Milliways: Taming Multiverses through Principled Evaluation of Data Analysis Paths

Abhraneel Sarma
abhraneel@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

Kyle Hwang
kylehwang2022@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

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

Matthew Kay
mjskay@u.northwestern.edu
Northwestern University
Evanston, Illinois, USA

ABSTRACT

Multiverse analyses involve conducting all combinations of reasonable choices in a data analysis process. A reader of a study containing a multiverse analysis might question—are all the choices included in the multiverse reasonable and equally justifiable? How much do results vary if we make different choices in the analysis process? In this work, we identify principles for validating the composition of, and interpreting the uncertainty in, the results of a multiverse analysis. We present Milliways, a novel interactive visualisation system to support principled evaluation of multiverse analyses. Milliways provides interlinked panels presenting result distributions, individual analysis composition, multiverse code specification, and data summaries. Milliways supports interactions to sort, filter and aggregate results based on the analysis specification to identify decisions in the analysis process to which the results are sensitive. To represent the two qualitatively different types of uncertainty that arise in multiverse analyses—probabilistic uncertainty from estimating unknown quantities of interest such as regression coefficients, and possibilistic uncertainty from choices in the data analysis—Milliways uses consonance curves and probability boxes. Through an evaluative study with five users familiar with multiverse analysis, we demonstrate how Milliways can support multiverse analysis tasks, including a principled assessment of the results of a multiverse analysis.

CCS CONCEPTS

• **Human-centered computing** → **Visualization systems and tools**; **Visual analytics**.

KEYWORDS

Multiverse analysis, Statistical analysis, Principled evaluation

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Abhraneel Sarma, Kyle Hwang, Jessica Hullman, and Matthew Kay. 2018. Milliways: Taming Multiverses through Principled Evaluation of Data Analysis Paths. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 15 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The numerous choices that researchers make during data analysis (e.g. how to measure variables? Which method to use for excluding outliers?) introduce uncertainty into the results of scientific research. When these *researcher degrees of freedom* are not well-motivated, the validity of findings is threatened. *Multiverse analysis* [53], and similar statistical procedures [42, 51, 61], have been proposed as a way to counteract the risks posed by undisclosed flexibility in conducting data analysis. In a multiverse analysis, a researcher implements and reports on all possible combinations of alternative data analysis decisions that might be considered reasonable. This allows readers to get a sense of the "fragility or robustness of a claimed effect" to arbitrary choices in the data analysis process [53].

While multiverse analyses are being conducted across a range of disciplines (e.g., [2, 7, 8, 10, 12, 14–17, 40, 44, 55, 56]), correct interpretation of a multiverse is difficult due to their potential size [25] and the complexity of correctly interpreting their uncertainty [26, 53]. In this work, we focus on helping readers of a multiverse analysis (e.g., as published in a research paper) interpret multiverse analyses in a principled way, avoiding common pitfalls in their interpretation. To do so, we first use the literature to establish **two principles for evaluating and interpreting a multiverse analysis:**

1. *Readers of a multiverse analysis should be able to assess whether the decisions in the analysis are all equally justifiable* in their expert opinion, and then examine whether (and why) making different choices regarding certain decisions in the analysis process lead to different outcomes. This will help them avoid misinterpreting large multiverses that swamp meaningful effects with unjustifiable choices [25].
2. *Readers of a multiverse analysis should be able to correctly distinguish between the probabilistic and possibilistic uncertainty inherent in such analyses*. Uncertainty arising from equally-justifiable analysis choices is *possibilistic* [26] (i.e. cannot be considered to be more or less likely based on their

frequency), yet most existing visual representations of multiverse analysis results (e.g. specification curves [51], histograms of *p*-values [53], etc.) emphasise the frequency of outcomes, thereby inviting misleading, *probabilistic* conclusions about the result.

Based on these principles, we built **Milliways, an interactive visualisation interface for principled exploration and sense-making of the results of a multiverse analysis**. Milliways is designed to help users perform multiverse analysis tasks identified in prior work [26]. Milliways visualises the decisions that make up a multiverse (commonly depicted as directed acyclic graphs) using a matrix-like tabular representation where each row corresponds to one universe (a unique analysis specification in the multiverse). Leveraging this linearised layout, Milliways provides complete distributional information for results from each universe using representations that distinguish between probabilistic and possibilistic uncertainty. In addition, Milliways presents users with relevant contextual information to understand the analysis, such as the raw data and analysis code. This allows users to assess whether all analysis paths seem equally justifiable.

We evaluate the design and usability of Milliways through a user study with five researchers who have experience with applied statistics and are familiar with multiverse analyses. Our evaluation reveals that Milliways enables participants to perform multiverse analysis tasks identified in prior work [26] and supports a principled approach for interrogating the validity of the results.

2 PRELIMINARIES

2.1 Multiverse Analysis

There are multiple ways to analyse a dataset, many of which could be considered equally reasonable. Yet, in most scientific studies only a singular analysis is reported, and other equally justifiable analysis paths which were considered or explored are not reported. This raises a pertinent concern for a scientific result—what were the results of other, equally justifiable analysis paths that were not reported? Multiverse analysis [53] (and other related statistical procedures, such as specification curve analysis [51] or vibration of effects analysis [42]) aims to address this methodological concern. By making explicit all of the possible decisions involved in data construction and model building, such approaches provide increased transparency as well as greater understanding of the sensitivity of outcomes to arbitrary, yet defensible, decisions that researchers might make during an analysis.

2.2 Example Scenario

We refer to the author of a multiverse analysis as the *analyst*, and the individual using Milliways as the *user* or the *reader*. In describing a multiverse analysis, we adopt terminology from prior work [18, 26, 49]: a multiverse analysis consists of decisions or *parameters*—a point in the analysis where an analyst must decide between two or more reasonable alternatives for performing an analysis step. Each parameter is characterised by one or more choices or *options*—the possible alternatives that the analyst has to choose from. A *universe*, also referred to as a *specification*, is one single analysis from the multiple analyses that make up a multiverse, obtained from a unique combination of analysis *options*.

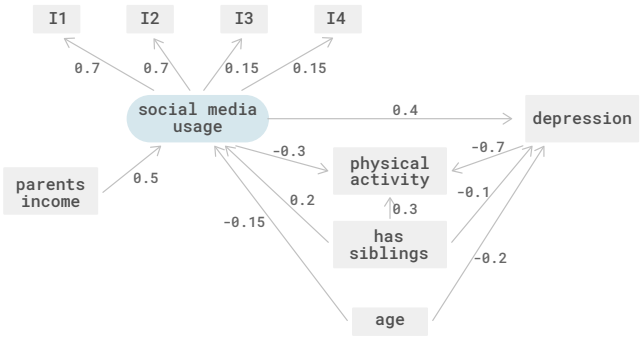

Figure 1: Causal model of a hypothetical study of the effect of social media usage on depression. Observed variables are represented by gray rectangles and unobserved latent variables are represented using blue rounded rectangles. Arrows indicate the direction of the effect, and numbers indicate the correlation coefficient between the two variables.

To make our discussion throughout the paper more concrete, we use a hypothetical multiverse analysis of a fictional study investigating the effects of social media usage on depression among adolescents as a running example. Note that this example is *not* meant to be a demonstration of how a multiverse analysis should be conducted, but rather an example of how multiverse analyses are conducted as seen in the literature (e.g., [51]). See `supplement ▸ dataset` for the code used to generate the data and implement this analysis.

Dataset. The data for this example is intended to measure the relationship between social media usage and depression in teenagers. We generated the data based on the causal model described in the directed acyclic graph in Figure 1, following closely the data generation process used by [25], though their data was generated for a different domain. In our hypothetical dataset, `depression` was measured using a five point Likert item. Four supposed indicators of social media usage were collected: hours spent on social media apps per week (`I1`), self-reported rating of social media usage on a 5-point Likert (`I2`), number of posts updated per week (`I3`), and total number of friends/followers on social media (`I4`). In addition, four other potential factors which may influence the relationship between social media usage and depression (covariates) were collected: `age`, `parent income`, `activity` (hours of physical activity per week), and `siblings` (binary variable indicating whether the participant has one or more siblings).

Specifying the Multiverse. We implement a multiverse analysis which includes decisions related to outlier exclusion, choice of predictor, and covariate selection. Specifically, we define four alternatives to **outlier exclusion**: no exclusion (i.e. analysing all observations), excluding observations using 2.5 SD from the mean, 3.5 SD from the mean, or first and third quartiles $\pm$ 1.5 times the interquartile range as the cutoffs (Tukey's fences). We consider six alternatives for choice of predictor (**sm predictor**): the indicators for social media usage (`I1`, `I2`, `I3` and `I4`) individually, a composite of `I1` and `I2` as these two variables had the highest correlation with the dependent variable (`depression`), and a composite of all four indicators. Three covariates were considered (**age**, **activity** and **siblings**)

representing three decisions with two alternatives each: including or not including the covariate. The result, in each universe, is obtained from a linear regression with `depression` as the dependent variable, and the measure of social media usage as the primary predictor. The coefficient for social media usage was used to draw conclusions from this analysis.

3 RELATED WORK

3.1 Multiverse Analysis Tasks

The primary goal of conducting a multiverse analysis is to assess outcome sensitivity and robustness—whether a particular result is sensitive to or stable across arbitrary decisions involved in the data analysis process [26]. Hall et al. [26] provide a taxonomy of tasks that analysts perform to achieve this goal:

[T1] Composition tasks are concerned with interpreting the composition of a multiverse, which involves understanding the steps in the data analysis process, the decision points, and the choices for each decision point.

[T2] Outcome tasks involve assessing the overall variability or stability of the primary outcome across the set of specifications.

[T3] Connect tasks are concerned with identifying the sources of sensitivity, if any, in the outcome of a multiverse analysis. More precisely, connect tasks can include determining which analytical choices or combinations of analytical choices lead outcome values to differ across analysis specifications, how often, and by how much. They also involve connecting specific outcome values to decisions that produced them.

[T4] Validation tasks focus on assessing the validity of the specified multiverse analysis, either quantitatively (e.g. using model fit metrics) or qualitatively (e.g. evaluating each analysis path in the context of the overall research question and existing domain knowledge to ensure they are valid).

