DOI: 10.1111/cgf.14443

COMPUTER GRAPHICS *forum*
Volume 41 (2022), number 1 pp. 402–426

A Survey of Tasks and Visualizations in Multiverse Analysis Reports

Brian D. Hall,[1] Yang Liu,[2] Yvonne Jansen,[3] Pierre Dragicevic,[4] Fanny Chevalier[5] and Matthew Kay[6]

[1]University of Michigan, Ann Arbor, MI, USA
[2]University of Washington, Seattle, WA, USA
[3]Sorbonne Universite, CNRS, ISIR, Paris, France
[4]Universite Paris-Saclay, CNRS, Inria, LISN, France
[5]University of Toronto, Toronto, ON, Canada
[6]Northwestern University, Evanston, IL, USA

Abstract

Analysing data from experiments is a complex, multi-step process, often with multiple defensible choices available at each step. While analysts often report a single analysis without documenting how it was chosen, this can cause serious transparency and methodological issues. To make the sensitivity of analysis results to analytical choices transparent, some statisticians and methodologists advocate the use of 'multiverse analysis': reporting the full range of outcomes that result from all combinations of defensible analytic choices. Summarizing this combinatorial explosion of statistical results presents unique challenges; several approaches to visualizing the output of multiverse analyses have been proposed across a variety of fields (e.g. psychology, statistics, economics, neuroscience). In this article, we (1) introduce a consistent conceptual framework and terminology for multiverse analyses that can be applied across fields; (2) identify the tasks researchers try to accomplish when visualizing multiverse analyses and (3) classify multiverse visualizations into 'archetypes', assessing how well each archetype supports each task. Our work sets a foundation for subsequent research on developing visualization tools and techniques to support multiverse analysis and its reporting.

Keywords: multiverse analysis, sensibility analysis, transparent reporting, statistical graphics

CCS Concepts: • Human-centered computing → Visualization; Visualization theory, concepts and paradigms; Mathematics of computing; Statistical graphics; • Human-centered computing → Visualization techniques

1. Introduction: Multiverse Analyses and Visualizations

Analysing data from experiments is a complex, multi-step process, with multiple choices available at each step, e.g. whether and how to exclude outliers, what approach to use to operationalize a variable, or what model and parameters to apply [2020]. While it is often possible to exclude some choices as invalid, often many alternatives remain that are equally valid. Faced with this complexity, analysts often try multiple analyses and report a single one without documenting how it was chosen. This practice, sometimes termed *undisclosed flexibility*, can cause serious transparency and methodological issues [SNS11, [2016, 2013] and has been identified as a major cause of the replicability crisis in psychology and other disciplines [NAB*15, MNB*17, Cum14]. Combined with a desire to report positive findings, undisclosed flexibility can be damaging because it substantially increases the chances of reporting erroneous findings, while being invisible to the reader.

One increasingly advocated solution to the issue of undisclosed flexibility is *pre-registration* [Ber12, 2018], whereby all analytical choices are made before the data are collected and submitted to a verifiable registry. Pre-registration eliminates undisclosed flexibility, but still hides *analytic uncertainty*: the extent to which results are dependent on the particular analytic choices made. Different researchers who analyse the same data will often make different choices and get slightly—and sometimes widely—different results [SUM*18a]. If one of these researchers were to pre-register their analysis, their report would still convey an incomplete picture.

Some statisticians and methodologists have promoted the use of *multiverse analysis* to convey a much fuller picture of analytic uncertainty [STGV16, SSN19]. This approach consists of identifying a set of defensible analytical choices, performing all analyses corresponding to the possible combinations of such choices

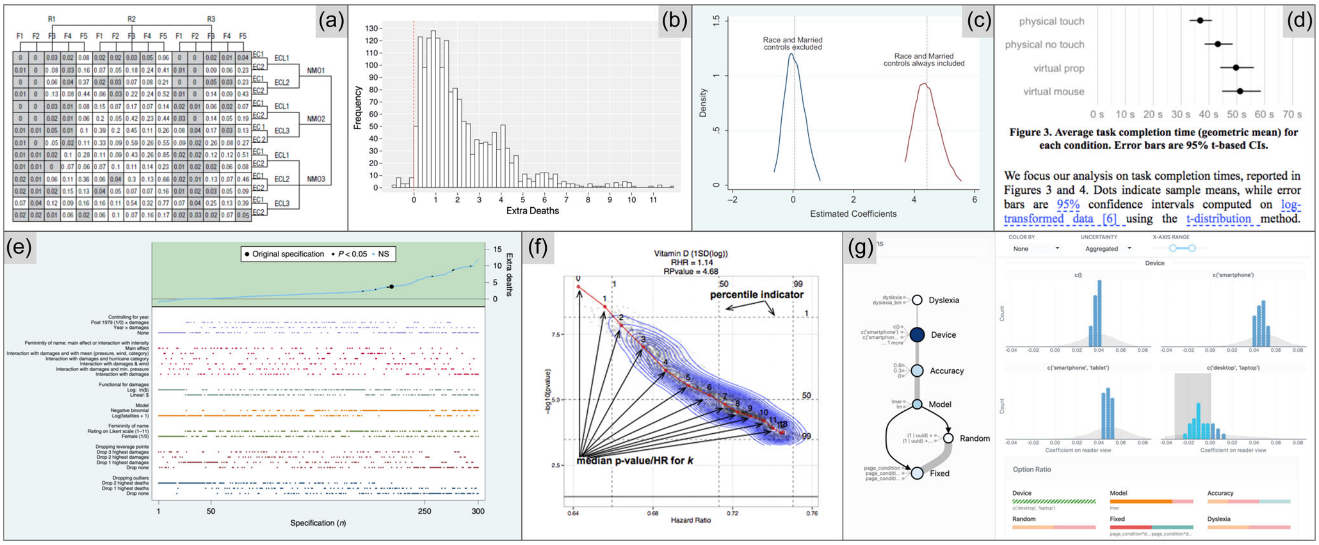


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data [6] using the t-distribution method.

Figure 1: *Examples of multiverse analysis visualizations discussed in this survey: (a) outcome matrix [STGV16], (b) outcome histogram [STGV16], (c) outcome density plot [YH17], (d) explorable multiverse analysis reports [2019], (e) specification curve [SSN20], (f) vibration of effects plot [PBI15], (g) Boba [LKAH20].*

(possibly hundreds, thousands or even millions) and reporting all outcomes, typically using summary visualizations. This idea is increasingly popular, with more and more academic papers reporting multiverse analyses; for example, a Google Scholar search for the term 'specification curve'—a type of multiverse analysis visualization [SSN20]—returns 217 papers for the years 2019–2020.

However, multiverse analyses still raise many challenges, three of which serve as primary motivators for this work: (1) explaining and reporting the outcomes of hundreds or thousands of statistical analyses is difficult, especially when some of those analyses do not all point towards the same general conclusions [SSN20, 2019]; (2) literature specifically discussing the methodology of multiverse analyses is scattered across several fields and uses inconsistent terminology and (3) visualization methods that have been proposed for helping to conduct and report multiverse analyses are similarly scattered across several fields and use inconsistent terminology. For example, considering the methodological literature (challenge 2), although the term *multiverse analysis* [STGV16] is recent, the core concept is found in older techniques under different names (e.g. *sensitivity analysis*, *robustness analysis*). These approaches have developed independently in different fields, leading to different terminology that often conflicts, which can make it difficult to communicate or reason about multiverse concepts. Similarly, static visual summaries of multiverse analyses (e.g. [STGV16, SSN19, PBI15]; Figure 1(a), (e), (f)) or interactive visualizations of multiverse analyses (e.g. [2019, LKAH20]; Figure 1(d), (g)) have been developed in different fields and under different names (challenge 3). Some more general visualization methods, like specification curve [SSN20], have been adopted in research papers, often with modifications, adaptations and improvements (e.g. [OP19a, BRRYD20]). Meanwhile, many papers use custom visualization methods for reporting multiverse analyses (e.g. [BNHC*20, BKB*20]). Some visualizations are domain-specific (e.g.: neuroimaging [Car12, BNHC*20]), or published in venues that may not be widely read outside of their field (e.g.: hydrology [Bie15]).

For a researcher who wants to report a multiverse analysis, these challenges make it hard to make informed choices about which visualizations to use; for a researcher who wants to study new multiverse visualization techniques, or teach the topic, it is hard to get a good overview of the state of the art. This article addresses the above challenges through a survey of academic articles that visualize multiverse analyses and related analyses. Importantly, our survey only covers ways visualization has been used to *report* multiple statistical analyses in an academic communication context. It does not discuss ways visualization has been used to help analysts *explore* multiple analyses, for example, in the context of model steering and selection [DCCE19, MLMP17, CPCS19], ensemble data analysis [WHLS18] and visual parameter space analysis [SHB*14]. Our scope is further clarified in Section 3.

In this survey, we (1) propose a conceptual framework and terminology for multiverse analyses that can be applied across fields, to support clarity when discussing this nascent family of concepts (Section 3); (2) identify the *tasks* researchers try to accomplish with multiverse analysis visualizations, the questions one can seek to answer, and the central goal related to each category (Section 5) and (3) classify multiverse visualizations into *archetypes*, assessing how well each *archetype* supports each *task*, their comparative limitations, key features and what role they can play in an analysis (Section 6). We close by discussing important design considerations surfaced by our survey—such as illusions of probability created by visualizing frequencies (Section 7.1) and the largely unmet need to support validation and interpretation of multiverses (Section 7.2)—as well as limitations and implications for future work (Section 7.5). For visualization researchers looking to develop multiverse analysis visualizations, our work provides a foundational set of tasks for subsequent tools and techniques to support; for practitioners of multiverse analysis, our work provides a mapping between tasks they wish to accomplish and archetypes they can use to accomplish them.

2. An Example of Multiverse Analysis: Are 'Female' Hurricanes More Deadly?

To make our discussion throughout the rest of the paper more concrete, we will be using the multiverse analysis by Simonsohn *et al.* [SSN19] as a running example. We introduce this example here. The terms **in bold** are from our proposed multiverse analysis terminology and will be defined more precisely in Section 3.

A 2014 study claimed that hurricanes whose names are female-gendered lead to more deaths, presumably because people do not take them as seriously as those with a male-gendered name [JSVH14]. However, later analyses of the same data called this finding into question [Mal14b, CC14, Mal14a]. It turns out that depending on how the analysis is carried out, it can be claimed that the data support the initial hypothesis, or the exact opposite. Simonsohn *et al.* [SSN19] conducted a multiverse analysis to investigate the space of possible analysis choices in more detail, and introduced the *specification curve* visualization (which we discuss in detail in Section 6.2) to better understand the influence of analytical decisions on outcomes.

The subject of the original study and its re-analyses is a dataset of hurricanes, with their name and information such as number of victims. The multiverse as set up by Simonsohn *et al.* [SSN19] focuses on two **outcomes**: (1) *extra deaths*, the number of extra deaths occurring for hurricanes with female names compared to those with male names, and (2) a *p*-value reflecting the degree to which those extra deaths are surprising. Ultimately, their question is whether or not there is a statistically significant effect of hurricane name gender on extra deaths (i.e. if $p < 0.05$). Any single analysis (a **universe**) gives rise to a specific number of extra deaths, a specific *p*-value, and a single answer to that question. The whole **multiverse analysis report** makes it possible to assess whether those results are **sensitive** to different ways of conducting the analysis. For example, one might handle outliers from the dataset in different ways: (i) do not exclude any hurricane; (ii) exclude the most deadly hurricane or (iii) exclude the two most deadly hurricanes. To reflect this, the multiverse has an *outliers* **parameter** that can take any of these **parameter values**. It also has other parameters, such as a *model* parameter for different model types that could be applied to the data, and a *femininity* parameter for different ways of operationalizing the gender of a hurricane name. Each universe is defined by a single combination of parameter values, which represents one unique way of analysing the dataset. Simonsohn *et al.* report 1728 universes, all produced by options they deemed to be reasonable.

If the outcomes of every universe were deemed to be practically equivalent, the multiverse analysis need proceed no further, and one could infer simply that any of the examined choices can be selected without impacting final conclusions. In contrast, Simonsohn *et al.* found the estimated number of extra deaths attributable to the gender of hurricane names to range from about -1 to $+12$ (mean of 1.63), while only 37 out of the 1728 universes (about 2%) yield $p < 0.05$. From those results, they concluded that the proposed relationship between the gender of hurricane names and their deadliness is not robust to defensible analytical choices, and thus should not be accepted as correct on the basis of this evidence alone.

3. Definitions of Key Concepts

In this section, we introduce definitions that will serve to outline the scope of our survey. These are stipulative [Pap64] and are not meant to be authoritative.

Central to our survey is our definition of a multiverse analysis report:

A **multiverse analysis report** is any statistical report that presents multiple analyses of the same raw dataset, which answer the same question, are reported with a similar level of detail, and whose purpose is to learn from—or communicate insights about—that dataset.

Our definition is consistent with the way the term *multiverse analysis* (without the word *report*) is used by Steegen and Gelman [STGV16], who first introduced it and defined it as '*performing the analysis of interest across the whole set of data sets that arise from different reasonable choices for data processing*' The only previous usage we know of this full term is in Dragicevic *et al.* [2019], though they do not explicitly define it. Our definition can be seen as a sharper version that more clearly distinguishes between multiverse analyses and related concepts.

Our definition has five key elements:

(1) *Any statistical report*: this includes any narrative describing the result of a data analysis, in any format, even though in this survey, we restrict ourselves to academic papers (see Section 4). Thus, the focus is on what is reported, not what is analysed. If multiple analyses are conducted but a single one is reported, as is commonly the case in empirical research [WVA*16b], then this cannot be considered to be a multiverse analysis report. Similarly, the process of building, selecting and tuning statistical models [DCCE19, MLMP17, CPCS19] is not within the scope of our definition, unless a report is written that uses multiple models to offer different perspectives on the same data.

(2) *Of the same raw dataset*: the multiple analyses must be carried out on the same raw dataset. Carrying out the same analysis on different raw datasets does not qualify as a multiverse analysis. Examples are (i) ensemble data analysis, where multiple simulations are computed with different parameter settings, and the results are summarized and analysed visually [WHLS18, SHB*14]; (ii) crowdsourced hypothesis testing, where multiple research teams conduct independent studies to answer the same research question [LJD*20] and (iii) meta-analysis [GHF14], except when multiple meta-analyses are performed on the same set of studies [DBH19]. If different raw datasets (e.g. different experiments in a study) are subjected to the same set of analyses, there are as many multiverse analyses as there are raw datasets. A multiverse analysis can, however, involve the analysis of different *processed* datasets, as long as they all arise from the same *raw* dataset (e.g. when collapsing the levels of a variable in different ways; see the DATAVERSE example in Drajicevic *et al.*[2019]). Resampling techniques (e.g. bootstrapping;

see the DANCE example in Drajicevic *et al.*[2019]) also generate multiple datasets from the same raw dataset, but in this survey, we do not consider them as multiverse analyses, because their goal is only to assess statistical uncertainty in the original raw dataset.

(3) *Answer the same question*: the multiple analyses need to answer the same question about the dataset. Statistical reports that use multiple analyses to answer different questions about a particular dataset (e.g. multiple subgroup analyses) do not qualify as multiverse analyses.

(4) *Similar level of detail*: the multiple analyses need to be reported with a similar level of detail. A detailed data analysis followed by a cursory mention of additional analyses (e.g. 'we redid the same analysis without outliers and obtained similar results') is not a multiverse analysis report. The outcomes from all analyses need to be reported with a similar level of detail. Similarly, a report that compares the goodness of fit of multiple statistical models but selects a single model to carry out the full data analysis does not qualify. However, we impose no lower limit on the number of analyses—a report with only two analyses can qualify as a multiverse analysis if the outcomes of both analyses are reported with a similar level of detail (e.g. [ESR17]). In addition, even if the analyses are heterogeneous in how they are conducted and reported (e.g. as in crowdsourced analyses [BKB*20, BNHC*20]), they still qualify as long as all outcomes are reported in a similar fashion.

