A Bayesian Cognition Approach
to Improve Data Visualization

Yea-Seul Kim
University of Washington
Seattle, Washington
yeaseul1@uw.edu

Logan A Walls
University of Washington
Seattle, Washington
loganwls@uw.edu

Peter Krafft
University of Washington
Seattle, Washington
pmkrafft@uw.edu

Jessica Hullman
Northwestern University
Evanston, Illinois
jhullman@northwestern.edu

ABSTRACT

People naturally bring their prior beliefs to bear on how they interpret the new information, yet few formal models exist for accounting for the influence of users' prior beliefs in interactions with data presentations like visualizations. We demonstrate a Bayesian cognitive model for understanding how people interpret visualizations in light of prior beliefs and show how this model provides a guide for improving visualization evaluation. In a first study, we show how applying a Bayesian cognition model to a simple visualization scenario indicates that people's judgments are consistent with a hypothesis that they are doing approximate Bayesian inference. In a second study, we evaluate how sensitive our observations of Bayesian behavior are to different techniques for eliciting people subjective distributions, and to different datasets. We find that people don't behave consistently with Bayesian predictions for large sample size datasets, and this difference cannot be explained by elicitation technique. In a final study, we show how normative Bayesian inference can be used as an evaluation framework for visualizations, including of uncertainty.

CCS CONCEPTS

• **Human-centered computing** → **Visualization**.

KEYWORDS

Visualization, Bayesian Cognition, Uncertainty Elicitation

1 INTRODUCTION

Data-driven presentations are used by the media, government, and private sector to inform and influence public opinion. For example, a journalist might present polling data prior to a midterm election in a choropleth map to convey to readers the probability of a Democratic majority in different areas. While visualization designers may acknowledge that users' expectations and prior knowledge (e.g., about the political sentiment within their district, or their own preferences for a given candidate) will influence what they conclude from the visualization, visualization design guidance and evaluation methods rarely acknowledge these factors. Most conventional visualization design guidance implies that finding an effective design means choosing the right combination of visual encodings and comparisons given the target task. Accordingly, evaluations often frame an ideal user as one who minimizes perceptual and other cognitive errors in extracting the information embedded in the visualization.

Opposing a "data-only" view of visualization, models of graphical comprehension from psychologists have described how top-down influences, including prior beliefs and expertise, influence what a person attends to [6]. Studies demonstrate how prior knowledge can lead to other "downstream" effects on visualization related outcomes, such as how effective an interactive visualization is for different users [20]. Recent visualization experiments demonstrate how eliciting people's beliefs about data directly through the interface [29, 30] can positively impact data recall and prompt critical thinking about data. While this work has examined how people update their beliefs after providing a prediction and seeing others' predictions about data, research in data visualization has yet to develop descriptive or normative cognitive models for predicting and evaluating how people update the prior beliefs they bring upon viewing data.

Outside of visualization research, psychologists have developed these types of models of how people update their beliefs or opinions about data or a proposition, given information about their prior beliefs [16, 48]. Bayesian models of cognition compare human cognition, which is assumed to draw on prior knowledge, to a normative standard for rational induction from noisy evidence [18]. By combining key components of Bayesian statistics—including a likelihood function describing the probability of the data given some assumed distribution, a description of the prior probability of different values, and laws of conditional probability—Bayesian cognitive modeling can *describe* how people update their beliefs given data. Bayesian models have provided explanatory accounts of how people make various real-world perceptual judgments, higher cognitive inferences, and learn and reason inductively [16, 33, 47, 48]. Bayesian cognitive models can also *prescribe* what updated beliefs are most consistent with one's prior beliefs and the data, providing a normative framework for evaluating interactions with data presentations.

We make several contributions. First, we **demonstrate a Bayesian cognitive model for assessing how people interpret data presentations** like simple visualizations. In contrast to other frameworks for studying visualization use, a Bayesian cognitive model can be used to examine how people *change their beliefs* in response to presented data.

Deploying the model we develop, we **characterize the extent to which the belief updating of users of a simple visualization of survey results resembles Bayesian inference** (Study 1). We find evidence that on average people update their beliefs rationally, but individuals often deviate from expectations of rational belief updating. These findings hold across multiple datasets and prior elicitation methods (Study 2). We find that people deviate considerably more from the predictions of Bayesian inference even in aggregate when presented with datasets of a very large sample size.

Finally, we **demonstrate how a Bayesian cognitive model can be used to evaluate data presentations** (Study 3). We show how Hypothetical Outcome Plots (HOPs), animated plots that show uncertainty via draws from a distribution, improve deviation from normative Bayesian responses relative to not presenting error information.

2 BACKGROUND

Interpreting Data Presentations

Cognitive psychologists proposed early models of visualization interpretation implying that "top down" factors relating to a user's information needs, prior knowledge, and graph literacy affect how visualized data is interpreted, for example, by guiding attention [31, 44]. Studies in graph comprehension provide evidence of such top-down effects [5, 41, 43, 56].

For example, static visualization of processes, which require use of internal representations to interpret, often outperform animations [20, 36]. Other studies show that externalizing one's internal representations leads to better understanding of visualized information [7, 8, 21, 39, 45].

The Impact of Prior Knowledge & Beliefs. In reflecting on the "value" of visualization, Van Wijk notes that the knowledge gained from a visualization will depend on the prior knowledge that a user brings [51]. Recent research demonstrates that while visualizations are slightly more likely to persuade people to change their attitudes about a data driven topic (e.g., to be more likely to believe that some factor X causes some symptom Y), the polarity of the person's original attitude influences the strength of the visualization's effect [42]. Going a step further toward understanding the role of prior beliefs, Kim et al. show that asking visualization users to "draw" their predictions in an interactive visualization prior to seeing the observed data can help them remember data 30% better [29, 30], perhaps by increasing their ability to compare the observed data to their expectations. Deviation between a predicted trend or value and the observed trend or value has been shown to be predictive of people's updated beliefs and ability to recall data [30, 38]. However, these works focused on eliciting a user's single best prediction of a trend, rather than a distribution over possible values, which is required for making use of the normative predictors that are possible from a Bayesian approach.

Bayesian Cognition

In cognitive science, Bayesian statistics has proven to be a powerful tool for modeling human cognition [16, 48]. In a Bayesian framework, individual cognition is modeled as Bayesian inference: an individual is said to have implicit beliefs about the world ("priors"); when the individual observes new data, their prior is "updated" to produce a new set of beliefs which account for the observed data (this new set of beliefs is referred to as the "posterior"). The prior is formalized as a probability distribution and Bayes' rule is used to obtain the posterior from the prior distribution and the likelihood function that the observed data is derived from.

This approach has been used to model many aspects of human cognition at various levels of complexity, such as object perception [28], causal reasoning [46], and knowledge generalization [47].

Particularly relevant to our work is a study conducted by Griffiths and Tenenbaum [18], which compared people's predictions for a number of everyday quantities to the predictions made by a model that used the empirical distribution as a prior (e.g., for human lifespans they used a model with a prior calculated from historical human lifespan data). The

study found that although there was variance between individuals, in aggregate people's judgments closely resembled the normative Bayesian posterior. We are similarly interested in how judgments that people make in everyday interactions with data presentations (like visualizations) compare to the expectations of normative Bayesian inference.

