Understanding, Communicating, and Reducing Analytical Uncertainty: Theory, Visualization Designs, and an Augmented Presentation System to Support Validation and Interpretation of a Multiverse Analysis

by

Brian D. Hall

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Information)
in the University of Michigan
2024

Doctoral Committee:

Professor Eytan Adar, Co-Chair
Professor Matthew Kay, Co-Chair
Professor Christopher Brooks
Professor Priti Shah

Brian D. Hall
briandh@umich.edu
ORCID iD: 0009-0008-8926-3742

DEDICATION

This work is dedicated to the loves of my life—my partner, children, and grandchild: Miranda, Collin, Kaitlin, Kristen, and Lincoln. You make work worth doing.

To my father Stephen Hall, and all my ancestors: your struggles were never in vain. Thank you.

ACKNOWLEDGMENTS

On consideration of who I should acknowledge for their professional assistance towards completion of this work and my doctorate, I find myself humbled by the sheer number of people who have helped me, and in so many different ways. Every one of them deserves more than a small mention in this text, but for want of a perfect memory and unlimited time to write, I can only name the few whose aid is most salient to me now. Regardless, my heartfelt thanks extends to each and every soul.

First, I must acknowledge my advisor Matt Kay, whose understanding and support throughout the years has been valuable to me beyond measure. It is devilishly difficult to provide just the right amount of encouragement, skepticism, positivity, constructive criticism, motivation, support, ambition, concern, and rationality—and deliver each at the appropriate time—yet he did so with such consistency that I cannot imagine how anyone could do it better. You made difficult work bearable, and collaborative intellectual work a pleasure.

Of course this work could also not have been completed without the help of all the other members of my great committee: co-chair Eytan Adar, Christopher Brooks, and Priti Shah. Your feedback, insightful questions, and advice have greatly improved this work.

I also want to acknowledge Michael Nebeling and Max Speicher, who taught me so many things about research and beyond. I treasure the memories of our struggles together to make sense of nonsense, bend bleeding-edge technology to our will, and make our own fun along the way. I am very glad we had the time to work together that we did, and without those experiences I would be so much the poorer.

It was a great pleasure to work with Matt Brehmer when I interned at Tableau Research. I often think back to what we accomplished in so little time, and it braces me to face new challenges with more confidence and less worry. You showed me how decisions can be structured to provide layers of meaning and value across multiple time scales, and reflecting on it still inspires *wonder*. You are a craftsman of the highest order.

Going back farther in time, I am no less indebted now to Tim Krause and Tim Kennedy than I was when I was finishing my undergraduate degree. I do not know how exactly to express how much working and talking with you both shaped my attitude and direction in my work. You helped me to have a far better idea of what exactly I wanted—and what I was getting myself into—than

any one else could have. With your help I was able to go into research and graduate school with my eyes open, and only because of your willingness to share your experience was I able to avoid so many troubles that could so easily have befallen me.

I also want to acknowledge three people particularly for their out-sized contribution to my education. It was because of Margaret Hedstrom that I finally got to experience what I believe to be true education: a genuine meeting of the minds in rigorous and respectful intellectual struggle to advance our individual and collective understanding of the world. Tiffany Veinot almost single-handedly helped me to finally see the value and proper role of theory and qualitative investigation. Mark Plonsky helped me to understand that: a truly great education is still attainable by those that want it most, no matter what the norm may be; statistics can be both understood and applied with all due skepticism, even as there is no substitute for knowing what you are doing and why; and that teaching is hard work, a joy, and a craft whose productiveness depends upon principles and arts I had never before considered.

I also want to explicitly acknowledge how much of this work was only possible because of the volunteered assistance of so many participants, reviewers, professors, classmates, and colleagues. So much is credited to so few, even when so much has been given by so many.

I must also acknowledge that this dissertation is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1256260. It is only because of the financial support provided by society that I was able to dedicate so many years of my life to working on problems that may not have solutions, and with no guarantee of profit.

Last yet not least, without the support of Miranda Gregory I could not have made it through—or at least, not without going entirely stark raving mad. I am forever in your debt.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

ABSTRACT

Analyzing data is a complex, multi-step process, with multiple choices available at each step, such as whether and how to exclude outliers, what approach to use to operationalize a variable, or what model and parameters to apply. While it is often possible to exclude some choices as invalid, often many alternatives remain that are equally valid, and when these alternatives lead to divergent conclusions there exists what I term *analytical uncertainty*. Faced with this complexity and the practical demands for professional productivity, analysts often report either a single analysis or a small number of non-divergent supporting analyses, which can result in what has been called *uncertainty laundering*: misrepresenting uncertainty as if it were a known quantity. Yet accurately communicating analytical uncertainty, which is often non-quantified or non-quantifiable, in a way that is both practically useful and professionally acceptable often proves to be an exceedingly difficult task.

In this dissertation, I explore ways to use data visualizations and interactive systems to support the assessment, communication, and reduction of analytical uncertainty. My aim is to pare away a slice of the ontological uncertainty present in empirical research and render it in such a way that not only can it be assessed and clearly communicated, but also so that it can be reduced.

A review of literature on topics related to uncertainty led me to reconceptualize the method of multiverse analysis as a way to assess and communicate analytical uncertainty. A systematic review and critical analysis of published multiverse analysis reports resulted in the derivation of a taxonomy of multiverse analysis tasks, as well as the identification and detailed description of multiverse visualizations archetypes. The systematic review also resulted in the observation that a specific set of multiverse analysis tasks—validation tasks and interpretation tasks—are critical to making a multiverse analysis useful and meaningful, yet are also the tasks that are the least supported by existing archetypes and interactive systems. I conclude that these tasks may be particularly difficult and require multiple perspectives on the data, such that they are perhaps best supported by a set or series of visualizations.

I then developed a prototype system is developed named *AugMeet*, which uses two techniques in combination to support validation and interpretation tasks: *augmented presentation*; and a specially designed *series of interactive visualizations*, including new designs, namely *parameter-faceted outcome curves*, *outcome-faceted outcome curves*, and *twin-faceted residual plots*. I evaluated

this system by recording an approximately 25 minute presentation I gave using the AugMeet system, which was based on my own novel analysis of the hurricane gender-name multiverse created by previous authors [89, 81]; I then showed the recorded demonstration to seven experienced researchers, who I judged to be representative of the intended users for this type of system. Based on these interviews, I offer the following four conclusions:

- The *augmented presentation* aspects of AugMeet seemed to serve primarily as a way to make it easier for the audience to listen to and understand the information being conveyed to them about and through the visualizations themselves.
- Comparing residuals of *twin universes* elicits divergent thinking, even though the design intent of *twin-faceted residual plots* was to avoid spurious conclusions about the superiority of one multiverse option over another.
- *Parameter-faceted outcome curves* give perspective and focus first by providing an overview of all the choices that have been identified as worth considering, and then by providing visual features to identify what is important (potentially impactful) and what is not.
- AugMeet supports the iterative group effort necessary to complete difficult validation and interpretation tasks, and participants found particular value in the demonstrated workflow that featured progressive iteration around a multiverse in a group setting.

Analytical uncertainty is far from trivial to assess, communicate, and reduce, but the techniques developed and explicated within this dissertation show how it is indeed possible and can be practicable.

CHAPTER 1

Introduction & Overview

1.1 Background

Analyzing data is a complex, multi-step process, with multiple choices available at each step, such as whether and how to exclude outliers, what approach to use to operationalize a variable, or what model and parameters to apply [57]. While it is often possible to exclude some choices as invalid, often many alternatives remain that are equally valid. When these alternatives can lead to divergent conclusions, there exists what I term *analytical uncertainty*. Faced with this complexity and the practical demands for professional productivity, analysts often report either a single analysis or a small number of non-divergent supporting analyses. This practice, sometimes termed *undisclosed flexibility*, can cause serious transparency and methodological issues [86, 98, 40] and has been identified as a major cause of the replicability crisis in psychology and other disciplines [68, 65, 23]. Undisclosed flexibility is but one example of a deeper underlying problem of *uncertainty laundering*: misrepresenting uncertainty as if it were a known quantity [38, 39]. Yet accurately communicating analytical uncertainty, which is often non-quantified or non-quantifiable, in a way that is both practically useful and professionally acceptable often proves to be an exceedingly difficult task.

In this dissertation, I explore ways to use data visualizations and interactive systems to support the assessment, communication, and reduction of analytical uncertainty. My aim is to pare away a slice of the ontological uncertainty present in empirical research and render it in such a way that it can not only be assessed and clearly communicated, but also so that it can be reduced. In the following sections I give a summary of each chapter of this dissertation, followed by the thesis statement for the dissertation as a whole (section 1.5).

1.2 Defining & Grounding Analytical Uncertainty (Chapter 2)

The replication crisis has highlighted the fact that scientific findings are inherently uncertain, and that the uncertainty is often greater than commonly anticipated [74, 49]. In chapter 2, I review literature related to the concept of *analytical uncertainty*, a term I define as: uncertainty about empirical knowledge acquired through the process of data analysis due to a lack of certainty about the data analysis process itself (see Figure 1.1).

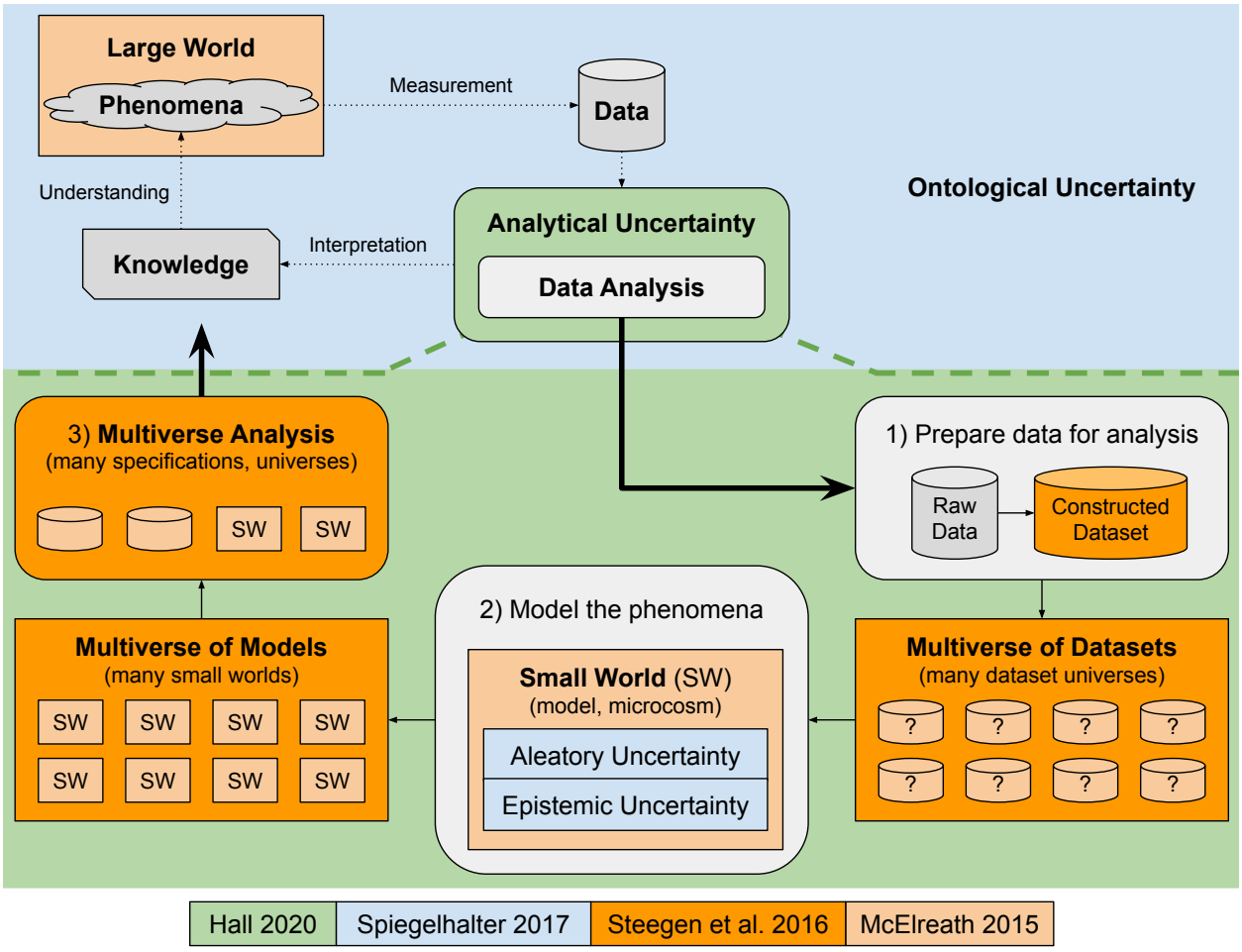


Figure 1.1: Diagram of the connections between analytical uncertainty and related concepts.

Chapter 2 addresses two research questions, summarized in the following subsections. Research questions are numbered according to their associated chapter in this dissertation.

RQ 2.1: What are the types of uncertainty that may be contributing to the replication crisis?

Figure 1.1 is a diagram of the empirical research process, which I depict as being set against a surrounding background of *ontological uncertainty* (section 2.7): from the Latin *ontologia*, meaning an account of being or existence, it is described by Spiegelhalter as "uncertainty about the entire modeling process as a description of reality ... not part of the modeled uncertainty, and can only be expressed as a qualitative and subjective assessment of the coverage of the model, conveying with humility the limitations of our knowledge. ... Such doubts about the whole modeling process arise from scientific uncertainty due to limitations in the available evidence, and this presents a complex communication challenge." [91]

Within the *small worlds* (section 2.6) used to model a phenomena of interest (bottom-center of Figure 1.1), it is possible to account for two other types of uncertainty (section 2.7): *aleatory uncertainty* and *epistemic uncertainty*. *Aleatory uncertainty*—from the Latin *alia*, meaning the rolling of dice or a game of chance—Kiureghian & Ditlevsen say is "presumed to be the intrinsic randomness of a phenomenon" [54], and Spiegelhalter describes it as "inevitable unpredictability of the future due to unforeseeable factors, fully expressed by classical probabilities" [91]. *Epistemic uncertainty*—from the Latin *episteme*, meaning knowledge— Kiureghian & Ditlevsen say is "presumed as being caused by lack of knowledge (or data)" [54]. The difference between aleatory and epistemic uncertainty is thus decided based on context, is defined relative to a given model, and the difference is decided by the person creating the model [54]. While aleatory uncertainty can theoretically be fully accounted for in even basic statistical models, the extent to which any given model accounts for epistemic uncertainty varies widely, with most approaches being primarily concerned with estimating intervals based on the variance in the collected samples.

RQ 2.2: Are there particular types of uncertainty that are not being adequately considered or routinely communicated?

While most statistical methodology focuses on the handling of aleatory and epistemic uncertainty, with ontological uncertainty either left mostly implicit or reported as a qualitative degree of confidence subjectively assessed by a study's authors [91], a different type of uncertainty is highlighted by methods like *multiverse analysis* [92]. Steegen et al. [92] observe that raw data usually needs to be converted into a form that is ready for analysis (creating a Constructed Dataset, center-right of Figure 1.1), such as combining multiple variables, transforming values, or applying exclusion criteria. However, there is often more than one reasonable, defensible (potentially valid) set of choices to make in this process, which leads to the existence of a Multiverse of Datasets (bottom-

right of Figure 1.1). Just as with datasets, defensible alternative analysis choices may exist for statistical models, implying a Multiverse of Models (bottom-left of Figure 1.1), and the combination of every defensible (potentially valid) model with every valid dataset produces the multiple statistical results that compose a multiverse analysis (center-left of Figure 1.1). In a multiverse analysis, every potentially valid combination of choices under consideration is calculated, and the entirety of the collected results is reported; the uncertainty is indicated by the extent to which results differ, indicating that any conclusions to be drawn from the results materially depend on decisions between defensible alternative analysis choices.

Chapter 2 Conclusions

The type of uncertainty highlighted by multiverse analysis is what I refer to as analytical uncertainty; it is uncertainty about the data analysis process which is not fully encompassed by modeled uncertainty (aleatory or epistemic), yet which can in principal be more concretely examined than other forms of ontological uncertainty. However, merely having the concept of analytical uncertainty in mind while performing a multiverse analysis is not sufficient to overcome the "complex communication challenge" [91]. What is next required is a detailed understanding of precisely what can be communicated by methods like multiverse analysis, and in what way can those details be usefully communicated in light of the inherent complexity involved with making sense of a multiplicity of conflicting quantitative outcomes.

1.3 Tasks & Visualization Archetypes for Communicating Analytical Uncertainty (Chapter 3)

In chapter 3, I further explore multiverse analysis as a way to concretely assess and communicate analytical uncertainty.[1] *Multiverse analysis* is an approach that consists of identifying a set of defensible analytical choices, performing all analyses corresponding to the possible combinations of such choices (possibly hundreds, thousands, or even millions) and reporting all outcomes, typically using summary visualizations. The approach has been increasingly popular in academic papers; for example, a Google Scholar search for the term "specification curve"—a type of multiverse analysis visualization [88, 89]— returns 217 papers for the years 2019–2020. However, multiverse analyses still raise many challenges, three of which are particularly relevant to this dissertation:

- Explaining and reporting the outcomes of hundreds or thousands of statistical analyses is

[1] This chapter is based on and contains previously published work [45]. I retain usage of the word 'we' to refer to the collective efforts of my colleagues and myself, and return to the use of 'I' and 'this author' to make clear when statements have not been reviewed by those who collaborated with me for this particular work.

difficult, especially when some of those analyses do not all point towards the same general conclusions [88, 30].

- Literature specifically discussing the methodology of multiverse analysis is scattered across several fields and uses inconsistent terminology, which makes it difficult to communicate and reason about multiverse concepts.

- Visualization methods that have been proposed for helping to conduct and report multiverse analyses are similarly scattered across several fields and use inconsistent terminology.

These challenges motivated the need for a systematic review and critical analysis of published multiverse analysis reports. This review was conducted on a corpus of 43 published research articles that included a non-trivial multiverse analysis report (see section 3.5 for methodology used to curate and analyze the corpus). The research questions addressed by this review are summarized in the following subsections.

RQ 3.1: What tasks or analytical questions do researchers aim to perform or answer when reporting a multiverse analysis visualization?

The corpus of multiverse analysis reports was used to derive a taxonomy of tasks, shown in Table 1.1 and briefly described below.

Composition tasks involve descriptions of the dataset source, how the data was processed, the included variables in the data, and what analytical choices are being considered (parameters[2] and their parameter values[3]). These tasks lay the groundwork necessary for the later sense-making process of drawing conclusions from the multiverse analysis.

Outcome tasks directly consider 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[4] 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. Importantly, how sensitive an outcome is depends upon context and expert judgment in the domain of the analysis. Assessing to what extent outcome values[5] vary across a multiverse typically requires judgments of practical magnitude that are domain-dependent and subject to the analyst's interpretation.

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

[3] A **parameter value** is a possible value taken by a parameter. A synonym is *option* [30], but we use here the term *parameter value* for consistency with the rest of the terminology.

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

[5] An **outcome value** is a possible value taken by an outcome.

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.
Interpret	Interpretation tasks are logical and rhetorical inferences made about the meaning of any given set of results.

Table 1.1: Overview of the derived taxonomy for multiverse analysis tasks. *The *Interpret* category was not formally included in the published version of this taxonomy, for reasons explained in subsection 3.8.2.

Connect tasks explore the potential relationships between individual parameters, parameter values, and outcome values. When outcomes have been determined to be sensitive to analytical choices, 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 analyzing these few choices in greater detail. Further attention could either involve additional tasks described in this framework, or deeper theoretical considerations.

Connect Combinations tasks explore and characterize the relationship between outcomes and analytical choices beyond what was considered in category *Connect*. 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.

Validate tasks critically evaluate 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.

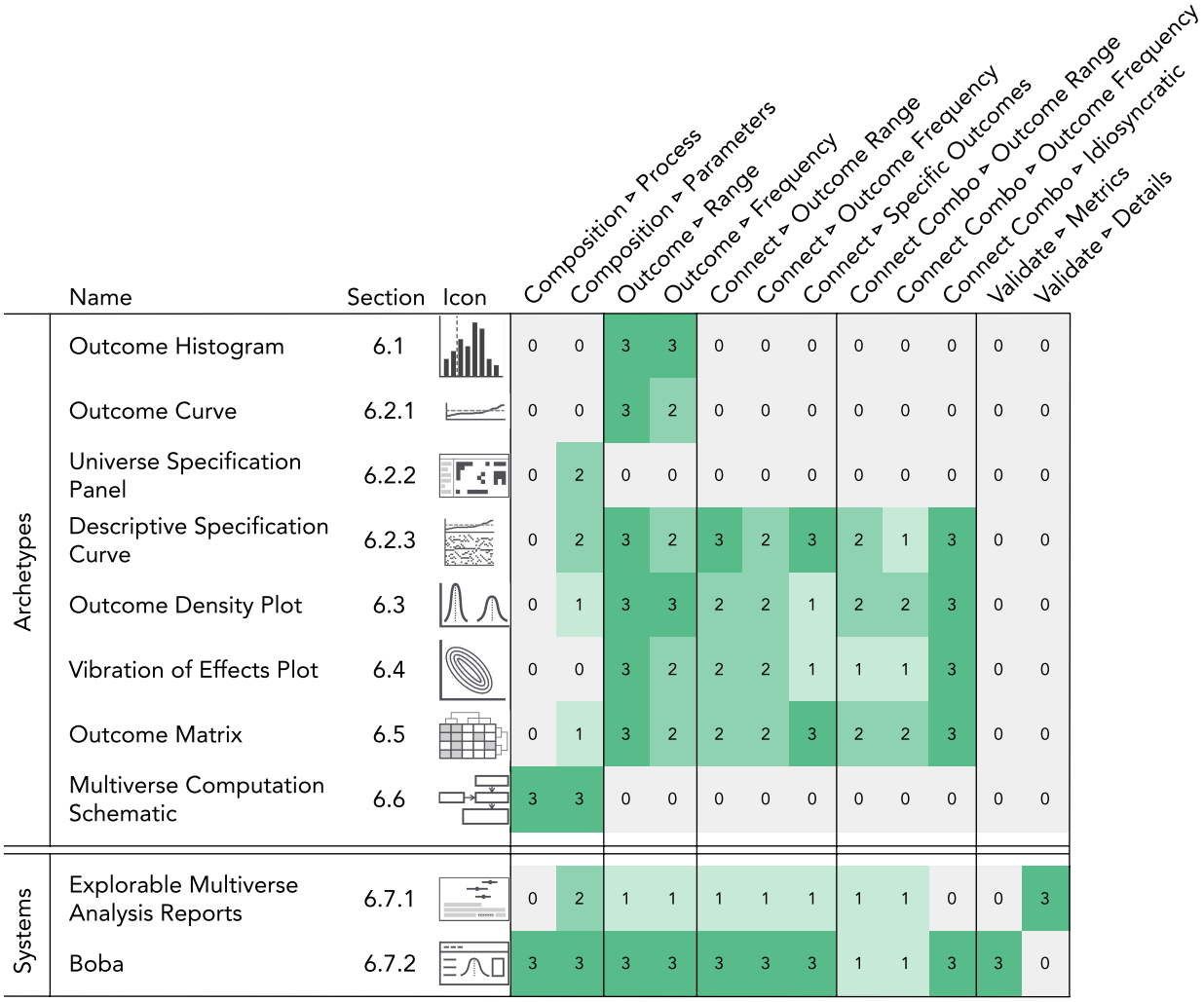
Interpret tasks are logical and rhetorical inferences made about the meaning of any given set of results, especially inferences about the original dataset or underlying phenomena of interest. In the corpus these tasks were carried out almost entirely in the textual narrative of papers, with little to no apparent visual representation or support.

RQ 3.2: What *archetypes* have been used to report multiverse analysis visualizations?

Most visualizations in the corpus can be classified as being composed of six *archetypes*[6], each of which are described in detail in section 3.7, and an overview is shown in Figure 1.2. Archetypes can be treated as modular panels, from which more complex visualizations or dashboards can be composed. The composability of the archetypes is illustrated in the two interactive systems for multiverse analysis that were found in the corpus, each of which used one or more archetypes alone, in series, or as linked visualizations to support a broader range of tasks. A few visualizations are not described as archetypes as they are highly domain-specific, mostly involving spatial data in weather and neuroscience (subsection 3.7.8), and their designs are not broadly applicable to most

[6]Definition of archetype from section 3.4: a class of multiverse analysis visualization designs that convey information about specific multiverse entities using a specific combination of visualization idioms [67].

multiverse analyses or domains.



	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
Archetypes	Outcome Histogram	6.1		0	0	3	3	0	0	0	0	0	0	0	0
	Outcome Curve	6.2.1		0	0	3	2	0	0	0	0	0	0	0	0
	Universe Specification Panel	6.2.2		0	2	0	0	0	0	0	0	0	0	0	0
	Descriptive Specification Curve	6.2.3		0	2	3	2	3	2	3	2	1	3	0	0
	Outcome Density Plot	6.3		0	1	3	3	2	2	1	2	2	3	0	0
	Vibration of Effects Plot	6.4		0	0	3	2	2	2	1	1	1	3	0	0
	Outcome Matrix	6.5		0	1	3	2	2	2	3	2	2	3	0	0
	Multiverse Computation Schematic	6.6		3	3	0	0	0	0	0	0	0	0	0	0
Systems	Explorable Multiverse Analysis Reports	6.7.1		0	2	1	1	1	1	1	1	1	0	0	3
	Boba	6.7.2		3	3	3	3	3	3	3	1	1	3	3	0

Figure 1.2: Overview of the archetypes and interactive systems described in section 3.7. 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).

RQ 3.3: Are there tasks that are not well supported by existing archetypes or systems (section 3.7)?

Tasks from the categories of Validate and Interpret are the least supported by existing archetypes. While the two interactive systems in this corpus (EMAR and Boba) each individually support different Validate tasks, there is very limited support overall for interpretation tasks among archetypes. In most papers interpretation is left entirely to narrative description, and often with only subjective qualitative descriptions of how relatively sensitive the outcomes are to analytical decisions. As the derivation of meaning from results is the ultimate goal of analysis, and inter-

pretation of a multiverse analysis is usually not a trivial task, why are there no visualizations that explicitly support interpretation tasks? What would it take to support such tasks?

Chapter 3 Conclusions

While multiverse analysis is a promising method for examining analytical uncertainty, substantial challenges remain. One particular challenge is the limited support for the Validate and Interpret categories of tasks, which are the tasks that are ultimately essential for deriving useful meaning from a multiverse analysis and communicating that meaning to others. So the question to be answered in the remainder of this work is: what will it take to support validation and interpretation tasks?

The first step towards answering this question is to consider what about the tasks might be making them so difficult. I believe the apparent difficulty and complexity of validation and interpretation tasks indicates that these tasks are fundamentally different from the other multiverse tasks, particularly in that they often require higher-level mental reasoning and connection of disparate facts, some of which may not even be encoded as specific data. For example, consider what it would take to answer this question: is a specific analytical choice statistically appropriate, given the data and the context of its specific multiverse? While this question is fundamental to a multiverse analysis and makes reference to data, it also requires logical and theoretical considerations that are not a feature of any given dataset. A task like this can still be supported by a visualization, but it may require multiple visualizations and views of the data at a universe-level or multiverse-level, and even then it cannot be decomposed to the same degree that a task like Assess Outcome Sensitivity can be. This suggests that there may not be a single visualization archetype that can support tasks like this, but instead there may be a set or series of inter-related visualizations that would be helpful.

These tasks may benefit from—or even require—collaboration between people with different areas of expertise. In my prior work [44], I built an augmented presentation system that supported nuanced discussion around a data analysis. In the next chapter, I extend this prior work to create a system that features a series of visualizations designed to collectively support validation and interpretation tasks in a research meeting about a multiverse analysis. This system represents one possible way of addressing the lack of support for these important yet difficult tasks.

1.4 AugMeet: Augmented Presentation System to Support Validation and Interpretation Tasks in a Multiverse Analysis (Chapter 4)

This chapter covers the culminating project of this dissertation, which involves the development and evaluation of a system that supports the use of augmented presentation techniques with a series of visualizations to collectively support validation and interpretation tasks.

RQ 4.1: How can validation and interpretation tasks be supported by using augmented presentation techniques together with a series of related data visualizations?

To answer this research question, I developed an interactive system named AugMeet. I recorded a presentation I created and delivered using AugMeet to serve as an applied demonstration of the system, and which depicts a person presenting the first results of a multiverse analysis to a mock-meeting of researcher collaborators. I then conducted an evaluation study by holding semi-structured interviews with seven participants that had relevant expertise, with each participant being highly-educated, having 5-20+ years of research experience, and having familiarity with multiverse analysis ranging from being somewhat familiar to having published papers about the topic. During the interviews all aspects of the system were discussed and critiqued.

