to tease apart the impact of chart type on query generation. Similarly, measuring the participants’ interest level toward the dataset after the experiment might have disrupted an accurate measurement.

The effect of individual traits on query asking patterns could be interesting to study. Besides the demographic information, traits such as locus of control and spatial ability have been shown to impact visualization comprehension of sighted people [46]. Their

effect on BLV users' interaction with QA system could be investigated further.

Our findings on blind people's queries about visualizations can also inform the authoring of alternative texts and pedagogical methods in math and science education. For example, the National Center for Accessible Media's (NCAM) guidelines that are frequently followed by school teachers [51, 59], and guidelines for alternative text for digital visualizations [33] can benefit from our characterization of BLV people's queries.

6 CONCLUSION

Our work offers a detailed look into how BLV people possibly interact with a chart QA system. We observed how they formulate queries and how they envision using a chart QA system through a Wizard of Oz study. We found that multiple interesting characteristics emerged from the queries that are possibly unique to BLV people. We also list multiple factors that influence their querying behavior and confirm sufficient motivation from BLV people to use the system. Based on our observations, we present design considerations for chart QA systems for BLV people. We envision that our study findings, as well as the released dataset, will pave the way for future research toward more accessible chart QA systems.

REFERENCES

- [1] Asma Ben Abacha and Pierre Zweigenbaum. 2015. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information processing & management* 51, 5 (2015), 570–594.
- [2] Shanza Abbas, Muhammad Umair Khan, Scott Uk-Jin Lee, Asad Abbas, and Ali Kashif Bashir. 2022. A Review of NLIDB with Deep Learning: Findings, Challenges and Open Issues. *IEEE Access* (2022).
- [3] Ali Abdolrahmani, Ravi Kubler, and Stacy M Branham. 2018. "Siri Talks at You" An Empirical Investigation of Voice-Activated Personal Assistant (VAPA) Usage by Individuals Who Are Blind. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 249–258.
- [4] Robert Amar, James Eagan, and John Stasko. 2005. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*. (Minneapolis, MN, USA). IEEE, 111–117.
- [5] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Salt Lake City, UT, USA). IEEE, 6077–6086.
- [6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision* (Santiago, Chile). IEEE, 2425–2433.
- [7] H. K. Ault, J. W. Deloge, R. W. Lapp, M. J. Morgan, and J. R. Barnett. 2002. Evaluation of Long Descriptions of Statistical Graphics for Blind and Low Vision Web Users. In *Computers Helping People with Special Needs (Lecture Notes in Computer Science)*, Klaus Miesenberger, Joachim Klaus, and Wolfgang Zagler (Eds.). Springer, Berlin, Heidelberg, 517–526. https://doi.org/10.1007/3-540-45491-8_99
- [8] Shiri Azenkot and Nicole B Lee. 2013. Exploring the use of speech input by blind people on mobile devices. In *Proceedings of the 15th international ACM SIGACCESS conference on computers and accessibility*. ACM, New York, NY, USA, 1–8.
- [9] Cristian Bernareggi, Dragan Ahmetovic, and Sergio Mascetti. 2019. μ Graph: Haptic Exploration and Editing of 3D Chemical Diagrams. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 312–317.
- [10] Stacy M Branham and Antony Rishin Mukkath Roy. 2019. Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 446–458.
- [11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [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*. ACM, New York, NY, USA, 131–138.
- [13] Davide Cervone, Peter Krautzberger, and Volker Sorge. 2016. New accessibility features in MathJax. *J Technol Pers Disabil* 4 (2016), 167–175.
- [14] Jinho Choi, Sanghun Jung, Deok Gun Park, Jaegul Choo, and Niklas Elmqvist. 2019. Visualizing for the Non-Visual: Enabling the Visually Impaired to Use Visualization. *Computer Graphics Forum* 38, 3 (2019), 249–260. <https://doi.org/10.1111/cgf.13686> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13686>