Approximate Inference & Sampling Behavior. While Bayesian models of cognition have seen wide applications, the idea that human cognition is accurately described as Bayesian inference is inconsistent with previous influential findings in cognitive psychology from authors such as Tversky and Kahneman [50]. Tversky and Kahneman found evidence that humans often use simple heuristics in their decisions, and that these heuristics lead to sub-optimal judgments. More recent research indicates that heuristics are adaptive and often lead to accurate judgments (e.g., [13]). A recently proposed explanation which reconciles the seemingly opposing findings between Bayesian models of cognition and the idea that heuristics lead to non-optimal judgments is motivated by Bayesian cognition [17]: what if human cognition is not *exact* Bayesian inference, but instead is *approximate* Bayesian inference? One such approach proposes that while people have a prior probability distribution which encodes their beliefs, they do not form judgments using the entire distribution at once [52]. Instead, they take a small number of samples from the distribution, and reason with these samples instead of the full distribution (we which refer to as *sample-based Bayesian*). Being a sample-based Bayesian can lead to sub-optimal individual inferences, but in aggregate, it can produce results very similar to exact Bayesian inference.

Application to Data Visualization. Recent work by Wu et al. [55] explored the application of the Bayesian framework to examine how people update their beliefs when viewing visualized data. However, Wu et al. prompted participants to internalize a provided prior, show them the observed data, and then ask for their posterior beliefs. Using a fixed prior is not ideal in cases where participants' pre-existing beliefs about a phenomenon will impact their ultimate beliefs. Our work demonstrates how to elicit and model participants' personalized priors for a more realistic application of Bayesian inference, including proposing and evaluating multiple elicitation techniques.

3 DEVELOPING RESEARCH QUESTIONS & GOALS

Prior beliefs clearly play a role in data interpretation. However, belief updating is rarely formally modeled in research related to data presentation and visualization. Studies of Bayesian cognition suggest that Bayesian inference can be used to characterize many aspects of learning and cognition. We apply a Bayesian cognitive modeling approach to a simple data interpretation task to understand where people align
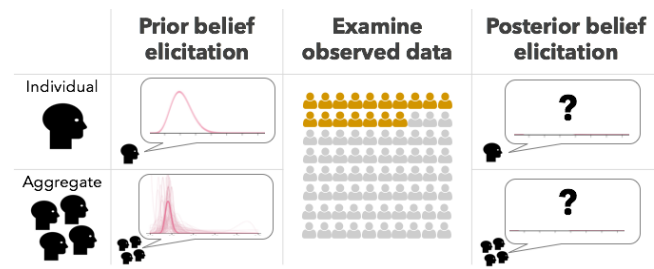

Figure 1: Bayesian inference at individual & aggregate level.

with, and deviate from, normative Bayesian inference individually and in aggregate. While the computational complexity of Bayesian inference makes it doubtful that cognition uses exact inference [25], in the context of interpreting presented data in everyday settings (such as in data journalism) we would expect under Bayesian assumptions to see that people (1) are capable of providing *priors* describing the uncertainty in their beliefs about a parameter, and (2) update these beliefs to incorporate observed data. Our work sheds light on the degree to which these assumptions hold for a simple data interpretation task.

In contrast to prior work in Bayeisan cognition that avoids obtaining priors directly from people [18, 55], we design and apply a paradigm in which we *elicit people's prior and posterior beliefs about the probability that a parameter takes various values* (Studies 1, 2, 3). Though it is commonly argued that people have difficulties reasoning about probability, the notion that people are capable of maintaining subjective probabilities is well-established in decision theory, congruent with canonical work in judgment and decision making like that of Tversky and Kahneman [50], and supported by a body of work in economics on subjective probability elicitation, including from laypeople (see [34] for a review. Having obtained prior beliefs, we fit a distribution to them then use Bayes' rule to compute the *normative posterior distribution* for each person, the posterior distribution that is expected if the person is a perfect Bayesian agent given the observed data and their prior distribution.

In a first study, we compare the distribution fit to the posterior beliefs we elicited from each person to the normative posterior beliefs computed using that person's prior Bayesian solutions (Fig. 1 top row). We also compare people's *aggregate posterior distribution* (i.e. the posterior distributions representing the aggregate of all people's posterior distributions) to the *normative aggregate posterior distribution* (i.e., the normative posterior distribution calculated using a prior distribution representing the aggregate of all people's prior distributions) (Fig. 1 bottom row). Alignment between people's responses and the normative Bayesian solution at this

aggregate level may suggest that people are "sample-based Bayesians" [52] performing approximately inference.

Prior work in visualization and judgment and decision making suggests that different subjective probability elicitation techniques can produce varying results, perhaps because some techniques (such as frequency framings) better align with people's internal representations of uncertainty [15, 22, 40]. In a second study, we assess *how sensitive people's responses are to different elicitation methods*, which vary in the input format for beliefs they use (i.e., continuous probability versus discrete samples).

In a third study, we show how a Bayesian cognitive model can be used to assess the effectiveness of design changes. One aspect of visualization design that is likely to be relevant to how people update beliefs is the presentation of uncertainty. If people see the observed data as more certain than it is (e.g., reflecting belief in the law of small numbers [49]), their posterior judgments may reflect overweighting of the observed data and underweighting of their prior. On the other hand, if people see the data as less certain that it is (e.g., non-belief in the law of large numbers [3]), their posterior judgments may reflect underweighting of the observed data and overweighting of their prior. To demonstrate how a Bayesian cognitive model can support visualization design decisions, we *compare the results of Bayesian modeling across a default static visualization typical of those found in the media and an animated hypothetical outcome plot (HOP [23]) uncertainty visualization.*

4 S1: BAYESIAN DATA INTERPRETATION

We demonstrate how a Bayesian model cognition of cognition can be used as a framework for assessing visualization interpretation. We evaluate the extent to which individuals' judgments are consistent with people being "fully" Bayesian by assessing how closely their individual posterior distributions align with the normative posterior distribution calculated given their prior. Secondly, we also consider whether people's judgments might instead be consistent with people being "sample-based" Bayesians (one form of being approximately Bayesian) by evaluating how closely the aggregate posterior distribution aligns with the normative aggregate posterior distribution.

Study Design

We designed a between-subjects experiment with 50 participants. We determined sample size via a prospective power analysis conducted on pilot data with a desired power of at least 0.8 (α=0.05) to detect a difference between the normative distribution and aggregated posterior distribution. We recruited participants from Amazon Mechanical Turk (AMT), rewarding their participation with $1.0. The average time to complete the task was 7.3 minutes (SD=5.2).

Dataset & Presentation. For the purposes of understanding how Bayesian cognitive modeling might provide insight into visualization interpretation, we sought a realistic yet relatively simple dataset similar those shown in the media or public facing reports. We selected a dataset with a single variable which represents a proportion. The dataset describes survey results intended to measure attitudes towards mental health in the tech workplace (N=747) [1]. We chose one question from the survey "how often do you feel that mental health affects your work?" to formulate our proportion parameter: "the proportion of women in the tech industry who feel that mental health affects their work often." To present the observed proportion to participants in our study, we created an "info-graphic" style visualization (Fig. 5 (a)) which shows this proportion using a grid format commonly used in the media to present proportions (e.g., [2, 24, 35]).

Prior & Posterior Elicitation. To elicit participants' prior and posterior distributions, we used a technique that asks participants about two properties of their internal distribution: the most probable value of the parameter (*mode (m)*) and their *subjective probability* (Fig. 4(b)) that the parameter falls into the interval around the mode ($[m - 0.25m, m + 0.25m]$). Prior research in probability elicitation for proportions indicates that this technique is less sensitive to noise which arises from externalizing subjective uncertainty compared to other techniques [54]. A second benefit of this approach is that estimates of Beta distribution parameters can be analytically computed from participants' answers [10].

Results

Fitting Individual Responses. We first converted participants' elicited responses of prior and posterior beliefs to Beta distributions using an optimization approach suggested in previous work [40]. The approach finds an optimal Beta distribution parameterized by α and β which minimizes the sum of two terms: (1) the square difference between the participants' mode and the estimated mode of the Beta distribution and (2) the square difference between the probability that each participant associated with the interval and the estimated probability of the interval in the distribution being optimized.