(5) *With the intent to learn from [...] that dataset*: several analysis types do not qualify multiverse analyses, as they do not have the goal to learn from the raw dataset itself: Such examples that are not multiverse analyses include an evaluation of the coverage of different confidence interval procedures [Wah83], a sensitivity analysis carried out for model evaluation purposes [MLMP17] or an educational simulation illustrating how different analytical choices yield different outcomes [Fiv15]. Similarly, reporting multiple analyses with the intent to learn from—or communicate insights about—the analyses (not the datasets) would also not qualify. Furthermore, the entity that is expected to learn from the data must be a human. Thus, systems that learn from data by analysing it in many different ways (e.g. ensemble learning algorithms [SR18]) are excluded, unless they explicitly convey the multiverse to human users. As stated initially, the multiple analyses need to be *reported*.

Figure 2 shows a Venn diagram where each ellipse stands for one of the criteria from our definition of multiverse analysis report. One criterion is not shown (i.e. that all analyses must answer the same question). The diagram regroups the edge cases we previously mentioned, and which fulfil most—but not all—of the criteria. We emphasize such edge cases because they help clarify the boundaries of our definition, and can speed up the classification of reports into multiverse or non-multiverse.

For the purpose of this survey, we additionally introduce the notion of trivial multiverse analysis report:

A **trivial multiverse analysis report** is a multiverse analysis report with very few analyses and very little detail about each analysis, and which can be fully reported in the text without the need for tables or figures.

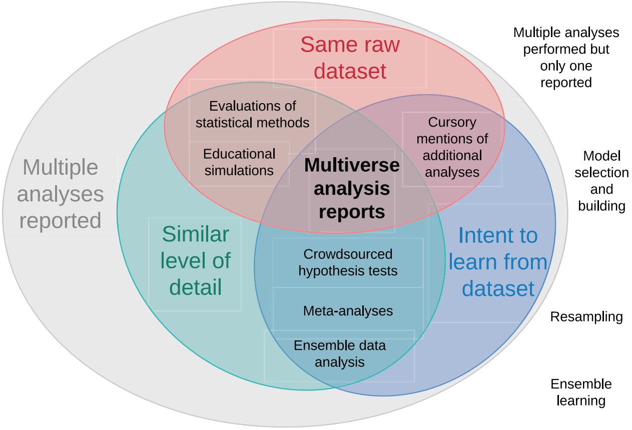

Same raw dataset
Multiple analyses performed but only one reported
Evaluations of statistical methods
Cursory mentions of additional analyses
Multiple analyses reported
Educational simulations
Multiverse analysis reports
Model selection and building
Similar level of detail
Intent to learn from dataset
Crowdsourced hypothesis tests
Meta-analyses
Resampling
Ensemble data analysis
Ensemble learning

Figure 2: *Overview of the four major criteria making up our definition of multiverse analysis report (each criterion is an ellipse), and examples of cases that fulfil some but not all criteria.*

An example of a trivial multiverse analysis report is a paper that reports a *p*-value after excluding outliers, and a *p*-value without excluding outliers. Such analyses formally meet our definition of multiverse analysis report but will be excluded from our survey nonetheless, because little can be gained from visualizing them.

We draw from previous work [2019] to define five basic elements that make up multiverse analysis reports, and which we will often refer to in this survey. In a multiverse analysis report:

A **universe** or **analysis** is one of the multiple analyses that are conducted and reported in the multiverse analysis report.

A **parameter** is a characteristic of the reported statistical analyses that varies across the multiverse.

A **parameter value** is a possible value taken by a parameter. A synonym is *option* [2019], but we use here the term *parameter value* for consistency with the rest of the terminology.

For example, suppose a paper uses three outlier exclusion methods to analyse data: (i) no exclusion; (ii) removing 3 standard deviations (SD) from the mean and (iii) removing 2 SD from the mean. Thus, *outlier exclusion procedure* is a parameter of the multiverse analysis, and this parameter has three possible parameter values, each defining a different analysis or universe.

Similarly, in a multiverse analysis report:

An **outcome** is a statistical result that is reported for all analyses in the multiverse.

An **outcome value** is a possible value taken by an outcome.

In the previous example, suppose the paper reports a point estimate and a *p*-value for the main effect size of interest, computed for each of the three outlier exclusion methods. In this case, the multiverse analysis reports two outcomes (a point estimate and a *p*-value), and a total of six outcome values (two per universe).

A primary goal of multiverse analysis is to assess outcome sensitivity and robustness:

Outcome sensitivity is the extent to which the values of an outcome vary across the multiverse.

Outcome robustness is the opposite of outcome sensitivity, i.e. it is the extent to which the values of an outcome are stable across the multiverse.

Now we can define our main focus of investigation, which is the multiverse analysis visualization:

A **multiverse analysis visualization** is any visual representation of the parameters, parameter values, outcomes and/or outcome values of multiple analyses in a multiverse analysis.

Visual representation means that at least some of the information is visually encoded [Mun14]. Thus, information conveyed exclusively via text and numerals (e.g. numerical tables) does not qualify, but hybrid representations that combine text or numerals with visual encodings (e.g. tabular visualization [PDF14]) qualify.

Finally, a last key concept central to this survey is the notion of *visualization archetype*:

A **visualization archetype** (or simply archetype) is a class of multiverse analysis visualization designs that convey information about specific multiverse entities (i.e. parameters, parameter values, outcomes and/or outcome values) using a specific combination of visualization idioms [Mun14].

A visualization archetype thus defines a family of visualization designs that encode the same type of information in (more or less) the same manner. For example, a histogram of *p*-values and a histogram of effect sizes belong to the same archetype because they are both histograms of outcome values. However, a histogram of outcome values and a histogram of parameter values belong to different archetypes because they do not encode the same type of information, despite using the same visualization idiom.

4. Methodology

Our goal was to understand:

1. What tasks or analytical questions do researchers aim to perform or answer when reporting a multiverse analysis visualization?
2. What multiverse analysis visualizations do researchers use, and how do these visualizations support those tasks?

To answer these questions, we curated a corpus of research articles. To be considered for inclusion into our corpus, each article had to contain at least one multiverse analysis report, as well as at least one multiverse analysis visualization. We performed a systematic analysis of our corpus, to (i) derive a task taxonomy for multiverse analysis visualization, (ii) identify a set of visualizations archetypes and (iii) analyse how well each archetype supports the tasks in our taxonomy.

Table 1: *Quantitative background on the corpus curation, including the number of search results per search type (serendipitous vs. systematic) and per search term, the number of papers that met our inclusion criteria, and the number of papers from the systematic search that were already in our initial corpus of seed articles.*

Search type — keyword	Search results	In final corpus	In both I. and II.
I. Serendipitous (seed articles)	53	36	12
II. Systematic:	>4893	19	12
— multiverse analysis	198	7	6
— specification curve	298	8	6
— vibration of effects	144	4	2
— crowdsourced analysis	264	3	2
— robustness analysis	>1000	0	0
— multimodel analysis	989	0	0
— perturbation analysis	>1000	1	0
— sensitivity analysis	>1000	0	0

4.1. Curating the corpus

Multiverse analysis reports are being used across a wide body of literature in many different areas of science. We addressed the challenge of reviewing such a heterogeneous body of literature using a two-step approach (Figure 3). We first collected articles in a serendipitous fashion during the conduct of other research or reading activity, through social networks, or suggested by recommendation systems like Mendeley. This resulted in 52 *seed articles*. Since there was no agreed-upon or widely used term to refer to the concept of multiverse analysis, we extracted the terms used by the article authors, resulting in eight terms (Table 1). We then used this list in a systematic literature search using the Google Scholar API through the *Publish or Perish* software [Har07] to find any documents with the terms appearing in the title, abstract or body text. We restricted the search to results published in 2015 or later, which for some keywords led to more results than the maximum of 1000 returned by the API. To keep the number of articles, we would need to analyse in detail manageable, we sorted each source list by the number of citations as counted by Google Scholar, and selected the first 20 items from each list. This led to 213 corpus candidates.

A second step consisted of checking whether each of the 213 corpus candidates was a research article. We replaced any item not passing this check with the next item from the respective source list. Using the definitions introduced in Section 3, we then checked for each of the 213 corpus candidates that it (1) included at least one multiverse analysis report, (2) was not of a trivial nature and (3) that the reported multiverse was visualized in some way. Thirty-six of the seed articles and 19 articles discovered through the systematic literature search passed these checks for a total of 43 articles, which form our final corpus (12 came up through multiple sources as detailed in Table 1).

More details on the corpus as well as the source lists from the systematic search are in the supplemental material.

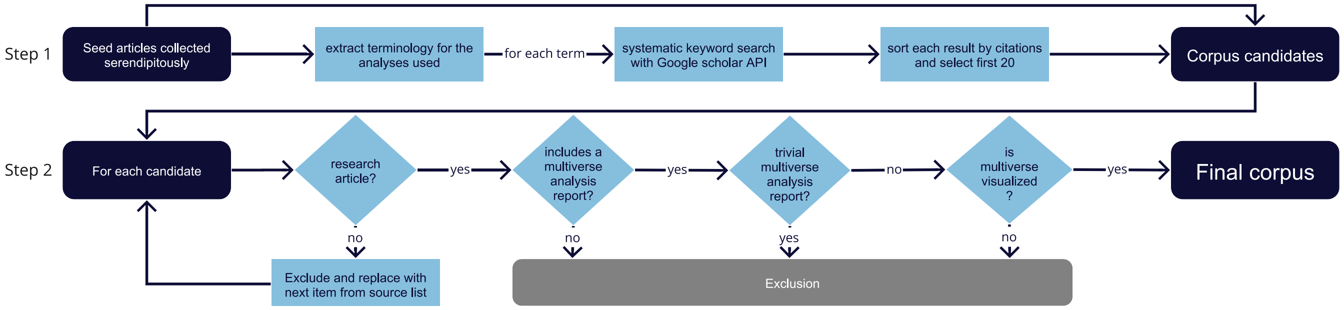

Step 1
Seed articles collected serendipitously
extract terminology for the analyses used
for each term
systematic keyword search with Google scholar API
sort each result by citations and select first 20
Corpus candidates
Step 2
For each candidate
research article?
yes
includes a multiverse analysis report?
yes
trivial multiverse analysis report?
no
is multiverse visualized ?
yes
Final corpus
no
no
yes
no
Exclude and replace with next item from source list
Exclusion

Figure 3: *Overview of our curation process. In step 1, we curated a corpus of candidates by combining serendipitously discovered articles with a systematic keyword search. In step 2, we analysed each candidate to identify all research articles that contain a non-trivial multiverse analysis report and illustrate that report with some form of visualization.*

4.2. Extracting tasks on multiverse analysis visualizations

To derive a taxonomy of the tasks researchers can perform with a multiverse analysis visualization, we performed a detailed analysis of parts of a subset of five articles in our corpus. These five were selected because their goal was to introduce a form of multiverse analysis as a *general method* rather than to use multiverse analysis to report specific findings [SSN19, STGV16, PBI15, YH17, SUM*18b]. Each paper analysed one or more datasets as a demonstration of the technique being introduced, as well as detailing reasoning and broader implications of their methodology.

For each of these articles, we extracted all figures that contained a multiverse analysis visualization, as well as any relevant text either present directly on the figure or in the figure caption. We also searched for all references to that figure in the article's main text and extracted all statements about the figure from the corresponding paragraphs, as well as the ones preceding and following it. Each captured passage was split into individual quotes, then copied onto the digital equivalent of sticky-notes in a collaborative whiteboard platform (Miro board, www.miro.com). Three authors conducted an affinity diagramming exercise to cluster the quotes into themes, which facilitated the identification of common tasks that could be performed using multiverse analysis visualizations. A selection of quotes relevant to each task is presented in Section 5.

Once all quotes from the initial articles were processed and a draft task taxonomy formed, we expanded and continued the analysis with additional articles from our corpus to ensure saturation was reached. Articles were chosen from reviewing the visualizations and discussion notes from our entire corpus, with a focus on selecting papers that were most likely to challenge our existing conceptions, judged from the distinctiveness of their associated visualizations and the topic of the articles themselves. The analysis of additional articles presenting interactive visualizations [2019, LKAH20] and theoretical considerations of multiverse analysis [DGGS20] inspired the definition of the last category added to our taxonomy (Validate, Section 5.5). Analysis of an additional set of seven articles [LKJA*19, ASGP18, PVB19, BKB*20, BNHC*20, Car12, ODT12], which featured distinctively different visualizations compared to the already included ones—and thus could likely challenge our task taxonomy—did not generate new tasks, categories or change our taxonomy structure.

We present the outcome of our task analysis in Section 5. The source material, including a PDF export of the Miro boards, can be found in the supplemental material.

4.3. Identifying visualization archetypes

To accomplish our second goal—identify multiverse analysis visualization archetypes and assess their capacity to support the tasks in our task taxonomy—we reviewed our full corpus of 43 articles, and extracted any figures and tables that initially appeared to satisfy our multiverse analysis visualization criteria. This resulted in a collection of 126 visualizations, which we trimmed so as to keep at least one representative figure for every distinct visual style present, as judged by all authors. The resulting set of 85 prospective archetypes was further reduced through closer review, with 16 being excluded as they were not actually multiverse analysis visualizations (e.g. visualizations of simulation studies, Sankey diagrams of a literature review), leaving 69 visualizations for further analysis.

To further distinguish between visualizations that supported different multiverse analysis tasks to some extent, from ones that only varied aesthetically, we conducted an in-depth iterative coding process. In each coding cycle, we picked one of the prospective archetypes, then reviewed the source paper. We then graded the visualization's support for each of the tasks in our taxonomy on a scale of 0–3 (as detailed below), assuming a multiverse of similar proportions than that featured in the visualization. In each cycle, if a visualization was found to be equivalent to a previously scored visualization, it was labelled to be a variant of the same archetype and excluded from re-scoring.

We defined a 0–3 grading system as: 0 = no support for this task; 1 = information required for task is present, but requires a large amount of effort or mental calculations, or supports the task minimally; 2 = tasks are sometimes well supported and sometimes not, depending on factors that naturally vary between multiverses; 3 = supports the task in a way that makes it reasonably fast and easy to complete, usually through clear visual features or explicit encoding of relevant information into distinct visual channels. All scores disregard the learning curve that may be required to use a visualization, and so adopt the perspective of a reader already familiar with that type of visualization. All scores were reviewed by at least two

Table 2: *Overview of the taxonomy for multiverse analysis tasks derived from the multiverse analysis visualizations in our corpus.*

Category	Task
Composition	**Composition** ▷ **Process**: understand the process that defines and creates the universes being considered. **Composition** ▷ **Parameters**: understand the definition and composition of universe parameters and parameter values.
Outcome	**Outcome** ▷ **Range**: assess range or spread of outcome values across all universes. **Outcome** ▷ **Frequency**: assess overall frequency of outcome values across all universes.
Connect	**Connect** ▷ **OutcomeRange**: connect parameters to outcomes by comparing similarity or range of outcome values across a subset of universes defined by a specific parameter value. **Connect** ▷ **OutcomeFrequency**: connect parameters to outcomes by comparing frequency of outcome values across a subset of universes defined by a specific parameter value. **Connect** ▷ **SpecificOutcomes**: connect parameters to outcomes by examining specific outcome values of interest and identifying parameter values that lead to those outcomes.
Connect Combinations	**ConnectCombo** ▷ **OutcomeRange**: connect combinations of parameters to outcomes by comparing range of outcome values across subsets of universes defined by parameter values. **ConnectCombo** ▷ **OutcomeFrequency**: connect combinations of parameters to outcomes by comparing frequency of outcome values across subsets of universes defined by parameter values. **ConnectCombo** ▷ **Idiosyncratic**: connect combinations of parameters to outcomes according to idiosyncratic patterns particular to a given visualization or analysis.
Validate	**Validate** ▷ **Metrics**: assess validity metrics of universes or compare metrics across parameter values. **Validate** ▷ **Details**: assess validity of universes by examining the underlying details of analyses in each universe to interrogate their validity.

authors after all visualizations were coded, with any disagreements resolved by discussion until consensus was reached.

