Would You Like A Chart With That?
Incorporating Visualizations into Conversational Interfaces

Marti Hearst*
UC Berkeley

Melanie Tory†
Tableau Software, Inc.

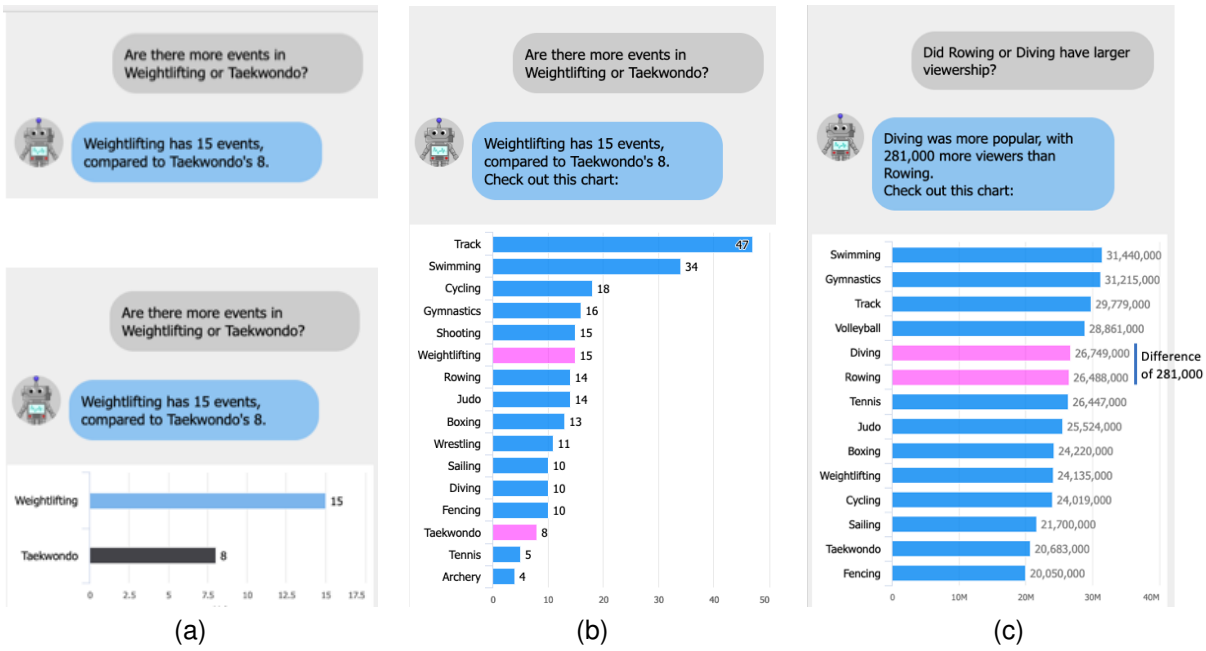

Figure 1: Four types of response to comparison queries. (a) [Top] A text only answer. (a) [Bottom] The same answer, augmented with one bar for each of the compared values. (b) Sixteen items of context, target items both highlighted. (c) Fourteen bars with values closer together, and differences called out with an annotation as well as highlighting.

Abstract

Conversational interfaces, such as chatbots, are increasing in prevalence, and have been shown to be preferred by and help users to complete tasks more efficiently than standard web interfaces in some cases. However, little is understood about if and how information should be visualized during the course of an interactive conversation. This paper describes studies in which participants report their preferences for viewing visualizations in chat-style interfaces when answering questions about comparisons and trends. We find a significant split in preferences among participants; approximately 40% prefer not to see charts and graphs in the context of a conversational interface. For those who do prefer to see charts, most preferred to see additional supporting context beyond the direct answer to the question. These results have important ramifications for the design of conversational interfaces to data.

Index Terms: Human-centered computing—Visualization;

1 Introduction

Conversational interfaces have become commonplace on mobile devices, helping users select music, get driving directions, and answer informational questions. Several recent studies have shown significant improvements in efficiency with conversational interfaces over conventional methods [1, 5, 14]. Recently, Folstad and Brandtzaeg [6] predicted that "in the not-too-distant future, chatbots may be the preferred user interface for many of the activities to which we have grown accustomed to performing through a webpage or a dedicated application," but note that this form of interaction is not receiving its due in the HCI research literature. This holds true as well in the field of information visualization where little is known about the most appropriate response to questions about data when posed in a conversational user interface.