Fitting Aggregate Responses. To obtain parameters for the aggregated prior/posterior distributions (α_{agg} and β_{agg}), we averaged participants' αs and βs respectively from the individual prior/posterior distributions: $\alpha_{\text{agg}} = (\alpha_1 + ... + \alpha_N)/N$, $\beta_{\text{agg}} = (\beta_1 + ... + \beta_N)/N$ (where N = # of participants).

Calculating Normative Posteriors. We can calculate a participant's normative posterior by using α and β estimates from

their prior distribution combined with the number of successes (e.g., the number of women who said their mental health affects their work often) and failures (e.g., the number of women who said their mental health affects their work not often) in the observed data (Eq. 1). The α and β for the aggregated normative posterior are calculated in the same manner using the aggregated prior α and β estimates.

$$\alpha_{\text{normative posterior}} = \#successes + \alpha_{\text{prior}}$$
$$\beta_{\text{normative posterior}} = \#failures + \beta_{\text{prior}} \tag{1}$$
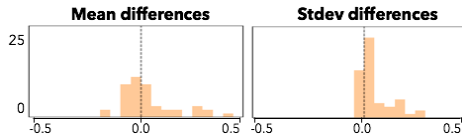

Figure 2: Distributions of residuals (observed - predicted) for participants' posteriors' means and standard deviations and the means and standard deviations of the normative posteriors.

We evaluate the degree to which individual and aggregate posterior distributions resemble the normative Bayesian posterior distributions by plotting residuals (*observed - predicted*) when predicting the means and standard deviations of participants' posterior distributions using normative Bayesian inference (Fig. 2). A distribution of residuals that is loosely centered around zero suggests "noisy" Bayesian inference, where each individual may deviate from the normative posterior due to approximate inference but in aggregate, the observed posterior resembles the normative posterior. Residuals for means are roughly centered around zero, with 95% of the values falling between -0.16 and 0.58). A small number of participants provided means that were much greater than predicted (i.e., believed that the true proportion of women in tech who feel that mental health affects their work often was much larger than predicted from the prior and the observed data). Residuals for standard deviation are also roughly distributed around zero, but show a stronger bias toward larger standard deviation in one's posterior judgments. This suggests a tendency among participants to provide posterior beliefs indicating more uncertainty than is rational given the observed data according to Bayesian inference.

Following this observation, we analyzed where each participant's posterior distribution was located relative to the normative posterior distribution (Fig. 3). We found that 44% of participants (22 out of 50) overweighted the mode of the observed data (i.e., their posterior distributions are closer to the observed data than they should be), while 34% of participants (17 out of 50) overweighted the mode of their prior distribution, and 18% of the participants (9 out of 50) provided posterior beliefs that moved further than the prior

from the observed data. Only two participants (4%) were within ±1% range of the mode of their normative posteriors.
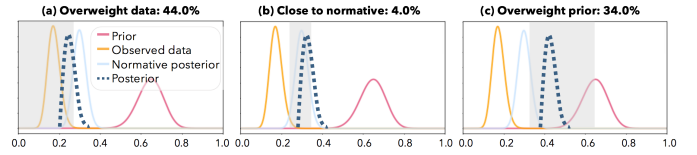

Figure 3: Proportions of participants whose posterior distributions (dotted line) imply overweighting of the mode of the observed data, reasonable alignment with the normative posterior, and overweighting of the mode of the prior distributions. An additional 18% of participants (not shown) provided posterior beliefs that were further than the prior from the observed data.

Per our pre-registration we also report log KL divergence (KLD) [32] between normative and observed posteriors. KLD is an information theoretic measure of the difference between two probability distributions. Examining log KLD at the individual and aggregate levels aligned with our observation from the residual plots: few individuals act "fully Bayesian", but in aggregate the responses are close to normative predictions. The mean log KLD for a participant at the individual level was 0.52 (SD=1.18; 3.31 in non-log terms). Normative behavior is represented by a smaller log KLD and non-log KLD close to 0. The aggregate log KLD was -2.18 (non-log KLD=0.11), which aligns with previous work that demonstrates people's collective reasoning is more consistent with Bayesian optimal behaviors even when individuals do not necessarily act as a fully Bayesian agent [18].

5 S2: ELICITATION TECHNIQUES & DATASET

Study 1 (S1) used an elicitation technique from the literature which was designed for fitting Beta distributions to participants' responses using a numerical solution [10]. While the technique has been shown to be less sensitive to noise from the elicitation process than other techniques [54], it is possible that the evidence for approximate or "sample-based" Bayesian inference that we observed was an artifact of the elicitation technique. For instance, by asking for a mode value, it is possible that the technique prompted people to consider only a single sample. We are interested in evaluating how robust our result in Study 1 is to changes in the dataset that is presented. In a second pre-registered study,[1] we therefore evaluate four different elicitation techniques (including interval elicitation from Study 1) and introduce a new dataset. The elicitation techniques vary in the degree to which they ask a participant to provide a full distribution

---

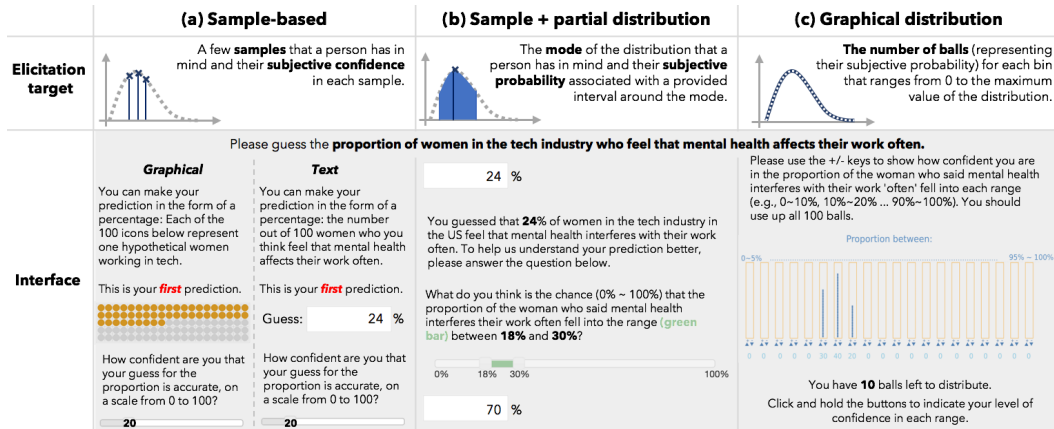[1]http://aspredicted.org/blind.php?x=4bf9ci

Figure 4: Elicitation target and interface associated with each elicitation technique. We developed two sample-based techniques (a), and used an interval technique [54] (b) and a graphical "balls and bins" technique [15] (c) from the literature.

versus a small set of samples. By manipulating both representation of uncertainty and the dataset, we aim to gain a better sense of how robust our observation of approximate Bayesian inference is.

Developing Elicitation Techniques & Conditions

We are interested in comparing a set of interfaces which vary in the format they use to elicit participants' responses. We describe two sample-based techniques of our own design, as well as two elicitation techniques from the literature. While our data interpretation task requires eliciting a Beta distribution specifically, we expect that the techniques we evaluate should generalize to other symmetric distributions.

Sample-based Elicitation. Evidence from research on reasoning with uncertainty (e.g., on classical Bayesian reasoning tasks [12]) and uncertainty visualization [9, 22, 23, 26, 27] indicates that people are often better at thinking about uncertainty when it is framed as frequency rather than probability. One way to elicit uncertainty is through a technique that asks people to provide one sample at a time until they have exhausted their mental representation. Imagine a person who is asked to provide their expectations for the proportion of women in tech who experience mental issues often. Several possible proportions seem salient to them, including 20% and 33%. We devise a sample-based elicitation method that asks a person to articulate a small set of samples (e.g., 5), one at a time (Fig. 4(a)).