In addition, Hall et al. [26] underline the need for visualisations to better support multiverse interpretation. As such we define an additional task:

[T5] Interpretation tasks involve drawing conclusions, based on the range of possible outcome values observed in the multiverse, regarding the overall effect being studied.

Multiverse analysis tasks and visualisations can be considered a specific case of the broader class of parameter space analysis problems [5, 50], with many of the tasks identified by Hall et al. [26] sharing similarities with analysis tasks defined for visual parameter space analysis (e.g., partitioning, outliers, uncertainty, and sensitivity).

3.2 Principled Evaluation of a Multiverse

Principled validation of the decisions in the multiverse. Del Giudice and Gangestad [25] lay out a framework to evaluate analytical decisions in a multiverse analysis, describing three types of decisions scenarios: *Type E* (equivalent), *Type N* (non-equivalent) and *Type U* (uncertain). In Type E decisions, every choice is equally justifiable, based on existing theoretical knowledge and understanding, and thus can be considered equivalent. For instance, in our hypothetical multiverse, **outlier exclusion** may be considered a Type E decision, as a priori, there is no reason to not consider each choice as equivalent. In Type N decisions, a subset of the alternatives may be considered to be more reasonable or justifiable than the others and thus all choices cannot be considered equivalent. In our example, the decision **sm predictor** includes two choices which have lower validity (`I3` and `I4`), which can be considered as Type N. In Type U decisions, the current information available to the analyst does not allow them to make the determination if all the alternative choices are equivalent or non-equivalent, even if there may be reasons to suspect non-equivalence. The decisions to include or not include covariates (**age**, **activity** and **siblings**) in a linear regression model indicates that the analyst is unsure about competing causal models of the data—the variables may be colliders, mediators, or completely superfluous to the causal model. Ideally, a multiverse should only include Type E decisions. However, in many cases, the challenge lies in the large number of Type U decisions that an analyst faces. Type U decisions often indicate a lack of theoretical understanding. In such scenarios, multiverse analyses can be constructed in an exploratory manner to include Type U decisions. Such analyses can potentially help surface gaps in existing theoretical understanding of a problem, and help direct future research efforts.

A **principled approach to validating the construction of a multiverse requires a reader to incorporate factors such as domain knowledge and statistical expertise** to construct arguments in favor of or against certain choices in the multiverse. While prior work [34] has adopted a metric-based, quantitative approach towards validation, there are drawbacks towards such approaches. Consider a dataset where the outcome distribution has heavier tails—one approach for analysing this data may be to exclude outliers, and use a Gaussian regression; alternatively, one can forego outlier exclusion and use a regression with a Student's t distribution for the outcome variable. While both methods may yield similar model fit metrics, other factors should be considered to determine which approach is more reasonable. For instance, were there systematic issues in data collection which led to an excess number of extreme values? Depending on the answer, we should consider one choice in the multiverse to be more reasonable than the other. However, designing a visualisation tool to fully support validation tasks may not be possible to the extent that a user is able to determine whether every decision is comprised of equivalent choices or non-equivalent choices. Instead, we design Milliways to support a principled validation insofar as the user is able to highlight which decisions require further review and may be deemed as Type U based on the information that is available to them.

Principled interpretation of uncertainty in multiverse results. Interpreting the results of a multiverse analysis requires a distinction between two qualitatively different forms of uncertainty—*probabilistic* and *possibilistic*—that arise in multiverse analyses [26]. *Probabilistic* uncertainty arises in each individual analysis that attempts to estimate a quantity whose true value is unknown using a finite sample, such as through a statistical model. Probabilistic uncertainty provides us with an estimated distribution of a random variable, which we can use to make statements about the likelihood of certain values. In contrast, *possibilistic* uncertainty in the outcome arises due to variation in the decisions involved in the data analysis process i.e. each outcome value in the multiverse is a possible result, and any

particular outcome value cannot be considered to be more or less likely based on their frequency [4, 22, 26].

Hall et al. [26] highlights this tendency to construe the results of multiverse analyses as *probabilistic*—readers (mistakenly) assuming that each analysis in the multiverse is equally likely to be correct and thus conclude that "outcome values that occur more frequently within the multiverse must be more likely to be correct" [26]. This assumption diverges from the original, *possibilistic*, interpretation [53]. Moreover, the belief that all universes are equally probable is an unfounded assumption as different researchers and readers will likely disagree on the validity of certain universes [26, 53]. Instead, according to Steegen et al. [53], if there are many branches in the multiverse and no strong arguments can be made for any of the analysis paths being more or less justifiable, *"the only reasonable conclusion [...] is that there is considerable scientific uncertainty."*

Thus, a **principled interpretation of the multiverse analysis results considers the variation in outcomes as possibilistic, and the uncertainty in each individual outcome as probabilistic**. Yet, most typical representations of uncertainty—histograms, dot plots, density plots etc.—are designed to only express probabilistic uncertainty [26]. In the design of Milliways, we explore representations such as consonance curves [1, 46] and probability boxes [21] which support accurate depiction of probabilistic and possibilistic uncertainty respectively.

3.3 Visualising the multiverse

Static visualisations have been used to summarise the results of a multiverse analysis and communicate the stability or sensitivity of an outcome to decisions in the data analysis process. These static approaches involve showing the distribution of outcomes (e.g. histogram of p-values [53], marginal distributions [61]) and highlighting which decisions or combinations of decisions lead to certain outcomes (e.g. outcome matrix [8, 53], specification curve [51]). However, these representations are generally not scalable, and only support a very limited number of tasks.

Boba [34] is an interactive visualisation tool designed for the analyst who has authored the multiverse analysis to assess and refine their analysis. Boba supports, fully or partially, the outcome (**T2**), connect (**T3**) and validation (**T4**) tasks described in §3.1. However, Boba aggregates the sampling uncertainty from each universe (probabilistic uncertainty) with the uncertainty arising due to alternative choices in the multiverse (possibilistic uncertainty); moreover, Boba only allows users to validate specifications based on certain metrics (e.g. model fit). Milliways adopts a fundamentally different design approach which places greater emphasis on *principled evaluation*, based on domain expertise, and *principled interpretation*, by distinguishing between possibilistic and probabilistic uncertainty, of the results (§3.2). We provide a detailed comparison between Milliways and Boba in §5.2.

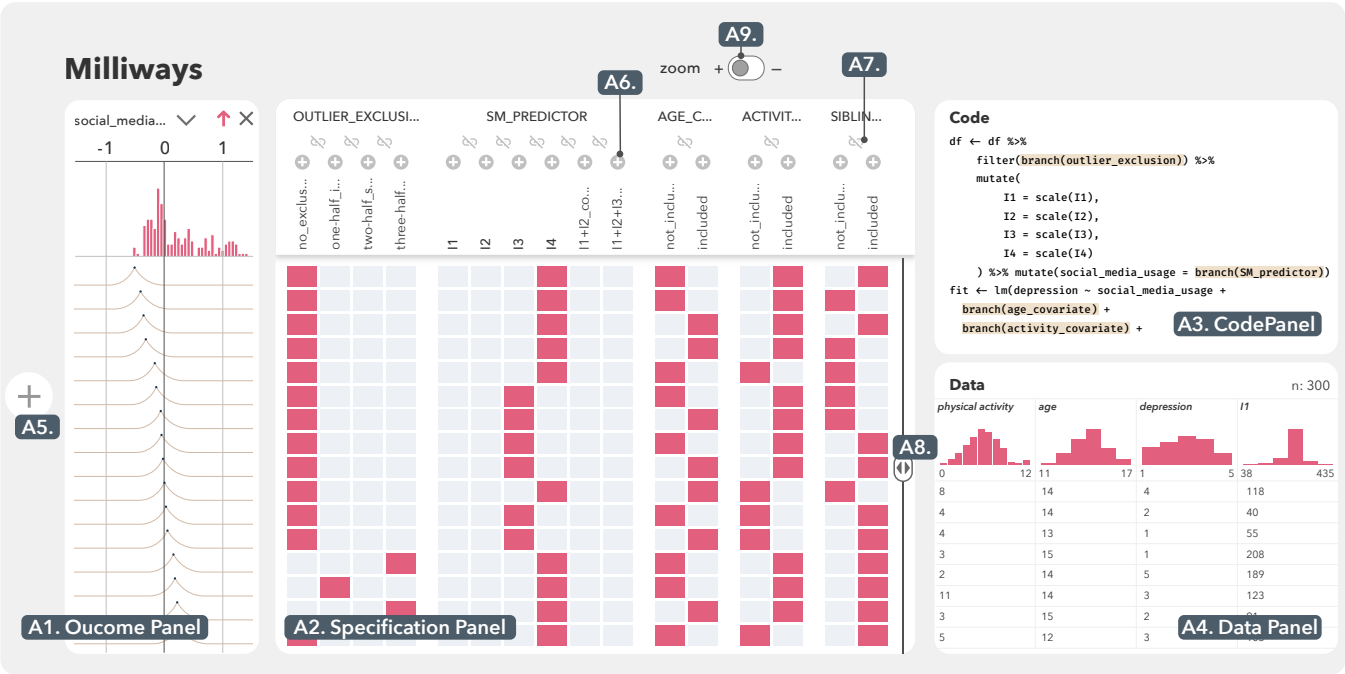

Figure 2: The Milliways interface consists of four panels to support the multiverse analysis tasks and a principled evaluation of the multiverse analysis. The outcome panel A1 presents the results of the multiverse—for each universe in the multiverse, we show the outcome. The specification panel A2 reveals the options for each parameter which the universe is composed of. The code panel A3 presents the code that was used to construct the multiverse (here, the code was from the multiverse R library). The data panel A4 shows the dataset used in the analysis. Other elements include the add button to add additional outcome variables A5, exclude button to remove options of a parameter A6, link button to link (aggregate) the outcomes across two or more options A7, slider for hierarchical sorting based on average effect A8 and the zoom toggle A9.

Exploratory multiverse analysis reports (EMARs) present the results of a multiverse analysis as an interactive report, allowing the reader to use embedded widgets to interactively explore alternative analysis paths. While EMARs provide limited support for other tasks (outcome, connect), they do provide details on how the analysis is conducted for each analysis path, thereby allowing the reader to understand the composition of, and validate the multiverse. We allow EMARs to be integrated into Milliways (§4.5) to support further validation of individual analysis paths.

