DOI: 10.1111/cgf.14573
EUROVIS 2022
S. Bruckner, C. Turkay, and K. Vrotsou
(Guest Editors)

Volume 41 (2022), Number 3
STAR – State of The Art Report

Chart Question Answering: State of the Art and Future Directions

E. Hoque[1] and P. Kavehzadeh[1] and A. Masry[1]

[1]Intelligent Visualization Lab, York University, Toronto, Canada

Abstract

Information visualizations such as bar charts and line charts are very common for analyzing data and discovering critical insights. Often people analyze charts to answer questions that they have in mind. Answering such questions can be challenging as they often require a significant amount of perceptual and cognitive effort. Chart Question Answering (CQA) systems typically take a chart and a natural language question as input and automatically generate the answer to facilitate visual data analysis. Over the last few years, there has been a growing body of literature on the task of CQA. In this survey, we systematically review the current state-of-the-art research focusing on the problem of chart question answering. We provide a taxonomy by identifying several important dimensions of the problem domain including possible inputs and outputs of the task and discuss the advantages and limitations of proposed solutions. We then summarize various evaluation techniques used in the surveyed papers. Finally, we outline the open challenges and future research opportunities related to chart question answering.

CCS Concepts

• ***Human-centered computing*** → ***Visualization;*** • ***Computing methodologies*** → ***Natural language processing;***

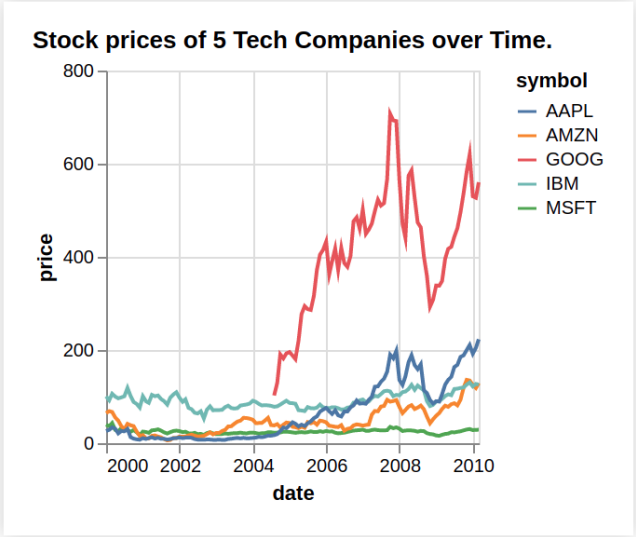
1. Introduction

Information visualizations such as bar charts and line charts are commonly used for analyzing data and making informed decisions. To analyze data, often people ask various complex questions about charts [KHA20]. However, answering such questions about charts is not always easy. In order to answer complex questions, one must have various analytical skills and would need to combine various low-level operations (e.g. retrieve values from bars, find extremes, aggregate values) which can be mentally taxing. For example, to answer the question *"When the difference between the Apple and Google stocks was the highest?"* in Figure 1, one needs to compute the differences between all data points of two red and blue lines and then find the highest one.

The goal of a chart question answering system is to automatically answer a natural language question about a chart to facilitate visual data analysis. The CQA problem belongs to the area of natural language interfaces (NLI) for visualizations, which has recently received a lot of attention in the research community [HSTD17, SS18] as well as in the industry [ask, FR16]. Unlike traditional approaches which typically support visual data analysis using mouse-based interactions, NLIs allow people to express their complex information needs easily through text or speech, thus lowering the threshold of analytical skills required for analyzing data. Moreover, it can significantly reduce the time and mental efforts. It is well known that performing elementary perceptual tasks like judging the length or area of a chart is not always easy [CM85]. Similarly, performing various arithmetic and logical operations such as comparisons and aggregations involve cognitive activities. Automatic question answering can perform these operations on behalf of users to reduce cognitive overload. Finally, CQA can enhance chart accessibility, where blind people can comprehend charts by asking questions. For example, blind people can get an overview of a chart using caption and data sonification and then they can ask simple questions [SWM*22] to further understand the given chart.

Automatic chart question answering is a challenging task because of the richness and ambiguities of natural language and complex reasoning that may be required to predict the answer [SBT*16]. The problem is also very inter-disciplinary in nature, falling in the intersection of information visualization, natural language processing, and human computer interactions. Some early approaches on NLIs and chart question answering applied natural language processing techniques by largely depending on heuristics or grammar-based parsing techniques [SBT*16, SS18, HSTD17, GDA*15]. Recently, there have been also some attempts for applying deep learning models for understanding natural language queries about visualizations [CSG*20, SS20a, RRDK19]. Several benchmark datasets have also been developed to evaluate the effectiveness of these approaches [KPCK18, KMA*18, CSG*20].

While there is a growing body of literature in the space of NLIs for visualizations in general and chart question answering in particular, there has not been any comprehensive survey on the topic of CQA. Some initial efforts summarize the prior works and highlight future challenges [SS20b, SSL*21], however, they cover NLIs



Question: ***When the difference between the Apple and Google stocks was the highest?***
Answer: ***2008***

Figure 1: *An example of a natural language question about a line chart that shows stock prices of some tech companies over time.*

for visualizations broadly without focusing on the specific chart question answering task. There also exist survey papers on question answering in other domains such as question answering on text data [ZLW*21], image with scenes and objects [HYY21, ZCW19] and knowledge base [LHJ*21]. However, to our knowledge there is no survey paper specifically focusing on chart question answering. This survey aims to fill that gap by systematically reviewing the state of the art on CQA and introducing a taxonomy of the problem.

In this paper, we present a formal taxonomy of the chart question answering task and categorize the existing solutions and evaluation techniques as outlined in Section 2. To narrow down the scope of the survey, we only focus on research works that take a question about a chart as input and produce the answer as output. We consider chart question answering as a sub-problem of developing NLIs ([SBT*16, SS18]), where NLIs focus on a broader range of tasks (e.g., manipulating visualizations through natural language commands, web-search like short queries) in addition to question answering [SNL*21]. Chart question answering is also related to natural language generation (NLG) for visualizations [OH20] and narrative storytelling as a CQA system may answer a question by generating visualizations and texts [TRB*18]. In order to describe the problem and the design space of proposed solutions, we identify several important dimensions of the problem domain including possible inputs and outputs of the task (see Figure 1). We also categorize the existing solutions and grouped the evaluation techniques used in the surveyed papers. Finally, we outline the open challenges and future research opportunities in this domain.

2. Structure of the Paper and Outline

To comprehensively survey the CQA domain and develop a taxonomy, we carried out exhaustive searches through Google Scholar using key phrases such as "chart question answering", "figure question answering", "natural language interfaces for visualizations", and "question answering with data visualizations". In addition, we browse through publications from the venues related to information Visualization and natural language processing fields in the last five years. We then reviewed the relevant papers and applied an iterative coding approach to discover the main categories which helped us to characterize the problem space. In particular, we analyze the CQA problem across the Input and output dimensions and review existing Evaluation techniques. Figure 1 visually summarizes the design space of the CQA problem along with major categories that we identified through our survey. Below we summarize the structure of the survey inspired by this categorization scheme.

(1) Inputs: Chart question answering systems may take a variety of visualization, text, and multimodal inputs as illustrated in Figure 1 (left). Most papers we reviewed on the CQA task considered that the system takes a visualization and a natural language question about it as input [CSG*20, SS20a, KPCK18, KSC*20, KHA20, RRDK19, MGKK19]. There are also some works on natural language interfaces that enable *multimodal* inputs by combining touch, speech and other modalities (e.g. Orko [SS18, SLHR*20]). The given question can be categorized in various ways, for example, based on complexity (e.g. simple vs. complex), or whether it refers to visual attributes of graphical marks in a chart (e.g. visual vs. non-visual). Similarly, charts can be presented in different formats (e.g. bitmap image vs. a SVG chart with access to underlying data table) as well as with various types (e.g. bar charts, line charts). When the underlying data table is available, the input chart can be presented in a declarative specifications such as using Vega-Lite [SMWH16]. Another possible problem variation could take multiple views as input with several underlying data tables, possibly created by visualization recommendation systems [HBL*19]. In Section 3, we will dive deep into the methods used for processing each input type in a CQA system.

(2) Outputs: Like inputs, outputs of a CQA system can be presented in different forms as shown in Figure 1 (right). Most CQA systems output textual answers to the given query but they can be characterized into different types depending on whether the answer comes from a fixed vocabulary with limited possible answers like 'yes' and 'no' (e.g., [KMA*18, KPCK18, CSG*20]) or from an open vocabulary with various possible answers (e.g., [MGKK19, KHA20]). Some other natural language interfaces produce a visualization as output [GDA*15] or highlight answers in an existing visualizations [SBT*16] or even combine both visualization and text/audio as multimodal output [SS18]. Some natural language interfaces also produce a response to the current question in the context of the previous questions asked by the user. Moving beyond the single query-response paradigm, these systems show the advantage of improving the flow of analytical conversation. We will discuss different possible types of output in CQA in Section 4.

(3) Evaluation: In recent years, researchers have released several benchmark datasets to evaluate the performance of the question answering [KMA*18, KPCK18, MGKK19] and NLI [LTL*21a,

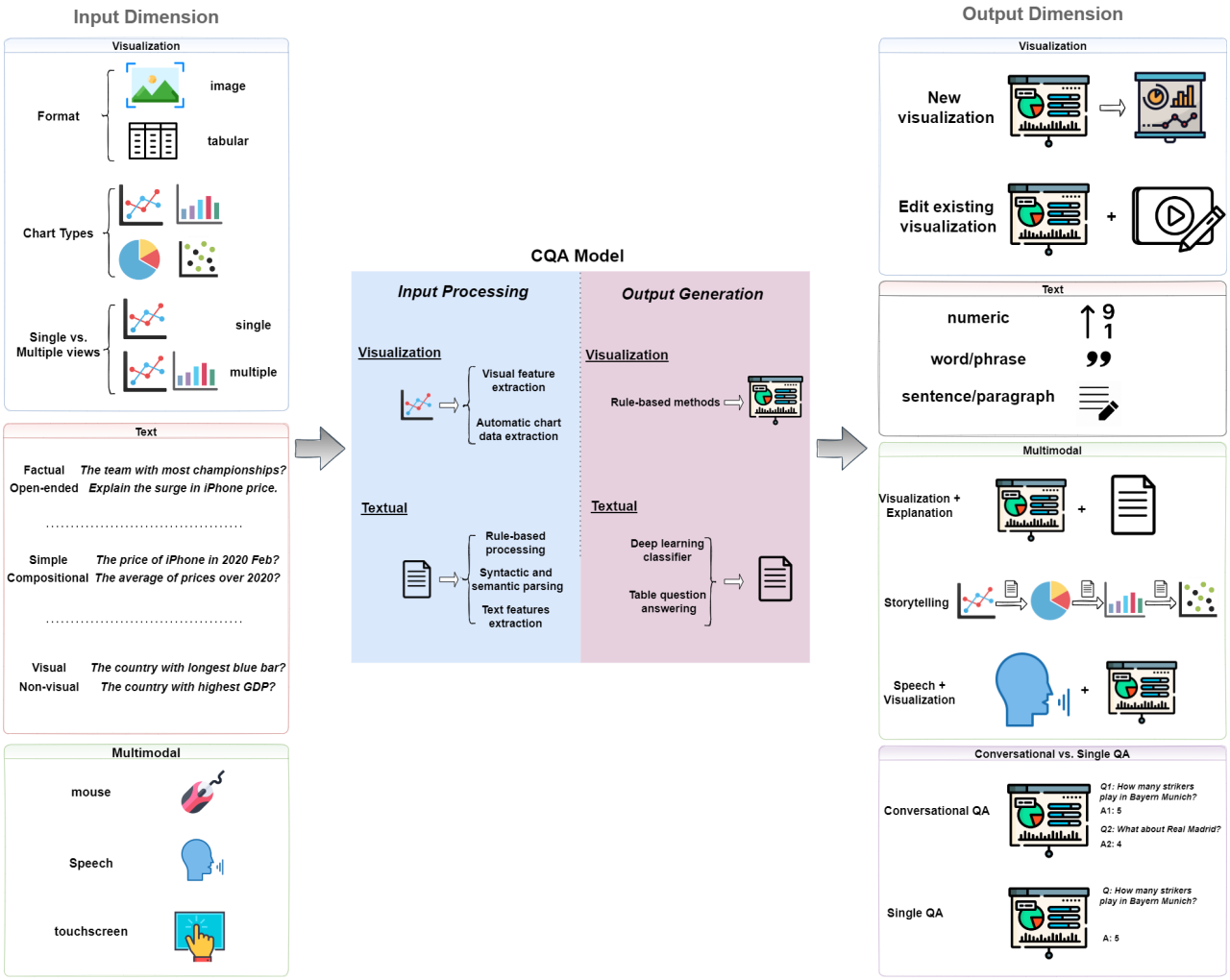


Figure 2: *An overview of the problem space of chart question answering, covering key categories of input and output dimensions* . A particular CQA problem setup may involve one or more categories of *input* and one or more categories of *output* dimensions.