Even if people find it easy to reason in the form of samples, we might still expect that they perceive some samples as more likely. A sample-based elicitation technique would not prevent a person from providing the same sample multiple times, proportional to its expected probability (i.e., resampling with replacement). However, articulating the same sample multiple times can be tedious. For each sample a

person provides, our technique asks for a corresponding judgment about the salience of the sample in the form of subjective confidence. Using this technique, the hypothetical person with two samples of 20% and 33% might provide 20% as a first estimate with a higher confidence (e.g., 70 on a scale of 0 to 100), and 33% as a second estimate with a lower confidence (e.g., 30). In practice, the confidence values do not need to sum to 100 as they can be normalized prior to using them to fit the responses to a distribution.

We created two versions of our sample-based elicitation technique. A **graphical sample-based elicitation interface** (Fig. 4 (a) left) allows participants to provide a predicted value (i.e., sample) by clicking icons in an icon array. This interface is nearly identical to the visual format used to present the observed data, except for the elicitation icon arrays, which use circles. An analogous **text sample-based elicitation interface** (Fig. 4 (a) right) allows participants to provide a predicted value by entering number in a text box.

As a participant provides their samples, each prior sample is appended to the bottom of the interface so that participants can review their samples and corresponding confidence values before submitting the response.

Graphical Distribution Elicitation. To conduct a Bayesian analysis in many domains (e.g., clinical trials, meteorology, etc.), analysts probe domain experts for uncertainty estimates, then use these to construct a prior distribution [40]. This approach generally assumes that people with domain knowledge possess a relatively complete mental representation of the uncertainty in a parameter. Graphical elicitation of a full subjective distribution has been proposed for use among lay people, such as to elicit preferred retirement outcomes [14]. Recent research indicates that a graphical interface that enables constructing a distribution via placing 100

hypothetical outcomes ("balls", or circles representing hypothetical outcomes) in multiple ranges ("bins") allows people to articulate a distribution that they have been presented with more accurately than a method that asks for quantiles of the distribution [15]. We implemented a **graphical "balls and bins" elicitation interface** (Fig. 4(c)). Participants are prompted to add exactly 100 balls in bins that span between 0% to 100% in increments of 5% to express the distribution they have in mind. Relative to the text and graphical sample-based techniques we developed, the graphical balls and bins interface encourages a person to consider their entire subjective probability distribution at once.

Sample + Partial Distribution Elicitation. The interval technique we used in Study 1 may be best considered a hybrid approach between approaches that emphasize small sets of samples and those that emphasize a full distribution (Fig. 4(b)). The mode that a participant provides can be thought of as the most salient sample in their priors. The subjective probability that a participant provides is analogous to the probability mass of a partial distribution.

As in Study 1, participants are first prompted to provide a prediction (m). Participants are then asked to provide the subjective probability (sp) that the true proportion falls into the range calculated based on the mode value that they entered ($[m - m * 0.25, m + m * 0.25]$).[2]
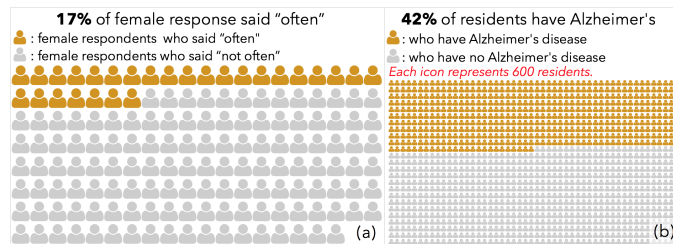

Figure 5: The data presentations for S1 (a) and S2 (a, b).

Study Design

Dataset and Presentation: We reuse the same proportion dataset used in Study 1 (mental health outcomes among women in the tech industry) and the same icon array visualization. However, we are also interested in understanding how robust our findings are to changes in the nature of the observed data. Specifically, the sample size of the observed data directly influences how closely the normative posterior is expected to align with the data. Intuitively, as the sample size of the observed data increases, the impact of the prior distribution on the normative posterior is reduced. With a very large sample, the normative posterior will be virtually indistinguishable from the data even with a reasonably concentrated prior distribution (Fig. 6).

---

[2]We elicited two additional random ranges to see how the response is impacted by the ranges. The analysis is in the supplemental material.
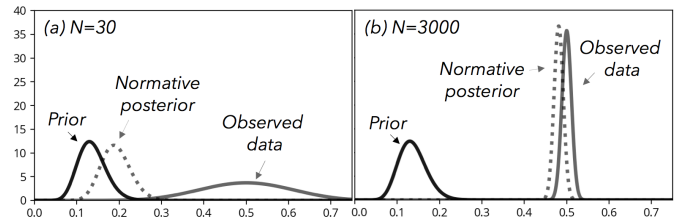

Figure 6: Illustration of the effect of sample size on normative posteriors given the same prior and observed mode.

We therefore chose one additional large sample dataset that has been visualized in the New York Times using icon-style visualizations [4]. This dataset depicts the results of a study of chronic health conditions among assisted living center residents in the U.S. (N=750,000). We chose one type of chronic health condition (Alzheimer's disease or another form of dementia) to formulate our target proportion. We asked participants to reason about "the proportion of residents who have Alzheimer's disease or another form of dementia" in the task. We created a visualization (Fig. 5 (b)) that shows this proportion in a similar icon array format to that used for the mental health in tech dataset. Because of the size of the sample, we tell participants that each icon represents 600 residents of assisted living centers.
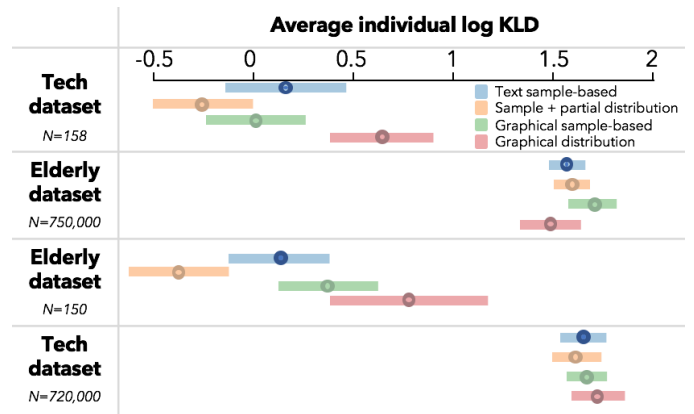

Figure 7: Bootstrapped 95% confidence intervals for average log KLDs.

Procedure. We used the same procedure as in Study 1 (eliciting priors, presenting observed data, eliciting posteriors). However, in Study 2 we randomly assigned participants to one of the four elicitation conditions, and one of the two datasets. On the last page of the experiment, we asked a pre-registered attention-check question about the numeric range in which the observed proportion fell to exclude participants who may not have paid attention to the observed data. Participants were asked to choose an answer among three ranges (0%-30%, 30%-60%, 60%-100%).

Participants. Based on a prospective power analysis conducted on pilot data with a desired power of at least 0.8 assuming α=0.05, we recruited 800 workers with an approval rating of 98% or more (400 per dataset, 200 per elicitation condition) in the U.S from Amazon Mechanical Turk (AMT). We disallowed workers who took part in Study 1. We excluded participants who did not respond correctly to our attention check question from the result according to our pre-registration. We posted the task to AMT until 800 participants who correctly answered the attention check question were recruited. Participants received $1.0 as a reward. The average time to complete the task was 4.8 minutes (SD=3.35).