The AugMeet system combines multiple ways to support the completion of validation and interpretation tasks, each of which are described in the results of the evaluation study (section 4.6). First, it uses *augmented presentation*[7] as a technique to make it easier for the audience to follow along and understand information provided about and through data visualizations, while simultaneously making it easier for people to pay attention and encouraging them to be interested and engaged. Second, it uses a series of visualizations to build up a shared understanding of the facts, provide an overview of what is being considered, focus effort on what is most potentially impactful, and support low-level interrogation of *twin universes* (4.4.2) to provide a new perspective and source of critical information about the multiverse.

[7] augmented presentation: the performance of a presentation aided by live video augmentation, which is a technique previously developed and published by this author [44]. The development of AugMeet used the prototype developed for that prior work as a starting point, extending and adapting the technique to apply to the multiverse context.

Chapter 4 Conclusions

The visualization series demonstrated by AugMeet integrates a variety of novel techniques and concepts to support validation and interpretation tasks, including: the concept of *twin universes* (subsection 4.4.2), *parameter-faceted outcome curves* (Figure 4.2), *two-dimensional option-faceted outcome curves* (Figure 4.3), and *twin-faceted residual plots* (Figure 4.6). AugMeet combines all of these individual aspects together with the technique of augmented presentation to form a single system, and the demonstration of the system also shows a workflow that participants identified as being of further value due to its approach of progressive, iterative interrogation of a multiverse to reduce uncertainty.

Overall, AugMeet takes a multi-pronged approach to support the tasks in a multiverse that are not only the hardest to perform, but also are the most essential to making a multiverse analysis useful and of more practical value when shared with others.

1.5 Thesis Statement & Claims

Through this dissertation I argue that visualizations and interactive systems can support the use of multiverse analysis as a way to assess analytical uncertainty, direct efforts towards ways to practically reduce it, and improve the accuracy and meaningfulness of the consequent assessment of whatever analytical uncertainty remains.

- Claim 1 (chapter 2, chapter 3): Assessing analytical uncertainty and reducing it requires the completion of distinct analytical tasks—which are supported to varying extents by identified multiverse visualization archetypes—but existing systems provide only very limited support for a category of tasks that are particularly necessary for reducing analytical uncertainty, namely validation and interpretation tasks.

- Claim 2 (chapter 4): A *series of visualizations* can be used to narrow and direct focus while still eliciting divergent insights, collectively supporting the completion of validation tasks to reduce uncertainty and interpretation tasks to derive useful meaning from the multiverse.

- Claim 3 (chapter 4): *Augmented presentation* of data visualizations is a way to facilitate and support nuanced discussions around a multiverse analysis, particularly by improving audience attention and engagement, aiding audiences in obtaining a clear understanding of complex information with less expenditure of mental effort.

I conclude this dissertation with final reflections on these thesis claims and offer suggestions for future work in chapter 5.

CHAPTER 2

Defining & Grounding Analytical Uncertainty

2.1 Chapter Introduction

Empirical research is the process of studying a phenomena through the recording of observations and experiences as data, and then analyzing that data to determine what it can tell us about a studied phenomenon. The data analysis process alone is rarely trivial, being composed of numerous steps, with multiple options available at each step. While it is often possible to exclude some options as invalid or inappropriate, there may still remain many defensible alternatives that cannot be excluded, and this presents a potential problem: if there is more than one reasonable way to perform an analysis on a given dataset, how is one to interpret multiple results that yield varying or contradictory conclusions?

This problem represents what I call analytical uncertainty: uncertainty about empirical knowledge acquired through the process of data analysis due to a lack of certainty about the data analysis process itself. This type of uncertainty is difficult to conceptualize, quantify, and communicate, and poses a hard problem for both researchers conducting analyses, and for researchers who read the published analyses of other researchers.

Let us consider a published example that illustrates what I mean by analytical uncertainty, and the difficulty it represents. In 2014, a paper entitled Female hurricanes are deadlier than male hurricanes [53] was published in the Proceedings of the National Academy of Sciences (PNAS). The authors used a statistical model to estimate a potential relationship between gendered hurricane names and deaths attributed to hurricanes. In the original published analysis, the authors concluded that feminine-named severe hurricanes are more deadly than masculine-named ones, based upon the deaths predicted from a model they fit to the archival data.

In *Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications*, Simonsohn et al. [88] describe five different major sets of decisions in the process of Jung et al. [52] that could have been done some other defensible way (ex: choice of data exclusion criteria, choice of regression model type). The full set of options combine to form 1,728 ways to analyze the data,

which the authors call *specifications*. Out these 1,728 specifications, only 37 resulted in statistically significant effects that indicated female-named hurricanes caused more deaths; all other results were not statistically significant, including 20 that indicated that it was actually masculine-named hurricanes that cause more deaths.

The specification curve for Jung et al. [52] is shown in Figure 2.1, and it can help us to better understand the analytical uncertainty present in this example. If you accept all the specifications examined as reasonable, the results show that there is so much analytical uncertainty that one cannot conclusively establish that either masculine or feminine hurricane names are more dangerous. Both the magnitude and direction of the effect seem especially dependent on two choices: the model form (negative binomial) and the exclusion criteria (dropping the 2 highest observations). Determining the proper model form to analyze this data is thus shown to be crucial, and the exclusion or inclusion of options in this category may have an outsized impact on the overall results of this analysis.

Other authors have also studied variations of the concept of performing and reporting analysis

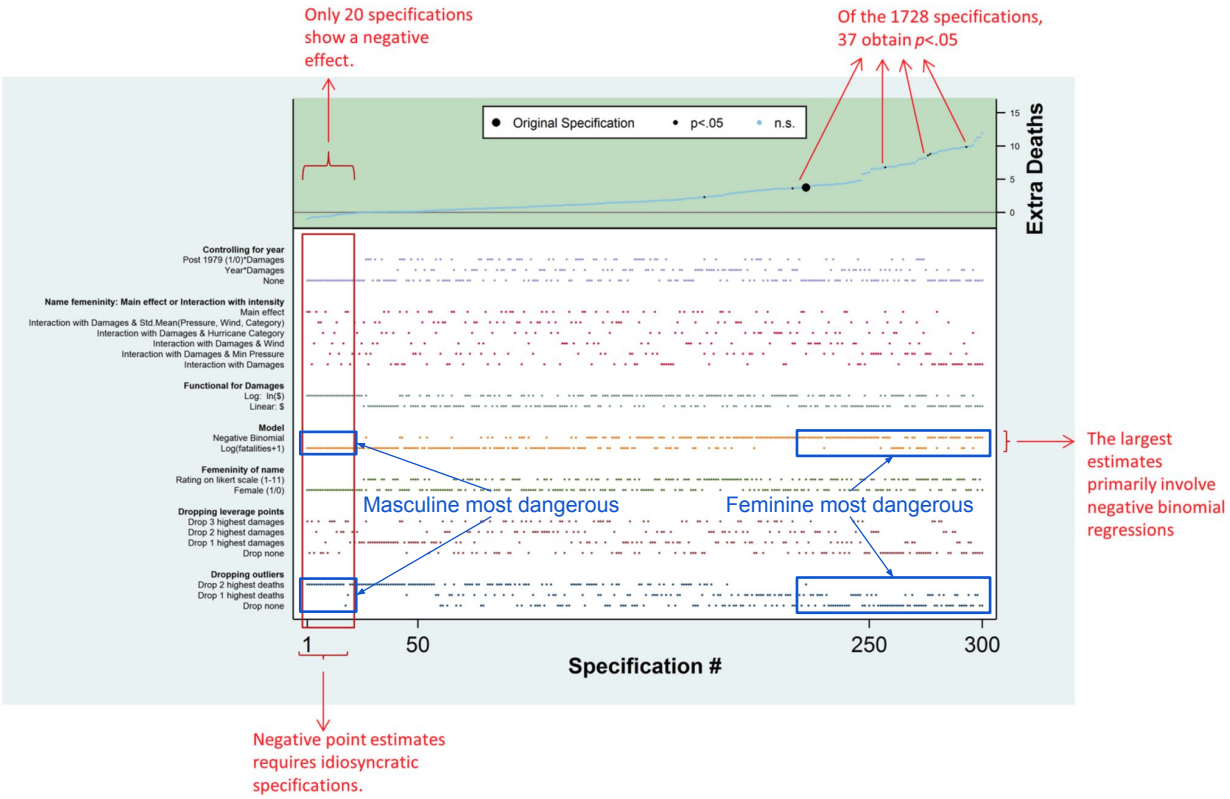


Figure 2.1: Illustrated example of a specification curve, reproduced from Simonsohn et al. [88], with original author annotations in red and my own added annotations in blue. Out of 1,728 specifications, this visualization shows the 100 specifications that produced the highest and lowest estimates (50 each), as well as 200 randomly sampled other specifications.

of many specifications, which Steegen et al. [92] have termed multiverse analysis[1] [92]. Those authors demonstrated how a dataset is itself constructed from raw data before being further analyzed, such as by combing variables or setting exclusion criteria. Just a few common choices combine to imply dozens or hundreds of possible datasets that could be constructed from the same raw data. Patel et al. [75] focuses on how covariate inclusion changes model estimates, examining 8,192 models for each of 417 variables of interest, resulting in over 3.4 million distinct results. The authors further calculated that if all available covariates in their dataset were included, the number of possible specifications would be 2417, even though they only used a single form of statistical model. Such combinatorial explosion demonstrates the need for a systematic approach.

While assessing and reporting a single analysis often takes up a substantial portion of a research paper, interpreting and communicating so many different analyses presents new challenges. Let us briefly consider the example of Patel et al. [75] shown in Figure 2.2 to illustrate some of the potential issues. In every panel the x-axis is a hazard ratio, while the y-axis is a transformed p-value, so each dot shows the output of 2 pieces of information produced by every model. Each panel shows the output of 8,192 models, with each model including a different set of 13 possible covariates, with the related variable of interest listed at the top of each panel.

Panel A and B of Figure 2.2 show two different visualization methods on different variables of interest, possibly because the authors did not agree on a single best way to handle the overplotting that results from having so many datapoints shown in a single scatterplot. Panels C, D, and E cover the same variable of interest, as the authors discovered that unique visual patterns often indicate that a single covariate may be responsible for divergent analysis results. To discover which covariates may be responsible for which patterns, Panel C uses color to show density, while Panels D and E use color to indicate whether a single covariate was included in a given model. The authors explain that interpretation of the visual patterns often requires subject-matter expertise, as for example the patterns in Panels C-D-E are related to the relationship between cardiometabolic indicators and diabetes [75].

The full workflow presented by Patel et al. [75] consists of 13 panels for each of their 417 variables of interest, composed of 1 panel as a combined summary (like Panel A, B, or C) and 13 panels for each of the 12 covariates analyzed (like Panels D and E). They also propose two novel summary statistics, the visual identification of 6 prototypical patterns identified from their data, and additional visualizations based on four metrics to compare analyses between variables of interest. Even with all the depth of the vibration of effects workflow, none of their visualizations allow a viewer to look at a single result and determine what set of analytical choices produced

[1] I adopt the term multiverse analysis to refer to this concept, and further unpack the meaning and origin of the term in later sections. For now, the term need only be understood to refer to systematically performing and reporting many analyses on a given source of data, as was done in Simonsohn et al. [88].

them, though this task is possible with a specification curve.

Taken together, the above examples show that while there are ways to try to assess and report analytical uncertainty in a way that can provide important insights to researchers, the remaining visualization challenges alone are considerable.

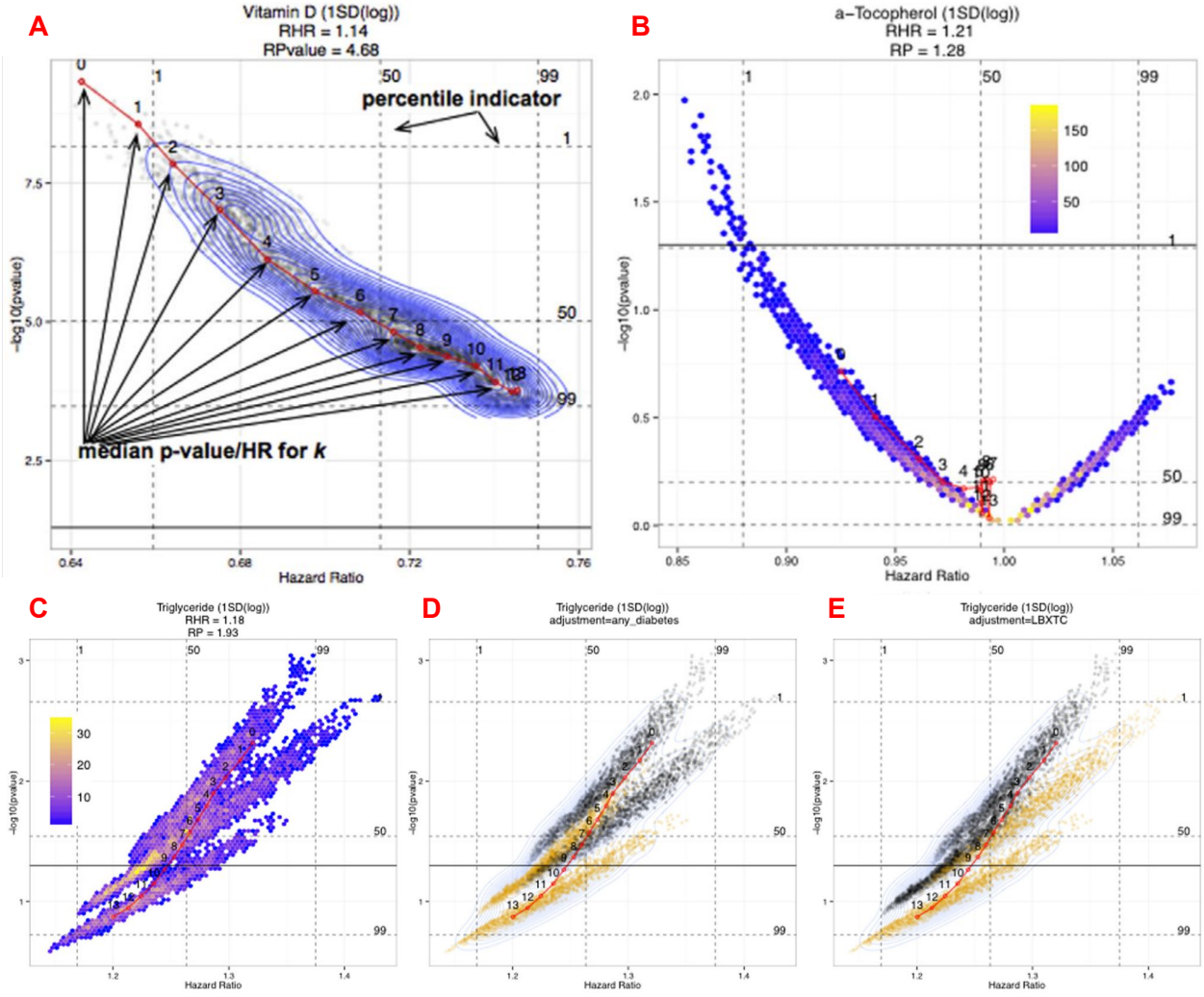


Figure 2.2: Example of Vibration of Effects visualization, original from [75], with addition of A–E panel labels in red for ease of reference in this chapter. This specialized version of a volcano plot illustrates how covariate inclusion impacts model estimates, examining 8,192 models for each of 417 variables of interest, ultimately representing over 3 million distinct results. Each panel represents a multiverse analysis for a single variable of interest (labeled at the top of each panel), as further explained in the text of this chapter. This type of plot is an example of a multiverse visualization archetype, as detailed later in subsection 3.7.4.

2.2 Research Questions

The following research questions are addressed in this chapter.

RQ 2.1: What are the types of uncertainty that may be contributing to the replication crisis?

The answer to this question is shown in Figure 2.4, the full explanation of which occurs throughout the remainder of this chapter.

RQ 2.2: Are there particular types of uncertainty that are not being adequately considered or routinely communicated?

I define a new term, *analytical uncertainty*, to refer to a particular type of uncertainty that I believe has not before been adequately considered and is not routinely communicated. The term is fully defined in the sections that follow. Specification curve analysis is used as an example throughout this chapter as a way of addressing analytical uncertainty. A broader class of methods for addressing analytical uncertainty—multiverse analysis—is the focus of the next chapter.

2.3 Common-use meaning of uncertainty

The *Guide to the Expression of Uncertainty in Measurement* states that "the word 'uncertainty' means doubt" [63]. Spiegelhalter goes a bit further, and states that many scientists use the term uncertainty to refer to "everything that is not certain, including a single coin flip ... only [distinguishing] the extent to which the uncertainty is quantifiable" [91]. In other words, there are two categories: certainty (or certain knowledge), and uncertainty (everything else).

I will not contradict these commonsensical definitions. However, the main problem with simply referring to uncertainty is that it is so ambiguous. Whose uncertainty? What specifically are they uncertain about? Does uncertainty differ only in magnitude, or does it differ also in kind? Are all types of uncertainty the same, such that they can be examined and communicated in precisely the same way?

Many alternative definitions and types of uncertainty have been proposed, with some terms (such as aleatory and epistemic) being used by hundreds of different authors over time. Unfortunately, the same terms are regularly used in different and often contradictory ways, sometimes even by a single author.

2.4 Defining analytical uncertainty in the context of empirical research

To explain how analytical uncertainty relates to the empirical research process, I illustrate a simplified version of this process in Figure 2.3. First, some observable phenomena (event, act, process, or similar object of interest) is identified. Second, some measurement (or observation) is made of the phenomena, recorded as data. Third, data processing provides input for data analysis. Fourth, interpretation of the data analysis results can produce knowledge. Understanding of this knowledge about the phenomena can have important implications for human behavior and additional research. While every step of the empirical process is important and complex in its own right, I will focus on one portion of this process: going from data to knowledge through data analysis.

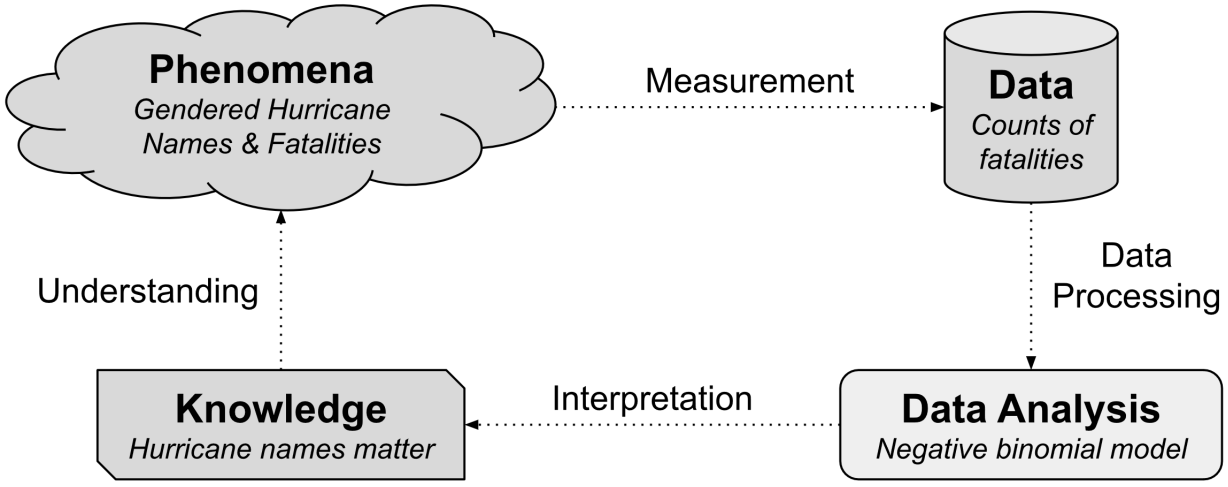


Figure 2.3: Illustration of the empirical research process using [52] as a simplified example.

When there is but one singular proper interpretation of a given dataset, then data analysis may be complex in practice, but it at least will still result in a singular outcome. Any inferences made about the studied phenomena may be flawed or simply wrong due to issues outside of the data analysis, such as problems with measurement or definition of the phenomena, but the problems and uncertainties are not with the analysis itself.

However, what if an analyst cannot justify a single proper analysis of a given dataset, and concludes that many different analyses are reasonable? This is the condition of the world that creates what I call analytical uncertainty:

analytical uncertainty: uncertainty about empirical knowledge acquired through the process of data analysis due to a lack of certainty about the data analysis process itself

The next sections review the theoretical works that are most relevant to the concept of analytical uncertainty, and I illustrate their connections in Figure 2.4. I will regularly refer back to this figure in the coming sections as each new concept is introduced.

2.5 Degrees of freedom, forking paths, and the multiverse

One way to deal with the fact that there are often so many different defensible ways to analyze data is to exploit it to obtain a desired result, such as statistical significance, an interpretation that is surprising or interesting, or that otherwise advances some agenda. Such a strategy has been variously called researcher degrees of freedom [86], specification searching [42], p-hacking [87], selective analyses and the pursuit of statistically significant results [50], and surely many other names as well. As an example of the general body of research, the work of Simmons et al. [86] demonstrated how even just four simple analytical choices could be combined to produce vastly inflated false-positive rates, which "allows presenting anything as significant" [86].

Overall, works related to ideas like researcher degrees of freedom were mostly concerned with how analytical uncertainty can be intentionally used to make findings look more certain and reliable than they really are. This has been hypothesized as being liable to contribute to a scientific literature that is distorted and more unreliable than it would otherwise be [87, 90, 98]. These terms thus tend to have an accusatory and strongly negative connotation, as they can imply knowing dishonesty of the involved researchers for personal gain.

Gelman & Loken [40, 41] argued that the problem of researcher degrees of freedom can be encountered without any intention or perceptibility of any bad behavior, which they explained through the allegory of the garden of forking paths. *The Garden of Forking Paths* is a short story by Jorge Luis Borges, which features a book that resembles a "choose your own adventure story" taken to its logical extreme. At each plot point where a character in the book would make a decision, and where a normal book would then follow only one such decision, Borges' storied book describes the outcome of every single possible decision each character could make. This concept was applied to data analysis to refer to what the authors termed data-dependent analysis [40, 41].

In a data-dependent analysis, "researchers can perform a reasonable analysis given their assumptions and their data, but had the data turned out differently, they could have done other analyses that were just as reasonable in those circumstances ... in this garden of forking paths, whatever route you take seems predetermined, but that's because the choices are done implicitly ... they are using their scientific common sense to formulate their hypotheses in a reasonable way, given the

data they have" [40]. To expand upon the allegory, analytical uncertainty is as a garden of forking paths, where many journeys are possible and will seem reasonable and coherent when viewed individually. Gelman and Loken point out that, from the perspective of a researcher who is not aware of structure of the garden, they took only one single pathway, ran only one test, and thus incorrectly conclude the problem of researcher degrees of freedom does not apply to them [40, 41].

The next major development in this line of thinking comes in the form of multiverse analysis, which the authors note is closely related to the garden of forking paths [92]. The authors observe that raw data usually needs to be converted into a form that is ready for analysis, called data construction, which can include combining multiple variables, transforming values, apply exclusion criteria, and so on. There are often more than one reasonable, defensible set of choices to make in this process, leading to there being a multiverse of datasets (many universes). This concept is illustrated in the upper-left of Figure 2.4, where to prepare data for analysis is to choose from many possible datasets that could be constructed. Similarly, there can a multiverse of models (or many worlds, illustrated lower-right of Figure 2.4).

Given a defined set of universes (specifications), multiverse analysis then runs all results and reports them together [92]. The same concept is at work with the specification curve discussed in the introduction [88]. In the terminology of the multiverse, each universe (or world of a single model) exists unto itself. A universe can be considered more or less synonymous with a specification, and refers to the entirety of the set of options that went into a data analysis, from raw data to results. World, universe, and specification are all words that can be used to describe the same general idea (though 'world' has additional meaning I consider in the next section). I illustrate this in Figure 2.4, with multiverse analysis feedback back into the data analysis process.

A multiverse is a collection of many—but not necessarily all—universes of analysis. To extend the terminology of Steegen et al. [92], I refer to a given multiverse specification as the uniquely identifying collection of universes that make up a given multiverse.

2.6 Small and large worlds

The terminology of referring to worlds in statistics appears to be heavily inspired by Savage's book, The Foundations of Statistics [82]. Savage originally defined a world as being "the object about which a person is concerned", then went on to define worlds as smaller if they contained less detail and context (as in a single egg within a carton), while larger worlds have more detail and context (the whole carton of eggs) [82]. Savage then goes on to to use the word "world" in many other ways and senses throughout his book, even mathematically defining a small world as states, which correspond to events in the grand world, while the grand world is defined as having consequences. Small worlds are even said to be partitions of a grand world (mathematically

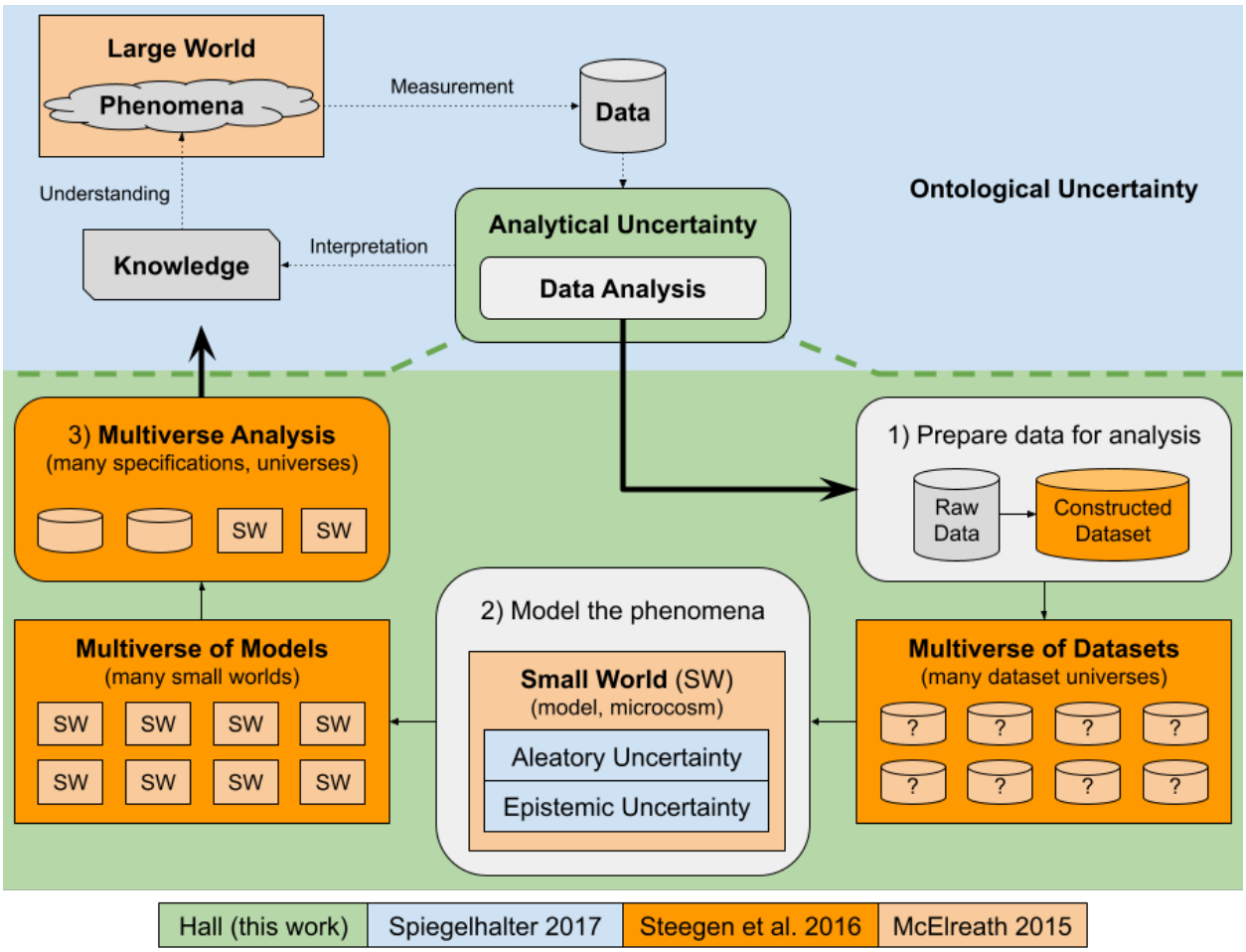


Figure 2.4: Diagram of the empirical research process and its relationship to analytical uncertainty. Data analysis is unpacked to show the process of conducting a multiverse analysis, which allows an assessment of analytical uncertainty. Colors cite source of concept, with legend at bottom.

defined), while a special version of a small world is defined as a microcosm if it behaves according to certain principles of decision-making that are not otherwise relevant here.

The myriad ways a "small world" was referred to by Savage is likely to be responsible for why some authors cite him as giving a definition for a small world that I am not able to locate in his book. As one notable example, Volz and Gigerenzer cite Savage as having defined a small world as being "for situations of perfect knowledge where all relevant alternatives, their consequences, and their probabilities are known for certain" [94]; though this is not far from what Savage called a real microcosm, this definition seems to me to be either apocryphal or from some other work. I mention this to warn the reader that if they have heard the term "small world" before, it may take on very different meanings between different authors.