YZY*18] systems. Other works have focused on conducting user studies to measure the performance and subjective feedback from participants to understand the effectiveness and limitations of different prototypes (e.g., [SBT*16,SS18]). In section 5, we will critically review these benchmark datasets and user study methods with respect to input and output dimensions to highlight the challenges and future directions in evaluating chart question answering systems.

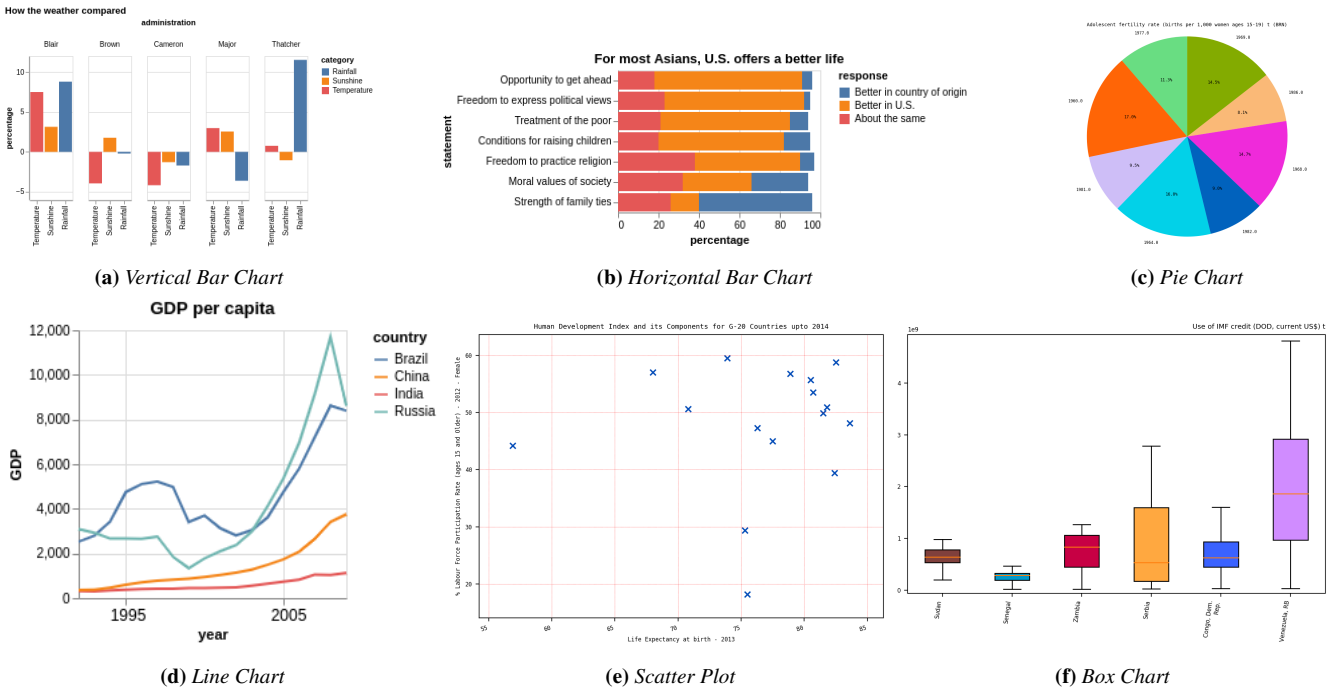
3. Input Dimension

In this section, we discuss the possible input dimensions of a CQA system including *Text*, *Visualization*, and *multimudal* inputs and review how existing research works analyze these inputs to answer the question.

3.1. Textual

In chart question answering, people express their information needs through a variety of textual queries as shown in Table 1. Note that these categories are not orthogonal; for example, a question can be factual, visual, and compositional simultaneously. We discuss the different types of textual queries along with how current CQA systems handle them below.

• **Factual vs. Open-ended** Most existing works focus on answering questions that require **factual** answers (e.g. *"Which country has the highest GDP?"*) [CSG*20, KHA20, RRDK19, KPCK18]. Factual questions require the system to compute the answer by analyzing the question and the chart and then subsequently performing some arithmetic and logical operations (e.g., compare values, find extremes). Several works have attempted to understand and analyze such questions using a combination of heuristic approaches and syntactic parsing techniques [SBT*16,SS18,NSS20]. Others have focused on leveraging various machine learning ap-



(a) *Vertical Bar Chart* **(b)** *Horizontal Bar Chart* **(c)** *Pie Chart*

(d) *Line Chart* **(e)** *Scatter Plot* **(f)** *Box Chart*

Figure 3: *Examples of different types of input charts (plots a, b, d are adapted from [KHA20] and plots c, e, f are adapted from [CSG*20])*

Table 1: *Different types of question inputs and possible answers. Each question is referring to a particular chart in Figure 3, as specified at the end of the question.*

Question Types	Answer Types				
	Fixed Vocabulary		Open Vocabulary		
	Numeric	Word(s)/Phrase(s)	Numeric	Word(s)/Phrase(s)	Explanatory Sentence(s)
Factual	How many years have fertility rate more than 11 percent? 3c	Which years have more than 11 percent of fertility? 3c	How much is life expectancy's mean bigger than labour force's? 3e	To which continent Venezuela belongs to? 3f	-
Open-ended	-	-	-	-	Explain the surge of Russia's GDP in the end of 1990's. 3d
Visual	What is the median of smallest box? 3f	Which administration has the longest blue component?3a	What is the ratio of the longest orange bar to the shortest? 3a	To which continent the red colored box belong to? 3f	Why Thatcher has the longest blue component? 3a
Non-visual	What is the median of Sudan? 3f	Which statement satisfies Asians most regarding living in U.S.? 3b	What is the ratio of Sudan's median to Serbia's? 3f	What is the population of the country with highest gdp? 3d	Why higher life expectancy included higher female labour? 3e
Simple	What is the percentage of fertility in 1982? 3c	Which country experienced the gdp higher than 10,000? 3d	What is the percentage of rainfall in Blair? 3a	Was the blue line increasing or decreasing from 2000 to 2005? 3d	What is the definition of GDP? 3d
Compositional	What is the overall percentage of fertility in 1982 and 1968? 3c	Which country has the highest median? 3f	What is the mean of total life expectancy rates? 3e	In which continent is the country with the highest GDP last year? 3d	Why is China's GDP less than Brazil's? 3d

proaches for table question answering [KHA20, MH21] and feature extraction from chart images [CSG*20, RRDK19, KPCK18]. For example, several research works utilize the recurrent neural model (RNN) with Long short-term memory (LSTM) architecture to encode the question [CSG*20, RRDK19, MGKK19, KSC*20, ZWXW20, KPCK18]. However, more recently transformer architecture has been found to be more effective than RNN for long input sequences, which inspired some researchers to adopt such architecture for answering factual questions with charts [SS20a, MH21, MLT*22, LHJY21]. Unlike factual questions, **open-ended** questions are exploratory in nature (e.g., *"Why the country with the highest GDP changed in 2016?"*) and typically the expected answer is an explanatory text. Generating such explanatory answers is very challenging and to our knowledge, there is no existing work that explores this problem. The closest works to this problem are the ones that automatically summarize the key insights from charts as text [DCM12, SC20, OH20, SRTKX*22]. For example, Obeid and Hoque [OH20] adopted a transformer-based model to generate an explanatory text describing the chart. Future works may explore such data-to-text generation approaches by taking the question as additional input.

• **Visual vs. Non-visual** Visual questions include reference to visual attributes such as *color*, *height*, and *length* of graphical marks (e.g., *bars*) in the chart. In contrast, non-visual questions do not

Table 2: *The table summarizes how existing research papers on CQA cover different categories of input and output dimensions. In the left most column, the papers coded with blue color represent those works that solely focus on CQA, while the ones with magenta color are the works that do cover some question answering aspects but focus more broadly on supporting natural language commands and web-search like queries for interactions with visualizations (such as filtering, sorting, zooming, and highlighting).*

Related Papers	Dimensions									
	Input					Output				
	Text			Visualization	Multimodal	Text		Visualization	Multimodal	Conversational
	Factual/ Open-ended	Visual/ Non-visual	Simple/ Compositional	Bitmap/ Data Table		Fixed Vocab Numeric/ Word	Open Vocab Numeric/ Word/ Sentence	New/ Existing		
LeafNet [CSG*20]	✓/✗	✓/✓	✓/✓	✓/✗	✗	✓/✓	✗/✗/✗	✗/✗	✗	✗
STL-CQA [SS20a]	✓/✗	✓/✓	✓/✓	✓/✗	✗	✓/✓	✗/✗/✗	✗/✗	✗	✗
DVQA [KPCK18]	✓/✗	✓/✓	✓/✓	✓/✓	✗	✓/✓	✗/✗/✗	✗/✗	✗	✗
PReFIL [KSC*20]	✓/✗	✓/✓	✓/✓	✓/✗	✗	✓/✓	✗/✗/✗	✗/✗	✗	✗
Kim et al. [KHA20]	✓/✗	✓/✓	✓/✓	✗/✓	✗	✓/✓	✓/✓/✓	✗/✗	✗	✗
FigureNet [RRDK19]	✓/✗	✓/✓	✓/✓	✓/✓	✗	✗/✓	✗/✗/✗	✗/✗	✗	✗
Affinity [ZWXW20]	✓/✗	✓/✓	✓/✓	✓/✗	✗	✓/✓	✗/✗/✗	✗/✗	✗	✗
FigureQA [KMA*18]	✓/✗	✓/✓	✓/✓	✓/✓	✗	✓/✓	✗/✗/✗	✗/✗	✗	✗
PlotQA [MGKK19]	✓/✗	✓/✓	✓/✓	✓/✗	✗	✓/✓	✓/✓/✗	✗/✗	✗	✗
ChartQA [MLT*22]	✓/✗	✓/✓	✓/✓	✓/✓	✗	✓/✓	✓/✓/✓	✗/✗	✗	✗
Eviza [SBT*16]	✓/✗	✗/✓	✓/✓	✗/✓	✓	✗/✗	✗/✗/✗	✗/✓	✗	✓
Evizeon [HSTD17]	✓/✗	✓/✓	✓/✓	✗/✓	✓	✗/✗	✗/✗/✗	✗/✓	✗	✓
DataTone [GDA*15]	✓/✗	✗/✓	✓/✓	✗/✓	✓	✗/✗	✗/✗/✗	✗/✓	✗	✗
Orko [SS18]	✓/✗	✗/✓	✓/✓	✗/✓	✓	✗/✗	✗/✗/✗	✗/✓	✓	✓
FlowSense [YS20]	✓/✗	✗/✓	✓/✓	✗/✓	✗	✗/✗	✗/✗/✗	✓/✓	✗	✓
NL4DV [NSS20]	✓/✗	✗/✓	✓/✓	✗/✓	✗	✗/✗	✗/✗/✗	✓/✗	✗	✗
ADVISor [LHJY21]	✓/✗	✗/✓	✗/✓	✗/✓	✗	✗/✗	✗/✗/✗	✓/✗	✗	✗
NL2VIS [LTL*21b]	✓/✗	✗/✓	✓/✓	✗/✓	✗	✗/✗	✗/✗/✗	✓/✗	✗	✗

contain such references. For instance, the question *"Which bar is the longest?"* is a visual question since it is referring to the *length* attribute of the mark type *bar*. Kim et al. [KHA20] handled the visual question by translating it to an equivalent non-visual question through a pipeline that first recognizes all phrases in the input question that refers to marks and their visual attributes and then replaces these phrases with equivalent non-visual terms. Then they pass the question and the data table of the chart to a table parsing model to obtain the answer. Others have extracted visual features from the chart and then build classification models to answer the visual questions [KMA*18, KPCK18].

• **Simple vs. Compositional** This categorization is based on the complexity of the analytical tasks [AES05] that the system needs to perform to answer the given question. To answer **simple** questions, the system needs to complete a single analytical task (e.g., retrieving a value). For instance, the question *"What percentage of people are Real Madrid's fan?"* is a simple question. On the other hand, **compositional** questions involve multiple mathematical/logical operations like *sum*, *difference* and *average*. Compositional questions are more challenging as the model needs to combine multiple operations together. For example, to answer this question "How many students achieved over 90%?" from a bar chart that shows the grades of students, the system needs to perform a *filtering* operation to find students who achieved over 90% followed by applying a *count* operation. To handle the compositional questions, some researchers have applied a compositional semantic parsing technique called Sempre [PL15] that was originally trained on a table question answering dataset [KHA20, YS20]. However, solving compositional questions still remains a challenging task as the accuracy is still very low for such questions. For example, Kim et al. [KHA20] found the accuracy of their model to be only 37% for compositional questions.