- [15] Albert T Corbett and Frederick R Chang. 1983. Pronoun disambiguation: Accessing potential antecedents. *Memory & Cognition* 11, 3 (1983), 283–294.
- [16] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—why and how. *Knowledge-based systems* 6, 4 (1993), 258–266.
- [17] Phong-Khac Do, Huy-Tien Nguyen, Chien-Xuan Tran, Minh-Tien Nguyen, and Minh-Le Nguyen. 2017. Legal question answering using ranking SVM and deep convolutional neural network. *arXiv preprint arXiv:1703.05320* (2017).
- [18] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Inkit Padhi, Mattia Rigotti, Jarret Ross, Yair Schiff, Richard A Young, and Brian Belgodere. 2020. Image captioning as an assistive technology: Lessons learned from vizwiz 2020 challenge. *arXiv preprint arXiv:2012.11696* (2020).
- [19] Zohar Eitan, Eitan Ornoy, and Roni Y Granot. 2012. Listening in the dark: Congenital and early blindness and cross-domain mappings in music. *Psychomusicology: Music, Mind, and Brain* 22, 1 (2012), 33.
- [20] Niklas Elmqvist and Ji Soo Yi. 2015. Patterns for visualization evaluation. *Information Visualization* 14, 3 (2015), 250–269.
- [21] Christin Engel, Emma Franziska Müller, and Gerhard Weber. 2019. SVGPlot: an accessible tool to generate highly adaptable, accessible audio-tactile charts for and from blind and visually impaired people. In *Proceedings of the 12th ACM International Conference on Pervasive Technologies Related to Assistive Environments*. ACM, New York, NY, USA, 186–195.
- [22] Leo Ferres, Gitte Lindgaard, Livia Sumegi, and Bruce Tsuji. 2013. Evaluating a Tool for Improving Accessibility to Charts and Graphs. *ACM Transactions on Computer-Human Interaction* 20, 5 (2013), 28:1–28:32. <https://doi.org/10.1145/2533682.2533683>
- [23] Leo Ferres, Petro Verkhogliad, Gitte Lindgaard, Louis Boucher, Antoine Chretien, and Martin Lachance. 2007. Improving accessibility to statistical graphs: the iGraph-Lite system. In *Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility (Assets '07)*. Association for Computing Machinery, New York, NY, USA, 67–74. <https://doi.org/10.1145/1296843.1296857>
- [24] American Foundation for the Blind. [n. d.]. Key Employment Statistics for People Who Are Blind or Visually Impaired. <https://www.afb.org/research-and-initiatives/statistics/key-employment-statistics> Accessed June 1, 2021.
- [25] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. 2015. DataTone: Managing Ambiguity in Natural Language Interfaces for Data Visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. ACM, Charlotte NC USA, 489–500. <https://doi.org/10.1145/2807442.2807478>
- [26] A. Jonathan R. Godfrey, Paul Murrell, and Volker Sorge. 2018. An Accessible Interaction Model for Data Visualisation in Statistics. In *Computers Helping People with Special Needs (Lecture Notes in Computer Science)*, Klaus Miesenberger and Georgios Kouroupetroglou (Eds.). Springer International Publishing, Cham, 590–597. https://doi.org/10.1007/978-3-319-94277-3_92
- [27] Jiangtao Gong, Wenyuan Yu, Long Ni, Yang Jiao, Ye Liu, Xiaolan Fu, and Yingqing Xu. 2020. "I can't name it, but I can perceive it" Conceptual and Operational Design of "Tactile Accuracy" Assisting Tactile Image Cognition. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 1–12.
- [28] Darren Guinness, Annika Muehlbradt, Daniel Szafir, and Shaun K Kane. 2019. Robographics: Dynamic tactile graphics powered by mobile robots. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 318–328.
- [29] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT, USA). IEEE, 3608–3617.
- [30] Highcharts 2022. Highcharts accessibility module. <https://www.highcharts.com/docs/accessibility/accessibility-module>
- [31] Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. 2022. Chart Question Answering: State of the Art and Future Directions. *arXiv preprint arXiv:2205.03966* (2022).
- [32] Shakila C Joyner, Amalia Riegelhuth, Kathleen Garrity, Yea-Seul Kim, and Nam Wook Kim. 2022. Visualization Accessibility in the Wild: Challenges Faced by Visualization Designers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 19.
- [33] Crescentia Jung, Shubham Mehta, Atharva Kulkarni, Yuhang Zhao, and Yea-Seul Kim. 2022. Communicating Visualizations without Visuals: Investigation of Visualization Alternative Text for People with Visual Impairments. *IEEE*

