Data Prophecy: Exploring the Effects of Belief Elicitation in Visual Analytics

Ratanond Koonchanok
Indiana University-Purdue University Indianapolis
rkoonch@iu.edu

Parul Baser
Indiana University-Purdue University Indianapolis
pbaser@iu.edu

Abhinav Sikharam
Indiana University-Purdue University Indianapolis
asikhar@iu.edu

Nirmal Kumar Raveendranath
Indiana University-Purdue University Indianapolis
niraveen@iu.edu

Khairi Reda
Indiana University-Purdue University Indianapolis
redak@iu.edu

ABSTRACT

Interactive visualizations are widely used in exploratory data analysis, but existing systems provide limited support for confirmatory analysis. We introduce *PredictMe*, a tool for belief-driven visual analysis, enabling users to draw and test their beliefs against data, as an alternative to data-driven exploration. PredictMe combines belief elicitation with traditional visualization interactions to support mixed analysis styles. In a comparative study, we investigated how these affordances impact participants' cognition. Results show that PredictMe prompts participants to incorporate their working knowledge more frequently in queries. Participants were more likely to attend to discrepancies between their mental models and the data. However, those same participants were also less likely to engage in interactions associated with exploration, and ultimately inspected fewer visualizations and made fewer discoveries. The results suggest that belief elicitation may moderate exploratory behaviors, instead nudging users to be more deliberate in their analysis. We discuss the implications for visualization design.

CCS CONCEPTS

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

KEYWORDS

Belief elicitation, confirmatory analysis, sensemaking

ACM Reference Format:
Ratanond Koonchanok, Parul Baser, Abhinav Sikharam, Nirmal Kumar Raveendranath, and Khairi Reda. 2021. Data Prophecy: Exploring the Effects of Belief Elicitation in Visual Analytics. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445798

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445798

1 INTRODUCTION

Visualization tools have become vital instruments to data science. These interactive analysis systems enable users to explore sets of data and look for patterns that might indicate new insights. However, existing visualization tools typically come only with data-driven interactions, providing no explicit support for confirmatory analyses. In particular, current tools do not provide affordances for users to share their working hypotheses, and test the accuracy of those hypotheses before peeking at the data.

Statisticians have long recognized a need for both exploratory and confirmatory analyses [48], with the choice of method dependent on the question at hand and the status of one's working knowledge. Research in cognitive science has also emphasized the importance of belief-driven reasoning, wherein people attempt to proactively test the fit of their mental models against observable data. For instance, Dunbar showed that scientific discovery usually occurs through a process of conceptual mismatch, whereby an analyst observes a discrepancy between their expectations and the evidence [7]. It is often by actively seeking to reconcile such mismatches that people begin to make new discoveries [8]. Similarly, Klein et al. observe that model-fit testing is key to sensemaking, arguing that most people seek to (dis)confirm and adapt their existing frames, as opposed to developing entirely new frames from scratch, even when faced with novel information [22]. This research suggests that, to be maximally effective, visualizations must also support a confirmatory approach to analysis, in addition to the traditional role as data-driven sensemaking tools. Addressing this gap could also serve to reduce the incidence of spurious discovery in visualizations [53], by fostering a healthy level of skepticism and grounding insights in prior beliefs.

Researchers have started to acknowledge the need to incorporate one's mental model as an essential aspect to reasoning with visualizations. For example, researchers tested the effect of eliciting prior knowledge from participants, and visualizing it alongside data to encourage reflection. This body of work suggests that knowledge externalization improves data recall [20], promotes normative Bayesian reasoning [21], and increases the communicative impact of narrative visualizations, if not their persuasiveness [15]. Yet, these studies were done under highly controlled experimental conditions, and using sparse datasets of a handful of data points. It is still unclear how belief elicitation can impact one's visual analysis

in realistic, open-ended scenarios. Furthermore, research is needed on how to design functional tools that can scaffold confirmatory analyses, while still providing the traditional suite of visualization interactions people have come to expect.

Our goal in this work is two-fold. First, we investigate how users structure their visual analysis while interacting with a system that supports belief externalization, as a way of learning about and testing one's knowledge against data. Second, we contribute a perspective on how to redesign exploratory, multi-view visualizations to also support hypothesis-driven analyses. To that end, we present *PredictMe*, a tool that enables users to sketch their predictions in a variety of charts. These custom interactions are blended with traditional visualization functionalities, allowing for a mix of exploratory and confirmatory analyses in one platform. We report on an exploratory, between-subjects study of participants' cognition and interaction patterns. We compare our design against a control condition of the same tool that lacks the ability to draw expectations. Our results show that, given the opportunity, participants frequently chose to share their data expectations with the system, despite the overhead involved. Analysis of their think-aloud statements showed that they developed more hypotheses before peeking at the data, and were more attentive to flaws in their mental models. However, those same participants inspected fewer visualizations on average, and ultimately developed fewer observations about the data. The results suggest that belief elicitation may have a moderating effect on exploratory behaviors, instead nudging participants to be more deliberate in their queries. We discuss these findings, and address the potential benefits and complications of incorporating belief-driven interactions in visual analytics tools.

2 BACKGROUND AND RELATED WORK

2.1 Exploratory versus Confirmatory Analyses

A hallmark of good science is the ability to attend to unexpected results. Indeed, some of the most prominent breakthroughs in the history of science, such as the discovery of Penicillin [11], occurred by chance when scientists saw surprising results, and were subsequently able to reinterpret those findings in new ways. To maintain an open perspective, analysts typically prescribe Exploratory Data Analysis (EDA) as an integral step in the data analysis pipeline [14, 48]. EDA is a process of looking for interesting distributions, outliers, and relationships, which can then be used to formulate new hypotheses or devise additional experiments [47]. Historically, EDA has relied heavily on visualization tools, which provide the sort of flexibility needed. Nevertheless, Tukey, who is largely credited with championing EDA, cautions against using it for "fishing expeditions" [10]. He notes that accepting findings from EDA as conclusive insights is "destructively foolish" [47]. This is because a hypothesis or a pattern suggested spontaneously by a dataset is unlikely to be refutable by that same data. Instead, findings from EDA should be considered preliminary, requiring confirmation with an independent data source.

By contrast, in confirmatory analyses, hypotheses are posited (and ideally preregistered [29]) before the data is seen. When data is tested against a preconceived prediction or model, and found to conform, that model (and its underlying hypothesis) can be said to be confirmed. Confirmatory analysis is considered the standard inferential method in science; inferences made are generally reliable, as long as quality of the data is controlled and the sample is reasonably representative of the underlying population. The key reliability indicator, however, is that hypotheses are posited prior to peeking at the data, (i.e., before the outcome is known) [19]. It is possible to view both exploratory and confirmatory analyses as instances of model check: the analyst compares the visualized data to an imagined dataset sampled from an (implicit) reference model [12, 16]. This comparison could then prompt a Bayesian update to revise the reference model, or, alternatively, a classical hypothesis test wherein the difference between the imagined and visualized data is adjudicated using a (visual) test statistic. However, others still maintain that exploration and confirmation should be conceptually separated in order to ensure the robustness of discoveries [6].

The distinction between exploration and confirmation (or lack thereof) is of specific concern for visual analytics. Visualization users appear to frequently accept results generated through EDA as conclusive, leading to spurious findings that may not generalize beyond the sample data at hand. For example, in a startling result, Zgraggen et al. found the majority of discoveries uncovered through interactive visual analysis to be false [53]. It has been suggested that the way visualization tools are currently designed serves to further blur the boundary between potentially robust confirmatory findings and preliminary, exploratory results [35]: as users interactively filter, bin, and slice-and-dice their data, they make a myriad inferences with just a few clicks, often without being aware of the effects of this multiplicity on the reliability of inferences [13]. Zhao et al. devised an "α-investing" approach to account for multiple comparisons during interactive analysis [54]. Jo et al. allow users to leave 'safeguard' annotations on uncertain visualizations, indicating that those visualizations need to be rechecked once the complete data is available [18]. These interventions may reduce the incidence of spurious discovery in visual analytics. However, the lack of clear hypothesis- and model-testing capabilities in visualization tools can still leave people overconfident in their analysis strategy. Preliminary evidence suggests that users could indeed benefit from such affordances [2, 36, 38]. Our work addresses this gap by proposing workflows and interactions that can be used for confirmatory and model-driven analyses in visualizations. We also study how the presence of these interactions affects user behavior and analysis patterns.