3.2. Visualization

The input visualization to the CQA system can have different types and storage format as shown in Figure 1. Below we discuss how current CQA methods process the input visualization.

• **Chart Types:** Most existing works on CQA focused on limited chart types. Bar charts [CSG*20, RRDK19, KSC*20, ZWXW20, MGKK19], Line charts [CSG*20, KSC*20, ZWXW20, MGKK19], Pie charts [CSG*20, RRDK19, KSC*20, ZWXW20], Box plots [CSG*20], Scatter plots [CSG*20, MGKK19] are most common chart types used in recent related works. Figure 3 shows examples of chart types. None of the CQA works in Table 2 (coded with blue color) supports map chart even though they are commonly used. However, some NLI systems such as Eviza support natural langauge queries with map charts [SBT*16]. Less common and unconventional chart types such as Parallel coordinates plot and radar charts are not supported in existing works.

• **Single vs. multiple views:** All the works on CQA listed in Table 2 focused on question answering with a single visualization while multiple coordinated views [JE12] and dashboards have rarely been considered [HSTD17, CMH*20]. When multiple views are introduced, there are several additional challenges that are introduced. For example, which view the user is referring in the question? How

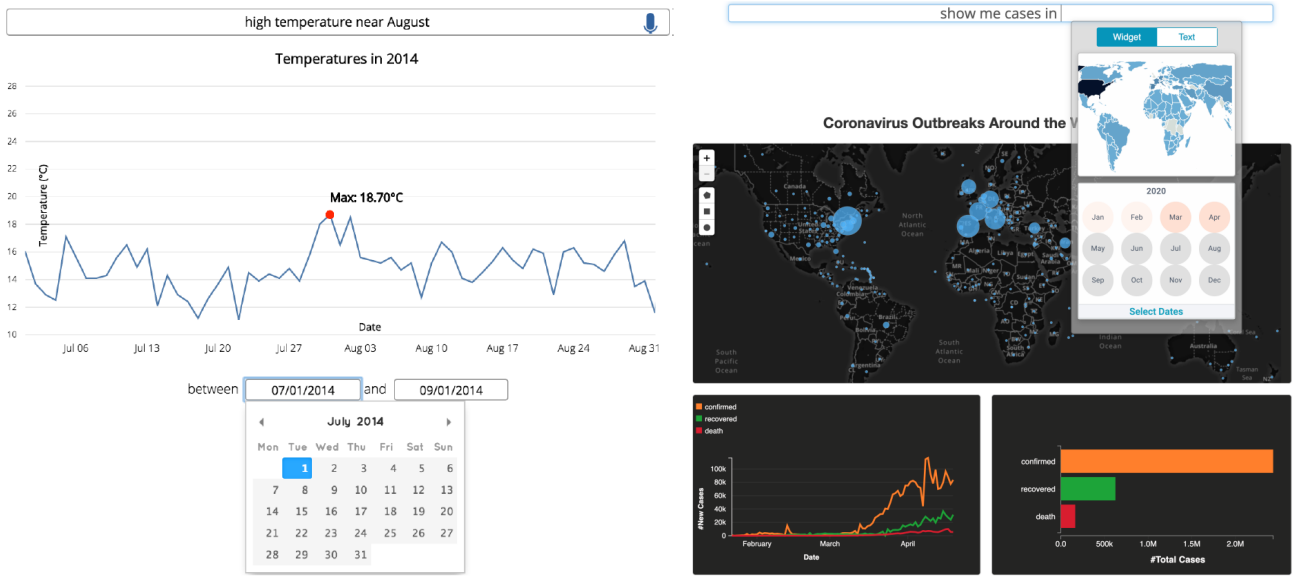


Figure 4: *Examples of how text and interactive visualization can be combined for inputting a query. In Figure 4a, the user types a vague query and then resolves the ambiguity by selecting more specific date from a calendar widget [SBT*16]. In Figure 4b, as the user types a partial query, two visualizations are popped up with previews of the dataset, thus supporting the user in completing the query [SHKC20].*

the system can combine reasoning over multiple charts? Such questions are not deeply investigated yet.

• **Storage Format:** Storage format is another important aspect of the input chart. Charts can be stored in different formats such as SVG vs. bitmap image. Several works take charts in bitmap image format which are more challenging to handle because the model does not have access to the underlying data table. They mainly utilized computer vision techniques to extract features from a bitmap chart image and then applied classification models [KMA*18,CSG*20,MGKK19,RRDK19,SS20a,KSC*20]. For example, FigureQA [KMA*18] extracts the features of the chart using a CNN. The problem with such CNN-based feature extraction is that the model treats the chart like a natural scene without extracting individual graphical marks (e.g., bars) and the underlying texts (e.g, x-axis-labels) in the chart. To overcome this limitation, others attempted to parse the chart by employing object detection models such as Mask-RCNN [HGDG17] to detect visual and textual elements [CSG*20, SS20a]. Subsequently, Optical Character Recognition (OCR) models are applied in order to dynamically encode the chart-specific tokens in the question in terms of the positional information of the textual elements in the chart image (e.g., 'x-axis-label-1', 'x-axis-label-2'). However, such dynamic encoding technique is prone to OCR errors and fails when the question refers to the chart texts using synonyms (e.g., 'US' vs 'United States').

One key question here is how the model should reason over both the question and the visualization together to generate the answer? FigureQA used Relation Network (RN) [SRB*17] to capture the relation between visual and textual features by concatenating all pairs of object representations from the chart image provided by a CNN with the question features [KMA*18]. However, the number of object pairs in RN can be very large, resulting in computational inefficiency and restricting the performance. Some researchers attempted to address the problem either by making relation features in RN more concise [ZWXW20] or by dividing the problem into various sub-tasks [RRDK19]. Others captured inter-relation between the chart and question features by applying attention mechanisms [CSG*20,KPCK18] or through fusing the low and high level features from the chart image and the question in parallel to facilitate multi-stage reasoning process. However, these models ignore the chart structure (e.g., relationships between axis labels, bars, and legends). STL-CQA [SS20a] attempted to better capture the relation between the chart element and the question using a multimodal transformer-based model, which can better exploit the structural properties by learning the intra-modality relationships (among the chart elements) and cross-modality relationships (between the question and the chart elements) [TB19].

A common limitation of all the above models is that they simply utilize the image features without extracting the underlying data table and visual encoding information of the given chart, which makes it difficult for their models to answer complex visual and compositional questions. Moreover, since they are classification based models, they could only handle fixed-vocabulary type answers and can not apply mathematical operations on the chart data.

In contrast to the above body of work, some systems simplify the problem by assuming that the underlying data and visual encodings of charts are available [SBT*16,SS18,HSTD17,LHJY21,NSS20]. Kim et al. [KHA20] propose a pipeline that recovers the data table and encodings from an input chart to derive the corresponding Vega-Lite specification [SMWH16]. It then extracts the data and transforms into a flat relational table and passes it to a ta-

ble question answering model. If the data can be extracted from the chart, leveraging the pre-trained transfer-based model for table question answering such as TaPas [HNM*20] or converting natural language question to SQL query [YZY*18] could be effective for CQA. However, directly applying table question answering method on chart data is not enough, rather it is necessary to effectively combine the chart features with the data table. Masry et al. [MLT*22] takes an initial step in this direction by extracting both the visual features extracted using ViT [DBK*21] and the underlying data table of the chart encoded by TaPas [HNM*20] and combining them using a cross modality encoder to infer the answer. Still, their model feeds the visual features and the data values separately to their model and does not relate between them. A promising direction here could be to create a better representation of the chart that combine and relate between the visual features and the data table is needed to solve complex visual reasoning questions.

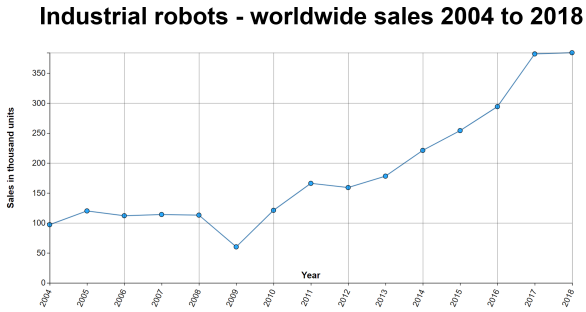


Figure 5: *Example outputs of an automatic chart summarization model. The textual summary attempts to explain important insights from the chart such as trends and extreme values [OH20].*

3.3. Multimodal

Multimodal interactions with visualizations have been actively explored with a focus on using touch and pen [WLJ*12], body movement in front of a large display [AEYN11], gestures [BAEI16], and coordinating between large displays and smartwatches [HBED18]. However, none of these works considered natural language as an input modality. On the other hand, many CQA systems only considered natural language question as input without considering other modalities like touch, gesture, and body movement (e.g., [KMA*18, KPCK18, CSG*20]).

Notable exceptions are some NLIs for visualizations which combine natural language with mouse/touch as input modalities [HSTD17, SS18, SLHR*20]. For example, Evizeon [HSTD17] and Orko [SS18] demonstrate multimodal interactions within a map chart and network diagram respectively. Eviza [SBT*16] and DataTone [GDA*15] demonstrate how combining natural language input with mouse-based input from ambiguity widgets can be helpful to resolve ambiguities and system's misunderstandings (see Figure 4a). Sneak Pique is another system that automatically pops up autocompletion widgets from which the user can select query criteria through direct manipulation in addition the textual input [SHKC20]. For example, in Figure 4b, as the user types a textual query *"show me case in"*, the system pops up a map and a calendar widgets that provide previews of data and allow users to click in the widgets to formulate the query.

Others have focused on multimodal interactions for tablet devices [KR18, SLHR*20]. For example, InChorus supports multimodal interactions with visualizations on tablet devices by combining pen, touch, and speech as input modalities [SLHR*20]. Saktheeswaran et al. [SSS20] ran a user study to confirm that compared to separate unimodal touch- and speech-based interfaces, an interface with both touch and speech modalities is preferred by participants because they felt it gave them more freedom to express the queries. Participants also felt comfortable with complementary of speech and touch modalities since whenever speech or touch was not sufficient, they could easily use another modality to illustrate what they want.

While the above body of work suggests that multimodal interaction offers strong promise, the potential challenges and opportunities in integrating natural language and other modalities have not been deeply investigated yet. In particular, it remains largely unknown how to synergistically integrate natural language interaction with other input modalities like touch, pen, and gestures in the context of different form factors ranging from large screen display, to desktop screen and mobile displays.

3.4. Discussion

In terms of the input dimension, there are multiple possible directions for future works. First, more advanced models can be designed to address the weaknesses of current models in comprehending complex and compositional questions about charts. Simple classification approaches or QA approaches for data tables usually lack the ability to handle most real-world questions. Using the combined information from both visual features and data table may overcome this issue. Second, most proposed works assumed that the underlying information of charts such as data table and visual encodings are already available; however, this assumption is not valid for most of the charts on Web that are stored in bitmap image format. Previously designed methods for extracting data values from chart images lack accuracy and devising more robust systems for chart data extraction can be helpful in improving CQA systems. Moreover, utilizing visual transformer models has been found to be effective for different computer vision tasks [LZW*21], employing such neural models can be a promising direction. Finally, existing approaches can only handle single visualization as input, while users tend to ask complicated questions about multiple charts (e.g. dashboard) in real-world scenarios. Designing models that are capable of taking multiple views as input can be an interesting direction. We will discuss these directions in detail in Section 6.

4. Output Dimension

As shown in Figure 1, similar to Input Dimension, outputs in CQA are shown in *Text*, *Visualization*, or *Multimodal*. We discuss each of these categories in detail.

