PriorWeaver: Prior Elicitation via Iterative Dataset Construction

YUWEI XIAO, University of California, Los Angeles, USA
SHUAI MA, Aalto University, Finland
EUNICE JUN, University of California, Los Angeles, USA
ANTTI OULASVIRTA, Aalto University, Finland

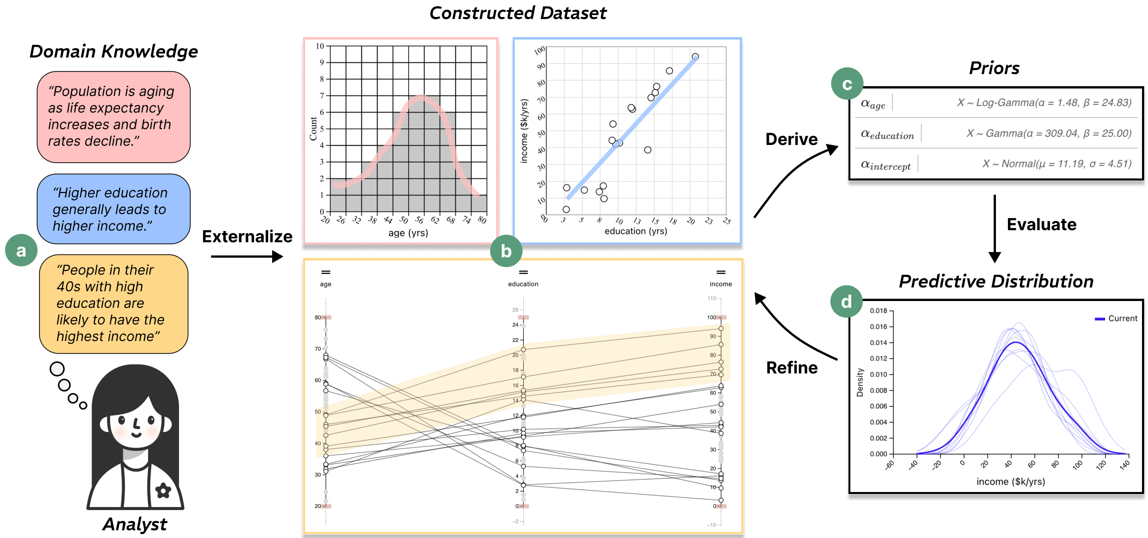

Fig. 1. PriorWeaver supports prior elicitation and iteration through the construction of a dataset representative of analysts' beliefs. (a) Analysts begin by considering their implicit domain knowledge about variable distributions (red), pairwise relationships (blue), and multivariate relationships (yellow). (b) Analysts express these assumptions through coordinated interactive visualizations, which simultaneously construct a representative dataset. (c) The dataset is used to derive statistical priors by fitting a predefined model. (d) Prior predictive checks visualize the predicted distribution of the outcome variable, which analysts can compare to their assumptions and use to iterate on their inputs.

In Bayesian analysis, prior elicitation, or the process of explicating one's beliefs to inform statistical modeling, is an essential yet challenging step. Analysts often have beliefs about real-world variables and their relationships. However, existing tools require analysts to translate these beliefs and express them indirectly as probability distributions over model parameters. We present PriorWeaver, an interactive visualization system that facilitates prior elicitation through iterative dataset construction and refinement. Analysts visually express their assumptions about individual variables and their relationships. Under the hood, these assumptions create a dataset used to derive statistical priors. Prior predictive checks then help analysts compare the priors to their assumptions. In a lab study with 17 participants new to Bayesian analysis, we compare PriorWeaver to a baseline incorporating existing techniques. Compared to the baseline, PriorWeaver gave participants greater control, clarity, and confidence, leading to priors that were better aligned with their expectations.

Authors' addresses: Yuwei Xiao, yuweix@ucla.edu, University of California, Los Angeles, Los Angeles, California, USA; Shuai Ma, shuai.ma@aalto.fi, Aalto University, Helsinki, Finland; Eunice Jun, emjun@cs.ucla.edu, University of California, Los Angeles, Los Angeles, California, USA; Antti Oulasvirta, antti.oulasvirta@aalto.fi, Aalto University, Helsinki, Finland.

1 INTRODUCTION

Researchers have been advocating for the adoption of *Bayesian analysis* methods across disciplines [6, 24, 42], including in HCI [32, 35]. Bayesian analysis can enable research communities to learn from small-sample studies, focus on effect sizes instead of statistical significance, and accrue knowledge across studies (e.g., meta-analyses) [35]. The mechanism by which Bayesian analysis realizes these benefits is through the integration of analysts' domain knowledge and assumptions (e.g., Figure 1a). Analysts imbue Bayesian analysis with their assumptions by specifying prior distributions for model parameters, or simply *priors* (e.g., Figure 1d).

The process of specifying these priors, referred to as *prior elicitation*, is critical yet difficult to do well [46, 51]. Prior elicitation requires not only expertise about a domain (e.g., *education year has a moderate positive impact to income*) but also the ability to express that knowledge using mathematical distributions over model parameters (e.g., *"the coefficient of education, given age, on income has a Normal(3, 0.5) distribution"*), which often do not have direct real-world analogs. When priors fail to capture analysts' domain knowledge, the subsequent inferences may be misleading or implausible, particularly in small-sample settings [14]. Indeed, a common practice is for domain experts to hire a Bayesian modeling expert to facilitate prior elicitation, a dependency that poses a barrier to wider adoption of Bayesian analysis.

Over the years, prior elicitation tools have made progress, from systems like MATCH [47] and PRELIZ [27] that rely on knowledge over parameters, to approaches that let analysts reason more directly about outcomes [3, 21]. Yet, gaps persist: most tools still rely on probability-format input, overlook relationships across variables, offer limited feedback for refinement, and provide little support for Bayesian novices.

We present PRIORWEAVER, an interactive tool that guides analysts through prior elicitation. The key idea of PRIORWEAVER is to approach prior elicitation as a dataset construction problem. Analysts directly express their assumptions about possible values one could observe about variables in the real-world (e.g., *"people in their 40s earn between $40k and $60k"*). This provides analysts a concrete and tangible representation of their domain knowledge. Under the hood, PRIORWEAVER constructs a concrete dataset from these inputs then derives prior distributions by fitting the statistical model to the dataset. PRIORWEAVER outputs prior predictive check visualizations that help analysts compare predictive outcomes and refine these priors. Figure 1 gives an overview of this interactive process.

To evaluate PRIORWEAVER, we conducted a controlled within-subjects experiment. Seventeen analysts experienced in statistical modeling but new to Bayesian analysis (i.e., Bayesian novices) specified priors for a statistical model using both PRIORWEAVER and a baseline parameter-based prior elicitation interface. Analysts reported feeling that they could express their knowledge more comfortably and clearly. They also produced initial priors that aligned closely with their beliefs and final priors. Analysts also reported that PRIORWEAVER made the feedback more actionable, enabling more effective and purposeful refinement. Our results suggest that shifting prior elicitation towards a constructive sensemaking process makes Bayesian analysis more approachable.

This paper contributes:

- A new perspective on prior elicitation as the process of constructing a dataset that captures analysts' assumptions without requiring direct parameter specification;
- PRIORWEAVER, an interactive system that supports iterative dataset construction through coordinated visualizations, derives statistical priors via bootstrapping, and provides feedback through prior predictive checks; and
- Evidence from a controlled lab study that PRIORWEAVER gave analysts helpful structure for externalizing domain knowledge, control over refining priors, and more positive attitudes toward Bayesian analysis.

Table 1. **Comparison of PriorWeaver with existing prior elicitation tools across key dimensions.** Elicitation space refers to what knowledge is elicited, either about parameter or observable. Elicitation modality, or how an analyst inputs their priors, is either graphical or textual. Elicitation format refers to the type of input required from the analysts. *Probability* denotes probabilistic input (e.g., mean, variance), while *Samples* denotes hypothetical samples. Feedback mechanism is how a tool shows analysts their priors. *PPC* refers to prior predictive checks, a common approach in Bayesian analysis. Multivariate refers to elicit beliefs about multiple parameters or variables simultaneously. Overall, existing tools typically operate in the parameter space, require probabilistic inputs, lack feedback mechanisms, and provide limited support for eliciting multiple parameters or variables simultaneously. In contrast, PriorWeaver integrates and extends these dimensions to offer more comprehensive support.

Tool	**Elicitation Space**	**Elicitation Modality**	**Elicitation Format**	**Feedback Mechanism**	**Multivariate** ($N > 2$)
MATCH [47]	Parameter	Graphical	Probability	–	×
SHELF [20]	Parameter	Graphical	Probability	–	×
Jones and Johnson [28]	Parameter	Textual	Probability	PPC	×
Sarma and Kay [54]	Parameter	Graphical	Probability	PPC	×
PreliZ [27]	Parameter	Textual	Probability	PPC	✓
Hartmann et al. [21]	Observable	Textual	Probability	–	✓
Bockting et al. [3]	Observable	Graphical	Probability	–	✓
Casement et al. [4]	Observable	Graphical	Samples	–	×
PriorWeaver	**Observable**	**Graphical**	**Samples**	**PPC**	✓

2 BACKGROUND AND RELATED WORK

Our work builds on and contributes to the literature on interactive tools for prior elicitation, perspectives on what makes a good prior, and the role of visualization in belief elicitation.

2.1 Tools for Prior Elicitation

The plurality of prior elicitation tools operate in the ***parameter space***, requiring analysts to express knowledge directly about statistical parameters. For example, SHELF [20] and MATCH [47] support this process by asking analysts to specify moments (e.g., mean, variance), quantiles, or histograms (trial-roulette method), and then fitting a distribution to these inputs. Additionally, the Preliz tool, the system from Jones and Johnson [28], and Sarma and Kay's tool [54] incorporate *predictive exploration* [31], allowing analysts to interactively change parameters and view the predicted distribution of the response variable implied by their parameter choices (i.e., prior predictive checks). While predictive exploration improves interpretability, these tools still require analysts to reason about abstract statistical parameters whose behavior and real-world implications are often difficult to intuit.

An alternative is to elicit priors in the ***observable space***, where analysts reason directly about measurable quantities. This aligns more closely with domain expertise, allowing analysts to specify patterns they have observed without translating them into parameters [46, 51]. For example, Casement et al. [4] ask analysts to iteratively select the most plausible histogram from a set of simulated samples. Then, their approach infers a prior from these selections. However, this method is limited to univariate models. More recent simulation-based approaches remove this requirement by letting experts specify expectations about multiple observable quantities and then automatically searching for priors that produce matching predictions. Techniques such as multi-objective Bayesian optimization [45] and stochastic gradient-based optimization [3, 21] learn priors that minimize the discrepancy between simulated and elicited information. Most observable-space tools still rely on probability-based inputs, provide limited feedback for validating priors, or are restricted to univariate specifications. In contrast, PriorWeaver enables analysts to externalize their knowledge in the observable space as a tangible dataset through coordinated visualizations. Analysts can then iteratively validate and refine these priors with the help of prior predictive checking.