2.2 Sensemaking with Visualizations

Sensemaking refers to a "class of activities and tasks in which there is a native seeking and processing of information to achieve understanding about some state of affairs" [23]. Theories of sensemaking have been a recurring theme in visual analytics, and have contributed heavily to the development of the field [4]. Among the most commonly cited models is Pirolli and Card's [33], which comprises the following sensemaking activities: analysts iteratively filter their data, select and highlight relevant evidence, and reorganize that evidence in a 'schema'. A schema can then be used to induce hypotheses to explain the data or to take decisions. Visualization designers have taken inspiration from this model. For example, Jigsaw divides its interface into several components, each

with interactions intended to support a specific sensemaking activity (e.g., 'evidence marshaling') [44]. Shrinivasan and Wijk provide a 'knowledge editor', enabling users to record their hypotheses and conclusions in the form of a concept graph [43]. Schemaline aids analysts in schematizing temporal events [28].

Although many visualization tools have been custom-designed to mirror empirical sensemaking models, these tools are primarily intended to facilitate 'bottom-up', data-driven sensemaking. By comparison, no tools exist to specifically support top-down, expectation-guided visual analyses (e.g., as espoused by Klein et al's data-frame theory [22]). Some research has sought to develop systems that adapt to user models in real-time. For instance, semantic interaction can deduce conceptual relationships by observing how users manipulate spatial layouts [9]. This information is then used to evolve the visualization to match analyst beliefs. Such techniques, however, are limited to inferring implicit, low-level features (e.g., pairwise multidimensional distance [50]), and are primarily meant to influence computational processes running in the background. As such, these techniques do not provide explicit hypothesis-testing affordances that people could use outright to validate their mental models and beliefs. Our work ultimately aims to re-architect visual sensemaking tools to equally support both data- and belief-driven (i.e., confirmatory) analyses.

2.3 Belief Elicitation in Visualization

Belief elicitation is the process of externalizing implicit knowledge (typically of experts) about some unknown quantity, and distilling that knowledge into a probability distribution [30]. These distributions are often used as prior models, which are then updated with new (typically empirical) data using a Bayesian framework. Despite the long history, the visualization community has only recently begun to incorporate user beliefs in data graphics. Practitioners have started experimenting with interactions that invite audience to externalize their beliefs by sketching in charts. For example, the New York Times featured a series of visualizations that invited the viewer to predict the impact of the Obama presidency on various socioeconomic indicators [31]. The viewer sketches the expected trend line by drawing in an initially blank chart. The actual timeseries are then revealed, enabling the viewer to compare the accuracy of their sketch and, accordingly, update their beliefs. Kim et al. studied this kind of interaction in a controlled study, finding that it improved participants' data recall [20]. They also proposed belief elicitation as an evaluation method by considering the degree to which visualizations promote normative Bayesian update among viewers [21]. Heyer et al. studied how people adjust their attitudes towards a message experienced through a narrative visualization [15]. They found that prior elicitation does not significantly impact attitudinal change, even though it is correlated with other knowledge acquisition metrics. Choi et al. conducted a Wizard-of-Oz study to explore whether natural language can be used to specify prior beliefs [2]. They subsequently developed a tool that allows users to frame hypotheses in natural prose, and accordingly receive visualizations tailored to their beliefs [3]. Sarma and Kay investigated how Bayesian statisticians set their priors [42]. They documented varying strategies and philosophies practitioners seem to draw upon when distilling subjective beliefs into prior distributions.

Empirical work on visual belief elicitation have so far utilized highly controlled experiments, surveys, or interview methods. Though informative, the results may not necessarily translate to visual analytics, where analysts engage in fluid, open-ended sensemaking, and have a choice to either specify priors or proceed in an exploratory fashion. Our work contributes insights on how users might behave in such contexts.

3 METHODOLOGY

Our goal is to broadly understand how users might interact with a visualization tool that supports belief-driven analysis. In this work, we specifically address two research questions:

- Given the opportunity to visually externalize their expectations, how often will people use this feature?
- How do users react to seeing their expectations represented alongside data? And how will the ability to test one's predictions affect their visual analytic process?

To investigate these two questions, we conducted a comparative, exploratory study. We recruited participants who have prior data analysis experience, and tasked them with visually analyzing two data sets that we provided. Participants were randomly assigned to one of two conditions. A **Prediction** condition consisted of an interface that provides belief elicitation affordances, optionally enabling participants to sketch their predictions into charts, and compare these sketches to visualized data. A second **Standard** condition provided all of the interactions available in the former condition, but otherwise lacked the prediction functionality. We describe the design of the visualization. We then discuss the study procedures and analysis methodology.

3.1 Visualization

Since there are no established visualization tools that support knowledge externalization, we created a custom-designed tool for this study, which we dub *PredictMe*. The design of PredictMe was inspired by existing visualization systems, such as Vizdom [5] and ExPates [17], and by results from formative studies on belief elicitation in visual analytics [2]. The interface allows users to create data views on demand; users drag data attributes from a side panel and release them onto an initially empty canvas to create charts. Multiple charts can be created, resized, and positioned freely within the canvas. Additional attributes can also be added to an existing chart by dragging onto placeholders. The tool supports five visualization types: bar charts, histograms, scatterplots, line graphs, and parallel coordinates plots. Chart type is determined based on the number of attributes and their types. For example, a single qualitative attribute produces a bar chart, whereas a quantitative attribute results in a histogram. Two quantitative attributes are visualized as a scatterplot. Combining a quantitative with a temporal attribute results in a line graph. Lastly, a parallel coordinates plot can be generated by incorporating two or more attributes. In addition to creating charts, users can brush-and-link by selecting data points from one chart and seeing their distribution highlighted in other charts. Figure 1 shows an overview of the interface.