Results

Data Preliminaries. For each technique, we aimed to use the simple and most direct technique to fit a Beta distribution, so as to minimize noise contributed by the fitting process. For sample-based elicitation conditions, we used the Method of Moments [19] to estimate distribution parameters (i.e., alpha and beta) using samples provided by each participant. This method provides an estimate using the mean of the samples that participants provided ($\bar{x}$) and the variance of the samples ($\bar{v}$) to calculate beta parameters: $\alpha = \bar{x}(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1), \beta = (1 - \bar{x})(\frac{\bar{x}(1-\bar{x})}{\bar{v}} - 1)$. Since we asked participants to provide their subjective confidence with each sample, we calculated weighted $\bar{x}$ by multiplying the value of each sample by the corresponding confidence value. This approach does not provide a unique solution when the participant provides the same values five times or 0 confidence for all samples. In this case, we gave the participant an uninformative uniform prior ($\alpha = 1, \beta = 1$). We provide a sensitivity analysis to different ways of interpreting these "deviant" cases in the supplemental material [3]. For the graphical distribution condition, we also used the Method of Moments approach by considering each ball as a sample known within a 5% (the bin width). For the sample & partial distribution condition, we used the same optimization approach we used in Study 1.

Residual Analysis and Log KLD. To assess the effect of elicitation technique on individual-level alignment with the normative Bayesian solution, we again plot residuals between normative (predicted) means and standard deviations for each participant and observed means and standard deviations (Fig. 8). For the tech dataset (N=158) used in Study 2, we observed a similar pattern as in Study 1, with errors roughly equally distributed about zero for means, and around zero but with a slight bias toward the degree of variance of priors (i.e., overestimating variance in the data).

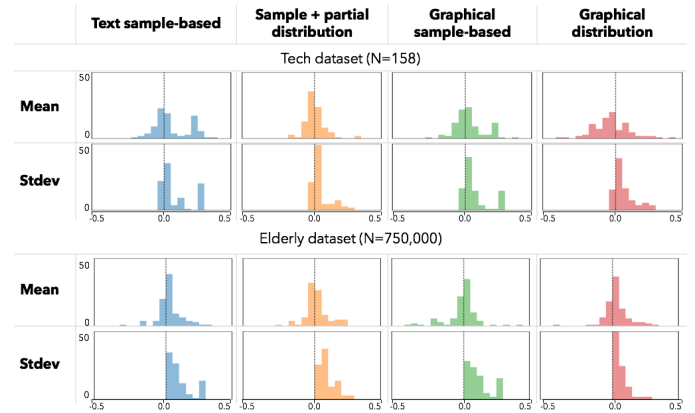For the elderly dataset (N=750,000), residuals for means are again roughly symmetric about zero, but residuals for

[3]https://github.com/yeaseulkim/Bayesian-Visualization

Figure 8: Distributions of residuals (observed-predicted) for participants' posteriors' means and standard deviations and the means and standard deviations of normative posteriors.

standard deviations are nearly entirely to the right of zero. This suggests a strong tendency for people to be more uncertainty about the true proportion than they should rationally be, given the size of the observed dataset.

We see some small differences in residual distributions between techniques. For example, those using the graphical balls and bins interface (Fig. 8 fourth column) appear to be slightly more consistent (i.e., more concentrated distribution) and slightly less likely to be biased in their estimates of standard deviation of the elderly dataset (Fig. 8 bottom row). We counted the participants whose responses spanned more than three bins, with the number of balls on either side of the center bin differing by less than two balls. 110 out of 200 participants in this condition attempted to create a symmetric distribution across more than three bins (totaling a 15% range) for their posterior distribution. Prior work on graphical elicitation has proposed that the axes ranges of an elicitation interface may implicitly influence the predictions that people "draw". In the case of the graphical distribution interface, it is possible that participants relied on a heuristic suggesting that distributions should be roughly centered and span more than one bin. The small differences in techniques, however, are far less pronounced that the more obvious differences between people's residuals for standard deviation for the (large) elderly dataset versus the (small) tech dataset.

Per our pre-registration we constructed bootstrapped 95% confidence intervals for the mean individual log KLD between participants' posteriors and the normative posteriors. Full results from this analysis are in supplemental material. We found that on average, the mean log KLDs from all conditions were larger than we would expect if people are "fully Bayesian" at an individual level, further aligning with what we see in Fig. 8. Across both datasets, we saw no consistent effects of the elicitation techniques on alignment between

people's posteriors and the normative posteriors as measured by log KLD.

To disambiguate whether the difference between the tech dataset and the elderly dataset is due to the different domains of the data or the different sample sizes, we introduced additional datasets by manipulating sample size. We reran the study with the sample sizes switched for the two datasets (tech dataset N=720,000, elderly dataset N=150). We observed the same pattern of results in residual plots (presented in supplemental material), where elicitation techniques did not appear to reliably impact individual's residuals in means or standard deviations, but the larger sample size datasets led to residual standard deviations that were strongly biased toward greatly overestimating the amount of uncertainty one should feel given their prior and the observed data. We again confirmed these results by examining log KLD (see supplemental material). We speculate that the deviation is caused by a well-documented tendency among people to show insensitivity to sample size and its relationship to variance (sampling error) [50]. Hence participants did not weight the value of information captured by the observed elderly dataset as much as they should, given its large sample size (N=750,000).

We examined the aggregate level log KLD results to confirm what the residual plots suggested regarding approximate Bayesian inference for the smaller sample datasets but not for the larger sample datasets. We found that while people's responses were consistent with an approximate or sample-based Bayesian hypothesis for the small sample size datasets, we don't see analogous evidence that people act as sample-based Bayesians for the large sample datasets (Fig. 9(a, b)).
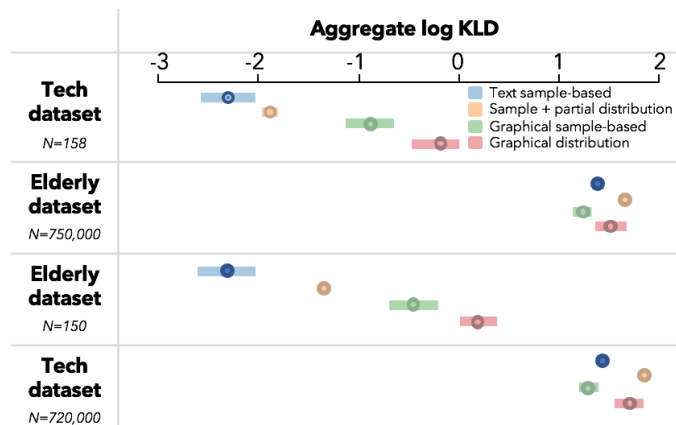
Figure 9: Bootstrapped 95% confidence intervals for aggregate KLDs.

Perceived Sample Size. To contextualize these results further, we analyzed perceived sample size. One benefit of obtaining

distributions rather than just expected values (e.g., [29, 30]) is that we can interpret the parameters of the fitted Beta distributions to gain insight into how participants perceived the data. For a Beta distribution, the two parameters α and β are associated with the sample size that the distribution represents. α stands for the number of successful trials, and β stands for the number of failed trials. By treating the participants' posteriors as normative posteriors and using the elicited priors, we reverse-calculated the perceived observed data distribution ($\alpha_{\text{perceived data}}$ and $\beta_{\text{perceived data}}$) for each participant (in other words, the counts that a Bayesian would have perceived in the data to arrive at that posterior), then summed these two parameters for sample size. Figure 10 shows how the perceived sample size of the observed data was roughly the same across elicitation techniques and datasets. The mean across all techniques for the tech dataset (N=158) was 212.47 (median=41.14) whereas the mean of elderly dataset (N=750,000) was 359.58 (median=51.51), despite the enormous actual difference in the sample sizes of the observed data (158 vs. 750,000).
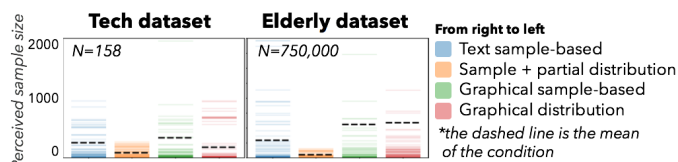
Figure 10: Perceived sample size as implied by participants' prior and posterior distributions. Participants perceived similar sample sizes between two very different sized datasets.