This new style of interaction opens an interesting area for investigation. Do information visualization guidelines change in the context of responding in a conversational setting? If so, how? People interact with other people via language regularly, but it is unusual for another person to insert graphs and charts into the conversation.

This work presents the first empirical studies exploring how people perceive the *appropriateness* of the presentation of charts and graphs in the context of a computer-mediated chat-style conversation. Our goal was to investigate the interplay of language and visualization, within a framework of *how much* information is appropriate in the form of charts and graphs *in the context of a chat-style interaction*.

We focused on two kinds of questions that are relevant to the information visualization literature: *comparisons* and *trends*. Comparisons are interesting because they are a fundamental activity within information visualization [2, 8, 22] and yet brief textual answers can be quite appropriate responses to comparison questions (see Figure

*e-mail: hearst@berkeley.edu; This work was conducted while Hearst was a visiting researcher at Tableau Research.
†e-mail: mtory@tableau.com

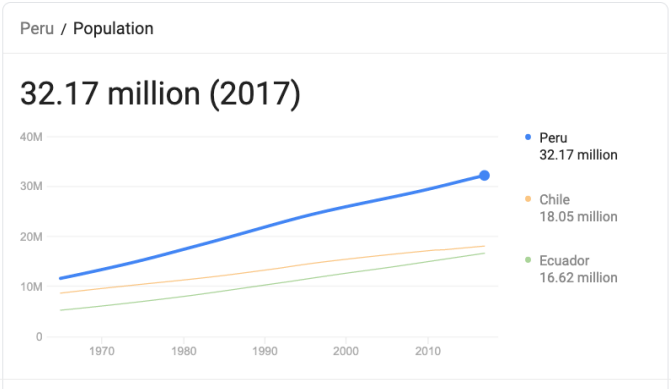

Figure 2: Answer box for Google's response to the query "What is the population of Peru?" Generated 2/28/2019.

1). Trends are also common in visualization and web search engines currently show graphs with additional context beyond what is asked for (see Figure 2), and we wanted to see if participants thought this extra information was appropriate. We also introduce a variation of a comparison question, a superlative. Superlative questions such as "Which athlete is tallest?" can be answered by a single word or a single bar, so we wanted to know if participants preferred a direct answer or additional context in a conversational setting.

To answer these questions, we conducted studies using a crowd sourcing platform. Instead of assessing how *accurately* or *efficiently* people process responses to questions, we assessed how *appropriate* they feel the interactions are within a conversation. This focus on appropriateness is done in the interest of creating natural, comfortable interactions [21].

2 Related Work

2.1 Natural Language Queries in Visualization Systems

Although there is voluminous research on how to automate natural language question answering, there is little work on the correct way to *show answers* to questions about information or data. Most work in the field of natural language interfaces to databases [14, 15, 18, 19] focuses on how to parse natural language statements into database queries; the method of display is usually a table.

Natural language interfaces to data sources that produce visualizations growing in interest [24]. Srinivasan and Stasko [24], in a thought piece on the space of NL and infoviz interfaces, note that this is an underexplored area and summarize the features covered by five existing systems. They do not discuss chatbot-style interactions specifically, and the design space they outline makes clear that systems either simply choose a default layout for a question type, or allow the user to choose a layout from a dropdown list interactively.

Research systems such as DataTone [7], Voder [23], Eviza [20] and Evizeon [10], have explored how to interpret user's natural language queries against data sources and express visualizations in response, and use mixed initiative interaction to help users refine an existing visualization. Recently released commercial products such as ThoughtSpot and Tableau's AskData allow users to type queries against data and see results expressed as visualizations, with inferencing to handle underspecification [21].

2.2 Search Results

Published research on visualization and chatbot interfaces does not seem to exist. However, question answering in a conversational interface bears some resemblance to interactive web search. Although not much research has been done on web search results that produce graphs and charts, there have been several experiments on what kind of information should be provided within search results depending on the type of query [4, 16, 17]. Kaisser et al. [13] found that judges could guess the appropriate search result length (word or phrase, sentence, paragaph(s), article, list) based on the expected answer type (person, number or quantity, product information, explanation, etc), and that getting the length right affected result quality.

Web search engines have become increasingly skilled at responding to queries posed in natural language, and often show an answer box at the top of the results listing, which in some cases contains extracted data [3].