I instead adopt a related but distinct use of these terms, based on the book *Statistical Rethinking* [62]. McElreath describes the metaphor of there being two types of worlds: a small world, and

a large world. A small world is the "self-contained logical world of the model" [62]. To use his example, a map of the Earth is a small world, in that it is a model that imperfectly and incompletely represents the thing it models. In the same way, using the mathematical construct of the Binomial probability distribution to model a coin flip is also a small world. No matter how complex a model is, it is a small world unto itself. As shown in Figure 2.4, I use the term small world to refer to a model of phenomena which may itself contain its own kind of uncertainty (discussed further in the following section), but which is distinct from the uncertainty the analyst themselves may have. The small world of a model can have a belief, in the Bayesian probability sense, but it is important to remain aware that there is no guarantee a small world of a model resembles the belief of the analyst, nor is there a guarantee the model faithfully represents the thing it was intended to resemble.

A large world is the "broader context in which one deploys a model" [62]. To continue the previous example, while a map of the Earth is a small world model, the Earth itself is the large world phenomenon the model was intended to represent. In Figure 2.4 I indicate this by placing phenomena within a large world to retain a clear distinction between the object of study, the data and analysis process used to attempt to learn about it, and the contextual environment that surrounds any given phenomena.

The terminology of large and small worlds is closely related to the idea of the macrocosm and the microcosm of philosophy (McDonough, 2020). The general historical concept was that some greater thing, the macrocosm, could be understood through the study of some smaller and easier to examine thing, the microcosm. If you believe in the unity of all of nature, this is especially convenient, because if all is one then anything about a macrocosm can be discovered via study of a valid microcosm, even if that macrocosm is the whole cosmos. If you believe that all human-crafted models are simplified representations that may fail to reflect important characteristics of the thing they are intended to model, things are much less convenient.

The key relevance of the small and large world concepts is that they help to clarify that any given model is, put plainly, just a model. As George Box said, "Since all models are wrong the scientist cannot obtain a 'correct' one by excessive elaboration" [11]. The idea of different worlds—or universes—helpfully emphasizes just how far apart and fundamentally different the world of our analytical models may be from the phenomena they were intended to help us learn about. If you accept these premises, it may then be easier to think about what alternate ways of modeling phenomena might help us learn.

2.7 Aleatory, epistemic, and ontological uncertainties

Aleatory and epistemic uncertainty are remarkably overloaded terms, so I will only consider here only the few the definitions that most contributed to the ideas of analytical uncertainty, and that contrast most with it.

Aleatory uncertainty comes from the Latin *alia*, meaning the rolling of dice or a game of chance. Kiureghian & Ditlevsen say that aleatory uncertainty is "presumed to be the intrinsic randomness of a phenomenon" [54], while Spiegelhalter calls it "inevitable unpredictability of the future due to unforeseeable factors, fully expressed by classical probabilities" [91]. I note that classical probability was developed, in part, by analyzing games of chance (gambling), which helps to explain the connection between alia, aleatory, and classical probability [7].

Epistemic uncertainty comes from the Latin *episteme*, meaning knowledge. Kiureghian & Ditlevsen say that epistemic uncertainty is "presumed as being caused by lack of knowledge (or data)" [54].

It is important to note that these terms, used in these classical senses, attribute the cause of uncertainty to two different sources. For example, consider the common example of a flip of a coin. Can you predict the exact outcome of a coin flip, even if only heads and tails are possible, and even if you know the coin flip is fair? It has been said that the inability to predict such an outcome is because of randomness or variability of this phenomenon (coin flips or dice rolls), and that is aleatory uncertainty. As such, the thinking is that the best you can do is learn the probabilities, but that is the absolute limit of predictability. On the other hand, consider that a coin is flipped and caught, but then held in such a way that it is not visible to you. As you can learn what the coin flip resulted in, but just do not know it at the moment, it is thus epistemic uncertainty (in the classic sense of the term).

The problem is that—outside of convenient games like coin flips and dice rolls—how can you know if something cannot be predicted because of its natural variability, or if it could be predicted but you do not know how to do so? This issue led Kiureghian & Ditlevsen to conclude that the difference between aleatory and epistemic is actually up to the person making a model (or doing an analysis) to decide [54]. In practice, they suggest that deciding that something is aleatory is equivalent to saying that you do not think there is any near-term thing that can be done better than estimating a probability, and this uncertainty is effectively treated as modeled and quantified but irreducible. If one thinks that there is some near-term action you can take to reduce the uncertainty, and thus deem it epistemic, one can consider additional model complexity and estimation procedures to try to decrease the uncertainty. The difference between aleatory and epistemic is thus decided based on context, is defined relative to a given model, and the difference is decided by the person creating the model [54].

In terms of doing the actual modeling, Spiegelhalter and Kiureghian & Ditlevsen seem to agree that aleatory uncertainty is treated as randomness of a variable in a way described by classical probability, while epistemic uncertainty is a variable in a model that is treated by more complex mathematical treatments [91, 54]. Epistemic uncertainty might be modeled with a probability distribution, or some learning or estimation process would be used to estimate a latent, hidden, or otherwise unknown value.

Given the above definitions, I have included both aleatory and epistemic uncertainty as properties of the small world models, as shown in Figure 2.4. However, what is uncertainty that is not in the model? Some might just consider it epistemic uncertainty, as it could be attributable to a lack of knowledge. Yet if epistemic is restricted to refer to quantified knowledge defined in the context of a modeling process (and Spiegelhalter, Kiureghian, and Ditlevsen all do restrict the term in just this way [91, 54]), then some other term is required. Instead, one last type of uncertainty is defined.

Ontological uncertainty comes from the Latin *ontologia*, meaning an account of being or existence. It is defined by Spiegelhalter as "uncertainty about the entire modeling process as a description of reality ... not part of the modeled uncertainty, and can only be expressed as a qualitative and subjective assessment of the coverage of the model, conveying with humility the limitations of our knowledge" [91]. He further notes that "Such doubts about the whole modeling process arise from scientific uncertainty due to limitations in the available evidence, and this presents a complex communication challenge" [91].

Of all the types of uncertainty considered here, ontological uncertainty is by far the most closely related to analytical uncertainty. Given that Spiegelhalter gives no further definition than noted above, it is not strictly clear what the boundaries for ontological uncertainty were meant to be. As such, in Figure 2.4 I render it simply as having no border and being in the background vaguely around the empirical process. Having no further formal definition, I use the term ontological uncertainty roughly to mean personal uncertainty of a type not otherwise specified. Analytical uncertainty is, in effect, an attempt to take a slice of ontological uncertainty and make it more intelligible, assessable, and able to be communicated in a way that is more practically useful to researchers.

2.8 Implications for design of analytical and ontological uncertainty

The theoretical foundation discussed previously suggests a potential interpretation and analysis concern related to the type of uncertainty represented by a multiverse analysis. To illustrate this problem, consider Simonsohn et al.'s [88] annotations of a specification curve in Figure 2.1. Those

authors point out that, out of 1,728 specifications, only 20 show a negative effect, and only 37 gave a result that was statistically significant. Other references to the 20 negative specifications called them "idiosyncratic" and "atypical", and in the details of a mathematical justification for a statistical hypothesis the authors say that considering many specifications "more closely approximates a random sample of [all reasonable specifications]" [88]. In short, Simonsohn et al. [88] define a mathematical model, and imply a conceptual model through their choice of words, that treats the analytical uncertainty of a multiverse analysis as either aleatory or epistemic uncertainty and thus probabilistic.

However, if the uncertainty represented by the multiverse is ontological, then modeling it as though it were aleatory or epistemic uncertainty may lead to results that are misleading, unreliable, or simply meaningless. The key questions, from an analytical standpoint, is whether considering many specifications in a multiverse analysis really does—or even theoretically could—approximate a random sample of all reasonable specifications, and if it is even meaningful to define the set of all reasonable specifications in the first place. It is not yet possible to answer such questions conclusively, but that they remain open questions implies two different design directions that can be considered going forward.

The first direction is to support the interpretation of a multiverse analysis in terms of classical probability (aleatory uncertainty). For example, comparing the 20 specifications that produce negative effects to the 1,728 total specifications in Figure 14 suggests a user might mentally calculate an approximate proportion, and then think of this as a probability, such as 20 / 1,728 = 0.01 (1%). If the user were to do this, they might then reason that a negative effect being the true effect seems very unlikely. If the designer of a visualization believes this interpretation is a good one, then it would make sense to use a visualization design that makes completing such a task easy.

On the other hand, if one believes that a multiverse analysis should not be interpreted this way, then one might want to avoid encouraging—or actively discourage—such an interpretation. Just as work describing an illusion of causality has shown that some visualizations cause more users to interpret an association or correlation as though it were causal [99], I hypothesize that some visualizations may encourage an illusion of probability, where users will improperly interpret counts or relative frequencies as though they represented probabilities. However, as I have not found any published work investigating such a concept, it is not clear at this point how one can encourage or discourage such an interpretation of a visualization.

To eliminate any ambiguity about my position, I believe that the proper interpretation of a multiverse is that it represents only possibilities, as it lacks the properties necessary to be a reliably treated as any kind of indicator of probability. The true outcome—to the extent that such a value can be usefully said to exist—may not even be present in any given multiverse, and commonality of error should not be mistaken as a likelihood of correctness. Further discussion of this point is

provided in the following chapter (subsection 3.8.1).

2.9 Final theoretical considerations

In their original specification curve paper, Simonsohn et al. [88] provide a diagram (reproduced here as Figure 2.5) to show how the set of analyses considered appropriate may align or differ. In addition to 'appropriate' they also use 'defensible' and 'reasonable', so I believe they intend all three words to be essentially synonymous in this context. To put the diagram into words, it indicates that not only may people agree or disagree on what set of specifications should be considered (that is, what multiverse specification is reasonable), but the extent of overlap in agreement may range from total to none at all.

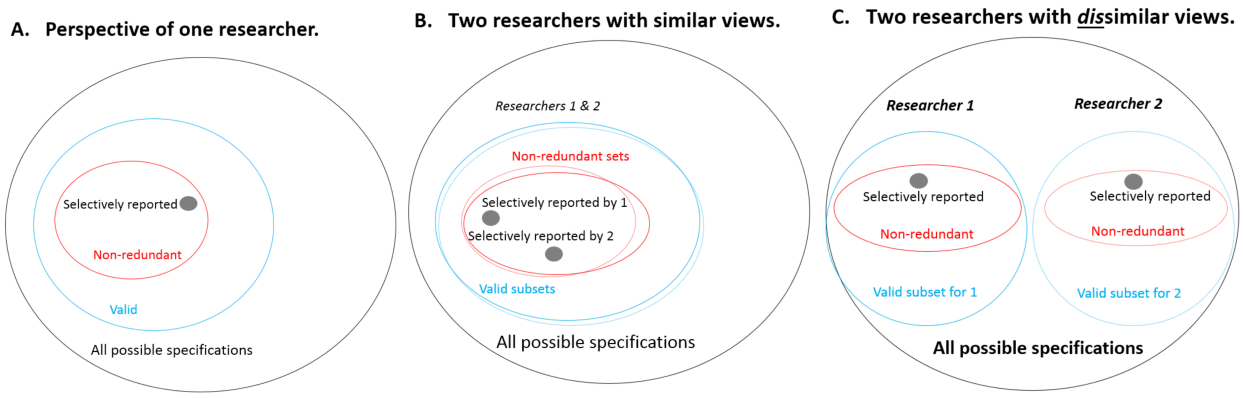


Figure 2.5: Different perspectives on what is reasonable, reproduced without alteration from Simonsohn et al. [88].

Ways to represent, visualize, and communicate some aspects of analytical uncertainty exist, such as the example of the Specification Curve given in the introduction [88], and many examples are examined in chapter 3. The authors of such works do not generally talk about their methods in terms of uncertainty, so that is my own reconceptualization of their work. As more ways are devised to examine analytical uncertainty, there may be the opportunity to learn more about this particular aspect of what we do and do not know.

2.10 Chapter 2 Conclusion

The type of uncertainty highlighted by multiverse analysis is what I refer to as analytical uncertainty; it is uncertainty about the data analysis process which is not fully encompassed by modeled uncertainty (aleatory or epistemic), yet which can in principal be more concretely examined than other forms of ontological uncertainty. However, merely having the concept of analytical uncertainty in mind while performing a multiverse analysis is not sufficient to overcome the "complex communication challenge" [91]. What is next required is a detailed understanding of precisely what can be communicated by methods like multiverse analysis, and in what way can those details be usefully communicated in light of the inherent complexity involved with making sense of a multiplicity of conflicting quantitative outcomes.

CHAPTER 3

Tasks & Visualization Archetypes for Communicating Analytical Uncertainty

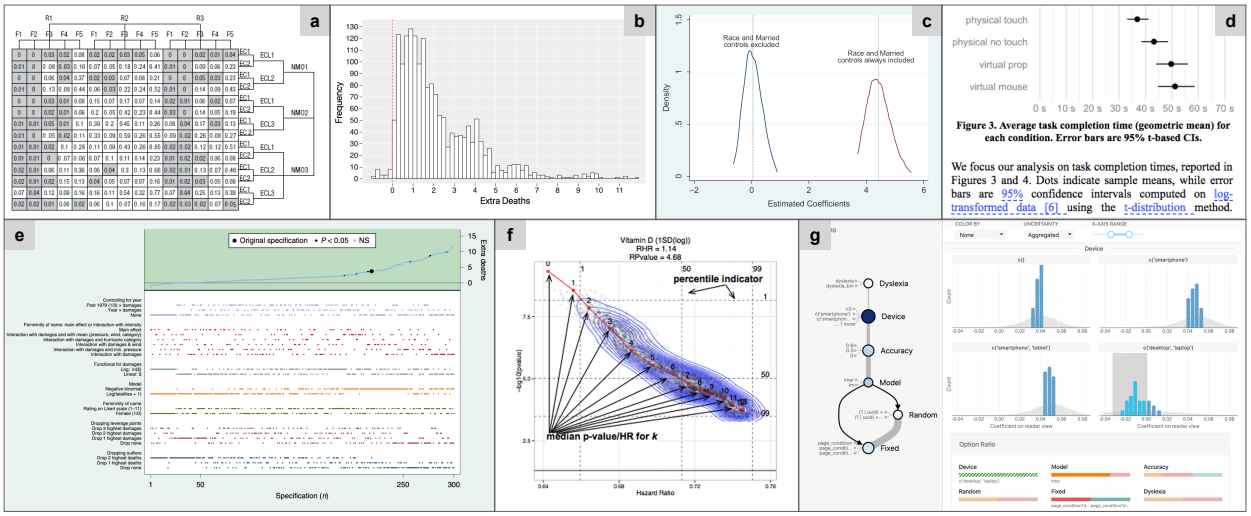


Figure 3.1: Examples of multiverse analysis visualizations discussed in this survey: (a) outcome matrix [92], (b) outcome histogram [92], (c) outcome density plot [101], (d) explorable multiverse analysis reports [30], (e) specification curve [89], (f) vibration of effects plot [75], (g) Boba [58].

3.1 Chapter Introduction

In the previous chapter I covered the definition and theoretical grounding for analytical uncertainty, as well as identifying multiverse analysis as a promising way to concretely assess and communicate analytical uncertainty. This chapter is entirely focused around multiverse analysis and its associated challenges.

Specifically, *multiverse analysis* is an approach that consists of identifying a set of defensible analytical choices, performing all analyses corresponding to the possible combinations of such choices (possibly hundreds, thousands, or even millions) and reporting all outcomes, typically using summary visualizations. The approach has been increasingly popular in academic papers;

for example, a Google Scholar search for the term "specification curve"—a type of multiverse analysis visualization [88, 89]— returns 217 papers for the years 2019–2020. However, multiverse analyses still raise many challenges, three of which are particularly relevant to this chapter:

- 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 [88, 30].
- Literature specifically discussing the methodology of multiverse analysis is scattered across several fields and uses inconsistent terminology, which makes it difficult to communicate and reason about multiverse concepts.
- 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 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 chapter addresses the above challenges through a survey of academic articles that visualize multiverse analyses and related analyses. Importantly, this 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 [24, 64, 17], ensemble data analysis [97], and visual parameter space analysis [83]. Our scope is further clarified in section 3.4.

In this survey, we[1] (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 3.4); (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 3.6); 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 3.7). We close by discussing important design considerations surfaced by our survey—such as illusions of probability created by visualizing frequencies (subsection 3.8.1) and the largely unmet need to support validation and interpretation of multiverses (subsection 3.8.2)—as well as limitations and implications for future work (subsection 3.8.5). For visualization researchers looking to develop multiverse analysis visualizations, our work provides

[1]This chapter is based on and contains previously published work [45]. I retain usage of the word 'we' to refer to the collective efforts of my colleagues and myself, and return to the use of 'I' and 'this author' to make clear when statements have not been reviewed by those who collaborated with me for this particular work.

a foundational set of tasks for subsequent research on developing visualization 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.

The research questions addressed by this chapter are summarized in the following subsections.

Note: This chapter is based on and contains previously published work [45]. I retain usage of the word 'we' to refer to the collective efforts of my colleagues and myself, and return to the use of 'I' and 'this author' to make clear when statements have not been reviewed by those who collaborated with me for this particular work.

3.2 Research Questions

RQ 3.1: What tasks or analytical questions do researchers aim to perform or answer when reporting a multiverse analysis visualization?

The corpus of multiverse analysis reports was used to derive a taxonomy of tasks, shown in Table 1.1. The taxonomy is explained in detail in section 3.6.

RQ 3.2: What *archetypes* have been used to report multiverse analysis visualizations?

With archetypes defined here (in section 3.4) as a class of multiverse analysis visualization designs that convey information about specific multiverse entities using a specific combination of visualization idioms [67], most visualizations in this corpus can be classified as being composed of six archetypes, each of which are described in detail in section 3.7, and an overview is shown in Figure 1.2. Archetypes can also be treated as modular panels, from which more complex visualizations or dashboards can be composed. The composability of the archetypes is illustrated in the two interactive systems for multiverse analysis that were found in the corpus, each of which used one or more archetypes alone, in series, or as linked visualizations to support a broader range of tasks. A few visualizations are not described as archetypes as they are highly domain-specific, mostly involving spatial data in weather and neuroscience (subsection 3.7.8), and their designs are not broadly applicable to most multiverse analyses or domains.

RQ 3.3: Are there tasks that are not well supported by existing archetypes or systems (section 3.7)?

Tasks from the categories of Validate and Interpret are the least supported by existing archetypes. While the two interactive systems in this corpus (EMAR and Boba) each individually support different Validate tasks, there is very limited support overall for interpretation tasks among archetypes. In most papers interpretation is left entirely to narrative description, and often with only subjective qualitative descriptions of how relatively sensitive the outcomes are to analytical decisions.

3.3 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. [88] 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 3.4.

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 [52]. However, later analyses of the same data called this finding into question [61, 19, 60]. 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. [88] 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 subsection 3.7.2) 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. [88] 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 to indicate whether the result is statistically significant. Ultimately, their question is whether or not there is a statistically significant effect of hurricane name gender on extra deaths (i.e., if $p \leq .05$). 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 analyzing the dataset. Simonsohn et al. [88] report 1,728 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. [88] 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 1,728 universes (about 2%) yield $p \leq .05$. 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.

3.4 Definitions of Key Concepts

In this section, we introduce definitions that will serve to outline the scope of our survey (and many of which are used accordingly throughout the remainder of this dissertation). These are stipulative [73] and are not meant to be authoritative.

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

> **multiverse analysis report**: 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 [92], 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. [30], 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 3.5). Thus, the focus is on what is reported, not what is analyzed. If multiple analyses are conducted but a single one is reported, as is commonly the case in empirical research [98], then this cannot

be considered to be a multiverse analysis report. Similarly, the process of building, selecting and tuning statistical models [24, 64, 17] 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 analyzed visually [97, 83]; *(ii)* crowdsourced hypothesis testing, where multiple research teams conduct independent studies to answer the same research question [56]; and *(iii)* meta-analysis [43], except when multiple meta-analyses are performed on the same set of studies [29]. 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 [30]). Resampling techniques (e.g. bootstrapping; see the DANCE example in [30]) 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. [32]). In addition, even if the analyses are heterogeneous in how they are conducted and reported (e.g. as in crowdsourced analyses [5, 10]), 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 [96], a sensitivity analysis carried out for model evaluation purposes [64], or an educational simulation illustrating how different analytical choices yield different outcomes [35]. 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 analyzing it in many different ways (e.g. ensemble learning algorithms [80]) are excluded, unless they explicitly convey the multiverse to human users. As stated initially, the multiple analyses need to be *reported*.

Figure 3.2 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 fulfill 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.

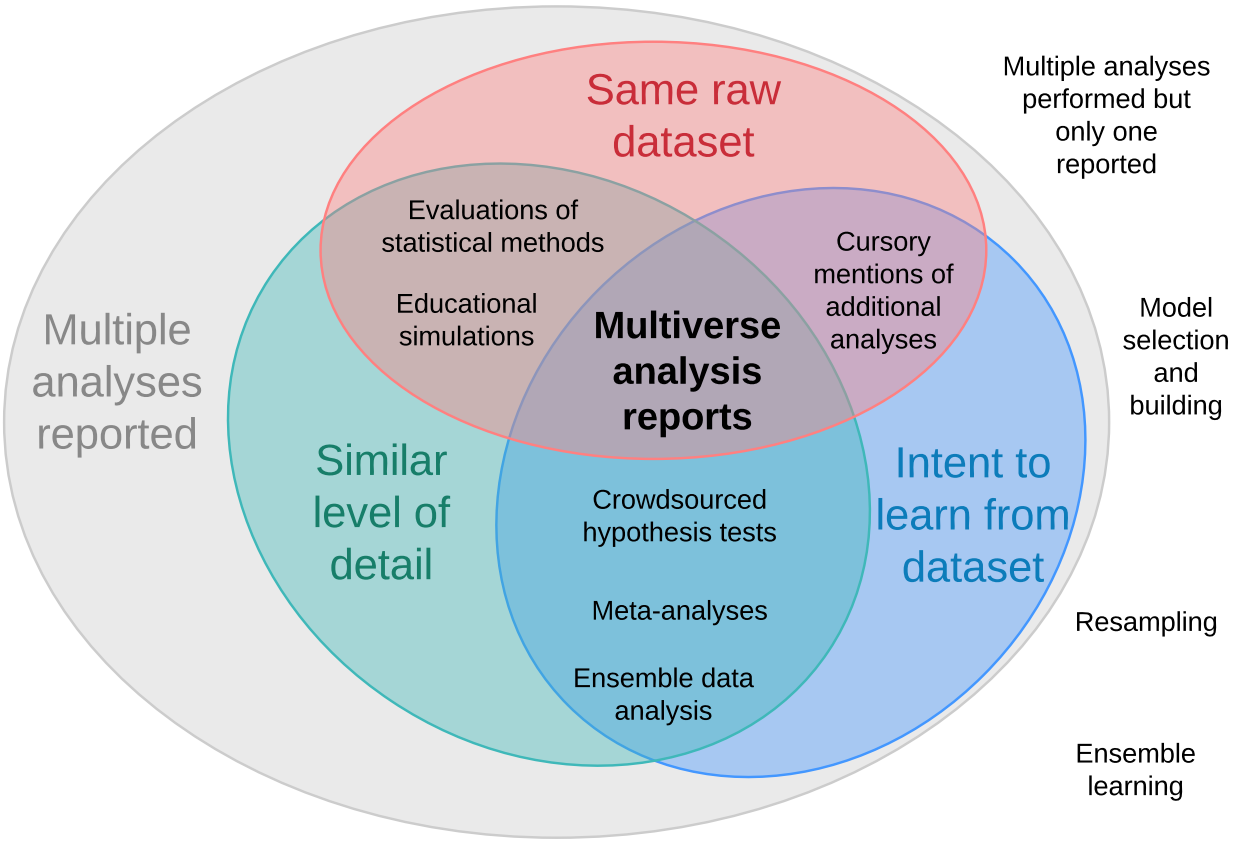


Figure 3.2: Overview of the four major criteria making up our definition of multiverse analysis report (each criterion is an ellipse), and examples of cases that fulfill some but not all criteria.

We additionally introduce the notion of a trivial multiverse analysis report:

> **trivial multiverse analysis report**: 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.

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 [30] 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:

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

> **parameter**: a characteristic of the reported statistical analyses that varies across the multiverse.

> **parameter value** or **option** [30]: a possible value taken by a parameter.

For example, suppose a paper uses three outlier exclusion methods to analyze 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:

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

> **outcome value**: 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:

> **multiverse analysis visualization**: any visual representation of the parameters, parameter values, outcomes, or outcome values of multiple analyses in a multiverse analysis.

Visual representation means that at least some of the information is visually encoded [67]. 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 [76]) 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 [67].

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.

3.5 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) analyze how well each archetype supports the tasks in our taxonomy.

3.5.1 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 3.3). 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 8 terms (Table 3.1). We then used this list in a systematic literature search using the Google Scholar API through the *Publish or Perish* software [47] 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 1,000 returned by the API. To keep the number of articles we would need to analyze 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 3.4, 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. 36 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 3.1).

More details on the corpus as well as the source lists from the systematic search are in the supplemental material of the original publication [45].

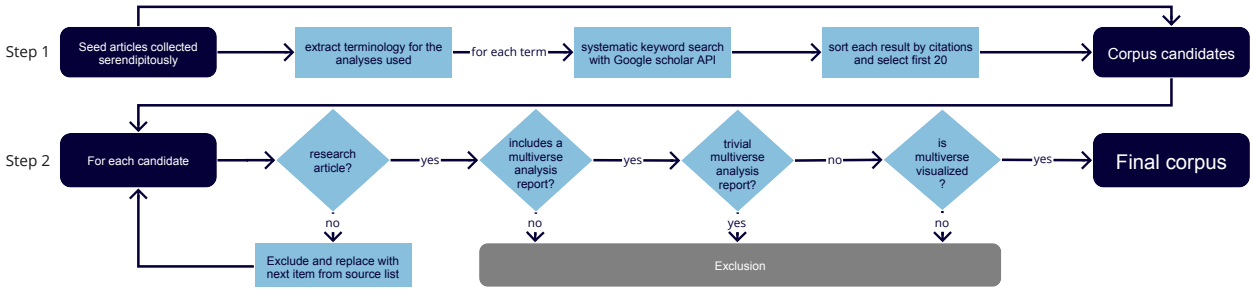


Figure 3.3: Overview of our curation process. In step 1, we curated a corpus of candidates by combining serendipitously discovered articles with a systematic keyword search. In step 2, we analyzed each candidate to identify all research articles that contain a non-trivial multiverse analysis report and illustrate that report with some form of visualization.

Search type – keyword	Search results	In final corpus	In both I. and II.
I. Serendipitous (seed articles)	53	36	12
II. Systematic:	>4,893	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	>1,000	0	0
– multimodel analysis	989	0	0
– perturbation analysis	>1,000	1	0
– sensitivity analysis	>1,000	0	0

Table 3.1: 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.

3.5.2 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 [88, 92, 75, 101, 85]. Each paper analyzed 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). 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 are presented in section 3.6.

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 [30, 58] and theoretical considerations of multiverse analysis [27] inspired the definition of the last category added to our taxonomy (Validate, subsection 3.6.5). Analysis of an additional set of 7 articles [59, 3, 78, 5, 10, 16, 69], 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 3.6. The source material, including a PDF export of the Miro boards, can be found in the supplemental material.

3.5.3 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 labeled 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 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 3.7. The full set of visualizations reviewed and scored are available as supplemental material.

3.6 Taxonomy of Analysis Tasks

We identified twelve tasks that can be performed using a multiverse analysis visualization, summarized in Table 3.2. 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. [88] as well as sample quotes taken from the corpus that were used to identify and synthesize these tasks.

3.6.1 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 was 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 sensemaking 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 3.6). 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. [75] (Figure 3.13), and elements of the Boba interactive interface [58] (Figure 3.15).

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.
Interpret	Interpretation tasks are logical and rhetorical inferences made about the meaning of any given set of results.*

Table 3.2: Overview of the taxonomy for multiverse analysis tasks derived from the multiverse analysis visualizations in our corpus. *The *Interpret* category was not formally included in the published version of this taxonomy, for reasons explained in subsection 3.8.2.

`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. [75] 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 1,000 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 3.13).

`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 3.3), 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.

3.6.2 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 definitions in section 3.4).

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 1,728 universes result

in a p-value below $.05$, 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 judgment in the domain of the analysis. Assessing to what extent outcome values vary across a multiverse typically requires judgments 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. [88] describe the results of completing this task: "The point estimates range from -1 to +12 additional deaths." Similarly, Steegen et al. [92] write: "For fiscal political attitudes ... the remaining choice combinations lead to p values across the entire range from .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. [88] state: "A 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. [88] 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. [92]: "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 subsection 3.8.1. Note that this task is closely matched to what Amar et al. [2] refer to as a "Characterize Distribution" task.