4 DESIGN AND IMPLEMENTATION OF MILLIWAYS

The Milliways interface and interactive elements are designed to support a principled evaluation of multiverse analyses (§3.2) and the four multiverse analysis tasks (§3.1). The interface consists of four panels—**code** (Figure 2 A3), **data** (Figure 2 A4), **outcome** (Figure 2 A1) and **specification** (Figure 2 A2). We describe the interface design of Milliways in detail below.

4.1 The Code and Data Panels

The **code** and **data** panels surface the steps involved in the construction of the multiverse (what were the analytical decisions that were made by the authors in this analysis? What were the choices identified for each decision?) without having to load and execute parts of the original analysis code in a separate R window. This enables the user to get an overview of how the multiverse analysis was authored (**T1**). These panels are also designed to provide contextual information regarding the analysis which may be useful to users when performing subsequent multiverse tasks (e.g., **T4**).

The **code** panel (Figure 2 A3) shows the formatted R code used to declare multiverse analysis using the multiverse R library. Parameters in the analysis, as declared by the analysts, are highlighted (for e.g., `branch(<SM_predictor>)`). We fold the options declared for each decision for readability and conciseness. Clicking on these highlighted decisions expands to reveal the options that were declared for each decision (Figure 3). Individual universes in the multiverse are constructed by substituting the branch statement for each decision node with the code for each option.

The **data** panel (Figure 2 A4) allows the reader to inspect whether an analysis step makes sense given the properties of the data. It presents both an overview of the dataset used for the analysis using

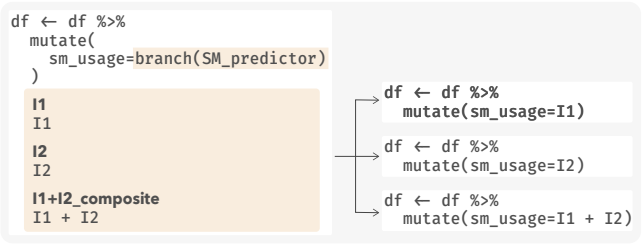

Figure 3: The code panel with one parameter expanded to show the corresponding options. The code on the left declares three options for the parameter `SM_predictor`, which represents three distinct analyses in the multiverse. The code for the resultant, distinct analyses are shown on the right.

histograms for each variable in the dataset, and the actual values as a table. We allow users to sort the dataset by a specific column. The design of the **data** panel was inspired by Kaggle, the online dataset repository, which allows users to get a high-level summary of the dataset, as well as relevant metadata.

4.2 The Outcome Panel

The **outcome** panel (Figure 2 A1) is designed to help analysts assess the range and distribution of outcomes among alternative specifications of the multiverse (**T2**). It visualises any result or variable of interest to answer the research question, such as a regression coefficient, estimate of effect size etc., from each universe of a multiverse analysis. In our example (§2.2), this is the regression coefficient of the predictor for social media usage. Typically, the results of a statistical analysis includes a mean or median point estimate, and the associated uncertainty from the estimation process.

Milliways provides an overview of the distribution of the point estimates as a histogram. Underneath this histogram, we show visualise the result from each universe in the multiverse analysis using consonance curves (Figure 4A). Consonance curves are a compact representation for depicting distributional information, and can be adapted to communicate both probabilistic uncertainty, and—unlike other representations such as densities—possibilistic uncertainty (see §4.4). Additionally, consonance curves allow us to accommodate estimates from both Bayesian and frequentist data analyses. In frequentist analyses, we can use consonance curves (also known as confidence distributions, *p*-curves or compatibility curves [1, 3, 45, 46, 52, 54, 60]) as a distribution function to estimate any parameter of interest [60]. The consonance curve represents the two-sided *p*-value function—the probability of the data given that the null hypothesis is x[1] (Figure 4A). Horizontal slices through this curve provide us the corresponding confidence interval, thereby supporting an interval-based interpretation. Alternatively, Bayesian analyses provide a posterior distribution for any parameter of interest, which can be used to obtain a Cumulative Distribution Function (CDF). As shown in Figure 4B, the consonance curve can be considered a variant of the CDF, $F\left(x\right)$. For Bayesian analyses, horizontal slices through the consonance curve represent the corresponding quantile credible intervals [29].

In cases where there are multiple outcome variables, on initialisation, Milliways shows only one outcome variable. A dropdown menu, present at the top of the outcome panel (Figure 2 A1), can be used to change which outcome variable (e.g., regression coefficients for other predictors in the linear model, effect size estimates etc.) is visualised. Additional outcome variables can be visualised at the same time using the **add** button (Figure 2 A5). This allows the reader to inspect and compare the impact of decisions across multiple outcomes of interest. Outcomes are initially sorted in ascending order based on the median point estimate. Users can change the sorting order or look at an "unsorted" view of the outcomes.

4.3 The Specification panel

The **outcome** panel is interlinked with the **specification** panel (Figure 2 A2), which depicts the composition of individual analyses

[1] in most cases, the p-value being referred to is the null p-value corresponding to the null hypothesis of no effect ($x = 0$); however, the p-values is defined for any value of x.

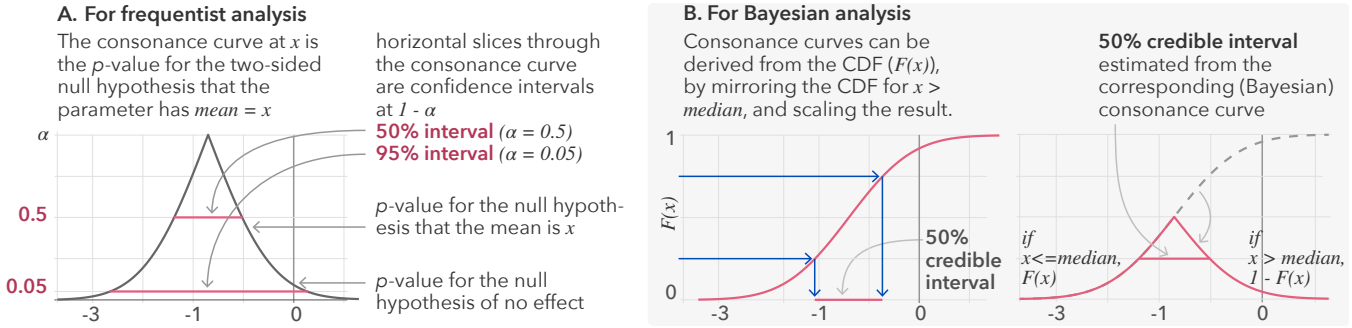

Figure 4: Consonance curves are used to represent the (probabilistic) uncertainty in estimates from each universe. Consonance curves have both frequentist (A) and Bayesian (B) interpretations. We show how they may be derived under both interpretations. Consonance curves allow users to identify interval information at various levels.

in the multiverse. The "tree of analysis" [39] is a common metaphor for depicting a multiverse analysis—an analysis proceeds from top to bottom with each decision represented as a node in the tree, and each option as branches from a node; every node at the same level represents the same conceptual decision (e.g., [18, 62]). Directed Acyclic Graph (DAG) or similar variants [33, 34] are used to visually represent such "trees of analyses."

Our design takes inspiration from the static specification curve visualisations of multiverses [51] and prior attempts at visualising graph-like structures in a linearised format [31, 36, 38]. We adopt a matrix-based representation to visualise multiverse composition. Matrix-like representations have three advantages over tree-like representations: (1) it preserves a one-to-one mapping between universe specification and outcome, (2) it is highly scalable and flexible [9, 59], and (3) many of the interactions (§4.4) to perform the tasks (**T1**-**T5**) are more easily supported within a tabular representation.

Each row of the matrix represents the specification of one universe. The columns of the matrix represent every option declared in the analysis, grouped by parameters. Filled-in cells indicate which option of a parameter was used in the construction of a particular universe. For example, consider the first specification (row) in Figure 2 A2 ( ) which indicates the universe is composed of the first option for the parameters outlier exclusion and age, the fourth option for the parameter sm predictor, and the second option for the parameters activity and sibling.

As the size of a multiverse analysis grows larger, it is not possible to show the results for every universe in this layout in a single view. Milliways supports a compressed layout through the zoom button (Figure 2 A9). The interface defaults to the zoomed-in view; the zoomed-out view collapses the height of the cells in the **specification** panel and only presents the median point estimates, hiding the distributional information in the **outcome** panel (Figure 7 C3).

The **specification** panel in conjunction with the **outcome** panel allows users to determine whether an outcome is sensitive to certain choices in the analysis (**T3**). If an outcome is sensitive to choices, the values of the outcome, sorted in either ascending or descending order, will be correlated to the levels of a decision (i.e. the choices) and the options for a decision would appear as clusters in the specification panel. We discuss this in more detail in our case study (§5.1).

4.4 Interactions

We designed several interactions into Milliways to enable principled evaluation of multiverse analyses. For implementing our interactions, we draw inspiration from prior work such as Bertifier which enables interactions to directly manipulate tabular representations to perform operations on the columns (and rows) of the table [43].

Filtering. When examining a multiverse analysis, a user may consider one (or more) options of a particular decision to be less reasonable or not equally justifiable compared to the other options for that decision (a Type N decision). Users can use the exclude button (Figure 2 A6) to systematically filter out specifications constructed with that option from the
multiverse which are deemed to be unreasonable or unjustifiable, and retain only options which can be considered either equally justifiable or uncertain. Alternatively, a user might be interested in connecting outcome values to decisions—*what decisions or combination of decisions lead to these specific outcome values?* (**T3**). To support this, Milliways allows brushing over a region in the histogram to filter the view to include only the specifications with outcomes within the brushed region.

Hierarchical Sorting based on average effect. In certain cases, the outcome may be sensitive to certain combinations of options across two or more parameters, rather than just the options of only a single parameter (**T3**). To facilitate users in identifying the effects of combinations of options on the outcome, we provide a *hierachical sorting* mechanism, which is controlled by a slider (Figure 5A).

The slider (Figure 2 A8) allows the user to configure how the outcomes are sorted. If the slider is all the way to the right, universes are sorted by the outcome, without grouping by parameters or options (Figure 5A). If the slider is moved in from the right, universes are sorted based on the average effect for each parameter to the right of the slider. Consider, in Figure 5, the specifications containing the option `not_included` ( ) for the activity parameter. We refer to the mean of this distribution of point estimates as the **average effect** for specifications with option `not_included` (Figure 5C).