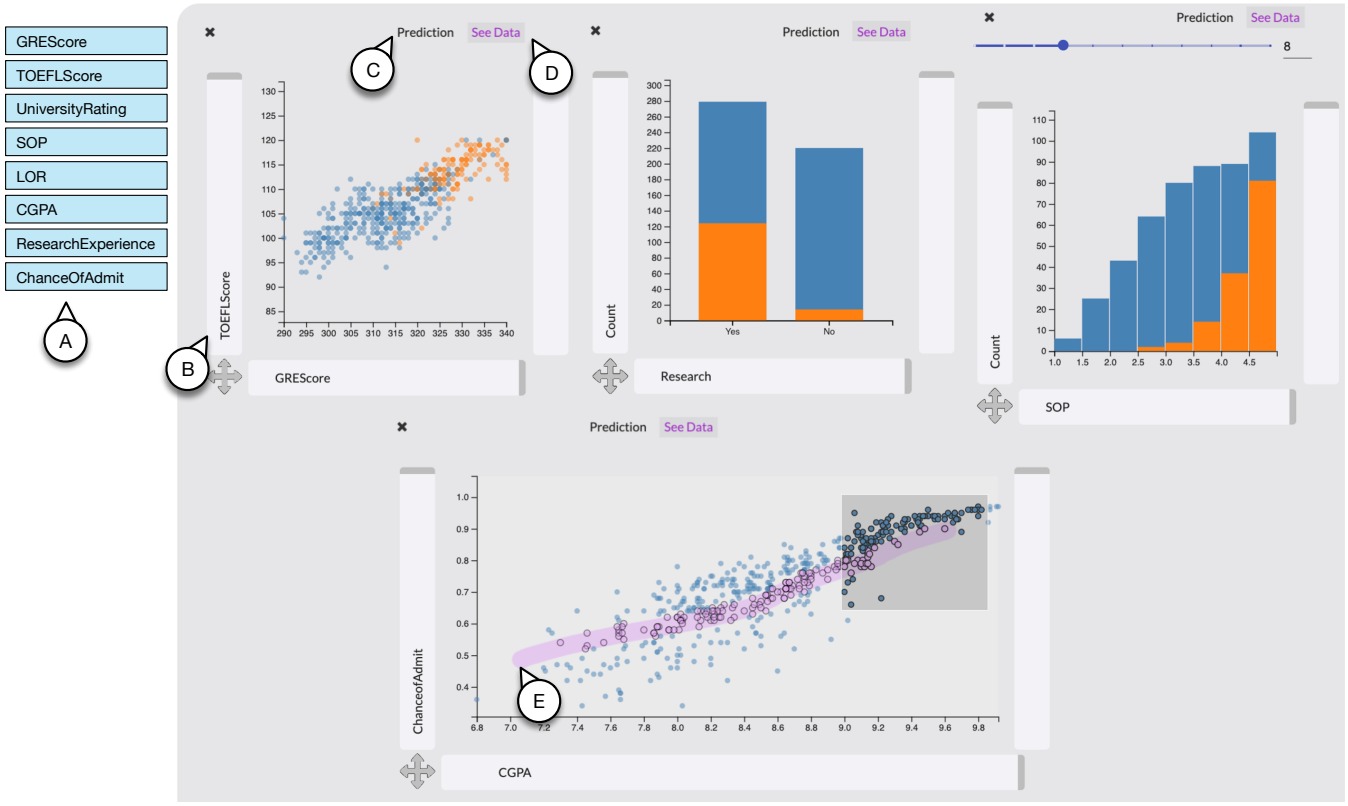

Figure 1: Overview of the *PredictMe* interface. Data attributes are displayed on the left (A). Charts can be created by dragging attributes onto a canvas (B). Newly created charts are initially set to a 'Prediction' mode (C), enabling the user to sketch their expectation, but can be then switched to a 'See Data' mode (D). The scatterplot at the bottom contains a sketched prediction in violet (E), which is contrasted with the actual point cloud (blue).

A key difference with existing tools is the ability to sketch one's expectations *prior* to seeing data. PredictMe then displays those expectations alongside the data. The sketching feature generally works by first presenting the user with initially blank charts: when creating a new chart, users see labeled axes and data ranges, but without actual data points. The user can then optionally sketch into the chart to outline the pattern they expect to observe. The precise sketch interaction is dependent on the chart type: for histograms and bar charts, predictions are specified by adjusting the length of bars, which are initially set at a baseline height. In doing so, the user specifies the frequency of individual bins in a histogram, or the value associated with a qualitative attribute. Figure 2 illustrates this interaction sequence. For line charts, the user draws with a pencil tool to outline the expected trend for a timeseries. Scatterplots come with a paintbrush that can be used to predict the density of the point cloud. Lastly, in parallel coordinates, the user predicts by specifying intervals on the parallel axes, effectively creating ribbons to designate the expected multi-variate pattern. Figure 3 illustrates these different sketching styles. In designing these interactions, we took inspiration from Kim et al's taxonomy [20], as well as from examples developed by practitioners [1, 31].

After entering their expectations, users click a 'See Data' button. This causes the actual data to be revealed in the chart and shown alongside the sketch (see Figures 1-E & 2). For distinction, expectations are consistently color-coded in violet, whereas data marks are always shown in blue. Specifying expectations is optional: the user may choose to skip this step by immediately clicking 'See Data'. All charts are initially set to a 'Prediction' mode, giving users the opportunity to specify expectations.

The sketch feature was only available in the Prediction condition. However, to give participants in the Standard condition an equal opportunity to reflect on their prior knowledge, data display is also delayed, with newly created charts shown blank. Participants in the Standard condition similarly had to click 'See Data' to reveal chart contents, even though they could not draw a prediction. This extra step enabled us to capture verbal predictions participants may have uttered prior to being exposed to the data.

3.2 Participants

We recruited 24 participants from a large, public university campus. All participants had prior data analysis experience (e.g., using Excel, R, Tableau, or SAP), and represented a range of analytic disciplines, including computer science, statistics, and data science. We compensated participants with a $20 gift card upon completing the

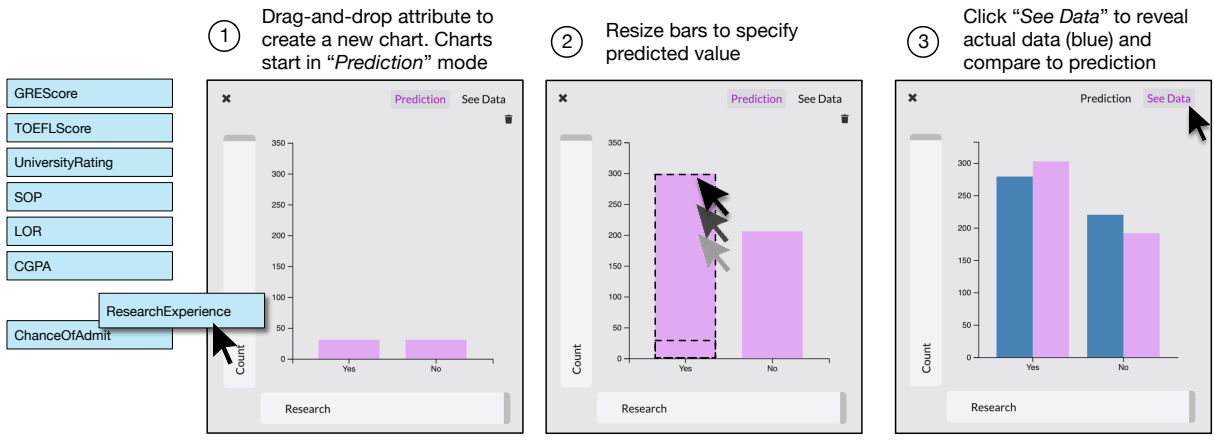

Figure 2: Illustration of the interaction sequence for specifying data expectations in a bar chart.