- Transactions on Visualization and Computer Graphics* 28, 1 (Jan. 2022), 1095–1105. <https://doi.org/10.1109/TVCG.2021.3114846>
- [34] Daekyoung Jung, Wonjae Kim, Hyunjo Song, Jeong-in Hwang, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. 2017. Chatsense: Interactive data extraction from chart images. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 6706–6717.
- [35] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Salt Lake City, UT, USA). IEEE, 5648–5656.
- [36] Kushal Kafle, Robik Shrestha, Scott Cohen, Brian Price, and Christopher Kanan. 2020. Answering questions about data visualizations using efficient bimodal fusion. In *Proceedings of the IEEE/CVF Winter conference on applications of computer vision* (Snowmass, CO, USA). IEEE, 1498–1507.
- [37] Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. FigureQA: An Annotated Figure Dataset for Visual Reasoning. *CoRR* abs/1710.07300 (2017). arXiv:1710.07300 <http://arxiv.org/abs/1710.07300>
- [38] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. 2018. FigureQA: An Annotated Figure Dataset for Visual Reasoning. *arXiv:1710.07300 [cs]* (Feb. 2018). <http://arxiv.org/abs/1710.07300> arXiv: 1710.07300.
- [39] Dae Hyun Kim, Enamul Hoque, and Maneesh Agrawala. 2020. Answering Questions about Charts and Generating Visual Explanations (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376467>
- [40] N. W. Kim, S. C. Joyner, A. Riegelhuth, and Y. Kim. 2021. Accessible Visualization: Design Space, Opportunities, and Challenges. *Computer Graphics Forum* 40, 3 (2021), 173–188. <https://doi.org/10.1111/cgf.14298> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14298>
- [41] Jonathan Lazar, Aaron Allen, Jason Kleinman, and Chris Malarkey. 2007. What Frustrates Screen Reader Users on the Web: A Study of 100 Blind Users. *International Journal of Human-Computer Interaction* 22, 3 (2007), 247–269. <https://doi.org/10.1080/10447310709336964> Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/10447310709336964>
- [42] Bongshin Lee, Eun Kyoung Choe, Petra Isenberg, Kim Marriott, and John Stasko. 2020. Reaching Broader Audiences With Data Visualization. *IEEE Computer Graphics and Applications* 40, 2 (March 2020), 82–90. <https://doi.org/10.1109/MCG.2020.2968244> Conference Name: IEEE Computer Graphics and Applications.
- [43] Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. 2015. More than telling a story: Transforming data into visually shared stories. *IEEE computer graphics and applications* 35, 5 (2015), 84–90.
- [44] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696* (2018).
- [45] Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2019. TVQA+: Spatio-Temporal Grounding for Video Question Answering. *CoRR* abs/1904.11574 (2019). arXiv:1904.11574 <http://arxiv.org/abs/1904.11574>
- [46] Zhengliang Liu, R. Jordan Crouser, and Alvitia Ottley. 2020. Survey on Individual Differences in Visualization. *Computer Graphics Forum* 39, 3 (2020), 693–712. <https://doi.org/10.1111/cgf.14033> _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14033>
- [47] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems* 29 (2016).
- [48] Alan Lundgard and Arvind Satyanarayan. 2022. Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. (Jan. 2022). <https://doi.org/10.1109/TVCG.2021.3114770>
- [49] Kim Marriott, Bongshin Lee, Matthew Butler, Ed Cutrell, Kirsten Ellis, Gagatay 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.
- [50] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (Snowmass, CO, USA). IEEE, 1527–1536.
- [51] Valerie S Morash, Yue-Ting Siu, Joshua A Miele, Lucia Hasty, and Steven Landau. 2015. Guiding novice web workers in making image descriptions using templates. *ACM Transactions on Accessible Computing (TACCESS)* 7, 4 (2015), 1–21.
- [52] Tomas Murillo-Morales and Klaus Miesenberger. 2017. Non-visually performing analytical tasks on statistical charts. In *Harnessing the Power of Technology to Improve Lives*. IOS Press, 339–346.
- [53] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Nick Schoelkopf, Riley Kong, Xiangru Tang, et al. 2022. FeTaQA: Free-form Table Question Answering. *Transactions of the Association for Computational Linguistics* 10 (2022), 35–49.