Table 1 summarizes the characteristics of prior elicitation tools and compares PriorWeaver against existing tools. There is little consensus on what abstractions (parameter-space vs. observable-space) or interfaces help analysts express their domain knowledge as priors. This work articulates and evaluates design considerations for facilitating domain knowledge expression for the purpose of eliciting priors.

2.2 What Makes a Good Prior?

What constitutes a good prior is highly contested within the Bayesian modeling community [2, 11]. The primary dimension along which priors differ is *informativeness*, or how much influence a prior exerts on the model fitting process. This depends on how narrow (e.g., low variance) or broad (e.g., high variance) a distribution is, which directly influences how much the collected data are prioritized when fitting a statistical model (See Figure 1 in [54]).

At one extreme is a *non-informative* prior, which has high variance and wide tails and, as a result, prioritizes the collected data during the model fitting process over an analyst's domain knowledge. At the other extreme is an *informative* prior, which is more opinionated, has a smaller variance, and exerts strong influence on the model fitting process. In between these extremes lies the *weakly informative* prior [12], which is intentionally specified to contain less information than what might be available. In practice, there is limited guidance about where the boundaries between these categories lie, how to use them, and why one is preferable over another in the same analysis setting.

Moreover, whether one type of prior distribution is better than another remains debated. Key considerations are whether priors lead to useful statistical predictions and to what extent they reflect the analyst's implicit understanding of the domain. For example, a prior may support strong predictive performance but fail to capture how an analyst conceptualizes the problem.

PriorWeaver's goal is to elicit priors that faithfully reflect analysts' knowledge [10], which depend on analysts' beliefs and may span across the informativeness spectrum. PriorWeaver prioritizes the specification of priors that would generate data consistent with the analyst's understanding of the world [14].

2.3 Effects of Visualization On Belief Elicitation

Prior elicitation is a form of belief elicitation, or how people form and externalize their assumptions. Research shows that analysts often ground their beliefs in conceptual models rather than data [5, 29, 39], and that making these beliefs explicit improves recall and reflection [30, 37]. Recent visualization experiments further demonstrate that users externalize more data assumptions when sketching than when verbalizing [40].

Building on these findings, researchers propose various visualization techniques for belief elicitation. Goldstein and Rothschild demonstrate that drawing full distributions yields more accurate results than providing quantiles [18]. Other studies show that frequency formats often support better reasoning than probability formats [17, 26, 34]. Kim et al. introduce a graphical sample-based elicitation interface in which users provide individual sample values that are aggregated into a distribution [38]. Karduni et al. propose the Line + Cone method for eliciting beliefs about bivariate correlation, where users draw a central trend and specify their uncertainty around it [33]. Koonchanok et al. introduce an interactive scatterplot to elicit bivariate beliefs where users can adjust the slope of a trendline to indicate the expected relationship, and adjust the expected uncertainty in the relationship using a slider [41]. These techniques highlight diverse approaches to belief elicitation. Mahajan et al. [44] synthesize this body of work into VIBE, a design space that organizes belief-driven visualizations by elicitation goals and input modalities.

Extending belief visualization research that emphasizes comparing beliefs to data, PriorWeaver anchors elicitation in the construction of a concrete, manipulable dataset that serves as the foundation for incorporating analysts' beliefs into mathematical models in Bayesian analysis.

3 DESIGN CONSIDERATIONS

Existing prior elicitation tools rarely focus on observable-space elicitation, often rely on probability-format input, overlook relational knowledge across variables, and offer limited feedback for refinement. To address these gaps and incorporate suggestions in prior work [9, 10, 15, 31, 46], we articulated four design considerations that informed the development of PriorWeaver:

DC1: Analysts should be able to express knowledge in the observable space. Analysts are more familiar with observable quantities (e.g., values of age, income, education) than with abstract model parameters (e.g., regression coefficients) [31, 51]. Analysts should be able to work with representations that map directly to their domain knowledge when specifying priors.

DC2: Analysts should be able to express both distributional and relational knowledge. Analysts have knowledge about not only distributions of individual variables but also relationships among variables (e.g., how education correlates with income) [31]. Yet existing observable-space tools often emphasize distributional specifications, while neglecting analysts' need to express relational knowledge [30]. Systems should therefore enable analysts to express both distributions and relationships. Systems should also ensure these forms of knowledge are connected to capture the multidimensional nature of analysts' beliefs.

DC3: Analysts should receive actionable feedback to support evaluation and iterative refinement of priors. Frequent feedback is essential for enabling evaluation and iterative refinement of priors [31]. While prior predictive checking [9] is a common method for evaluating priors, the outputs (i.e., predictive distributions) are often disconnected from analysts' inputs, offering little guidance for correction. Instead, systems should connect analysts' externalized knowledge with the resulting priors explicitly.

DC4: Analysts should receive visualization support to conduct the elicitation process. Interactive visualizations provide interpretable ways to elicit inputs and evaluate priors [9, 46]. As such, systems should incorporate multiple visualization types to support flexible elicitation strategies analysts may have [54]. In addition, frequency-based visualizations may be most effective given that prior work has found that people interpret and reason with frequency formats more effectively than probability formats [19, 26, 50, 51].

4 USAGE SCENARIO

To illustrate how PriorWeaver supports iterative prior construction, we present a usage scenario involving Jane, a social scientist investigating how age and education years influence income.

Jane has a predefined linear model written in R ($income = \alpha \cdot age\ + \beta \cdot education\_years\ + \sigma$). She uploads the model script into PriorWeaver, which parses the code and initializes blank visualizations for the involved variables (i.e., age, education years, income) and the parameters to be inferred (Figure 2a).

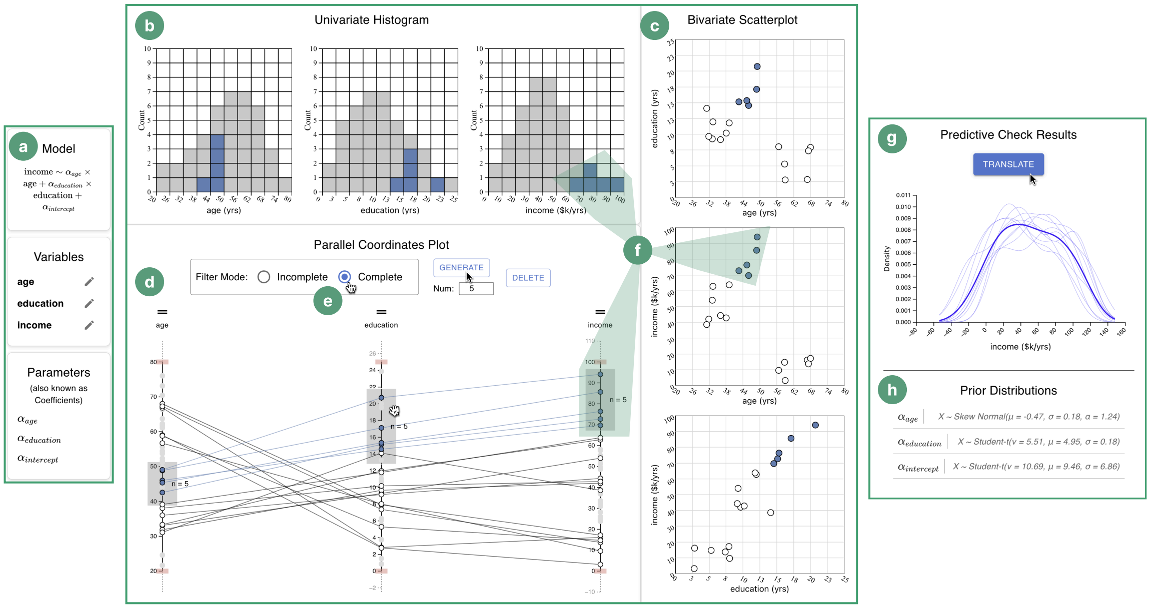

Fig. 2. **PriorWeaver's user interface.** (a) An information panel displays the model formula, variables, and parameters. To externalize their knowledge for priors, analysts work in the central coordinated visualizations panel, which includes (b) univariate histograms for variable distributions, (c) bivariate scatterplots for pairwise relationships, and (d) a parallel coordinates plot for multivariate relationships. (f) Brushing on the parallel coordinates plots' axes serves as a cross-filter, with selections (blue dots) synchronized across all visualizations. (e) Analysts can toggle between displaying complete or incomplete entities (white dots), and hide the others (gray dots). In *Complete* mode, they can use the GENERATE function to define multivariate assumptions within brushed regions and add the generated entities to the constructed dataset. To derive and evaluate the priors, analysts can click TRANSLATE to view (g) prior predictive checks and (h) suggested prior distributions.

Jane begins by drawing on her expertise. From past research, she expects most individuals in her population to be between 25 and 55 years old, with education clustering around high school, and income broadly distributed but skewed toward lower brackets. She first records individual values for each variable in the **univariate histograms** (Figure 2b), adding samples that capture typical ages, education levels, and income ranges. These entries are not connected across variables. Next, Jane turns to the **parallel coordinates plot** (Figure 2d) and uses the **CONNECT** (Figure 3) function to assemble these univariate samples into multivariate examples. This allows her to represent meaningful subgroups she has in mind, such as younger graduates with modest starting salaries or older adults with stable mid-range incomes. By switching between adding samples in the histograms and connecting entries in the parallel coordinates plot, Jane externalizes her domain knowledge as a synthetic dataset that embodies both distributions and relationships.

Once Jane has constructed this initial dataset, she clicks on the **TRANSLATE** button. PriorWeaver then derives prior distributions from the constructed dataset and provides a **prior predictive distribution** (Figure 2g) of income as feedback. To verify the behavior of derived priors, Jane compares this simulated distribution with the income histogram she specified earlier. While the two share a similar overall shape, she notices important discrepancies: the predictive distribution places too much probability on extremely high incomes. The distribution also exhibits a negative tail. These mismatches suggest that some of her earlier examples may have unintentionally overemphasized certain patterns.

Guided by this feedback, Jane revisits the visualizations. She switches to *Complete* mode and adds new examples with the **GENERATE** (Figure 2e) function. To mitigate the implausible negative tail, she generates cases of older adults with low education but still positive incomes. This reinforces her belief that income should remain above zero even with extreme cases. To counter the excessive weight on very high incomes, she adds examples of middle-aged adults with high education but only moderate incomes. This reflects her belief that high education does not always yield extreme earnings.

After this new round of externalization, Jane applies **translate** again. This time, the predictive distribution of income better reflects her expectations: Most of the income values lie in moderate ranges. The right tail captures economic disparity without dominating the distribution. All income values are positive. Upon generating a few more potential values, Jane arrives at priors she considers both credible and well-aligned with her domain expertise.