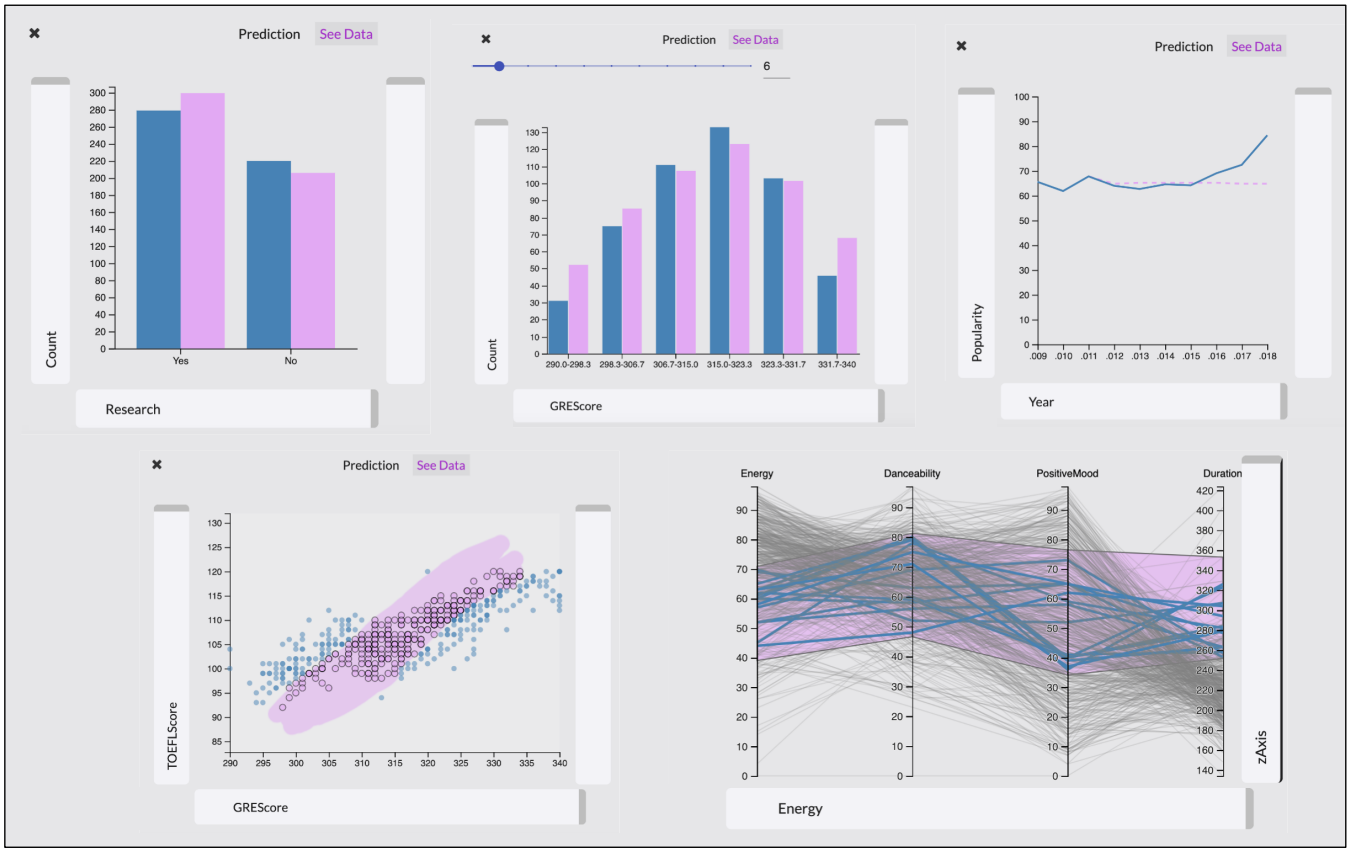

Figure 3: PredictMe supports different data sketching styles based on the chart. In histogram and bar charts, expectations are specified by adjusting bar lengths. Line charts provide a pencil tool to draw the expected shape of a timeseries. In a scatterplot, the expected point cloud density can be specified using a paintbrush. Lastly, in parallel coordinates, the expected multi-variate pattern is designated by specifying intervals on the vertical axes. Expectations are color-coded in violet to distinguish from data marks (blue). Data points that fall within the expectations are also visually differentiated from those that deviate.

study. In addition to the 24 participants, we piloted the study with 3 participants whose data were excluded from the analysis.

Participants were assigned randomly to one of the two conditions (Prediction or Standard), for a total of 12 participants in each. We refer to participants in the Prediction condition by P_i and those in Standard by S_j. Thirteen participants completed the study in-person; they interacted with the visualization through a standard desktop setup (i.e., mouse, keyboard, and a full-HD monitor). For the remainder 11 participants (7 in Standard and 4 in Prediction), the study was conducted remotely (a change prompted by the COVID-19 pandemic). Those latter participants were provided with a web link to the visualization. They completed the study using their own computers, sharing their screen content with the experimenter via Zoom. Notwithstanding the change in format, we maintained identical procedures across the in-person and remote sessions.

3.3 Procedures

Our goal was to place participants in an open-ended visual analysis context. We therefore adopt the setup employed in insight-based evaluation methodologies [34, 41]. As such, we did not provide participants with specific tasks or questions to answer. Rather, participants were instructed to freely analyze the provided data, by developing their own hypotheses and lines of question. We told participants that they may share their beliefs (either verbally or through sketch) particularly if they had expectations of what the data might look like, but that they may also skip this step if they wish. Recall that in both conditions, charts are initially blank, which gave subjects in the Standard condition an opportunity to verbally externalize their beliefs.

Participants were first given a demonstration of the visualization tool using brief example scenarios. During this demonstration, the experimenter showed participants examples of how they might externalize their beliefs. This was done by sketching in the Prediction condition, or by verbalizing in Standard (e.g., "I predict X might increase with Y...") prior to clicking the 'See Data' button. To avoid biasing participants, the demonstration employed a different dataset from those participants were tasked with analyzing. Following the demonstration, participants conducted two separate analysis sessions using two different datasets. The datasets were acquired from *kaggle.com*, an open source dataset repository. The first dataset contained statistics of student admissions to select US graduate programs, comprising attributes such as the student's GPA, test scores, and research experience, among others. The second dataset comprised statistics about top music songs in the past decade, with attributes such as genre, danceability, loudness, and popularity. The datasets were chosen as they represent common knowledge to a university community (admission process) as well as data about popular culture (music), thus providing participants with attributes they are likely to have some prior knowledge about.

At the beginning of each analysis session, participants were given a data sheet containing a brief description of each set, including size, data types, and column definitions. Participants were given a few minutes to read the data sheet before starting their analysis. We allocated 30 minutes per dataset, although participants were at liberty to stop earlier if they ran out of ideas. Alternatively, they could extend the session longer if they felt they needed more time for their analysis. An experimenter was present throughout to answer participants' questions and proctor the study. The experiment was audio recorded and the contents of participants' screen were captured in video. We instructed participants to think aloud and verbalize their thought throughout.

3.4 Analysis, Segmentation, and Coding

We first transcribed participants' utterances and segmented them using standard verbal protocol analysis methods [46]. Segments consisted of independent clauses that can be understood on their own. We then grouped related segmented into 'queries'. A single query comprised a self-contained line of analysis with one or more associated visualizations, and with typically multiple verbal statements. The segmentation process resulted in a total of 2,728 segments, and 651 unique queries. The average number of queries per participant was 27.

To analyze participants' verbal utterances and reactions, we developed a coding scheme using a grounded theory approach [45]. Two coders inductively coded the segmented data. The coders consulted the video recording to resolve any ambiguities in the process. Throughout, the emerging coding scheme was revised iteratively and discussed regularly with members of the research team. After finalizing the code book, the entire dataset was then re-coded using the final scheme. We subsequently measured coding reliability by having the two coders redundantly and independently code 60 segments (one entire analysis session from a randomly selected participant). Inter-coder agreement was measured at 92.64%, with a Cohen's kappa of 0.9144, indicating excellent agreement between the two coders [26].

The codes were divided into three orthogonal categories: Expectations, Assessments of Data-Expectation Fit, and Reactions. Expectations comprised three codes designating the point at which a participant supplied predictions: *before* or *after* inspecting the data, or whether they chose to not provide a prediction for a particular query. Data-Expectation Fit indicates the degree to which a participant's expectation was confirmed or contradicted by data, as self-assessed by the participant. Lastly, Reactions comprised verbal statements uttered either before or after inspecting visualizations. This latter category included insight-related codes, such as *Observations* and *Hypotheses* [41]. We also distinguish between hypotheses verbalized *before* or *after* the relevant data is seen by a participant. Lastly, we coded Reactions that are indicative of certain cognitive activities, including *Goals*, *Reasoning*, *Surprises*, and *Belief Updates*. The complete coding scheme is available in the supplementary materials. We also include the transcribed and coded data.

4 RESULTS

We first report on differences in analytic behaviors and insight acquisition across the two conditions. We then analyze variations in participants' interaction patterns. Given that our study is *exploratory* in nature, we refrain from making generalizable statistical inferences. Instead, we present our results (with confidence intervals) as exploratory findings requiring confirmation in future experiments.

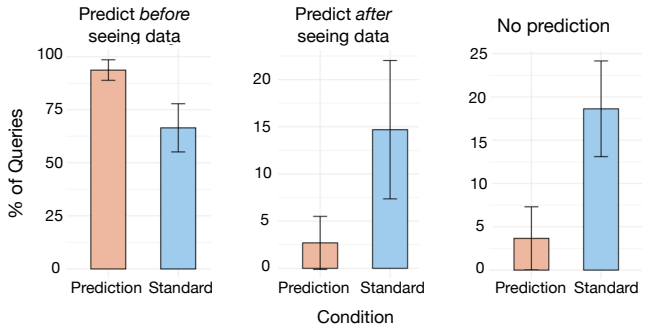

Figure 4: Average percentage of queries in which predictions were made *before* (left) or *after* seeing the data (center). Queries in which no predictions were made are shown on the right. Error bars represent 95% confidence intervals.

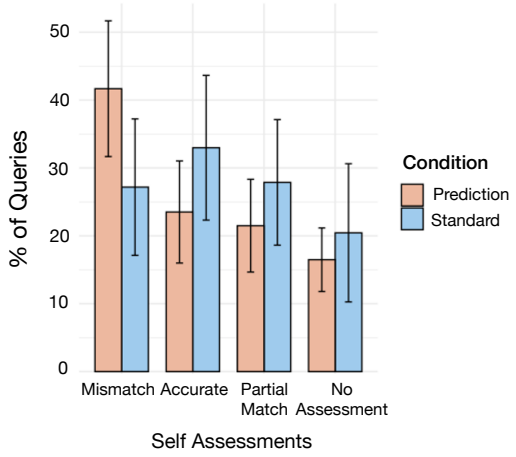

Figure 5: We coded three types of self-assessments to capture how often participants thought their predictions were Accurate, Mismatched, or Partially Matched the data. Error bars are 95% CI.