- [54] Arpit Narechania, Arjun Srinivasan, and John Stasko. 2021. NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 369–379. <https://doi.org/10.1109/TVCG.2020.3030378> tex.ids=narechania2021a arXiv: 2008.10723.
- [55] John Neuhoff. 2019. Is Sonification Doomed to Fail? *International Conference on Auditory Display* 52, 327–330. <https://doi.org/10.21785/icad2019.069>
- [56] Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305* (2015).
- [57] Revanth Reddy, Rahul Ramesh, Ameer Deshpande, and Mitesh M Khapra. 2019. FigureNet: A deep learning model for question-answering on scientific plots. In *2019 International Joint Conference on Neural Networks (IJCNN)* (Budapest, Hungary). IEEE, 1–8.
- [58] Melanie A Revilla, Willem E Saris, and Jon A Krosnick. 2014. Choosing the number of categories in agree–disagree scales. *Sociological methods & research* 43, 1 (2014), 73–97.
- [59] L Penny Rosenblum, Li Cheng, Kim Zebeha, Robert Wall Emerson, and Carole R Beal. 2020. Teachers’ descriptions of mathematics graphics for students with visual impairments: A preliminary investigation. *Journal of Visual Impairment & Blindness* 114, 3 (2020), 231–236.
- [60] Doug Schepers. 2022. <http://diagramcenter.org/the-future-of-accessible-charts.html>.
- [61] 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*. ACM, New York, NY, USA, 1–16.
- [62] Ather Sharif and Babak Fourougrahi. 2018. evoGraphs – A jQuery plugin to create web accessible graphs. In *2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. 1–4. <https://doi.org/10.1109/CCNC.2018.8319239>
- [63] Ather Sharif, Olivia H. Wang, and Alida T. Muongchan. 2022. “What Makes Sonification User-Friendly?” Exploring Usability and User-Friendliness of Sonified Responses. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) (ASSETS '22). Association for Computing Machinery, New York, NY, USA, Article 45, 5 pages. <https://doi.org/10.1145/3517428.3550360>
- [64] 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 *Proceedings of the 2022 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>
- [65] Ather Sharif, Andrew Mingwei 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 *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) (ASSETS '22). Association for Computing Machinery, New York, NY, USA, Article 61, 5 pages. <https://doi.org/10.1145/3517428.3550363>
- [66] Monika Sharma, Shikha Gupta, Arindam Chowdhury, and Lovekesh Vig. 2019. Chartnet: Visual reasoning over statistical charts using mac-networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–7.
- [67] Leixian Shen, Enya Shen, Yuyu Luo, Xiaocong Yang, Xuming Hu, Xiongshuai Zhang, Zhiwei Tai, and Jianmin Wang. 2022. Towards Natural Language Interfaces for Data Visualization: A Survey. *IEEE Transactions on Visualization and Computer Graphics* (2022), 1–1. <https://doi.org/10.1109/TVCG.2022.3148007>
- [68] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. 2021. Calliope: Automatic Visual Data Story Generation from a Spreadsheet. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (Feb. 2021), 453–463. <https://doi.org/10.1109/TVCG.2020.3030403> Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- [69] Hrituraj Singh and Sumit Shekhar. 2020. STL-CQA: Structure-based Transformers with Localization and Encoding for Chart Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 3275–3284. <https://doi.org/10.18653/v1/2020.emnlp-main.264>
- [70] Alexa F. Siu, Danyang Fan, Gene S-H Kim, Hrishikesh V. Rao, Xavier Vazquez, Sile O’Modhrain, and Sean Follmer. 2021. COVID-19 highlights the issues facing blind and visually impaired people in accessing data on the web. In *Proceedings of the 18th International Web for All Conference (W4A '21)*. Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3430263.3452432>
- [71] Arjun Srinivasan and Vidya Setlur. 2021. Snowy: Recommending Utterances for Conversational Visual Analysis. 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, 864–880. <https://doi.org/10.1145/3472749.3474792>
- [72] Abigale Stangl, Nitin Verma, Kenneth R Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*.