If more than one parameter is to the right of the slider (Figure 5F), we use a hierarchical approach to calculate the average effect. The specifications are first sorted based on the average effect of the rightmost parameter. Then, for the next parameter, the specifications are

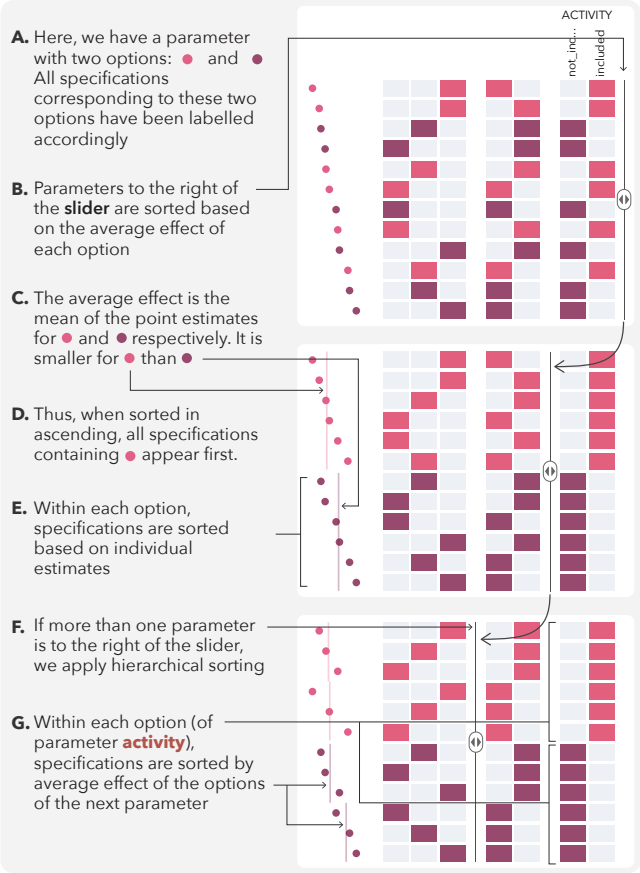

Figure 5: Hierarchical sorting based on the average marginal effect allows users to see the variation in the outcome due to the options of a parameter. It also allows users to inspect the outcomes conditional on the options of parameters

sorted separately, by average effect, for each option (Figure 5G). This continues recursively until we have exhausted all the parameters to the right of the slider. After that, the universes are sorted by individual outcome values. Thus, re-ordering parameters and adjusting the position of the sort slider allows users to progressively investigate the effects of more complex combinations of parameters on the outcome.

Aggregation and Probability Boxes. A user may determine that two or more options of a parameter are equivalent (a Type E decision) or decide they cannot establish non-equivalence between them (a Type U decision). For instance, if a user does not have clear expectations regarding what may be considered atypical values, all the choices of outlier treatment may be considered Type E. Based on this determination, the user may want to understand—*what is the extent of variation in the outcome values across these choices?* The link button (Figure 2 A7) allows users to link two adjacent options of a parameter together which presents the results aggregated across the "linked" specifications.

Visual representations for multiverse analyses should preserve and distinguish between the two qualitative different forms of uncertainty—probabilistic and possibilistic—that arise in multiverse analyses. In prior work [32], the aggregated results across several specifications of a multiverse analyses have been typically represented using superimposed cumulative distribution functions, which are similar to spaghetti plots or ensemble representations of uncertainty. However, such representations may be ill-suited for communicating possibilistic uncertainty as ensemble representations communicate the likelihood of certain events—readers tend to associate the density of the curves with events which are more probable [41].

Instead, to accurately depict possibilistic uncertainty, we consider the outcomes (which are themselves probability distributions) aggregated across specifications to represent probability bounds for the uncertainty in the analysis process. Ferson and Siegrist [21] recommend using p-boxes which are "specified by left and right bounds on the cumulative probability distribution function" (Figure 6) to provide a unified representation of probabilistic and possibilistic uncertainty. As consonance curves are merely variants of CDFs, we use an analogous approach to construct p-boxes for consonance curves. Similar to the consonance curves, horizontal slices through the p-boxes represent the uncertainty in the limits of the uncertainty interval i.e. where the possible upper and lower bounds of the intervals could be (Figure 6B).

Reordering. We support manual column reordering of parameters through a drag and drop interaction. Users can use this reordering interaction to apply the sorting based on **average effect** to different parameters. We also support manual reordering of options within each parameter, which allows users to **aggregate** different combinations of options.

4.5 Integrated EMARs

As described in §3.3, Exploratory Multiverse Analysis Reports (EMARs) are statistical reports which allow readers to explore alternative analysis options through widgets embedded in interactive

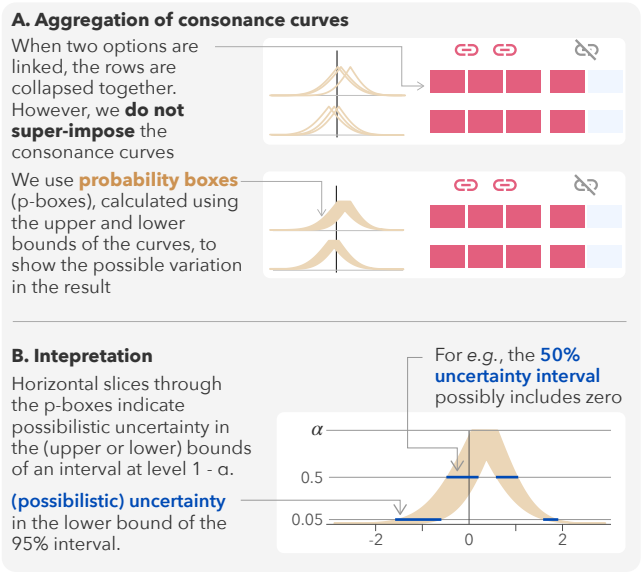

Figure 6: On aggregation, we represent uncertainty in the result using probability boxes

papers. This narrative-rich format can provide additional details to the reader about how and why each alternative analysis path is included. As Milliways only supports visualising outcomes as 1D uncertainty visualisations, EMARs can help overcome some of the challenges in representing multi-dimensional uncertainty [48], as they can support more complex visualisations of results. We support the integration of EMARs, which can be created using the `multiverse` R library, into Milliways to allow the reader to peer into individual analyses and inspect the validity of specific choices in the multiverse. As each row in the **specification** panel corresponds to a distinct analysis, a user can click on any row to bring up the corresponding EMAR in a separate window.

4.6 Implementation and Usage

We implement Milliways as a Svelte.js[2] application that runs in the browser,[3] and use Svelte and D3 [6] to generate the interface elements and visualisations. Our implementation allows Milliways to be compiled into standalone HTML files, making it easier to share and distribute the results of the analysis. We anticipate that authors of multiverse analysis may find it useful to include such interactive visualisations as part of a paper's supplementary materials.

To visualise the results of a multiverse analysis using Milliways, users would need to provide as input the results of a multiverse analysis, the analysis code, the dataset used for the analysis and, optionally, an EMAR. The results file should contain, for each universe, the option name for every parameter and the mean point estimate, and x and y values for the consonance curve. The code file should contain the (formatted) code used to implement the multiverse analysis. The results, analysis code, and dataset provided as input should be in the form of JSON files with a well-defined schema (see `supplement ▶ milliways ▶ README.md`).

To provide an integrated pipeline for analysts to implement multiverse analyses, obtain results, and visualise these results within the same analysis workflow, we developed an R interface for the visualisation system. The R interface facilitates the steps between authoring and executing a multiverse analysis (which is supported by libraries such as `multiverse` [49]) and initialising the visualisation interface for Milliways. The R interface provides functions to allow users to extract results from a multiverse object, as well as other necessary files[4] and save them in the required JSON format. Users can pass these outputs of into a function which initialises the interactive visualisation system on a local server. Multiverse analysis implemented using other authoring tools can also be visualised using Milliways, provided the results are exported using our JSON schema.

5 CASE STUDY

Through a case study, we demonstrate how an analyst might perform a principled exploration and evaluation of a multiverse analysis. Consider a fictional scenario: Alice, an HCI researcher with domain expertise on social media usage among adolescents, was asked to review a scientific article that contains a multiverse analysis (§2.2). We describe how, as a reviewer, Alice can interact with Milliways to understand the composition of the multiverse (**T1**), assess the range and distribution of possible outcome values (**T2**), identify any possible sources of sensitivity (**T3**), interrogate the validity of the multiverse (**T4**), and interpret the results of the multiverse (**T5**). We provide the example as compiled HTML files (`supplement ▶ user-study ▶ task▶ template.html`), and a video demonstrating all the interactions described in the case study (`supplement ▶ video-demo.mp4`).

5.1 Using Milliways to Evaluate and Interpret the Results of a Multiverse Analysis on the Effect of Social Media Usage on Depression

As a reviewer, Alice wants to make sure that the multiverse only consists of equally justifiable (to her) decisions before interpreting the results of the multiverse analysis. She begins her evaluation by importing the results into Milliways:

Composition of the multiverse. Alice first explores the **code** panel (Figure 7 C1) to familiarise herself with the analysis. From reading the code used to construct the analysis, she notices that the multiverse is composed of five decisions related to **outlier exclusion**, **sm predictor**, **age**, **sibling** and **activity**. She expands on the parameter names which reveals the options defined for each parameter. Alice observes that the model coefficient of social media usage when regressed on `depression` is the result of this multiverse analysis.

Determining the Fragility or Robustness of the Result. Alice inspects the **outcome** panel (Figure 7 C2), which displays the regression coefficient for social media usage for each individual analysis in the multiverse. From the histogram at the top, she notices that a range of outcomes—both positive and negative—are possible. This extent of variation in the outcome leads her to ask how the outcome values depend on decisions made in the data analysis process.