The primary results of this analysis are reported in Section 6. The full set of visualizations reviewed and scored are available as supplemental material.

5. Taxonomy of Analysis Tasks

We identified twelve tasks that can be performed using a multiverse analysis visualization, summarized in Table 2. We organize these tasks into five analytical categories, with each category encompassing a general class of questions and goals that are common to most multiverse analyses. We denote each task definition as follows:

Category Name ▷ **Task Name**: definition of this task.

In text, we use the notation Category Name ▷ Task Name to refer to specific tasks. We have given the categories and tasks a logical ordering primarily to make them easier to describe and understand; this order does not necessarily reflect the order in which these tasks are carried out or reported. For each category, we provide an *Example question* based on our running example from Simonsohn *et al.* [SSN19] as well as sample quotes taken from the corpus that were used to identify and synthesize these tasks.

5.1. Composition: understand composition of the multiverse

Example question: What are the different methods used to exclude outliers in this multiverse?

Goal: Understand the components and processes that define and makeup this multiverse.

Tasks in this category can involve descriptions of the dataset source, how the data were processed, the included variables in the data, and what analytical choices are being considered (parameters and their parameter values). These tasks lay the groundwork necessary for the later sense-making process of drawing conclusions from the multiverse analysis. This category is unique in that it does not consider the outcomes of any analyses. We refer to this category as **Composition**.

In most published reports, this category of tasks is addressed solely through narrative descriptions, often in the form of lists in the text itself or as a table (see Figure 7). But as the composition of a multiverse grows in complexity, some authors choose to use visualizations to facilitate navigation and understanding of that complex structure. Two notable examples are the computation schematic of Patel *et al.* [PBI15] (Figure 13), and elements of the Boba interactive interface [LKAH20] (Figure 15).

Composition: Process. understand the process that defines and creates the universes being considered.

This task concerns the details and processes involved in creating individual universes, and thus the multiverse altogether. This can generally include data sources and data collection procedures, any processing of the data that is common to all universes, criteria for selecting outcomes of interest, and any other contextually relevant and important information of this kind.

For example, Patel *et al.* [PBI15] used the following narrative description to explain a few key steps in their process: *'First, we downloaded 417 self-reported, clinical and molecular measures with linked all-cause mortality information in participants from NHANES 1999–2004. [...] We chose variables of interest that had data on at least 1000 participants and at least 100 death events during follow-up'.* In that work, the authors both described the process in the text and illustrated the steps in a diagram—the computation schematic visualization (Figure 13).

Composition: Parameters. understand the definition and composition of universe parameters and parameter values.

This task involves understanding how parameters and parameter values included in the multiverse are defined, as well as how they can combine to form universe specifications. In the hurricane multiverse (Section 2), one parameter is *model*, with two parameter values: *negative binomial* and *log-normal*. In that multiverse, every combination of parameter values is considered valid, so there are no complex relationships between parameters and parameter values that need to be communicated. However, some multiverse analyses include more complex parameter contingencies, e.g. selecting one value for parameter A could render some available values for parameter B invalid. Communicating such relationships falls within the scope of this task as well.

5.2. Outcome: assess outcome sensitivity

Example question: Is the relationship between hurricane name genders and model-predicted fatalities stable across combinations of defensible analytical choices?

Goal: Assess the extent to which important outcomes vary among alternative analytical choices (sensitivity or robustness—see (see definitions in Section 3).

The topic of this category is the fundamental concern of multiverse analysis: If all considered analytical choices lead to effectively the same conclusions, then there is no need to proceed any further in the multiverse analysis. If outcomes are not sensitive, one can conclude that which of the considered choices one prefers does not matter, as the ultimate conclusions one would reach are the same regardless. For example, in the hurricane study, only 37 of the 1728 universes result in a p-value below 0.05, which indicates that some universes produce outcome values that differ substantially from the majority.

Importantly, how sensitive an outcome is depends upon context and expert judgement in the domain of the analysis. Assessing to what extent outcome values vary across a multiverse typically requires judgements of practical magnitude that are domain-dependent and subject to the analyst's interpretation. For example, if an analyst considers a certain range of effect sizes to be practically equivalent, then the effect size outcome is robust if it remains within that range. Similarly, if an analyst hinges their interpretation of p-values on a statistical significance threshold, then the p-value outcome is sensitive if, across universes, outcome values fall on both sides of that threshold.

Outcome: Range. assess range or spread of outcome values across all universes.

One way to assess outcome sensitivity is to examine the similarity (or spread, or range) of outcome values that occur within the multiverse, which is the goal of this task.

Simonsohn *et al.* [SSN19] describe the results of completing this task: '*The point estimates range from −1 to +12 additional deaths'.* Similarly, Steegen *et al.* [STGV16] write: '*for fiscal political attitudes … the remaining choice combinations lead to p values across the entire range from 0.05 to 1.0'.*

Outcome: Frequency. assess overall frequency of outcome values across all universes.

Another way to assess outcome sensitivity is by examining the frequency or proportion of specific outcome values that occur within the multiverse. However, there is more than one way to interpret outcome frequencies, which necessitates a nuanced consideration of this task.

The first interpretation of outcome frequency is *probabilistic*; i.e. treating frequencies as estimates of relative likelihood, with outcomes that occur in more universes deemed more plausible than ones that occur in fewer universes. For example, Simonsohn *et al.* [SSN19] state: '*researcher motivated to show a negative point estimate would be able to report twenty different specifications that do so, but the specification curve shows that a negative point estimate is atypical'.* Simonsohn et al. [SSN19] even introduce a technique for calculating a p-value of statistical significance for the multiverse as a whole, which treats the selection of analytical choices as a probabilistic sampling process.

Alternatively, a *possibilistic* interpretation of outcome frequency is illustrated in Steegen *et al.*[STGV16]: '*If no strong arguments can be made for certain choices, we are left with many branches of the multiverse that have large p values. In these cases, the only reasonable conclusion on the effect of fertility is that there is considerable scientific uncertainty. […] When only one choice is clearly and unambiguously the most appropriate one, variation [in outcomes] across this choice is uninformative'.* In other words, frequency information can indicate the possibility that something could be true, but cannot be used to determine what outcomes are more or less likely. The second part of this quote goes even further, implying that relative frequency of outcomes for some options should not be interpreted as encoding any relevant meaning.

Consideration for how a reader could, or should, interpret outcome frequencies is important for visualization design, as we suspect different visualizations may invite incorrect probabilistic interpretations. We discuss this issue further in Section 7.1. Note that this task is closely matched to what Amar *et al.* [2005] refer to as a 'characterize distribution' task.

5.3. Connect: connect parameters to outcome values to identify sources of sensitivity

Example question: Do some values within the 'dropping outliers' parameter lead to consistently larger outcome values of model-predicted fatalities?

Goal: Identify which analytical choices cause outcomes to differ across universes.

This category explores potential relationships between individual parameters, parameter values and outcome values. When outcomes have been determined to be sensitive to analytical choices (Section 5.2), one can seek to determine which choices produce this sensitivity. For instance, it could be that only some small subset of parameter values produces a divergent outcome, in which case one might wish to focus on critically analysing these few choices in greater detail. Further attention could either involve additional tasks described in this framework, or deeper theoretical considerations.

Connect: Outcome Range. connect parameters to outcomes by comparing similarity or range of outcome values across a subset of universes defined by a specific parameter value.

As with the previously described Outcome ▷ Range task, this task examines the similarity or overall range of outcome values within a multiverse, but with the added detail of conditioning (subsetting) on a parameter or parameter value. It is this additional point that allows for sources of sensitivity to be identified, and for the impact of different parameter values to be compared.

An example from Steegen *et al*. [STGV16] describes two parameters identified as not being the primary drivers of outcome sensitivity: '*The different exclusion criteria and cycle day estimation options do not seem to have a large impact on fluctuation in the statistical conclusion'.* In contrast, Silberzahn *et al*. [SUM*18b] describe the identification of two parameters that are sources of outcome sensitivity: '*The teams also varied in their approaches to handling the non-independence of players and referees, and this variability also influenced both median estimates of the effect size and the rates of significant results'.*

Connect: Outcome Frequency. connect parameters to outcomes by comparing frequency of outcome values across a subset of universes defined by a specific parameter value.

As with the previously described Outcome ▷ Frequency task, this task examines the frequency of outcome values, but now conditioned (subsetted) on a parameter or parameter value.

Silberzahn *et al*. [SUM*18b] compare the frequency of outcomes across the parameter *model form*: '*Fifteen teams used logistic models, and 11 of these teams found a significant effect [...] Six teams used Poisson models, and four of these teams found a significant effect'.* Steegan *et al*. [STGV16] use a more roughly estimated proportion: '*For religiosity [...] most data sets constructed under the second option for relationship assessment (R2) yield a nonsignificant interaction effect'.*

Connect: Specific Outcomes. connect parameters to outcomes by examining specific outcome values of interest and identifying parameter values that lead to those outcomes.

Another approach to identifying sources of sensitivity is to instead focus on specific outcome values, and find what parameter values produce them. This can be particularly important when some outcome values are more consequential than others, such as when some outcome values imply a therapeutic intervention is harmful.

Simonsohn *et al*. [SSN19] considered negative effect sizes in this way: '*[...] we can see that obtaining a negative point estimate requires a fairly idiosyncratic combination of operationalizations'.*

5.4. Connect combinations: connect combinations of parameters to outcome values to identify complex relationships that lead to sensitivity

Example question: Do the outcomes associated with the choice of model form strongly depend upon the choice of dropping outliers? In other words, do the parameters interact?

Goal: Identify which combinations of analytical choices cause outcomes to differ across universes.

In this category, the relationship between outcomes and analytical choices is further explored and characterized in ways that go beyond what was considered in category Connect (see Section 5.3).

The primary additional factor is considering combinations of parameters and parameter values. As a simplified example, if some model forms are more sensitive to outliers, then any parameter value related to excluding outliers could theoretically have a combined effect that would not be noticeable when examining the parameter values individually.

Connect Combo: Outcome Range. connect combinations of parameters to outcomes by comparing range of outcome values across subsets of universes defined by parameter values.

This task extends task Connect ▷ OutcomeRange by considering combinations of parameter values, rather than treating parameters as effectively independent from one another. While we primarily consider the combination of only two parameter values at a time, conceptually there is no reason that more complex relationships might exist with even more parameter values, just as in a traditional multivariate analysis. However, just as in traditional multivariate analysis, it is extremely difficult to cognitively and intuitively consider higher-order interaction effects, and a three-way interaction is the most complex relationship we have an example for in our corpus.

Steegen *et al*. [STGV16] describe a two-way interaction effect between parameters thusly: '*Using the third option for relationship status assessment (R3) leads to more fluctuation, depending on the choices for the other processing steps'.* In the report from Young *et al*. [YH17], the combined effect of two choices is a centrally important finding: '*Why do these estimates vary so much? Why is the distribution so non-normal? What combinations of control variables are critical to finding a positive and significant result? [...] In order to draw robust conclusions from these data, one must make a substantive judgement about two key modelling assumptions: the inclusion of race and marital status'.*

Connect Combo: Outcome Frequency. connect combinations of parameters to outcomes by comparing frequency of outcome values across subsets of universes defined by parameter values.

This task similarly extends task Connect ▷ OutcomeFrequency by adding the consideration of a combination of multiple analytical choices, with a focus on the relative frequency of outcomes.

Steegen *et al*. [STGV16] provide an example of this task where proportion is considered with rough approximations: '*The first and third options (R1 and R3) consistently lead to a significant interaction effect in combination with the first and second option for fertility* *assessment (F1 and F2) and to a nonsignificant interaction effect in combination with F5, whereas data sets constructed under R1 or R3 in combination with F3 or F4 lead to more fluctuating conclusions, depending on the other choices for data processing'.*

Connect Combo: Idiosyncratic. connect combinations of parameters to outcomes according to idiosyncratic patterns particular to a given visualization or analysis.

This task encompasses a variety of special relationships and patterns that are described throughout the corpus. These patterns are generally specific to certain visualizations, and we discuss these cases in greater detail in Section 6. However, as a brief example, we consider here the most commonly described concept of modality/multi-modality of the outcome value distribution.

In a univariate analysis, distributions can have one or more modes, which are the value(s) that occur most often in that distribution. When all outcome values from a multiverse are analysed as a distribution, there can be a single mode representing the value that the largest number of universes produce, or the distribution can be multi-modal. In Young *et al*.'s report [YH17], multi-modality is considered to possibly indicate that some parameter value, or combination of parameter values, are responsible for disparity of the outcome values. Having identified such parameter values, the authors state: '*In essence, there are two distinct modelling distributions to consider'*. This concept of modality is also described by Patel *et al*. [PBI15], referred to as 'modality in the Vibration of Effects', and is given an equivalent interpretation: '*W e observed three modes in the association between triglyceride levels and mortality […] The multimodal plots indicated that total cholesterol and diabetes were driving these modes'.*

5.5. Validate: validate the multiverse

Example question: Are all combinations of parameter values equally reasonable or defensible? For instance, does model fit, or other statistical diagnostic metrics, suggest one model type may provide more reasonable estimates?

Goal: Determine the validity, reasonableness, plausibility or defensibility of the multiverse overall.

This category is concerned with critically evaluating the validity of the constructed multiverse. Analytical choices and associated universes can be re-examined in light of additional insights gained from the multiverse analysis process itself. This can include examining model fits, statistical/predictive diagnostic criteria, re-evaluation of the handling of the underlying dataset or other investigation of individual universes or sets of universes.

Conducting an analysis can lead one to reconsider some of the decisions that were included in the multiverse, or to realize other parameters and parameter values should be considered as well. Early work in multiverse analysis, such as that of Simonsohn *et al*. [SSN19] and Steegen *et al*. [STGV16], primarily considered analytical choices that could be considered defensible prior to examining the data, or at least without using the data to evaluate the appropriateness of the analytical choices themselves. However, an analyst could reasonably come to question whether some outcomes should be given greater weight than others, which would mean that some universes are not considered equally defensible, even if they cannot be definitely excluded as inappropriate.

This category ultimately represents a stage of reflection that would ideally come before final interpretation of the multiverse analysis results. While conducting an analysis, this might lead one to reconsider some of the decisions that were included, or to realize other parameters and parameter values that should be considered. It could also suggest that some outcomes should be given greater weight than others, which would mean that some universes are not considered equally defensible, even if they cannot be definitely excluded as inappropriate.

This category and its associated tasks are described in a broader and less exacting way, as there were fewer examples of these tasks and visualizations to support them.

Validate: Metrics. assess validity metrics of universes or compare metrics across parameter values.

This task considers the validity of universes that make up the multiverse using some form of metric, such as model fit metrics. For example, some model types may produce better model fits overall, or the model fits may vary across parameter values. Model fit is the only specific example of this task we identified in the corpus, but other metrics could certainly be used for a similar purpose.

In Boba [LKAH20], support for this task is described: '*Do we have evidence that certain outcomes are less trustworthy? We toggle the colour-by drop-down menu so that each universe is coloured by its model quality metric […]. The large estimates are almost exclusively coming from models with a poor fit. We further verify the model fit quality by picking example universes and examining the model fit view […]. The visual predictive checks confirm issues in model fit, for example the models fail to generate predictions smaller than three deaths, while the observed data contains plenty such cases. […] We have reasons to be sceptical of the large estimates'.*

Validate: Details. assess validity of universes by examining the underlying details of analyses in each universe to interrogate their validity.

This general task is about investigating universes in a level of depth that may be more typical with traditional analyses, but which is difficult to do with an entire multiverse. This task is instead concerned with diving into either single universes or small sets of universes in greater detail, to allow for the richness and detail of a traditional analysis to be able to inform the construction and assessment of validity of the multiverse overall.