3.6.3 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 (subsection 3.6.2), 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 analyzing these few choices in greater detail. Further attention could either involve additional tasks described in this framework, or deeper theoretical considerations.

> **`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.

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. [92] 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. [85] describe the identification of two parameters that are sources of outcome sensitivity: "The teams also varied in their approaches to handling the nonindependence of players and referees, and this variability also influenced both median estimates of the effect size and the rates of significant results."

`Connect` ▷ `OutcomeFrequency`: 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. [85] 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 4 of these teams found a significant effect." Steegan et al. [92] 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` ▷ `SpecificOutcomes`: 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. [88] considered negative effect sizes in this way: "... we can see that obtaining a negative point estimate requires a fairly idiosyncratic combination of operationalizations."

3.6.4 Connect Combinations: Connect Combinations of Parameters to Outcome Values to Identify Potential Relationships

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 subsection 3.6.3).

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.

`ConnectCombo` ▷ `OutcomeRange`: 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. [92] 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. [101], 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 judgment about two key modeling assumptions: the inclusion of race and marital status."

`ConnectCombo` ▷ `OutcomeFrequency`: 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. [92] 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."

ConnectCombo ▷ 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 3.7. 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 analyzed 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 [101], 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 modeling distributions to consider". This concept of modality is also described by Patel et al. [75], referred to as "modality in the Vibration of Effects", and is given an equivalent interpretation: "We observed three modes in the association between triglyceride levels and mortality… The multimodal plots indicated that total cholesterol and diabetes were driving these modes."

3.6.5 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. [88] and Steegen et al. [92], 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 [58], support for this task is described: "Do we have evidence that certain outcomes are less trustworthy? We toggle the color-by drop-down menu so that each universe is colored 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 3 deaths, while the observed data contains plenty such cases. … we have reasons to be skeptical of the large estimates."

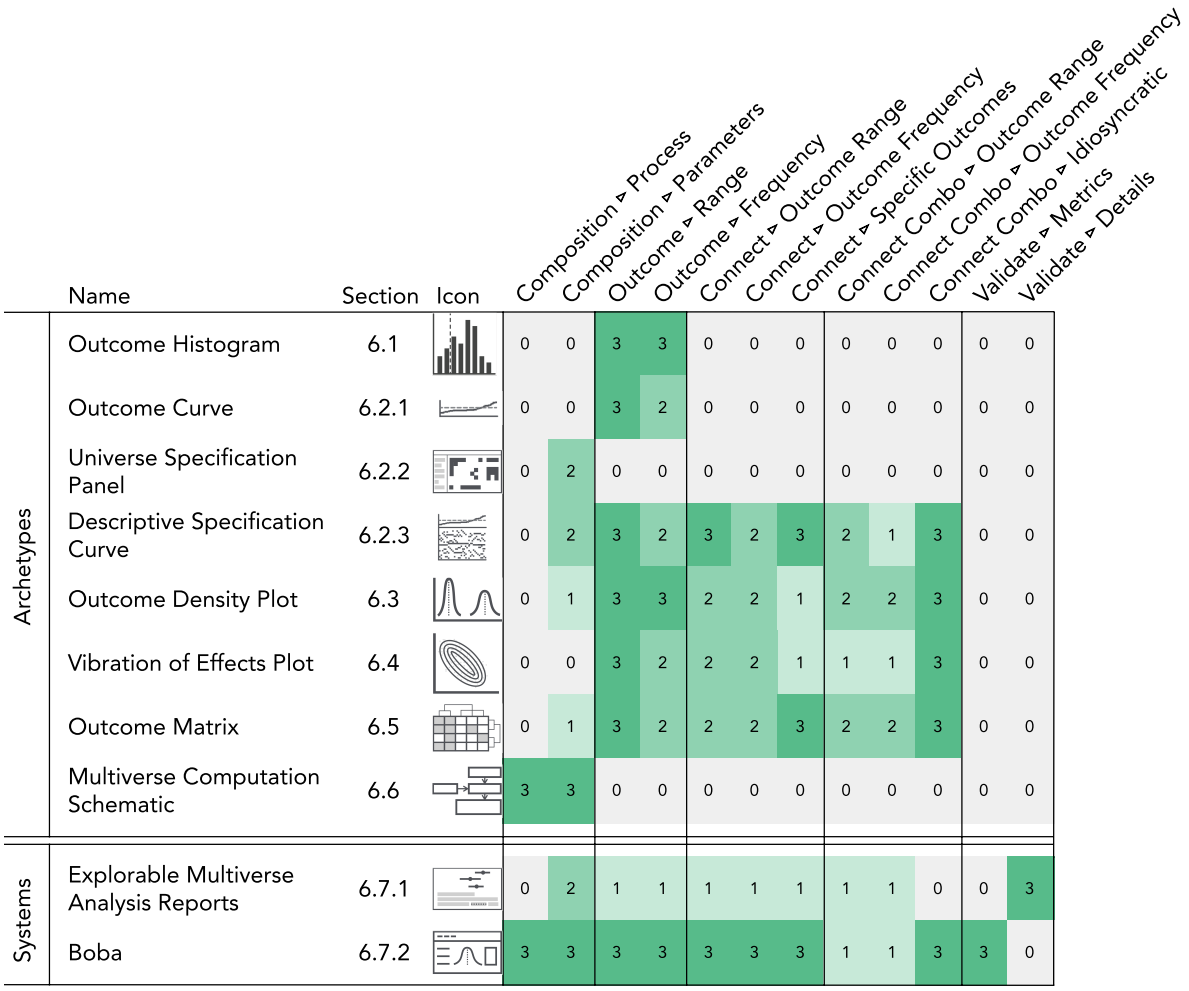
> **`Validate` ▷ `Details`**: assess validity of universes by examining the underlying details of analyses in each universe to interrogate their validity.

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 (EMAR) [30], 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 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.

3.7 Multiverse Visualization Archetypes and Systems



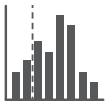
	Name	Section	Icon											
Archetypes	Outcome Histogram	6.1		0	0	3	3	0	0	0	0	0	0	0	0
	Outcome Curve	6.2.1		0	0	3	2	0	0	0	0	0	0	0	0
	Universe Specification Panel	6.2.2		0	2	0	0	0	0	0	0	0	0	0	0
	Descriptive Specification Curve	6.2.3		0	2	3	2	3	2	3	2	1	3	0	0
	Outcome Density Plot	6.3		0	1	3	3	2	2	1	2	2	3	0	0
	Vibration of Effects Plot	6.4		0	0	3	2	2	2	1	1	1	3	0	0
	Outcome Matrix	6.5		0	1	3	2	2	2	3	2	2	3	0	0
	Multiverse Computation Schematic	6.6		3	3	0	0	0	0	0	0	0	0	0	0
Systems	Explorable Multiverse Analysis Reports	6.7.1		0	2	1	1	1	1	1	1	1	0	0	3
	Boba	6.7.2		3	3	3	3	3	3	3	1	1	3	3	0

Figure 3.4: Overview of the archetypes and interactive systems described in section 3.7. 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).

We describe the set of multiverse visualization archetypes[2] identified in our analysis, along with the tasks they support. We also describe two interactive visualization systems designed to support multiverse analysis. A visual summary of task support is shown in Figure 3.4, and the process for arriving at these results is detailed in section 3.5).

[2]archetype: a class of multiverse analysis visualization designs that convey information about specific multiverse entities using a specific combination of visualization idioms [67].

3.7.1 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 3.5, the x-axis encodes the outcome values (here, point estimates of extra deaths for female hurricanes in the example of section 3.3), 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.

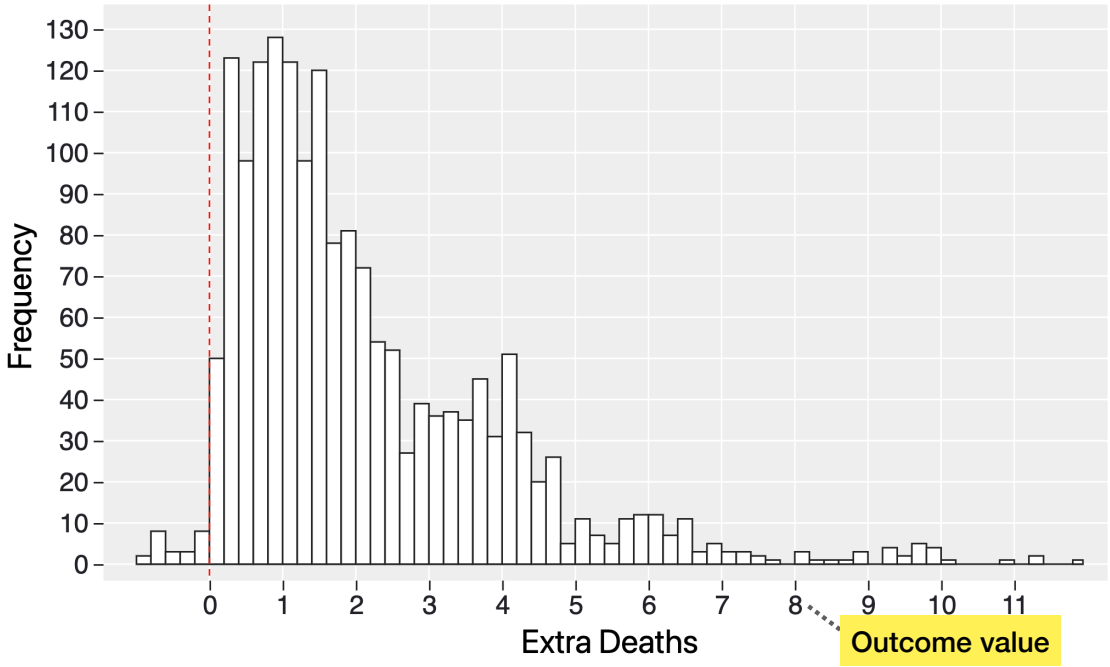


Figure 3.5: Example of an outcome histogram. Recreated after Steegen et al. [92], but using the hurricane dataset (section 3.3). The x-axis encodes outcome values (effect size estimates), while the y-axis shows the count across the multiverse.

The outcome histogram allows a viewer to easily and simultaneously complete both Outcome (subsection 3.6.2) 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 3.5 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 versus -1). However, with no mapping of parameters and options to outcomes, the viewer cannot explore which analytical choices are responsible for this variation.

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 subsection 3.6.2), 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 subsection 3.8.1.

The outcome histogram is a general approach that we encountered frequently in our corpus, e.g. [92, 27, 78, 12, 26, 95]. We also note one variation where the outcome is a p-curve [12], while Cirillo et al. [20] reported multiple varieties of this type.

3.7.2 Descriptive Specification Curve

The *descriptive specification curve* is an example of a *composite visualization*, which is a visualization that is made up of two or more linked components, each of which could individually function as stand-alone 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 3.7), or just the top panel (Figure 3.7a). Following Simonsohn et al. [88], we use *descriptive specification curve* to refer to the full composite. We first review each component individually before discussing the composite.

3.7.2.1 Outcome Curve (Component)

The core component of the descriptive specification curve is the *outcome curve* (Figure 3.7a). 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. [88] shown in Figure 3.7, dot color 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 1,728 universes: only those with the top and bottom 50 outcome values are shown, along with 200 other randomly sampled universes.

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

The outcome curve is commonly presented as a stand-alone visualization, e.g. [21, 51, 95, 13, 71, 28], especially in papers explicitly reporting a *specification curve analysis*. Simonsohn et al. [88] include 3 examples of the curve presented alone, with only one example of the full descriptive specification curve.

3.7.2.2 Universe Specification Panel (Component)

The second component of the descriptive specification curve is the *universe specification panel* (Figure 3.7b). It consists of a tabular visualization [76] 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 3.6 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.

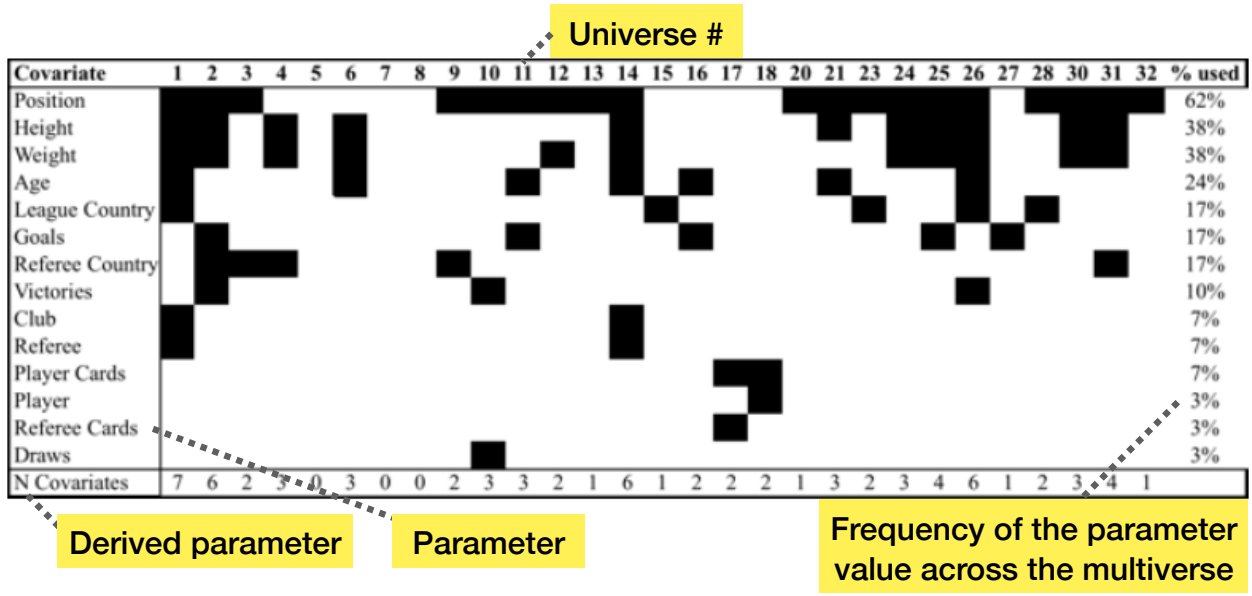


Figure 3.6: A variant of a universe specification panel [84]. Each column is a team of analysts (i.e. a universe) having analyzed the same dataset using different analytical choices, as defined by black cells indicating the selection of parameters values. The bottom row is the number of parameter values in each universe, and the rightmost column indicates the frequency of a given parameter value across the sparse multiverse.

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. [46, 43], but all have equivalent task support.

3.7.2.3 Descriptive Specification Curve (Composite)

Combined together on a common x-axis, the components above form the full composite *descriptive specification curve* (Figure 3.7), which allows the viewer to connect outcome values to analytical choices.

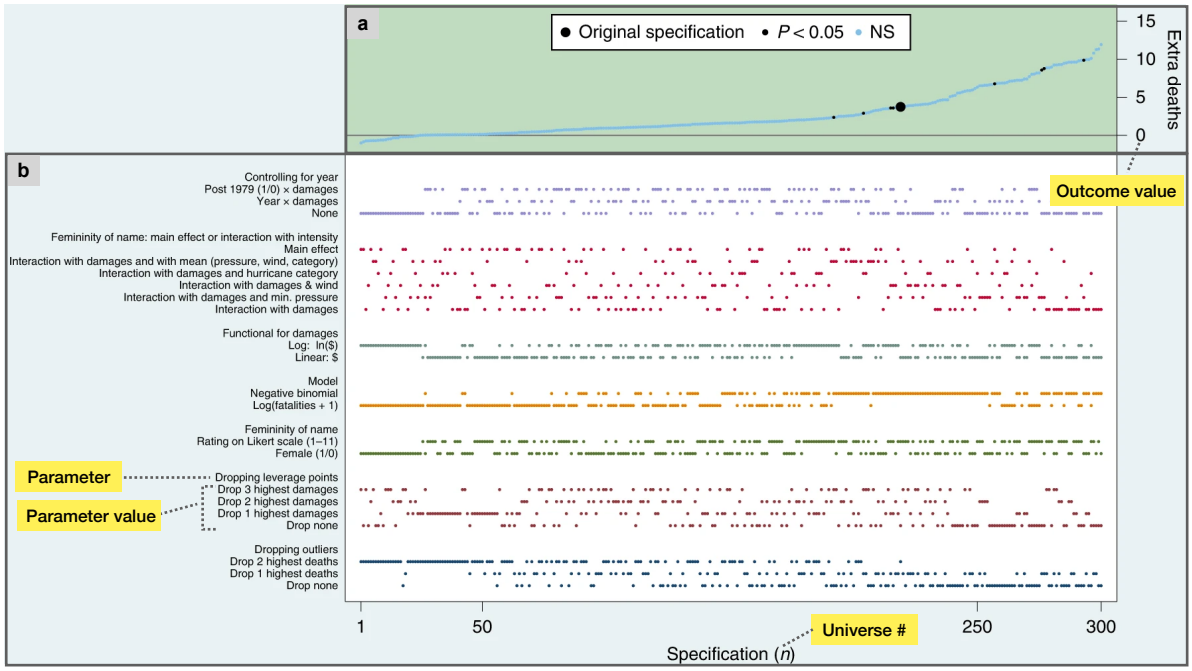


Figure 3.7: Example of a descriptive specification curve [88, 89]. We treat the full figure as a composite visualization that is made up of two components: (a) an outcome curve, (b) an universe specification panel. The composite visualization has super-additive functionality, enabling tasks that neither component supports by itself. 300 universes are shown here, out of the full multiverse of 1,728. The 50 universes with the smallest and largest outcome values are shown, along with a random sample of 200 other universes.

The composite supports all the tasks that its individual components support, but also supports all tasks in the Outcome and Connect categories (subsection 3.6.2 and subsection 3.6.3). Consider, for instance the *dropping outliers* parameter: at the bottom of the specification panel (Figure 3.7b), 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 3.7a) 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 form patterns.

This visualization also enables identification of Simonsohn et al. [88] 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 subsection 3.8.1.

3.7.2.4 Variants of the Descriptive Specification Curve

Figure 3.8 shows notable variants featuring interesting adaptations and improvements. In Figure 3.8a, statistical significance is color-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 3.8b maps significance to color but uses a three-color scheme that also indicates the sign of the effect.

Figure 3.8c also uses an error band and a different three-color 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 (subsection 3.6.4) easier compared to Figure 3.7 (so long as the desired combinations of parameter values are clustered together) while making tasks in Connect (subsection 3.6.3) more difficult (by disrupting the sorting within single parameter values).

Figure 3.8d presents a multiverse of meta-analyses, where each universe is one meta-analysis. The number of studies within each universe is color-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 color 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 3.8e 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

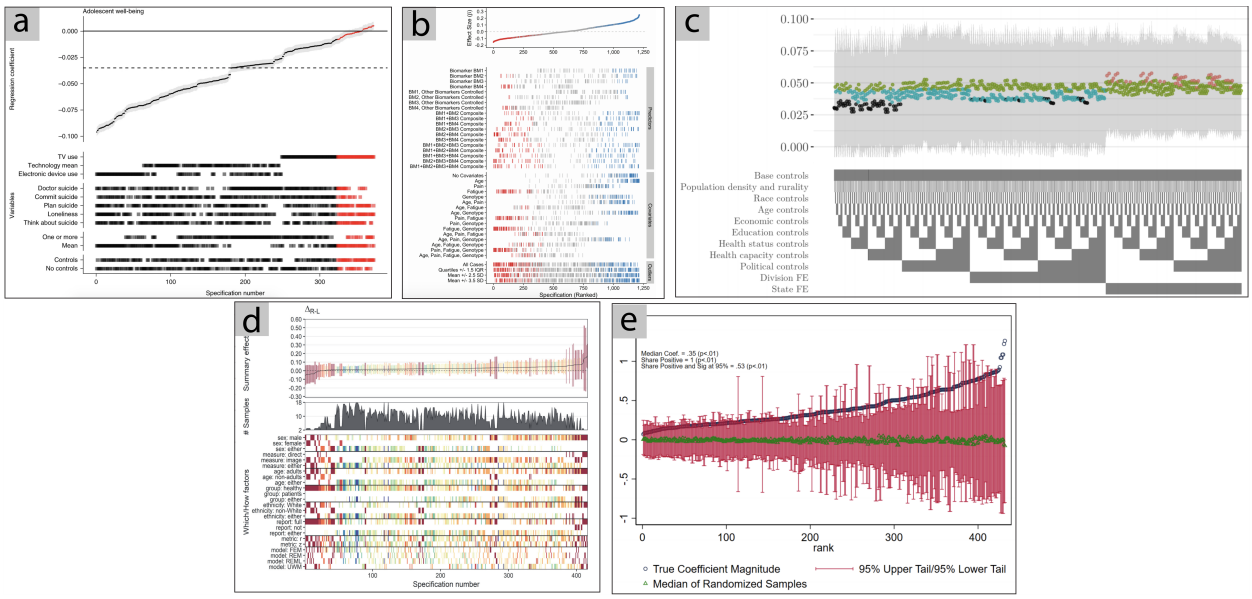
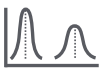


Figure 3.8: Example variants of the specification curve archetype, notable for their alternative mappings of the color channel and integration of uncertainty quantification metrics. (a) Figure 1 from Orben et al. [71], (b) Figure 5 from Del Giudice et al. [27], (c) Figure 7 from Burstyn et al. [14], (d) Figure 2 from Voracek et al. [95], (e) Figure 5 from Jelveh et al. [51]

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 [3], and dot-interval plot, e.g. Silberzahn's Figure 2 [85] 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 [71, 72, 70, 14, 27, 79, 95].

3.7.3 Outcome Density Plot



The *outcome density plot* shows the distribution of outcome values as a density plot. In Figure 3.9, 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 (subsection 3.6.3) and Connect Combination (subsection 3.6.4), by isolating subsets of the parameter space of interest. This also means that these tasks are only supported for the particular parameter

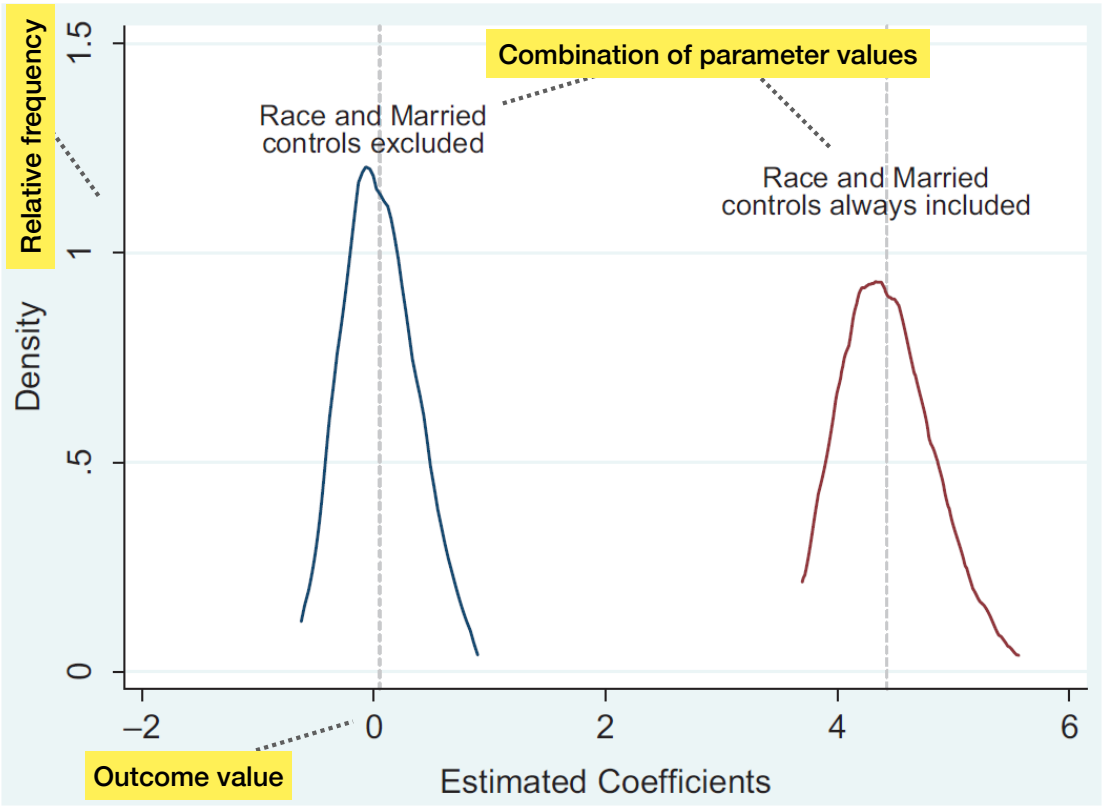


Figure 3.9: Example of an outcome density plot, from Young et al. [101]. Here, each density curve represents the relative frequency of outcome values across a subset of universes, defined by combinations of parameter values.

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 [59, 48, 69, 100, 66].

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. [101] identify variables for race and marital status as being especially important in their study, and use Figure 3.9 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 (subsection 3.7.4).

3.7.4 Vibration of Effects Plot

Figure 3.10 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 8,192 universes. In this *vibration of effects plot* (also called a *volcano plot* by Patel et al. [75]), 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. [75] use 2D binned heatmaps instead of scatterplots.

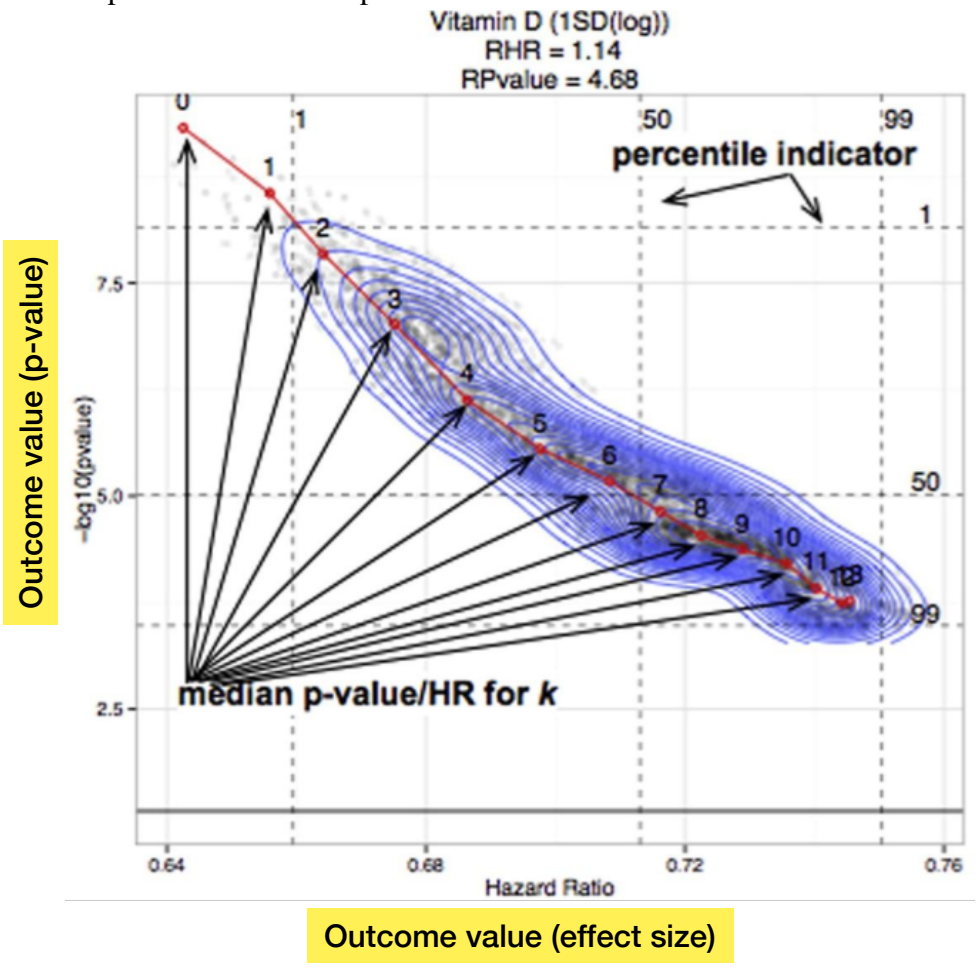


Figure 3.10: Example of a vibration of effects plot [75]. The x-axis encodes outcome values (effect size estimates), and the y-axis encodes the statistical significance (negative log transform of p-value). Blue contour lines show the relative frequency of outcomes within the multiverse.

All tasks in Outcome (subsection 3.6.2) are well-supported by this plot to the extent that density contours and overplotted scatterplots support frequency estimation. All tasks in Connect (subsec-

tion 3.6.3) are supported with comparable ease, and in much the same way, as the Outcome Density plot (subsection 3.7.3). Similar caveats apply: generally only a small set of combinations of parameter values can be compared at once, e.g. by mapping parameters to colors (Figure 3.11). 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.