5 PRIORWEAVER: PRIOR SPECIFICATION AS DATASET CONSTRUCTION

The central contribution of PriorWeaver is the perspective that prior elicitation can be understood as a dataset construction problem. This perspective mirrors real-world practices, where analysts often reason about priors by referencing published datasets or concrete examples in mind [13, 50, 55]. An analyst-constructed dataset serves as a tangible artifact of the analyst's knowledge: columns represent distributional knowledge of individual variables, while rows capture relational knowledge across variables. Through constructing this dataset, analysts externalize implicit assumptions as concrete values in the observable space. This conceptual reframing comes with three technical challenges: (i) how to support analysts in effectively constructing the dataset, (ii) how to derive statistical priors from the constructed dataset, and (iii) how to ensure that the derived priors accurately reflect analysts' beliefs.

Below, we describe the system's design and implementation, including how PriorWeaver addresses these three technical challenges. PriorWeaver currently supports prior elicitation for generalized linear models (GLMs) involving only continuous variables. We discuss challenges and opportunities for expanding support for more statistical models and variable types in Section 9.

5.1 Externalizing Domain Assumptions via Coordinated Interactive Visualizations

To make dataset construction feasible, a key challenge is to design elicitation input methods that are both feasible for analysts to express their knowledge and sufficiently informative to derive meaningful prior distributions [46]. More specifically, analysts need interaction techniques that allow them to express assumptions about both individual variables and relationships across multiple variables in details [15, 51].

PriorWeaver addresses this problem through three coordinated interactive visualizations (Figure 2b-f). Each view highlights a different aspect of the dataset: histograms for distributional knowledge, scatterplots for bivariate relational knowledge, and the parallel coordinates plot for multi-variable relational knowledge. Together, these views support analysts to incrementally build a dataset that reflects their assumptions (see Figure 4). Importantly, changes in one view are immediately reflected in the other views.

5.1.1 Univariate Histogram. Analysts can externalize distributional assumptions of individual variables, such as plausible ranges, skew, or concentration around certain values, using univariate histograms. PriorWeaver uses the histogram for specifying hypothetical samples [26, 38] instead of allocating probabilities [18, 49] given that analysts may find it easier to think in terms of concrete values rather than probabilities [17, 26, 34]. Analysts can add observations by clicking directly within bins, and adjust the number of bins (default: 10) or the variable's range (default: 0–100) for finer granularity. In this way, abstract assumptions about variable distributions become tangible dataset entries, directly encoded in observable counts through frequency-based reasoning.

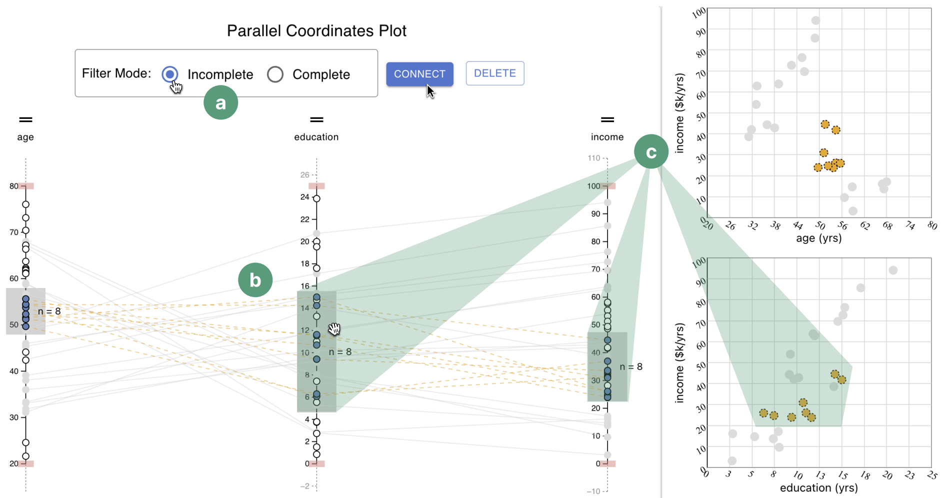

Fig. 3. **Building relationships across multiple variables using the parallel coordinates plot.** (a) When analysts select the INCOMPLETE mode, PRIORWEAVER displays only the incomplete entities (white dots) and hide the complete entities (gray dots). When analysts brush regions on axes to select and connect entities, PRIORWEAVER automatically identifies the maximum possible connections and previews (b) potential connections (orange dashed lines) and (c) corresponding potential entities (orange dots). Analysts can then click on CONNECT to establish connections and merge these entities in the underlying dataset under construction.

5.1.2 Bivariate Scatterplot. Moreover, analysts often have and want to express expectations about how two variables relate [29, 30]. PRIORWEAVER provides an interactive scatterplot where analysts can brush rectangular regions and use the GENERATE function to populate new examples within that region. These interactions allow tacit expectations about relationships to be encoded as concrete dataset entries, while automatically linking back to the univariate histograms and forward to the parallel coordinates view.

5.1.3 Parallel Coordinates Plot. Assumptions often span multiple variables (e.g., "older adults with high education but moderate income"), but higher-dimensional relationships are difficult to express with only univariate or bivariate views. As a result, analysts may lose track of how their marginal or pairwise inputs interact within the dataset, producing incoherent specifications.

PRIORWEAVER provides an interactive parallel coordinates plot (Figure 2d), where each axis corresponds to a variable and each polyline corresponds to a synthetic case (i.e., a dataset row). Analysts can select ranges on multiple axes and use the GENERATE function to add new cases that satisfy all selected constraints (see Figure 2e). They can also drag axes to reorder them. This global view extends the expressiveness of scatterplots to higher dimensions while helping analysts verify that their distributional and relational assumptions cohere at the dataset level.

5.1.4 Coordinating Multiple Visualizations. Since analysts externalize their assumptions in different views, the constructed dataset may contain both incomplete and complete entries. For example, adding values in a histogram creates incomplete rows with only one variable specified, while adding a case in the parallel coordinates plot creates a full row with values for all variables. As a result, analysts risk fragmenting their specifications into inconsistent pieces.

PRIORWEAVER provides two modes—*Incomplete* and *Complete*—and a CONNECT function to help analysts reconcile their specifications. By switching modes, analysts are able to focus on either incomplete entities (rows with missing values) or complete entries (i.e., rows with full values). Entities in the unselected mode visually fade into the background.

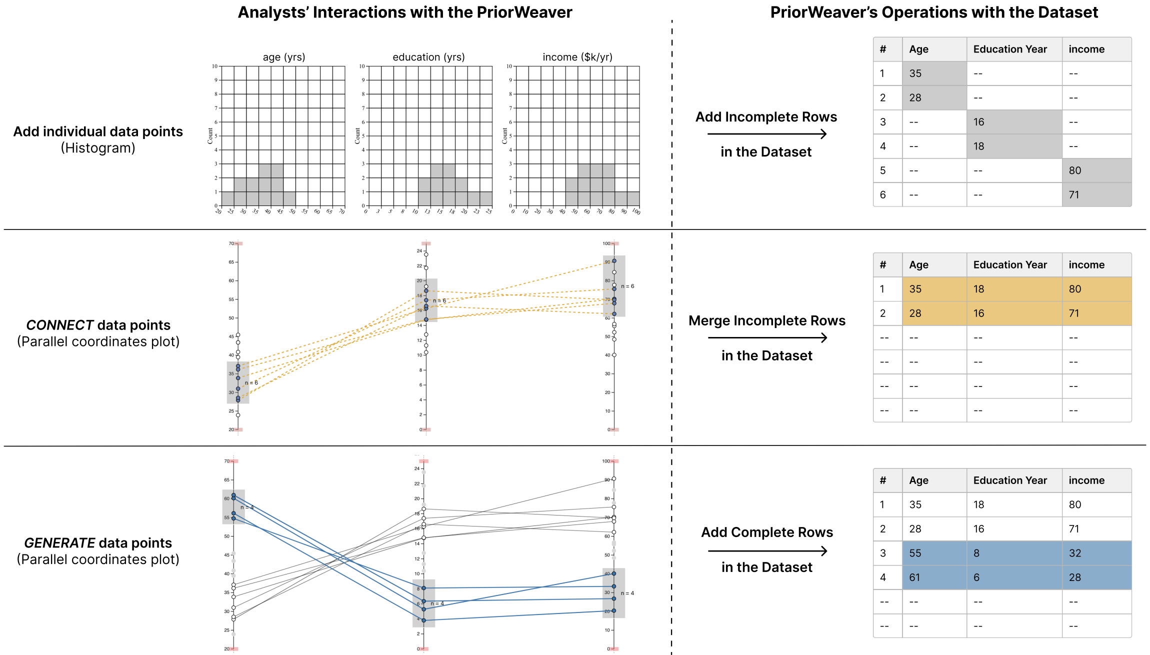

Fig. 4. **Interactive dataset construction.** As analysts interact with the visualizations to externalize their knowledge, PriorWeaver simultaneously constructs a dataset that represents this knowledge behind the scenes. The constructed dataset embodies analysts' knowledge in two dimensions: columns record distributional assumptions about each variable, while rows link these values together, reflecting relational knowledge across variables.

In the *Incomplete* mode, when analysts select compatible incomplete entries in the parallel coordinates plot, PriorWeaver automatically previews possible merges by connecting values with orange dashed lines and dots across the visualizations. This real-time visual preview enables users to validate and finalize connections (see Figure 3). Moreover, analysts can trace entries selected in one view in the other views when brushing-and-linking (Figure 2f).

Figure 4 shows how analysts leverage PriorWeaver's interactions to externalize and reconcile their knowledge into a coherent, representative dataset.

5.2 Deriving Statistical Priors

Analysts' observable-level assumptions must ultimately be transformed into formal statistical priors in order to be usable in Bayesian inference. However, deriving priors from observable assumptions poses two difficulties. First, user-constructed datasets may contain incomplete rows with missing values, which cannot reliably contribute to parameter estimation. Second, fitting priors directly from a single constructed dataset risks overfitting and fails to capture the inherent uncertainty of priors.