This task has some degree of limited support in a few different visualizations, but the primary source for the identification of this task is from the Explorable Multiverse Analysis Reports (EMARs) [2019], an interactive media where balancing depth and richness with the comprehensiveness of a multiverse analysis is a primary design goal; for example: '*four aspects of the analysis can be changed by the reader, which has the effect of immediately updating the two plots and some text elements such as explanations and figure captions'.* While the technique was not designed or explicitly

	Name	Section	Icon	Composition ▹ Process	Composition ▹ Parameters	Outcome ▹ Range	Outcome ▹ Frequency	Connect ▹ Outcome Range	Connect ▹ Outcome Frequency	Connect ▹ Specific Outcomes	Connect Combo ▹ Outcome Range	Connect Combo ▹ Outcome Frequency	Connect Combo ▹ Idiosyncratic	Validate ▹ Metrics	Validate ▹ Details	References
Archetypes	Outcome Histogram	6.1		0	0	3	3	0	0	0	0	0	0	0	0	[STGV16] [DGGS20] [PVB19] [BI16] [DMH∗18] [VKT19] [BI16] [CT16]
	Outcome Curve	6.2.1		0	0	3	2	0	0	0	0	0	0	0	0	[SSN20] [Coo18] [JKN18] [VKT19] [BYO19] [OP19a] [DS18] [SSN19]
	Universe Specification Panel	6.2.2		0	2	0	0	0	0	0	0	0	0	0	0	[SSN20] [HCM13] [GHF14] [SUM∗17]
	Descriptive Specification Curve	6.2.3		0	2	3	2	3	2	3	2	1	3	0	0	[SSN20] [JKN18] [OP19a] [OP19b] [ODP19] [DGGS20] [BRRYD20] [VKT19] [RES17]
	Outcome Density Plot	6.3		0	1	3	3	2	2	1	2	2	3	0	0	[YH17] [LKJA∗19] [HS06] [ODT12] [You18] [MY18]
	Vibration of Effects Plot	6.4		0	0	3	2	2	2	1	1	1	3	0	0	[PBI15] [DGGS20]
	Outcome Matrix	6.5		0	1	3	2	2	2	3	2	2	3	0	0	[STGV16] [CJT19] [DGH∗18] [DKBK19] [DS18]
	Multiverse Computation Schematic	6.6		3	3	0	0	0	0	0	0	0	0	0	0	[PBI15]
Systems	Explorable Multiverse Analysis Reports	6.7.1		0	2	1	1	1	1	1	1	1	0	0	3	[DJS∗19]
	Boba	6.7.2		3	3	3	3	3	3	3	1	1	3	3	0	[LKAH20]

Figure 4: *Overview of the archetypes and interactive systems described in Section 6. Shaded cells indicate how well an archetype or system supports an analysis task in our taxonomy, on a scale of 0 (not supported) to 3 (fully supported).*

described with the goal of examining the validity of a multiverse, it is one of few multiverse visualizations that demonstrate how this task might be supported.

6. Multiverse Visualization Archetypes and Systems

We describe the set of multiverse visualization archetypes—families of similar visualizations designs-identified in our analysis, along with the tasks they support. We also discussed two interactive visualization systems designed to support multiverse analysis. A visual summary is shown in Figure 4 (see definition in Section 3, and process in Section 4).

6.1. Outcome histogram

The *outcome histogram* conveys the frequency of the different outcome values that occur within a multiverse for a particular outcome, so that each individual universe outcome value is counted once. In Figure 5, the *x*-axis encodes the outcome values (here, point estimates of extra deaths for female hurricanes in the example of Section 2), while the *y*-axis encodes the number of times binned outcome values occur within the multiverse. The dotted line serves as a visual aid to highlight the effect size of zero, which can be interpreted in the context of our running example as implying that there is no net effect of hurricane name femininity on predicted fatalities.

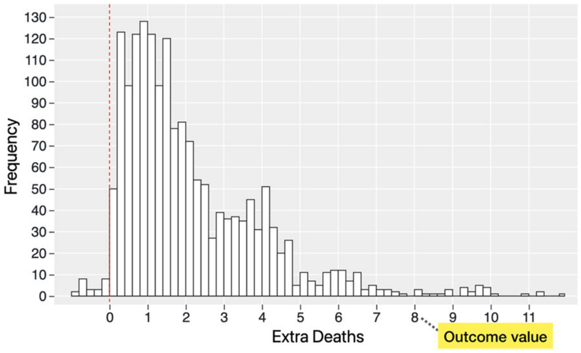


Figure 5: *Example of an outcome histogram. Recreated after Steegen* et al. *[STGV16], but using the hurricane dataset (Section 2). The* x*-axis encodes outcome values (effect size estimates), while the* y*-axis shows the count across the multiverse.*

The outcome histogram allows a viewer to easily and simultaneously complete both outcome (Section 5.2) tasks: Outcome ▷ Range and Outcome ▷ Frequency. This is made possible because both the full range of outcome values, as well as their proportions, are explicitly encoded in the plot. For instance, Figure 5 allows to identify that the most common outcome values are near zero, and that there are also many more results above zero than below it. One can also see that the the positive effect sizes go to greater magnitudes than the negative ones (+12 vs. −1). However, with no mapping of parameters and options to outcomes, the viewer cannot explore which analytical choices are responsible for this variation.

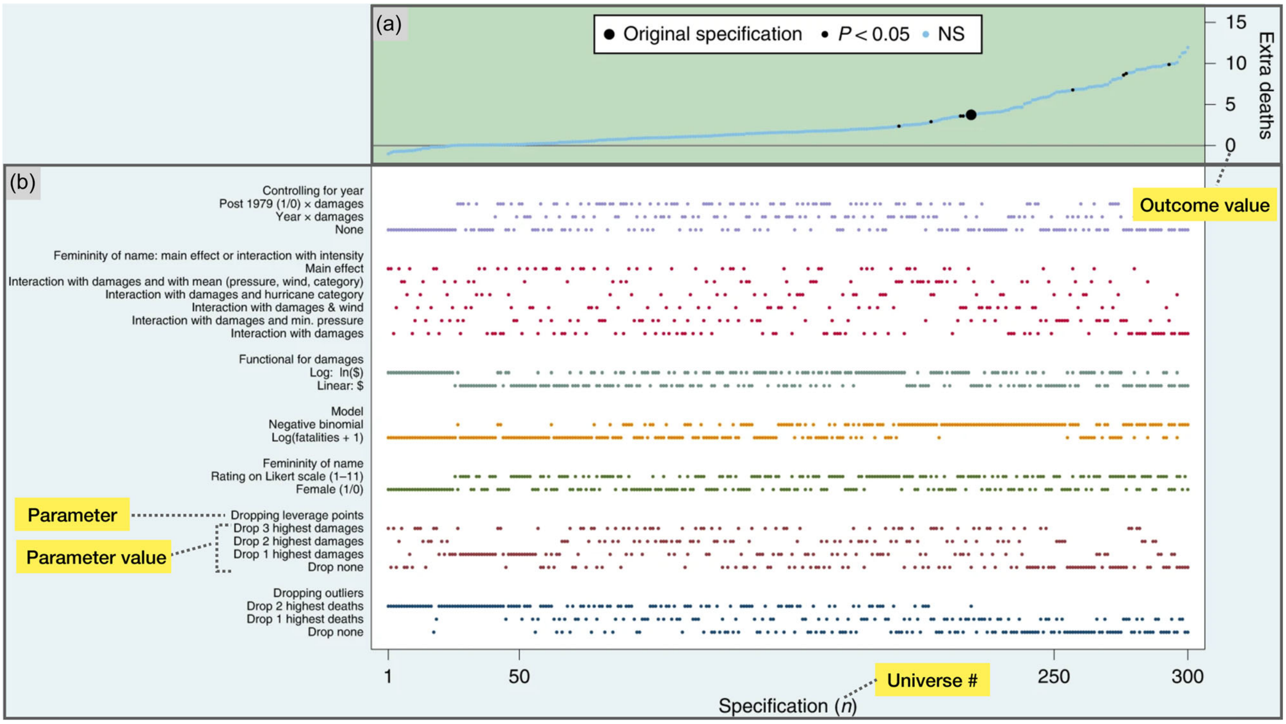


Figure 6: *Example of a specification curve [SSN19]. We treat the full figure as a composite, made up of two components: (a) an outcome curve, (b) an universe specification panel. The composite visualization has super-additive functionality, enabling tasks that neither component supports by itself. Three hundred universes are shown here, out of the full multiverse of 1728. The 50 universes with the smallest and largest outcome values are shown, along with a random sample of 200 other universes.*

Though frequency is a fundamental feature of the outcome histogram, and Steegen *et al*. themselves were clearly aware of the dangers of a probabilistic interpretation (as described in Section 5.2), under a strictly possibilistic interpretation, the existence of even one seemingly valid universe with a given outcome value is evidence that outcome cannot be ruled out. This suggests a potential issue with this (and other) frequency-based encodings: they may invite unintended or incorrect interpretations of multiverse outcomes. We discuss this further in Section 7.1.

The outcome histogram is a general approach that we encountered frequently in our corpus, e.g. [STGV16, DGGS20, PVB19, BI16, DMH*18, VKT19]. We also note one variation where the outcome is a *p*-curve [BI16], while Cirillo *et al*. reported multiple varieties of this type [CT16].

6.2. Descriptive specification curve

The *descriptive specification curve* is an example of a *composite visualization*—a visualization that is made up of two or more linked components, each of which could individually function as standalone visualizations on their own. Some composites feature *super-additive functionality*, which is when a composite visualization supports more tasks than all of the individual components considered separately, and this archetype is the primary example of this concept. Note that the term *specification curve* has been ambiguously used in the literature to refer to a multiverse analysis, the full composite (Figure 6), or just the top panel (Figure 6(a)). Following Simonsohn *et al*. [SSN20], we use *descriptive specification curve* to refer to the full composite. We first review each component individually before discussing the composite.

6.2.1. *Outcome curve (component)*

The core component of the descriptive specification curve is the *outcome curve* (Figure 6(a)). The *y*-axis encodes the outcome values (here, extra deaths), and universes are sorted along the *x*-axis according to outcome value, giving this visualization its distinctive shape. In the design of Simonsohn *et al*. [SSN20] shown in Figure 6, dot colour encodes a second outcome (black for statistically significant and blue for non-significant). In addition, due to limited horizontal space and the large size of their multiverse, the authors chose to only display a subset of the 1728 universes: Only those with the top and bottom 50 outcome values are shown, along with 200 other randomly sampled universes.

This visualization supports the same two tasks as the histogram of outcomes, i.e. Outcome ▷ Range and Outcome ▷ Frequency. The outcome curve resembles a cumulative distribution function (CDF) with the axes swapped. Because frequency is not explicitly encoded, the task Outcome ▷ Frequency is more difficult and less precise, especially when values being compared are not adjacent.

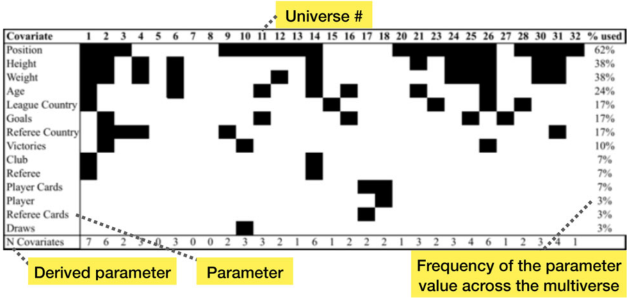
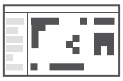

Universe #
Covariate 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 20 21 23 24 25 26 27 28 30 31 32 % used
Position 62%
Height 38%
Weight 38%
Age 24%
League Country 17%
Goals 17%
Referee Country 17%
Victories 10%
Club 7%
Referee 7%
Player Cards 7%
Player 3%
Referee Cards 3%
Draws 3%
N Covariates 7 6 2 3 0 3 0 0 2 3 3 2 1 6 1 2 2 2 1 3 2 3 4 6 1 2 3 4 1
Derived parameter
Parameter
Frequency of the parameter value across the multiverse

Figure 7: *A variant of a universe specification panel [SUM*18a]. Each column is a team of analysts (i.e. a universe) having analysed the same dataset using different analytical choices, as defined by black cells indicating the selection of parameters values. The bottom row indicated the number of parameter values in each universe, and the rightmost column indicates the frequency of a given parameter value across the sparse multiverse.*

The outcome curve is commonly presented as a stand-alone visualization, e.g. [Coo18, JKN18, VKT19, BYO19, OP19a, DS18], especially in papers explicitly reporting a *specification curve analysis*. Simonsohn *et al*. [SSN19] include three examples of the curve presented alone, with only one example of the full descriptive specification curve.

6.2.2. *Universe specification panel (component)*



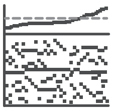
The second component of the descriptive specification curve is the *universe specification panel* (Figure 6(b)). It consists of a tabular visualization [PDF14] where columns are individual universes and rows are parameter values clustered by parameter. Columns may be sorted by outcome value, although outcome values themselves are not shown in this component. A cell in this table indicates when a universe (column) includes a given parameter value (row) in its specification. As this visualization shows no outcome values, it only supports task Composition ▷ Parameters.

Figure 7 shows a variant of this archetype, designed for sparse multiverses, i.e. where not every combination of parameter values is used. The plot indicates on the far-right column how many of the universes has each parameter value enabled. In this example, columns are also sorted by the number of covariates included in the analysis performed by a team.

Note that the number of covariates is not a free parameter, but is instead a function of other parameter values (which would be, for example, one parameter per covariate that indicates if it was used in the analysis). We refer to this as a **derived parameter**. Derived parameters can be visualized the same as any other parameter.

There are a number of other examples of this archetype in our corpus, e.g. [HCM13, GHF14], but all have equivalent task support.

6.2.3. *Descriptive specification curve (composite)*



Combined together on a common x-axis, the components above form the full composite *descriptive specification curve* (Figure 6), which allows the viewer to connect outcome values to analytical choices.

The composite supports all the tasks that its individual components support, but also supports all tasks in the Outcome and Connect categories (Sections 5.2 and 5.3). Consider, for instance the *dropping outliers* parameter: at the bottom of the specification panel (Figure 6(b)), the eye is drawn to a continuous pattern of dark blue dots indicating that the *drop 2 highest deaths* parameter value (i.e. exclude the two deadliest hurricanes Katrina and Audrey) leads to the all of the lowest outcome values. The viewer can read up to see that all of the outcome values below zero are associated with this parameter value (Connect ▷ OutcomeRange). Alternatively, if the viewer were interested in outcome values below zero, they could have started in the Outcome Curve (Figure 6(a)) and read down (Connect ▷ SpecificOutcomes), leading to the same observation, with other similar patterns observed for the *controlling for the year* parameter value *none* (purple), or *model* parameter value *log(fatalities +1)* (yellow).

The tasks ConnectCombo ▷ OutcomeRange and ConnectCombo ▷ OutcomeFrequency can be completed in the same manner, but with less ease because columns that satisfy a combination of more than one parameter values (e.g. *controlling for year = year × damages* and *feminity of name = rating on Likert scale (1–11)*) are not clustered together, making it difficult to identify whether the corresponding outcome values exhibit any particular pattern.

This visualization also enables identification of Simonsohn *et al.* termed *idiosyncratic specifications* (ConnectCombo ▷ Idiosyncratic), e.g. pointing out that only a particular, small subset of the available parameter values lead to negative effect sizes. We discuss such interpretations of outcome frequencies in more depth in Section 7.1.

6.2.4. *Variants of the descriptive specification curve*

Figure 8 shows notable variants featuring interesting adaptations and improvements. In Figure 8(a), statistical significance is colour-coded on both the outcome curve and the universe specification panel (red is significant), and standard error is shown using an error band around the outcome values. This places more visual emphasis on statistical significance and confidence within each universe. Figure 8(b) maps significance to colour but uses a three-colour scheme that also indicates the sign of the effect.