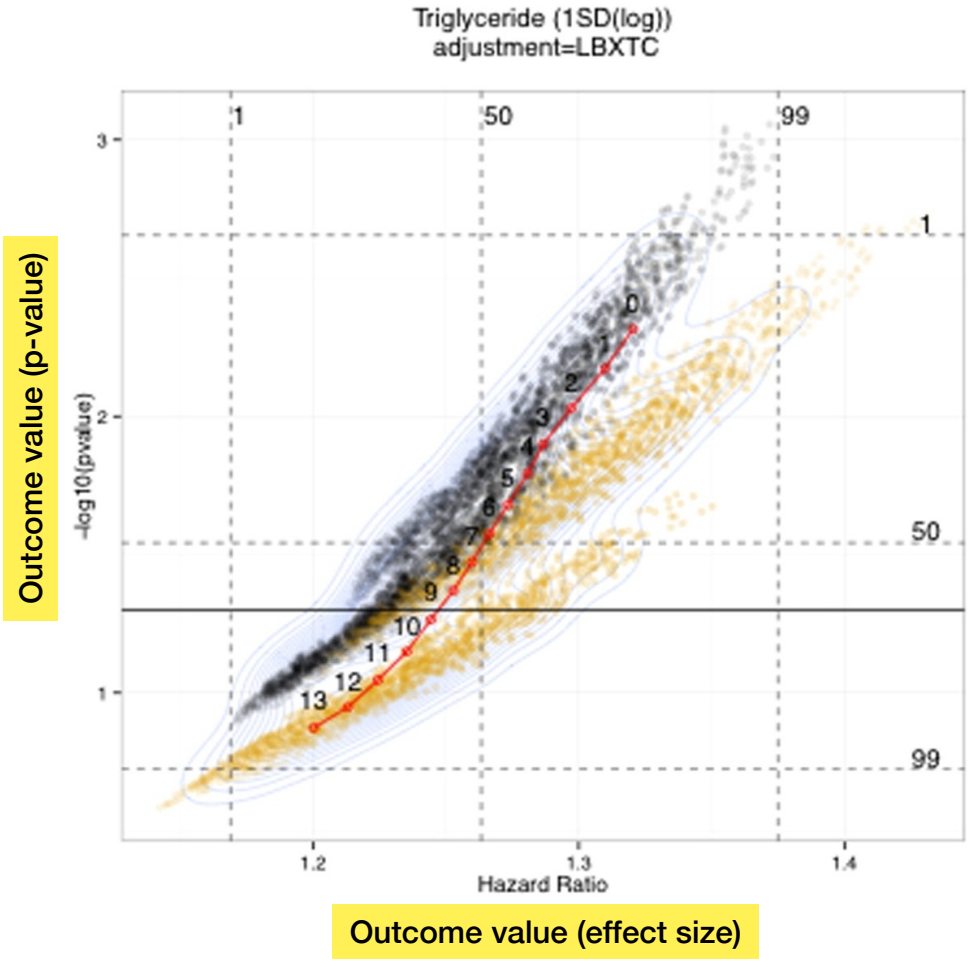


Figure 3.11: Variant of the vibration of effects plot [75] where each universe mark is color-coded to indicate the value of the parameter value for that universe. Marks that are gold indicate the parameter (triglyceride) was included in that model, while black indicates exclusion.

The identification of potentially important clusters in outcome values is an example of the task `ConnectCombo` ▷ `Idiosyncratic`. Patel et al. [75] dedicate extensive discussion of visual patterns exhibited by vibration of effects plots and their interpretation. For example, while the color coding of parameter values in Figure 3.11 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 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. [75] describe how hundreds of such figures are to be generated to this end.

Patel et al. [75] 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. [75] 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 labeled 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 subsubsection 3.7.2.2). 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. [75] reported finding cases where more adjusting variables were associated with smaller effect sizes, larger effect sizes, 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. [27], but this was a near-exact reproduction of the style of this archetype that differed primarily in color choice.

3.7.5 Outcome Matrix

An outcome matrix is a tabular visualization [76] where both rows and columns are parameter values, and each cell reports an outcome value. In Figure 3.12 each cell reports a p-value, both using numerals and a color (statistically significant in gray). 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 3.12 is an example of the *outcome matrix* from Steegen et al. [92], 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* (subsection 3.7.1). 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 color coding of the outcome values supports `Outcome ▷ Frequency` (here, the more gray 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 color coding, although known issues with heatmaps may make certain tasks difficult [55, 37].

Tasks in Connect (subsection 3.6.3) 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 3.12 spans 5 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 3 non-adjacent columns). The ease of connecting

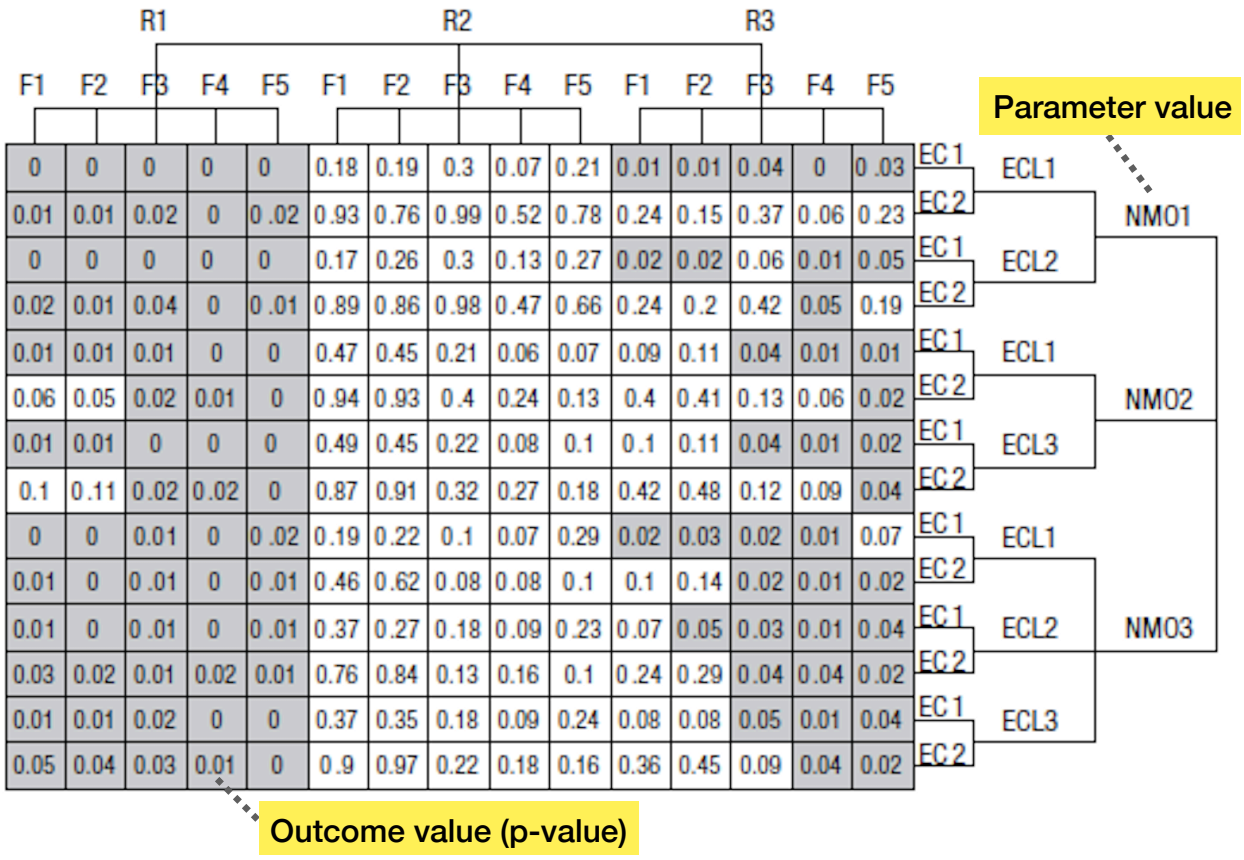


Figure 3.12: Example of an outcome matrix [92]. The double-dendrogram structure encodes parameter specification: each level is a parameter, and each node at a given level is a parameter value. Each cell in the matrix thus corresponds to a universe, and indicates the outcome value for this universe (also color-coded).

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 3.12 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 were 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 [6].

3.7.5.1 Variants of the Outcome Matrix

Variants of the outcome matrix in our corpus were generally less structurally complex than the example shown in Figure 3.12, 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 [18, 25, 28]. Multiple variants used continuous outcomes and applied different color maps (e.g. diverging palette for positive-negative effect and magnitude), illustrating how this archetype is not fundamentally limited to binary outcomes types [31, 28].

3.7.6 Multiverse Computation Schematic

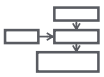


Figure 3.13, also from Patel et al. [75], is an example of the *multiverse computation schematic* archetype. This is one of the few archetypes whose focus is on Composition (subsection 3.6.1)—as opposed to reporting outcome values—providing the most support for the tasks Composition ▷ Process and Composition ▷ Parameters in our corpus.

Each panel of Figure 3.13 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 outcome 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 (subsection 3.7.4). 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.

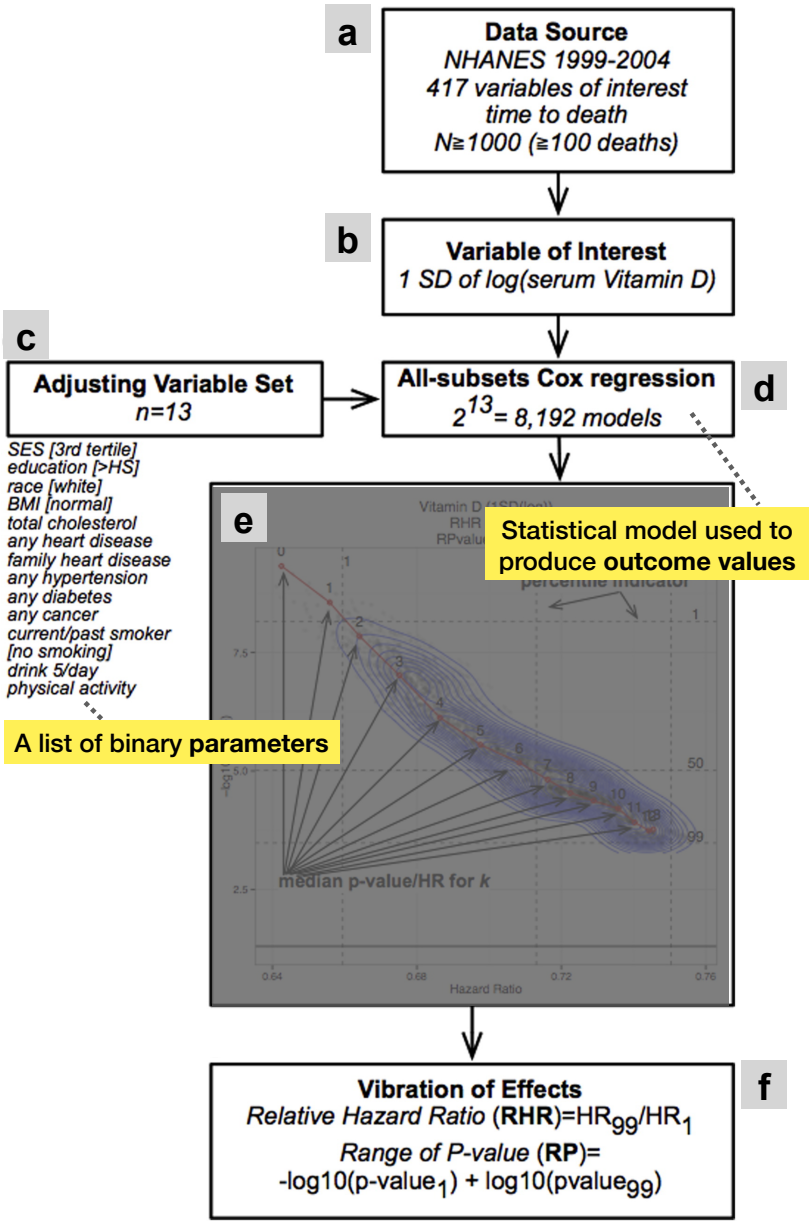
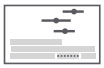


Figure 3.13: Example of a multiverse computation schematic [75], describing data source (a), variable of interest (b), and parameters (c,d) composing the multiverse; and elements of the multiverse analysis report: a vibration of effects plot (e); and measures of outcome value spread (f).

3.7.7 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 (subsection 3.6.5), as these tasks are largely unsupported by the other visualizations in our corpus.

3.7.7.1 Explorable Multiverse Analysis Reports (EMAR)



Explorable Multiverse Analysis Reports (EMARs) (Figure 3.14) are interactive variants of academic articles inspired by *explorable explanations* [93]. 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 3.14 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.

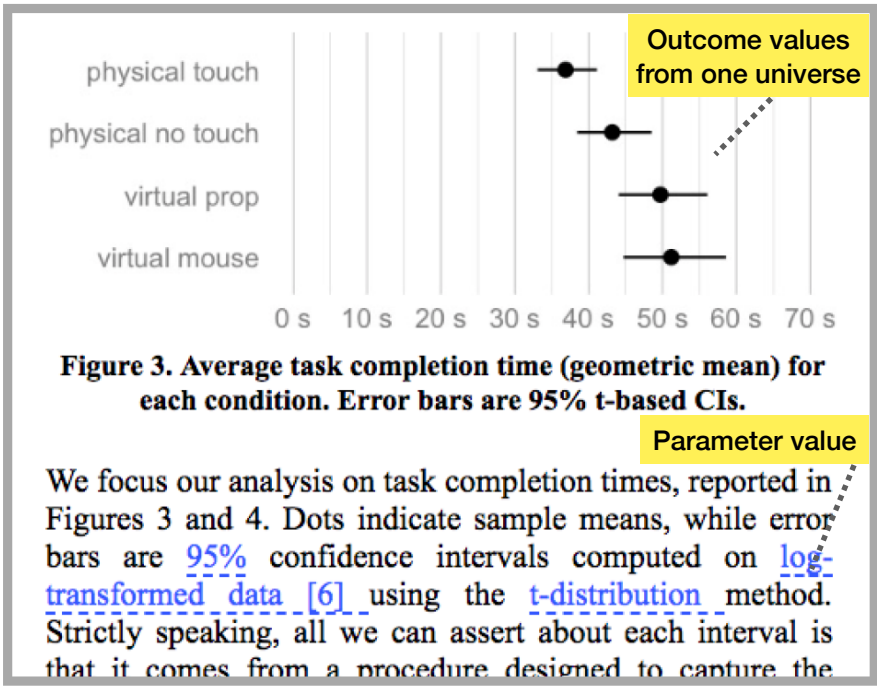


Figure 3.14: Excerpt from an explorable multiverse analysis report [30], where parameter values can be selected dynamically through interactive text widgets, resulting in figures, numerals and text updating accordingly in the report.

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 judgments about the validity of each universe (`Validate` ▷ `Details`). However, this can make it more difficult to gain a higher-level understanding of outcome sensitivity (subsection 3.6.2). 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`).

3.7.7.2 Boba

Boba (Figure 3.15) 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 (subsection 3.6.3) by allowing viewers to interactively select parameters of interest (Figure 3.15c), which it uses to show dotplots of outcome values faceted by parameter values (Figure 3.15d). It has some support for Connect Combination (subsection 3.6.4) 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.

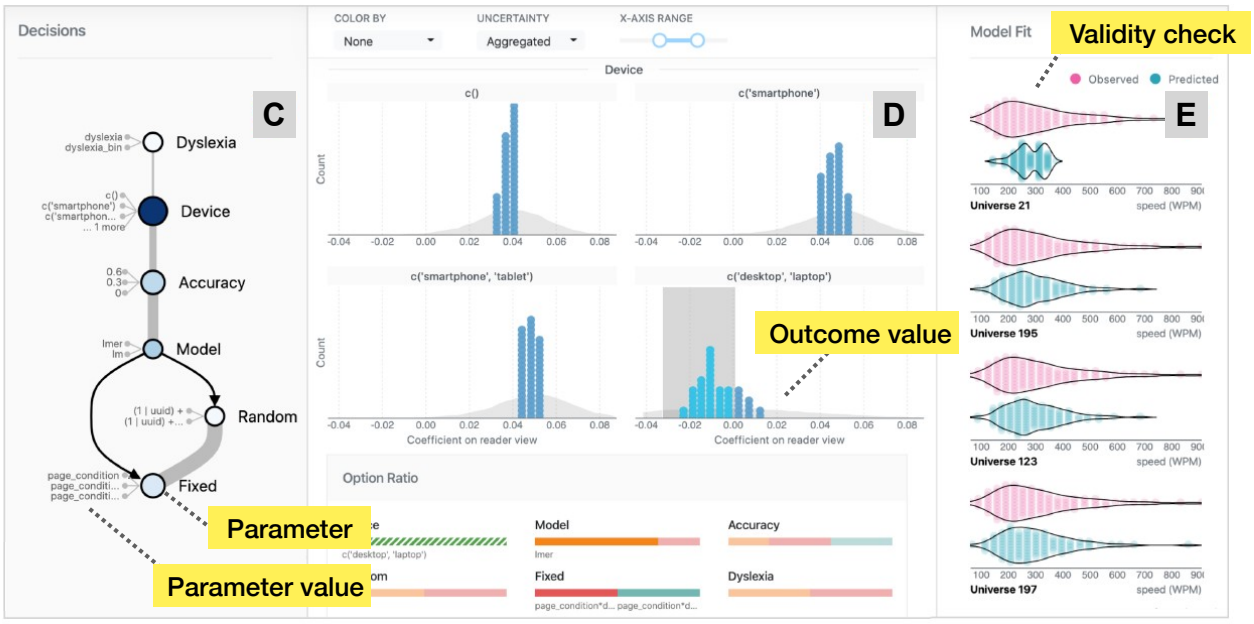


Figure 3.15: Screenshot of the Boba system [58]. Panel C shows the design space of parameters and their relationships; parameters that are source of sensitivity are in a darker color. Panel D is a trellis of dotplots of outcome values, subsetted by parameter values. Panel D shows predictive distributions from each universe compared to the observed data.

A unique contribution of this system is that it explicitly considers model fit (Figure 3.15e) 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 [27]. Support for this task is provided by allowing the viewer to examine model fit (Figure 3.15e) and exclude outcome values from poor-fitting models in the final interpretation.

3.7.8 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 [16, 4, 77, 5]. 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.

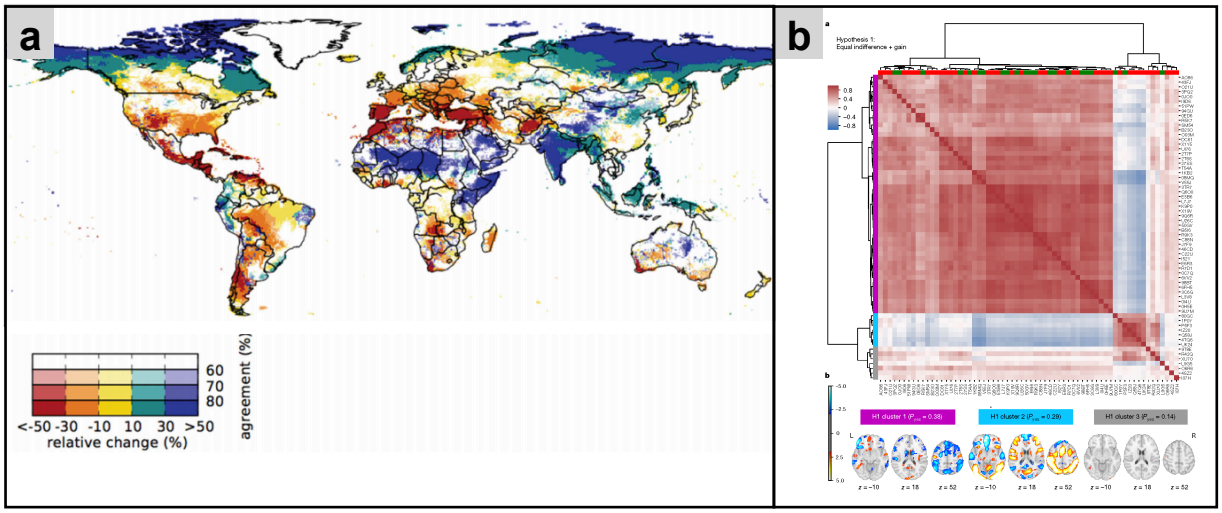


Figure 3.16: Two examples of domain-specific visualizations of multiverse analyses. (a) outcome values are contextualized in a geographical map [8], (b) correlation matrix of outcome values [10].

Figure 3.16a shows the output of water runoff (discharge) predictions from 55 climate models [8]. Outcome values and sensitivity are encoded on a bivariate color 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 3.16b 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 (subsection 3.7.5), but are the result of a clustering algorithm rather than a direct representation of parameters. The color-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, subsection 3.6.3).

3.8 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.

3.8.1 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 [88]. Authors and readers may even disagree on the validity of some universes [88]. 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* [36]: 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 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 [34]; see Bonneau et al. [9] for further discussion of possibilistic versus probabilistic uncertainty visualization. We have not seen examples of possibilistic uncertainty visualizations applied to multiverse analysis as yet.

3.8.2 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. [27] 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 [88, 92], but most multiverse reports we reviewed did not go beyond tasks from the Outcome (subsection 3.6.2) and Connect (subsection 3.6.3) categories, or at least not in a way that explicitly referenced a visualization.

Only 2 visualizations provided support for tasks under the Validate category (subsection 3.6.5). 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. [58] suggest doing so by examining model fit and provide some support for this task in Boba (Figure 3.15). They suggest 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. [27] 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. [88] 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.

3.8.3 Multiplexing and Interaction to Investigate Parameter Combinations

Few visualizations provided substantial support for tasks in the Connect Combinations category (subsection 3.6.4). For most visualizations, this support comes with the caveat that meaningful combinations of parameters have been selected ahead of time (e.g. Figure 3.9), 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 (subsection 3.7.4) can only compare across a small number of parameter values at once, Patel et al. [75] 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. [78] faceted by both variables of interest and parameters, producing an 8-by-6 of outcome histograms (subsection 3.7.1) 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 (subsection 3.6.3 and subsection 3.6.4).

Boba (Figure 3.15) 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 3.12), for example, could benefit from interactive row and column reordering to aid in cluster identification [76]; similar functionality could also help reduce tradeoffs in fixed column ordering on specification curve charts (e.g. Figure 3.8a versus Figure 3.8b). Such approaches could be used in interactive systems aimed at analysts, like Boba [58], or incorporated into interactive reports aimed at readers, like EMARs [30].

3.8.4 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. [85]). Some archetypes explicitly visualize this structure (e.g. the dendrograms in outcome matrices; Figure 3.12) 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 3.5).

Part of the inherent difficulty of multiverse analysis is that the data is 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 (subsection 3.7.1). As discussed previously (subsection 3.7.1 & subsection 3.8.1), 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 (subsection 3.6.3). 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 (subsection 3.7.4) 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.

3.8.5 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 3.4. Since adjacent concepts, such as model comparison, or parameter space exploration (also see Figure 3.2) likely entail different tasks, we curated our corpus by strictly applying the definitions presented in section 3.4. The eight relevant keywords identified from our list of 53 seed articles resulted in a total of 213 corpus candidates. In analyzing 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 3.2). 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 [24, 64, 17] and for ensemble data analysis [97, 83] 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 was analyzed, 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 analyzing 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 analyzed 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 modeling 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 (subsection 3.7.8), considerable innovation may be required.

3.9 Chapter 3 Conclusion

This chapter serves as a state of the art report reviewing 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. It covers a survey of literature across multiple fields and disciplines, considering visualizations from areas as diverse as psychology, statistics, economics, and visualization.

As a matter of general contributions, this chapter includes a coherent and operational terminology to provide researchers with a common vocabulary so they can better communicate and reason about multiverse analyses (section 3.4). It also provides an assembled a taxonomy of analysis

inspection tasks that multiverse visualizations should support, grounded in an extensive analysis of the curated corpus (section 3.6). It also provides a detailed discussion of the design and functionality of major multiverse visualization archetypes, including an assessment of how well each of them supports multiverse tasks (section 3.7), in order to guide analysts in the selection of appropriate visualizations to use when conducting or reporting multiverse analyses. 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.

All this said, if the derivation of meaning from results is the ultimate goal of analysis, and interpretation of a multiverse analysis is usually not a trivial task, why are there no visualizations that explicitly support interpretation tasks? What would it take to support such tasks?

While multiverse analysis is a promising method for examining analytical uncertainty, substantial challenges remain. One particular challenge is the limited support for the Validate and Interpret categories of tasks, which are the tasks that are ultimately essential for deriving useful meaning from a multiverse analysis and communicating that meaning to others. So the question to be answered in the remainder of this work is: what will it take to support validation and interpretation tasks?

The first step towards answering this question is to consider what about the tasks might be making them so difficult. I believe the apparent difficulty and complexity of validation and interpretation tasks indicates that these tasks are fundamentally different from the other multiverse tasks, particularly in that they often require higher-level mental reasoning and connection of disparate facts, some of which may not even be encoded as specific data. For example, consider what it would take to answer this question: is a specific analytical choice statistically appropriate, given the data and in the context of this specific multiverse? While this question is fundamental to a multiverse analysis and makes reference to data, it also requires logical and theoretical considerations that are not a feature of any given dataset. A task like this can still be supported by a visualization, but it may require multiple visualizations and views of the data at a universe-level or multiverse-level, and even then it cannot be decomposed to the same degree that a task like Assess Outcome Sensitivity can be. This suggests that there may not be a single visualization archetype that can support a task like that, but instead there may be a set or series of inter-related visualizations that would be helpful instead.

The conclusion of this chapter is two-fold. First, validation and interpretation tasks are higher-level tasks that can require multiple perspectives on the data, the multiverse itself, and consideration of things that are not even in the data itself (e.g. statistical methodology, theory specific to the subject). As a consequence, these tasks may benefit from or even require collaboration between

people with different areas of expertise, and as they may be especially cognitively demanding any reduction in effort required to complete dependent tasks in support of them could be especially helpful. Second, given that validation and interpretation tasks may be require multiple perspectives on the data, it might be better to consider a series of visualizations to better support all dependent tasks collectively.

The first conclusion motivates the search for ways to ease the cognitive burden necessary to explain, understand, and have discussions around data in general. The second conclusion motivates the special attention on exploring a series of visualizations as a way to collectively support validation and interpretation tasks. These two threads are explored simultaneously in the penultimate chapter of this work.

CHAPTER 4

AugMeet: Augmented Presentation System to Support Validation and Interpretation Tasks in a Multiverse Analysis

4.1 Chapter Introduction

The multiverse analysis tasks of validation and interpretation are necessary for deriving meaning from the multiverse, and the degree to which they are successfully completed essentially determines how much value can be obtained from the process of conducting and sharing it with others. The support previously reviewed multiverse visualizations have for these tasks is particularly limited (as detailed in chapter 3), which testifies to just how challenging these types of tasks can be in practice. While this could be a limitation merely of published visualizations, reconsideration of the papers associated with those visualizations finds that most show no indication that validation and interpretation tasks were completed by their authors through any means. The difficulty of these tasks stems from the fact that validation and interpretation requires a higher level of thinking than other multiverse tasks (like assessing outcome sensitivity), and often involve multiple lines of reasoning informed by multiple perspectives on the data and phenomena of interest. Rather than trying to overload a single visualization with information, which could make otherwise simple tasks harder to perform while causing high cognitive load, it may instead be fruitful to explore how a series of visualizations can be designed to collectively build up the network of connections necessary to convey and support arguments common among multiverse analyses.

When considering the practical usefulness of multiverse analysis, it is also important to consider that multiverse analyses may often be conducted by a team of professionals, as evidenced by the fact that no multiverse-related papers in the previously surveyed corpus (chapter 3) are single-authored. Analysis may also benefit from experience in statistical methodology as well as additional subject matter expertise [92], which would be an additional potential benefit of collaboration; in some topic areas, teams are even composed of such highly specialized sets of skills that

group coordination is a necessity [33, 1]. Insights from group discussion can be used to guide what analyses are performed next, and conclusions are either devised collectively or require assent of the group before they are reported externally [57, 33, 1]. All these observations suggest that it may prove fruitful to intentionally design systems to support multiverse analysis done by a group.

The other line of observation that motivates this work is that validation and interpretation tasks are inherently difficult and cognitively demanding (as discussed in section 3.9). Using the example of a validation task that asks 'is this set of choices statistically appropriate?', completing this task may require multiple perspectives on the data at both a universe and multiverse level, in addition to consideration of theory-based concerns that are are not encoded in the data itself. When tasks are as demanding as this it can be be particularly helpful to reduce the cognitive burden of prerequisite tasks as much as possible, to preserve mental energy for where it is needed most. A previous work of this author [44] established that a technique I now call *augmented presentation*[1] is a way to make it easier for people to pay attention and understand information provided in data visualizations, and in a way that encourages engagement and can help to facilitate discussion. However, the evaluation of that system focused on only the most common types of visualizations, concepts that were much less demanding than the ones considered here, and was not designed or evaluated in the specific context of a multiverse analysis.