PriorWeaver addresses both of these challenges through a three-step procedure (see Figure 5). First, PriorWeaver filters out incomplete rows from the constructed dataset, ensuring that only fully specified cases contribute to translation. Second, PriorWeaver bootstraps 100 datasets, each created by randomly sampling 50 rows with replacement from the constructed dataset. Each bootstrapped dataset is then fit to the predefined statistical model to obtain a set of parameter estimates. This bootstrapping captures variability in the translation process and yields a more robust basis for estimation. Finally, PriorWeaver aggregates these parameter estimates and fits continuous probability distributions to

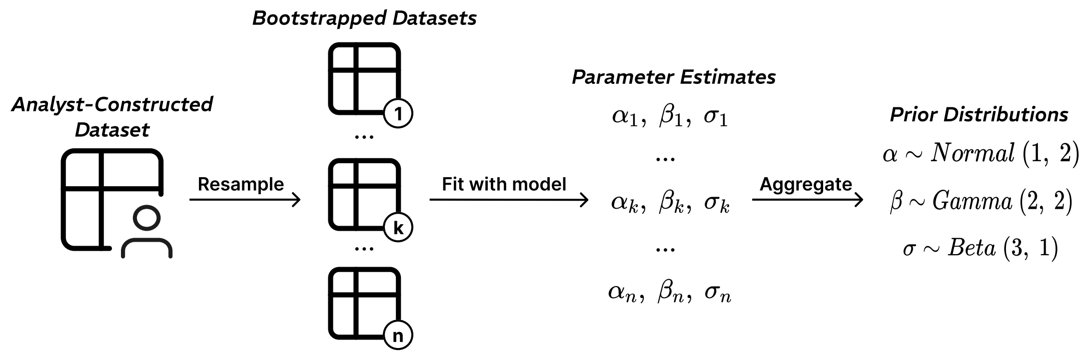

Fig. 5. **Deriving priors.** PriorWeaver derives priors in three steps. (1) First, generate multiple datasets by sampling with replacement from the analyst-constructed dataset. (2) Next, fit the pre-specified statistical model to each bootstrapped dataset and obtain parameter estimates. (3) Finally, aggregate these estimates across samples and smooth them to form continuous prior distributions.

each parameter's possible values using Maximum Likelihood Estimation (MLE). The resulting distributions constitute the derived statistical priors that can be used in a Bayesian analysis.

5.3 Evaluating and Refining Derived Priors

Evaluating whether derived priors reflect analysts' knowledge is challenging because priors are defined in parameter space, but assumptions are in the observable space. To address this, PriorWeaver employs prior predictive checks, a standard Bayesian technique for examining the consequences of prior distributions [9]. Prior predictive checks use simulations to generate predictive distributions in the observable space [15, 51].

In traditional workflows, prior predictive checks draw predictor values from an observed dataset and combine them with parameters sampled from the prior to generate predictions [9]. This requires analysts to have a dataset at hand when specifying priors, which is often unrealistic and contradicts some best practices of Bayesian statistics [13]. Instead, in PriorWeaver, predictor values are sampled directly from the analyst-constructed dataset. For each predictor, PriorWeaver independently draws 100 values with replacement from its marginal distribution, as specified by the analyst's histogram, and combines them into a simulated dataset of 100 cases. Then, PriorWeaver draws 10 parameter sets from the derived priors, each of which is paired with the simulated dataset to generate a predictive distribution for the response variable. As a result, PriorWeaver provides 10 predictive distributions along with their average, as illustrated in Figure 6. More importantly, analysts can directly compare the predictive distribution against the histogram of the response variable they constructed (e.g., the histogram of income in Figure 6).

In this way, PriorWeaver not only offers an interpretable evaluation but also creates clear pathways for refinement. For example, predictive checks may reveal discrepancies, such as implausible negative incomes or extremely high average income. Analysts can return to the coordinated visualizations to adjust their assumptions by adding examples to enforce plausible ranges or re-balancing distributions. As a result, prior elicitation becomes an iterative loop rather than a one-off specification.

6 USER STUDY

PriorWeaver introduces a new interactive process for specifying priors via iterative construction of a dataset representative of analysts' domain knowledge. Our goal is to assess the acceptability and impact of using observable space expressions of domain knowledge to derive statistical priors. The following research questions guided our study:

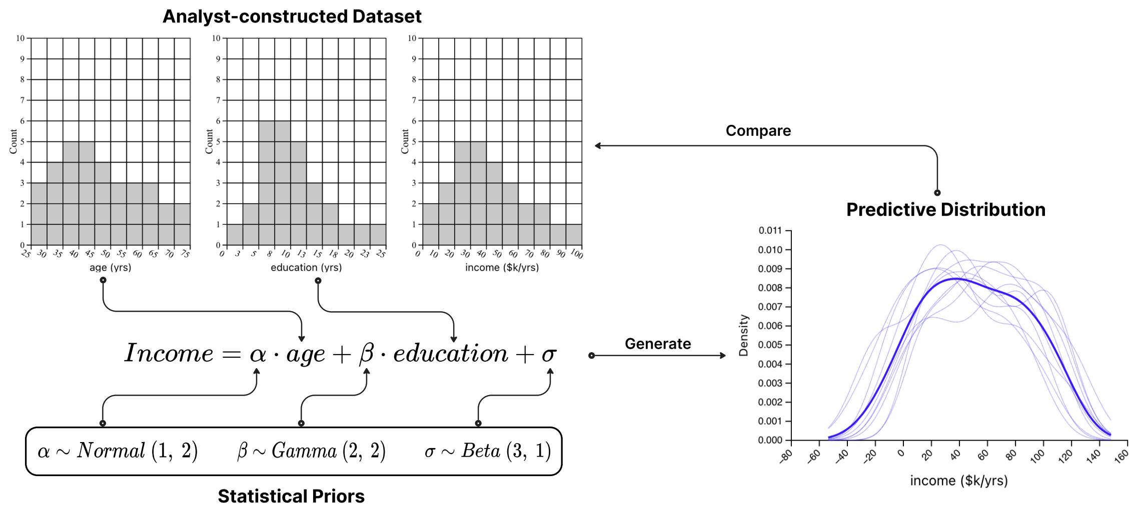

Fig. 6. **Feedback through prior predictive checks.** PriorWeaver samples predictor values (e.g., age, education) from the analyst-constructed dataset (top left), draws parameter sets (e.g., α, β, σ) from the derived priors (bottom left), and combines them to generate predictive distributions (right). On the rigth, each predictive distribution is shown as a faded blue line, with the average depicted as a solid blue line. Analysts can detect discrepancies between these predictive distributions and the histogram of the response variable (e.g., income) (top right).

- **RQ1: Externalizing knowledge for priors.** What strategies do analysts adopt when expressing assumptions in the observable space versus parameter space? How does PriorWeaver support analysts in externalizing their implicit domain knowledge compared to a parameter-space tool?
- **RQ2: Evaluating and Refining priors.** How does PriorWeaver affect how analysts evaluate and iterate on priors compared to a parameter-space tool?
- **RQ3: Attitudes towards Bayesian analysis.** How does PriorWeaver impact analysts' perspectives on Bayesian analysis? Specifically, are they more likely to adopt Bayesian methods using a tool like PriorWeaver? Why or why not?

6.1 Experimental Design

We conducted a within-subjects design with interface (PriorWeaver vs. baseline) and task (student performance vs. gym member weight) as the independent variables. The two interface conditions were

- **PriorWeaver**: Analysts interact with the full set of features offered by our system, including expressing priors in the observable space and receiving feedback via prior predictive checks (See Figure 2).
- **Parameter space baseline**: Analysts directly express their priors in the parameter space via trial-roulette method [49] and are shown prior predictive check results. We created this interface (See Figure 9 in the Appendix) to incorporate prevailing practices of prior elicitation [46].

We chose an within-subjects design in order to control for individual differences in domain and statistical knowledge between interface conditions.

Table 2. **Participant demographics and background information.**

PID	Age	Gender	Field	Study/Job
P1	27	Male	HCI	PhD student
P2	22	Male	Computer Science	Master student
P3	23	Male	Computer Science	PhD student
P4	23	Male	Computer Science	PhD student
P5	23	Female	HCI	Master student
P6	35	Male	HCI	Postdoc
P7	23	Male	Computer Science	Master student
P8	21	Male	Data Science	Undergraduate
P9	30	Female	HCI	Researcher
P10	22	Female	HCI	Master student
P11	30	Male	Communications	Master student
P12	23	Male	Computer Science	PhD student
P13	18	Male	Data Science	Undergraduate
P14	27	Female	Design	Designer
P15	22	Female	Data Science	Undergraduate
P16	27	Female	HCI	PhD candidate
P17	28	Male	Robotics	Master student

6.1.1 Analysis Tasks. We selected two analysis tasks from prior studies on human decision-making. They are comparable in complexity and rely minimally on specialized expertise.

- **Student Performance Prediction**: Analysts predict college students' exam scores based on their hours of study per week and attendance rate to class. This task is from previous decision-making studies [7, 8, 16, 56]. The corresponding dataset we used for this task is publicly available on Kaggle [1].
- **Gym Member Weight Prediction**: Analysts predict gym members' weight based on their height and exercise level (i.e., hours of exercise per week). The corresponding dataset we used for this task is publicly available on Kaggle [36].

6.2 Participants

Participants were representative of our intended user group for PriorWeaver, analysts familiar with statistical modeling but new to Bayesian analysis. With approval from our institution's IRB, we recruited 17 participants [1] (6 female, 11 male) through email and social media outreach at local universities. They came from diverse fields, including computer science, communication, and design. Participants were undergraduates, graduate students, and recent graduates currently in the workforce. On a five-point scale, participants self-reported moderate to substantial familiarity with the concepts of frequentist statistics ($mean = 4.0, std = 0.4$) and linear modeling ($mean = 3.6, std = 0.6$). Participants also rated their familiarity with the concepts behind Bayesian analysis (e.g., Bayes' theorem) as slight to moderate ($mean = 2.7, std = 0.2$), but, importantly, none of the participants had performed a Bayesian analysis before. The study lasted approximately 80 minutes, and participants received a $25 USD gift card for their time.

6.3 Procedure

After obtaining informed consent, a researcher introduced the study's objectives and overall procedure. Participants then completed a pre-task survey assessing their background in statistics and gathering demographic data.

[1]We included the pilot participant (P1) to provide additional data. The remaining 16 participants followed the counterbalanced experimental design.

Participants completed two analysis tasks, each using a different interface. Task order and interface assignment were counterbalanced and randomly assigned to participants. For each task, participants followed the same five-step process:

1. **Tutorial & Practice**. Participants first watched a tutorial video introducing the interface (PriorWeaver or baseline). Participants then freely explored the interface until they felt familiar with it. The tutorials for both interfaces used the same analysis example task, which was to predict income based on age and years of education. We chose this task for the tutorials since it is distinct enough from the experimental tasks and has been widely used in previous decision-making research [16, 22, 53, 57].
2. **Task Introduction**. A researcher introduced the analysis task (student performance or gym member weight).
3. **Prior Elicitation**. After confirming their understanding of the task context and objective, participants specified their priors using the interface. Participants self-elected to conclude when they judged the priors to adequately reflect their knowledge.
4. **Post-Task Survey**. Participants completed a questionnaire related to their experience with the current condition.
5. **Post-Task Interview.** A researcher asked participants open-ended questions about their experience with the system and the analysis task.

After completing all five steps for the first task, participants proceeded to the second task. After both tasks, a researcher engaged participants in a final semi-structured exit interview about additional reflections and feedback on their experiences. All study materials are included as supplemental material.

6.4 Measurements and Analysis

We analyzed data from system interaction logs, surveys, and semi-structured interview transcriptions.

Quantitative metrics are based on participants' survey responses and their interaction logs. The data violated the parametric assumption of normality, so we conducted Wilcoxon signed-rank tests for paired comparisons between interface conditions. We applied the Benjamini–Hochberg correction to account for multiple comparisons.

We conducted a thematic analysis of the open-ended survey responses and semi-structured interview transcriptions. Our pre-defined interview questions guided the coding process. Two authors independently coded the transcripts, developed a shared codebook, and resolved discrepancies through discussion.