Figure 8(c) also uses an error band and a different three-colour scheme for statistical significance (blue: $\alpha = 0.05$, red: $\alpha = 0.10$, and black: non-significant). Note also that the columns in this variant are the result of a depth-first sorting across the parameter values. This makes some tasks in Connect combinations (Section 5.4) easier compared to Figure 6 (so long as the desired combinations of

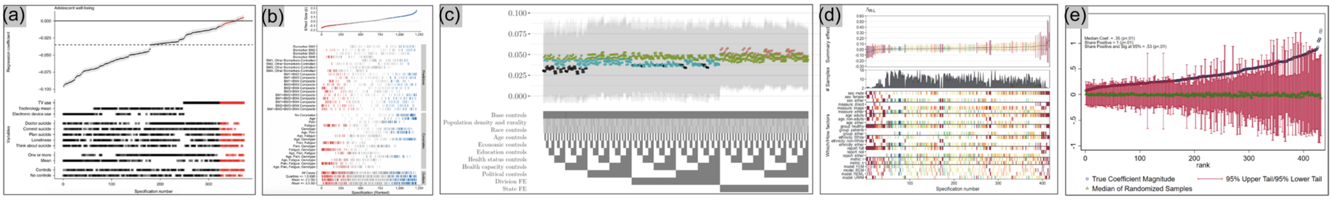


Figure 8: *Example variants of the specification curve archetype, notable for their alternative mappings of the colour channel and integration of uncertainty quantification metrics. (a) Figure 1 from Orben* et al. *[OP19a], (b) Figure 5 from Del Giudice* et al. *[DGGS20], (c) Figure 7 from Burstyn* et al. *[BRRYD20], (d) Figure 2 from Voracek* et al. *[VKT19], (e) Figure 5 from Jelveh* et al. *[JKN18].*

parameter values are clustered together) while making tasks in Connect (Section 5.3) more difficult (by disrupting the sorting within single parameter values).

Figure 8(d) presents a multiverse of meta-analyses, where each universe is one meta-analysis. The number of studies within each universe is colour-coded (red = 2, blue = 18), and plotted as a frequency plot as an additional middle panel. More generally, this is mapping an additional outcome variable onto colour in the specification curve. This allows a task specific to multiverse meta-analysis (ConnectCombo ▷ Idiosyncratic): reasoning about the validity of individual universes based on the number of studies included in their meta-analyses.

Figure 8(e) is a variant of the outcome curve component (stand alone). It uses confidence intervals around a bootstrapped null distribution instead of around the outcome value, but is otherwise similar to other variants that use error bands.

While not strictly variants of this archetypal family, the standard forest plot, e.g. Arslan's Figure 4 [ASGP18], and dot-interval plot, e.g. Silberzahn's Figure 2 [SUM*18b] could be considered as ancestors of the outcome curve, and have some similar visual features and functionality, though to show only a very small number of universes. See the supplemental material for more detail, including a number of other examples of this archetype [OP19a, OP19b, ODP19, BRRYD20, DGGS20, RES17, VKT19].

6.3. Outcome density plot

The *outcome density plot* shows the distribution of outcome values as a density plot. In Figure 9, the outcomes of a multiverse analysis examining potential racial and gender bias in a mortgage-lending dataset are shown. The parameters in this universe indicate whether a specific variable (such as a mortgage applicant's race, marital status) was included as a covariate in a statistical model. The x-axis encodes outcome values of estimated effect size, while the y-axis encodes the relative proportion of universes with the associated effect size.

While similar in function to the outcome histogram, this archetype splits the multiverse into two distribution lines (blue and red) corresponding to two different subsets of the multiverse defined by chosen parameter values. This allows it to support additional task categories, Connect (Section 5.3) and Connect combination (Section 5.4), by isolating subsets of the parameter space of interest. This also means that these tasks are only supported for the particular parameter value(s) or subsets that are directly encoded. While one can easily imagine plotting more than two curves in one plot, it can quickly become cluttered. See the supplemental material for more examples of this archetype [LKJA*19, HS06, ODT12, You18, MY18].

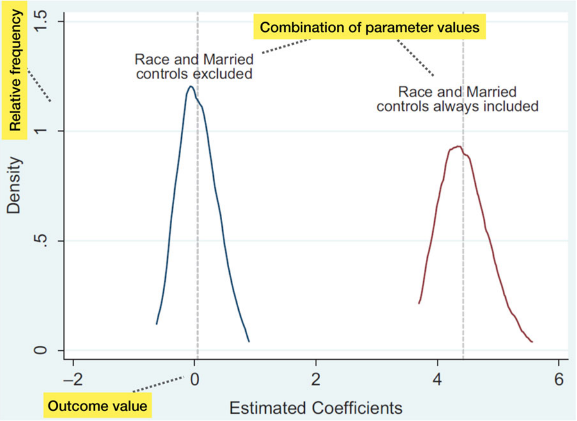


Figure 9: *Example of an outcome density plot, from Young* et al. *[YH17]. Here, each density curve represents the relative frequency of outcome values across a subset of universes, defined by combinations of parameter values.*

The limited scalability of this archetype in terms of the number of parameters that can be supported is emblematic of an important tradeoff in multiverse visualization design: Some visualizations are better for identifying the source of sensitivity in a multiverse overall, while visualizations like the outcome density plot can effectively show the sensitivity of a small selection of parameters after having identified them by other means.

Multi-modality in a density curve of outcome values may indicate that a small subset of parameter values, or combination of parameters, are especially important as they are uniquely responsible for widely different outcome values. As an example of task ConnectCombo ▷ Idiosyncratic, Young *et al.* [YH17] identify variables for race and marital status as being especially important in their study, and use Figure 9 to illustrate the effect of these decisions on outcome sensitivity. The distribution is multi-modal: All outcome values are close to zero (left curve) or they span large positive effect sizes (right curve). The importance of modality of the outcome distribution is also emphasized in vibration of effects plots (Section 6.4).

6.4. Vibration of effects plot

Figure 10 depicts a multiverse analysis concerned with the reliability of hazard ratios (an effect size) associated with various health factors, like blood levels of vitamin D. The parameters in this analysis are thirteen covariates that can individually be included or excluded, resulting in 8192 universes. In this *vibration of effects plot* (also called a *volcano plot* by Patel et al. [PBI15]), the effect size is plotted on the x-axis and statistical significance is plotted on the y-axis of a scatter plot with density contour lines. Some other variants in Patel *et al*. [PBI15] use 2D binned heatmaps instead of scatterplots.

All tasks in Outcome (Section 5.2) are well-supported by this plot to the extent that density contours and overplotted scatterplots support frequency estimation. All tasks in Connect (Section 5.3) are supported with comparable ease, and in much the same way, as the Outcome density plot (Section 6.3). Similar caveats apply: generally only a small set of combinations of parameter values can be compared at once, e.g. by mapping parameters to colours (Figure 11). However, the 2D density of statistical significance and effect size may allow additional clusters of outcomes to be visible that would not be visible in a 1D density chart, potentially aiding identification of interesting clusters of parameter values.

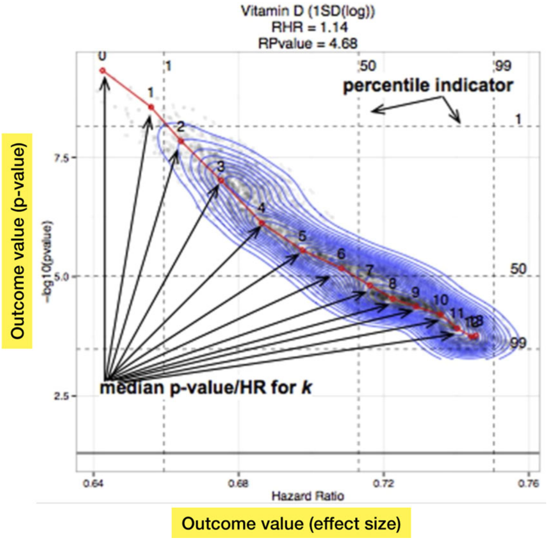


Figure 10: *Example of a vibration of effects plot [PBI15]. The* x*-axis encodes outcome values (effect size estimates), and the* y*-axis encodes the statistical significance (negative log transform of* p*-value). Blue contour lines are used to show the relative frequency of outcomes within the multiverse.*

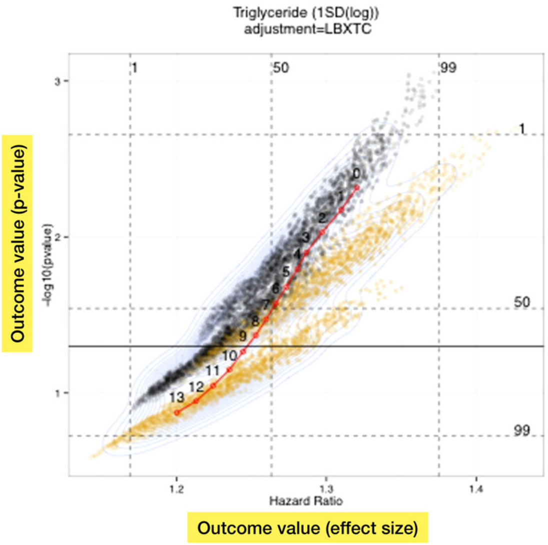
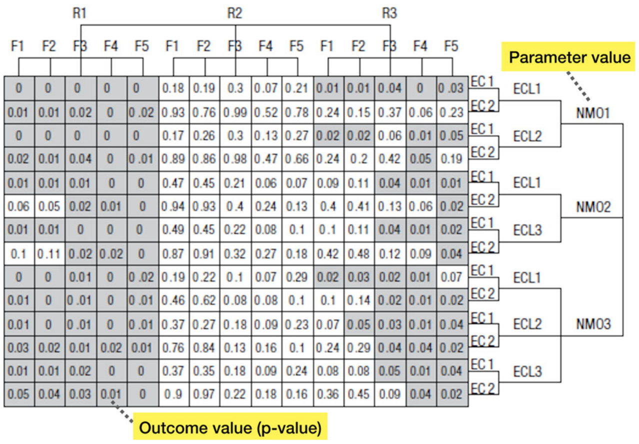


Figure 11: *Variant of the vibration of effects plot [PBI15] where parameter values for a given parameter (inclusion vs. exclusion) are colour-coded.*

The identification of potentially important clusters in outcome values is an example of the task ConnectCombo ▷ Idiosyncratic. Patel *et al*. [PBI15] dedicate extensive discussion of visual patterns exhibited by vibration of effects plots and their interpretation. For example, while the colour coding of parameter values in Figure 11 shows this parameter is part of the cause of multimodality in outcomes, there are still at least two visually distinct regions within the outcomes associated with this parameter. This suggests that this parameter is not the only cause of multimodality, and that there may be an interaction with another parameter. This ability to identify interaction effects is a unique feature of this archetype, though identifying what specific parameters are responsible (ConnectCombo ▷ OutcomeFrequency or ConnectCombo ▷ OutcomeRange) requires creating additional charts—Patel *et al*. [PBI15] describe how hundreds of such figures are to be generated to this end.

Patel *et al*. [PBI15] also describe many idiosyncratic visual patterns and corresponding relationships that can be identified with vibration of effects plots. As an example, outcomes may form a U-shape around 0, which indicates that there are universes that show opposite effect sizes, which Patel *et al*. call the *Janus effect* (after the Roman god with two faces). Other patterns feature when all universes had the same direction of effect, but disagreed only on magnitude or statistical significance of the effect.

Another common feature of vibration of effects plots is the red line with numerically labelled points, where each points is the median outcome value of all universes with the corresponding number of covariates included (a *derived parameter* as defined in Section 6.2.2). This allows identification of patterns concerned with the joint combination of effect size, statistical significance and the number of covariates used (ConnectCombo ▷ Idiosyncratic). For example, Patel *et al*. reported finding cases where more adjusting variables were associated with smaller effect sizes, larger effect sizes



			R1					R2					R3				
			F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
NMO1	ECL1	EC1	0	0	0	0	0	0.18	0.19	0.3	0.07	0.21	0.01	0.01	0.04	0	0 .03
		EC2	0.01	0.01	0.02	0	0 .02	0.93	0.76	0.99	0.52	0.78	0.24	0.15	0.37	0.06	0.23
	ECL2	EC1	0	0	0	0	0	0.17	0.26	0.3	0.13	0.27	0.02	0.02	0.06	0.01	0.05
		EC2	0.02	0.01	0.04	0	0 .01	0.89	0.86	0.98	0.47	0.66	0.24	0.2	0.42	0.05	0.19
NMO2	ECL1	EC1	0.01	0.01	0.01	0	0	0.47	0.45	0.21	0.06	0.07	0.09	0.11	0.04	0.01	0.01
		EC2	0.06	0.05	0.02	0.01	0	0.94	0.93	0.4	0.24	0.13	0.4	0.41	0.13	0.06	0.02
	ECL3	EC1	0.01	0.01	0	0	0	0.49	0.45	0.22	0.08	0.1	0.1	0.11	0.04	0.01	0.02
		EC2	0.1	0.11	0.02	0.02	0	0.87	0.91	0.32	0.27	0.18	0.42	0.48	0.12	0.09	0.04
NMO3	ECL1	EC1	0	0	0.01	0	0 .02	0.19	0.22	0.1	0.07	0.29	0.02	0.03	0.02	0.01	0.07
		EC2	0.01	0	0 .01	0	0 .01	0.46	0.62	0.08	0.08	0.1	0.1	0.14	0.02	0.01	0.02
	ECL2	EC1	0.01	0	0 .01	0	0 .01	0.37	0.27	0.18	0.09	0.23	0.07	0.05	0.03	0.01	0.04
		EC2	0.03	0.02	0.01	0.02	0.01	0.76	0.84	0.13	0.16	0.1	0.24	0.29	0.04	0.04	0.02
	ECL3	EC1	0.01	0.01	0.02	0	0	0.37	0.35	0.18	0.09	0.24	0.08	0.08	0.05	0.01	0.04
		EC2	0.05	0.04	0.03	0.01	0	0.9	0.97	0.22	0.18	0.16	0.36	0.45	0.09	0.04	0.02

Figure 12: *Example of an outcome matrix [STGV16]. The double-dendrogram structure encodes parameter specification: Each level is a parameter, and each node at a given level is a parameter value. Each cell in the matrix thus corresponds to a universe, and indicates the outcome value for this universe (also colour-coded).*

and cases where the effect size appeared to have no dependence on this parameter.

Overall, this archetype represents an effective way of getting an overview of the outcome of a multiverse where two outcome metrics are jointly important. The only other example we found of this archetype was in del Guidace *et al*. [DGGS20], but this was a near-exact reproduction of the style of this archetype that differed primarily in colour choice.

6.5. Outcome matrix

An outcome matrix is a tabular visualization [PDF14] where both rows and columns are parameter values, and each cell reports an outcome value. In Figure 12, each cell reports a p-value, both using numerals and a colour (statistically significant in grey). In this figure, the axes are dendrograms where each level of the tree is a parameter and each branch a parameter value, thus a path through the tree shows the combinations of parameter values defining each universe. Insofar as the size of the tree is able to scale to the size of multiverse, there is good support for Composition ▷ Parameters in that the structure and relationships within and between parameter values can be derived easily.

Figure 12 is an example of the *outcome matrix* from Steegen *et al*. [STGV16], a work of the authors who coined the term *multiverse analysis* itself. They chose to visualize their analyses both with this archetype (the *outcome matrix*) and the previously described *outcome histogram* (Section 6.1). They examined data that explored the relationship between human fertility and religious and political attitudes, across a multiverse defined by data exclusion and operationalization parameters. The outcome of interest is a p-value.

The colour coding of the outcome values supports Outcome ▷ Frequency (here, the more grey cells the more occurrences of a significant outcome). Outcome ▷ Range for other types of outcomes (e.g. effect size) could be supported given a more granular colour coding, although known issues with heatmaps may make certain tasks difficult [2020, GW12].