4.1 Analytic Behaviors and Insights

We consider differences in the number of queries, predictions, hypotheses, and observations generated by participants. To mitigate the effects of inter-participant variation, we compare averaged, normalized rates where appropriate, by counting code occurrences per subject and dividing by the total number of coded segments for that subject.

4.1.1 Queries & Predictions: We report the number of queries participants made as an approximation to the unique lines of analysis developed during the study. Participants in the Prediction condition made fewer queries on average (22.6, 95% CI: 18.5–26.6), compared to those in Standard (33.1, CI: 25.8–40.3). In each query, a participant can decide to provide a prediction *before* seeing the data, state their prediction *after* seeing the data, or simply explore the data *without* supplying any prediction. Figure 5 depicts the average tendency for these three alternatives. On average, 93.6% of queries (CI: 88.8–98.5%) in the Prediction condition included a prediction prior to data revelation, compared to 66.4% (CI: 55.1–77.8%) in Standard. By contrast, participants in the Standard condition were approximately 6 times more likely to predict after inspecting data (14.7% of queries, CI: 7.4–22% versus 2.7%, CI: 0–5.5% in Prediction). Similarly, participants in the Standard condition were 5 times more likely to not specify predictions (18.6%, CI: 13.1–24.2% versus 3.7%, CI: 0–7.3% in Prediction).

4.1.2 Assessing Data-Expectations Fit: Participants making predictions (either by sketching or by verbalizing their expectations) usually follow with an assessment of how accurate their predictions were. We identified and coded three types of self-assessments: Accurate, Partial Match, and Mismatch. A forth category (No Assessment) indicates no explicit assessment. Figure 4 depicts the average frequency of these codes across the two conditions.

Participants in the Prediction condition declared a Mismatch between their expectations and the data more frequently (41.7% of queries, CI: 31.7–51.7%) compared to those in the Standard condition (27.2%, CI: 17.1–37.2%). As an example of a Mismatch, P10 used a histogram to test their knowledge of the TOEFL scores distribution. Upon inspecting the data, the participant observed that "actual scores [in the 110 range] are higher than what [they had] predicted." By contrast, participants in the Standard condition stated that their predictions were accurate in 33% (CI: 22.3–43.7%) of the time, compared to only 23.5% (CI: 16–31%) in Prediction. For example, after inspecting the numbers for male and female singers, S2 stated: "as I have guessed, there are more male than female singers." Similarly, there were more Partial Matches in Standard than in the Prediction condition (21.5%, CI: 14.7–28.3% in Prediction versus 27.9%, CI: 18.6–37.1% in Standard). A Partial Match indicates that, at least, some aspects of the prediction were realized. For example, participant S2 hypothesized that personal statement ratings have an identical effect on the chance of admission as do letters of recommendation. They later discovered that, while there were similarities, there were also differences in patterns that did not seem to align with their mental model: "the trend is similar, but with letter of recommendation ratings, the range is very wide compared to statement of purpose ratings."

4.1.3 Hypotheses & Post-Data Hypotheses: Hypotheses occur when a participant verbalizes a clear conjecture before seeing see the data. One key criteria for coding a statement as a hypothesis is the inclusion of an *explanation*, such as a justification for an expected correlation or a causal mechanisms through which one attribute influences another. For instance, participant P8 stated while sketching their expectation in a scatter plot: "These two variables [positive mood and dance-ability] should be correlated. If you are happy, you'd want to dance to the music." On the other hand, a Post-Data Hypothesis occurs when the verbalized conjecture is stated after the participant had seen the relevant data, typically as an explanation to something that had not necessarily been expected. For instance, participant P3 stated: "I actually believed if the University rating is good then chances to get a higher CGPA are more. I did not expect this result." Figure 6-top compares the rate of pre- and post-data hypotheses. On average, Hypotheses amounted to 19.7% (CI: 13.6–25.8%) of the total reactions in the Prediction condition

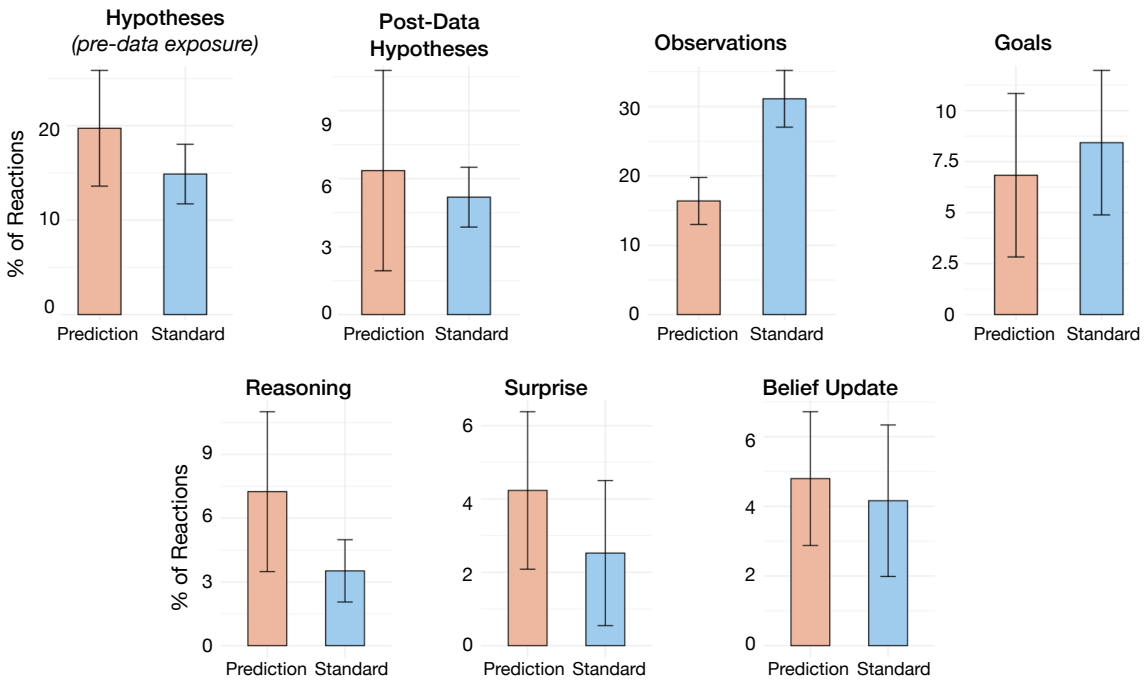

Figure 6: Observed rates for cognitive indicators by condition. The top codes that are typically associated with 'insights' [41], including Observations and Hypotheses. We specifically distinguish between Pre- and Post-Data exposure Hypotheses. The bottom codes include additional data cognition indicators, such as Belief Update, Surprise, and miscellaneous Reasoning. Error bars are 95% CI.

compared to 14.9% (CI: 11.7–18%) in Standard. Post-Data Hypotheses also occurred more frequently in Prediction than in Standard, even though the difference amounts to merely 1% of reactions (6.4%, CI: 1.9–10.8% in Prediction versus 5.2%, CI: 3.9–6.5% in Standard). The confidence interval for that former estimates are especially wide, suggesting wide variation among participants.

4.1.4 Goals. At the onset of a query, participants sometimes chose to verbalize a specific goal they have in mind. Goals can be seen as 'questions' the participant sought to answer, but without concrete expectations to be considered hypotheses. For example, participant S3 stated: "I want to see what genre is the most popular," before exploring the relationship between music genre and popularity. On average, 6.8% (CI: 2.8–10.9%) of reactions were coded as Goal in the Prediction condition compared to 8.4% (CI: 4.9–12%) in Standard (see Figure 6-top-right).

4.1.5 Observations. Observations occur when a participant attempts to draw an insight while inspecting a visualization. As such, and by definition, observations occur solely after the data is revealed. As an example, participant S2 explored the number of singers by gender throughout the last decade, observing "an increase in the number of female singers from 2010 to 2017." Figure 6-top shows the mean observations rate. Participants in the Standard condition made more frequent data-driven remarks, with 31.1% (CI: 27–35.2%) of their reactions coded as Observation, compared to 16.4% (CI: 13–19.8%) in Prediction.