3 Study Design

We conducted two studies to investigate the appropriateness of visualization responses within a chatbot style interaction with data. We examined text versus text plus visualization responses within the context of a chat-style conversation. We investigated this within a framework of how much information is appropriate, as described in Section 1.

3.1 Development of Study Stimuli

The two studies were structured identically, but presented different sets of stimuli to participants. Experiment 1 asked *comparison* questions, while Experiment 2 asked *trend* and *superlative* questions. Participants were asked to imagine that they were engaged in a conversation with a chatbot, and share their views of the appropriateness of the responses.

We refined the stimuli, tasks and questions over a series of pilot studies. In these pilots, we noticed a split in preferences when asking participants about the appropriateness of showing charts in response to questions. We coded the most frequent reasons and compressed them into two questions, presented on a Likert agreement scale from strongly disagree through strongly agree:

- The view I selected provides exactly the information I asked for and nothing else.
- The view I selected provides additional information beyond what I asked for that is potentially relevant.

Also common in the pilot studies were valence statements about charts – positive or negative. In the final studies, participants were asked to respond to this statement to capture this aspect:

- I made my selection in part because I like charts and graphs.

3.2 Participant Recruitment

Much contemporary research in information visualization recruits participants from the crowdsourcing platform Amazon Mechanical Turk (e.g., [9, 11]) as we do here, both for convenience and because the platform participants reflect our target demographic. We restricted participation to English speakers in the U.S. with at least a 95% acceptance rate and 500 approved tasks. We payed a rate equivalent to $1.50 for 10 minutes of effort. We carefully read all textual responses received and eliminated any that appeared not to be conforming to the requested information. The stimuli did not require excluding participants for color deficiencies.

3.3 Study 1: Comparison Questions

The first study compared text only and text plus bar chart responses to the comparison question "Are there more events in Weightlifting or Taekwondo?" in Block B1 and to "Did Rowing or Diving have larger viewership?" for Block B2. Each block compared two text and two bar chart responses as shown (partially) in Figure 1. B2 differed from B1 in that it examined the role of comparing values that have small differences between them. Choices were shown in the same order to all participants, from the least text to the most bars/lines/annotations, for all blocks B1 - B4.

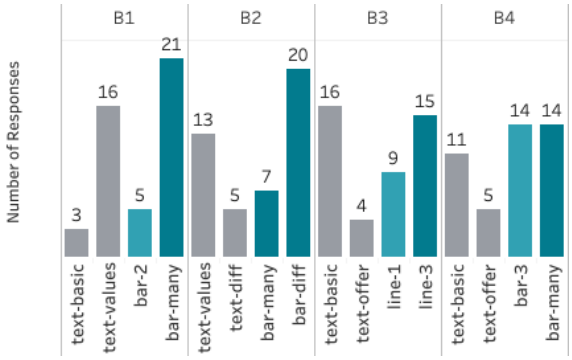

Figure 3: Preferences for Studies 1 (B1, B2) and 2 (B3, B4) (grey = text, aqua = charts). Lighter aqua = the option with less context.

	Description	B1	B2
Text - basic	Direct answer only: "Weightlifting has more events."	x	
Text - values	Provides values (Figure 1a, top): "Weightlifting has 15 events, compared to Taekwondo's 8."	x	x
Text - diff	Provides values and difference: "Diving was more popular, with 281,000 more viewers than Rowing, for a total of more than 26 million."		x
Bar - 2	Bar chart with only the two entities in the input question (Figure 1a, bottom).	x	
Bar - many	Bar chart with many bars; focus entities highlighted (Figure 1b for B1).	x	x
Bar - diff	Bar - many, annotated with the difference between target values (Figure 1c).		x

Table 1: Conditions in Block 1 (B1) and Block 2 (B2) of Study 1. Bar-many had 16 bars in B1 and 14 bars in B2.

3.3.1 Method

Participants completed a brief survey containing two blocks of questions via Mechanical Turk.

Tasks and Procedure: For each block, participants were told to assume they had asked a chatbot a given question, were shown 4 different views and were asked to indicate which view they preferred. They then reported the reason for their choice as free-form text. Finally, they were asked to agree / disagree (on a 5 point rating scale) with 3 possible reasons for their answer: (1) *Exact Info* - "provides exactly the information I asked for and nothing else," (2) *Additional Info* - "provides additional information beyond what I asked for that is potentially relevant," and (3) *Like Charts* - "I made my selection in part because I like charts and graphs." The survey is available as supplemental material. The conditions shown in each block are summarized in Table 1. Average completion time was 4.5 minutes.