Tasks in Connect (Section 5.3) are generally well supported, with a few qualifiers. Tasks Outcome ▷ Range and Outcome ▷ Frequency are relatively easily accomplished when the specified parameter is at the top of the hierarchical axis (e.g. *R1* Figure 12 spans five adjacent columns), but require more mental effort otherwise as all the relevant universes are not found within adjacent columns or row (e.g. *F1* spans three non-adjacent columns). The ease of connecting specific outcomes to parameters (Connect ▷ SpecificOutcomes) depends on the hierarchical structure of the parameters as it impacts how outcome values cluster with parameter values: In Figure 12, one can easily observe that all significant p-values are in R1 and R3, but if the axes were ordered differently (e.g. swap the order of the R and F parameters), or if the viewer was interested in a more specific outcome value, the task can become difficult. Similarly, ConnectCombo ▷ OutcomeRange and ConnectCombo ▷ OutcomeFrequency may be well-supported for some combinations of outcome values and axis orderings, making this one of the few visualizations that can support these tasks (at least in some cases). However, the difficulty of all of these tasks depends heavily on row and column ordering and the resulting clusters, as with matrix visualization in general [BBHR*16].

6.5.1. *Variants of the outcome matrix*

Variants of the outcome matrix in our corpus were generally less structurally complex than the example shown in Figure 12, as they omitted the use of a hierarchical axis on either columns or rows. Multiple examples used only one axis to represent parameters, while the other axis was used to show outcomes of interest [CJT19, DKBK19, DS18]. Multiple variants used continuous outcomes and applied different colour maps (e.g. diverging palette for positive-negative effect and magnitude), illustrating how this archetype is not fundamentally limited to binary outcomes types [DGH*18, DS18].

6.6. Multiverse computation schematic

Figure 13, also from Patel *et al*. [PBI15], is an example of the *multiverse computation schematic* archetype. This is one of the few archetypes whose focus is on Composition (Section 5.1)—as opposed to reporting outcome values—providing the most support for the tasks Composition ▷ Process and Composition ▷ Parameters in our corpus.

Each panel of Figure 13 denotes a single major stage of the analysis pipeline for creating this multiverse analysis. Panel (a) describes the data source and panel (b) describes the dependent variable in the analysis. Supporting Composition ▷ Parameters, panel (c) lists parameters (here, parameter values are either include or exclude) and panel (d) describes the statistical model used to produce out-

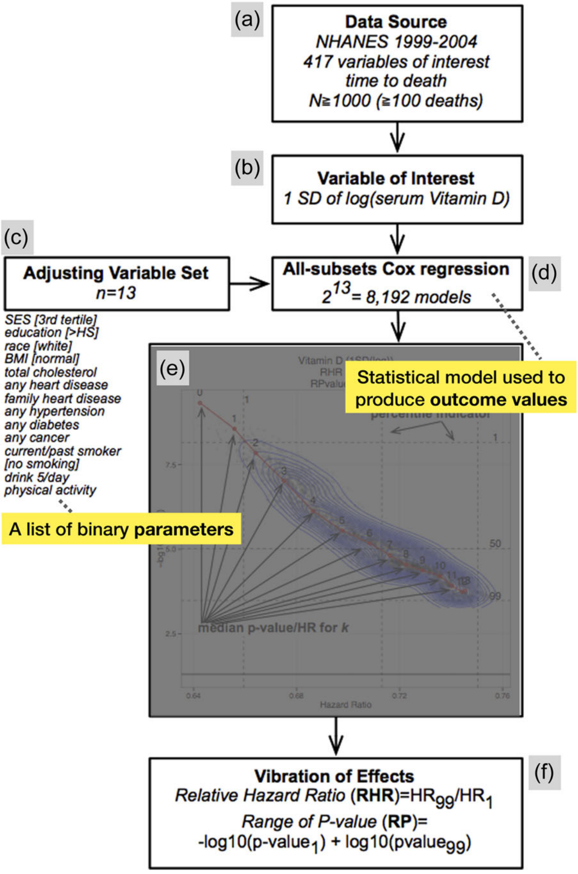


Figure 13: *Example of a multiverse computation schematic [PBI15], describing data source (a), variable of interest (b) and parameters (c, d) composing the multiverse; and elements of the multiverse analysis report: a vibration of effects plot (e); and measures of outcome value spread (f).*

come values for each universe (in some multiverses this would be a parameter if there were more than one model type). Panel (e) is a miniature vibration of effects plot (Section 6.4). Panel (f) contains two metrics the authors use to quantify the spread of outcome values of a multiverse (Outcome ▷ Range), though this is not an essential part of this archetype and the vast majority of multiverse analyses in our corpus do not use such metrics. The illustrated pipeline helps a viewer gain a high level understanding of the multiverse structure (Composition ▷ Process) and the process of analysis.

6.7. Interactive visualization systems

While most of the visualizations in our corpus are static, we identified two interactive visualization systems designed to support multiverse analysis. These systems are the primary inspiration for category Validate (Section 5.5), as these tasks are largely unsupported by the other visualizations in our corpus.

6.7.1. *Explorable multiverse analysis reports (EMARs)*

EMARs (Figure 14) are interactive variants of academic articles inspired by *explorable explanations* [Vic11]. EMARs allow readers to interactively explore individual universes by selecting combinations of parameter values directly in the report, and see the full analysis report resulting from the corresponding universe update accordingly. For example, the dot-interval plot in Figure 14 is not itself a multiverse visualization; instead, each parameter value in the text is an interactive widget that allows the reader to select different values for that parameter, which updates the body text and all visualizations in the report to describe the analysis resulting from the selected universe. For example, clicking on the *t-distribution* widget allows the reader to switch to bootstrapped confidence intervals.

Unlike the summary visualizations in our corpus, EMARs allow the reader to inspect the full statistical report for a single universe. This allows a reader to make more informed judgements about the validity of each universe (Validate ▷ Details). However, this can make it more difficult to gain a higher-level understanding of outcome sensitivity (Section 5.2). EMARs address this by allowing the reader to animate over all of the universes to see how much individual visualizations of outcomes change depending on the active universe (Outcome ▷ Frequency).

6.7.2. *Boba*

Boba (Figure 15) is an interactive system designed to support multiverse analysis. As a full system it supports many tasks in our taxonomy, but the support for some tasks are limited. It supports tasks in Connect (Section 5.3) by allowing viewers to interactively select parameters of interest (Figure 15(c)), which it uses to show dotplots

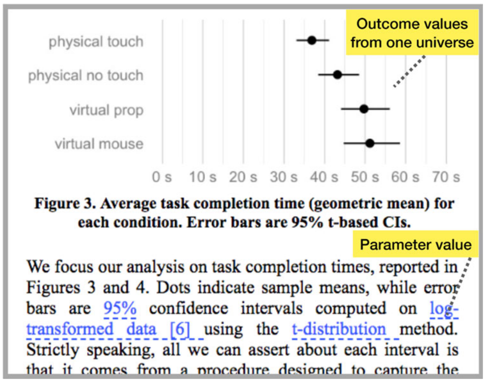


Figure 3. Average task completion time (geometric mean) for each condition. Error bars are 95% t-based CIs.

We focus our analysis on task completion times, reported in Figures 3 and 4. Dots indicate sample means, while error bars are 95% confidence intervals computed on log-transformed data [6] using the t-distribution method. Strictly speaking, all we can assert about each interval is that it comes from a procedure designed to capture the

Figure 14: *Excerpt from an explorable multiverse analysis report [2019], where parameter values can be selected dynamically through interactive text widgets, resulting in figures, numerals and text updating accordingly in the report. See https://explorablemultiverse.github.io/.*

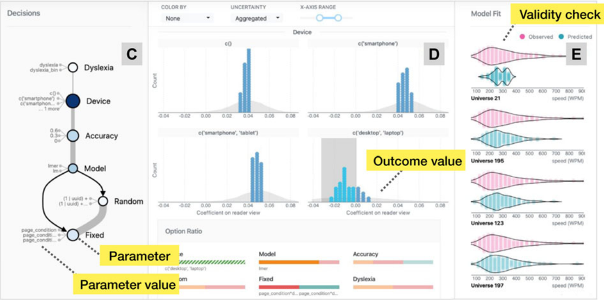


Figure 15: *Screenshot of the Boba system [LKAH20]. Panel C shows the design space of parameters and their relationships; parameters that are source of sensitivity are in a darker colour. Panel D is a trellis of dotplots of outcome values, subsetted by parameter values. Panel D shows predictive distributions from each universe compared to the observed data.*

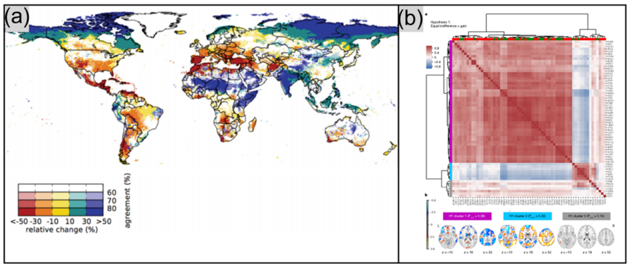


Figure 16: *Two examples of domain-specific visualizations of multiverse analyses. (a) Outcome values are contextualized in a geographical map [Bie15], (b) correlation matrix of outcome values [BNHC*20].*

of outcome values faceted by parameter values (Figure 15(d)). It has some support for Connect Combination (Section 5.4) tasks by allowing the viewer to select multiple parameters, though the scalability of these tasks is limited by the fact that faceting is itself limited to two axes. It does not support Validate ▷ Details as it mainly relies on summary visualizations.

A unique contribution of this system is that it explicitly considers model fit (Figure 15(e)) as a component of assessing multiverse validity (Validate ▷ Metrics). This is because a cross-product of *a priori* reasonable parameters may produce many universes with poor model quality, and some universes may not provide a sound basis for inference [DGGS20]. Support for this task is provided by allowing the viewer to examine model fit (Figure 15(e)) and exclude outcome values from poor-fitting models in the final interpretation.

6.8. Domain-specific visualizations

We selected the archetypes above for full description as we believe they are likely to be widely applicable to multiverse analyses, regardless of domain. Some of the visualizations in our corpus are instead highly domain-specific [Car12, BZ08, PV17, BKB*20]. A common example is spatial data, such as encountered in geographic and medical research. We present two examples of this type of visualization that both employ heatmaps to encode multiverse outcome data together with domain-specific visualizations that would otherwise only show the result of a single analysis.

Figure 16(a) shows the output of water runoff (discharge) predictions from 55 climate models [Bie15]. Outcome values and sensitivity are encoded on a bivariate colour scale: Mean predicted change in water runoff (outcome value) is mapped to hue, and percentage agreement between universes (outcome sensitivity) is mapped to saturation, helping the viewer assess the range of outcomes in each region on the map (Outcome ▷ Range).

Figure 16(b) shows the correlation between outcomes across universes in a neuroimaging analysis multiverse. The top panel is a correlation matrix: Rows and columns are universes, and each cell shows the correlation of outcome values between two universes. The dendrogram axes are similar to those of the outcome matrix (Section 6.5), but are the result of a clustering algorithm rather than a direct representation of parameters. The colour-coding on the rows links the results in the matrix to the models of human brains in the lower panel. Each brain model uses a heatmap to show the averaged relative activation of certain brain areas. This aids in assessing the sensitivity of outcomes to different analysis choices (Connect, Section 5.3).

7. Discussion

In this section, we discuss some difficulties and limitations of current multiverse analysis visualizations, implications for design following from our survey, and directions for future work.

7.1. The illusion of probability in multiverse visualizations

One issue with existing multiverse visualizations that show outcome values stems from the subtle yet important distinction between *probabilistic* and *possibilistic* interpretations of frequencies. Although this is a general difficulty when interpreting any multiverse analysis, it may be exacerbated by visualizations.

Under a *probabilistic* interpretation, all specified universes would be assumed to be equally likely to be correct, so outcome values that occur more frequently within the multiverse must be more likely to be correct. Yet the set of reported universes in a multiverse analysis is not a random sample of all reasonable specifications, and universes are themselves not statistically independent [SSN20]. Authors and readers may even disagree on the validity of some universes [SSN20]. It follows that when interpreting visualizations such as outcome histograms, the relative frequencies of outcomes should *not* be treated probabilistically.

Instead, variation in outcomes should be treated *possibilistically* [GK75]: The presence of an outcome in a multiverse indicates that it is a possible result of reasonable analytical choices. Under this interpretation, no outcome value can be considered any more or less likely to be correct solely based on how frequently it occurs; one must instead examine the validity of universes leading to particular outcomes.

However, the probabilistic interpretation is very tempting: We suspect that many readers may interpret visualized outcome frequencies as probabilities or likelihoods, and we have encountered such interpretations while reading multiverse analysis reports. This relates to classic notions of visualization *expressiveness*: density plots, histograms, dotplots and so forth. All invite a probabilistic interpretation even though that interpretation is not intended for multiverse data. One might consider this misinterpretation a kind of **illusion of probability**. This illusion puts designers of visualizations for multiverse analysis in a bind, as the frequency information that creates the illusion is still useful for many tasks (e.g. Connect ▷ OutcomeFrequency and ConnectCombo ▷ OutcomeFrequency). How can we visualize this frequency information while preventing erroneous probabilistic interpretations? One potential direction may be to use visualization types explicitly designed for possibilistic uncertainty, such as probability boxes [2011]; see Bonneau *et al.* [2014] for further discussion of possibilistic versus probabilistic uncertainty visualization. We have not seen examples of possibilistic uncertainty visualizations applied to multiverse analysis as yet.

7.2. Visualizations to better support multiverse validation and interpretation are needed

We considered proposing a sixth task category, 'Interpret the Multiverse' as the logical final step in a multiverse analysis: to make some inference about the original dataset (not about the sensitivity of that inference). We decided against doing so as we did not find examples of tasks in this category that were substantially supported by multiple sources in the corpus, generalizable and explicitly a feature of a visualization. Overall, we found that interpretations of a multiverse vary widely between authors, are often domain-dependent, and are not strongly tied to specific features of any visualization.

Del Giudice *et al.* [DGGS20] stated that, '*going forward, multiverse-style methods should not be narrowly thought of as a means to promote transparency in reporting, but rather as an analytic tool that can profitably aid the interpretation of data and inform the development of theoretical models'.* This echoes similar suggestions made in earlier works [SSN20, STGV16], but most multiverse reports we reviewed did not go beyond tasks from the Outcome (Section 5.2) and Connect (Section 5.3) categories, or at least not in a way that explicitly referenced a visualization.

Only two visualizations provided support for tasks under the Validate category (Section 5.5). Two recent threads of research have suggested the need to more carefully validate the universes in a multiverse, possibly pruning some universes. Liu *et al.* [LKAH20] suggest doing so by examining model fit and provide some support for this task in Boba (Figure 15). They suggest that an analyst might wish to iteratively redefine the multiverse itself as a result of a previous round of multiverse analysis, given that some analytical choices may no longer be considered equally defensible after having run them on the data.

Relatedly, Del Giudice *et al.* [DGGS20] argue that analysts should explicitly consider whether analytical choices have *principled equivalence*, *principled non-equivalence*, or if there is *uncertainty* about their equivalence; each conclusion leads to different choices about whether to include a parameter value in the multiverse. They argue that if poor analysis choices were truly excluded, most multiverses would be much smaller than ones seen in practice. Simonsohn *et al.* [SSN20] note that, '*while all included specifications should be theoretically justified, statistically valid and non-redundant, researchers may nevertheless consider some specifications superior to others and that some should be given greater weight than others'.* However, to date, we are not aware of reported multiverse analyses that attempt such relative weightings.