Identifying Sources of Sensitivity. To get a complete view of the multiverse, Alice clicks on the *zoom out* toggle (Figure 2 A9). In this view (Figure 7 C3), she discovers that the outcome—the coefficient for social media usage regressed on `depression`—appears to be correlated with the options of the parameter **activity**: the option `not_included` generally leads to smaller values for the coefficient, while the option `included` leads to larger values for the coefficient. This suggests that the outcome is sensitive to the choices of this decision. Although the option names provide some indication to what the decision and the choices entail, Alice refers to the **code** panel and confirms that the choices for this decision involve whether the variable `activity` has been included as a covariate in the regression model or not. Based on her domain knowledge, Alice realises that the variable `activity` is a collider, and hence should not included in the regression model [11, 35]. She excludes the option `included` from the multiverse and continues her inspection.

Identifying Combinations of Options that the Outcome is Sensitive To. Inspecting the histogram, Alice observes that while most of the choices result in positive estimates, negative estimates for the coefficient are still possible (Figure 7 C4). She wants to determine if there are any other sources of sensitivity. She notices that these outcomes are associated with the `no_exclusion` option of the **outlier exclusion** parameter. However, this option alone is not the

[2] https://svelte.dev/
[3] we will add a URL here once the paper is accepted (currently omitted for anonymity)
[4] EMARs can be created using the `multiverse` [49] R library.

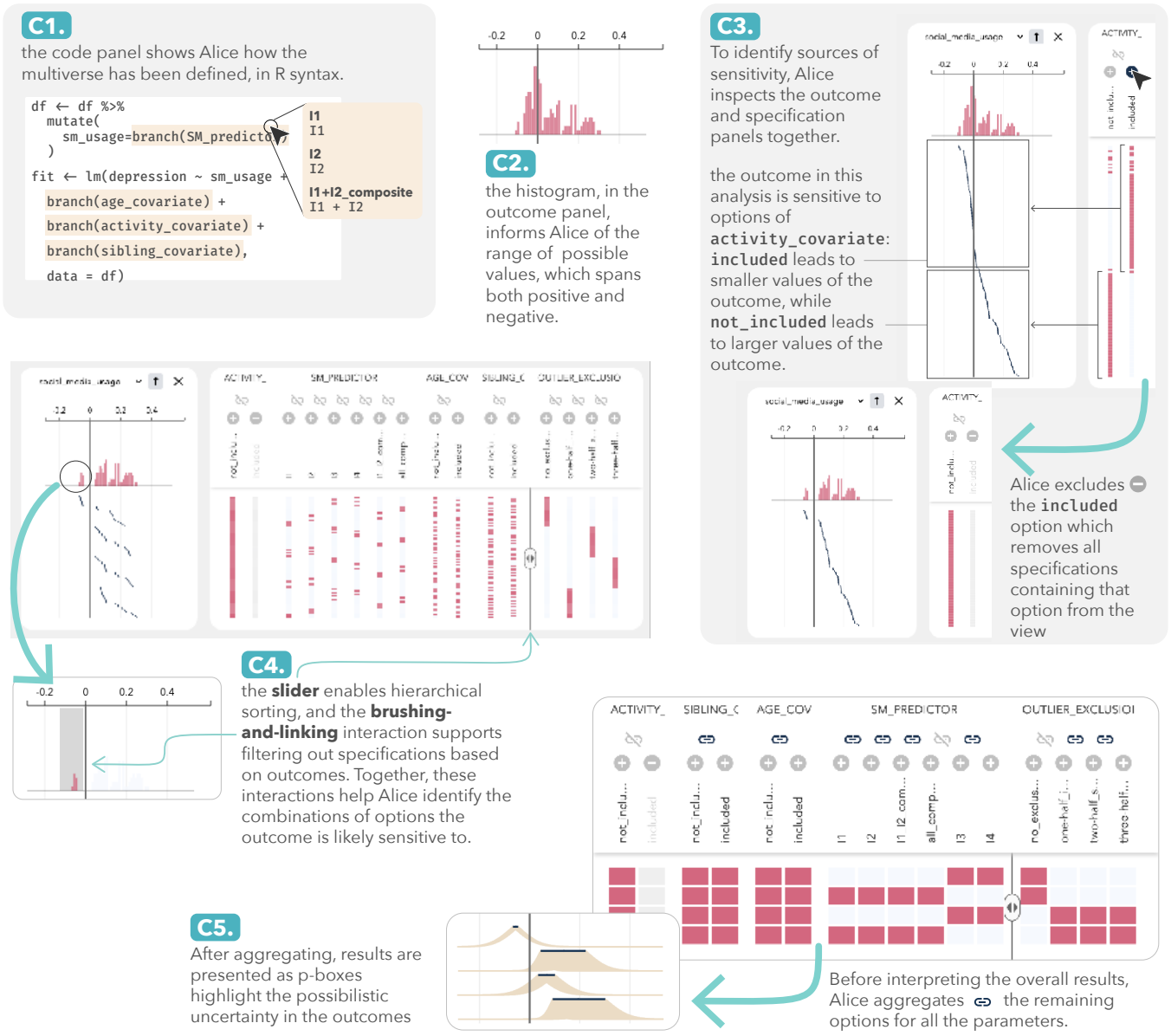

Figure 7: A walkthrough of the steps that Alice performs during her principled evaluation of a multiverse analysis using Milliways which is described in our case study. For a video of this demonstration, see `supplement ▸ video-demo.mp4`

source of the sensitivity in the outcomes, as there exist universes where the `no_exclusion` option leads to positive estimates. To determine which combinations of options the outcome is sensitive to, Alice drags the **outlier exclusion** parameter all the way to the right, and moves the slider to the left of the this parameter. This sorts the outcomes based on the average effect of each option of **outlier exclusion**. This view (Figure 7 C4) suggests to her that these negative estimates may be linked to the options `I3` and `I4` of the **sm predictor** parameter. Alice uses the brushing-and-linking interaction on the histogram and filters to only show coefficients that are negative. This confirms that these negative values correspond to universes where outliers are not excluded and the variables `I3` and `I4` are used as measures of social media usage, suggesting that the outcome may be sensitive to these combinations of choices.

Interrogating the validity of the multiverse. She next takes a look at the **data** panel to inspect histograms of the data variables `I3` and `I4`, which reveals that both of these variables have long tails. Alice knows that extreme values of a predictor may have a large influence on the fit of the regression model as these can be *leverage* points [19]. She feels that if the options `I3` and `I4` are included in the multiverse, some form of outlier exclusion may be desirable. Hence, the choice of options `I3` and `I4` in conjunction with the `no_exclusion` option may not be equally justifiable as the other specifications declared in the multiverse. Based on her evaluation, there are reasons to believe

that some of specifications stemming from the **outlier exclusion** and **sm predictor** decisions are perhaps non-equivalent; however, lacking additional information, Alice decides to treat them as Type U decisions, and interpret the results separately.

Interpreting the results of the multiverse. Alice decides to aggregate the remaining options, which she considers equivalent (Figure 7 C5). Combining possibilistic and probabilistc uncertainty, the resultant view allows her to draw the following interpretations regarding the results of the multiverse: (1) in the set of specifications which result from `I3` and `I4` being used as predictors and outliers are not excluded (`no_exclusion`) (first row in Figure 7 C5), the 95% confidence interval of the estimated outcome *definitely* overlaps zero; (2) in the set of specifications in the second and third rows, the 95% confidence interval *possibly* overlaps zero; and (3) in the final set of specifications, the 95% confidence interval *definitely does not* overlap zero.

Alice's Takeaways from the Principled Evaluation. After her evaluation, Alice found the possible outcome values of the resultant multiverse to be both positive and negative. Conditional on whether certain choices in the analysis are considered equally justifiable or not (including options `I3` and `I4` in conjunction with `no_exclusion`), the conclusions from the multiverse will vary. If these specifications are considered not equally justifiable, the possible outcome values are all positive, indicating effects are in the same direction. However, the possible outcomes have a large range, suggesting that the effect could possibly be quite negligible or quite large, depending on how the data is analysed. Alice's analysis highlights the need for domain knowledge on whether using `I3` and `I4` as predictors and `no_exclusion` of outliers constitute valid analytical decisions.

5.2 Comparison with Boba

We implement the same analysis using Boba [34], an alternate multiverse visualisation tool, and contrast the conclusions drawn using Milliways from those that can be made with Boba (see `supplement ▶ example-boba`). While Boba supports many of the multiverse analysis tasks (§3.1), it is primarily designed for the analyst who has authored the multiverse analysis to assess and refine their analysis. Boba partially supports composition tasks (**T1**) by visualising an overview of the decisions declared in a multiverse analysis using a node-link graph (see Fig. 5a in [34]). However, this graph omits the details of the construction of the multiverse. This may likely be because Boba assumes that its users are already familiar with the composition of the multiverse analysis. As such, Boba may not be as readily usable by users unfamiliar with the implemented multiverse analysis.

In Boba, users can view the distribution of point estimates from each universe, using a dot plot (**T2**). The uncertainty in the point estimates are shown through a hover interaction using density plots. Boba visualises an "end-to-end" uncertainty distribution—it aggregates sampling uncertainty and the variation arising due to alternative choices in the tree of decisions comprising the multiverse—and visualises it as an area chart (see Fig. 5 in [34]), but does not distinguish between probabilistic and possibilistic uncertainty. These plots show the extent of variation in the results. Boba calculates the marginal sensitivity for each decision and encodes this metric directly in the visualisation interface. For this dataset, it highlights the parameter **activity**. To further support connect tasks (**T3**), Boba allows the user to facet the outcome distribution by options of up to two parameters, one each along the x and y axes (see Fig. 6 in [34]).

	Boba	Milliways
T1 Composition tasks	partially supported through the decision graph	supported through the code and data panels
T2 Outcome tasks	fully supported	fully supported
T3 Connect tasks	partially supported through faceting	fully supported
T4 Validation tasks	supports model fit metric-based validation only	supports validation based on domian knowledge and statistical expertise
T5 Interpretation tasks	supports inference–to conclude whether the there is an effect or not	supports possibilistic interpretation–to identify the range of possible outcomes

Table 1: Comparison of the extent to which multiverse analysis tasks are supported by Milliways and Boba

Boba supports validation (**T4**) of the paths specified within the multiverse using metrics such as quality of model fit (RMSE or R^2). For this analysis, pruning based on the R^2 metric results in only retaining the specifications where the variable `activity` has been included as a covariate; however, as `activity` is a collider (by construction), it should not be included. This highlights one possible limitation of a metric-based validation approach. In contrast, Milliways treats such metrics as another outcome variable, and instead emphasises assessing validity through a principled evaluation based on the readers' statistical expertise and domain knowledge. To help the analyst interpret (**T5**) the results, Boba provides a distinct "inference view," which compares the aggregated uncertainty distribution with a null distribution or null effect (see Fig. 9 in [34]). This view is intended to help the user determine whether an observed effect is reliable or not—a *probabilistic* conclusion. On the other hand, Milliways allows the reader to identify the range of *possible* outcomes, supporting principled possibilistic interpretations such as those described at the end of S5.1

6 EVALUATION

We conducted an hour-long user study with five researchers to: (a) understand how participants interact with the tool; and (b) assess the system's usability for performing the multiverse analysis tasks (§3.1).