4.1.6 Reasoning, Surprises, and Belief Updates. After inspecting the data, some participants provided rationale to substantiate or explain their conclusions. For instance, participant P9 discovered that students with a research experience seem to have a higher chance of being admitted to a graduate school. The participant subsequently provided a reason for this observation, stating that "since it's a graduate school, top universities would probably expect some research experience from applicants prior to joining their program." Figure 6-bottom illustrates the rates of statements coded as Reasoning. Participants in the Prediction condition were roughly twice as likely to provide rationale to support their discoveries than those in Standard (7.2%, CI: 3.5–11% versus 3.5%, CI: 2.1–5% of verbal reactions). In addition to providing rationale, participants could also update their belief to incorporate any new information they had uncovered. As an example, Participant P6 previously expected songs with lower 'loudness' to be more popular. After observing data to the contrary, they stated that "low noise songs are not at all popular. That's the different thing which I learnt." We did not find a difference between the two conditions in terms of Belief-Updates (4.8%, CI: 2.9–6.7% of verbal reactions in Prediction versus 4.2%, CI: 2–6.3% in Standard). In a few occasions, participants explicitly expressed surprise at the data. For instance, after finding out that songs by both male and female singers had roughly equal loudness, P9 said, "Surprisingly, they are the same." On average, 4.2% (CI: 2.1–6.4%) of reactions in the Prediction condition were coded as Surprise, compared to 2.5% (CI: 0.5–4.5%) in Standard.

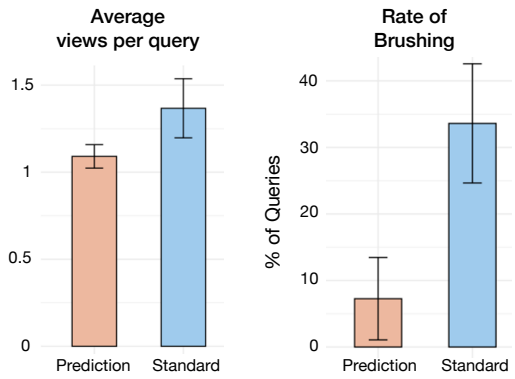

Figure 7: Average number of views created per query (left). The rate of brushing-and-linking interactions (right).

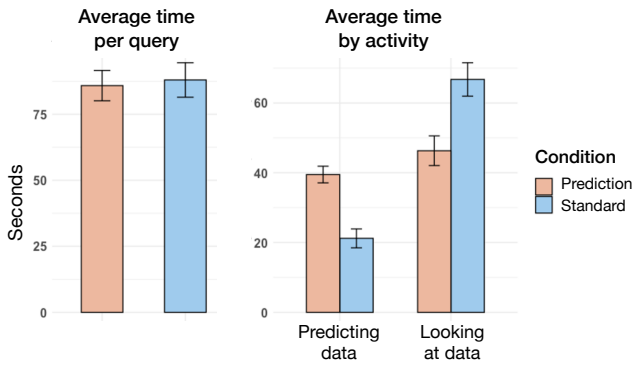

Figure 8: Average time spent per query (left). The same data is broken down by activity (predicting versus looking at data).

4.2 Interaction Patterns

We report differences in how participants utilized the interface. We focused on indicators that can be used as proxies to gauge participants' stance (i.e., confirmatory versus exploratory). Specifically, we consider the following metrics: number of views created, frequency of brushing-and-linking, and the amount of time spent looking at or predicting the data. These events were identified and coded manually from the video recordings.

4.2.1 Number of views: We counted the number of charts participants created as an indicator to the breadth of their analysis. Recall that participants had the freedom to create as many charts as they needed during a particular line of analysis. Having two or more views affords an opportunity to look for multi-variate relationships, either through visual comparison alone or by brushing-and-linking. Figure 7-left shows the average number of views created per query in the two conditions. On average, participants in the Prediction condition utilized 1.1 (CI: 1–1.2) views compared to 1.4 (CI: 1.2–1.5) in Standard. The latter group were thus more likely to look at multiple charts and, by extension, potentially consider a larger number of attributes in their queries.

4.2.2 Brushing-and-linking: A standard visualization feature that has come to be associated with exploratory analysis is brushing-and-linking [39, 52]. We measured the rate of brushing to understand how this feature might be used in a system that emphasizes belief-driven analysis. Figure 7-right shows the percentage of queries in which brushing was activated at least once. Participants in Standard utilized this feature five times more frequently compared to those in the Prediction condition (33.6%, CI: 24.6–42.6% versus 7.3%, CI: 1.1–13.4%). This may indicate a higher tendency to look for relationships across multiple views in the former. It may also reflect the fact that those in Standard were more likely to create multiple views, and hence activate the brush. Collectively, however, these two metrics (number of views and the rate of brushing) may indicate a higher propensity for data-driven exploration in the Standard condition.

4.2.3 Analysis time: Lastly, we measured the time spent by participants on each query. On average, participants in both conditions spent virtually equal amounts of time addressing a single query (85.8 seconds, CI: 80.1–91.6 in Standard versus 88, CI: 81.4–94.5 in Prediction), as shown in Figure 8. We broke these down into times spent 'predicting' or 'looking over the data'. Recall that even though participants in the Standard condition could not draw their predictions, they similarly saw blank charts initially, giving them to the opportunity to verbalize their expectations. Those in the Prediction condition spent nearly twice as much time making a prediction compared to subjects in the Standard condition (39.5 seconds, CI: 37.1–41.9 in Prediction versus 21.2, CI: 18.5–23.9 in Standard). However, and conversely, participants in the Standard condition spent approximately 20 seconds in extra time looking at the data (46.3 seconds, CI: 42.1–50.6 in Prediction versus 66.8, CI: 62–71.5 in Standard).

5 DISCUSSION

The results suggest marked differences in behavior and interaction patterns across the two conditions. We discuss the emerging variations, highlighting implications for design where possible.

5.1 Exploration versus Confirmation

The distinction between exploratory and confirmatory analyses is important for proper inference [10, 27, 47]. However, for users of interactive visualizations, it is often quite difficult to distinguish between the two styles of analysis [35]. Eliciting prior beliefs can help both users and systems discriminate between exploratory and confirmatory activities. Extant work suggests that sketching data expectations into charts is intuitive for most users [15, 20]. However, these earlier studies have been conducted under highly constrained settings and on very small datasets. A natural follow-up question is whether this kind of interaction might work in an open-ended, visual analytics context. Our study sheds light on this question. The results show that participants utilized the *PredictMe* feature in the majority of their queries. Specifically, 93.6% of queries in the Prediction condition came with concrete data expectation, which were externalized in the form of a graphical sketch. By comparison, in only 66.4% of queries in the Standard condition did participants verbalize their expectations prior to inspecting chart contents (recall that participants in Standard lacked the ability to sketch, but were

otherwise prompted and given the opportunity to verbally state their beliefs, should they want to). Those same participants were also more likely to state their beliefs *after* seeing the data (14.7% of queries). By contrast, only 2.7% of expectations were verbalized post-data exposure in the Prediction conditions.

Design implication: Our results suggest that belief elicitation through sketching is viable in the context of visual analytics, and could perhaps become a standard feature of interactive visualization systems. While such interaction can be burdensome in multi-view environments, as users would need to repeatedly sketch their beliefs in multiple charts, our study suggests that analysts may still embrace this feature. The *PredictMe* feature could, in turn, provide a way to help people discriminate between confirmatory and exploratory queries—the latter are distinguished by charts that lack concrete expectations, or having expectations that are formed after the data is seen. It is important to note, however, that our study does not distinguish between expectations that reflected substantive beliefs and those that might represent a participant's 'best guess'. Several participants commented during the study that they were unsure about their predictions. Thus, in addition to capturing beliefs, designers may also prompt users to specify the confidence in their predictions. This information can help differentiate true hypotheses from guessing. The former could be labeled as confirmatory with potential to generate robust conclusions from a visualization, whereas the latter may be flagged as exploratory, requiring further confirmation by independent sources.

5.2 Hypothesizing versus HARKing

A related behavioral difference between the two conditions is the number of Pre- versus Post-Data Hypotheses. Recall that the former represent hypotheses a participant verbalizes *before* exposing the relevant data, whereas the latter reflect attempts to hypothesize after the results are known (sometimes referred to HARKing [19]). Participants in the Prediction condition exhibited higher rates of Pre-Data Hypotheses (19.7% of total reactions) than in the Standard condition (14.9%). It appears that the prediction feature may have encouraged participants to frame their hypotheses prior to inspecting data, which could indicate more willingness to adopt a normative, confirmatory stance. That said, the rate of HARKing in the two conditions was quite similar, which suggests that both groups engaged in exploratory analyses, conceiving hypotheses after encountering patterns that seem interesting. Although HARKing is often seen as problematic [27], it is a perfectly reasonable outcome of exploratory analysis. However, it is vital to distinguish between hypotheses that are posited a priori from those that are formulated to fit observed data [40]. To that end, giving people the opportunity to predict may serve to establish such distinction in visual analytics.