Participants: 48 crowdworkers participated; we removed 3 of these from the analysis based on off-topic free-form responses.

3.3.2 Results

Figure 3 shows a summary of preferred system responses for blocks B1 and B2. Participants split into two distinct groups: those who preferred text alone and those who preferred bar charts in addition to the text; a statistical test comparing the second choice given the first significantly divides participants into two groups $\chi^2(1, N = 45) = 26.8, p < .001$). As shown in Figure 4, the distribution of ratings differed between people who preferred charts versus text. Interestingly, most people were consistent in their choice across

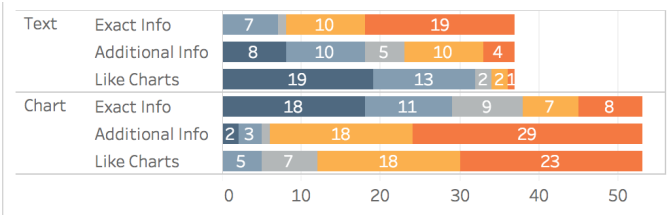

Figure 4: Reasons for ratings in Study 1 for participants who preferred text versus chart responses (dark blue = strongly disagree, light grey = neutral, dark orange = strongly agree). Each bar compiles counts across both B1 and B2.

blocks: 40 (89%) kept the same preference, 3 switched from text to chart, and 2 switched from chart to text. People who preferred text appreciated seeing a more concise answer to their question (the information requested and little more), whereas those who preferred charts appreciated both the form of presentation and the extra information provided.

Among the text group, there was a clear preference for showing the raw values of the two target items (Text - Values condition) over the other two text variants. In the bar chart group, participants preferred to see many bars over just two (Block 1) and additionally preferred to see the difference annotated on the chart (Block 2).

Free form answers agreed with the rating scale findings and offered additional insight. Participants who chose text options liked the simplicity and felt additional information was unnecessary or excessive. For example, "*It's precise and gives me enough info without too many details or too few*" and "*It was easiest to understand and the answer was not overly complicated*." Text - Values was the most preferred text format in both blocks. Text - Basic was considered too simple and Text - Diff was more complicated to understand (e.g. "*[Text - Values] answers the question perfectly. [Text - Diff] was awkwardly worded*)."

Participants who chose chart answers usually preferred to see many bars over just two. The extra information provided context and pertained to future potential questions. For example, "*At first I thought I just wanted a simple answer - 15 and 8. But after seeing [Bar - Many], I realized I really like getting the answer in context. And one question often leads to another, so I already have answers about other sports that have more or fewer events than the two I originally asked about.*". In Block 2, more participants chose the chart with the annotated difference. These participants found the extra information helpful and appreciated not having to do the math themselves. The size and similarity of the numbers may also play a role in people's preferences. For instance, "*The numbers are so close I don't want to have to do the math myself, I like it that the bot gives me the difference, and I like comparing to other events since I plan on attending more than one event.*" and "*With such large numbers, the graph and annotation help.*"

3.4 Study 2: Trend and Superlative Questions

The second study was structured identically to Study 1, but examined different questions and stimuli. For B3, we examined how participants would react to the additional information that is shown in major web search engines for queries such as "What is the population of Peru?" (see Figure 2). We compared two versions of purely textual responses, a single line graph for the population over time, and three line graphs, one with the target country and two others providing context.

The purpose of B4 was to determine an appropriate response for the superlative form of a comparison [12]. When someone asks for the best, worst, most expensive, least dangerous, etc., should a single item be the answer, or is this an implicit request for a comparison? The question asked was "You and a friend are planning a vacation

	Description	B3	B4
Text - basic	Direct answer only: B3: "In 2017, the population of Peru was 32 million." B4: "According to Wikipedia, Russia has the largest forested area, at more than 8 million squared km."	x	x
Text - offer	Adds to Text-Basic an offer of more information: B3: "Would you like to know the population of nearby countries?" B4: "Would you like to know the next largest countries?"	x	x
Line - 1	Text-Basic with line chart showing a trend across time.	x	
Line - 3	Adds context to Text-Basic "Check out this chart which also shows 2 nearby countries." Shows three line charts across time and 3 country names (Figure 2) .	x	
Bar - 3	Text - Basic with Bar chart with 3 values focus entity highlighted.		x
Bar - many	Text - Basic with Bar chart with 14 values with top bar highlighted.		x

Table 2: Conditions in Block 3 (B3) and Block 4 (B4) from Study 2.