7.3. Multiplexing and interaction to investigate parameter combinations

Few visualizations provided substantial support for tasks in the Connect combinations category (Section 5.4). For most visualizations , this support comes with the caveat that meaningful combinations of parameters have been selected ahead of time (e.g. Figure 9), which does not address how visualization might be used to discover these interesting relationships in the first place. Two strategies in the corpus were used to help analysts discover the impact of arbitrary parameter combinations on outcome sensitivity: multiplexing in space (e.g. faceting), and interactivity.

While the vibrations of effects plot (Section 6.4) can only compare across a small number of parameter values at once, Patel *et al.* [PBI15] describe a full analysis workflow in which an analyst reviews potentially hundreds of vibration of effects plots representing combinations of parameter values. On a smaller scale, Poarch *et al.* [PVB19] faceted by both variables of interest and parameters, producing an 8-by-6 of outcome histograms (Section 6.1) to report their multiverse analysis. In theory, faceting by parameter can be performed with any base-plot type, but in our corpus, faceting was primarily used with archetypes that did not otherwise support connecting parameters to outcome values (Sections 5.3 and 5.4).

Boba (Figure 15) combined faceting with interactivity, allowing viewers to facet according to interactively selected parameters. Interactivity removes the need to present all faceted plots at once and could aid in more focused exploration. However, there is an untapped potential to enhance the value of other plot types in our corpus through interactivity, beyond just interactively selecting facets: the outcome matrix plot (Figure 12), for example, could benefit from interactive row and column reordering to aid in cluster identification [PDF14]; similar functionality could also help reduce tradeoffs in fixed column ordering on specification curve charts (e.g. Figure 8(a) vs. Figure 8(b)). Such approaches could be used in interactive systems aimed at analysts, like Boba [LKAH20], or incorporated into interactive reports aimed at readers, like EMARs [2019].

7.4. Importance of multiverse scale and structure

Multiverses vary in their scale, in terms of both the number of parameters and the number of universes those parameters form in combination. Some multiverses are *dense*, if most or all combinations of parameter values are included, while some are *sparse*, if many theoretically possible combinations of parameters are not included. Sparse multiverses are typical in analyses constructed by using only the specifications found in previous work, or when specifications are crowdsourced (e.g. [SUM*18b]). Some archetypes explicitly visualize this structure (e.g. the dendrograms in outcome matrices; Figure 12) and may not scale well to large numbers of parameters or complex relationships between them, while others do not depict any particular structure and are thus usable regardless (e.g. the outcome histogram; Figure 5).

Part of the inherent difficulty of multiverse analysis is that the data are not easily reduced or summarized without losing information that is critical for supporting important tasks, such as Connect ▷ OutcomeRange or ConnectCombo ▷ OutcomeRange. Summarization of outcome values can appear trivial at first, such as when stating the proportion of universes with outcomes values that were statistically significant, or presenting outcomes with a histogram (Section 6.1). As discussed previously (Sections 6.1 and 7.1), under a possibilistic interpretation even this task is fraught with the danger of misinterpretation. While frequency can also serve as an indicator for how much of the examined choice space is connected to any given outcome, summarizing outcomes severs the threads that connect outcomes to parameter values, thus preventing one from performing any Connect-related tasks (Section 5.3). It may be that supporting some tasks better will tend to reduce support for other tasks. This implies that designers and researchers may be best served by building up a toolbox of multiverse visualizations that support their desired tasks, rather than trying in vain to create an all-in-one solution.

Given this, the design of visualizations must take into account the scale of the multiverses they are to support. In the visualization table in the supplement, we provide our estimation of the scale of multiverses that are supported by each archetype, both in terms of number of parameters and number of universes. As an example, the vibration of effects plot (Section 6.4) scales to an unlimited number of universes, but is only able to show one (or very few) parameter values in a single plot. By contrast, an interactive system is not limited in the amount of parameters it can support overall, but the component visualizations are still limited to simultaneously displaying a number of parameters on the order of tens. Future work might investigate ways to scale multiverse visualizations that already have good support for some tasks to larger multiverses.

7.5. Limitations of this survey and future work

There are several ways in which our survey is limited. We set out to survey tasks and visualizations for multiverse analysis reports, as detailed in Section 3. Since adjacent concepts, such as model comparison, or parameter space exploration (also see Figure 2) likely entail different tasks, we curated our corpus by strictly applying the definitions presented in Section 3. The eight relevant keywords identified from our list of 53 seed articles resulted in a total of 213 corpus candidates. In analysing these candidates, we only found a total of 43 articles fulfilling our criteria. Consequently, our survey may have missed some potentially relevant visualizations.

Our survey only covers multiverse visualizations reported in academic papers, most of which are static. We had to exclude many visualization designs and tools—some of which are interactive—that have been designed for related purposes (see Figure 2). Future work should examine how such tools can inspire the design of multiverse analysis reports, while remaining aware of differences in goals. For example, interactive visualization tools for model building [DCCE19, MLMP17, CPCS19] and for ensemble data analysis [WHLS18, SHB*14] focus on using data visualization to help analysts prune vast spaces of possibilities, often with the goal of identifying one optimal model or set of parameters. In contrast, in a typical multiverse analysis, the entire multiverse is reported as it was decided before the data were analysed, irrespective of the outcomes of those analyses. Nevertheless, pruning tools require effective data overview techniques, which can be re-purposed for multiverse analysis reporting. In addition, adding interactive pruning tools to multiverse analysis reports could help readers navigate them.

Our survey covers how multiverse visualizations have been used across disciplines, but few of the papers we examined are from within the field of information visualization. This is because such visualizations are not broadly used, and we know of very few examples in information visualization. Our focus is however less on helping information visualization researchers *use* such visualizations in their own papers, and more on helping them *study* them as a research subject. We however expect that many of the insights gained by looking at practices across disciplines can transfer to visualization papers, as methodologies for analysing and reporting experiments and transparency criteria are very similar across research areas.

None of the tasks discussed in this work are unique to any single domain or discipline, and the vast majority of datasets being analysed are well expressed in tabular data structures familiar to all quantitative analysts. Major challenges to be addressed by future researchers will involve finding ways to effectively communicate multiverse results of data and analyses with additional structural complexity. For example, hierarchical data and modelling techniques can require multiple visualizations to adequately communicate the results of a single analysis. Similarly, there is no reason why multiverse analysis techniques cannot be applied to analyses of other data structures, such as networks. While domain-specific techniques applied to spatial data may provide some inspiration (Section 6.8), considerable innovation may be required.

8. Conclusion

This state of the art report has reviewed the development and advances made in the visual design and communication of multiverse analysis results, starting with related techniques that go back long before the term *multiverse analysis* was first coined, and carried through the year 2020. We surveyed literature across multiple fields and disciplines, considering visualizations from areas as diverse as psychology, statistics, economics and visualization.

We contributed a coherent and operational terminology to provide researchers with a common vocabulary so they can better communicate and reason about multiverse analyses (Section 3). We assembled a taxonomy of analysis inspection tasks that multiverse visualizations should support, grounded in an extensive analysis of the curated corpus (Section 5). Finally, we discussed the design and functionality of major multiverse visualization archetypes and assessed how well each of them supports our tasks (Section 6), in order to guide analysts in the selection of appropriate visualizations to use when conducting or reporting multiverse analyses.

Our work was motivated by the fact that visualization solutions to multiverse analysis and reporting have, to date, been largely explored in isolation. We contribute a conceptual framework and reflections that can help shed light on this rich design space. Ultimately, no single multiverse visualization has dominant support for all tasks, and there is ample opportunity for future work to investigate improvements to existing visualizations, new visualizations or even combinations of visualizations to better support the range of tasks needed for a complete reporting of a multiverse analysis.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1256260.

References

[AES05] Amar R., Eagan J., Stasko J.: Low-level components of analytic activity in information visualization. In *INFOVIS: Proceedings of the IEEE Symposium on Information Visualization*, (2005), IEEE, pp. 111–117. http://doi.org/10.1109/INFVIS.2005.1532136

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

[BBHR*16] Behrisch M., Bach B., Henry Riche N., Schreck T., Fekete J.-D.: Matrix reordering methods for table and network visualization. *Computer Graphics Forum 35*, (2016), 693–716.

Berry D. (2012) Multiplicities in Cancer Research: Ubiquitous and Necessary Evils. *JNCI Journal of the National Cancer Institute*, *104*, (15), 1125–1133. https://doi.org/10.1093/jnci/djs301

[BHJ*14] Bonneau, G.-P., Hege, H.-C., Johnson C. R., Oliveira M. M., Potter K., Rheingans P., Schultz T.: Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization*. London: Springer (2014), pp. 3–27. https://doi.org/10.1007/978-1-4471-6497-5_1

Bruns Stephan B., Ioannidis John P. A. (2016) p-Curve and p-Hacking in Observational Research. *PLOS ONE*, *11*(2), e0149144. https://doi.org/10.1371/journal.pone.0149144

Bierkens Marc F. P. (2015) Global hydrology 2015: State, trends, and directions. *Water Resources Research*, *51*(7), 4923–4947. https://doi.org/10.1002/2015wr017173

Bastiaansen Jojanneke A., Kunkels Yoram K., Blaauw Frank J., Boker Steven M., Ceulemans Eva, Chen Meng, Chow Sy-Miin, de Jonge Peter, Emerencia Ando C., Epskamp Sacha, Fisher Aaron J., Hamaker Ellen L., Kuppens Peter, Lutz Wolfgang, Meyer M. Joseph, Moulder Robert, Oravecz Zita, Riese Harriëtte, Rubel Julian, Ryan Oisín, Servaas Michelle N., Sjobeck Gustav, Snippe Evelien, Trull Timothy J., Tschacher Wolfgang, van der Veen Date C., Wichers Marieke, Wood Phillip K., Woods William C., Wright Aidan G.C., Albers Casper J., Bringmann Laura F. (2020) Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *Journal of Psychosomatic Research*, *137*, 110211. https://doi.org/10.1016/j.jpsychores.2020.110211

Botvinik-Nezer Rotem, Holzmeister Felix, Camerer Colin F., Dreber Anna, Huber Juergen, Johannesson Magnus, Kirchler Michael, Iwanir Roni, Mumford Jeanette A., Adcock R. Alison, Avesani Paolo, Baczkowski Blazej M., Bajracharya Aahana, Bakst Leah, Ball Sheryl, Barilari Marco, Bault Nadège, Beaton Derek, Beitner Julia, Benoit Roland G., Berkers Ruud M. W. J., Bhanji Jamil P., Biswal Bharat B., Bobadilla-Suarez Sebastian, Bortolini Tiago, Bottenhorn Katherine L., Bowring Alexander, Braem Senne, Brooks Hayley R., Brudner Emily G., Calderon Cristian B., Camilleri Julia A., Castrellon Jaime J., Cecchetti Luca, Cieslik Edna C., Cole Zachary J., Collignon Olivier, Cox Robert W., Cunningham William A., Czoschke Stefan, Dadi Kamalaker, Davis Charles P., Luca Alberto De, Delgado Mauricio R., Demetriou Lysia, Dennison Jeffrey B., Di Xin, Dickie Erin W., Dobryakova Ekaterina, Donnat Claire L., Dukart Juergen, Duncan Niall W., Durnez Joke, Eed Amr, Eickhoff Simon B., Erhart Andrew, Fontanesi Laura, Fricke G. Matthew, Fu Shiguang, Galván Adriana, Gau Remi, Genon Sarah, Glatard Tristan, Glerean Enrico, Goeman Jelle J., Golowin Sergej A. E., González-García Carlos, Gorgolewski Krzysztof J., Grady Cheryl L., Green Mikella A., Guassi Moreira João F., Guest Olivia, Hakimi Shabnam, Hamilton J. Paul, Hancock Roeland, Handjaras Giacomo, Harry Bronson B., Hawco Colin, Herholz Peer, Herman Gabrielle, Heunis Stephan, Hoffstaedter Felix, Hogeveen Jeremy, Holmes Susan, Hu Chuan-Peng, Huettel Scott A., Hughes Matthew E., Iacovella Vittorio, Iordan Alexandru D., Isager Peder M., Isik Ayse I., Jahn Andrew, Johnson Matthew R., Johnstone Tom, Joseph Michael J. E., Juliano Anthony C., Kable Joseph W., Kassinopoulos Michalis, Koba Cemal, Kong Xiang-Zhen, Koscik Timothy R., Kucukboyaci Nuri Erkut, Kuhl Brice A., Kupek Sebastian, Laird Angela R., Lamm Claus, Langner Robert, Lauharatanahirun Nina, Lee Hongmi, Lee Sangil, Leemans Alexander, Leo Andrea, Lesage Elise, Li Flora, Li Monica Y. C., Lim Phui Cheng, Lintz Evan N., Liphardt Schuyler W., Losecaat Vermeer Annabel B., Love Bradley C., Mack Michael L., Malpica Norberto, Marins Theo, Maumet Camille, McDonald Kelsey, McGuire Joseph T., Melero Helena, Méndez Leal Adriana S., Meyer Benjamin, Meyer Kristin N., Mihai Glad, Mitsis Georgios D., Moll Jorge, Nielson Dylan M., Nilsonne Gustav, Notter Michael P., Olivetti Emanuele, Onicas Adrian I., Papale Paolo, Patil Kaustubh R., Peelle Jonathan E., Pérez Alexandre, Pischedda Doris, Poline Jean-Baptiste, Prystauka Yanina, Ray Shruti, Reuter-Lorenz Patricia A., Reynolds Richard C., Ricciardi Emiliano, Rieck Jenny R., Rodriguez-Thompson Anais M., ROMYN Anthony, SALO Taylor, SAMANEZ-LARKIN Gregory R., SANZ-MORALES Emilio, SCHLICHTING Margaret L., SCHULTZ Douglas H., SHEN Qiang, SHERIDAN Margaret A., SILVERS Jennifer A., SKAGERLUND Kenny, SMITH Alec, SMITH David V., SOKOL-HESSNER Peter, STEINKAMP Simon R., TASHJIAN Sarah M., THIRION Bertrand, THORP John N., TINGHÖG Gustav, TISDALL Loreen, TOMPSON Steven H., TORO-SEREY Claudio, TORRE TRESOLS Juan Jesus, TOZZI Leonardo, TRUONG Vuong, TURELLA Luca, VAN 'T VEER Anna E., VERGUTS Tom, VETTEL Jean M., VIJAYARAJAH Sagana, VO Khoi, WALL Matthew B., WEEDA Wouter D., WEIS Susanne, WHITE David J., WISNIEWSKI David, XIFRA-PORXAS Alba, YEARLING Emily A., YOON Sangsuk, YUAN Rui, YUEN Kenneth S. L., ZHANG Lei, ZHANG Xu, ZOSKY Joshua E., NICHOLS Thomas E., POLDRACK Russell A., SCHONBERG Tom (2020) Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

[BRRYD20] BURSZTYN L., RAO A., ROTH C. & YANAGIZAWA-DROTT D.: Misinformation During a Pandemic. Working paper, University of Chicago, Becker Friedman Institute for Economics, 2020, 1–114. https://www.doi.org/10.3386/w27417

BRYAN Christopher J., YEAGER David S., O'BRIEN Joseph M. (2019) Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, *116*, (51), 25535–25545. https://doi.org/10.1073/pnas.1910951116

[BZ08] BARALDI P., ZIO E.: A combined Monte Carlo and possibilistic approach to uncertainty propagation in event tree analysis. *Risk Analysis: An International Journal 28*, 5 (2008), 1309–1326.

[Car12] CARP J.: On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience 6* (2012), 149.

[CC14] CHRISTENSEN B., CHRISTENSEN S.: Are female hurricanes really deadlier than male hurricanes? *Proceedings of the National Academy of Sciences 111*, 34 (2014), E3497–E3498.

[CGD18] COCKBURN A., GUTWIN C., DIX A.: Hark no more: On the preregistration of CHI experiments. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), ACM, pp. 141. https://doi.org/10.1145/3173574.3173715

[CJT19] CESARIO J., JOHNSON D. J., TERRILL W.: 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.

[Coo18] COOKSON J. A.: When saving is gambling. *Journal of Financial Economics 129*, 1 (2018), 24–45.