6.1 Method

Recruitment. We recruited five participants from three sources—attendees of a session on multiverse analysis at a conference (two), convenience sampling (two) and via Twitter (one). All participants had significant experience in statistics, strong familiarity with R, and possessed conceptual familiarity with multiverse analysis. None of the participants belonged to the same institution as, or were involved

in any prior research collaboration with the authors. We asked participants how they would rate their level of expertise if they were asked to review a paper (in their field of expertise) containing a multiverse analysis.[5] All participants responded they would consider themselves either being "knowledgeable" or having "passing knowledge."[6] These participants reflect the population of intended users of Milliways—researchers who may implement a multiverse analysis, or review a study containing a multiverse analysis. Participants were compensated $60 USD for a session of approximately one hour.

Study setup. The study was conducted remotely over Zoom with participants sharing their screen. The study consisted of two sessions—a training session followed by a task session. Since Milliways can be compiled into standalone webpages, we uploaded the training and task interface to a web server, allowing participants to use the tool on their own computers. The training session involved an interactive step-through tutorial which introduced the participants to the various interface elements in Milliways. This was followed by a demonstration of how the features of the interface may be used to interpret the results of the multiverse through three videos. Participants were informed that the steps shown in the video were not prescriptive, but rather one way of performing exploratory analysis of a multiverse. To ensure that participants understood the visual representations used—how the representations were constructed, what they encoded etc.—participants were encouraged to ask questions regarding any of the interface elements and visual representations used during training, which were then discussed and clarified before proceeding to the tasks.

Tasks. After the training session, participants were introduced to the scenario, dataset (§2.2), and high-level goal of the analysis:

> *Imagine you are a researcher who has been asked to review a (fictional) study because of your statistical expertise. Your goal is to understand and validate the composition of the multiverse analysis and interpret the results with the help of the visualisation interface.*

Participants were then asked to perform the multiverse analysis tasks (**T1**-**T5**) [26] with Milliways. Specifically, they were asked to describe the decisions and options that were included in the analysis (**T1**), identify the range of outcomes and whether they consider the outcome to be sensitive to choices in the data analysis process (**T2**), identify the decisions and options (if any) the outcome is sensitive to (**T3**), identify decisions or options which they may consider not theoretically justifiable or require further investigation (**T4**), and describe their overall conclusions of the presented analysis (**T5**). Participants were encouraged to think aloud while performing these tasks. Once participants completed the tasks, we asked them a few post-study questions to get their takeaways from performing the tasks and to receive feedback on the design of the interface. We recorded participants' screens and audio for the duration of the study, and transcribed the audio recordings (see `supplement ▶ user-study ▶ interview-protocol.pdf` for the full interview protocol).

[5] this question is intended to be similar to the expertise question asked in reviewer forms for ACM CHI and IEEE VIS.

[6] For details on individual participants responses to this questionnaire, see `supplement ▶ user-study ▶ participant-details.pdf`

6.2 Results

Overall, four out of the five participants were able to successfully perform each of the multiverse analysis tasks. In addition, these participants raised concerns about the validity of certain alternative paths in the multiverse analysis based on their own statistical expertise. This suggests that Milliways can support each of the multiverse analysis tasks and allow the user to adopt a principled approach towards evaluating the validity of the analysis. However, one participant (P5) failed to complete the tasks, as they misinterpreted various interface elements, which points to various learnability concerns that can inform future improvements for Milliways. We first analyse how the four participants approached the tasks using Milliways, followed by a discussion of the usability and learnability issues that our study revealed.

The code panel surfaces multiverse composition. To perform the **composition** tasks, participants tended to use either the **specification** or the **code** panels. They generally found the **code** panel to be intuitive and helpful for understanding the decisions and choices declared in the multiverse. This was likely assisted by participants' proficiency in R.

Participants assess outcome sensitivity based on inconsistency in the direction of effects. Most participants referred to the histogram to determine whether the **outcome** is sensitive to analytical choices, while some explored the **outcome** panel in more detail. All participants described how the outcomes spanned both positive and negative values, which suggested to them that the direction of effects is unclear and that perhaps certain analysis paths might lead to a *Type S error*—when the estimated effect, if statistically significantly different from zero, has the incorrect sign [23]. This led them to conclude that the outcome is likely sensitive to choices in the analysis process.

Participants engaged in a visual pattern search to identify sources of sensitivity. All participants successfully identified that the outcome was sensitive to choices of the **activity** decision (**T3**), mentioning that this was "*visually obvious*." Two participants said that the outcome may also be sensitive to certain choices of **sm predictor**, but they were more hesitant about this. Participants engaged in a visual search for patterns across the outcome and specification panels to perform this task. In performing this task, participants used the slider to sort by **average effect** extensively; however, they did not engage much with the filter and aggregation interactions. P4 appreciated the exploratory nature of the interface, and mentioned that they would have "*played around more*" with the interactive elements if they had more time.

When engaged in this visual search, participants interacted with and explored many views of the multiverse analysis. Participants wanted the ability to keep track of progress: through interaction logs (P1), undo (and redo) interactions (P3), and saving specific views or configurations to compare across them (P3); they felt that such features would support provenance tracking and make performing these tasks more efficient.

Milliways allows users to assess validity based on their respective principles of statistical modeling. During almost every task, most participants raised questions regarding the "reasonable-ness" or "justifiability" of choices in the multiverse. This was often pertaining to the causal processes being investigated in the study. For instance:

> *"for most [decisions] [...] how do the covariates relate to other factors like SES, how does it relate to the causal model [in terms of] moderators, confounds etc.?"* (P1)

Other participants also raised similar questions: "*[I] have questions about variables in the dataset which are not included in the multiverse, and the rationale for excluding them*" (P2) and "*how to [best] measure social media usage as it's quite a broad concept*" (P4). We also observed a tendency in participants towards including predictors or covariates in the linear regression model, as opposed to leaving them out. For instance, while participants wanted to better understand the justification of including or not including activity as a covariate, they were more skeptical of the choice of leaving it out. Participants constantly sought justification for options which they suspected to be not equally reasonable, particularly those for the **sm predictor** parameter. All four participants said they would want justification from the authors regarding why only certain combinations of the measures were used to create composites.

As all of our participants were experienced researchers with extensive statistical training, they appeared to approach the validation process based on their own principles of statistical modeling. For instance, P1 tends to be more skeptical of any form of outlier exclusion:

> *"the tail is a part of the distribution, so what is the justification? [...] are you artificially reducing the variance [in the data using outlier exclusion]?"* (P1)

Other participants shared different perspectives regarding the same outlier exclusion decision in the multiverse analysis: one participant wanted to know "*how commonly are these used in the literature? There are other like, more robust methods that sometimes get used*" (P2), while another participant believes that "*people can think about the [outlier exclusion] criteria and just declare that in a transparent manner, it will be fine"* (P3). These varying perspectives regarding the same decision suggest that, for many decisions in the multiverse, there may not be a consensus on which choices will be considered "equivalent." Milliways was able to support these various notions of equivalence, helping participants apply their own criteria.

Supporting a possibilistic interpretation of the results. Participants offered a number of interpretations of the overall results of the multiverse analysis, some of which pertain to the validity of the multiverse. However, responses from two participants in particular suggests that they may be interpreting the results, visualised using p-boxes of consonance curves, in a possibilistic manner:

> *"since [the values] are passing from negative to positive effects, that it's definitely inconclusive [...] even though there seems to be a few more seems to suggest that there is a positive effect, if I can say that."* (P4)

Offering both a possibilistic and probabilistic interpretation of the overall result of the multiverse, P4 seemed unsure whether a strictly possibilistic interpretation (the results are "inconclusive") or a probabilistic interpretation—viewing more frequent values as more likely to reflect the actual effect—is valid. On the other hand, P1 "typically use(s) multiverse analysis [...] as the opposite of p-hacking [...] where I want to make sure there is no reasonable choice or variant of the analysis that could be done, that would not show the same effect." While this reflects a possibilistic interpretation of the results, it also suggests that they had prior notions of interpreting the results of a multiverse in this manner.

Even though the other participants found Milliways to be helpful, they were not very confident regarding their assessment of the multiverse analysis. They primarily discussed the sensitivity of the outcomes to analytical decisions, and did not offer interpretations of the overall uncertainty in the results of the multiverse. While their responses did not suggest a probabilistic interpretation of the result, we cannot determine whether P2 and P3 interpreted the results in the desired possibilitic sense. We attribute this to both the complexity of understanding a multiverse analysis, and of becoming familiar with a new user interface within a short span of time.

Learnability issues and mental models. One participant initially failed to understand the matrix-based representation of the decisions in the multiverse, and did not interpret the rows of the **specification** panel correctly. During the post-task discussion, they revealed that they were used to viewing and reasoning about multiverse analyses as decision trees. We believe that this resulted in a gulf between the user's mental model and our representation of multiverse analysis decisions. This gulf, however, meant they were not able to perform the tasks. This raises potential concerns regarding the learnability of Milliways. While we describe what each interface element represents in the tutorial, and specifically connect the "universes" to the terminal nodes in the tree representation, as evidenced by this participant, it is still possible for users of the tool to be confused by the linearised layout used in Milliways. We discuss potential ways to address this gulf in §7.3.