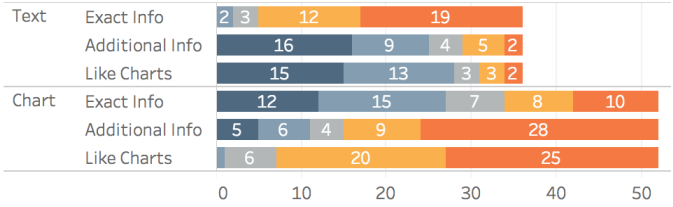

Figure 5: Reason ratings in Study 2 for participants who preferred text versus chart responses (dark blue = strongly disagree, light grey = neutral, dark orange = strongly agree).

and are trying to decide what country to visit. Assume you ask: Which country has the most forested land?"

3.4.1 Method

Tasks and Procedure: The same as for Study 1. The conditions shown in each block are summarized in Table 2. Average completion time was 4.6 minutes.

Participants: 50 crowdworkers participated; we removed 6 of these from the analysis based on off-topic free-form responses, resulting in 44 responses for analysis.

3.4.2 Results

A summary of preferred system responses for Blocks B3 and B4 are shown in Figure 3. As in Study 1, participants split into two distinct groups: those who preferred text alone and those who preferred bar or line charts in addition to the text $\chi^2(1, N = 44) = 17.9, p < .001$). As in Study 1, most people were consistent in their choices: 36 (82%) kept the same preference, 6 switched from text to chart, and 2 switched from chart to text. The distribution of ratings is also strikingly similar to that of Study 1 (see Figure 5), as are the reasons given in the text responses.

Again, free form answers agreed with the rating scale findings and offered additional insight. For the direct answer question ("What is the population of Peru?"), respondents who selected Text-basic for B3 made comments such as "*While the additional information is nice, it isn't necessary for the question.*" and "*If I wanted extra information like a chart or comparisons, I'd ask for that.*" Comments associated with Text-offer included "*I don't need to see a chart, but I do appreciate the bot asking if I want to know the population of nearby countries.*"

Block B3 contrasted a graph with just one line versus showing an additional two lines of context. Line-3 was twice as popular as Line-1. One person who selected Line-1 wrote as justification "*It offers slightly more info that option a*" while another wrote "*C answered the question ... and the bonus graph showing growth was interesting. Don't care about the other countries.*" Those who chose Line-3 appreciated the extra context, using justifications such as "*some context on nearby countries, more detailed, shows some other countries for comparison, population of nearby countries is really helpful, more info than I asked for, but it's still useful, more information is always better.*"

The superlative question of B4 ("Which country has the most forested land?") received a somewhat lower proportion of selections for text-only responses. Most of the difference can be attributed to five participants switching from a text option in B3 to Bar-3 in B4. An example of justification for switching is: [B4] *I don't mind the chart in this instance. It adds some context to how much forested area Russia has.* Participants were evenly split between Bar-3 and Bar-Many. Those who preferred Bar-3 stated that Bar-Many was overwhelming or otherwise showed too much information. Those who preferred Bar-Many mentioned statements of the more-information-is better variety.

4 Discussion

Overall, these studies found a strong difference in preferences for seeing charts in conversational interfaces across question types, with text alone preferred 41% of the time, and almost consistently within individuals. The results also found, for those who appreciate charts, a preference for more rather than less context for the choices shown. For instance, for B1, more than half the participants wanted to see additional sports beyond the two named in the question.

Further work is needed to better uncover the nuances of context. For instance, did participants in B4 switch from text to charts because the top answer (Russia) was not acceptable to them in this case? Experimenting with a different top answer could be enlightening.

The studies themselves have several limitations. We assumed the chatbot understood and answered the users' questions correctly, but in real life usage, an automated interface often makes errors. We asked participants to assume they were using a chatbot when formulating their responses, but their preferred answers to our predefined questions may not reflect actual use, where people bring their own questions and information needs. However, the fact that we observed a consistent distribution of answers across participants and question types suggests that our method of questioning is likely to be sound and reproducible. Our results also may not extend to conversational styles other than chat or when there is more screen real estate (our stimuli suggested a phone-sized device) or to a more realistic dialogue with multiple turns between bot and human. More studies could determine the limits to which people wish to see additional context, i.e., how many bars are too many for a chatbot versus for a more standard graphical user interface.