This chapter describes the design and evaluation of a system to support teams conducting a multiverse analysis, with particular focus on the stage of analysis where the results of a provisional exploration are presented to the working group. Here the multiverse analysis will not be considered a singular end-point, but rather an iterative process of questioning and learning from all of the data—both the underlying data used within universes and the multiverse of results itself. The group will be interrogating the multiverse, which is to critically re-examine the analytical choices in light of what can be learned from the multiverse analysis. Validation and interpretation tasks are central to the focus of this chapter, and other multiverse tasks are featured only to the extent they are necessary to the completion of the focal tasks. Special consideration is also given to the pragmatic concerns of productivity and progress, and the efficient use of meeting time and direction of group effort is also given substantial attention.

[1] augmented presentation: the performance of a presentation aided by live video augmentation

4.2 Research Question

RQ 5.1: How can validation and interpretation tasks be supported by using augmented presentation techniques together with a series of related data visualizations?

To answer this research question, I developed an interactive system named AugMeet. I recorded a presentation I created and delivered using AugMeet to serve as an applied demonstration of the system, and which depicts a person presenting the first results of a multiverse analysis to a mock-meeting of researcher collaborators. I then conducted an evaluation study by holding semi-structured interviews with seven participants that had relevant expertise, with each participant being highly-educated, having 5-20+ years of research experience, and having familiarity with multiverse analysis ranging from being somewhat familiar to having published papers about the topic. During the interviews all aspects of the system were discussed and critiqued.

The AugMeet system combines multiple ways to support the completion of validation and interpretation tasks, each of which are described in the results of the evaluation study (section 4.6). First, it uses augmented presentation as a technique to make it easier for the audience to follow along and understand information provided about and through data visualizations, while simultaneously making it easier for people to pay attention and encouraging them to be interested and engaged. Second, it uses a series of visualizations to build up a shared understanding of the facts, provide an overview of what is being considered, focus effort on what is most potentially impactful, and support low-level interrogation of twin universes to provide a new perspective and source of critical information about the multiverse.

The visualization series demonstrated by AugMeet integrates a variety of novel techniques and concepts to support validation and interpretation tasks, including: the concept of *twin universes* (subsection 4.4.2), *parameter-faceted outcome curves* (Figure 4.2), *two-dimensional option-faceted outcome curves* (Figure 4.3), and *twin-faceted residual plots* (Figure 4.6). AugMeet combines all of these individual aspects into a single system, and the demonstration of the system also shows a workflow that participants identified as being of further value due to its approach of progressive, iterative interrogation of a multiverse to reduce uncertainty.

4.3 Use Case: Research group critically evaluating hurricane gender-name effects

The use case described below is used throughout this chapter, and centers on a fictional remote meeting of a team of professional researchers collaborating on an ongoing research project. This use case provides the context for the presentation that was created and performed as a demonstration of AugMeet. The demonstration presentation was recorded and shown to participants in the evaluation study of this system (section 4.5), and the use-case was described similarly to participants for shared context. This use case deals with the same topic of hurricane gender-naming as used in section 3.3, but is approached from a different angle here so review of that earlier example is not necessary for the purposes of this chapter.

In this use case, a team of researchers are currently re-evaluating the statistical results taken from a real published paper titled *Female hurricanes are deadlier than male hurricanes* by Jung, et al. [53]. The dataset being re-analyzed includes information about past named hurricanes in the USA, including their associated death tolls, dollar damages, and the femininity of their names. The original study estimated the association of hurricane name femininity with deaths not otherwise predicted by the dollar value of damages sustained, finding that female-named hurricanes were associated with more deaths regardless of dollar damages. Such a finding is consistent with a hypothesis that people take female-named storms less seriously, and thus may be inclined to take more risks than they otherwise would in the face of such storms.

The presenter leading the meeting has planned to discuss the results of a multiverse analysis on the aforementioned dataset, which is a method that involves running multiple different analyses using justifiable alternate data analysis choices. Each category of potential analysis choices is referred to as a *parameter* and each choice in that category is called an *option*. The analysis discussed in the meeting involves 6 parameters with 2-4 options each. Combining one option from each parameter results in a single statistically estimated effect size (called an *outcome*), and each unique combination of options is called a *universe*. The multiverse initially discussed in this meeting is made up of 2016 distinct universes, giving a wide range of estimated outcomes.

The multiverse specification in this use case is a slightly modified version of a multiverse published by other authors [81], which was itself based on the example used in the specification curve publication [88]. The goal of the depicted meeting is to discuss the multiverse, and to further discuss whether all the included choices should still be considered reasonable.

4.4 AugMeet

In order to support validation and interpretation tasks in a multiverse I developed AugMeet, a system for live augmented presentation of interactive multiverse visualizations in remote meetings. The AugMeet system is the result of an iterative design and development process which used the previously developed AugChiro [44] system as a starting point, and all new design concepts described here were developed specifically for supporting communication around a multiverse analysis to complete the focal tasks of validation and interpretation. As with AugChiro [44] before it, AugMeet is implemented as a browser-based presentation environment developed in JavaScript that composites presenter webcam video with interactive visualization overlays that respond to a presenter's bodily actions, which is a technique I will refer to more generally as *augmented presentation*. The presenter can use their hands to interact with the data visualizations to highlight or show additional information about a data point, transform the visualization, or transition between visualizations. The system is designed to be controlled by a single presenter in a live remote meeting, with the single composite view of themselves behind the data visualizations being shared with meeting attendees via standard video conferencing software (e.g. Zoom, Google Meet).

The following subsection explains the system as it was demonstrated to participants for the evaluation study (section 4.5), and is illustrated using screenshots taken from that same demonstration video. Subsequent subsections describe the aspects of AugMeet that are new and distinct from the previous AugChiro [44] system.

4.4.1 Demonstration

The demonstration of AugMeet described here depicts a presenter leading a group meeting whose agenda is to interrogate the multiverse—that is, to complete validation and interpretation tasks, with a focus on questioning whether all initially included options are indeed appropriate. The description of the demonstration of AugMeet that follows is divided into seven scenes to match the seven video segments shown to participants during the evaluation study. As a brief overview, the presenter first describes the overall results of the initial multiverse analysis, then leads the group step-by-step through some arguments the presenter had initially planned in advance, followed by depicting alternative considerations and digressions suggested by colleagues. The group ultimately comes to a decision about two parameters, and the meeting closes with a summary of findings and suggestions for next steps.

4.4.1.1 Scene 1: Assess initial multiverse of outcomes overall

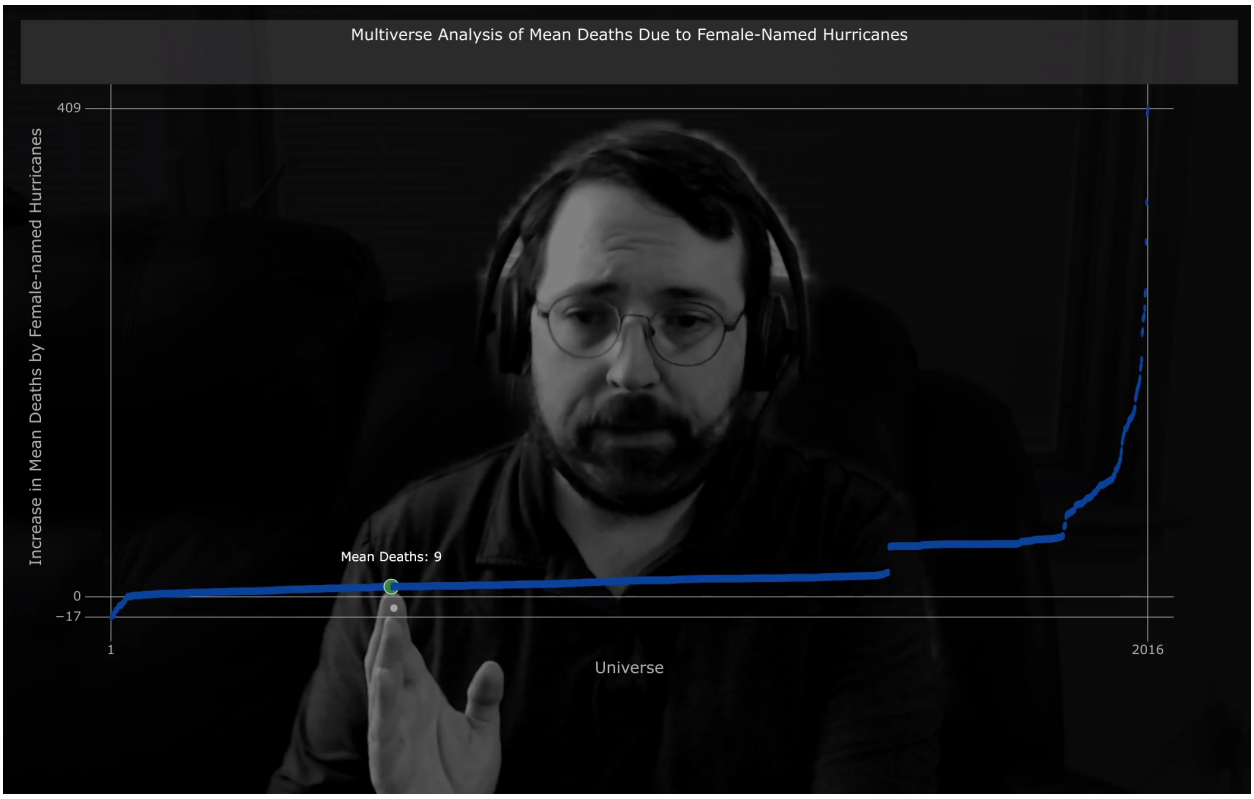


Figure 4.1: Screenshot from AugMeet demonstration Scene 1, featuring an outcome curve (commonly called a specification curve) that shows all of the effect sizes that are outcomes of the example multiverse analysis. For detailed description of this plot, see subsubsection 4.4.1.1.

Using an outcome curve visualization (also known as a specification curve, subsubsection 3.7.2.1), the initial multiverse of results are visualized as shown in Figure 4.1. Each universe is a single mark, with the universes ordered by effect size and distributed evenly along the x-axis so they can all be seen. The outcome shown here on the y-axis is an effect size that is defined as the estimated difference in mean deaths associated with female-named hurricanes compared to male-named ones, and ranges from -17 to 409. The presenter notes the large range of effects sizes and that some seem implausibly large, and that this multiverse includes both universes that indicate male names are a deadlier (effect sizes that are negative numbers) and ones that indicate female names are deadlier. Figure 4.1 is also used to point out a few potentially interesting features of the distribution: the comparatively small tail of negative effect sizes on the far left of the plot; the discontinuity in effect sizes visible as a gap appearing on the right third of the plot; and the large upper tail of outcomes on the far right of the plot.

4.4.1.2 Scene 2: Comparing potential impactfulness of parameters

Figure 4.2: Screenshot from Scene 2, featuring an example of *parameter-faceted outcome curves*, where each mark is colored according to the option level values within that parameter. For detailed description of this plot, see subsubsection 4.4.1.2.

The presenter states that the multiverse is now visualized in a different way (Figure 4.2), which is a grid of six outcome curves where each cell of the grid is assigned to a single parameter of the multiverse; I will call this visualization design *parameter-faceted outcome curves*. Within each individual cell the universe marks are colored according to level value of each option in that parameter, such that the first option for that parameter receives the first color, the second option receives the second color, and so on. Each cell thus is a duplication of the full multiverse (as seen previously in Figure 4.1), but the use of faceting, color, and the assignment of options to arbitrary level values allows this plot to support a task that is not possible with the original (described next).

The presenter states that the parameters do not have the same potential for reducing uncertainty at this stage, and thus it may be more productive to focus on the ones that have the greatest potential impact. Damage Outliers (upper-left cell of Figure 4.2), for example, is described as looking like a crayon box due to how the colors are all mixed; it is explained that, because no single color is clearly associated with any of the interesting features previously identified, it is not worth spending time discussing this parameter at this point because even if one option could be excluded it would not substantially change the overall range of outcomes in this multiverse. By contrast, the Death

Outliers parameter (upper-right of Figure 4.2) has one single option that is associated with the large upper-tail and upper portion of the discontinuity; this shows that it is possible that excluding even one option from this parameter could substantially change the multiverse as a whole, and thus this parameter is worth exploring more at this time. The Model parameter (lower-right of Figure 4.2) is similarly indicated as important, due to having a single option associated with both the upper-tail and small lower-tail of negative effect, which also happens to be colored blue in this case.[2] The remaining parameters are then briefly described for completeness: Femininity Calculation is not impactful; Damage Transform could be impactful; Main Interaction also could be impactful, but less so than the others as it seems only cleanly associated with the small tail of negative effect sizes and is otherwise mixed. The presenter then acknowledges that there are a few valid options for which parameter to pursue further, but that he thinks Death Outliers and Model are the two parameters that should be selected for further discussion in the meeting.

4.4.1.3 Scene 3: Consider interactions between two select parameters

The scene begins with the presenter selecting the two parameters discussed previously using the touchless interaction of a two-hand grab, as seen in in Figure 4.2. With a shake of the hands and pulling motion, the visualization transforms into the one shown in Figure 4.3. I call the plot shown in Figure 4.3 *two-dimensional option-faceted outcome curves*, which is the result of faceting rows according to the options of the Model parameter and faceting columns according to the options of the Death Outliers parameter. Figure 4.3 thus divides the multiverse into unique subsets, where each cell in the plot is an outcome curve of only the universes that have the relevant combination of options indicated by its row and column position.

The presenter points out that one cell looks much unlike the others (upper-left of Figure 4.3), which is the cell containing only universes that have these two facts in common: use of the No Death-Based Exclusions option for the Death Outliers parameter, which excludes no hurricanes excluded from the original dataset based on how many deaths they were associated with; and the Linear option for the Model parameter, where a gaussian-family generalized linear model is used. Nearly the entire range of outcomes in the full multiverse occur within this cell (3–409 here versus -17–409 overall), yet every other pair of combinations of these two parameters result in a far smaller range (the next largest range being -16–21 in the upper-center of Figure 4.3). This observation is described as showing that there is an interaction effect between these two selected parameters, which means that the sensitivity of one specific option or parameter depends upon what option is used for the other parameter. In particular, choosing No Exclusions for Death Criteria, and only

[2]Note that there is no shared meaning of colors between cells in Figure 4.2, due to the fact that what option in a parameter is assigned the first level of color is essentially meaningless here (e.g.: random, alphabetical, first occurrence in the data).

Figure 4.3: Screenshot from Scene 3, featuring an example of *option-faceted outcome curves*, where the multiverse is divided into a grid of cells in a way that allows for interaction effects between parameters to be considered. Rows and columns are each assigned a respective parameter, so that there will be one cell for each unique pair of options that occur for the two given parameters. For detailed description of this plot, see subsubsection 4.4.1.3.

when also combined with a choice to use the Linear model, is the only pair of these decisions that result in very large effect sizes and such a large range of outcomes.

There are a few other non-obvious implications supported by the design of Figure 4.3 worth noting. First, it shows that the Linear model option—but not the Poisson model option—is sensitive to a single specific datapoint, which happens to be hurricane Katrina as it was the most deadly hurricane in this dataset. Realization about this dataset-level fact based on a multiverse-level visualization is made possible because one of the parameters involves data exclusion criteria, and the names of the options for that parameter allow one to infer that it must be a single datapoint that is responsible. However, it is also important to observe that while the upper-left cell contains large effect sizes, it also contains small effect sizes, so it would not be true to conclude that hurricane Katrina alone causes large effect sizes; rather, it is the presence of hurricane Katrina in this dataset, together with the use of the Linear model, and even then only when using some other subset of analytical choices, that leads to some large effect size estimates.

4.4.1.4 Scene 4: Interrogating the residuals of twin universes

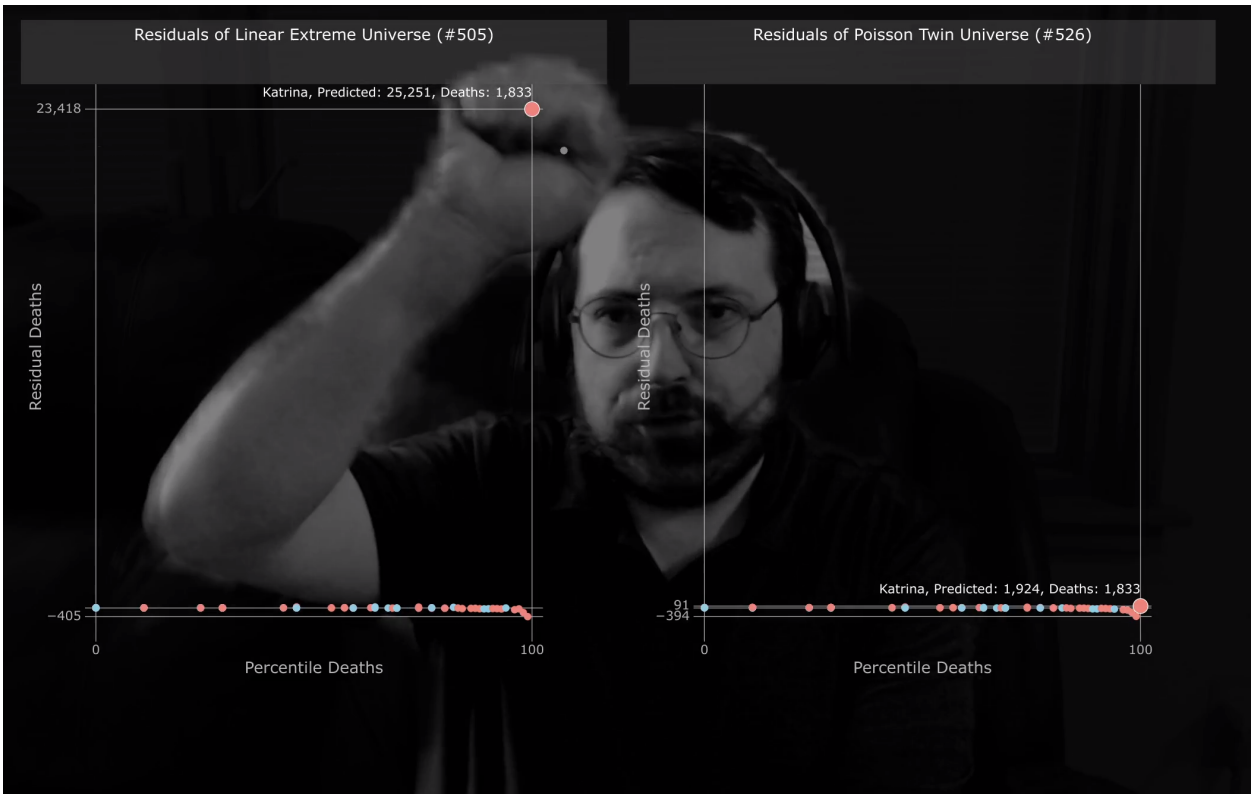


Figure 4.4: First screenshot from Scene 4, featuring an example of a *twin-faceted residual plot*. Each side of the plot depicts a separate universe, where the two cells here all for comparison between *twin universes*. For detailed description of this plot and explanation of the twin universe concept, see subsubsection 4.4.1.4.

The scene begins with a mention that during a break for discussion, an analyst asked a specific question the presenter wants to respond to (which the presenter says while pointing to the phone they are holding, indicating the source of the question), which is: "why are any of the universes so extreme?" The presenter highlights the largest effect size universe in the upper portion of the same visualization featured in the previous scene (Figure 4.3), and notes that its corresponding *twin universe* is also selected in the lower portion of the plot. It is explained that the initially highlighted universe's *twin universe* is a universe that has all the same option selections except for the Model parameter, which is the parameter that the rows are faceted by in this figure.[3]

Comparison of the twin universes highlighted in Figure 4.3 shows that the universe with the largest outcome would instead have an outcome near the center of the distribution if the only fact about that universe that was different was the use of a Poisson model instead of a Linear one; specifically the most extreme Linear model universe has an outcome of 409, but its twin that uses Poisson has an outcome of only 44. The presenter then says the posed question of "why so

[3]The *twin universe* concept is further discussed in subsection 4.4.2.

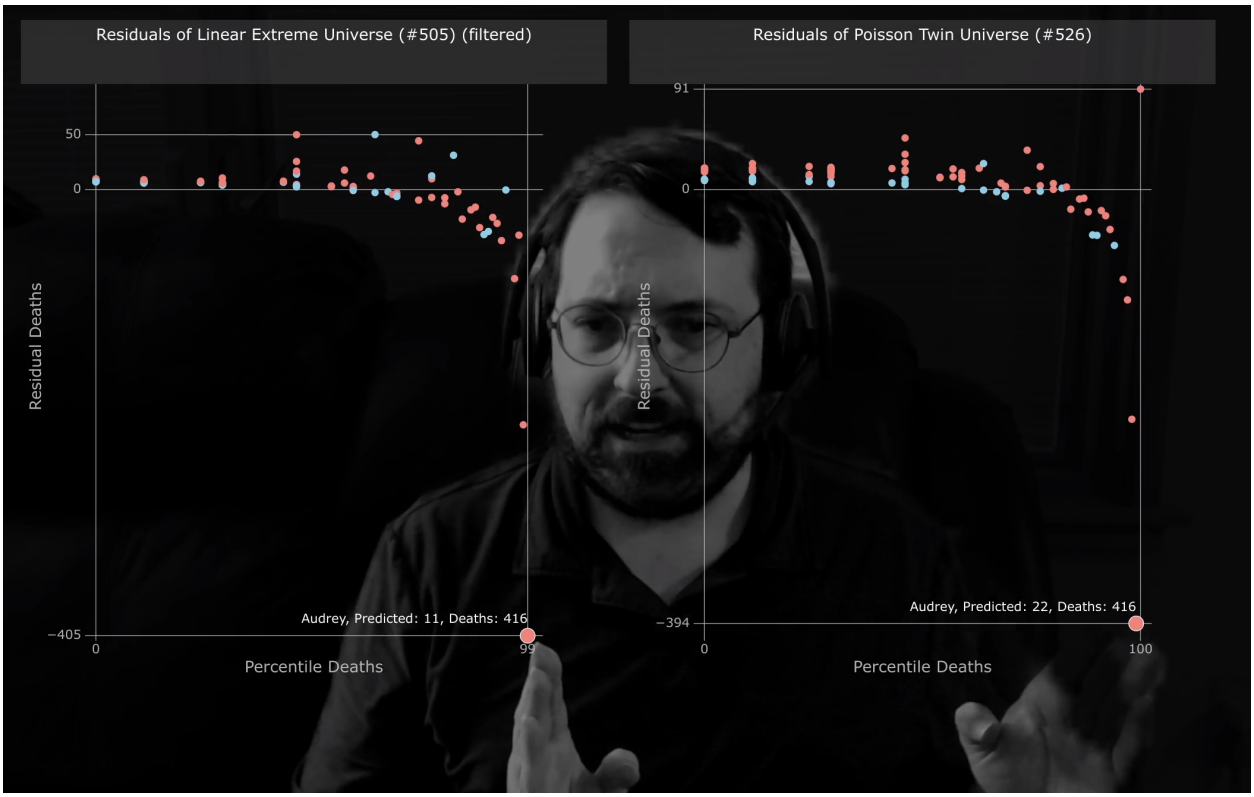


Figure 4.5: Second screenshot from Scene 4, featuring the same plot as shown in Figure 4.4, but with the single datapoint associated with hurricane Katrina hidden from the left side of the plot so the y-axis can be re-scaled to better view the rest of the data. For detailed description of this plot, see subsubsection 4.4.1.4.

extreme" will be explored using these twin universes, and visibly leans in and looks down (towards their keyboard) and says the visualization will be changed to something entirely different. The plot shown on screen then changes to the one shown in Figure 4.4.

Figure 4.4 is not using marks to indicate universes, and instead each mark represents a hurricane from the underlying dataset, and I call this visualization design a *twin-faceted residual plot*. The figure shows a plot that is duplicated side-by-side to form two cells, with each side of the visualization dedicated to one of the two twin universes identified previously. Each cell has the same axes, where the y-axis is residual deaths (difference in model predicted deaths and historically recorded death for that hurricane), and the x-axis is the percentile of deaths attributed to the historical hurricanes.

The presenter says that one extreme data point is throwing off the entire chart, which is hurricane Katrina, and in the linear universe the model over-predicted associated deaths by a factor of more than 10 (left-side of Figure 4.4). The presenter says this one data point is making the rest of the chart too hard to see, then performs a one-handed grab-and-scrub gesture over the datapoint to filter (hide) this one mark, which results in the plot transforming into the one shown in Figure 4.5.

Figure 4.5 is described as being the same as the previous plot, except the y-axis is re-scaled

because the mark for hurricane Katrina is hidden only on the left-side of the plot. It is noted that both models were over-predicting the amount of deaths for hurricane Katrina, though in the Poisson universe this was only a minor over-prediction. However, it is also noted that both models under-predicted the deadliness of many of the most deadly hurricanes (marks seen on the left two-thirds of each cell), while also under-predicting the deadliness of many the most deadly hurricanes, before then again under-predicting the deadliest hurricane (as previously noted).

4.4.1.5 Scene 5: Interrogating the percentile-residuals of twin universes

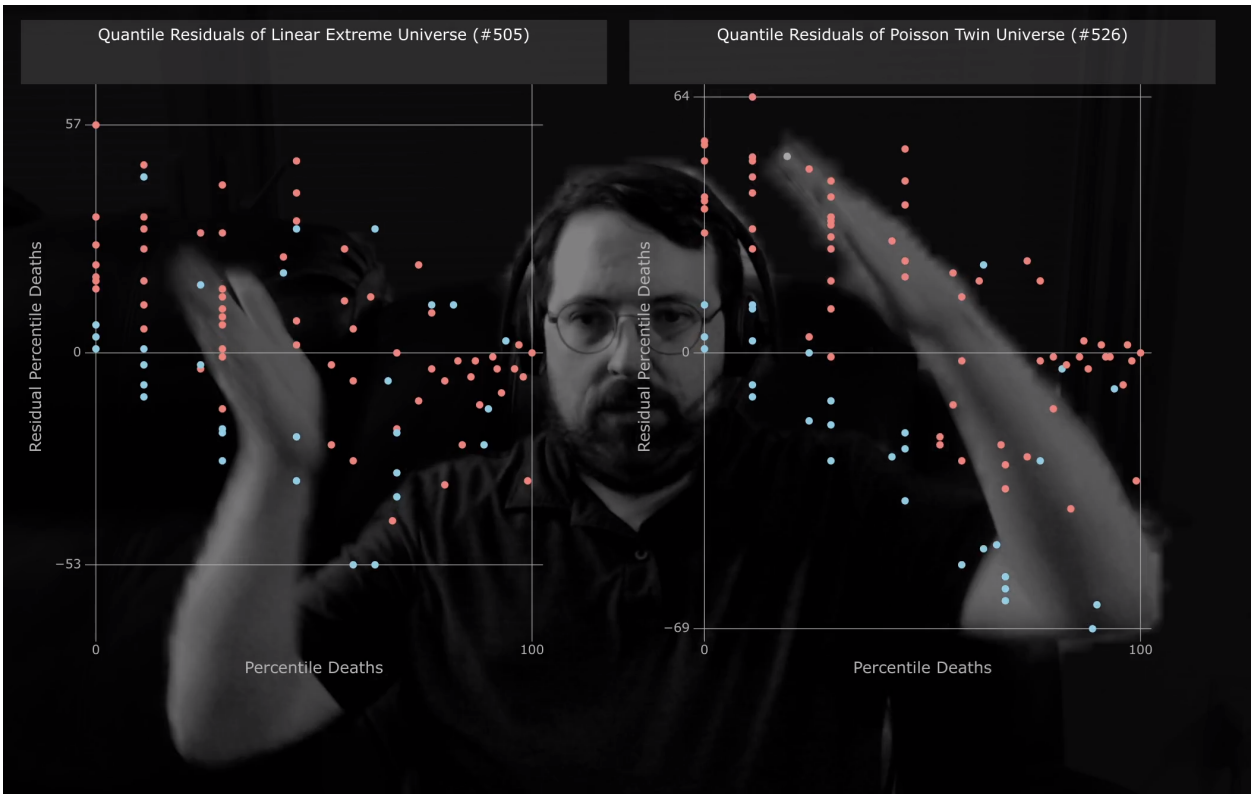


Figure 4.6: Screenshot from Scene 5, featuring another example of a *twin-faceted residual plot*, but this time the y-axis shows percentile-residuals instead of regular residuals. For detailed description of this plot, see subsubsection 4.4.1.5.

This scene continues the examination of residuals, starting with the presenter noting that another way of examining residuals is to consider how well models predict the relative ranking of the deadliness of hurricanes. To do this the previous plot (Figure 4.5) has the y-axis changed from the residual deaths to percentile-residuals, which is the difference in the predicted percentile deadliness of hurricanes and their historical percentile of deadliness, as shown in Figure 4.6.