There is also evidence that participants in the Standard condition engaged in more exploratory behavior. For instance, the rate of Observations, which correspond to post-hoc patterns interpreted while examining the data, is approximately twice as high in the Standard condition (31.1%) as in Prediction (16.4%). Similarly, there were more brushing-and-linking interactions in Standard (33.6% of queries) than in Prediction (7.3%). Since brushing is often classified as an exploratory activity [37, 52], the difference may reflect a focus on EDA in Standard. We speculate that those exploratory tendencies where moderated by the PredictMe feature.

Design implication: While it is not necessary nor desirable to restrict EDA in visual analytics, an opportunity exists to design more balanced systems that place equal emphasis on exploratory and confirmatory sensemaking. Belief elicitation could encourage participants to incorporate more confirmatory activities in their visual analysis. With proper distinction, balancing these two styles of analysis may allow for more normative visual inference. Externalizing prior beliefs could in turn reduce people's tendency to overinterpret the data. In effect, sketching one's predictions may serve a similar purpose to regularization in Bayesian inference and machine learning [25, 35]; by deliberately limiting learning to 'regular' data features that are within a well-informed prior distribution, one reduces the risk of overfitting and, potentially, the incidence of false discovery.

Interactive visualization systems can also be designed to actively facilitate proper inference. For example, systems could track user hypotheses, along with their history of data exposures. This analytic provenance can then be audited (either manually or by the system) to discriminate between hypotheses that were 'preregistered' before the results are known and those that were formed after. With this information, systems can provide feedback on the reliability of discoveries made in interactive analyses.

5.3 Reflections on Prior Beliefs

Externalizing beliefs and receiving visual feedback on the accuracy of those beliefs has been found to promote reflection in communicative visualizations [20]. Our results suggest that those effects may generalize to visual analysis. Participants appeared to engage in this kind of reflection more frequently when given the opportunity to sketch their beliefs. Specifically, those in the Prediction condition declared that their beliefs did not match the data at a rate that is approximately 50% higher than in Standard. A possible explanation is that the former group, having created a concrete representation of their working knowledge, could more easily relate those beliefs to the data. By contrast, participants in the Standard condition were convinced that their beliefs were accurate 33% of the time, compared to only 23.5% in Prediction. Participants who predicted the data also expressed Surprise at roughly twice the rate. On the other hand, we found minimal differences in the rate of Belief Update between the two conditions. Such statements would reflect active attempts by participants to reformulate their knowledge or amend their beliefs in light of new or contradictory data.

Overall, belief elicitation may lead to more active processing of visualizations—an effect that appears to hold for communicative [15, 20] as well as analytical visualizations, as per this study. However, we saw no evidence that visualizing belief-data gaps would translate to outright conceptual change, as we had speculated based on early research in cognitive science [7].

5.4 Breadth of Analysis

While there appears to be cognitive benefits to externalizing one's beliefs in analytical visualizations, there are also potential side effects to be considered. Among those is a reduction in the number of unique queries; participants in the Prediction condition addressed

22.6 queries on average, whereas those in Standard managed 33.1. Those who externalized their priors also created fewer visualizations on average, with 1.1 charts per query in Prediction compared to 1.4 in Standard. There were also fewer brushing-and-linking events in Prediction. These interactions are often considered essential to exploratory visualization [32, 52], with heightened exploration typically encouraged as a desirable benchmark [24]. It seems, however, that prior elicitation may have a dampening effects on these behavioral and interaction markers. This effect could be attributable to the extra effort in drawing one's expectations in the Prediction condition or, alternatively, may reflect a deeper change to one's analysis behavior. For instance, a recent study suggests that being driven by a hypothesis may inadvertently reduce one's propensity to detect unexpected patterns in data [51].

On the other hand, belief externalization appears to encourage more thoughtful interaction with a visualization. For example, participants in the Prediction condition spent approximately equal amounts of time predicting (39.5 seconds on average) and looking at the data (46.3 seconds). Qualitatively, we observed participants carefully inspecting charts in the Prediction condition, paying close attention to outliers that deviate from their expectations. However, the increased focus on individual visualizations may impede wider exploration. This in turn could prevent people from noticing unexpected relationships or features. We find evidence of this phenomenon in the rate of Observations, which was approximately half as much in the Prediction condition as in Standard. Participants whose beliefs were elicited seemed more concerned with how their priors related to the data, than in discovering new patterns they had not thought about.

Design implication: A challenge for data analysts is to maintain a degree of skepticism while being open to seeing new patterns. An exploratory stance can help surface unexpected insights, but, at its extreme, may cause one to see spurious structures in random noise. A confirmatory approach, on the other hand, aids analysts in asking relevant questions and testing plausible hypotheses, but an emphasis on prior knowledge could also lead to confirmation bias. Designers of visual analytics tools have traditionally adopted a laissez-faire approach, providing analysts with maximum flexibility, but leaving them free to adopt their own strategies. We suggest that designers should think about how to actively foster a balanced analytic experience with their interaction design. A potential research avenue is to create models that can infer analyst intents, and accordingly provide feedback on their performance. Prior work, for instance, has proposed techniques for detecting certain cognitive biases in real-time [49]. Similarly, it may be possible to utilize analyst prior beliefs to classify their behaviors on an exploratory-confirmatory spectrum. With such classification, it may be possible to provide tailored feedback. For examples, systems can nudge users to explore outside the purview of their existing knowledge, if they seem to be following a purely confirmatory approach, and vice versa when they appear to adopt an overly aggressive exploratory strategy.

6 LIMITATIONS AND FUTURE WORK

Our study provides a first look onto the effects of belief elicitation in open-ended visual analysis. However, there are several limitations that should be taken into considerations when interpreting our findings. First, this study is exploratory in nature; we specifically utilized a grounded-theory approach to observe participants and quantify their emerging analytic activities. Our findings are thus primarily data-driven and, therefore, should be considered preliminary. The generalizability of these insights should be validated in future confirmatory studies. Second, although our findings suggest differences in analytic behaviors between the two experimental conditions, the effects on the discovery process are still unclear. In particular, we did not seek to evaluate the correctness of insights reported by participants. We speculate that belief elicitation, combined with appropriate feedbacks, can decrease the incidence of false discovery in visual analytics. However, this and other hypothesized effects with respect to inference should be evaluated in future studies. Third, our subjects were limited by features available in PredictMe. For instance, the prototype did not allow participants to predict conditionally (e.g., by predicting for a subset of the data). Relatedly, our prototype did not enable participants to express their priors in the form of probability distributions as is typical in normative Bayesian inference. These limitations may have affected the way participants externalized their beliefs or their willingness to use this feature. Future work is needed to improve our design and test the effects of such improvements.

7 CONCLUSION

Interactive visualization tools are almost exclusively designed for exploratory data analysis. This narrow focus on data-driven sensemaking has led to little support for hypothesis-driven (i.e., confirmatory) analyses. We introduced PredictMe, a fully functional visualization tool that incorporates belief elicitation, in addition to supporting a range of traditional visualization features. We sought to understand how users behave in this kind of visual analytic environment. In an exploratory study, we compared this design to a Standard condition that mimics how existing visualizations work. Our results show noticeable differences in user behavior between the two conditions. Analysis of participants cognitive and interaction patterns suggest that users adopt a distinct analytic style, when given the opportunity to externalize and test the accuracy of their beliefs. This shift is marked by an increased confirmatory behavior and decreased exploration. Our findings indicate benefits but also suggest side effects and challenges to incorporating belief elicitation in general-purpose visual analytic tools. We discussed the implications for visualization design, and proposed future research directions.

ACKNOWLEDGEMENT

We thank our study participants. We also acknowledge the anonymous reviewer for their helpful feedback on an earlier version of this manuscript. This paper is based upon research supported by the National Science Foundation under awards 1942429 and 1755611.