7 FINDINGS

7.1 RQ1: Strategies for domain knowledge externalization

7.1.1 There were three strategies for externalizing domain knowledge in the observable space using PriorWeaver. Participants often began with a single strategy, commonly distribution-driven or example-driven, and then flexibly switched across strategies, moving back and forth as needed to articulate their knowledge.

- **Distribution-driven strategy: Matching marginal shapes.** When participants had clear expectations about how individual variables should behave, they focused on shaping the marginal distributions of those variables. For instance, they adjusted data points in histograms until a distribution's center, spread, or skew, aligned with their beliefs.
- **Association-driven strategy: Matching pairwise relationships.** When participants thought in terms of how variables relate to one another, they concentrated on expressing these pairwise associations. For example, they used the parallel coordinates plot and scatterplots to represent and verify their assumptions about the direction and form of relationships (e.g., positive, negative, linear, or nonlinear) between variables.

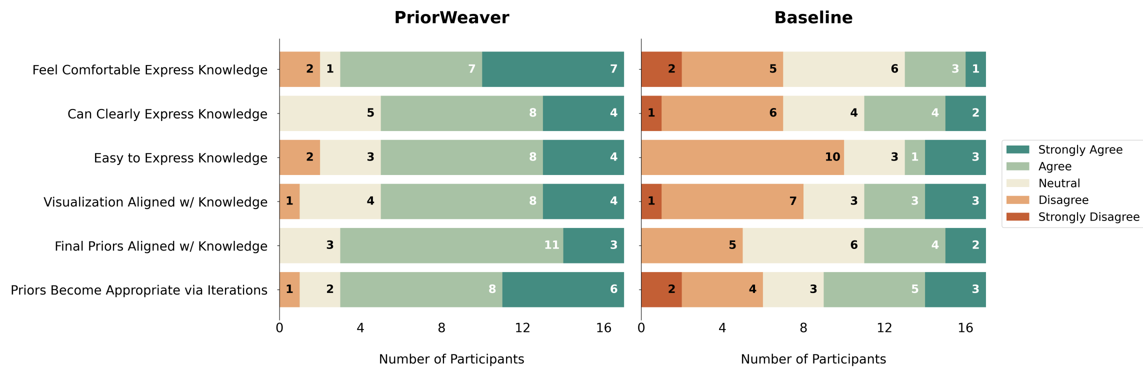

Fig. 7. **Participants' survey responses comparing PriorWeaver and the baseline.** Participants reported that PriorWeaver provided stronger support for articulating their knowledge, obtaining priors aligned with their expectations, and refining priors effectively.

- **Example-driven strategy: Matching multivariate patterns.** When participants anchored their reasoning in specific, concrete real-world examples, they specified data points across variables to represent an individual example. For instance, they drew out a single plausible case. They used the parallel coordinates plot to encode multivariate patterns that captured how variables combine to form one entity observable in the real-world.

The vast majority of participants (14/17) began with the distribution-driven strategy, focusing first on histograms. As P7 explained, *"it felt natural to start with the histogram and think about each variable in isolation."* Similarly, P1 noted that *"I drew the histogram first, so that I could be more purposeful when connecting different variables."* From there, some participants (P1, P12, P13) moved to the relationship-driven strategy, emphasizing that they *"felt confidence in pairwise relationships"* (P13) and that *"pairwise relationships were very direct"* (P12). Others (P3, P5–8, P10–17) transitioned to the example-driven strategy. As P7 described, *"when the [data points] were established from histograms, I would then think about the different groups of real-world examples that I had in mind, such as low X + high Y = high Z."*

A few participants (P2, P4, P9) started with the example-driven strategy, creating multivariate examples in the parallel coordinates plot. P9 described how relational knowledge guided his process: *"Rather than distributions, some relational knowledge (i.e., examples) came to mind right away when I saw the task. So I expressed them in the parallel coordinates plot, then worked backward to the histograms to see if the distributions made sense."* Similarly, P4 explained his reasoning in detail: *"To my knowledge, students who study more hours and have a higher attendance rate usually get better final exam scores. So I used GENERATE to create these examples first. And I have an approximate mean value and shape of distribution for study hours in mind, so I built the distribution next."*

7.1.2 With the baseline tool, participants struggled to follow a clear strategy for externalizing their knowledge. The majority of participants (10/17) relied on guessing through trial and error when using the baseline tool. As P5 described, *"[I] would randomly guess a mean value and a distribution shape, then tweaked them until the predictive distribution looked appropriate."* The other seven participants vocalized beliefs in observable space terms first then manually translated these into the parameter space. P7, P13, P14, and P17 performed *"rough mental conversion"* (P14). P3, P11, and P12 completed *"manual calculation[s] on paper"* (P3) to estimate plausible mean values and ranges for parameters. P3 explained, *"I set up an equation that if someone has maximum study hours and attend[s] every class [they] should have a 100 exam score. Based on that, I estimated that both parameters' mean should be around 0.5."* These approaches demonstrate how, without PriorWeaver, the cognitive burden of encoding their beliefs into statistical parameters was on the participants.

Echoing these observations, participants reported that the visualizations in PriorWeaver were well aligned with their knowledge (Z = -2.280, p < 0.01), which reduced the cognitive effort required to express their knowledge and allowed them to apply purposeful strategies.

7.2 RQ1: General support for knowledge externalization

7.2.1 The parameter space lacks support for translation of domain knowledge. In the baseline condition, most participants (15/17) struggled to express their knowledge about parameters. They described the parameter space as *"too abstract"* (P17) and admitted being *"confused about the meaning of parameter distributions"* (P3). As P4 explained, they *"have knowledge about the variables, but it is hard to correctly and clearly transform the knowledge about the variables into knowledge about the parameters."* Similarly, P11 remarked that *"there is a gap between what I knew and what I need to express."* Beyond understanding individual parameters, participants (P2, P3, P7, P9, P17) also found it difficult to reason about them jointly. Even with a background in machine learning, P2 explained, *"it is difficult to set parameters individually and consider the combined effects at the same time. This often lead[s] to unexpected results."*

Interestingly, when asked how to cope with these challenges, seven participants (P3, P7, P11-14, P17) reported that they would reason about the variables rather than the parameters themselves. For example, some described that they would *"think about the correlation between a variable and the outcome"* (P12) or *"do a rough mental estimate of how each variable would affect the outcome"* (P17). Similarly, P3 explained that he would *"set up an equation and fill in variable values to calculate the possible values of parameters."* In other words, participants were imagining and simulating manipulations in the observable space.

7.2.2 The observable space makes knowledge externalization more direct. With PriorWeaver, all participants appreciated being able to externalize their knowledge in the observable space directly. They explained that this approach *"aligns better with [their] natural thinking process"* (P11) and *"abstracts away the need to deal with parameters"* (P7). Several described the process as *"intuitive"* (P3, P4, P6, P10, P12, P17) because they can *"input real values from examples I encounter in daily life"* (P10) and *"reason with frequency format"* (P6). For instance, P3, who shared that they frequent the gym, reasoned about the weight prediction task by recalling *"several figures and samples in my mind, of people who I regularly see in the gym."* In this way, PriorWeaver allowed participants to focus on their domain knowledge instead of figuring out how to represent that as parameters.

Indeed, as shown in Figure 7, participants gave significantly higher ratings to PriorWeaver than to the baseline in terms of comfortable of expression (Z = -2.866, p < 0.05), clarity of expression (Z = -2.684, p < 0.01) and ease of expression (Z = 3.022, p < 0.01).

7.2.3 PriorWeaver provides better support for expressing knowledge about variable relationships. Ten participants (P2-5, P7-9, P11, P12, P16) valued how PriorWeaver enabled them to see and build relationships between variables more *"clearly"* (P7) and *"easily"* (P8). Participants found PriorWeaver's parallel coordinates plot particularly helpful. P2 elaborated on how PriorWeaver supported them: *"It's hard to express the final outcome based on each variable separately. You have to combine them together and express that kind of knowledge. I think the parallel coordinates plot, which allows me to connect or generate the data points, helps me effectively express the combined effects of multiple variables."* P5 also noted how the parallel coordinates plot *"helped capture knowledge that might otherwise be overlooked if expressed only through histograms".* Furthermore, three participants (P7, P9, P13) highlighted the usefulness of the

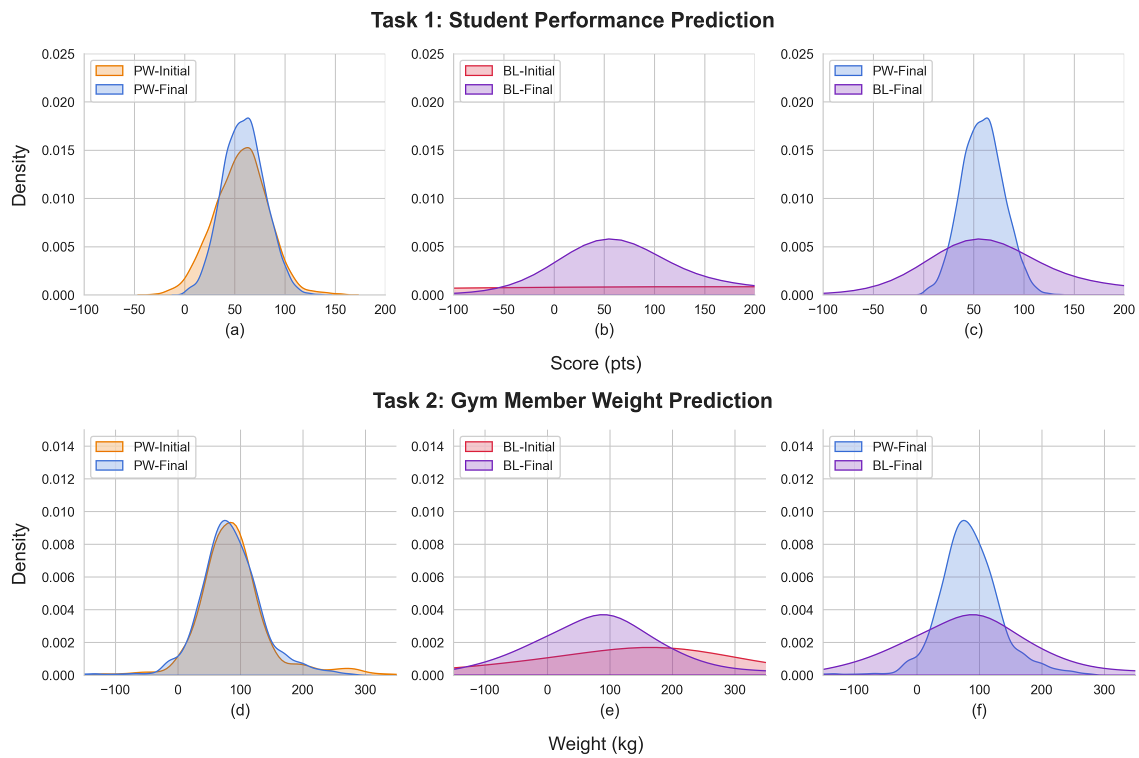

Fig. 8. **Participants specified priors that produced more reasonable predictive distributions when using PriorWeaver.** *PW* denotes PriorWeaver and *BL* denotes the baseline. Predictive distributions are aggregated across participants. Figures (a) and (d) compare initial and final priors in PriorWeaver. Figures (b) and (e) compare initial and final priors in the baseline. Figures (c) and (f) compare the final priors between PriorWeaver and the baseline. In both tasks, participants obtained priors reasonably aligned with their knowledge on the first attempt with PriorWeaver and eventually yielded more reasonable priors than the baseline.

bivariate scatterplot. P7 found that it was *"a great breakdown of what the parallel coordinates plot is trying to show"* and allowed them to *"focus on two variables at a time and validate: Does this relationship make sense?"*

At the same time, the requirement to express variable relationships was challenging for some participants (P11, P12, P13, P15). For instance, P13 explained that they had beliefs about only some subset of the variables, so *"it was hard to find clear relationships between all variables."* Related, four participants (P4, P7, P15, P17) remarked that the parallel coordinates plot and its functions made PriorWeaver more tedious for externalizing knowledge and more challenging to learn initially compared to the baseline tool.