CASHMAN Dylan, PERER Adam, CHANG Remco, STROBELT Hendrik (2020) Ablate, Variate, and Contemplate: Visual Analytics for Discovering Neural Architectures. *IEEE Transactions on Visualization and Computer Graphics*, *26*, (1), 863–873. https://doi.org/10.1109/tvcg.2019.2934261

CIRILLO Pasquale, TALEB Nassim Nicholas (2016) On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications*, *452*, 29–45. https://doi.org/10.1016/j.physa.2016.01.050

CUMMING Geoff (2014) The New Statistics. *Psychological Science*, *25*, (1), 7–29. https://doi.org/10.1177/0956797613504966

[DBH19] DONNELLY S., BROOKS P. J., HOMER B. D.: 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.

[DCCE19] DAS S., CASHMAN D., CHANG R., ENDERT A.: BEAMES: Interactive multimodel steering, selection, and inspection for regression tasks. *IEEE Computer Graphics and Applications 39*, 5 (2019), 20–32.

DEL GIUDICE Marco, GANGESTAD Steven W. (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), 251524592095492. https://doi.org/10.1177/2515245920954925

[DGH*18] DUBOIS J., GALDI P., HAN Y., PAUL L. K., ADOLPHS R.: Resting-state functional brain connectivity best predicts the personality dimension of openness to experience. *Personality Neuroscience 1* (2018), e6.

[DJS*19] DRAGICEVIC P., JANSEN Y., SARMA A., KAY M., CHEVALIER F.: Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York City, NY, USA, 2019), Association for Computing Machinery, pp. 1–15. https://doi.org/10.1145/3290605.3300295

[DKBK19] DEJONCKHEERE E., KALOKERINOS E. K., BASTIAN B., KUPPENS P.: Poor emotion regulation ability mediates the link between depressive symptoms and affective bipolarity. *Cognition and Emotion 33*, 5 (2019), 1076–1083.

DEJONCKHEERE Egon, MESTDAGH Merijn, HOUBEN Marlies, ERBAS Yasemin, PE Madeline, KOVAL Peter, BROSE Annette, BASTIAN Brock, KUPPENS Peter (2018) The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology*, *114*, (2), 323–341. https://doi.org/10.1037/pspp0000186

[DS18] DENNY M. J., SPIRLING A.: Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis 26*, 2 (2018), 168–189.

[ESR17] ELSHERIF M. M., SABAN M. I., ROTSHTEIN P.: The perceptual saliency of fearful eyes and smiles: A signal detection study. *PloS One 12*, 3 (2017), e0173199.

FiveThirtyEight [Fiv15] : Hack your way to scientific glory, 2015. https://projects.fivethirtyeight.com/p-hacking/

[FS11] FERSON S., SIEGRIST J.: Verified computation with probabilities. In *Proceedings of the IFIP Working Conference on Uncertainty Quantification* (New York City, NY, USA, 2011), Springer, pp. 95–122. https://doi.org/10.1007/978-3-642-32677-6_7

GILDERSLEEVE Kelly, HASELTON Martie G., FALES Melissa R. (2014) Do women's mate preferences change across the ovulatory cycle? A meta-analytic review. *Psychological Bulletin*, *140*, (5), 1205–1259. https://doi.org/10.1037/a0035438

[GK75] GAINES B. R., KOHOUT T. L.: Possible automata. In *Proceedings of the International Symposium on Multiple-Valued Logic* (1975).

[GL13] GELMAN A., LOKEN E.: *The Garden of Forking Paths: Why Multiple Comparisons can be a Problem, Even when there is No "Fishing Expedition" or "p-hacking" and the Research Hypothesis was Posited ahead of Time*. Department of Statistics, Columbia University, 2013. http://www.stat.columbia.edu/~gelman/research/unpublished/forking.pdf

GEHLENBORG Nils, WONG Bang (2012) Heat maps. *Nature Methods*, *9*, (3), 213–213. https://doi.org/10.1038/nmeth.1902

[Har07] HARZING A.: Harzing, A.W. (2007) Publish or Perish, available from https://harzing.com/resources/publish-or-perish

HARRIS Christine R., CHABOT Aimee, MICKES Laura (2013) Shifts in Methodology and Theory in Menstrual Cycle Research on Attraction. *Sex Roles*, *69*, (9-10), 525–535. https://doi.org/10.1007/s11199-013-0302-3

HEGRE Håvard, SAMBANIS Nicholas (2006) Sensitivity Analysis of Empirical Results on Civil War Onset. *Journal of Conflict Resolution*, *50*, (4), 508–535. https://doi.org/10.1177/0022002706289303

JELVEH Zubin, KOGUT Bruce, NAIDU Suresh Political Language in Economics. *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.2535453

JUNG K., SHAVITT S., VISWANATHAN M., HILBE J. M. (2014) Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, *111*, (24), 8782–8787. https://doi.org/10.1073/pnas.1402786111

[KAB*20] KRAUS M., ANGERBAUER K., BUCHMÜLLER J., SCHWEITZER D., KEIM D. A., SEDLMAIR M., FUCHS J.: Assessing 2D and 3D heatmaps for comparative analysis: An empirical study. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), pp. 1–14. https://doi.org/10.1145/3313831.3376675

[LAH20] LIU Y., ALTHOFF T., HEER J.: 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* (2020), pp. 1–14. https://doi.org/10.1145/3313831.3376533

LANDY Justin F., JIA Miaolei (Liam), DING Isabel L., VIGANOLA Domenico, TIERNEY Warren, DREBER Anna, JOHANNESSON Magnus, PFEIFFER Thomas, EBERSOLE Charles R., GRONAU Quentin F., LY Alexander, VAN DEN BERGH Don, MARSMAN Maarten, DERKS Koen, WAGENMAKERS Eric-Jan, PROCTOR Andrew, BARTELS Daniel M., BAUMAN Christopher W., BRADY William J., CHEUNG Felix, CIMPIAN Andrei, DOHLE Simone, DONNELLAN M. Brent, HAHN Adam, HALL Michael P., JIMÉNEZ-LEAL William, JOHNSON David J., LUCAS Richard E., MONIN Benoît, MONTEALEGRE Andres, MULLEN Elizabeth, PANG Jun, RAY Jennifer, REINERO Diego A., REYNOLDS Jesse, SOWDEN Walter, STORAGE Daniel, SU Runkun, TWOREK Christina M., VAN BAVEL Jay J., WALCO Daniel, WILLS Julian, XU Xiaobing, YAM Kai Chi, YANG Xiaoyu, CUNNINGHAM William A., SCHWEINSBERG Martin, URWITZ Molly, THE CROWDSOURCING HYPOTHESIS TESTS COLLABORATION, UHLMANN Eric L. (2020) Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*, (5), 451–479. https://doi.org/10.1037/bul0000220

LIU Yang, KALE Alex, ALTHOFF Tim, HEER Jeffrey (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

LONSDORF Tina B, KLINGELHÖFER-JENS Maren, ANDREATTA Marta, BECKERS Tom, CHALKIA Anastasia, GERLICHER Anna, JENTSCH Valerie L, MEIR DREXLER Shira, MERTENS Gaetan, RICHTER Jan, SJOUWERMAN Rachel, WENDT Julia, MERZ Christian J (2019) Navigating the garden of forking paths for data exclusions in fear conditioning research. *eLife*, *8*, https://doi.org/10.7554/elife.52465

MALEY Steve (2014) Statistics show no evidence of gender bias in the public's hurricane preparedness. *Proceedings of the National Academy of Sciences*, *111*, (37), E3834–E3834. https://doi.org/10.1073/pnas.1413079111

MALTER D. (2014) Female hurricanes are not deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, *111*, (34), E3496–E3496. https://doi.org/10.1073/pnas.1411428111

MUHLBACHER Thomas, LINHARDT Lorenz, MOLLER Torsten, PIRINGER Harald (2018) TreePOD: Sensitivity-Aware Selection of Pareto-Optimal Decision Trees. *IEEE Transactions on Visualization and Computer Graphics*, *24*, (1), 174–183. https://doi.org/10.1109/tvcg.2017.2745158

MUNAFÒ Marcus R., NOSEK Brian A., BISHOP Dorothy V. M., BUTTON Katherine S., CHAMBERS Christopher D., PERCIE DU SERT Nathalie, SIMONSOHN Uri, WAGENMAKERS Eric-Jan, WARE Jennifer J., IOANNIDIS John P. A. (2017) A manifesto for reproducible science. *Nature Human Behaviour*, *1*, (1), https://doi.org/10.1038/s41562-016-0021

[Mun14] MUNZNER T.: *Visualization Analysis and Design*. CRC Press, 2014.

MUÑOZ John, YOUNG Cristobal (2018) We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model

Robustness. *Sociological Methodology*, *48*, (1), 1–33. https://doi.org/10.1177/0081175018777988

Nosek B. A., Alter G., Banks G. C., Borsboom D., Bowman S. D., Breckler S. J., Buck S., Chambers C. D., Chin G., Christensen G., Contestabile M., Dafoe A., Eich E., Freese J., Glennerster R., Goroff D., Green D. P., Hesse B., Humphreys M., Ishiyama J., Karlan D., Kraut A., Lupia A., Mabry P., Madon T., Malhotra N., Mayo-Wilson E., McNutt M., Miguel E., Paluck E. Levy, Simonsohn U., Soderberg C., Spellman B. A., Turitto J., VandenBos G., Vazire S., Wagenmakers E. J., Wilson R., Yarkoni T. (2015) Promoting an open research culture. *Science*, *348*, (6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Orben Amy, Dienlin Tobias, Przybylski Andrew K. (2019) Social media's enduring effect on adolescent life satisfaction. *Proceedings of the National Academy of Sciences*, *116*, (21), 10226–10228. https://doi.org/10.1073/pnas.1902058116

Olkin Ingram, Dahabreh Issa J., Trikalinos Thomas A. (2012) GOSH - a graphical display of study heterogeneity. *Research Synthesis Methods*, *3*, (3), 214–223. https://doi.org/10.1002/jrsm.1053

Orben Amy, Przybylski Andrew K. (2019) The association between adolescent well-being and digital technology use. *Nature Human Behaviour*, *3*, (2), 173–182. https://doi.org/10.1038/s41562-018-0506-1

Orben Amy, Przybylski Andrew K. (2019) Screens, Teens, and Psychological Well-Being: Evidence From Three Time-Use-Diary Studies. *Psychological Science*, *30*, (5), 682–696. https://doi.org/10.1177/0956797619830329

[Pap64] Pap A.: Theory of definition. *Philosophy of Science 31*, 1 (1964), 49–54.

Patel Chirag J., Burford Belinda, Ioannidis John P.A. (2015) Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*, (9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

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

[PV17] Piironen J., Vehtari A.: Comparison of Bayesian predictive methods for model selection. *Statistics and Computing 27*, 3 (2017), 711–735.

[PVB19] Poarch G. J., Vanhove J., Berthele R.: The effect of bidialectalism on executive function. *International Journal of Bilingualism 23*, 2 (2019), 612–628.

[RES17] Rohrer J. M., Egloff B., Schmukle S. C.: Probing birth-order effects on narrow traits using specification-curve analysis. *Psychological Science 28*, 12 (2017), 1821–1832.

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

Simmons Joseph P., Nelson Leif D., Simonsohn Uri (2011) False-Positive Psychology. *Psychological Science*, *22*, (11), 1359–1366. https://doi.org/10.1177/0956797611417632

Sagi Omer, Rokach Lior (2018) Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*, (4), https://doi.org/10.1002/widm.1249

Simonsohn Uri, Simmons Joseph P., Nelson Leif D. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. *SSRN Electronic Journal*, https://doi.org/10.2139/ssrn.2694998

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

Steegen Sara, Tuerlinckx Francis, Gelman Andrew, Vanpaemel Wolf (2016) Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, *11*, (5), 702–712. https://doi.org/10.1177/1745691616658637

Silberzahn R., Uhlmann E. L., Martin D. P., Anselmi P., Aust F., Awtrey E., Bahník Š., Bai F., Bannard C., Bonnier E., Carlsson R., Cheung F., Christensen G., Clay R., Craig M. A., Dalla Rosa A., Dam L., Evans M. H., Flores Cervantes I., Fong N., Gamez-Djokic M., Glenz A., Gordon-McKeon S., Heaton T. J., Hederos K., Heene M., Hofelich Mohr A. J., Högden F., Hui K., Johannesson M., Kalodimos J., Kaszubowski E., Kennedy D. M., Lei R., Lindsay T. A., Liverani S., Madan C. R., Molden D., Molleman E., Morey R. D., Mulder L. B., Nijstad B. R., Pope N. G., Pope B., Prenoveau J. M., Rink F., Robusto E., Roderique H., Sandberg A., Schlüter E., Schönbrodt F. D., Sherman M. F., Sommer S. A., Sotak K., Spain S., Spörlein C., Stafford T., Stefanutti L., Tauber S., Ullrich J., Vianello M., Wagenmakers E.-J., Witkowiak M., Yoon S., Nosek B. A. (2018) Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*, (3), 337–356. https://doi.org/10.1177/2515245917747646

Silberzahn R., Uhlmann E. L., Martin D. P., Anselmi P., Aust F., Awtrey E., Bahník Š., Bai F., Bannard C., Bonnier E., Carlsson R., Cheung F., Christensen G., Clay R., Craig M. A., Dalla Rosa A., Dam L., Evans M. H., Flores Cervantes I., Fong N., Gamez-Djokic M., Glenz A., Gordon-McKeon S., Heaton T. J., Hederos K., Heene M., Hofelich Mohr A. J., Högden F., Hui K., Johannesson M., Kalodimos J., Kaszubowski E., Kennedy D. M., Lei R., Lindsay T. A., Liverani S., Madan C. R., Molden D., Molleman E., Morey R. D., Mulder L. B., Nijstad B. R., Pope N. G., Pope B., Prenoveau J. M., Rink F., Robusto E., Roderique H., Sandberg A., Schlüter E., Schönbrodt F. D., Sherman M. F., Sommer S. A., Sotak K., Spain S., Spörlein C., Stafford T., Stefanutti L., Tauber S., Ullrich J., Vianello M., Wagenmakers E.-J., Witkowiak M., Yoon S., Nosek B. A. (2018) Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*, (3), 337–356. https://doi.org/10.1177/2515245917747646

[Vic11] Victor B.: Explorable explanations. Online. http://worrydream.com/ExplorableExplanations/ (2011).

[VKT19] Voracek M., Kossmeier M., Tran U. S.: Which data to meta-analyze, and how? *Zeitschrift für Psychologie 227* (2019), 64–82.

[Wah83] Wahba G.: Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society: Series B (Methodological) 45*, 1 (1983), 133–150.

Wang Junpeng, Hazarika Subhashis, Li Cheng, Shen Han-Wei (2019) Visualization and Visual Analysis of Ensemble Data: A Survey. *IEEE Transactions on Visualization and Computer Graphics*, *25*, (9), 2853–2872. https://doi.org/10.1109/tvcg.2018.2853721

Wicherts Jelte M., Veldkamp Coosje L. S., Augusteijn Hilde E. M., Bakker Marjan, van Aert Robbie C. M., van Assen Marcel A. L. M. (2016) Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*, https://doi.org/10.3389/fpsyg.2016.01832

Wicherts Jelte M., Veldkamp Coosje L. S., Augusteijn Hilde E. M., Bakker Marjan, van Aert Robbie C. M., van Assen Marcel A. L. M. (2016) Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology*, *7*, https://doi.org/10.3389/fpsyg.2016.01832

Young Cristobal, Holsteen Katherine (2017) Model Uncertainty and Robustness. *Sociological Methods & Research*, *46*, (1), 3–40. https://doi.org/10.1177/0049124115610347

Young Cristobal (2018) Model Uncertainty and the Crisis in Science. *Socius: Sociological Research for a Dynamic World*, *4*, 237802311773720. https://doi.org/10.1177/2378023117737206

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure 1: All publications in our corpus

Supporting Information