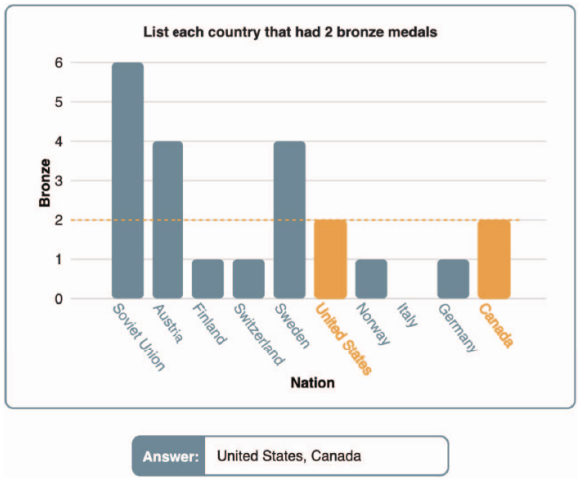


Figure 6: *An example of visualization as output from the Advisor system [LHJY21]. The query "List each country that had 2 bronze medals?" is asked regarding a data table. In response, the system returns the textual answer in the "Answer:" bar besides providing a visualization (bar chart) in which the number of bronze medals of each country is exhibited and the two countries that are answers of the query are highlighted [LHJY21]*

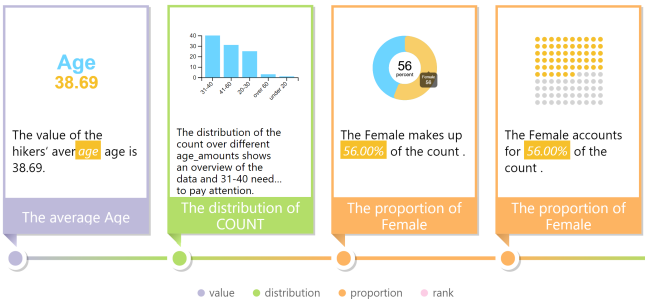


Figure 7: *An example output from the Calliope system which automatically generates a visual data story as a sequence of visualizations and texts [SXS*20]. Here the visual story summarizes data about hiking using a combination of multiple different visualizations including bar and pie charts.*

4.1. Textual

Most CQA models focus on answering questions that require a textual answer. Usually, such an answer is either a **numeric token** or a **word/phrase** [CSG*20, KPCK18, KMA*18, KSC*20]. Such textual outputs can be categorized into **fixed vocabulary** vs. **open vocabulary**. In case of the fixed vocabulary output setting, the answer comes from a small fixed-sized vocabulary (chart axis labels, bar values, 'yes', 'no'). For example, early systems outputs only `yes` or `no` to the questions about the chart [RRDK19, KMA*18]. Some systems support other possible output textual elements in chart and certain numeric values as they add them to the fixed vocabulary [KPCK18, KSC*20, CSG*20, SS20a, ZWXW20]. These models that only support *fixed vocabulary* questions usually treat the task as a classification problem and rely on dynamic encoding techniques where the questions and answers are encoded in terms of spatial positions of chart elements (*x-axis-label-1*, *x-axis-label-2* and so on). Such approaches do not produce correct outputs when the OCR model for extracting text from a chart generates errors or when the question refers to chart elements using synonyms (e.g., `US` vs. `United States`). Overall, treating the CQA problem as a classification task with a fixed vocabulary output is very limited for practical purposes, as it ignores many complex reasoning questions where the answer is derived through various mathematical operations such as aggregation and comparison.

Open vocabulary questions could be more challenging as the system no longer treats the problem as a classification task and instead needs to derive the answer through analytical operations. However, the works focusing on generating open vocabulary outputs are still very limited [KHA20, MGKK19, MH21, HSTD17]. Most of these works apply a table question answering model named Sempre [PL15]. However, Kim et al. simplify the problem by assuming that the underlying data table is available [KHA20] where others try to extract the data using computer vision techniques [MGKK19, MH21]. One limitation with PlotQA is that after automatically extracting the data table it directly applies the Sempre [PL15] while answering questions which does not consider any visual features of a chart. As a result their model accuracy can be low for visual questions that makes references to graphical marks and their attributes in chart.

Moving beyond outputting a word/phrase, an open-ended question that requires explanatory sentences or paragraphs has not been supported by existing works. For example, to answer the question, *"How has the GDP of Brazil changed over time?"* from the line chart Figure 3 (d), the output needs to be an explanatory paragraph summarizing the major trends. Recent works on automatic caption generation from charts (e.g., [OH20, SRTKX*22] can serve as a starting point for generating explanatory answers (see Figure 5). Another way a descriptive sentence/paragraph might be helpful is by explaining how the model computes the answer to enhance the transparency of the model to the user. Kim et al. take the intermediate logical form representation of the table question answering system and then translate it into an explanation using pre-defined templates by applying a rule-based approach [KHA20]. Overall, generating explanatory answers to a question about charts remains extremely under-explored.

4.2. Visualization

Some natural language interfaces for visualizations generate an answer by creating a new visualization [GDA*15, LHJY21, NSS20] based on the given question and a data table (see Figure 6). Others considered that a visualization is already given as input and the goal is to highlight answers within that existing visualization [SBT*16, SS18, SHKC20] (see the example in Figure 4a). In an early work, DataTone generates simple charts to answer questions about data tables using dependency parsing and a rule based approach [GDA*15]. Evizeon demonstrates a method for adding multiple visualizations using a hand-crafted grammar-based parsing approach. NL4DV [NSS20] is a Python toolkit that can take a question and a table as input and then output related data attributes,

analytical tasks and Vega-Lite specifications. This enables developers that may not be experts in natural language processing to automatically create the desired visualization using the specifications in the system output. The toolkit interprets the query using a combination of dependency parsing and a rule-based approach.

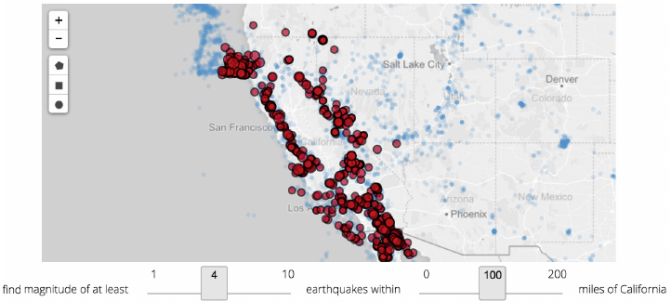
The above body of work generally relies on some heuristics which may put restrictions on possible user input, therefore more robust methods are needed to interpret natural language with rich syntactic and semantic structure. In this direction, Luo et al. [LTL*21b] utilize a transformer-based sequence-to-sequence (seq2seq) model that translates a natural language statement to visualization. Another system named ADVISor [LHJY21] attempts to address the problem by applying a deep learning model that used pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to encode the question and the table header and then feeding them to two fully-connected neural networks in order to determine the aggregation operation and related attributes. Finally, the model decides which type of visualization to show based on the aggregation operation and required attribute types extracted from the tabular data and the user's question. Figure 6 shows an example output of this approach.

One aspect that has been rarely covered in terms of outputting visualizations in CQA is presenting multiple views as an answer. Providing different views from the same dataset may help users to grasp different perspectives of the data [Rob98]. In this regard, combining automatic visualization recommendation [ZSJ*20, ZMD*21, WMA*15, VRS*22] with chart question answering can be a promising direction. Future work may explore how to generate multiple views as an answer to provide more perspective perhaps by incorporating visualization recommendation algorithms within the CQA system, especially when the question is open-ended in nature.

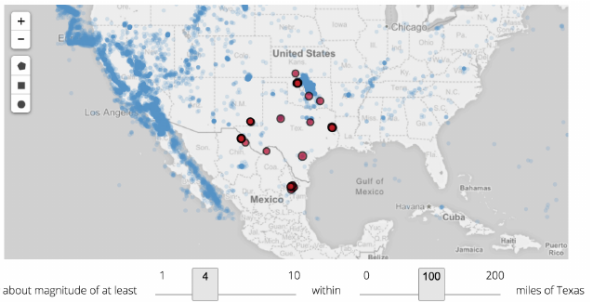
4.3. Multimodal

A multimedia output that leverages the complementary power of text and visualization could facilitate users to comprehend the answer more effectively than either visualizations or texts. Given the question *"How have the house prices in Toronto changed over time?"*, a line chart could show the average house price over time while the text could describe the price trends. However, this problem scenario has been virtually unstudied. For example, Orko [SS18] provides audio feedback respectively while responding to a query. But none of the existing CQA approaches explain the important patterns, trends, or outliers with respect to a query.

The multimodal output of a CQA system could also be presented as a sequence of scenes consisting of several visualizations and texts, thus resulting in a narrative story. There have been some recent efforts on automatically generating visual stories consisting of texts and visualizations [SXS*20, CWW*19, CZW*19, WSZ*19]. For example, Calliope explores the data space given by the input spreadsheet to generate data facts and organize them in a story sequence [SXS*20] (see Figure 7). Some works took images or charts as input and produce visual stories and interactive charts to elaborate the input in another perspective [TLW*20, ZXC*21, WMA*15]. There were also papers focusing on tabular data and spreadsheets [SXS*20, WSZ*19] or text



(a) Previous query: "Large earthquakes near California"



(b) Subsequent query: "how about near Texas?"

Figure 8: *An example of an analytical conversation regarding a visualization [SBT*16].*

data [MZJS18, HDA13, CZW*19, FBM15, Gov21] as input in timeline or story generation models. For instance, Cui et al. [CZW*19] designed a system that takes a simple factual statement with numeric data and outputs various possible infographics to illustrate the statement visually. However, these works do not take any questions as part of the input. While these works are not solving the chart question answering they can provide us with future directions on how to answer questions by combining visualizations with explanatory texts, which has not been investigated yet.

4.4. Conversation vs. Single QA pairs

People have tendencies in engaging in conversations involving a series of related question-answer pairs instead of a single question-answer pair. While the conversational question answering task has been explored in the domain of text data [RCM18, QYQ*19] as well as image data [DKG*17, MSDT18], it has not been explored deeply for chart question answering. Eviza took a first step towards supporting follow-up conversation [SBT*16], while Evizeon models pragmatic behaviour by executing the current query in the context of past interactions [HSTD17]. Figure 8 shows an example of a conversation in Eviza where the user asks a query *"Large earthquakes near California"* followed by *"how about near Texas?"*. In response, the system shows the large earthquakes near Texas based on the context from the last query. Orko [SS18] also attempted to handle follow-up questions regarding a visualization by keeping track of context through a conversational centering approach. For instance, to handle two follow-up questions *"Show the strikes of*

Real Madrid." and *"Now show the defenders."*, the model keeps *"Real Madrid"* as the center to use it in the following questions. Heart and Tory [HT19] conducted user studies to evaluate the influence of showing charts and graphs in the context of a conversational interface. They found that most of the people who wanted to see charts as part of analytical conversation preferred to see additional supporting context beyond the direct answer to the question.

Overall, the conversational question answering is a promising direction as existing works demonstrate the benefits of enabling pragmatics in improving analytical workflow while exploring visualization. However, these investigations are preliminary and have not been tested on any benchmark datasets. In future, it is necessary to build large-scale benchmark datasets to systematically evaluate different methods using both quantitative and qualitative measures. Moreover, advanced sequence-to-sequence deep learning method may capture the complex pragmatic aspects of the conversational interfaces more effectively than simple conversational centering [HSTD17, SS18] or finite state machine based approach [SBT*16].

4.5. Discussion

In this section, we identified several key limitations and future opportunities along the output dimensions. First of all, many CQA systems are limited to outputting textual tokens with a fixed vocabulary. Future work may explore how to effectively solve open-vocabulary questions and open-ended questions that require explanatory answers. Second, using both texts and visualizations together as the output of the CQA model could contribute to the comprehensiveness and transparency of the answers for users' queries. Third, answering a question by generating multiple views or by generating a visual story can be an interesting direction. Section 6 explains more details about the possible challenges and directions for the output dimension of CQA systems.

5. Evaluation

In this section, we discuss the various evaluation techniques for chart question answering. We group the evaluation methods used in the surveyed papers based on the major dimensions of the CQA problem space: (1) Input Dimension, (2) Output Dimension.

5.1. Input Dimension

5.1.1. Text

In terms of the text as input dimension, most benchmarks contain questions about charts that are created using pre-defined templates (see Table 3). For example, earlier CQA datasets, namely FigureQA [KMA*18] and DVQA [KPCK18], consisted of questions generated from a small number of templates. Such template-based questions lack the lexical/syntactic variations and nuances/erratum that are prevalent in the human-written questions. For example, the question in Figure 9a can be written as *''Is the area colored in orange smaller than the green area?"*. Although LeafQA [CSG*20] and LeafQA++ [SS20a] applied paraphrasing tools to introduce lexical variations into their questions, they still do not stand on par with the human-authored questions. PlotQA attempted to alleviate the issue through by gathering a limited number of human-authored questions through crowdsourcing to create more realistic templates for the questions in their datasets [MGKK19].