6 S3: UNCERTAINTY VIS & PRIOR ELICITATION

In Study 2 (S2), we observed a pervasive insensitivity to the sample size of the observed data. We turn now to how a Bayesian approach can be used to evaluate how well different visualization alternatives encourage normative interpretations. One natural way to attempt to better calibrate people to the potential for sampling error as a function of sample size is to explicitly present uncertainty information. Visualizing uncertainty may help people make better-calibrated judgments on how much they should weigh the observed data when they incorporate it with their prior beliefs to formulate their posterior beliefs. We also consider the possibility that the insensitivity to large sample sizes that we observe is exaggerated by people anchoring to their prior beliefs because they have been made more salient by elicitation. Our goals in our third pre-registered[4] study are to use Bayesian cognition to evaluate the effect of uncertainty visualization and to better understand the extent to which the act of eliciting people's priors might alter how they update their beliefs.

---

[4]http://aspredicted.org/blind.php?x=496ri9

Elicitation Technique & Dataset

To evaluate our questions, we used the tech dataset (N=158) and the elderly dataset (N=750,000) that we used in Study 2 (Fig. 5). We used the text sample-based technique from Study 2. To show uncertainty around the observed data, we used Hypothetical Outcome Plots (HOPs) [23]. HOPs convey uncertainty by animating set outcomes randomly drawn from a target distribution. To create HOPs for each dataset, we constructed a binomial distribution using parameters of the dataset (e.g., $\beta(n = 158, p = 0.17)$ for the tech dataset), then sampled multiple hypothetical modes from the distribution to present as hypothetical outcomes, using a frame rate of 400ms as suggested by prior work [23, 26] (Fig. 11).
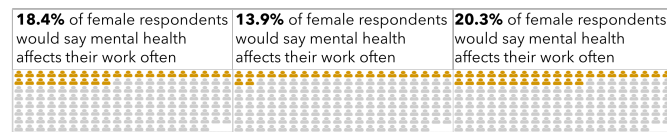

Figure 11: The example frames from the HOPs (tech dataset).

Conditions & Participants

We crossed two interventions (uncertainty visualization, prior elicitation) to arrive at four study conditions (Fig. 12). Participants in the **Elicitation-Uncertainty condition** were prompted to externalize their priors before seeing the observed data, then to examine the observed data as HOPs. Participants in the **Elicitation-No uncertainty condition** were prompted to externalize their priors before seeing the observed data, then to examine the observed data as a static icon array as in Study 2. Participants in the **No elicitation-Uncertainty condition** were asked to examine the observed data presented with HOPs but were not prompted to externalize their prior beliefs. Lastly, participants in the **No elicitation-No uncertainty condition** were asked to examine the static observed data without being prompted to externalize their prior beliefs beforehand. Participants in all conditions provided their posterior beliefs after examining the observed data.

Figure 12: Table of Study 3 conditions.

The Elicitation-No uncertainty condition responses consisted of participants' responses from the text sample-based conditions from Study 2 (responses from a total of 200 participants, 100 per dataset). For the remaining conditions, we recruited an additional 600 participants (100 per condition, a total of 300 per dataset) in the U.S from AMT. We disallowed

workers who took part in Study 1 or Study 2. We excluded participants who did not respond correctly to our attention check question. We posted the task to AMT until 300 participants who correctly answered the attention check questions were recruited. Participants received $1.0 as a reward.

Analysis Approach

Per our pre-registration we used a Bayesian linear regression implemented in R's *rethinking* package to evaluate the effect of prior elicitation and uncertainty visualization using a single measure (log KLD). We examined residual plots for mean and variance of participants' posterior distributions for all conditions (see supplemental material) to confirm our model interpretations below.

To compute the normative posterior for No-elicitation conditions, we used the aggregate priors from participants in the text sample-based condition in Study 2 (Tech dataset: $\alpha = 10.79$, $\beta = 18.99$, Elderly dataset: $\alpha = 31.25$, $\beta = 39.59$). We specified a model to regress the mean effect in individual log KLD on dummy variables indicating whether uncertainty visualization was shown, whether prior elicitation was prompted and which dataset was presented (tech vs. elderly). We specified identical weakly regularizing Gaussian priors for mean effects (μ: 0, σ: 1) and half-Cauchy priors (Cauchy distributions defined over positive real numbers) for scale parameters (μ: 0, σ: 1). The thick tailed Cauchy distribution tends to be slightly preferable to Gaussian distributions as a weakly regularizing prior for standard deviations [37]. We included the (mean-centered) time that the participant spent to examine the observed data to interpret whether time spent alone impacted the results. We present posterior mean estimates of effects with 95% confidence intervals.

Results

Mean task completion time was 3.8 minutes (SD:2.4) for No-elicitation and 4.7 (SD:3.2) for Elicitation conditions.

Impacts on Individuals' Updated Beliefs. Figure 13 shows the posterior mean estimates for effects on log KLD. Prior elicitation reduced log KLD of individuals' posterior beliefs relative to the normative Bayesian posteriors but not reliably so (mean:-0.04, 95% CI:[-0.15,0.1]). Log KLD reliably improved when participants were exposed to uncertainty visualization, with log KLDs relative to the normative posteriors for those who viewed HOPs being on average lower by -0.15 (95% CI:[-0.29,-0.04]). Being assigned to view the large sample size dataset (i.e., elderly dataset) still had a large impact on results at the individual level, with the average log KLD for those who viewed the large sample dataset being on average 1.54 log KLD units larger than those for the small sample

size dataset (95% CI:[1.42,1.67]). Spending more time examining the observed data reduced log KLD but not reliably (mean=-0.07, 95% CI:[-0.14,0.01]).

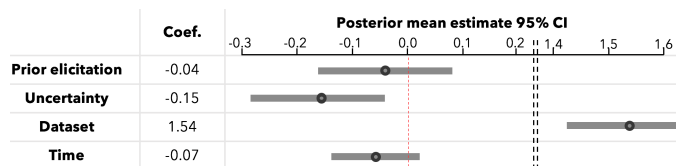|	Coef.	Posterior mean estimate 95% CI
Prior elicitation	-0.04	
Uncertainty	-0.15	
Dataset	1.54	
Time	-0.07	

Figure 13: Posterior mean estimates of effects with 95% confidence intervals from a model regressing the mean effect on individual log KLD on whether uncertainty visualization was shown, whether prior elicitation was prompted and which dataset was presented. Lower values indicate a greater effect toward lowering log KLD.

Perceived Sample Size. Even though participants assigned to examine the large sample size dataset had relatively high log KLDs relative to the small sample size dataset, viewing HOPs did have some impact on how accurately they perceived the sample size of the observed data. Figure 14 shows how the predicted perceived sample size of the observed data based on the dataset and whether uncertainty (HOPs) was presented. For the tech dataset (N=158), while the means of the No uncertainty and Uncertainty conditions were similar (326.0 vs. 327.3), the median was much closer to the actual sample size of the dataset for the Uncertainty conditions (median perceived: 166.3, true sample size: 158) than the No uncertainty conditions (median perceived: 97.2). For the elderly dataset (N=750,000), both the mean and median of the Uncertainty conditions were closer to the true observed sample size (mean perceived: 60,268.9, median perceived: 734.0, true sample size: 750,000) than the No uncertainty conditions (mean perceived: 809.54, median perceived: 216.1). These results suggest that presenting uncertainty information helps people arrive at normative Bayesian inferences. However, our results also suggest that uncertainty presentation may help some participants more than others.
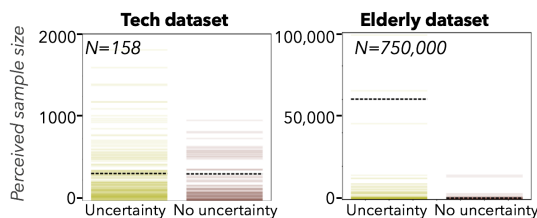

Figure 14: Perceived sample size for the tech and the elderly datasets. The uncertainty visualization helps participants more accurately perceived sample size in the both datasets.