- ACM, New York, NY, USA, 1–15.
- [73] Huan Sun, Hao Ma, Xiaodong He, Wen-tau Yih, Yu Su, and Xifeng Yan. 2016. Table cell search for question answering. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 771–782.
- [74] Visa Chart 2022. Visa Chart Components. <https://github.com/visa/visa-chart-components>
- [75] Data visualization guidelines CFPB Design System. 2020. CFPB Design System. <https://cfpb.github.io/design-system/guidelines/data-visualization-guidelines>.
- [76] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W White. 2019. Verse: Bridging screen readers and voice assistants for enhanced eyes-free web search. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, New York, NY, USA, 414–426.
- [77] B. N. Walker and L. M. Mauney. 2010. Universal design of auditory graphs: A comparison of sonification mappings for visually impaired and sighted listeners. *ACM Transactions on Accessible Computing* 2, 3 (2010), 1–16. <https://doi.org/10.1145/1714458.1714459>
- [78] Steven Wall and Stephen Brewster. 2006. Feeling what you hear: tactile feedback for navigation of audio graphs. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, New York, NY, USA, 1123–1132.
- [79] David L. Waltz. 1978. An English language question answering system for a large relational database. *Commun. ACM* 21, 7 (1978), 526–539.
- [80] Ruobin Wang, Crescentia Jung, and Yea-Seul Kim. 2022. Seeing Through Sounds: Mapping Auditory Dimensions to Data and Charts for People with Visual Impairments. *Computer Graphics Forum* 41, 3 (2022), 71–83. <https://doi.org/10.1111/cgf.14523>
- [81] WebAIM 2018. WebAIM: Survey of Users with Low Vision #2 Results. <https://webaim.org/projects/lowvisionsurvey2/#types>
- [82] Dunwei Wen, John Cuzzola, Lorna Brown, and Dr Kinshuk. 2012. Instructor-aided asynchronous question answering system for online education and distance learning. *International Review of Research in Open and Distributed Learning* 13, 5 (2012), 102–125.
- [83] Markus Weninger, Gerald Ortner, Tobias Hahn, Olaf Drümmer, and Klaus Miesenberger. 2015. ASVG Accessible Scalable Vector Graphics: intention trees to make charts more accessible and usable. *Journal of Assistive Technologies* 9, 4 (Jan. 2015), 239–246. <https://doi.org/10.1108/JAT-10-2015-0027> Publisher: Emerald Group Publishing Limited.
- [84] Ericksonm William, Camille Lee, and Sarah von Schrader. [n. d.]. Disability Statistics from the 2011 American Community Survey (ACS). www.disabilitystatistics.org. Accessed June 1, 2021.
- [85] WolframAlpha 2022. WolframAlpha. <https://www.wolframalpha.com/>
- [86] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* 163 (2017), 21–40.
- [87] Hongyang Xue, Zhou Zhao, and Deng Cai. 2017. Unifying the video and question attentions for open-ended video question answering. *IEEE Transactions on Image Processing* 26, 12 (2017), 5656–5666.
- [88] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718* (2019).
- [89] Yalong Yang, Kim Marriott, Matthew Butler, Gagatay Goncu, and Leona Holloway. 2020. Tactile presentation of network data: Text, matrix or diagram?. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12.
- [90] Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning. *CoRR abs/1709.00103* (2017). arXiv:1709.00103 <http://arxiv.org/abs/1709.00103>
- [91] Jonathan Zong, Crystal Lee, Alan Lundgard, JiWoong Jang, Daniel Hajas, and Arvind Satyanarayan. 2022. Rich Screen Reader Experiences for Accessible Data Visualization. *arXiv preprint arXiv:2205.04917* (2022).