Lack of real-world human-annotated questions remains a key bottleneck in solving the chart question answering question. Kim et al. [KHA20] ran a formative study with a very small human-authored dataset consisting of 52 charts and 629 QA pairs to understand how people ask questions about charts and explain answers. Srinivasan et al. [SNL*21] runs another study to curate a dataset of 893 utterances that are used to generate visualizations, out of which 114 were questions while the remaining ones include natural language commands and web search-like phrasal queries. However, both of these datasets are quite small and were curated for understanding how people interact with visualizations through natural language rather than utilizing for training models. Masry et al. [MLT*22] attempted to addressed the limitation by introducing a large number of human written questions (9.6K). Since collecting human-authored questions is time-consuming and costly, they have augmented their dataset with 23.1K machine-generated questions from the Statista chart human-written summaries using the T5 model [RSR*19]. While such data augmentation through machine-generated question is promising, they are also limited in various ways (e.g, unanswerable questions, lack of visual reasoning questions that refer to chart elements).

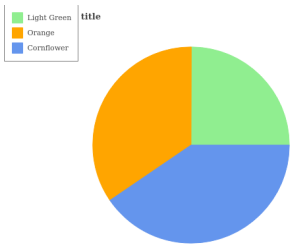
5.1.2. Visualization

When we consider visualization as input, most benchmark datasets constructed charts synthetically using randomly generated data. Among them, FigureQA [KMA*18] covers five different **chart types** (e.g., bar, line, and pie) while DVQA [KPCK18] contains only bar charts. All benchmarks except [KHA20] contain **bitmap images** of charts as input visualizations. FigureQA [KMA*18], DVQA [KPCK18], and Masry et al. [MLT*22] provide **underlying data** of input charts alongside the bitmap images. Kim et al [KHA20] only consider underlying data table as the input chart to CQA model. To generate the charts' textual labels (e.g., *x-axis* labels) synthetically, FigureQA utilized the X11 color set and DVQA [KPCK18] used the most common 1000 vocabulary from the Brown Corpus. While such approaches can help to develop a large dataset automatically, the randomly generated data and textual labels caused the charts and the questions to lack realism and meaningfulness. For example, Figure 9b shows a bar chart in the DVQA dataset with three completely unrelated categorical labels ('silver', 'month', and 'reason') along with the corresponding question *"Did the item month sold less units than reason?"*, which sounds unnatural and semantically meaningless. LeafQA [CSG*20] and LeafQA++ [SS20a] datasets entail new chart types such as box and scatter plots. Unlike the previous datasets, they utilized real-world tabular data extracted from online sources to synthetically plot their chart images.

Overall, most of the above datasets suffer from several common issues due to their synthetic nature. First, synthetically plotted charts lack the diversity in visual styles and charts' structure that are available in real-world charts. Although some works like PlotQA tried to randomly set visual parameters (e.g., font size, presence of grids), the charts visual styles were still limited due to using only

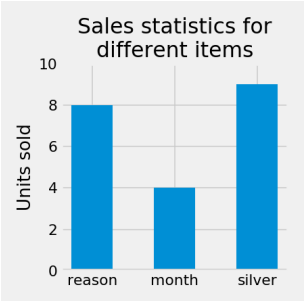
Table 3: *Comparisons among some existing datasets for chart question answering. The top five datasets contain automatically generated questions through template-based approach and were developed for the purpose of training and evaluating CQA models. The last two rows introduce datasets with human-written questions.*

Related Papers	Dimensions												#Charts/ #QA pairs
	Input								Output				
	Text				Visualization				Text		Visualization		
	Factual/ Open-ended	Visual/ Non-visual	Simple/ Compositional	Question Types	Bitmap/ Data Table	Real-world Charts	#Chart Types	Real-world Data	Fixed Vocab Numeric/ Word	Open Vocab Numeric/ Word/ Sentence	New/ Existing		
LEAFQA [CSG*20]	✓/✗	✓/✓	✓/✓	Template Based	✓/✗	✗	6	✓	✓/✓	✗/✗/✗	✗/✗		240K/2M
LEAFQA++ [SS20a]	✓/✗	✓/✓	✓/✓	Template Based	✓/✗	✗	6	✓	✓/✓	✗/✗/✗	✗/✗		244K/2.5M
DVQA [KPCK18]	✓/✗	✓/✓	✓/✓	Template Based	✓/✓	✗	1	✗	✓/✓	✗/✗/✗	✗/✗		300K/3.4M
PlotQA [MGKK19]	✓/✗	✓/✓	✓/✓	Template Based	✓/✗	✗	3	✓	✓/✓	✓/✓/✗	✗/✗		224K/28M
FigureQA [KMA*18]	✓/✗	✓/✓	✓/✓	Template Based	✓/✓	✗	4	✗	✓/✓	✗/✗/✗	✗/✗		180K/2.3M
Kim et al. [KHA20]	✓/✗	✓/✓	✓/✓	Human Authored	✗/✓	✓	2	✓	✓/✓	✓/✓/✓	✗/✗		52/629
Masry et al. [MLT*22]	✓/✗	✓/✓	✓/✓	Human Authored	✓/✓	✓	3	✓	✓/✓	✓/✓/✗	✗/✗		4.8K/9.6K



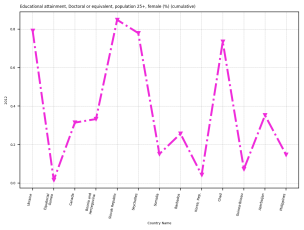
Q: Is Orange less than Light Green?
A: No

(a) *FigureQA Pie Chart [KMA*18]*



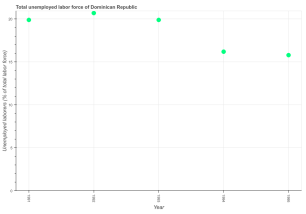
Q: Did the item month sold less units than reason?
A: Yes

(b) *DVQA Bar Chart [KPCK18]*



Q: Indicate the Country Name that has the Educational attainment, Doctoral or equivalent, population 25+, female (%) (cumulative) of 2012 less than the year Guinea-Bissau?
A: Korea, Rep.

(c) *LeafQA Line Chart [CSG*20]*



Q: What is the difference between the total unemployed labour force in 1993 and 1994 ?
A: 3.7

(d) *PlotQA dot line chart [MGKK19]*

Figure 9: *Some example charts from existing CQA datasets along with some of their QA pairs. Due to using randomly-generated data, the questions in 9a and 9b lack naturalness. 9c shows a line chart from LeafQA [CSG*20] where the x-axis consists of categorical variables that gives incorrect sense of value change. 9d shows a dot line chart from the PlotQA [MGKK19] dataset with open-vocabulary question.*

one software tool, namely Matplotlib †. Moreover, data attributes were plotted on charts without considering their types, resulting in poorly designed charts. For example, it is well known that plotting categorical data on a line chart is a bad idea [Mun14]. However, Figure 9c shows a line chart from LeafQA with categorical attributes and gives a false sense of perceptual change of values from one country to another.

Very recently, Masry et al. [MLT*22] constructed the ChartQA dataset, consisting of 4.8K real-world charts. To ensure variations in the visual styles, they crawled chart images from four different online sources. Still, their dataset is relatively small compared to the previous large-scale synthetic datasets and only covers three different chart types. Therefore, there is a pressing need for benchmarks covering a large collection of charts with diverse styles.

† https://matplotlib.org/

5.1.3. Multimodal

While all the benchmark CQA datasets take only textual query as input, research works related to more broad topics about natural language interfaces for visualizations conducted user studies to evaluate multimodal input features (e.g. [SS18, SBT*16, SLHR*20]). In these approaches, experimenters ask participants to have interactions with the designed system (or multiple systems in case of comparison). By analyzing some quantitative measurements (e.g., task completion time) and subjective feedback from participants through questionnaires, interviews regarding their interactions, experimenters gained both quantitative and qualitative insights about the system. Eviza [SBT*16] followed two goals of (1) observing qualitative feedback, and (2) comparing NLI and direct manipulation (DM). For qualitative measurement, they followed the think-aloud protocol where participants explained their thoughts during the tasks. To compare natural language interface

with a direct manipulation-based interface, they designed the tasks on both their system and Tableau Desktop. Orko [SS18] did the same about qualitative measurement by asking participants to express their thoughts. They also evaluated the multimodal interaction in their system. Orko [SS18] and Inchorus [SLHR*20] used the *jeopardy evaluation* method which allows experimenters to give participants some *facts* instead of explicit queries. In this way, participants have to interact with the system in a way that they can show the fact by visualization output. Finally, participants were given a questionnaire to tell their opinions and score the system based on different measurements.

In general, many of these research works are limited in that their systems are not compared with their counterparts, perhaps because many of these NLI systems are not publicly available. Furthermore, most studies have been based on qualitative measurements depending on participants' subjective feedback without much objective quantitative measures. Finally, user studies often took place informally in laboratories by asking a limited number of participants to use the target system which lack realism [Car08, LBI*11].

5.2. Output Dimension

5.2.1. Text

In all existing benchmarks, outputs are limited to texts only and in most datasets possible answers are also limited (Table 3). **Fixed-vocab** datasets such as FigureQA, DVQA, LeafQA and LeafQA++ have a limited number of possible answers that often refer to the charts' textual elements or common answers such as 'yes', 'no', and 'all' (see Figure 9b, 9a, and 9c). For example, FigureQA only supports 'yes'/'no' answers. DVQA has 25 different templates with 1576 distinct possible answers while LeafQA [CSG*20] and LeafQA++ [SS20a] had 75 possible answers, which mainly consist of textual labels in charts (see Figure 9c) or common answers. Such datasets are limited in that they do not consider questions where the answer cannot be found directly in the chart and instead it is required to be computed through operations such as sum, average, count etc.

To address the limitation, PlotQA [MGKK19] introduces **open-vocabulary** questions that require an answer that can be obtained by aggregating over the underlying data values of the charts (see Figure 9d). Still, all the existing CQA datasets have factual **factual** questions (e.g. Figure 9d). None of the benchmarks contains **open-ended** questions that require textual explanatory answers.

5.2.2. Visualization

Some NLIs which generate an answer by creating a new visualization [GDA*15, LHJY21, NSS20] mainly focused on demonstrating the capability of the method through analyzing example outputs. NL4DV demonstrated it's capabilities through several application scenarios (e.g., create visualizations in Jupyter notebooks) while AdVISor was compared with NL4DV for various sample queries over a dataset. DataTone [GDA*15] was evaluated through a user study following the *jeopardy evaluation*, where it was compared with IBM's Watson. However, as the authors acknowledged that directly comparing these two interfaces was difficult as the two systems looked very different. Overall, a key weakness of evaluating the above system is that they were not evaluated on any benchmark dataset by comparing with any competing systems.

5.2.3. Multimodal

As we have discussed in Section 4.3, existing CQA systems do not produce multimodal output combining explanatory text and visualizations as an answer. A starting point could be to build a benchmark dataset to facilitate development and evaluation of new CQA systems that generate multimodal output. Since a multimodal output can be presented as a data story (possibly as a sequence of scenes) it is also worth visiting the evaluation techniques for data driven storytelling. Amini et al. identified several key criteria for evaluating data driven stories [ABB*18], such as comprehension, memorability, and engagement which can be useful in the context of answering questions via data-driven stories. If the output is interactive, quantitative metrics such as time spent by users on explanatory story consumption and interaction statistics (e.g., number of clicks) can give us valuable hints about the efficacy of the generated output. Similarly, self-reported measures such as post-viewing questionnaires and interviews can provide subjective insights about the effectiveness of the generated answer.

5.3. Discussion

In this section, we identified the characteristics and limitations of evaluation techniques for CQA systems. Existing benchmark datasets lack realism in various ways and there is a pressing need for large-scale benchmark datasets consisting of human-authored questions and real-world charts in solving the chart question answering problem. A starting point could be to utilize several existing real-world large-scale datasets containing visualizations that cover a variety of chart types and diverse visual styles [BDM*18, HA19, CLL*21, DWS*20, CZL*20]. For example, Beagle is a dataset with 41K real-world charts crawled from the Web that were created using five different tools [BDM*18]. Hoque and Agrawala [HA19] built a search engine for visualizations by crawling a collection of 7.8K D3 charts from the Web and deconstructing each one to recover its data. Others have collected visualizations and figures from scientific papers [SHL*16, CZL*20, DWS*20, CLL*21]. Moreover, we may consider converting existing TableQA datasets [PL15, ZXS17, IYC17] into ChartQA datasets by translating their queries to SQL commands [YZY*18] and plotting the relevant portions of the data tables into chart images using SQL2Visualization methods [Han06, LTL*21a]. With respect to user studies, future works need to focus on field trials and longitudinal studies where the participants can ask their own questions with their own datasets. This will help to understand the utilities, trade-offs and adoption rate of CQA systems in a more realistic way.