Overall, the four participants who were able to complete the tasks did not report facing any issues. However, one participant did mention that "*some parts of the tool would still require me to [take time] playing around with*" (P4). P1 and P2 performed the composition task primarily based on the **specification** panel, and required prompting to explore the code panel. Once they started using the code panel, however, they found the task easier to perform. Similarly, while performing the range task, P1 did not initially realise that the histogram plot depicted the median point estimate from each universe. These highlight potential learnability issues with the interface of Milliways—as participants were introduced to several interface features within such a short period of time, it was likely cognitively demanding for them to keep track of each feature. Moreover, as the tasks were presented to the participants in a particular order, P2, P4 and P5 thought that it may have been helpful if they were only provided with the specific panel relevant for performing the tasks, and the other panels were progressively revealed. Although Milliways does not share the same complexity as feature-rich softwares such as Adobe Photoshop, these concerns point towards the potential for incorporating principles of task-centric interface design [30] in Milliways to ease initial learnability issues. For example, as all participants first start with the **composition** task, the interface might initialise with only the code panel. Participants also desired the ability to collapse panels which are not actively being used.

7 DISCUSSION AND FUTURE WORK

7.1 Principled Evaluation of Multiverse Analysis

The ability to reason about the validity of alternatives in the analysis requires the participants to have a clear understanding of how the analysis was performed. Our usability study revealed that participants were constantly assessing the validity of the choices in the multiverse, even while they were performing other analysis tasks. For instance, we noticed that some participants, based on observations in the **outcome** panel, would refer to the **code** panel to assess the validity of choices.

This suggests that Milliways surfaces enough information about the construction of the multiverse and its results for participants to be able to perform validation tasks. We believe that this is primarily supported through the **code** panel. This was further supported by comments made by participants: "*a domain expert who knows [...] which factors should be included, they don't even need to see this [outcome plot]*" (P3). This seems to suggests the value of integrating the code used to declare the analysis within multiverse visualisations, and fulfils an intended design goal for Milliways—allowing users to evaluate the composition of multiverses using a principled approach, without having to switch between applications and files.

7.2 The Design of Possibilistic Representations of Uncertainty

The design of Milliways makes a deliberate distinction between probabilistic and possibilistic uncertainty that arises in multiverse analysis. This resulted in the use of consonance curves to depict probabilistic uncertainty and p-boxes to depict possibilistic uncertainty. P-boxes, by design, are less expressive than other comparable representations such as super-imposed consonance curves or cumulative distribution functions, as they only provide distributional information in the form of upper and lower bounds, but do not provide any information on frequencies or probability densities. Thus, p-boxes allow us to express an appropriate level of incertitude about the results. Novel uncertainty representations such as quantile dotplots [28] or ensemble plots [13] have been evaluated in various decision-making scenarios (*e.g.* [20, 28, 41, 47]) to determine their effectiveness. However, interpretation of results depicted using p-boxes, to the best of our knowledge, not been studied in large-scale controlled experiments. While a subset of the users in our evaluation interpreted the results, visualised using p-boxes, in the desired possibilistic sense, we hope to conduct a more focused, controlled experiment exploring the effectiveness of such possibilistic representations of uncertainty for interpreting and making decisions in the future [27].

Additionally, Milliways uses a histogram to provide an overview of the distribution and range of the median point estimates from each universe. The histogram also enables a brushing-and-linking interaction for filtering specifications based on their median point estimate. However, a histogram is essentially a graph used to represent the frequency of values, which has the undesired effect of being susceptible to a probabilistic interpretation. On the other hand, representations such as p-boxes likely provide the desired possibilistic interpretation, yet interactions such as brushing-and-linking may be less intuitive and theoretically unsound. This raises an interesting design challenge—how do we create a summary visualisation that correctly, and distinctly, represents probabilistic and possibilistic uncertainty, and supports intuitive interactions for brushing-and-linking or filtering?

7.3 Reducing the gulf between tree and matrix-based representations

While tabular or matrix-based representations are commonly used to represent graphs [37], one participant in our user study failed to complete the tasks due to misinterpreting the data being presented in the specification panel. Prior work has found that participants are able to read the data encoded in matrix-based representations of graphs, despite being unfamiliar with them [24]. However, the multiverse analysis tasks [26] differ significantly from the multivariate network analysis tasks [37] that a reader performs with graphs. As such, we believe that the connection between our representation and decision graphs can be made more concrete and explicit, which we hope to include in Milliways in the future.

7.4 Challenges to adoption of Multiverse visualisation interfaces

Milliways requires the input data to be specified in a certain schema (§4.6). One potential challenge for adoption of visualisation systems such as Milliways is if the authors of multiverse analyses do not provide data that is compatible (or easily adaptable to be compatible) with Milliways (§4.6). However, statistical procedures like multiverse analysis [51, 53] have been proposed as part of the movement of increasing transparency in research practices, which also calls for sharing research materials, data, analysis scripts etc. [57, 58]. With the goal of transparency in mind, we expect, at the very least, authors of multiverse analyses to share their data analysis scripts. As we provide a pipeline for adapting the results of a multiverse to be used with Milliways using the Milliways R interface (§4.6), we hope that eager readers of such analyses can use the tools provided to create the necessary files, and visualise the results of any multiverse analysis easily using Milliways.

8 CONCLUSION

We contribute Milliways, a novel interactive visualisation system to communicate the results of a multiverse analysis. In the design of Milliways, we adopt a principled approach to communicating the results of a multiverse analysis, by supporting possibilistic reasoning and allowing the user to interrogate the validity of the multiverse based on domain knowledge and statistical expertise. Milliways provides several interactive features that allow users to perform multiverse analysis tasks identified in prior work [26]. Through a user study with five researchers, we found that the design of Milliways encouraged participants to engage in a principled evaluation of the results of the multiverse analysis shown to them—the contextual information provided in the **code** and **data** panels allowed the participants to understand the composition of the analysis and validate the construction of the multiverse; the linearised layout and distributional information presented in the **specification** and **outcome** panels enabled the participants to understand the distribution of outcomes and reason about sources of sensitivity. The findings from our user studies also revealed areas for future work to improve the design of tools to evaluate the results of multiverse analyses.

REFERENCES

[1] Valentin Amrhein and Sander Greenland. 2022. Discuss practical importance of results based on interval estimates and p-value functions, not only on point estimates and null p-values. *Journal of Information Technology* 37, 3 (Sept. 2022), 316–320. https://doi.org/10.1177/02683962221105904 Publisher: SAGE Publications Ltd.

[2] Ruben C. Arslan, Katharina M. Schilling, Tanja M. Gerlach, and Lars Penke. 2021. Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *Journal of Personality and Social Psychology* 121, 2 (2021), 410–431. https://doi.org/10.1037/pspp0000208

[3] Allan Birnbaum. 1961. A Unified Theory of Estimation, I. *The Annals of Mathematical Statistics* 32, 1 (1961), 112–135. https://www.jstor.org/stable/2237612 Publisher: Institute of Mathematical Statistics.

[4] Georges-Pierre Bonneau, Hans-Christian Hege, Chris R. Johnson, Manuel M. Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. 2014. Overview and State-of-the-Art of Uncertainty Visualization. In *Scientific Visualization*, Charles D. Hansen, Min Chen, Christopher R. Johnson, Arie E. Kaufman, and Hans Hagen (Eds.). Springer London, London, 3–27. https://doi.org/10.1007/978-1-4471-6497-5_1 Series Title: Mathematics and Visualization.

[5] Maryam Booshehrian, Torsten Möller, Randall M. Peterman, and Tamara Munzner. 2012. Vismon: Facilitating Analysis of Trade-Offs, Uncertainty, and Sensitivity In Fisheries Management Decision Making. *Computer Graphics Forum* 31, 3pt3 (2012), 1235–1244. https://doi.org/10.1111/j.1467-8659.2012.03116.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2012.03116.x

[6] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. 2011. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics* 17, 12 (2011), 2301–2309.

[7] Christopher J Bryan, David S Yeager, and Joseph M O'Brien. 2019. Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences* 116, 51 (2019), 25535–25545.

[8] Joseph Cesario, David J Johnson, and William Terrill. 2019. Is there evidence of racial disparity in police use of deadly force? Analyses of officer-involved fatal shootings in 2015–2016. *Social psychological and personality science* 10, 5 (2019), 586–595.

[9] Chun-Houh Chen, Hai-Gwo Hwu, Wen-Jung Jang, Chiun-How Kao, Yin-Jing Tien, ShengLi Tzeng, and Han-Ming Wu. 2004. Matrix visualization and information mining. In *COMPSTAT 2004—Proceedings in Computational Statistics: 16th Symposium Held in Prague, Czech Republic, 2004*. Springer, 85–100.

[10] Pasquale Cirillo and Nassim Nicholas Taleb. 2016. On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications* 452 (2016), 29–45.

[11] Stephen R Cole, Robert W Platt, Enrique F Schisterman, Haitao Chu, Daniel Westreich, David Richardson, and Charles Poole. 2009. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology* 39, 2 (11 2009), 417–420. https://doi.org/10.1093/ije/dyp334 arXiv:https://academic.oup.com/ije/article-pdf/39/2/417/18480441/dyp334.pdf

[12] J Anthony Cookson. 2018. When saving is gambling. *Journal of Financial Economics* 129, 1 (2018), 24–45.

[13] Jonathan Cox, Donald House, and Michael Lindell. 2013. Visualizing uncertainty in predicted hurricane tracks. *International Journal for Uncertainty Quantification* 3, 2 (2013), 143–156.

[14] Marcus Credé and Leigh A Phillips. 2017. Revisiting the power pose effect: How robust are the results reported by Carney, Cuddy, and Yap (2010) to data analytic decisions? *Social Psychological and Personality Science* 8, 5 (2017), 493–499.

[15] Egon Dejonckheere, Elise K Kalokerinos, Brock Bastian, and Peter Kuppens. 2019. Poor emotion regulation ability mediates the link between depressive symptoms and affective bipolarity. *Cognition and Emotion* 33, 5 (2019), 1076–1083.

[16] Egon Dejonckheere, Merijn Mestdagh, Marlies Houben, Yasemin Erbas, Madeline Pe, Peter Koval, Annette Brose, Brock Bastian, and Peter Kuppens. 2018. The bipolarity of affect and depressive symptoms. *Journal of personality and social psychology* 114, 2 (2018), 323.

[17] Seamus Donnelly, Patricia J Brooks, and Bruce D Homer. 2019. Is there a bilingual advantage on interference-control tasks? A multiverse meta-analysis of global reaction time and interference cost. *Psychonomic bulletin & review* 26, 4 (2019), 1122–1147.