The presenter mimes a linear regression line with their hands (as shown in Figure 4.6), while noting both universes show a downward slope, which agrees with the previous observation about over/under-prediction of hurricanes in both models. The presenter deselects all data points and

says to step back and look at the results overall (while then visibly leaning back and getting farther away from the screen), and points out that the dots are colored according to the hurricane name genders (male names in blue and female names in pink). It is pointed out that male hurricanes tend to be under-predicted, while female hurricanes were over-predicted, indicating a bias in the models that would lead to an exaggerated difference in estimated effect size beyond what the data itself indicates.

The presenter reasons that consideration of residuals suggests that perhaps both models are particularly poor, and indicates that perhaps neither universe should be taken seriously. If the goal of the project were to identify individual universes for potential exclusion, it is said that these universes would be good candidates.

4.4.1.6 Scene 6: Considering interaction effects of exclusion parameters

Figure 4.7: Screenshot from Scene 6, featuring another example of *option-faceted outcome curves* (also in shown in Figure 4.3), but this time faceted by the options for Death Exclusion and Damage Exclusion parameters. For detailed description of this plot, see subsubsection 4.4.1.6.

The presenter states that after more discussion (implied to have occurred between scenes or offline), it was suggested that the Model parameter will require more investigation later. However, the importance of data exclusion criteria led to a new observation: there are two parameters that deal with data exclusion criteria, so perhaps they should be considered at the same time. The presenter

concurs and shows a new visualization in response, which is a plot of *two-dimensional option-faceted outcome curves* just like the one described previously (Scene 3 in subsubsection 4.4.1.3 with Figure 4.3), but now the columns are faceted by options for the Damage Exclusion parameter and rows are faceted by options for the Death Exclusion parameter to form a total of 12 cells (as shown in Figure 4.7). The universe marks are colored according to their Model option, but this is not shown on screen or mentioned by the presenter.

The presenter notes that the two twin universes discussed previously both come from the same cell of this visualization (upper-right cell of Figure 4.7), as both of them had the option "no death-based exclusions" and "exclude 3 most-damaging hurricanes"; the source of this information is not depicted visually or mentioned earlier, and is merely offered as an observation by the presenter. In considering all of the cells of the figure, the presenter notes that every possible pair of options for the two exclusion criteria parameters would result in a smaller range of outcome values than the set of options associated with both of the twin universes. It is pointed out that this alone does not provide information about what choices are the right ones, but that it does show the importance of this set of decisions and thus would be a good use of the group's time.

4.4.1.7 Scene 7: Reflect on the impact of tentative decisions

The final scene of the demonstration begins with the presenter stating they will restate and summarize the conclusions that have been arrived at through previous discussion (implied to have occurred after the previous scene). The presenter says that it seems that the entire reasoning for excluding hurricane data seems to be because some data points caused some models to fit poorly. However, no one seems to claim that the hurricanes being excluded are inaccurate or that they are not valid examples of the subject being studied. It is reasoned that the decision to exclude any data is a case of changing the data to fit a model rather than changing the model to fit the data, and is thus inappropriate. The result of this is that one option for each of the two exclusion-based parameters are selected as explicitly appropriate and all other options are deemed inappropriate, thus effectively eliminating two parameters from further consideration in this multiverse.

The impact of these decisions is then plotted as in Figure 4.8. The left-side of the figure shows the original multiverse the group started with, and the right-side shows the multiverse that would result from implementing the discussed decisions. The presenter notes that these decisions are subject to any new objections, but notes tentative next steps in the case that all are in agreement.

The presenter prepares to end the meeting by restating the previous consideration of the Model parameter is still important, though probably will require further analysis and discussion outside of the meeting. They also point out the universe marks in Figure 4.8 are colored according to the Model option associated with them, which indicates the clear potential importance the Model parameter still has. One final point of note is that the size of the multiverse has been reduced from

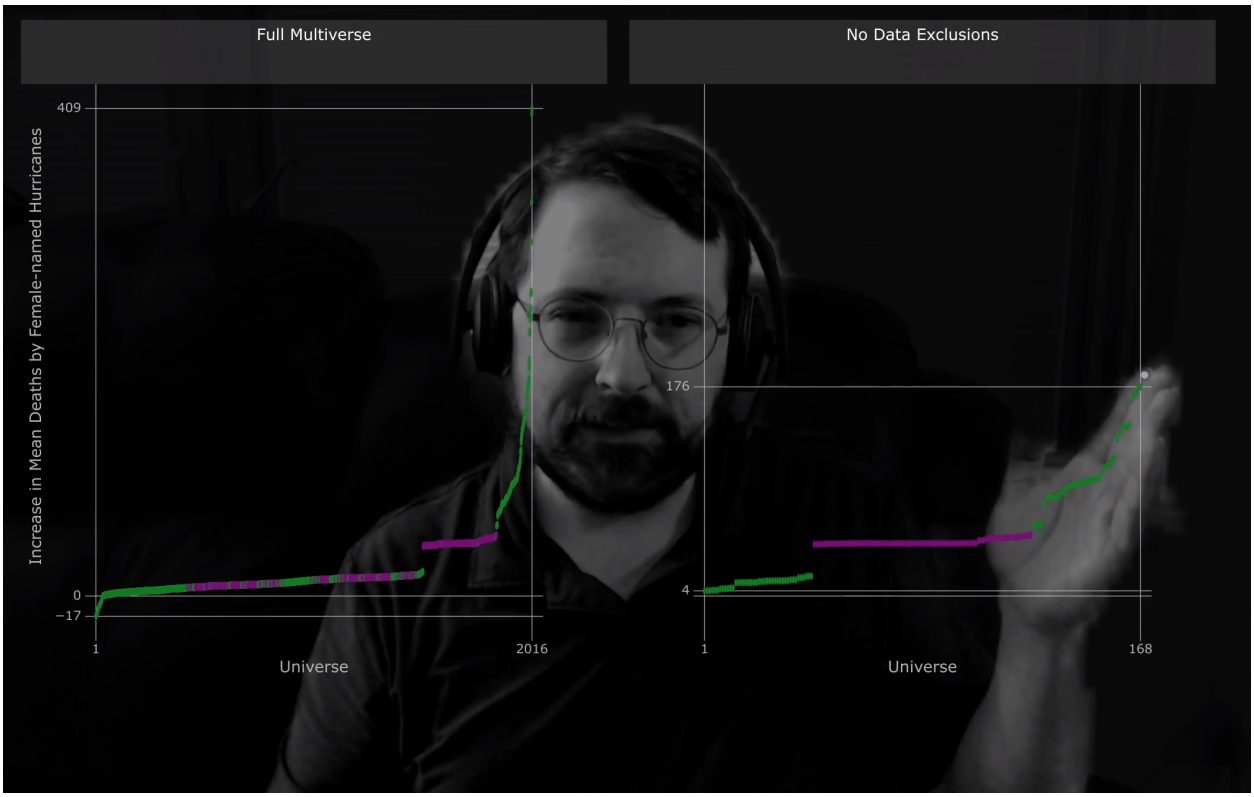


Figure 4.8: Screenshot from Scene 7, depicting the initial multiverse on the left and the after-decision multiverse that would be left if the decision not to exclude any data were adopted. For detailed description of this plot, see subsubsection 4.4.1.7.

2016 universes to only 168 universes appearing on the right-side of the figure and that the range of outcomes has been more than cut in half and eliminated negative effect sizes, which the presenter notes is still a larger range than desired but that this is still pretty good progress for one day.

4.4.2 Twin Universes for Contrast in Statistical Examination

The concept of *twin universes* is first described in demonstration *Scene 4: Interrogating the residuals of twin universes* (subsubsection 4.4.1.4), but the utility of this concept is substantial enough to warrant further description here.

Twin universes (or alternatively *sibling universes*) are universes that share all the same options other than the value of a specified parameter. In the example case shown in Figure 4.3, two twin universes are selected and highlighted, with the universe highlighted in the bottom-left cell being the only universe that shares every option with the universe highlighted in the top-left cell other than their option for the Model parameter. With respect to a selected parameter and a target universe to use as a reference point, the number of all sibling universes will be the same as the number of unique option values for the selected parameter.

The clearest value provided by the concept of twin universes is that it facilitates learning about

parameter-level considerations while simultaneously engaged in individual universe-level investigations of statistical validity. For example, the original inspiration for the topic of Scene 4 (subsubsection 4.4.1.4) and Scene 5 (subsubsection 4.4.1.4) was my own suspicion that the linear model option was a poor choice for this data, and an examination of the residuals for that universe agreed with that hypothesis. However, when one simultaneously considers the twin universe where the only option that differs is the Model option, the idea that the poor model fit previously observed was solely a feature of the linear option was quickly dispelled.

It appears that the use of twin-faceted residual plots as illustrated in Figure 4.4 and Figure 4.5 can also encourage divergent thinking, as discovered during the evaluation study and further described in section 4.6.

4.4.3 Augmented Presentation for the Multiverse

The primary role of augmented presentation in AugMeet is to ease the cognitive burden required to complete tasks that are prerequisites to being able to complete validation and interpretation tasks. Put simply, if it is easier and takes less time to get everyone on the same page, then more time and energy can be spent discussing what matters most. AugMeet thus benefits from all the core features of AugChiro [44], but some modifications and additional considerations were necessary to adapt augmented presentation techniques to the multiverse.

While augmenting deictic gestures (e.g.: pointing to a specific data point, touching the margins to show an axis value reference line) was still helpful in simpler plots like the first large outcome curve in Figure 4.1, they are harder to perform in a useful way in faceted plots like the one in Figure 4.3. To help with this a few new features were introduced, like the ability to toggle margin and datapoint selection features individually using a keyboard command, and most importantly keyboard commands to freeze, hide, and re-select data points. While these types of modal features were intentionally avoided in the AugChiro [44] system due to the cognitive overhead it places on the performer/presenter, I personally found the added difficulty I experienced while demonstrating the system to be a small price to pay for the ability to switch between augmented and non-augmented use of my hands at any time.

Two new noteworthy touchless interactions were also depicted during the demonstration: two-hand grab selection to trigger a transformation to a new plot based on those selections (shown in Figure 4.2), and a one-handed grab-and-scrub gesture to filter or hide a specific data point (shown in Figure 4.4). The two-hand gesture to trigger a transformation is intended as a way to help an audience follow a transition between two quite different visualizations, but where some aspect of the first chart is being retained—even if only as a logical connection—and used in the second visualization. In the demonstration this was the selection of parameters for further discussion, which

were then split apart into their constituent options, yet the logical connection between these sets of facts is not immediately obvious from reviewing the before (Figure 4.2) and after (Figure 4.3) visualizations and this can be distracting until the connection is understood by the audience. Transitions like this can also be supported through visual means like animation, but the gesture seemed to provide additional support regardless. The grab-and-scrub to filter gesture had essentially the same role, but with the added detail that the transition was not between different plots but between states of the same plot.

4.4.4 Multiverse Visualization Designs: Multi-Faceted Faceting and Archetype Variations

Adapting augmented presentation to a multiverse context required the implementation of additional types of interactive visualizations, ultimately resulting in multiple distinct types of visualizations that are all distinctly different from the ones used in the AugChiro [44] system. Uses of faceting and archetype variation are only briefly described below, as detailed descriptions and example tasks are provided for every visualization used in this system in their respective portion of the Demonstration section (subsection 4.4.1).

4.4.4.1 Multiple uses of faceting to support different tasks

Faceting is a technique used in multiple ways in AugMeet, and each way of applying faceting supports different types of tasks. In order of their use in the demonstration, the first use of faceting is *parameter-faceted outcome curves* (Figure 4.2), which uses faceting to create one grid cell for each parameter in the multiverse, and where the additional use of color allows the potential impactfulness of each parameter to be compared.

The second use of faceting creates the *two-dimensional option-faceted outcome curves* shown in Figure 4.3, where faceting is used to assign the option value of each of two select parameters to individual rows and columns. The cells here are also not reproductions of the entire multiverse of outcomes, but rather each cell depicts a subset of universes associated with that cell's respective pair of option values. This application of faceting allows for any potential interaction effect between the options of two parameters to examined directly, which is also an example of a Connection Combination task (subsection 3.6.4). This type of plot can enable many non-obvious insights, examples of which are described in subsubsection 4.4.1.3.

The third and final noteworthy use of faceting results in a *twin-faceted residual plot*, as shown in Figure 4.6 (also in Figure 4.4 and Figure 4.5). By faceting across twin universes, there will be one cell per universe, and one universe for every option of the reference parameter. This use of faceting integrates universe-level analysis into a way of performing a multiverse analysis, which is

a feature of only two other previously reviewed systems ([30, 58]); however, the simultaneous use of the twin universe concept with universe-level analysis allows for parameter-level insights to be obtained from universe-level examination, which is a novel feature unique to AugMeet.

4.4.4.2 Variations of the outcome curve archetype

AugMeet focuses on variations of the outcome curve archetype for all of its multiverse-level visualizations. The reasons for this decision are entirely pragmatic. First, the outcome curve is one of the most commonly used types of multiverse visualization, so audiences are more likely to be familiar with it already. Second, it is one of the most visually simple archetypes that also supports a large variety of multiverse tasks even in its basic form. And finally, it turned out to be so flexible that varied uses of faceting and mark color mapping was sufficient to provide good support for every multiverse-level task required for the full demonstration of this system. Put simply, variations of a single archetype performed well enough in all included tasks that no other archetypes were needed at this time.

4.5 Evaluation Method

To evaluate this system, this author conducted an interview study with human participants, conducted through remote video-meeting software. Interviews were each approximately one hour and fifteen minutes in length. The protocol for this study was reviewed by the Institutional Review Board (IRB) of the University of Michigan and approved as exempt from ongoing review.

As a brief summary of the process, participants were asked to view an online form and screen-share their view of the form with the interviewer. After completing a consent form and responding to a few self-rated experience questions, participants watched video clips (seven in all) showing a demonstration of the AugMeet system depicted as a live remote meeting of a research group. After each clip the interviewer solicited participants for their thoughts, critique, and assessment of the content of that video clip. Once all video clips were reviewed in this way, participants completed a final multi-choice rating form and the interviewer solicited their critical reflection on the entirety of what they viewed.

The following subsections provide details on each aspect of the study, and the results of the interviews are provided in section 4.6.

4.5.1 Participant Recruitment

Participants were recruited from a list of people selected for their prior knowledge and experience with multiverse analysis specifically, or data analysis communication and statistics knowledge in general. The list itself was seeded with people known to the author as having previously worked with multiverse analyses or closely related concepts (such as specification curves), then extended by recommendations from members of that list (snowball sampling). The purpose of this recruitment process was to focus on participants who possess relevant experience sufficient to allow them to make substantial critiques of all aspects of the system, as well as serving as representative users. Participants were asked to self-rate their experience and prior familiarity with the subject, the results of which are reported in section 4.6.

4.5.2 Interview Form Prompt

The focal prompt for the interviews was an online form (Qualtrics), which consisted of the five sections briefly described below.

1. Consent form: used to obtain informed consent and authorization to participate (required to be a legal adult and not currently physically located within the EU, and thus not subject to the GDPR), as well as additional authorization to record the interview (optional).

2. Overview of interview process and topic: a brief description of the interview content and process, along with a two paragraph description of the mock meeting scenario context for the videos that will be reviewed (described in subsection 4.5.3).

3. Familiarity and experience self-rating form: a three-question multi-choice rating form for the participant to self-rate their familiarity with general method of multiverse analysis, the subject of gendered-naming of hurricanes, and their level of experience with interpreting the results of statistical models commonly used in the social sciences.

4. Video vignette review: repeated for each of the seven vignettes, each page consisted of an embedded video vignette to be watched first, followed by a request to say what were the key points being communicated in the video, a multi-choice question to rate the mental effort to understand these points, and then a semi-structured oral discussion with the interviewer about the vignette.

5. Final evaluation form and concluding feedback: form with six multi-choice questions which asked the participant to rate different aspects of the system as presented in the previous

vignettes, followed by a final oral discussion with the interviewer guided by a set of semi-structured oral interview questions (not listed on the form). Ratings were requested for the reasonableness and justification for arguments made in the presentation, helpfulness of see-through aspect of the system, helpfulness of the interactive gesturing with the data visualizations, potential usefulness in other multiverse analyses, and potential usefulness in other types of projects.

4.5.3 Video Vignette Structure and Purpose

The primary focus of the interviews were video vignettes, which served as an applied demonstration of the AugMeet system, the content of which is described in subsection 4.4.1. The participants were told that the videos depict a remote meeting of a team of professional researchers collaborating on an ongoing research project, and were given the context of the meeting as previously detailed in section 4.3. Videos were edited for length by trimming the beginning and end points, but all audio and video content within the videos were as originally recorded without alteration. Each scene was 2-4 minutes in length, with a total run-time of 23 minutes and 50 seconds.

Participants were asked to share their critiques and opinions on any aspect of what they were seeing, especially the subject matter content and arguments depicted, how data was visualized, and the overall workflow for conducting a multiverse analysis in a group setting.

4.5.4 Qualitative Analysis Process

Interviews were recorded for review and transcription for participants who consented to be recorded. Any participants who did not consent to digital recording were interviewed in an alternate form, wherein the interviewer shared the videos one at a time and took additional notes during and after the session to use in lieu of a transcript for that participant. Manual notes were taken by the interviewer during all interview sessions, and memos were written immediately after each session to collect observations and reflections for later review. Transcripts, notes, and memos were reviewed by the author during an iterative open-coding process to flag all notable participant feedback and identify overall themes to organize and summarize the content of the interviews. All participants who agreed to be recorded submit their responses to multi-choice questions through the online form, in addition to discussing with the interviewer the response to each question as it was made, so the responses to these questions will be analyzed and reported holistically rather than only quantitatively summarized.

4.6 Evaluation Results

Direct quotations from participants are drawn from transcriptions and hand-written notes of oral conversations with participants, and have only been lightly edited to adapt them to the written form, such as by the removal of conversational artifacts like repeated clauses and the use of filler words.

4.6.1 Description of Participants

A total of seven people were interviewed for this study, six of which agreed to be recorded and with the other consenting to a non-recorded alternate form for the interview. All participants were highly educated, having attained a PhD or being in the latter stage of completing a PhD program, including: three professors, two post-doctoral researchers, one PhD student, and one PhD graduate employed by a technology company. All participants had at least five years of research experience in fields of technology or social science, with many having a decade or more. All participants were authors of multiple research papers, at least three had co-authored published research papers explicitly about the subject of multiverse analysis (two of which included a multiverse analysis of the same subject as the use case of this chapter), and all had co-authored papers involving the use and interpretation of statistical models.

In response to the multi-choice item asking each participant to rate their experience in interpreting the results of statistical models commonly used in the social sciences, all participants described themselves as being either moderately experienced (4 participants) or extremely experienced (3 participants). Participants that rated themselves as only moderately experienced gave reasons for their rating such as stating they were not a statistician, that some people have greater expertise than they do, or that they did not consider it their primary area of expertise.

With regards to familiarity with the general method of multiverse analysis, all participants rated themselves as at least somewhat familiar (3 participants) or very familiar (4 participants). The participants also also rated their familiarity with the subject of gendered-naming of hurricanes as at least somewhat familiar (5 participants) or very familiar (2 participants). The participants who rated themselves as somewhat familiar with the hurricane example generally described hearing about it prior to the interview but that they had not studied the topic closely. Both participants who rated themselves as very familiar with the hurricane example also rated themselves as very familiar with multiverse analysis, and specifically they both had been separately involved with at least one multiverse analysis project that used it as an example for analysis.

Overall, all participants are considered by the author as representative of the intended audience for the AugMeet system. Their self-ratings were also consistent with the content of their interviews

and the interviewer's knowledge of the background of each participant.

4.6.2 Augmented presentation makes it easier for people to understand and pay attention

The most common theme across all interviews was how easy it was to pay attention, follow along with explanations, and understand the key points being communicated in each scene. These descriptions also often ran together with mentions of feelings of engagement, interest, and human connection. The participants most commonly described the augmented presentation aspects of the system as the cause of these judgements and feelings.

For example, P1 contrasted how poor the experience is of trying to talk through a visualization using a mouse cursor with the use of touchless interactions: "I especially like that, specifically pointing as you talked through things like that. Now I think this is quite helpful, because you have a good sense of where and what you're talking about. Often, you know, hovering [over a data visualization with] the cursor is not great." P3 compared the experience with what is typical and expected in most meetings: "The whole thing, using your hands with the data and all of that, its cool and fun and interesting ... it is just different from everything else you see. ... makes things very easy to pay attention to, and very clear." P3 also thought the fun-factor would impact their desire to participate in such a meeting: 'It makes me want to ask a question. ... You know they might do something interesting to answer it." P5 pointed out how even non-augmented gestures were helpful, referring to a moment in scene 5 where the presenter uses their arms to point out a relationship in the data (as shown in Figure 4.6): "I really like the moment when your hands were trying to draw the regression line visually. I think that's very helpful." They further explained how the use of touchless interactions helped them specifically: "When your hand was tracking along with where you pointed, it helped me to follow what you were saying. When the hands were not there it was harder for me to pay attention and follow along."

As to exactly how and why the augmented presentation elements of the system were beneficial, P6 explained it this way: "I appreciate I'm not getting bored. I think the component of seeing you on top of seeing the visualizations, I don't have divided attention. I look at it and I have everything there, I look at the person and they keep me engaged. I'm not doing that kind of thing where you are listening and you try to look at the person, and then I have to look at the visualization and try to recombine whatever they were saying with the visualization. So I find, it works. It works well." P7 described how he had trouble figuring out why the experience felt so different, then suggested: "It's almost like you're using a tool versus having a visual effects person come through ... it's just like it's me, and [you], and this [system] is the tool. It's not me, [you], and someone else who came in to polish." P6 also reflected on how the feeling of human connection with the presenter can help

improve receptiveness to the information: “This is great. I love this tutorial thing. I praise the way, you know, it’s hand-holding—but not in a manner that makes me feel like I’m some dumb person you are explaining the concept to.”

Overall, the augmented presentation aspects of AugMeet seemed to serve primarily as a way to make it easier for the audience to listen to and understand the information being conveyed to them about and through the visualizations themselves. At the same time, augmented presentation elements also provide additional visual stimulation, novelty, and a sense of human connection which promotes attention, engagement, and interest in the material—and all of these would naturally also contribute to reduce the effort required for an audience to listen and understand.

4.6.3 Comparing residuals of twin universes elicits divergent thinking

The comparison of residuals between twin universes that occurred in scenes 4 and 5 prompted multiple participants to engage in divergent thinking, specifically suggesting additional options or concerns that were not directly mentioned or implied in the videos themselves. For example, after viewing these scenes P1 said: “Let’s compare what’s going on with the universes, and I think seeing the individual diagnostic plots made me realize it’s like, okay, maybe both of them [linear and poisson model options] are not great choices. Maybe we need some different model, which accounts for the variance as well as the impact of other predictors in the model … things I hadn’t really considered [in their previous thinking about this multiverse example]”. P1 suggested it was possible all the options for the Model parameter were inappropriate and should be excluded, which is an example of a validation task with a possibly dramatic impact on the multiverse. If this suggestion was adopted, all universes considered so far would be eliminated and it would be required to find at least one new model option to be able to generate an entirely new multiverse. Regardless, this is an exemplar of the type of discussion AugMeet aims to elicit.

Another example of divergent thinking is provided by P4, after viewing scene 4: “Do we even have the right features [in the data] to answer this research question about female versus male names being more deadly? Because we’re just measuring deaths, not evacuation counts or anything like that… I started thinking, are there other factors that make it hard to potentially get at the research question?” Omitted variable bias and the potential impact of confounding variables is a major concern in analyzing any data, yet it has not received substantial attention in most treatments of multiverse analysis. Even though it was not the subject of the presentation, the examination of twin universes and their respective model fits elicited deeper consideration about the nature and completeness of the data, and this participant wondered if the models could not predict hurricanes accurately because it might not be possible to do so when the most important information is not available to them.

P2 also noted how this universe-level focus and examination of what is appropriate is a point of departure from how multiverse analysis has generally been described: "especially this part [Scene 4], I think, is where it really starts to transition in a way, and this is not something you would have automatically seen from knowing about a multiverse. This is a different kind of approach." P1 concurred with this, and even though they had previously published at least one paper with a multiverse analysis of the same subject as the one in this presentation, they noted particularly after scene 4: "Now this is information that I hadn't seen before. … These are things I have never considered before."

If multiverse analysis led to less attention on the most important of questions in exchange for more attention being paid to salient trivialities, it would not be a worthwhile endeavour. While I initially conceived of the use of twin universes as a way to avoid spurious conclusions about the superiority of one option over another, the fact that multiple participants offered divergent insights in response to them suggests this technique can support validation tasks in more ways than one. It also illustrates how focusing on the low-level details examined in twin universe residual plots need not distract from the most important high-level questions, and can even support high-level validation tasks by eliciting divergent insights about what options are appropriate.

4.6.4 Parameter-faceted outcome curves give perspective and focus

The parameter-faceted outcome curves shown in scene 2 (Figure 4.2) were described as giving perspective by first providing an overview of all the things that are being considered, and then providing focus on what is most important by providing the details of comparative importance of parameters; essentially, this visualization complies with "The Mantra" of "overview first … details-on-demand" [22, 15]. P4 described the effect the scene had on them this way: "It's that these [parameters] seem to be the 2 biggest factors we need to make a decision on this, and we need to spend time on that, say over something else. … It's so much nicer that this stuff is getting explored, and then this is also effectively—very effectively—saying, 'Hey, let's spend time in a meeting talking about our reasoning for excluding [data] … as opposed to arguing about how you measure femininity of a name.' You know, that makes me happy. Knowing there's things that I don't need to be anxious about, or at least far less anxious about. … that sort of narrowing and process has been convincing and is very valuable."

At the same time, even what was described as narrowing and focusing still elicited divergent thinking from P4, as shortly after the previous comment they suggested another parameter or option should be considered if it had not been already: "I am starting to anticipate [what the next point in the presentation will be] … [for example,] did the deaths get log transformed? Is it data that should be log transformed? And I would, I would guess, given most things that never go

negative need to be log transformed. But obviously, that's probably a point that you're leading to." Though P4 considered this detail obvious enough that it surely would be the next point in the presentation, indeed it was not. While the log transformation of the damage variable is a feature of this multiverse, the log transformation of deaths is not and for a relatively complex reason well beyond what was discussed in the video. The linear model option treats hurricane deaths as following a log-normal distribution by having a generalized linear model predict the log of deaths plus one and assume a gaussian distribution (post-transformation); however, in the poisson model option a log link function is used for the calculations, but the predictor is not itself explicitly log transformed. I would suggest the addition of a separate log transformation parameter, which might only be valid for the linear model and not the poisson, but in any case this is another good example of a validation and interpretation task being accomplished based on a divergent insight.

On a related note, after viewing this same scene P2 asked about the underlying hurricane data, specifically how many people hurricanes usually kill, and how the effect size estimates compare to a naive analysis of historical averages. While the presentation makes some reference to the general historical performance of hurricanes, and this author had considered the addition of minimal references to historical averages in a few of the visualizations, I had decided to leave out these details. On further consideration of P2's questions, I realized that basic facts about the underlying data are important for building intuition about what model outputs seem reasonable and which are more indicative of a procedural error than an estimate that should be given serious consideration. This oversight is not inconsequential, and more or better ways to include dataset-level consideration into multiverse analysis reports and workflows would be a worthy subject for future research.

P7 also noted that the step-by-step progression of information caused them to engage strongly with the content of the presentation, saying: "I was actually thinking really hard about it, you know. Now, there's like 6 things to consider ... I guess it's more related to the the data problem and the complexity of what we're talking about itself. Less so than with the graphs. I think you know these are still quite understandable, that interface. ... It's like, you know, there's a puzzle in front of us. And we're thinking about it."

Parameter-faceted outcome curves provide an overview of all the choices that have been identified as worth considering, and then in the context of AugMeet their visual features are used to identify what is important (potentially impactful) and what is not. Narrowing the focus to what is most important can help people to not feel overwhelmed by the multiplicity of options that can be considered, and thus there can be more energy available to apply to a smaller 'surface'. At the same time, it may be that the act of trying set other considerations aside contributes to the elicitation of some divergent insights, as some people may be compelled to make sure everything has been considered before they are able to 'relax' and focus. Divergent insights can essentially complete some validation tasks, while focusing effort and mental energy can make it easier to perform

many others, so this technique supports target tasks in more ways than one.

4.6.5 AugMeet supports the iterative, progressive group effort necessary to complete difficult validation and interpretation tasks

Participants also called attention to an aspect of the demonstration that has not been described as an explicit feature of the AugMeet system, yet which multiple people considered interesting and valuable. This aspect is the overall workflow depicted in the demonstration, which is the general process and way of organizing the meeting, as well as the idea of an iterative and progressive approach to multiverse analysis in a group setting.