6. Challenges and Research Opportunities

So far, we have presented a formal taxonomy of the chart question answering task and categorized the existing solutions and evaluation techniques around the CQA problem space we identified. This

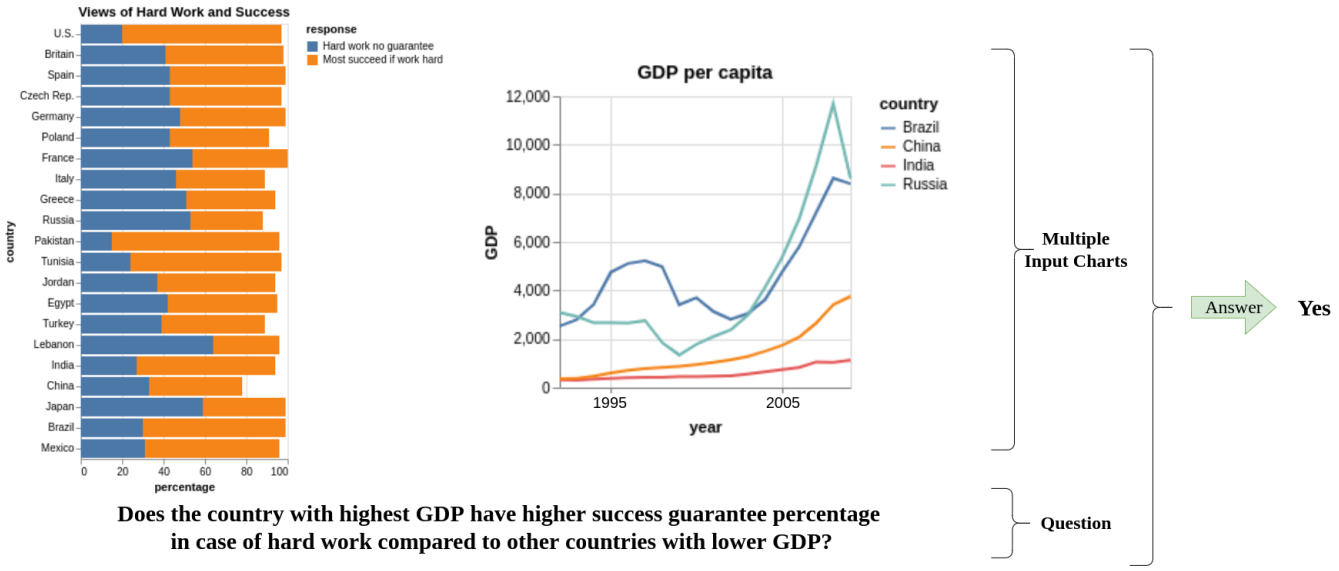


Figure 10: *An example of multiple charts as input. The charts are from the dataset used in [KHA20]. The question requires information from both charts. First, the information for "the country with highest GDP" should be found in the right chart and then, by analyzing the left chart, the Most succeed if work hard component of all countries existing in the right chart are compared to each other in order to reach the final answer.*

systematic review has helped us identify the major gaps in current literature and future opportunities in this exciting area at the intersection of information visualization and AI. We now present these major challenges and possible future directions.

• **Constructing more benchmark datasets**: As we have discussed in Section 5.1, most existing CQA datasets mainly provide template-based questions and synthetically plotted chart images which lack realism. To address this key limitation, it is necessary to develop human authored questions and real-world chart sources with a variety of chart types and visual styles. Human-authored questions can pose several challenges for the CQA models due to the rich semantics, language variations, informal languages and typos. Moreover, there are several types of questions from Table 1 that are underexplored in the existing datasets. While the PlotQA dataset [MGKK19] supported open-vocab questions, they were mostly data-related questions without focusing on questions that refer to the visual attributes (e.g., *color*, *position*) of the chart elements. Finally, there is no existing datasets for open-ended questions that require explanatory paragraphs. Similarly, while there have been some work on conversational interfaces for visualizations [SKC*11], there is no existing dataset that would allow us to train and evaluate conversational question answering interfaces quantitatively. Constructing new datasets on such under-explored problems could open up exciting new avenues of chart question answering.

• **Leveraging more advanced deep learning models**: There are significant rooms for improving the existing methods for chart question answering. Many natural language interfaces for charts often rely on simple heuristics that fail to understand questions that are compositional and require a deeper understanding of the syntactic, semantic, and pragmatic aspects [SBT*16, HSTD17, SS18, YS20]. This is a serious limitation, as the user becomes very frustrated if the system frequently fails to understand the query correctly. In contrast, existing works that solely focus on chart question answering apply classification techniques on chart images [KMA*18, KPCK18, CSG*20] or table question answering on the data table of the chart [MGKK19, KHA20]. As we discussed in Section 4.1, both approaches are limited in handling many real-world questions. To address the limitation, future models need to effectively combine both the visual features and the data values of a chart in a unified representation. In this regard, adapting models from the Vision-Language domain [TB19, DBK*21, SMV*19, CLTB21, LSG*21], which unify visual and textual features effectively through transformer-based models can be a starting point.

• **Addressing chart data extraction challenges**: Many CQA tasks involving perceptual [HTP18] and arithmetic reasoning with charts, which require accurate extraction of chart data and visual encodings (i.e., how data is mapped to visual attributes of graphical marks). Many natural language interfaces assume that the underlying data table and visual encodings of the charts are readily available [SBT*16, SS18, HSTD17, SHKC20]. This assumption becomes invalid for most real-world charts on the Web which are available in bitmap image format without the underlying data. For other charts that are available in SVG format (e.g., D3 charts), there are some techniques for extracting data and visual encodings, however they are also error-prone in some situations (e.g., cannot accurately extract map chart data) [HA17, HA19].

In general, chart data extraction from bitmap images is more challenging and requires us to apply computer vision techniques to automatically extract the underlying data. PlotQA [MGKK19] attempts to use automatically extracted data from chart images to perform question answering, however, it suffers from low accuracy due

to data extraction errors. The key problem here is that existing solutions for chart data extraction are quite limited. Some of them are not fully automatic [SKC*11, JKS*17] which is not very useful for practical purposes. Some focused on recovering color encodings from various charts [PMH17, YZF*21]. Others provided automatic solutions but they rely on various heuristics which do not work for many real-world charts and the performance is still not high enough [CJP*19, LKB19] . ChartOCR automatically extracts data from real-world charts with reasonably high accuracy [LLWL21] but the model only predicts the raw data values of marks (e.g., bars) without associating them with their corresponding axis or legends. Overall, current approaches for automatic data extraction are usually modular and rely on assumptions which are error-prone. An end-to-end deep learning approach could help improve the performance and generalize well to different chart styles.

• **Answering questions with multiple views**: So far, chart question answering has been explored mainly in the context of a single chart as input while only a few natural language interfaces have demonstrated some interactions with multiple views [HSTD17, SHKC20]. In reality, people often interactively analyze dashboards and multiple coordinated views. Analytical questions with multiple views can be more challenging as questions can be more compositional, involving references to multiple charts. Figure 10 shows an example of a question about multiple input visualizations. As we can see, the question requires information from both charts and the model needs to first retrieve the country with *"Maximum GDP value"* in the related chart (the right one) and then find the *"Success guarantee percentage in case of hard work "* of that country and other countries in the left chart to make the comparison. In this context, a starting point could be to collection large-scale real-world human-annotated questions involving multiple charts to better understand and characterize the problem.

• **Combining texts and visualizations as answers**: Existing question answering interfaces for data visualizations typically convey answers in the form of either textual (e.g., [KMA*18, KPCK18, CSG*20, MGKK19]) or visual representations (e.g. [GDA*15, LHJY21] [LTL*21b]) only. Srinivasan et al [SDES18] provided users with a system whereby they can interact with automatically generated textual data facts to search for possible visualizations. Automatically generating a multimedia output combining both text and visualization would not only facilitate users to comprehend the results more effectively by explaining key points but also enhance the transparency of the algorithm by conveying how the results were computed. For example, given the question *"How have the house prices in Toronto changed over time?"* a line chart could show the average house price over time while the text could summarize the price trends. While there have been some works for automatically generating chart summary to describe the patterns, trends and outliers in the chart [OH20], more effort is needed to generate such text in the context of an open-ended question, where generating a chart and the related explanatory text is very helpful to users. Building on deep learning models with encoder-decoder architectures from the NLP and Vision-Language domains [TB19, DBK*21, SMV*19, CLTB21] can be a promising future direction to generate such open-ended explanatory answers.

• **Answering questions via visual data storytelling**: Another future avenue could be combining chart question answering with visual data storytelling to answer the user's questions. The idea here is that rather than showing all the insights about a high-level question (e.g., *"what is really warming the world?"*) at once, a story creates suspense and makes it easy to comprehend insights by transitioning through a sequence of visualizations [RHDC18, GP01, KSJ*14, SH10]. For instance, given the above question, a manually designed data story went through a sequence of line charts to explain which factors have been argued to contribute to global warming to lead to the conclusion that greenhouse gas is indeed the main factor for global warming [‡]. But can we generate such data-driven stories automatically? Recently, there have been some attempts to automatically generate stories from data, however, they do not take any questions into account [SXS*20, CWW*19, CZW*19, WSZ*19]. Generating an interactive data story as a sequence of scenes would require us to answer to questions: *(i)* what to say in the story? *(ii)* how to say it? Solving this problem is challenging as it would be very difficult for a machine to build a creative narrative automatically that highly professional data journalists build manually. To address this challenge, future work may investigate how the system can automatically build a narrative structure with a sequence of 'scenes' combining text and visualizations in a coherent way [HD11, RHDC18]. In future, machine learning models may learn narrative structures from a large collection of annotated human-authored visual data stories. As a starting point, the model could focus on how to learn a simple linear narrative, then followed by generating complex non-linear narrative structures. Linking texts and visualizations [ZOM19, KHKA18] can be also helpful to readers while answering questions through storytelling.

• **Leveraging question answering for chart accessibility** Although the field of visualization has grown dramatically in recent years, the research on inclusive and accessible visualization design remains under-explored [KJRK21]. Specific visualization tools such as SAS [§] and HighCharts [¶] provide limited support by reading data values from charts based on keyboard interactions which can be difficult and mentally taxing when the number of data points is large. While encountering visualizations on Web, blind users largely depend on screen-reader tools that read the alternative text or caption embedded to the chart. However, such alternative texts and captions are often not much helpful or not available at all [MJBC18]. A recent study found that chart description should explain important trends and key statistics rather than simply saying how the data is encoded [LS22]. In this context, introducing chart question answering could significantly advance the field of accessible data visualizations [SWM*22]. In particular, future research could combine automatic chart summarization [OH20] with chart question answering so that people who are blind or have impaired vision can use the audio description of a chart and then simply ask various questions about the chart via speech rather than navigating through numerous data points.

‡ https://www.bloomberg.com/graphics/2015-whats-warming-the-world/
§ www.sas.com
¶ https://www.highcharts.com/docs/accessibility/accessibility-module

7. Conclusion

In this paper, we have presented a survey on chart question answering by analyzing relevant papers from the field of information visualization, human computer interactions, and natural language processing. Through this analysis, we derived a taxonomy with the possible input and output dimensions that illustrate the problem space. We synthesized the findings from existing works along these dimensions and identified key knowledge gaps in this domain. Finally, we outlined the open challenges and opportunities to inform future research in the domain. We hope that this survey will help other researchers in initiating future contributions in this relatively new area of visualization research.

Acknowledgement

This work was supported by the Natural Sciences and Engineering Research Council (NSERC), Canada. We thank anonymous reviewers for their valuable comments and suggestions.