7 DISCUSSION

Data Interpretation as Bayesian Cognition

Through three studies, we demonstrated how a Bayesian cognitive modeling approach can be used to interpret and evaluate how people update their beliefs after being exposed to a data presentation. Our work represents an important step forward for visualization research for several reasons.

First, our work provides evidence to suggest that in a naturalistic scenario wherein people bring prior beliefs, visualization cognition can be interpreted as a Bayesian process. While no single study can definitively establish whether people reason about presented data in a Bayesian way, our experiments showed that on average, people's responses were consistent with a sample-based Bayesian account when examining small sample size datasets. On average, people's responses deviated from Bayesian reasoning when presented with large sample size datasets. Future work might test a larger range of sample sizes to further characterize this bias. We found that the act of eliciting prior beliefs reduces deviation from the normative Bayesian posterior but not reliably so, and that aggregated prior distributions can be used in place of individual level priors to predict a person's deviation from the normative posterior.

While we used Bayesian cognition as an inferential model to examine how rationally people updated their beliefs at an individual and aggregate level, a Bayesian cognitive model like ours would support studying other influences in visualization or data interpretation. A Bayesian cognitive model could enable designers and researchers to derive more indepth insights about how graphical, contextual, and individual factors drive the difference between observed versus expected belief change. For example, preferences for certain states of the world are related to, but distinct from, subjective probabilities. Preferences over parameter values (e.g., what should the percentage of women responding positive to the survey question be?) could be elicited as a means of explaining deviation, as could other subjective beliefs like one's trust in the data source.

Our second contribution was to demonstrate how Bayesian cognitive modeling can be used prescriptively to evaluate how well visualization alternatives promote rational behavior. We used the alignment between participants' posteriors and the normative Bayesian posterior to reason about how "helpful" it is to add uncertainty representation to a visualization intended for a lay audience.

An important implication of our findings stems from the evidence we find that visualization users do not necessarily perceive the data they are presented with as "absolute truth" with regard to a phenomenon. In the case of very large samples, we found that may deviate substantially from changing their beliefs to match the data. This evidence provides a counterpoint to the implicit assumptions behind many design and evaluation techniques for visualizations. The prevalence of evaluations that rely on accuracy by comparing interpretations to the data directly and the prevalence of design guidelines that prioritize minimizing perceptual error are

more congruent with a normative view in which users *should* match their beliefs to the data than one that emphasizes how visualizations should be used to update users' existing beliefs. The Bayesian approach we demonstrate provides a way to quantify how much and in what direction people adjust their beliefs based on new data. Rather than having to carefully craft questions that one thinks might show a difference in accuracy between designs, an evaluator making use of Bayesian models can simply elicit priors and posteriors from a representative sample of users to understand how helpful a visualization is relative to another. We demonstrate a graphical sample method that can be used to elicit beliefs for arbitrary visualizations. The statistical literature offers techniques for fitting distributions of varying types to samples (including via Bayesian inference). By representing beliefs as distributions, as we did in fitting participants' responses to Beta distributions, an evaluator can infer additional information about how users perceived presented data from properties of the fit distributions. For example, we demonstrated how one could calculate the sample size that a rational Bayesian updater would have required to arrive at the same posterior beliefs as the user did, providing a valuable perspective on how a person was off in their interpretation.

Eliciting Mental Representation of Uncertainty

Our work did not identify a single, optimal technique for eliciting an untrained visualization user's internal representation of uncertainty for a proportion. With the exception of the graphical "balls-and-bins" interface, which tended to produce posterior beliefs that deviated slightly more from the normative solution, the techniques were difficult to distinguish. However we suspect that for more complex datasets and visual representations, the graphical sample-based technique is likely to have advantages. Imagine observed data visualized as a line chart. A graphical sample-based technique will allow people to simply draw more lines to represent their prior distribution, while other techniques would be more cumbersome (e.g., eliciting intercepts and slopes).

Our demonstrations involved presenting a proportion statistic. However even this simple scenario required reflecting on the best way to frame the elicitation of a prior. We chose to elicit prior and posterior distributions directly. These distributions are defined over parameter values (i.e., in model space). We chose to change the icons in the icon array format we used for elicitation and presentation to circles, rather than human icons, to better align with the notion of eliciting a population proportion. Alternatively, we could have elicited the prior and posterior predictive distributions by asking participants to think about the specific value (e.g., number of women) given some sample size.

In applying Bayesian modeling to interpretations of more complex datasets and visualizations, we believe that an important consideration will be identifying the appropriate level of "resolution" for the prior. For example, if the data is a spatially distributed statistic shown in a choropleth map, the relevant prior might involve subjective distributions over values for individual states, or it may be best represented by distributions over values at a national or regional level. In evaluation scenarios involving Bayesian cognition, the evaluator might consider which prior best matches the intended messages of the data presentation.

The insensitivity toward sample size of the observed data that we reported in Study 2 may partially be because the icon array did not encode sample size directly for the large data. One takeaway is that visualizations should do more to convey sample size, or impacts of sampling error.

One potential critique of using Bayesian cognition for designing or evaluating visualizations might be that it is unrealistic to expect lay visualization users to possess reasonable (or meaningful) priors on phenomena that tend to be presented to the public in outlets like data journalism. However, we cite multiple forms of evidence to the contrary. If people did not possess priors or were not able to articulate them, we might expect that with the sample-based elicitation techniques, which required providing samples with confidence, we would see a number of unidentifiable distributions due to all zero confidence values, or no variation in the sample values, for example. However, over 85% of participants who used sample-based techniques provided valid probability distributions. Moreover, across all elicitation techniques we saw that these distributions had non-trivial predictive power for posterior beliefs. A Bayesian model constructed with personal priors achieved a better fit (using Watanabe-Akaike Information Criterion (WAIC) [11, 53]) than did a model constructed with the aggregate priors or assuming a uniform prior (WAIC = 2315.7, 3159.8, 3159.8, respectively; see supplemental material for full details). Finally, several decades worth of work on probability elicitation in large surveys of the public have led economists to believe that laypeople are capable of providing useful information about real world phenomena via subjective probabilities [34].

8 CONCLUSION

Our work started by asking "Can Bayesian inference be used to explain how people update their beliefs upon viewing a data presentation?" In sum, our results demonstrate the promise of using Bayesian cognitive modeling to understand how data presentations like visualizations shape beliefs. Our work demonstrates a path toward better aligning studies of data interpretation with the undeniable effects of prior knowledge, and provides a valuable framework for evaluating new presentation methods.

9 ACKNOWLEDGEMENTS

REFERENCES

[1] 2016. OSMI Mental Health in Tech Survey 2016. https://www.kaggle.com/osmi/mental-2016%2011%2014ealth-in-tech-2016. Accessed: 2018-05-01.

[2] Chris Alcantara and Chiqui Esteban. 2016. 2016 Election exit polls. *The Washington Post*, Nov. 29, 2016, https://www.washingtonpost.com/graphics/politics/2016-election/exit-polls/,.

[3] Daniel J Benjamin, Matthew Rabin, and Collin Raymond. 2016. A model of nonbelief in the law of large numbers. *Journal of the European Economic Association* 14, 2 (2016), 515–544.