5 Conclusions

This work provides preliminary evidence via empirical studies that appropriateness for including visualizations in conversational interfaces may be determined more by personal preferences than other factors. The studies also showed that, for the types of questions asked, participants who do wish to see charts tended to prefer additional context beyond the exact answers to the questions asked.

These results have implications for the design of data-oriented conversational interfaces. Charts containing contextual information might be a reasonable default since they were preferred by more people. However, given the diversity of user preferences, such systems should offer personalization options, perhaps learning from user feedback on both the presentation format and information quantity.

REFERENCES

[1] P. B. Brandtzaeg and A. Følstad. Why people use chatbots. In *International Conference on Internet Science*, pp. 377–392. Springer, 2017.

[2] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013.

[3] L. B. Chilton and J. Teevan. Addressing people's information needs directly in a web search result page. In *Proceedings of the 20th international conference on World wide web*, pp. 27–36. ACM, 2011.

[4] E. Cutrell and Z. Guan. What are you looking for?: an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 407–416. ACM, 2007.

[5] E. Fast, B. Chen, J. Mendelsohn, J. Bassen, and M. S. Bernstein. Iris: A conversational agent for complex tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 473. ACM, 2018.

[6] A. Følstad and P. B. Brandtzæg. Chatbots and the new world of hci. *interactions*, 24(4):38–42, 2017.

[7] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pp. 489–500. ACM, 2015.

[8] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[9] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 203–212. ACM, 2010.

[10] E. Hoque, V. Setlur, M. Tory, and I. Dykeman. Applying pragmatics principles for interaction with visual analytics. *IEEE transactions on visualization and computer graphics*, 24(1):309–318, 2018.

[11] J. Hullman, E. Adar, and P. Shah. The impact of social information on visual judgments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1461–1470. ACM, 2011.

[12] N. Jindal and B. Liu. Identifying comparative sentences in text documents. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 244–251. ACM, 2006.

[13] M. Kaisser, M. A. Hearst, and J. B. Lowe. Improving search results quality by customizing summary lengths. *Proceedings of ACL-08: HLT*, pp. 701–709, 2008.

[14] E. Kaufmann and A. Bernstein. Evaluating the usability of natural language query languages and interfaces to semantic web knowledge bases. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(4):377–393, 2010.

[15] F. Li and H. Jagadish. Constructing an interactive natural language interface for relational databases. *Proceedings of the VLDB Endowment*, 8(1):73–84, 2014.

[16] J. Lin, D. Quan, V. Sinha, K. Bakshi, D. Huynh, B. Katz, and D. R. Karger. What makes a good answer? the role of context in question answering. In *Proceedings of the Ninth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2003)*, pp. 25–32, 2003.

[17] T. Paek, S. Dumais, and R. Logan. Wavelens: A new view onto internet search results. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 727–734. ACM, 2004.

[18] R. A. Pazos R, J. J. González B, M. A. Aguirre L, J. A. Martínez F, and H. J. Fraire H. Natural language interfaces to databases: an analysis of the state of the art. *Recent Advances on Hybrid Intelligent Systems*, pp. 463–480, 2013.

[19] A.-M. Popescu, O. Etzioni, and H. Kautz. Towards a theory of natural language interfaces to databases. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pp. 149–157. ACM, 2003.

[20] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 365–377. ACM, 2016.

[21] V. Setlur, M. Tory, and A. Djalali. Inferencing underspecified natural language utterances in visual analysis. In *IUI*, 2019.

[22] A. Srinivasan, M. Brehmer, B. Lee, and S. M. Drucker. What's the difference?: Evaluating variations of multi-series bar charts for visual comparison tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 304. ACM, 2018.

[23] A. Srinivasan, S. M. Drucker, A. Endert, and J. Stasko. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE transactions on visualization and computer graphics*, 25(1):672–681, 2019.

[24] A. Srinivasan and J. Stasko. Natural language interfaces for data analysis with visualization: Considering what has and could be asked. In *Proceedings of EuroVis*, vol. 17, pp. 55–59, 2017.