References

[ABB*18] AMINI F., BREHMER M., BOLDUAN G., ELMER C., WIEDERKEHR B.: Evaluating data-driven stories and storytelling tools. In *Data-Driven Storytelling*. AK Peters/CRC Press, 2018, pp. 249–286. 12

[AES05] AMAR R., EAGAN J., STASKO J.: Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* (2005), IEEE, pp. 111–117. 5

[AEYN11] ANDREWS C., ENDERT A., YOST B., NORTH C.: Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Information Visualization 10*, 4 (2011), 341–355. 7

[ask] Tableau ask data, howpublished = https://help.tableau.com/current/pro/desktop/en-us/ask_data.htm. 1

[BAEI16] BADAM S. K., AMINI F., ELMQVIST N., IRANI P.: Supporting visual exploration for multiple users in large display environments. In *Visual Analytics Science and Technology (VAST), 2016 IEEE Conference on* (2016), IEEE, pp. 1–10. 7

[BDM*18] BATTLE L., DUAN P., MIRANDA Z., MUKUSHEVA D., CHANG R., STONEBRAKER M.: Beagle: Automated extraction and interpretation of visualizations from the web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–8. 12

[Car08] CARPENDALE S.: Evaluating information visualizations. In *Information visualization*. Springer, 2008, pp. 19–45. 12

[CJP*19] CHOI J., JUNG S., PARK D. G., CHOO J., ELMQVIST N.: Visualizing for the non-visual: Enabling the visually impaired to use visualization. *Computer Graphics Forum 38* (2019). 14

[CLL*21] CHEN J., LING M., LI R., ISENBERG P., ISENBERG T., SEDLMAIR M., MOLLER T., LARAMEE R. S., SHEN H.-W., WUNSCHE K., ET AL.: Vis30k: A collection of figures and tables from ieee visualization conference publications. *IEEE Transactions on Visualization and Computer Graphics* (2021). 12

[CLTB21] CHO J., LEI J., TAN H., BANSAL M.: Unifying vision-and-language tasks via text generation. In *ICML* (2021). 13, 14

[CM85] CLEVELAND W. S., MCGILL R.: Graphical perception and graphical methods for analyzing scientific data. *Science 229*, 4716 (1985), 828–833. 1

[CMH*20] CHOWDHURY I., MOEID A., HOQUE E., KABIR M. A., HOSSAIN M. S., ISLAM M. M.: Designing and evaluating multimodal interactions for facilitating visual analysis with dashboards. *IEEE Access 9* (2020), 60–71. 5

[CSG*20] CHAUDHRY R., SHEKHAR S., GUPTA U., MANERIKER P., BANSAL P., JOSHI A.: Leaf-qa: Locate, encode attend for figure question answering. *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020* (2020), 3501–3510. doi:10.1109/WACV45572.2020.9093269. 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14

[CWW*19] CHEN Z., WANG Y., WANG Q., WANG Y., QU H.: Towards automated infographic design: Deep learning-based auto-extraction of extensible timeline. *IEEE transactions on visualization and computer graphics 26*, 1 (2019), 917–926. 9, 14

[CZL*20] CHEN X., ZENG W., LIN Y., AI-MANEEA H. M., ROBERTS J., CHANG R.: Composition and configuration patterns in multiple-view visualizations. *IEEE Transactions on Visualization and Computer Graphics 27*, 2 (2020), 1514–1524. 12

[CZW*19] CUI W., ZHANG X., WANG Y., HUANG H., CHEN B., FANG L., ZHANG H., LOU J.-G., ZHANG D.: Text-to-viz: Automatic generation of infographics from proportion-related natural language statements. *IEEE transactions on visualization and computer graphics 26*, 1 (2019), 906–916. 9, 14

[DBK*21] DOSOVITSKIY A., BEYER L., KOLESNIKOV A., WEISSENBORN D., ZHAI X., UNTERTHINER T., DEHGHANI M., MINDERER M., HEIGOLD G., GELLY S., USZKOREIT J., HOULSBY N.: An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations* (2021). URL: https://openreview.net/forum?id=YicbFdNTTy. 7, 13, 14

[DCM12] DEMIR S., CARBERRY S., MCCOY K. F.: Summarizing information graphics textually. *Computational Linguistics 38*, 3 (2012), 527–574. URL: https://www.aclweb.org/anthology/J12-3004, doi:10.1162/COLI_a_00091. 4

[DKG*17] DAS A., KOTTUR S., GUPTA K., SINGH A., YADAV D., MOURA J. M., PARIKH D., BATRA D.: Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 326–335. 9

[DWS*20] DENG D., WU Y., SHU X., WU J., XU M., FU S., CUI W., WU Y.: Visimages: a corpus of visualizations in the images of visualization publications. *arXiv preprint arXiv:2007.04584* (2020). 12

[FBM15] FULDA J., BREHMER M., MUNZNER T.: Timelinecurator: Interactive authoring of visual timelines from unstructured text. *IEEE transactions on visualization and computer graphics 22*, 1 (2015), 300–309. 9

[FR16] FERRARI A., RUSSO M.: *Introducing Microsoft Power BI*. Microsoft Press, 2016. 1

[GDA*15] GAO T., DONTCHEVA M., ADAR E., LIU Z., KARAHALIOS K. G.: Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software amp; Technology* (New York, NY, USA, 2015), UIST '15, Association for Computing Machinery, p. 489–500. URL: https://doi.org/10.1145/2807442.2807478, doi:10.1145/2807442.2807478. 1, 2, 5, 7, 8, 12, 14

[Gov21] GOVE R.: Automatic narrative summarization for visualizing cyber security logs and incident reports. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (2021), 1182–1190. 9

[GP01] GERSHON N., PAGE W.: What storytelling can do for information visualization. *Communications of the ACM 44*, 8 (2001), 31–37. 14

[HA17] HARPER J., AGRAWALA M.: Converting basic d3 charts into reusable style templates. *IEEE transactions on visualization and computer graphics 24*, 3 (2017), 1274–1286. 13

[HA19] HOQUE E., AGRAWALA M.: Searching the visual style and structure of d3 visualizations. In *IEEE Transactions on Visualization and Computer Graphics (Proc IEEE InfoVis 2019)* (2019), vol. 26, IEEE, pp. 1236–1245. 12, 13

[Han06] HANRAHAN P.: Vizql: A language for query, analysis and

visualization. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, 2006), SIGMOD '06, Association for Computing Machinery, p. 721. URL: https://doi.org/10.1145/1142473.1142560, doi:10.1145/1142473.1142560. 12

[HBED18] HORAK T., BADAM S. K., ELMQVIST N., DACHSELT R.: When david meets goliath: Combining smartwatches with a large vertical display for visual data exploration. In *Conference on Human Factors in Computing Systems - Proceedings* (apr 2018), vol. 2018-April, Association for Computing Machinery. doi:10.1145/3173574.3173593. 7

[HBL*19] HU K., BAKKER M. A., LI S., KRASKA T., HIDALGO C.: Vizml: A machine learning approach to visualization recommendation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019), pp. 1–12. 2

[HD11] HULLMAN J., DIAKOPOULOS N.: Visualization rhetoric: Framing effects in narrative visualization. *IEEE transactions on visualization and computer graphics 17*, 12 (2011), 2231–2240. 14

[HDA13] HULLMAN J., DIAKOPOULOS N., ADAR E.: Contextifier: automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on human factors in computing systems* (2013), pp. 2707–2716. 9

[HGDG17] HE K., GKIOXARI G., DOLLÁR P., GIRSHICK R.: Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision* (2017), pp. 2961–2969. 6

[HNM*20] HERZIG J., NOWAK P. K., MÜLLER T., PICCINNO F., EISENSCHLOS J. M.: TAPAS: weakly supervised table parsing via pre-training. *CoRR abs/2004.02349* (2020). URL: https://arxiv.org/abs/2004.02349, arXiv:2004.02349. 7

[HSTD17] HOQUE E., SETLUR V., TORY M., DYKEMAN I.: Applying pragmatics principles for interaction with visual analytics. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2017), 309–318. 1, 5, 6, 7, 8, 9, 10, 13, 14

[HT19] HEARST M., TORY M.: Would you like a chart with that? incorporating visualizations into conversational interfaces. In *2019 IEEE Visualization Conference (VIS)* (2019), IEEE, pp. 1–5. 10

[HTP18] HAEHN D., TOMPKIN J., PFISTER H.: Evaluating 'graphical perception'with cnns. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 641–650. 13

[HYY21] HUANG T.-Y., YANG Y.-L., YANG X.-J.: A survey of deep learning-based visual question answering. *Journal of Central South University 28*, 3 (2021), 728–746. 2

[IYC17] IYYER M., YIH W.-T., CHANG M.-W.: Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vancouver, Canada, July 2017), Association for Computational Linguistics, pp. 1821–1831. URL: https://aclanthology.org/P17-1167, doi:10.18653/v1/P17-1167. 12

[JE12] JAVED W., ELMQVIST N.: Exploring the design space of composite visualization. In *2012 ieee pacific visualization symposium* (2012), IEEE, pp. 1–8. 5

[JKS*17] JUNG D., KIM W., SONG H., IN HWANG J., LEE B., KIM B., SEO J.: Chartsense: Interactive data extraction from chart images. ACM. URL: https://www.microsoft.com/en-us/research/publication/chartsense-interactive-data-extraction-chart-images/. 14

[KHA20] KIM D. H., HOQUE E., AGRAWALA M.: Answering questions about charts and generating visual explanations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13. 1, 2, 3, 4, 5, 6, 8, 10, 11, 13

[KHKA18] KIM D. H., HOQUE E., KIM J., AGRAWALA M.: Facilitating document reading by linking text and tables. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (2018), pp. 423–434. 14

[KJRK21] KIM N. W., JOYNER S. C., RIEGELHUTH A., KIM Y.: Accessible visualization: Design space, opportunities, and challenges. *Computer Graphics Forum 40* (2021). doi:10.1111/cgf.14298. 14

[KMA*18] KAHOU S. E., MICHALSKI V., ATKINSON A., ÁKOS KÁDÁR, TRISCHLER A., BENGIO Y.: Figureqa: An annotated figure dataset for visual reasoning. *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings* (2018), 1–20. 1, 2, 5, 6, 7, 8, 10, 11, 13, 14

[KPCK18] KAFLE K., PRICE B., COHEN S., KANAN C.: Dvqa: Understanding data visualizations via question answering. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018), 5648–5656. doi:10.1109/CVPR.2018.00592. 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 13, 14

[KR18] KASSEL J.-F., ROHS M.: Valletto: A multimodal interface for ubiquitous visual analytics. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), pp. 1–6. 7

[KSC*20] KAFLE K., SHRESTHA R., COHEN S., PRICE B., KANAN C.: Answering questions about data visualizations using efficient bimodal fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (2020), pp. 1498–1507. 2, 4, 5, 6, 8

[KSJ*14] KWON B. C., STOFFEL F., JÄCKLE D., LEE B., KEIM D.: Visjockey: Enriching data stories through orchestrated interactive visualization. In *Poster compendium of the computation+ journalism symposium* (2014), vol. 3, p. 3. 14

[LBI*11] LAM H., BERTINI E., ISENBERG P., PLAISANT C., CARPENDALE S.: Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics 18*, 9 (2011), 1520–1536. 12

[LHJ*21] LAN Y., HE G., JIANG J., JIANG J., ZHAO W. X., WEN J.-R.: Complex knowledge base question answering: A survey. *arXiv preprint arXiv:2108.06688* (2021). 2

[LHJY21] LIU C., HAN Y., JIANG R., YUAN X.: Advisor: Automatic visualization answer for natural-language question on tabular data. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)* (2021), pp. 11–20. doi:10.1109/PacificVis52677.2021.00010. 4, 5, 6, 8, 9, 12, 14

[LKB19] LIU X., KLABJAN D., BLESS P. N.: Data extraction from charts via single deep neural network. *ArXiv abs/1906.11906* (2019). 14

[LLWL21] LUO J., LI Z., WANG J., LIN C.-Y.: Chartocr: Data extraction from charts images via a deep hybrid framework. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2021), 1916–1924. 14

[LS22] LUNDGARD A., SATYANARAYAN A.: Accessible Visualization via Natural Language Descriptions: A Four-Level Model of Semantic Content. *IEEE Trans. Visualization & Comp. Graphics (Proc. IEEE VIS)* (2022). URL: http://vis.csail.mit.edu/pubs/vis-text-model. 14

[LSG*21] LI J., SELVARAJU R. R., GOTMARE A. D., JOTY S., XIONG C., HOI S.: Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems* (2021), Beygelzimer A., Dauphin Y., Liang P., Vaughan J. W., (Eds.). URL: https://openreview.net/forum?id=OJLaKwiXSbx. 13

[LTL*21a] LUO Y., TANG N., LI G., CHAI C., LI W., QIN X.: *Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks*. Association for Computing Machinery, New York, NY, USA, 2021, p. 1235–1247. URL: https://doi.org/10.1145/3448016.3457261. 2, 12

[LTL*21b] LUO Y., TANG N., LI G., TANG J., CHAI C., QIN X.: Natural language to visualization by neural machine translation. *IEEE Trans-*

actions on Visualization and Computer Graphics 28, 1 (2021), 217–226. 5, 9, 14

[LZW*21] LIU Y., ZHANG Y., WANG Y., HOU F., YUAN J., TIAN J., ZHANG Y., SHI Z., FAN J., HE Z.: A survey of visual transformers. *arXiv preprint arXiv:2111.06091* (2021). 7