REFERENCES

[1] Gregor Aisch, Amanda Cox, and Kevin Quealy. 2015. You draw it: How family income predicts children's college chances. *The New York Times* (2015).

[2] In Kwon Choi, Taylor Childers, Nirmal Kumar Raveendranath, Swati Mishra, Kyle Harris, and Khairi Reda. 2019. Concept-driven visual analytics: an exploratory study of model-and hypothesis-based reasoning with visualizations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 1–14.

[3] In Kwon Choi, Nirmal Kumar Raveendranath, Jared Westerfield, and Khairi Reda. 2019. Visual (dis)Confirmation: Validating Models and Hypotheses with Visualizations. In *2019 23rd International Conference in Information Visualization–Part II*. IEEE, 116–121.

[4] Kristin A Cook and James J Thomas. 2005. *Illuminating the path: The research and development agenda for visual analytics*. Technical Report. Pacific Northwest National Lab, Richland, WA.

[5] Andrew Crotty, Alex Galakatos, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. 2015. Vizdom: interactive analytics through pen and touch. *Proceedings of the VLDB Endowment* 8, 12 (2015), 2024–2027.

[6] Ulrich Dirnagl. 2020. Resolving the Tension Between Exploration and Confirmation in Preclinical Biomedical Research. *Good Research Practice in Non-Clinical Pharmacology and Biomedicine* (2020), 71–79. https://doi.org/10.1007/164_2019_278

[7] Kevin Dunbar. 1993. Concept discovery in a scientific domain. *Cognitive Science* 17, 3 (1993), 397–434.

[8] Kevin Dunbar. 2000. How scientists think in the real world: Implications for science education. *Journal of Applied Developmental Psychology* 21, 1 (2000), 49–58.

[9] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2879–2888.

[10] Luisa T Fernholz, Stephan Morgenthaler, et al. 2000. A conversation with John W. Tukey and Elizabeth Tukey. *Statist. Sci.* 15, 1 (2000), 79–94.

[11] Robert Gaynes. 2017. The discovery of penicillin—new insights after more than 75 years of clinical use. *Emerging Infectious Diseases* 23, 5 (2017), 849.

[12] Andrew Gelman. 2003. A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review* 71, 2 (2003), 369–382.

[13] Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* (2013).

[14] Garrett Grolemund and Hadley Wickham. 2014. A cognitive interpretation of data analysis. *International Statistical Review* 82, 2 (2014), 184–204.

[15] Jeremy Heyer, Nirmal Kumar Raveendranath, and Khairi Reda. 2020. Pushing the (Visual) Narrative: the Effects of Prior Knowledge Elicitation in Provocative Topics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[16] Jessica Hullman and Andrew Gelman. 2020. Interactive Analysis Needs Theories of Inference. (2020).

[17] Waqas Javed and Niklas Elmqvist. 2013. ExPlates: spatializing interactive analysis to scaffold visual exploration. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 441–450.

[18] Jaemin Jo, Sehi L'Yi, Bongshin Lee, and Jinwook Seo. 2019. ProReveal: Progressive Visual Analytics with Safeguards. *IEEE Transactions on Visualization and Computer Graphics* (2019).

[19] Norbert L Kerr. 1998. HARKing: Hypothesizing after the results are known. *Personality and social psychology review* 2, 3 (1998), 196–217.

[20] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2017. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1375–1386.

[21] Yea-Seul Kim, Logan A Walls, Peter Krafft, and Jessica Hullman. 2019. A bayesian cognition approach to improve data visualization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[22] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems* 21, 5 (2006), 88–92.

[23] Christian Lebiere, Peter Pirolli, Robert Thomson, Jaehyon Paik, Matthew Rutledge-Taylor, James Staszewski, and John R Anderson. 2013. A functional model of sensemaking in a neurocognitive architecture. *Computational intelligence and neuroscience* 2013 (2013).

[24] Zhicheng Liu and Jeffrey Heer. 2014. The effects of interactive latency on exploratory visual analysis. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2122–2131.

[25] Richard McElreath. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

[26] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[27] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. 2017. A manifesto for reproducible science. *Nature human behaviour* 1, 1 (2017), 1–9.

[28] Phong H Nguyen, Kai Xu, Rick Walker, and BL William Wong. 2014. Schemaline: Timeline visualization for sensemaking. In *2014 18th International Conference on Information Visualisation*. IEEE, 225–233.

[29] Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. 2018. The preregistration revolution. *Proceedings of the National Academy of Sciences* 115, 11 (2018), 2600–2606.

[30] Anthony O'Hagan, Caitlin E Buck, Alireza Daneshkhah, J Richard Eiser, Paul H Garthwaite, David J Jenkinson, Jeremy E Oakley, and Tim Rakow. 2006. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons.

[31] Larry Buchanan Haeyoun Park and Adam Pearce. 2017. You Draw It: What Got Better or Worse During Obama's Presidency. https://nyti.ms/2jS9b4b.

[32] William A Pike, John Stasko, Remco Chang, and Theresa A O'connell. 2009. The science of interaction. *Information visualization* 8, 4 (2009), 263–274.

[33] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[34] Catherine Plaisant, Jean-Daniel Fekete, and Georges Grinstein. 2007. Promoting insight-based evaluation of visualizations: From contest to benchmark repository. *IEEE Transactions on Visualization and Computer Graphics* 14, 1 (2007), 120–134.

[35] Xiaoying Pu and Matthew Kay. 2018. The garden of forking paths in visualization: A design space for reliable exploratory visual analytics. (2018).

[36] Khairi Reda, Andrew E Johnson, Jason Leigh, and Michael E Papka. 2014. Evaluating user behavior and strategy during visual exploration. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*. ACM, 41–45. https://doi.org/10.1145/2669557.2669575

[37] Khairi Reda, Andrew E Johnson, Michael E Papka, and Jason Leigh. 2015. Effects of display size and resolution on user behavior and insight acquisition in visual exploration. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2759–2768.

[38] Khairi Reda, Andrew E Johnson, Michael E Papka, and Jason Leigh. 2016. Modeling and evaluating user behavior in exploratory visual analysis. *Information Visualization* 15, 4 (2016), 325–339. https://doi.org/10.1177%2F1473871616638546

[39] Jonathan C Roberts. 2007. State of the art: Coordinated & multiple views in exploratory visualization. In *Fifth International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2007)*. IEEE, 61–71.

[40] Mark Rubin. 2017. When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology* 21, 4 (2017), 308–320.

[41] Purvi Saraiya, Chris North, and Karen Duca. 2005. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics* 11, 4 (2005), 443–456.

[42] Abhraneel Sarma and Matthew Kay. 2020. Prior Setting In Practice: Strategies and rationales used in choosing prior distributions for Bayesian analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[43] Yedendra Babu Shrinivasan and Jarke J van Wijk. 2008. Supporting the analytical reasoning process in information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1237–1246.

[44] John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization* 7, 2 (2008), 118–132.

[45] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology. *Handbook of qualitative research* 17, 1 (1994), 273–285.

[46] SB Trickett and J Gregory Trafton. 2009. A primer on verbal protocol analysis. *The PSI handbook of virtual environments for training and education* 1 (2009), 332–346.

[47] John W Tukey. 1977. *Exploratory data analysis*. Vol. 2. Reading, MA.

[48] John W Tukey. 1980. We need both exploratory and confirmatory. *The American Statistician* 34, 1 (1980), 23–25.

[49] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 104–115.

[50] John Wenskovitch, Michelle Dowling, and Chris North. 2020. With respect to what? simultaneous interaction with dimension reduction and clustering projections. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 177–188.

[51] Itai Yanai and Martin Lercher. 2020. A hypothesis is a liability. *Genome Biology* 21, 231 (2020). https://doi.org/10.1186/s13059-020-02133-w

[52] Ji Soo Yi, Youn ah Kang, John Stasko, and Julie A Jacko. 2007. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1224–1231.

[53] Emanuel Zgraggen, Zheguang Zhao, Robert Zeleznik, and Tim Kraska. 2018. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[54] Zheguang Zhao, Lorenzo De Stefani, Emanuel Zgraggen, Carsten Binnig, Eli Upfal, and Tim Kraska. 2017. Controlling false discoveries during interactive data exploration. In *Proceedings of the 2017 ACM International Conference on Management of Data*. 527–540. https://doi.org/10.1145/3035918.3064019