[18] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. 2019. Increasing the Transparency of Research Papers with Explorable Multiverse Analyses. 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, 1–15. https://doi.org/10.1145/3290605.3300295

[19] Julian Faraway. 2015. *Diagnostics* (2nd ed.). CRC, 83–89.

[20] 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, Montreal QC Canada, 1–12. https://doi.org/10.1145/3173574.3173718

[21] Scott Ferson and Jack Siegrist. 2012. Verified Computation with Probabilities. In *Uncertainty Quantification in Scientific Computing*, Andrew M. Dienstfrey and Ronald F. Boisvert (Eds.). IFIP Advances in Information and Communication Technology, Vol. 377. Springer Berlin Heidelberg, Berlin, Heidelberg, 95–122. https://doi.org/10.1007/978-3-642-32677-6_7

[22] Brian R. Gaines and Ladislav J. Kohout. 1976. The Logic of Automata. *International Journal of General Systems* 2, 4 (Jan. 1976), 191–208. https://doi.org/10.1080/03081077608547469

[23] Andrew Gelman and John Carlin. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* 9, 6 (Nov. 2014), 641–651. https://doi.org/10.1177/1745691614551642 Publisher: SAGE Publications Inc.

[24] Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. 2005. On the Readability of Graphs Using Node-Link and Matrix-Based Representations: A Controlled Experiment and Statistical Analysis. *Information Visualization* 4, 2 (June 2005), 114–135. https://doi.org/10.1057/palgrave.ivs.9500092 Publisher: SAGE Publications.

[25] Marco Del Giudice and Steven W. Gangestad. 2021. A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Advances in Methods and Practices in Psychological Science* 4, 1 (2021), 2515245920954925. https://doi.org/10.1177/2515245920954925 arXiv:https://doi.org/10.1177/2515245920954925

[26] Brian D. Hall, Yang Liu, Yvonne Jansen, Pierre Dragicevic, Fanny Chevalier, and Matthew Kay. 2022. A Survey of Tasks and Visualizations in Multiverse Analysis Reports. *Computer Graphics Forum* 41, 1 (2022), 402–426. https://doi.org/10.1111/cgf.14443

[27] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2019. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 903–913. https://doi.org/10.1109/TVCG.2018.2864889

[28] 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, San Jose California USA, 5092–5103. https://doi.org/10.1145/2858036.2858558

[29] Will Kurt. 2019. *Bayesian statistics the fun way: understanding statistics and probability with Star Wars, Lego, and Rubber Ducks* (1st ed.). No Starch Press, Chapter 13.

[30] Benjamin Lafreniere, Andrea Bunt, and Michael Terry. 2014. Task-centric interfaces for feature-rich software. In *Proceedings of the 26th Australian Computer-Human Interaction Conference on Designing Futures: the Future of Design*. ACM, Sydney New South Wales Australia, 49–58. https://doi.org/10.1145/2686612.2686620

[31] Alexander Lex, Nils Gehlenborg, Hendrik Strobelt, Romain Vuillemot, and Hanspeter Pfister. 2014. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 1983–1992. https://doi.org/10.1109/TVCG.2014.2346248 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[32] Le Liu, Lace Padilla, Sarah H. Creem-Regehr, and Donald H. House. 2019. Visualizing Uncertain Tropical Cyclone Predictions using Representative Samples from Ensembles of Forecast Tracks. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 882–891. https://doi.org/10.1109/TVCG.2018.2865193 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[33] Yang Liu, Tim Althoff, and Jeffrey Heer. 2020. Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data 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, 1–14. https://doi.org/10.1145/3313831.3376533

[34] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. 2021. Boba: Authoring and Visualizing Multiverse Analyses. *IEEE Transactions on Visualization and Computer Graphics* 27, 2, 1753–1763. https://doi.org/10.1109/TVCG.2020.3028985

[35] Richard McElreath. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC.

[36] Carolina Nobre, Nils Gehlenborg, Hilary Coon, and Alexander Lex. 2019. Lineage: Visualizing Multivariate Clinical Data in Genealogy Graphs. *IEEE Transactions on Visualization and Computer Graphics* 25, 3 (March 2019), 1543–1558. https://doi.org/10.1109/TVCG.2018.2811488 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[37] Carolina Nobre, Miriah Meyer Meyer, Marc Streit, and Alexander Lex. 2019. The State of the Art in Visualizing Multivariate Networks. *Computer Graphics Forum* 38, 3 (2019), 807–832. https://doi.org/10.1111/cgf.13728 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.13728.

[38] Carolina Nobre, Marc Streit, and Alexander Lex. 2019. Juniper: A Tree+Table Approach to Multivariate Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan. 2019), 544–554. https://doi.org/10.1109/

TVCG.2018.2865149

[39] Deborah Nolan and Duncan Temple Lang. 2007. Dynamic, interactive documents for teaching statistical practice. *International Statistical Review* 75, 3 (2007), 295–321.

[40] Amy Orben and Andrew K Przybylski. 2019. The association between adolescent well-being and digital technology use. *Nature Human Behaviour* 3, 2 (2019), 173–182.

[41] Lace M. Padilla, Ian T. Ruginski, and Sarah H. Creem-Regehr. 2017. Effects of ensemble and summary displays on interpretations of geospatial uncertainty data. *Cognitive Research: Principles and Implications* 2, 1 (Oct. 2017), 40. https://doi.org/10.1186/s41235-017-0076-1

[42] Chirag J Patel, Belinda Burford, and John PA Ioannidis. 2015. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of clinical epidemiology* 68, 9 (2015), 1046–1058.

[43] Charles Perin, Pierre Dragicevic, and Jean-Daniel Fekete. 2014. Revisiting Bertin Matrices: New Interactions for Crafting Tabular Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (Dec. 2014), 2082–2091. https://doi.org/10.1109/TVCG.2014.2346279 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[44] Gregory J Poarch, Jan Vanhove, and Raphael Berthele. 2019. The effect of bidialectalism on executive function. *International Journal of Bilingualism* 23, 2 (2019), 612–628.

[45] C Poole. 1987. Beyond the confidence interval. *American Journal of Public Health* 77, 2 (Feb. 1987), 195–199. https://doi.org/10.2105/AJPH.77.2.195

[46] Zad Rafi and Sander Greenland. 2020. Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. *BMC Medical Research Methodology* 20, 1 (Sept. 2020), 244. https://doi.org/10.1186/s12874-020-01105-9

[47] Ian T. Ruginski, Alexander P. Boone, Lace M. Padilla, Le Liu, Nahal Heydari, Heidi S. Kramer, Mary Hegarty, William B. Thompson, Donald H. House, and Sarah H. Creem-Regehr. 2016. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation* 16, 2 (April 2016), 154–172. https://doi.org/10.1080/13875868.2015.1137577

[48] Abhraneel Sarma, Shunan Guo, Jane Hoffswell, Ryan Rossi, Fan Du, Eunyee Koh, and Matthew Kay. 2023. Evaluating the Use of Uncertainty Visualisations for Imputations of Data Missing At Random in Scatterplots. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 602–612. https://doi.org/10.1109/TVCG.2022.3209348

[49] Abhraneel Sarma, Alex Kale, Michael Jongho Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. 2023. Multiverse: Multiplexing Alternative Data Analyses in R Notebooks. 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 148, 15 pages. https://doi.org/10.1145/3544548.3580726

[50] Michael Sedlmair, Christoph Heinzl, Stefan Bruckner, Harald Piringer, and Torsten Möller. 2014. Visual Parameter Space Analysis: A Conceptual Framework. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 2161–2170. https://doi.org/10.1109/TVCG.2014.2346321

[51] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. 2020. Specification curve analysis. *Nature Human Behaviour* 4, 11 (Nov 2020), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

[52] Kesar Singh, Minge Xie, and William E. Strawderman. 2007. Confidence Distribution (CD): Distribution Estimator of a Parameter. *Lecture Notes-Monograph Series* 54 (2007), 132–150. https://www.jstor.org/stable/20461464 Publisher: Institute of Mathematical Statistics.

[53] Sara Steegen, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science* 11, 5 (2016), 702–712.

[54] Kevin M. Sullivan and David A. Foster. 1990. Use of the Confidence Interval Function. *Epidemiology* 1, 1 (1990), 39–42. https://www.jstor.org/stable/20065621 Publisher: Lippincott Williams & Wilkins.

[55] Rachel Visontay, Louise Mewton, Matthew Sunderland, Steven Bell, Annie Britton, Bridie Osman, Hayley North, Nisha Mathew, and Tim Slade. 2022. A comprehensive evaluation of the longitudinal association between alcohol consumption and inflammation: Multiverse and vibration of effects analyses. (2022).

[56] Martin Voracek, Michael Kossmeier, and Ulrich S Tran. 2019. Which Data to Meta-Analyze, and How? *Zeitschrift für Psychologie* (2019).

[57] Chat Wacharamanotham, Matthew Kay, Steve Haroz, Shion Guha, and Pierre Dragicevic. 2018. Special Interest Group on Transparent Statistics Guidelines. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (, Montreal QC, Canada,) *(CHI EA '18)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3170427.3185374

[58] Chat Wacharamanotham, Fumeng Yang, Xiaoying Pu, Abhraneel Sarma, and Lace Padilla. 2022. Transparent Practices for Quantitative Empirical Research. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 122, 5 pages. https://doi.org/10.1145/3491101.3503760

[59] Han-Ming Wu, ShengLi Tzeng, Chun-houh Chen, Wolfgang Karl Härdle, and Antony Unwin. 2008. *Matrix Visualization*. 681–708. https://doi.org/10.1007/978-3-540-33037-0_26

[60] Min-ge Xie and Kesar Singh. 2013. Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. *International Statistical Review* 81, 1 (2013), 3–39. https://doi.org/10.1111/insr.12000 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/insr.12000.

[61] Cristobal Young and Katherine Holsteen. 2017. Model uncertainty and robustness: A computational framework for multimodel analysis. *Sociological Methods & Research* 46, 1 (2017), 3–40.

[62] Anđela Šoškić, Suzy Styles, Emily Kappenman, and Vanja Kovic. 2022. Garden of forking paths in ERP research – effects of varying pre-processing and analysis steps in an N400 experiment. https://doi.org/10.31234/osf.io/8rjah