[MGKK19] METHANI N., GANGULY P., KHAPRA M. M., KUMAR P.: Data interpretation over plots. *CoRR abs/1909.00997* (2019). URL: http://arxiv.org/abs/1909.00997, arXiv:1909.00997. 2, 4, 5, 6, 8, 10, 11, 12, 13, 14

[MH21] MASRY A., HOQUE E.: Integrating image data extraction and table parsing methods for chart question answering. *Chart Question Answering Workshop, in conjunction with the Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 1–5. 4, 8

[MJBC18] MORRIS M. R., JOHNSON J., BENNETT C. L., CUTRELL E.: Rich representations of visual content for screen reader users. vol. 2018-April. doi:10.1145/3173574.3173633. 14

[MLT*22] MASRY A., LONG D. X., TAN J. Q., JOTY S. R., HOQUE E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *ArXiv abs/2203.10244* (2022). 4, 5, 7, 10, 11

[MSDT18] MASSICETI D., SIDDHARTH N., DOKANIA P. K., TORR P. H.: Flipdial: A generative model for two-way visual dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 6097–6105. 9

[Mun14] MUNZNER T.: *Visualization Analysis and Design*. CRC Press, 2014. 11

[MZJS18] METOYER R., ZHI Q., JANCZUK B., SCHEIRER W.: Coupling story to visualization: Using textual analysis as a bridge between data and interpretation. In *23rd International Conference on Intelligent User Interfaces* (2018), pp. 503–507. 9

[NSS20] NARECHANIA A., SRINIVASAN A., STASKO J.: Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics 27*, 2 (2020), 369–379. 3, 5, 6, 8, 12

[OH20] OBEID J., HOQUE E.: Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference on Natural Language Generation* (2020), Association for Computational Linguistics, pp. 138–147. URL: https://www.aclweb.org/anthology/2020.inlg-1.20. 2, 4, 7, 8, 14

[PL15] PASUPAT P., LIANG P.: Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (Beijing, China, July 2015), Association for Computational Linguistics, pp. 1470–1480. URL: https://www.aclweb.org/anthology/P15-1142, doi:10.3115/v1/P15-1142. 5, 8, 12

[PMH17] POCO J., MAYHUA A., HEER J.: Extracting and retargeting color mappings from bitmap images of visualizations. *IEEE transactions on visualization and computer graphics 24*, 1 (2017), 637–646. 14

[QYQ*19] QU C., YANG L., QIU M., CROFT W. B., ZHANG Y., IYYER M.: Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval* (2019), pp. 1133–1136. 9

[RCM18] REDDY S., CHEN D., MANNING C. D.: Coqa: A conversational question answering challenge. *arXiv* (2018). doi:10.1162/tacl_a_00266. 9

[RHDC18] RICHE N. H., HURTER C., DIAKOPOULOS N., CARPENDALE S.: *Data-driven storytelling*. CRC Press, 2018. 14

[Rob98] ROBERTS J. C.: On encouraging multiple views for visualization. In *Proceedings. 1998 IEEE Conference on Information Visualization. An International Conference on Computer Visualization and Graphics (Cat. No. 98TB100246)* (1998), IEEE, pp. 8–14. 9

[RRDK19] REDDY R., RAMESH R., DESHPANDE A., KHAPRA M. M.: Figurenet : A deep learning model for question-answering on scientific plots. *Proceedings of the International Joint Conference on Neural Networks 2019-July* (2019). doi:10.1109/IJCNN.2019.8851830. 1, 2, 3, 4, 5, 6, 8

[RSR*19] RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., PETER W. L., LIU J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv 21* (2019), 1–67. 10

[SBT*16] SETLUR V., BATTERSBY S. E., TORY M., GOSSWEILER R., CHANG A. X.: Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (New York, NY, USA, 2016), UIST 2016, ACM, pp. 365–377. 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 13

[SC20] SPREAFICO A., CARENINI G.: Neural data-driven captioning of time-series line charts. In *Proceedings of the International Conference on Advanced Visual Interfaces* (New York, NY, USA, 2020), AVI '20, Association for Computing Machinery. URL: https://doi.org/10.1145/3399715.3399829, doi:10.1145/3399715.3399829. 4

[SDES18] SRINIVASAN A., DRUCKER S. M., ENDERT A., STASKO J.: Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE transactions on visualization and computer graphics* (2018). 14

[SH10] SEGEL E., HEER J.: Narrative visualization: Telling stories with data. *IEEE transactions on visualization and computer graphics 16*, 6 (2010), 1139–1148. 14

[SHKC20] SETLUR V., HOQUE E., KIM D. H., CHANG A. X.: Sneak pique: Exploring autocompletion as a data discovery scaffold for supporting visual analysis. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (New York, NY, USA, 2020), UIST '20, Association for Computing Machinery, p. 966–978. URL: https://doi.org/10.1145/3379337.3415813, doi:10.1145/3379337.3415813. 6, 7, 8, 13, 14

[SHL*16] SIEGEL N., HORVITZ Z., LEVIN R., DIVVALA S., FARHADI A.: Figureseer: Parsing result-figures in research papers. In *European Conference on Computer Vision* (2016), Springer, pp. 664–680. 12

[SKC*11] SAVVA M., KONG N., CHHAJTA A., FEI-FEI L., AGRAWALA M., HEER J.: Revision: automated classification, analysis and redesign of chart images. *Proceedings of the 24th annual ACM symposium on User interface software and technology* (2011). 13, 14

[SLHR*20] SRINIVASAN A., LEE B., HENRY RICHE N., DRUCKER S. M., HINCKLEY K.: Inchorus: Designing consistent multimodal interactions for data visualization on tablet devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–13. 2, 7, 11, 12

[SMV*19] SUN C., MYERS A., VONDRICK C., MURPHY K. P., SCHMID C.: Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 7463–7472. 13, 14

[SMWH16] SATYANARAYAN A., MORITZ D., WONGSUPHASAWAT K., HEER J.: Vega-lite: A grammar of interactive graphics. *IEEE transactions on visualization and computer graphics 23*, 1 (2016), 341–350. 2, 6

[SNL*21] SRINIVASAN A., NYAPATHY N., LEE B., DRUCKER S. M., STASKO J.: Collecting and characterizing natural language utterances for specifying data visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), pp. 1–10. 2, 10

[SRB*17] SANTORO A., RAPOSO D., BARRETT D. G., MALINOWSKI M., PASCANU R., BATTAGLIA P., LILLICRAP T.: A simple neural network module for relational reasoning. *arXiv preprint arXiv:1706.01427* (2017). 6

[SRTKX*22] SHANKAR K., RIXIE TIFFANY KO L., XIANG L., AHMED M., MEGH T., ENAMUL H., SHAFIQ J.: Chart-to-text: A large-scale benchmark for chart summarization. In *In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2022* (2022). 4, 8

[SS18] SRINIVASAN A., STASKO J.: Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. *IEEE transactions on visualization and computer graphics 24*, 1 (2018), 511–521. 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13

[SS20a] SINGH H., SHEKHAR S.: 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)* (Online, Nov. 2020), Association for Computational Linguistics, pp. 3275–3284. URL: https://www.aclweb.org/anthology/2020.emnlp-main.264, doi:10.18653/v1/2020.emnlp-main.264. 1, 2, 4, 5, 6, 8, 10, 11, 12

[SS20b] SRINIVASAN A., STASKO J.: How to ask what to say?: Strategies for evaluating natural language interfaces for data visualization. *IEEE Computer Graphics and Applications 40*, 4 (2020), 96–103. 1

[SSL*21] SHEN L., SHEN E., LUO Y., YANG X., HU X., ZHANG X., TAI Z., WANG J.: Towards natural language interfaces for data visualization: A survey. *CoRR abs/2109.03506* (2021). URL: https://arxiv.org/abs/2109.03506, arXiv:2109.03506. 1

[SSS20] SAKTHEESWARAN A., SRINIVASAN A., STASKO J.: Touch? speech? or touch and speech? investigating multimodal interaction for visual network exploration and analysis. *IEEE transactions on visualization and computer graphics 26*, 6 (2020), 2168–2179. 7

[SWM*22] SHARIF A., WANG O. H., MUONGCHAN A. T., REINECKE K., WOBBROCK J. O.: Voxlens: Making online data visualizations accessible with an interactive javascript plug-in. 1, 14

[SXS*20] SHI D., XU X., SUN F., SHI Y., CAO N.: Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Transactions on Visualization and Computer Graphics 27*, 2 (2020), 453–463. 8, 9, 14

[TB19] TAN H., BANSAL M.: Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (2019). 6, 13, 14

[TLW*20] TANG T., LI R., WU X., LIU S., KNITTEL J., KOCH S., YU L., REN P., ERTL T., WU Y.: Plotthread: Creating expressive storyline visualizations using reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics 27*, 2 (2020), 294–303. 9

[TRB*18] TONG C., ROBERTS R., BORGO R., WALTON S., LARAMEE R. S., WEGBA K., LU A., WANG Y., QU H., LUO Q., MA X.: Storytelling and visualization: An extended survey. *Information 9*, 3 (2018). URL: https://www.mdpi.com/2078-2489/9/3/65, doi:10.3390/info9030065. 2

[VRS*22] VIJ R., RAJ R., SINGHAL M., TANWAR M., BEDATHUR S.: Vizai: Selecting accurate visualizations of numerical data. In *5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)* (2022), pp. 28–36. 9

[WLJ*12] WALNY J., LEE B., JOHNS P., RICHE N. H., CARPENDALE S.: Understanding pen and touch interaction for data exploration on interactive whiteboards. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (2012), 2779–2788. 7

[WMA*15] WONGSUPHASAWAT K., MORITZ D., ANAND A., MACKINLAY J., HOWE B., HEER J.: Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics 22*, 1 (2015), 649–658. 9

[WSZ*19] WANG Y., SUN Z., ZHANG H., CUI W., XU K., MA X., ZHANG D.: Datashot: Automatic generation of fact sheets from tabular data. *IEEE transactions on visualization and computer graphics 26*, 1 (2019), 895–905. 9, 14

[YS20] YU B., SILVA C. T.: Flowsense: A natural language interface for visual data exploration within a dataflow system. *IEEE Transactions on Visualization and Computer Graphics 26*, 1 (2020), 1–11. doi:10.1109/TVCG.2019.2934668. 5, 13

[YZF*21] YUAN L.-P., ZENG W., FU S., ZENG Z., LI H., FU C.-W., QU H.: Deep colormap extraction from visualizations. *arXiv preprint arXiv:2103.00741* (2021). 14

[YZY*18] YU T., ZHANG R., YANG K., YASUNAGA M., WANG D., LI Z., MA J., LI I., YAO Q., ROMAN S., ZHANG Z., RADEV D.: Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (Brussels, Belgium, Oct.-Nov. 2018), Association for Computational Linguistics, pp. 3911–3921. URL: https://aclanthology.org/D18-1425, doi:10.18653/v1/D18-1425. 2, 7, 12

[ZCW19] ZHANG D., CAO R., WU S.: Information fusion in visual question answering: A survey. *Information Fusion 52* (2019), 268–280. 2

[ZLW*21] ZHU F., LEI W., WANG C., ZHENG J., PORIA S., CHUA T.-S.: Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv preprint arXiv:2101.00774* (2021). 2

[ZMD*21] ZENG Z., MOH P., DU F., HOFFSWELL J., LEE T. Y., MALIK S., KOH E., BATTLE L.: An evaluation-focused framework for visualization recommendation algorithms. *IEEE Transactions on Visualization and Computer Graphics 28*, 1 (2021), 346–356. 9

[ZOM19] ZHI Q., OTTLEY A., METOYER R.: Linking and layout: Exploring the integration of text and visualization in storytelling. In *Computer Graphics Forum* (2019), vol. 38, Wiley Online Library, pp. 675–685. 14

[ZSJ*20] ZHU S., SUN G., JIANG Q., ZHA M., LIANG R.: A survey on automatic infographics and visualization recommendations. *Visual Informatics 4*, 3 (2020), 24–40. 9

[ZWXW20] ZOU J., WU G., XUE T., WU Q.: An affinity-driven relation network for figure question answering. *Proceedings - IEEE International Conference on Multimedia and Expo 2020-July* (2020). doi:10.1109/ICME46284.2020.9102911. 4, 5, 6, 8

[ZXC*21] ZHAO J., XU S., CHANDRASEGARAN S., BRYAN C., DU F., MISHRA A., QIAN X., LI Y., MA K.-L.: Chartstory: Automated partitioning, layout, and captioning of charts into comic-style narratives. *arXiv preprint arXiv:2103.03996* (2021). 9

[ZXS17] ZHONG V., XIONG C., SOCHER R.: Seq2sql: Generating structured queries from natural language using reinforcement learning, 2017. arXiv:1709.00103. 12