7.3 RQ2: Evaluating and refining Priors

7.3.1 Participants used similar criteria for assessing prior quality in both tools. Participants (P2, P3, P4, P7, P9, P11-14, P17) paid attention to the ranges and tails of the predictive distributions. They checked for any unrealistic extreme values. In the exam score prediction task, five participants (P3, P7, P9, P13, P17) mentioned *"if there is notable density of the distribution exceed 100, then [they] would refine another round"* (P7). When visually assessing the prior predictive distribution, five participants (P4, P5, P6, P8, P14) looked to see that *"the majority of [the] density should fall between (or* *around) [certain] interval[s] that approximately matches with what [they] expressed"* (P4). Three participants (P7, P12, P17) focused more on the *"overall shape of the distribution"* (P17).

7.3.2 In the baseline, refinement was frustrating and unpredictable. The majority of participants (P1, P3-6, P8, P10, P14, P16, P17) struggled to identify what to adjust and how even after recognizing that the predictive distributions were not aligned with their expectations. Participants ultimately resorted to trial-and-error. For example, P3 explained, *"The results still show scores over 100. I know I should have eliminated that part, but I didn't know how to do it."*

Six participants (P2, P7, P9, P11, P12, P13) tried *"narrowing parameter ranges to reduce unrealistic values"* (P13) or *"lowering peak parameter values to lower the effect"* (P7). However, these adjustments often failed to produce the expected results. P2 explained how *"even small changes could lead to large and unintended shifts in the predictive distribution."*

7.3.3 In PriorWeaver, iteration was more purposeful. When participants noticed undesired tails or ranges in the predictive distribution, they (P1, P4, P5, P6, P9, P10, P15) responded by adding specific data points to capture extreme scenarios or by removing unrealistic cases. A couple participants (P10, P14) also added points to counterbalance results. P4's reflection on their refinement process with PriorWeaver succinctly summarized its main benefit for iteration: *"[I] could clearly locate where anomalies were and have a concrete direction for refinement."*

For instance, to shift distributions with undesired shapes, six participants (P2, P6, P11, P12, P13, P17) modified specific intervals. P11 described, *"The predicted average score was too low, so I linked some hardworking but low-attendance students with high scores, which might increase the overall exam scores."* P2 had a similar approach: *"The average grade should be between 70 and 80, but the distribution was uniform. So I added more points around that specific grade interval."* In addition, participants (P2, P4, P7, P13) refined the strength of relationships. As P7 put it, *"The number of points can represent the 'weight' of relationships in the model."* Using the generate function, P7 emphasized how it allowed them to *"quickly establish new relationships or emphasize certain relationships on the fly."*

When comparing the baseline and PriorWeaver, participants reported that the final elicited priors in PriorWeaver were more aligned with their knowledge (Z = -2.397, p < 0.05). Furthermore, we observed that the quality of prior predictive distributions was more appropriate at the end when participants used PriorWeaver than when they used the baseline. For instance, most predicted scores fall between 20 and 100, and most predicted weights between 40 and 150. The initial priors that analysts specified were also more reasonable and closer to their final priors with PriorWeaver than with the baseline. Figure 8 illustrates this. Put simply, PriorWeaver enabled analysts to specify reasonable priors right away and refine them meaningfully, resulting in priors that were more reasonable overall.

7.4 RQ3: Attitudes towards Bayesian analysis

Participants reported that PriorWeaver made the complex task of Bayesian prior elicitation more approachable for them (P2-8, P10, P14). P14 summarized that the tool *"breaks down Bayesian analysis into bite-sized chunks, making the process far less overwhelming than learning from textbooks."* Similarly, P7 appreciated that the system allowed him to *"quickly get started exploring relationships without heavy technical setup,"* highlighting its support for exploratory, iterative thinking. Participants found PriorWeaver significantly more helpful for eliciting their knowledge compared to the baseline (Z = -3.169, p < 0.01). They also reported higher confidence in using it (Z = -2.939, p < 0.01) and found it less cumbersome to use (Z = -2.359, p < 0.05).

Participants (P2, P3, P11, P14) also expressed that PriorWeaver transformed Bayesian analysis from a theoretical concept into a practical tool for real-world use. P3 emphasized, *"If I don't have these tools, Bayesian analysis would only exist in textbooks for me."* Additionally, P11 shared, *"This process really gave me an introduction to Bayesian analysis. It* *made me realize it's something I can actually apply in my future research."* Participants were significantly more inclined to apply Bayesian methods in future projects if they had access to PriorWeaver than the baseline (Z = -2.676, $p < 0.01$).

7.5 Opportunities to improve PriorWeaver

Participants gave feedback on ways to improve PriorWeaver. First, a couple participants (P6, P11) were unfamiliar with parallel coordinates plots and asked for clearer onboarding materials, simpler explanations of system features (e.g., the distinction between complete and incomplete modes), and lightweight tutorial videos to support familiarization. Second, participants (P1, P10, P15) wanted more insight into and control over the entire process. Participants asked for more transparency into the process of turning the visual inputs into statistical priors (P1), finer histogram bins (P10), and recommended parameter ranges to support refinement (P15). Third, participants hoped for more flexible ways to express their knowledge, including incorporation of categorical variables (P7), nonlinear variable relationships (P16), and multiple input modalities (e.g., rule-based logic, natural language) (P9).

7.6 Key takeaways

PriorWeaver scaffolded the elicitation process by guiding participants through knowledge externalization and prior iteration. For both, PriorWeaver gave participants a greater sense of control in formulating goals, working towards them, and assessing their progress. During knowledge externalization, PriorWeaver's use of the observable space allowed participants to accurately express their knowledge, while employing multiple possible strategies. As a result, participants reported feeling more comfortable and clearer about what and how to express, leading to initial priors that were more reasonable and better aligned with their expectations. While iterating on priors, participants leveraged prior predictive checks to evaluate priors in both tools. Yet, PriorWeaver made the feedback more actionable and facilitated more goal-oriented refinements. Without PriorWeaver, participants relied on trial-and-error changes. Overall, PriorWeaver lowered the barriers to prior elicitation and Bayesian analysis for participants.

8 DISCUSSION

This paper develops and evaluates PriorWeaver, an interactive system for prior elicitation via iterative dataset construction. PriorWeaver scaffolds elicitation by guiding participants to externalize knowledge in the observable space and structures prior iteration around predictive checks. In a lab study, we found that PriorWeaver helped participants express their knowledge more systematically and directly. It also helped participants refine priors more effectively than with trial-and-error approaches common with the parameter-space baseline. Our design process and the evaluation results give insight into the design of abstractions for eliciting beliefs and the role of constructed datasets as knowledge representations.

8.1 Balancing usability and control in eliciting beliefs

With PriorWeaver, participants in the user study could easily specify distributions and relationships in the observable space. The underlying parameterization of the model was abstracted away. As a result, they could not directly change exact coefficients or distributional forms. In contrast, the parameter-space baseline system gave participants fine-grained control, but most participants struggled to wield this control effectively due to a lack of Bayesian analysis expertise. Many reported that adjusting parameters felt like *"guessing"* (P5), and changes often led to unintended consequences in the predictive check results. Based on these results, we conclude that trading off control for ease of expression is the right design decision for Bayesian novices.

Generalizing beyond Bayesian analysis, our evaluation results suggest that interfaces should require users to translate their beliefs as little as possible. Systems should also provide feedback in a form that matches the input in order to make it actionable. Without this kind of support, users have the burden of manually translating their implicit beliefs into terms appropriate for the task. This process is not only likely to be error-prone but also lead some to resort to trial-and-error.

8.2 Constructed datasets as knowledge representations

PriorWeaver's design intends for users to engage with their implicit beliefs (and gaps therein) through interacting with the coordinated visualizations. In the user study, participants adopted different strategies based on what they knew about the domain. For example, some started with distributions of variables. Others thought through concrete examples across all variables. In other words, PriorWeaver scaffolded how analysts recollected their domain knowledge. Through iteration, the constructed datasets came to capture a data generating process for the domain.

Furthermore, the dataset that analysts create through the interactive visualizations serve as an intermediate representation between analysts' abstract domain knowledge and their statistical priors. Viewed in this light, the constructed dataset need not only "target" statistical priors but could also be useful for other data-related tasks. For example, the constructed dataset may be useful for a larger data analysis pipeline, such as the Bayesian Workflow [15]. In these applications, the constructed dataset could serve as a shared representation [23] between analysts and a system.

8.3 Benefits of concretizing implicit assumptions

PriorWeaver grounded analysts' thinking in concrete, familiar representations. Specifically, its interactive visualization interface provided a higher level of abstraction than existing parameter-space prior elicitation tools. The result was that prior elicitation felt less like performing abstract statistical modeling and more like expressing analysts' knowledge. In the study, participants felt greater confidence and expressed less confusion when using PriorWeaver, even without additional guidance or previous Bayesian analysis experience [52]. These observations align with prior research showing that interfaces which reflect users' mental models and everyday reasoning can foster greater confidence and engagement, particularly in complex or technical domains [25, 43].

8.4 Lowering the barrier to Bayesian analysis

Beyond supporting prior elicitation process, PriorWeaver helped shift participants' broader attitudes toward Bayesian analysis. Participants, all of whom initially had little or no experience with Bayesian methods, expressed greater willingness and confidence to apply Bayesian approaches in the future after using the system. They emphasized how the visual and interactive nature of the interface in the observable space made prior elicitation feel approachable and actionable. This is in stark contrast to the baseline system focused on the parameter-space that participants found to be abstract and theoretical. Several participants noted that without tools like PriorWeaver, Bayesian methods would have remained confined to textbooks and theoretical discussions.