[4] Matthew Bloch and Hannah Fairfield. 2013. For the Elderly, Diseases That Overlap. *The New York Times*, Apr 15, 2013, https://archive.nytimes.com/www.nytimes.com/interactive/2013/04/16/science/disease-overlap-in-elderly.html,.

[5] Matt Canham and Mary Hegarty. 2010. Effects of knowledge and display design on comprehension of complex graphics. *Learning and instruction* 20, 2 (2010), 155–166.

[6] Patricia A Carpenter and Priti Shah. 1998. A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied* 4, 2 (1998), 75.

[7] Leda Cosmides and John Tooby. 1996. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *cognition* 58, 1 (1996), 1–73.

[8] Richard Cox. 1999. Representation construction, externalised cognition and individual differences. *Learning and instruction* 9, 4 (1999), 343–363.

[9] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty Displays Using Quantile Dotplots or CDFs Improve Transit Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 144.

[10] Bennett L Fox. 1966. A Bayesian approach to reliability assessment. (1966).

[11] Andrew Gelman, Jessica Hwang, and Aki Vehtari. 2014. Understanding predictive information criteria for Bayesian models. *Statistics and computing* 24, 6 (2014), 997–1016.

[12] Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review* 102, 4 (1995), 684.

[13] Daniel G Goldstein and Gerd Gigerenzer. 2009. Fast and frugal forecasting. *International Journal of Forecasting* 25, 4 (2009), 760–772.

[14] Daniel G Goldstein, Eric J Johnson, and William F Sharpe. 2008. Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *Journal of Consumer Research* 35, 3 (2008), 440–456.

[15] Daniel G Goldstein and David Rothschild. 2014. Lay understanding of probability distributions. *Judgment & Decision Making* 9, 1 (2014).

[16] Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008. Bayesian models of cognition. (2008).

[17] Thomas L Griffiths, Falk Lieder, and Noah D Goodman. 2015. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science* 7, 2 (2015), 217–229.

[18] Thomas L Griffiths and Joshua B Tenenbaum. 2006. Optimal predictions in everyday cognition. *Psychological science* 17, 9 (2006), 767–773.

[19] Lars Peter Hansen. 1982. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* (1982), 1029–1054.

[20] Mary Hegarty. 2004. Diagrams in the mind and in the world: Relations between internal and external visualizations. In *International Conference on Theory and Application of Diagrams*. Springer, 1–13.

[21] Mary Hegarty and Kathryn Steinhoff. 1997. Individual differences in use of diagrams as external memory in mechanical reasoning. *Learning and Individual differences* 9, 1 (1997), 19–42.

[22] Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha. 2018. Imagining Replications: Graphical Prediction & Discrete Visualizations Improve Recall & Estimation of Effect Uncertainty. *IEEE transactions on visualization and computer graphics* 24, 1 (2018), 446–456.

[23] Jessica Hullman, Paul Resnick, and Eytan Adar. 2015. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one* 10, 11 (2015), e0142444.

[24] Christopher Kaeser. 2015. A day in the Life. *The Wall Street Journal*, June 24, 2015, https://www.wsj.com/articles/were-working-more-hoursand-watching-more-tv-1435187603?cb=logged0.2694279181305319,.

[25] Daniel Kahneman and Patrick Egan. 2011. *Thinking, fast and slow*. Vol. 1. Farrar, Straus and Giroux New York.

[26] Alex Kale, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data. *IEEE transactions on visualization and computer graphics* (2018).

[27] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5092–5103.

[28] Daniel Kersten and Alan Yuille. 2003. Bayesian models of object perception. *Current opinion in neurobiology* 13, 2 (2003), 150–158.

[29] 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*. ACM, 1375–1386.

[30] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. 2018. Data Through Others' Eyes: The Impact of Visualizing Others' Expectations on Visualization Interpretation. *IEEE transactions on visualization and computer graphics* 24, 1 (2018), 760–769.

[31] Stephen M Kosslyn. 1989. Understanding charts and graphs. *Applied cognitive psychology* 3, 3 (1989), 185–225.

[32] Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics* 22, 1 (1951), 79–86.

[33] Stephan Lewandowsky, Thomas L Griffiths, and Michael L Kalish. 2009. The wisdom of individuals: Exploring people's knowledge about everyday events using iterated learning. *Cognitive science* 33, 6 (2009), 969–998.

[34] Charles F Manski. 2018. Survey measurement of probabilistic macroeconomic expectations: progress and promise. *NBER Macroeconomics Annual* 32, 1 (2018), 411–471.

[35] Bill Marsh. 2012. Are We in the Midst Of a Sixth Mass Extinction? *The New York Times*, June 1, 2012, https://archive.nytimes.com/www.nytimes.com/interactive/2012/06/01/opinion/sunday/are-we-in-the-midst-of-a-sixth-mass-extinction.html,.

[36] Richard E Mayer. 2014. Cognitive theory of multimedia learning. *The Cambridge handbook of multimedia learning* 43 (2014).

[37] Richard McElreath. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan.* CRC Press.

[38] Edward Munnich, Micheal Andrew Ranney, and Mirian Song. 2007. Surprise, surprise: The role of surprising numerical feedback in belief change. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 29.

[39] Hedwig M Natter and Dianne C Berry. 2005. Effects of active information processing on the understanding of risk information. *Applied Cognitive Psychology* 19, 1 (2005), 123–135.

[40] 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.

[41] Lace M Padilla, Sarah H Creem-Regehr, Mary Hegarty, and Jeanine K Stefanucci. 2018. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive research: principles and implications* 3, 1 (2018), 29.

[42] Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. 2014. The persuasive power of data visualization. *IEEE transactions on visualization and computer graphics* 20, 12 (2014), 2211–2220.

[43] Steven Pinker. 1990. A theory of graph comprehension. *Artificial intelligence and the future of testing* (1990), 73–126.

[44] Priti Shah, Richard E Mayer, and Mary Hegarty. 1999. Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of educational psychology* 91, 4 (1999), 690.

[45] Elsbeth Stern, Carmela Aprea, and Hermann G Ebner. 2003. Improving cross-content transfer in text processing by means of active graphical

representation. *Learning and Instruction* 13, 2 (2003), 191–203.

[46] Mark Steyvers, Joshua B Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum. 2003. Inferring causal networks from observations and interventions. *Cognitive science* 27, 3 (2003), 453–489.

[47] Joshua B Tenenbaum, Thomas L Griffiths, and Charles Kemp. 2006. Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences* 10, 7 (2006), 309–318.

[48] Joshua B. Tenenbaum Thomas L. Griffiths and Charles Kemp. 2012. Bayesian Inference. In *The Oxford handbook of thinking and reasoning*, Keith J. Holyoak and Robert G. Morrison (Eds.). Oxford University Press, Oxford.

[49] Amos Tversky and Daniel Kahneman. 1971. Belief in the law of small numbers. *Psychological bulletin* 76, 2 (1971), 105.

[50] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[51] Jarke J Van Wijk. 2005. The value of visualization. In *Visualization, 2005. VIS 05. IEEE*. IEEE, 79–86.

[52] Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum. 2014. One and done? Optimal decisions from very few samples. *Cognitive science* 38, 4 (2014), 599–637.

[53] Sumio Watanabe. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 11, Dec (2010), 3571–3594.

[54] Yujun Wu, Weichung J Shih, and Dirk F Moore. 2008. Elicitation of a beta prior for Bayesian inference in clinical trials. *Biometrical Journal* 50, 2 (2008), 212–223.

[55] Yifan Wu, Larry Xu, Remco Chang, and Eugene Wu. 2017. Towards a Bayesian Model of Data Visualization Cognition.

[56] Jeff Zacks and Barbara Tversky. 1999. Bars and lines: A study of graphic communication. *Memory & Cognition* 27, 6 (1999), 1073–1079.