9 LIMITATIONS AND FUTURE WORK

Our longer term goal is to make Bayesian analysis more approachable for analysts of various backgrounds. Towards this goal, this work has three limitations that offer opportunities for future work.

Explore alternative methods for deriving statistical priors. Using the dataset that analysts construct through the interactive visualizations, PriorWeaver repeatedly samples and fits a previously specified statistical model. We developed this method based on the definition of a prior as the set of "reasonable" possible values that a coefficient could take on, based on the analyst's beliefs. Future work could explore additional ways to use the constructed dataset, histogram, or relationships expressed visually to derive statistical priors.

Moreover, only data points that have been connected across all variables are included as part of the dataset used to derive statistical priors. It can be tedious for analysts to connect data every variable. Also, it may be especially challenging when analysts do not know how some subset of variables relate. Alternative approaches for deriving statistical priors may be able to take advantage of more data points that are not connected. One possibility may be to impute missing values before deriving statistical priors.

Give analysts feedback on their externalized beliefs. PriorWeaver focuses on faithfully representing analysts' expressed beliefs. It does not assess the quality or validity of expressed beliefs. We focused on faithful representation because we suspected, and our user study confirmed, that analysts are often unsure whether their specifications in current prior elicitation tools reflect their intent. We found that PriorWeaver does indeed bridge the gulfs of execution and evaluation [48] more effectively than existing tools.

Future work should explore ways to assess the veracity of expressed beliefs. For instance, an LLM could check to what degree the assumptions analysts express are likely in the real-world and suggest changes. Similarly, given that analysts in our lab study expressed gaps in their domain knowledge, perhaps future systems could allow analysts to specify weights for the beliefs they hold strongly and delegate the other unknown areas to the system to fill in. These ideas could build up to intelligent "linters" for Bayesian priors and analysis.

Support a wider range of statistical models. Here, we scoped the statistical models supported in PriorWeaver to generalized linear models (without mixed effects) involving only continuous variables. Now that we have gathered evidence that viewing prior elicitation as an iterative dataset construction process is promising, we are excited to test the bounds of this approach. Future work should explore ways to support categorical variables and more complex statistical models. Categorical variables will require use of interactive visualizations, such as the Sankey diagram, to externalize frequency-based beliefs. Also, more complex statistical models raise new opportunities and challenges for eliciting and deriving more complex priors, such as joint prior distributions.

Moreover, iteration on both the priors and statistical models are often necessary in Bayesian analysis [15]. While PriorWeaver supports refining priors through prior predictive checks, there is no support for iterating on the statistical models themselves. Future work should explore ways to support iterations on both.

10 CONCLUSION

PriorWeaver transforms prior elicitation into an iterative process of constructing a dataset that represents the domain assumptions of analysts. Through interactive visualizations, PriorWeaver enables analysts to express their assumptions about observable variables, such as their distributions and relationships with other variables. These interactions construct an underlying dataset that PriorWeaver uses to derive priors for a statistical model. In a controlled lab study with Bayesian novices, we found that PriorWeaver lowered barriers to prior elicitation by helping participants externalize their knowledge and refine priors through actionable predictive feedback. Compared to trial-and-error approaches in the baseline, PriorWeaver gave participants greater control, clarity, and confidence, leading to priors that better aligned with their expectations. Also, PriorWeaver made Bayesian analysis feel approachable and increased participants' willingness to use it in the future. These results suggest that interactive dataset construction is a promising step toward wider adoption of Bayesian analysis methods.

11 AVAILABILITY

Source code and additional information are available at https://github.com/ucla-cdl/prior-weaver.

REFERENCES

[1] [n. d.]. Student Performance Factors Dataset. https://www.kaggle.com/datasets/lainguyn123/student-performance-factors

[2] James Berger. 2006. The case for objective Bayesian analysis. (2006).

[3] Florence Bockting, Stefan T Radev, and Paul-Christian Bürkner. 2024. Simulation-based prior knowledge elicitation for parametric Bayesian models. *Scientific Reports* 14, 1 (2024), 17330.

[4] Christopher J Casement and David J Kahle. 2018. Graphical prior elicitation in univariate models. *Communications in Statistics-Simulation and Computation* 47, 10 (2018), 2906–2924.

[5] 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*. 1–14.

[6] Zoltan Dienes. 2011. Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science* 6, 3 (2011), 274–290.

[7] Berkeley Dietvorst, Joseph Simmons, and Cade Massey. 2014. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *SSRN Electronic Journal* (01 2014). https://doi.org/10.2139/ssrn.2466040

[8] Berkeley Dietvorst, Joseph Simmons, and Cade Massey. 2018. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* 64 (03 2018), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

[9] Jonah Gabry, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. Visualization in Bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society* 182, 2 (2019), 389–402.

[10] Paul H Garthwaite, Joseph B Kadane, and Anthony O'Hagan. 2005. Statistical methods for eliciting probability distributions. *Journal of the American statistical Association* 100, 470 (2005), 680–701.

[11] Andrew Gelman and Christian Hennig. 2017. Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society Series A: Statistics in Society* 180, 4 (2017), 967–1033.

[12] Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* 2 (2008), 1360–1383. https://api.semanticscholar.org/CorpusID:1592811

[13] Andrew Gelman and Cosma Rohilla Shalizi. 2013. Philosophy and the practice of Bayesian statistics. *Brit. J. Math. Statist. Psych.* 66, 1 (2013), 8–38.

[14] Andrew Gelman, Daniel P. Simpson, and Michael Betancourt. 2017. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy* 19 (2017), 555. https://api.semanticscholar.org/CorpusID:36516383

[15] Andrew Gelman, Aki Vehtari, Daniel Simpson, Charles C. Margossian, Bob Carpenter, Yuling Yao, Lauren Kennedy, Jonah Gabry, Paul-Christian Bürkner, and Martin Modrák. 2020. Bayesian Workflow. arXiv:2011.01808 [stat.ME] https://arxiv.org/abs/2011.01808

[16] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

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

[18] Daniel G. Goldstein and David Rothschild. 2014. Lay understanding of probability distributions. *Judgment and Decision Making* 9, 1 (2014), 1–14. https://doi.org/10.1017/S1930297500004940

[19] Daniel G Goldstein and David Rothschild. 2014. Lay understanding of probability distributions. *Judgment and Decision making* 9, 1 (2014), 1–14.

[20] John Paul Gosling. 2018. *SHELF: The sheffield elicitation framework*. 61–93. https://doi.org/10.1007/978-3-319-65052-4_4

[21] Marcelo Hartmann, Georgi Agiashvili, Paul Bürkner, and Arto Klami. [n. d.]. Flexible Prior Elicitation via the Prior Predictive Distribution. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)* (2020-08-27). PMLR, 1129–1138. https://proceedings.mlr.press/v124/hartmann20a.html

[22] Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831* (2020).

[23] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.

[24] George S Howard, Scott E Maxwell, and Kevin J Fleming. 2016. The proof of the pudding: an illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. (2016).

[25] Jessica Hullman. 2019. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 130–139.

[26] Jessica Hullman, Matthew Kay, Yea-Seul Kim, and Samana Shrestha. 2017. Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 446–456.

[27] Alejandro Icazatti, Oriol Abril-Pla, Arto Klami, and Osvaldo A Martin. 2023. PreliZ: A tool-box for prior elicitation. *Journal of Open Source Software* 8, 89 (Sept. 2023), 5499. https://doi.org/10.21105/joss.05499

[28] Geoffrey Jones and Wesley O Johnson. 2014. Prior elicitation: Interactive spreadsheet graphics with sliders can be fun, and informative. *The American Statistician* 68, 1 (2014), 42–51.

[29] Eunice Jun, Melissa Birchfield, Nicole De Moura, Jeffrey Heer, and Rene Just. 2022. Hypothesis formalization: Empirical findings, software limitations, and design implications. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 1 (2022), 1–28.

[30] Eunice Jun, Edward Misback, Jeffrey Heer, and René Just. 2024. rTisane: Externalizing conceptual models for data analysis prompts reconsideration of domain assumptions and facilitates statistical modeling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[31] Joseph Kadane and Lara J Wolfson. 1998. Experiences in elicitation. *Journal of the Royal Statistical Society Series D: The Statistician* 47, 1 (1998), 3–19.

[32] Maurits Kaptein and Judy Robertson. 2012. Rethinking statistical analysis methods for CHI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1105–1114.

[33] Alireza Karduni, Douglas Markant, Ryan Wesslen, and Wenwen Dou. 2020. A bayesian cognition approach for belief updating of correlation judgement through uncertainty visualizations. *IEEE Transactions on Visualization and Computer Graphics* 27, 2 (2020), 978–988.

[34] 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*. 5092–5103.

[35] Matthew Kay, Gregory L Nelson, and Eric B Hekler. 2016. Researcher-centered design of statistics: Why Bayesian statistics better fit the culture and incentives of HCI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4521–4532.

[36] Seyed Vala Khorasani. [n. d.]. Gym Members Exercise Dataset. https://www.kaggle.com/datasets/valakhorasani/gym-members-exercise-dataset/data

[37] 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* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 1375–1386. https://doi.org/10.1145/3025453.3025592

[38] Yea-Seul Kim, Logan A. Walls, Peter Krafft, and Jessica Hullman. 2019. A Bayesian Cognition Approach to Improve Data Visualization *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300912

[39] David Klahr and Kevin Dunbar. 1988. Dual space search during scientific reasoning. *Cognitive science* 12, 1 (1988), 1–48.

[40] 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 *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 18, 12 pages. https://doi.org/10.1145/3411764.3445798

[41] Ratanond Koonchanok, Gauri Yatindra Tawde, Gokul Ragunandhan Narayanasamy, Shalmali Walimbe, and Khairi Reda. 2023. Visual Belief Elicitation Reduces the Incidence of False Discovery. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 466, 17 pages. https://doi.org/10.1145/3544548.3580808

[42] JK Kruschke. 2010. Bayesian data analysis. WIREs Cognitive Science, 1 (5), 658–676.

[43] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*. 126–137.

[44] Shambhavi Mahajan, Bonnie Chen, Alireza Karduni, Yea-Seul Kim, and Emily Wall. 2022. Vibe: A design space for visual belief elicitation in data journalism. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 477–488.

[45] Andrew A Manderson and Robert JB Goudie. 2023. Translating predictive distributions into informative priors. *arXiv preprint arXiv:2303.08528* (2023).

[46] Petrus Mikkola, Osvaldo A Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, et al. 2024. Prior knowledge elicitation: The past, present, and future. *Bayesian Analysis* 19, 4 (2024), 1129–1161.

[47] David E. Morris, Jeremy E. Oakley, and John A. Crowe. 2014. A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software* 52 (2014), 1–4. https://doi.org/10.1016/j.envsoft.2013.10.010

[48] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[49] Jeremy E Oakley, Alireza Daneshkhah, and Anthony O'Hagan. 2010. Nonparametric prior elicitation using the Roulette method. *School of Mathematics and Statistics, University of Sheffeld, UK* (2010).

[50] Anthony O'Hagan. [n. d.]. Expert Knowledge Elicitation: Subjective but Scientific. 73 ([n. d.]), 69–81. Issue sup1. https://doi.org/10.1080/00031305.2018.1518265

[51] 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. (2006).

[52] Chanda Phelan, Jessica Hullman, Matthew Kay, and Paul Resnick. 2019. Some Prior(s) Experience Necessary: Templates for Getting Started With Bayesian Analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300709

[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[54] 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* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376377

[55] Kert Viele, Scott Berry, Beat Neuenschwander, Billy Amzal, Fang Chen, Nathan Enas, Brian Hobbs, Joseph G Ibrahim, Nelson Kinnersley, Stacy Lindborg, et al. 2014. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceutical statistics* 13, 1 (2014), 41–54.

[56] Hilde J. P. Weerts, Werner van Ipenburg, and Mykola Pechenizkiy. 2019. A Human-Grounded Evaluation of SHAP for Alert Processing. arXiv:1907.03324 [cs.LG] https://arxiv.org/abs/1907.03324

[57] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

APPENDIX

A BASELINE INTERFACE

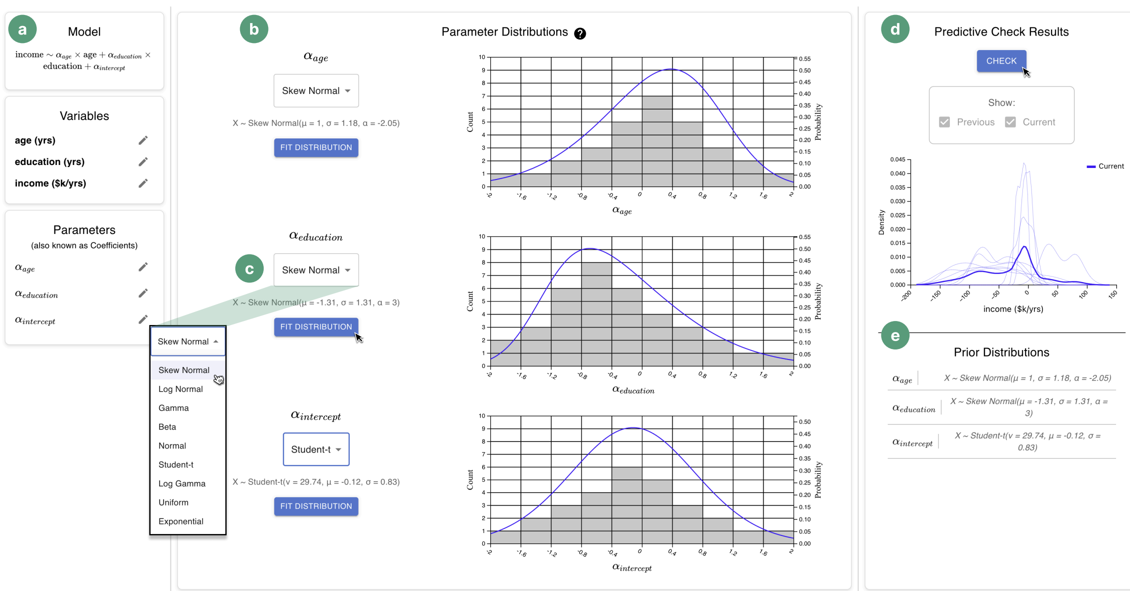

Fig. 9. User interface of the parameter space baseline: (a) an information panel displaying the model formula, variables, and parameters; (b) histograms where users can click and sketch parameter distributions; (c) after sketching, users can click FIT DISTRIBUTION to obtain candidate continuous distributions matching their sketch and select the desired one; (d) users can click CHECK to receive feedback on expressed knowledge through prior predictive checking; and (e) the final prior distributions selected by users.

B DETAILED MEASUREMENTS

Table 3 shows our detailed metrics and questions used in the user study.

Table 3. Measurements used in our user study. For the survey items, a 5-point Likert scale was used, with 1 indicating “Strongly disagree” and 5 indicating “Strongly agree”.

Data Source	Detailed Metric/Question
System log	Metrics of interaction patterns (e.g., number of link uses, generate uses, histogram edits)
	Number of iterations; predictive check results between iterations
Survey	[*Clear Expression*]: "I can clearly express my knowledge using this system."
	[*Visualization Alignment w/ Knowledge*]: "I think the visualization results align with my knowledge."
	[*Comfortable*]: "I feel comfortable expressing my knowledge using this system."
	[*Ease of Expression*]: "This system makes it easy for me to express my knowledge."
	[*Translation Accuracy*"]: I believe my expressed knowledge is accurately translated into prior distributions."
	[*Final Prior Alignment w/ Knowledge*]: "The final prior distributions I choose are aligned with my knowledge."
	[*Understanding of Final Prior*"]: I understand the meaning of the finally elicited priors."
	[*Becoming Appropriate*]: “I think my expressed knowledge became more and more appropriate through the iterations.”
Survey (System Usability Scale)	[*Helpfulness*]: "Overall, I think this system is helpful in supporting prior elicitation."
	[*Complexity*]: "I found the system unnecessarily complex."
	[*Easy of Use*]: "I thought the system was easy to use."
	[*Technical Support Needs*]: "I would need support of a technical person."
	[Function Integration]: "Functions are well integrated."
	[*Quick to Learn*]: "I would imagine that most people would learn to use this system very quickly."
	[*Confidence in Use*]: "I feel very confident using this system."
	[*Cumbersome to Use*]: "I found the system cumbersome to use."
	[*Learning Curve*]: "I needed to learn a lot before getting started."
	[*Future Use*]: "I am inclined to use Bayesian analysis in my future experiments and analyses if I have access to this system."
Interview	How do you express your knowledge using these visualizations?
	How do you decide that your expressed knowledge needs refinement, and what strategies do you use to do so?
	How would you compare your experiences with eliciting priors in the parameter space versus the observable space?
	Which of the two tools would make you more inclined to use Bayesian analysis in the future? Why?

C DETAILED EVALUATION RESULTS

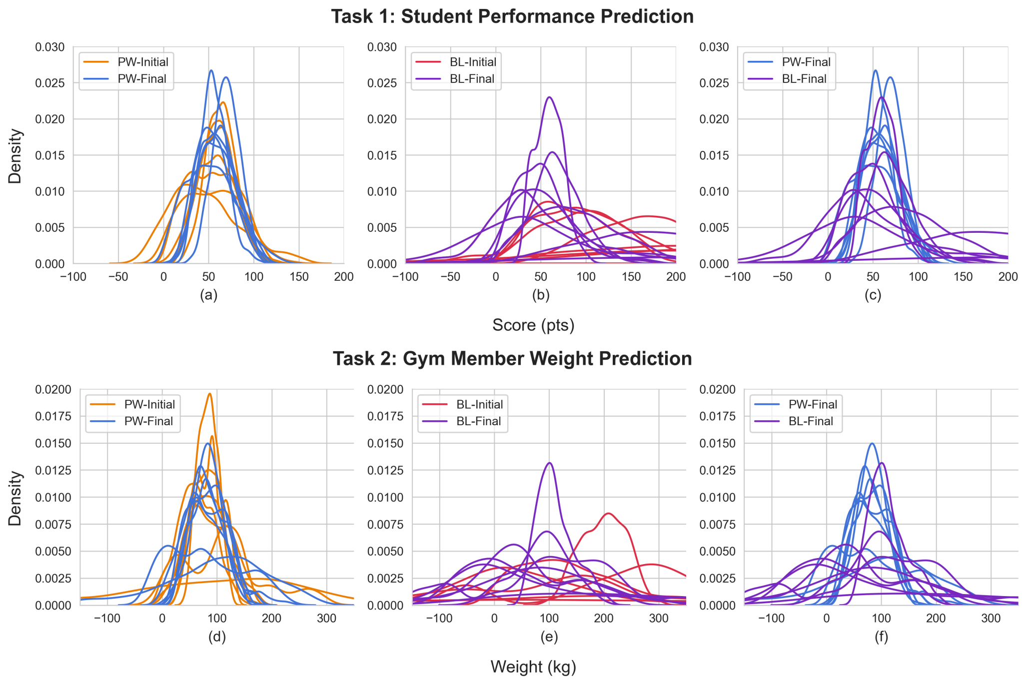

Fig. 10. Predictive distributions of all participants across two tasks. *PW* denotes PriorWeaver and *BL* denotes the baseline. Figures (a) and (d) compare initial and final priors in PriorWeaver. Figures (b) and (e) compare initial and final priors in the baseline. Figures (c) and (f) compare the final priors between PriorWeaver and the baseline.

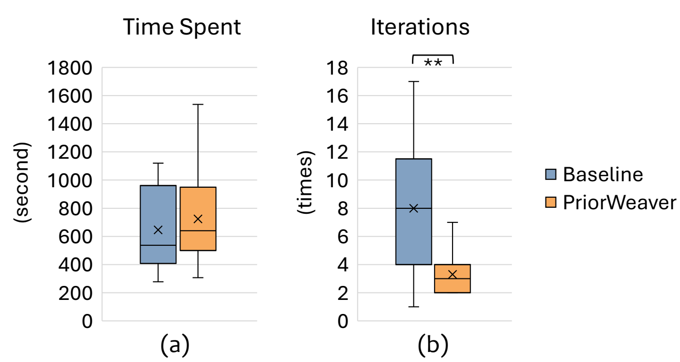

Fig. 11. Participants' engagement in the prior elicitation process with the two systems: (a) time spent across the process and (b) number of refinement iterations. These results suggest that although participants spent similar amounts of time in both conditions, the additional iterations in the baseline condition reflected trial-and-error adjustments rather than purposeful refinement of priors. (*: p < 0.05; **: p < 0.01; ***: p < 0.001).

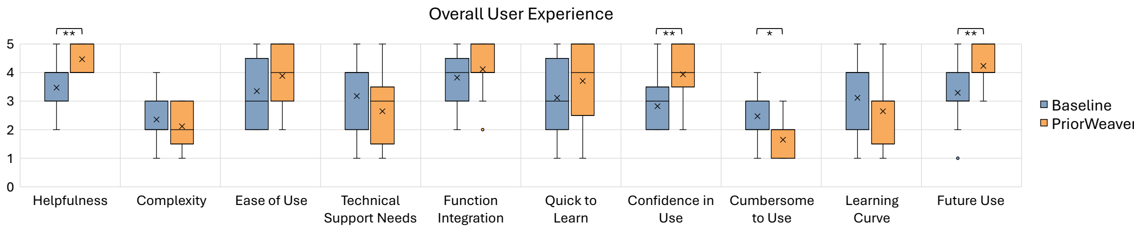

Fig. 12. Survey results from the System Usability Scale (SUS). Participants' overall experience with the two systems indicated that they found PriorWeaver more helpful, felt more confident using it, and were more inclined to use it for future Bayesian analysis. (*: p < 0.05; **: p < 0.01; ***: p < 0.001).