For example, P2 stated she was currently working on a new multiverse analysis with a colleague, and was encountering problems she realized cold be addressed by applying some of the techniques shown in the demonstration: "We are putting together preliminary results, and we're trying to interpret things. And there are all these universes that don't make sense... And applying that iterative process to that project, I can see how it would work. I almost wanna do that now."

P2 also described how the demonstration changed her mind about the purpose of a multiverse analysis and what it can be used to achieve: "Depending on what parameter options you choose for an analysis, you could end up with wildly different results, and I think that that in itself is interesting, and so maybe hogs the spotlight. But actually, the thing that you want is to have some kind of substantive understanding about the research question at hand. How much uncertainty do we have? Can we say, we know with some certainty what we think the relationship between these specific variables are? And probably that is the more useful endeavor—and, I think, what this kind of workflow allows for." In other words, she wants to go beyond a description of sensitivity to give a more focused, accurate, and useful description of what findings are uncertain and how much uncertainty there really is. At the same time, are there conclusions that have substantially less uncertainty than others? For example, if it is highly uncertain what the health impact is of a small exposure to a toxic substance, but there is very little uncertainty on the health impact of a large dose, then it might be far more valuable to be able to communicate that clearly even while additional efforts can be used to try to unravel why the impact of a small dose is so much more uncertain.

P4 specifically noted the appropriateness of the iterative aspect of the workflow, saying: "Iterative is something I just think is fitting for a multiverse analysis." P7 also considered how the system might be applicable to projects beyond multiverse analysis, and said: "Try to learn something from it, iterate based on what we have, and then wrap up the meeting. That kind of tactic, as in do that again and again and again until we have something. It's not just that it's useful, I feel like it's necessary."

Overall, participants thought the workflow demonstrated in AugMeet is well suited to multiverse analysis, and they emphasized the importance of progressive iteration and group effort. If people with different areas of expertise are able to work together in the way demonstrated by AugMeet, doing so in a way they find manageable and allows them make progress a step at a time, then validation and interpretation tasks of even the most challenging variety could be made practically achievable.

4.7 Limitations

The intentional selection of participants who are representative users of a system like AugMeet brings with it the limitation that no presumptions should be made about generalizability to other populations; all participants were highly-educated, and more statistically-literate than even the norm among highly-educated people overall. Interviews were also all conducted one-on-one, so this evaluation was not able to examine how larger group dynamics might develop or influence participant behavior and perception. Future work might consider studying how a system like AugMeet could support groups with a more substantial diversity in experience levels and areas of expertise, and with groups of varying size.

While the augmented presentation aspects of this system have been studied with a larger variety of participants during the previous study of AugChiro [44], and all findings in this study are consistent with what was found previously, there is a common factor to both studies: the presenter performing the presentations for both evaluations was I, your humble author. Given this, it cannot be determined what the impacts of augmented presentation would be when performed by other people. From my own personal experience I can say with certainty that augmented presentation involves multiple skills that can improve with practice, and people will certainly vary in their interest and ability to give augmented presentations.

AugMeet was also demonstrated using a single multiverse, in a single presentation that represented only one iterative cycle of analysis. The presentation and data analysis were also created by me, and other people might want to tell a different type of story or explain it in a different way, which could lead to different system requirements or call for features I have not considered. While participants thought AugMeet would apply well to any multiverse they had been involved with, only future work can determine what changes or new features might be inspired by applying these techniques to an entirely different multiverse, or repeatedly through many iterations on the same multiverse.

Finally, many aspects of the demonstration were a novel experience for participants. Some participants had seen some version of augmented presentation before, but it was still essentially unfamiliar to all participants. Perceptions of systems like AugMeet are also likely to be subject

to acculturation, and experience with similar systems could eventually impact how people receive and react to systems that appear similar—for good or ill.

4.8 Chapter 4 Conclusion

To return to the research question of this chapter (section 4.2), the AugMeet system provides a collection of ways to support the completion of validation and interpretation tasks. First, it uses augmented presentation as a technique to make it easier for the audience to follow along and understand information provided about and through data visualizations, while simultaneously making it easier for people to pay attention and encouraging them to be interested and engaged. Second, it uses a series of visualizations to build up a shared understanding of the facts, provide an overview of what is being considered, focus effort on what is most potentially impactful, and support low-level interrogation of twin universes to provide a new perspective and source of critical information about the multiverse.

The visualization series demonstrated by AugMeet integrates a variety of novel techniques and concepts to support validation and interpretation tasks, including: the concept of *twin universes* (subsection 4.4.2), *parameter-faceted outcome curves* (Figure 4.2), *two-dimensional option-faceted outcome curves* (Figure 4.3), and *twin-faceted residual plots* (Figure 4.6). AugMeet combines all of these individual aspects into a single system, and the demonstration of the system also shows a workflow that participants identified as being of further value due to its approach of progressive, iterative interrogation of a multiverse to reduce uncertainty.

Overall, AugMeet takes a multi-pronged approach to support the tasks in a multiverse that are not only the hardest to perform, but also are the most essential to making a multiverse analysis useful and of more practical value when shared with others. Perhaps a more eloquent summation of this chapter is this assessment of AugMeet offered by P6 during their interview: “This is, ‘here’s a way to do it’, which I think is meaningful. Until there is another proposal of how to do it better, this is kind of saying that, you know, this works, and some areas don’t work as well. But you know, good luck solving these areas, and that is for future people to pick up.”

CHAPTER 5

Conclusions and Future Work

5.1 Reflecting on Thesis Statement and Claims

I will now return to the thesis statement and claims made way back in section 1.5, to reflect on what I have achieved, how it was achieved, and what remains. I ask the reader's indulgence as I engage in a bit of story telling in this penultimate section of the final chapter of this work, which has been so many years in the making.

Thesis statement: Visualizations and interactive systems can support the use of multiverse analysis as a way to assess analytical uncertainty, direct efforts towards ways to practically reduce it, and improve the accuracy and meaningfulness of the consequent assessment of whatever analytical uncertainty remains.

Thesis claims:

- Claim 1 (chapter 2, chapter 3): Assessing analytical uncertainty and reducing it requires the completion of distinct analytical tasks—which are supported to varying extents by identified multiverse visualization archetypes—but existing systems provide only very limited support for a category of tasks that are particularly necessary for reducing analytical uncertainty, namely validation and interpretation tasks.

- Claim 2 (chapter 4): A *series of visualizations* can be used to narrow and direct focus while still eliciting divergent insights, collectively supporting the completion of validation tasks to reduce uncertainty and interpretation tasks to derive useful meaning from the multiverse.

- Claim 3 (chapter 4): *Augmented presentation* of data visualizations is a way to facilitate and support nuanced discussions around a multiverse analysis, particularly by improving audience attention and engagement, aiding audiences in obtaining a clear understanding of complex information with less expenditure of mental effort.

When I first identified multiverse analysis as a potential way of dealing with analytical uncertainty, it seemed to me to be a method with great potential even if it were not yet practical for most

work. At the time there were essentially no tools to help actually perform a multiverse analysis, and even if there were, I was plagued by one particular line of questions in my mind: if an outcome is sensitive to analytical choices that seemed reasonable to begin with, then what? In the end, do we just so often know nothing at all?

The survey of multiverse visualizations (chapter 3) was essentially a way to try to answer these questions for myself under the implicit theory that perhaps deriving meaning from a multiverse is just very hard and we lack the tools to make it easier. For that matter, it was not even clear what kind of tasks one can or might want to do, so I set out on a quest to closely read and deeply consider as many interesting-looking multiverse analyses as my collaborators and I could manage to find. Given how complicated and hard to understand so many multiverse visualizations seemed at first, perhaps if I just stared at them long enough I could find enlightenment. While not quite enlightenment, I did at least learn a lot. The findings from that work also provide the support for the first thesis claim and a piece of my thesis statement: visualizations can certainly support multiverse tasks to varying degrees, making tasks easier or essentially impossible depending on the information they encode and details of their design.

However, after the multiverse survey was complete the questions I had from before bothered me even more. The question I most wanted an answer to—then what?—remained unanswered. In the multiverse survey this problem manifested itself as the prospective category of tasks I called "interpretation", which was the tasks required to make meaning from the multiverse, and for which I could find so few clear and useful examples. If you cannot say what the meaning of an analysis is, why bother to do it? Is sensitivity all there is?

As no direct answer to the question that seemed most important to me was forthcoming, I advanced in what I initially thought was an entirely different direction. My experience in doing data analyses had repeatedly given me the feeling that I learned a lot from performing a data analysis, but then when I wanted to discuss what I had learned with others I had a new problem. Specifically, it was often so difficult and time consuming to build up the set of facts and network of related observations necessary to have a discussion that there ended up being too little time or energy left to have a useful discussion. This was especially true in online meetings, where my usual strategy of pointing and gesturing at visualizations proved to be of far less use than they were in person. This line of motivation eventually led to AugChiro [44], which is how I found that I could work with visualizations in a way that people described as easy to understand and follow, and even resulted in an experience people thought was fun and interesting.

Afterwards I returned to again thinking more about the multiverse, and still felt I had not made progress on the questions that bothered me most. I ended up deciding that it would just have to be enough if I could only find a way to make it easier to talk about a multiverse analysis, to share and explain the sensitivity of results to others. Given my extensive study of multiverse

visualizations, and with augmented presentation as a technique that made comparatively more simple discussions of data easier, perhaps I could combine the two to find ways to make it easier to assess and communicate analytical uncertainty to others than ever before? Perhaps, if it were easier to understand and share, it would leave more time and energy for discussion, which could at least be pleasant. This resulted in the last project of this dissertation, AugMeet (chapter 4), which brought augmented presentation into the multiverse context. Augmented presentation of a multiverse required some tweaks and adjustments to support some of the more complex and information-dense multiverse visualizations, but it worked. This project provides support for my claim 3: augmented presentation can perhaps be even more effective at making it easier for people to understand and engage in critical discussion in the context of a multiverse than it is in simpler contexts, if for no other reason then there is more difficulty to be reduced.

What surprised me was that, while conducting the interviews to evaluate AugMeet, it was the reactions and feedback of the people I interviewed that made me realize I had succeeded in doing a bit more than making it easier to communicate analytical uncertainty and discuss it with others. This ultimately resulted in support for Claim 2: a series of data visualizations—including some new designs I created for this project—can be used to manage focus, elicit divergent insights, and work together in a way that supports not only assessing analytical uncertainty, but supports a way to interrogate the multiverse through the completion of validation tasks to reduce the uncertainty and interpret whatever remains. What shocks me still is that the answer to my earlier questions now seem so blindly obvious, I can no longer understand why it took me so long to figure it out. I believe the answer to "now what?" is this: now you try to figure out how the data actually should be analyzed. Just because you have a bunch of choices you thought were reasonable, or other people thought were reasonable, doesn't mean they are actually appropriate to this specific context and your specific dataset. Rather than the work ending with a multiverse analysis, it is actually just the starting point for digging towards the truth. The advantage of the multiverse analysis is you won't have to do all the digging in the dark. And the benefit of a system like AugMeet (and any system whatsoever that can provide support) is that everything in a multiverse is hard enough as it is, so anything that lightens any portion of the load can be a welcome relief that frees up resources that will probably be applied right back into the multiverse anyhow.

The above observations are not a claimed finding of this dissertation, just as the question I had would not pass as a proper research question. I share it, however, because it is a question that has kept me awake and been the center of countless hours of my thoughts, and I doubt that I am the only one who it has so afflicted. I also share my answer because, at least for me personally, it puts to bed a question I have had for more years than I care to count.

I have now made full use of all the story-time indulgence I requested. In the the remaining section of this final chapter I share my suggestions for future work.

5.2 Suggestions for Future Work

After reflecting on what I have learned in the course of completing this dissertation, and on the limitations of that knowledge, I now humbly suggest a few directions for future research. I have limited this list to the things that I think I would would like to do, were it not time for me to move on to whatever it is that comes next in life.

The first subsection is dedicated to my musings on augmented presentation, while the subsequent subsection holds a list of other ideas in no discernible order and without a singular guiding theme.

5.2.1 What is the role of augmented presentation?

I believe the practice of augmented presentation, which was demonstrated in AugChiro and AugMeet, represents a fantastic way to *potentially* make discussions of data easier to have, more interesting, and more pleasant. I emphasize 'potential', because it is important to consider that it is beyond a doubt possible to use augmented presentation to make things so much worse that people long to see bullet point text explanations in Comic Sans on a bright green background.

On the other hand, I personally have had a dismissive attitude in the past towards the supposed line of thinking in HCI that assumes things will be better because they are easier or simpler. Yet when working with complex statistical analyses or multiverse analyses, where cognitive load is often maxed out because there is so much to look for and think about, I have gained a new appreciation for the value of seeking out ways to make lower-level tasks easier so as to support higher-level ones. In short: first find what works; remove anything not clearly essential to making it work; add things only when their value is resoundingly positive and substantial in magnitude.

Given all this, I think augmented presentation is more than a minor or unrelated add-on. People have limited mental energy, nearly everyone has to expend at least some effort to pay attention and be an active listener, and people naturally find it easier to think about and be involved with experiences they find pleasant. As obvious as this might seem, if these details are not attended to then the true value any other beneficial aspects of the system have can be completely lost.

While augmented presentation requires technological tools to perform it, it is nonetheless a *performance*. Some people are going to be better at on their first try, no one will be as good as they can be with practice, some people are going to want to do it, and some people will not. However, as a performance, prior work has hardly scratched the very surface of what can and should be done, and there is very little directly relevant work to inform how systems can help people get better, support the achievement of excellent performance, and so on. Now that it is at least technologically possible to give an augmented presentation of data, both as demonstrated in AugChiro and AugMeet

and as publicly announced as a future feature of Tableau, a whole new avenue of research is now possible. How else can it be done? What are the effects of doing it different ways, and do these effects vary by person—how and why? How can it be better supported? How can people be supported in getting better at it? What happens when you try to apply it to a new context or use case? What can be done to enable and support the performance of really great augmented presentations?

Finally, the recurring themes of attention, focus, and engagement that have been mentioned by so many participants (and other people who have commented on the topic) in relation to augmented presentation suggests to me that there might be a disproportionate effect on some subsets of the neurodivergent population. Pleasantly, I suspect some effects might be disproportionately positive for people with characteristics similar to those that are common with ADHD, and perhaps also some on the autism spectrum, though I would only have encountered people who are generally considered high-functioning in respect to this particular topic. If some uses of augmented presentation are helpful for most people, and in addition are especially helpful to some people who otherwise have trouble in contexts like this, that would really be a really nice thing to know. Of course if the reality were less happy, and is almost certainly more complex and nuanced, that would be important to know too.

5.2.2 Other Ideas and Unanswered Questions

How can convergent thinking be encouraged without suppressing divergent thinking? Prior to evaluating AugMeet, I was concerned that the proximate goal of reducing uncertainty might unintentionally discourage people from divergent insights, such as realizing there were other options or parameters that should be considered. I wagered that people who were interested in multiverse analysis like me might already be naturally quite good at divergent thinking, but could use help to make progress towards convergence. Participants still came up with great divergent insights, so my original wager at least was not completely wrong, but I still feel like there is far more that needs to be understood about supporting convergent thinking in contexts like this, and especially without discouraging or preventing the type of divergent thinking that is still critical.

What happens when you try to apply the uncertainty reduction workflow to other multiverses? While simple enough in concept, and with the proviso that I did at least have multiple multiverses in mind when I designed AugMeet and participants thought it would be applicable and useful in cases they know about, there is still no substitute to what can be learned by a full experience of trying something with a new use case. I know of at least one person other than me who will give it a try, but there is no way of knowing what new interesting patterns, strategies, or visualizations might suggest themselves.

What do you do with a multiverse that has multiple outcomes? I am aware of prior work that

integrates a few extra details for outcomes that have both point estimates and upper and lower ranges like uncertainty intervals. But what do you do—what can you do—when there may be multiple related yet discrete outcomes of interest? Naively I would first try generating a multiverse for each and look at them separately, but I would think there must be some better way to consider this problem. This problem in particular was pointed out to me by someone who wanted to consider more complex models, especially ones that can model all sorts of non-linear relationships. Perhaps some work on the topic of model comparisons could be guidance about what could be applicable to a multiverse? But what about other cases where multiple outcome metrics are of interest?

How about an actual tool to help interrogate a multiverse in some of the ways demonstrated in AugMeet, and beyond? Multiple participants expressed their wish to have a tool that actually made this as easy to do as it looked in the video. For example, the multiverse package in R is a tremendous help in executing the analyses, but pulling out individual universes is still a chore that involves a lot of copy-pasting and manual editing, and even then I myself spent many hours on what certainly could be done in a matter of minutes with the help of a tool (generating a visualization to compare parameter A and B, creating plots comparing a specific universe and its twin). With a tool to make doing a multiverse interrogation as easy as it looked in the video, it would really expand the opportunities for further study.

Visualizations for Multiverse Interrogation: The visualizations I used to demonstrate AugMeet were as simple as I thought could possibly work, and adjusted until they actually worked, and that is just about all. I would think there must be all sorts of visualization designs to support the stages of the uncertainty reduction workflow that are better than what I came up with.

5.3 Dissertation Conclusion

In this dissertation I define and ground the concept of analytical uncertainty, propose multiverse analysis as a way of assessing and communicating this distinct type of uncertainty, and ultimately demonstrate an interactive system designed to support the use of multiverse analysis as a way to reduce analytical uncertainty. The aim of this work is to pare away a slice of the ontological uncertainty present in all empirical research, and render it in such a way that it can be assessed, communicated, and reduced. Analytical uncertainty is far from trivial to assess, communicate, and reduce, but the techniques developed and explicated within this dissertation show how it is indeed possible and can be practicable.

BIBLIOGRAPHY

[1] Dharma Akmon, Margaret Hedstrom, James D. Myers, Anna Ovchinnikova, and Inna Kouper. Building tools to support active curation: Lessons learned from sead. *International Journal of Digital Curation*, 12:76–85, 1 2018.

[2] Robert Amar, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*, pages 111–117, 2005.

[3] Ruben C. Arslan, Katharina M. Schilling, Tanja M. Gerlach, and Lars Penke. Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *Journal of Personality and Social Psychology*, 8 2018.

[4] Piero Baraldi and Enrico Zio. A combined monte carlo and possibilistic approach to uncertainty propagation in event tree analysis. *Risk Analysis: An International Journal*, 28(5):1309–1326, 2008.

[5] Jojanneke A Bastiaansen, Yoram K Kunkels, Frank J Blaauw, Steven M Boker, Eva Ceulemans, Meng Chen, Sy-Miin Chow, Peter de Jonge, Ando C Emerencia, Sacha Epskamp, et al. 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, 2020.

[6] Michael Behrisch, Benjamin Bach, Nathalie Henry Riche, Tobias Schreck, and Jean-Daniel Fekete. Matrix reordering methods for table and network visualization. *Computer Graphics Forum*, 35(3):693–716, 2016.

[7] Peter L. Bernstein. *Against the Gods: The Remarkable Story of Risk*. John Wiley & Sons, 1996.

[8] Marc F. P. Bierkens. Global hydrology 2015: State, trends, and directions. *Water Resources Research*, 51(7):4923–4947, 2015.

[9] Georges-Pierre Bonneau, Hans-Christian Hege, Chris R Johnson, Manuel M Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization*, pages 3–27. Springer, 2014.

[10] Rotem Botvinik-Nezer, Felix Holzmeister, Colin F Camerer, Anna Dreber, Juergen Huber, Magnus Johannesson, Michael Kirchler, Roni Iwanir, Jeanette A Mumford, R Alison Ad-

cock, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 582:84–88, 2020.

[11] George E. P. Box. Science and statistics. *Journal of the American Statistical Association*, 71:791–799, 12 1976.

[12] Stephan B. Bruns and John P. A. Ioannidis. p-curve and p-hacking in observational research. *PLOS ONE*, 11:e0149144, 2 2016.

[13] Christopher Bryan, David S. Yeager, and Joseph O'Brien. Replicator degrees of freedom allow publication of misleading 'failures to replicate'. *SSRN Electronic Journal*, pages 1–56, 2019.

[14] Leonardo Bursztyn, Aakaash Rao, Christopher Roth, and David Yanagizawa-Drott. Misinformation during a pandemic. *SSRN Electronic Journal*, 2020.

[15] Stuart Card, Jock Mackinlay, and Ben Shneiderman. *Readings in Information Visualization: Using Vision To Think*. 01 1999.

[16] Joshua Carp. On the plurality of (methodological) worlds: Estimating the analytic flexibility of fmri experiments. *Frontiers in Neuroscience*, 6:1–13, 2012.

[17] Dylan Cashman, Adam Perer, Remco Chang, and Hendrik Strobelt. Ablate, variate, and contemplate: Visual analytics for discovering neural architectures. *IEEE transactions on visualization and computer graphics*, 26(1):863–873, 2019.

[18] Joseph Cesario, David J Johnson, and William Terrill. 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):586–595, 2019.

[19] Björn Christensen and Sören Christensen. Are female hurricanes really deadlier than male hurricanes? *Proceedings of the National Academy of Sciences*, 111(34):E3497–E3498, 2014.

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

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

[22] B. Craft and P. Cairns. Beyond guidelines: what can we learn from the visual information seeking mantra? In *Ninth International Conference on Information Visualisation (IV'05)*, pages 110–118, 2005.

[23] Geoff Cumming. The new statistics: Why and how. *Psychological science*, 25(1):7–29, 2014.

[24] Subhajit Das, Dylan Cashman, Remco Chang, and Alex Endert. Beames: Interactive multimodel steering, selection, and inspection for regression tasks. *IEEE computer graphics and applications*, 39(5):20–32, 2019.

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

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

[27] Marco Del Giudice, Steven W Gangestad, and W Steven. A traveler's guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions, 2020.

[28] Matthew J. Denny and Arthur Spirling. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189, 2018.

[29] Seamus Donnelly, Patricia J. Brooks, and Bruce D. Homer. Is there a bilingual advantage on interference-control tasks? a multiverse meta-analysis of global reaction time and interference cost. *Psychonomic Bulletin & Review*, 26:1122–1147, 8 2019.

[30] Pierre Dragicevic, Yvonne Jansen, Abhraneel Sarma, Matthew Kay, and Fanny Chevalier. Increasing the transparency of research papers with explorable multiverse analyses. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA, 2019. Association for Computing Machinery.

[31] Julien Dubois, Paola Galdi, Yanting Han, Lynn K Paul, and Ralph Adolphs. Resting-state functional brain connectivity best predicts the personality dimension of openness to experience. *Personality neuroscience*, 1, 2018.

[32] Mahmoud Medhat Elsherif, Muhammet Ikbal Saban, and Pia Rotshtein. The perceptual saliency of fearful eyes and smiles: A signal detection study. *PloS one*, 12(3):e0173199, 2017.

[33] Sebastian S. Feger, Sünje Dallmeier-Tiessen, Albrecht Schmidt, and Pałel W. Woźniak. Designing for reproducibility: A qualitative study of challenges and opportunities in high energy physics. pages 1–14. ACM Press, 2019.

[34] Scott Ferson and Jack Siegrist. Verified computation with probabilities. *Working Conference on Uncertainty Quantification in Scientific Computing (WoCoUQ)*, pages 95–122, 2011.

[35] FiveThirtyEight. Hack your way to scientific glory, 2015.

[36] Brian R Gaines and T L Kohout. Possible automata. *Proceedings of the International Symposium on Multiple-Valued Logic*, 1975.

[37] Nils Gehlenborg and Bang Wong. Points of view: Heat maps. *Nature Methods*, 9(3):213, 2012.

[38] Andrew Gelman. The problems with p-values are not just with p-values. *The American Statistician*, pages 1–2, 2016.

[39] Andrew Gelman and Sander Greenland. Are confidence intervals better termed "uncertainty intervals"? *BMJ*, 366, 2019.

[40] Andrew Gelman and Eric Loken. 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, 2013.

[41] Andrew Gelman and Eric Loken. The statistical crisis in science. *American Scientist*, 102:460, 2014.

[42] Alan Gerber and Neil Malhotra. Do statistical reporting standards affect what is published? publication bias in two leading political science journals. *Quarterly Journal of Political Science*, 3:313–326, 10 2008.

[43] Kelly Gildersleeve, Martie G. Haselton, and Melissa R. Fales. Do women's mate preferences change across the ovulatory cycle? a meta-analytic review. *Psychological Bulletin*, 140:1205–1259, 2014.

[44] Brian D Hall, Lyn Bartram, and Matthew Brehmer. Augmented chironomia for presenting data to remote audiences. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14, 2022.

[45] Brian D Hall, Yang Liu, Yvonne Jansen, Pierre Dragicevic, Fanny Chevalier, and Matthew Kay. A survey of tasks and visualizations in multiverse analysis reports. In *Computer Graphics Forum*, volume 41, pages 402–426, 2022.

[46] Christine R. Harris, Aimee Chabot, and Laura Mickes. Shifts in methodology and theory in menstrual cycle research on attraction. *Sex Roles*, 69:525–535, 2013.

[47] A.W. Harzing. Publish or perish, 2007.

[48] Håvard Hegre and Nicholas Sambanis. Sensitivity analysis of empirical results on civil war onset. *Journal of conflict resolution*, 50(4):508–535, 2006.

[49] John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2:e124, 8 2005.

[50] John P.A. Ioannidis and Thomas A. Trikalinos. An exploratory test for an excess of significant findings. *Clinical Trials*, 4:245–253, 2007.

[51] Zubin Jelveh, Bruce Kogut, and Suresh Naidu. Political language in economics, 2018.

[52] Kiju Jung, Sharon Shavitt, Madhu Viswanathan, and Joseph M Hilbe. Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, 111(24):8782–8787, 2014.

[53] Kiju Jung, Sharon Shavitt, Madhu Viswanathan, and Joseph M. Hilbe. Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, 111:8782–8787, 6 2014.

[54] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31:105–112, 2009.

[55] Matthias Kraus, Katrin Angerbauer, Juri Buchmüller, Daniel Schweitzer, Daniel A Keim, Michael Sedlmair, and Johannes Fuchs. Assessing 2d and 3d heatmaps for comparative analysis: An empirical study. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[56] Justin F. Landy, Miaolei (Liam) Jia, Isabel L. Ding, Domenico Viganola, Warren Tierney, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Charles R. Ebersole, Quentin F. Gronau, Alexander Ly, Don van den Bergh, Maarten Marsman, Koen Derks, Eric-Jan Wagenmakers, Andrew Proctor, Daniel M. Bartels, Christopher W. Bauman, William J. Brady, Felix Cheung, Andrei Cimpian, Simone Dohle, M. Brent Donnellan, Adam Hahn, Michael P. Hall, William Jiménez-Leal, David J. Johnson, Richard E. Lucas, Benoît Monin, Andres Montealegre, Elizabeth Mullen, Jun Pang, Jennifer Ray, Diego A. Reinero, Jesse Reynolds, Walter Sowden, Daniel Storage, Runkun Su, Christina M. Tworek, Jay J. Van Bavel, Daniel Walco, Julian Wills, Xiaobing Xu, Kai Chi Yam, Xiaoyu Yang, William A. Cunningham, Martin Schweinsberg, Molly Urwitz, and Eric L. Uhlmann. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, 1 2020.

[57] Yang Liu, Tim Althoff, and Jeffrey Heer. 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*, pages 1–14, 2020.

[58] Yang Liu, Alex Kale, Tim Althoff, and Jeffrey Heer. Boba: Authoring and visualizing multiverse analyses, 2020.

[59] Tina B. Lonsdorf, Maren Klingelhöfer-Jens, Marta Andreatta, Tom Beckers, Anastasia Chalkia, Anna Gerlicher, Valerie L. Jentsch, Shira Meir Drexler, Gaetan Mertens, Jan Richter, Rachel Sjouwerman, Julia Wendt, and Christian J. Merz. Navigating the garden of forking paths for data exclusions in fear conditioning research. *eLife*, 8:1–36, 12 2019.

[60] Steve Maley. Statistics show no evidence of gender bias in the public's hurricane preparedness. *Proceedings of the National Academy of Sciences*, 111(37):E3834–E3834, 2014.

[61] Daniel Malter. Female hurricanes are not deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, 111(34):E3496–E3496, 2014.

[62] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC, 1 edition, 2015.

[63] Joint Committee For Guides In Metrology. Evaluation of measurement data — guide to the expression of uncertainty in measurement. *International Organization for Standardization Geneva ISBN*, 50:134, 2008.

[64] Thomas Mühlbacher, Lorenz Linhardt, Torsten Möller, and Harald Piringer. Treepod: Sensitivity-aware selection of pareto-optimal decision trees. *IEEE transactions on visualization and computer graphics*, 24(1):174–183, 2017.

[65] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):0021, 2017.

[66] John Muñoz and Cristobal Young. We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology*, 48(1):1–33, 2018.

[67] Tamara Munzner. *Visualization Analysis and Design*. A K Peters/CRC Press, 12 2014.

[68] Brian A Nosek, George Alter, George C Banks, Denny Borsboom, Sara D Bowman, Steven J Breckler, Stuart Buck, Christopher D Chambers, Gilbert Chin, Garret Christensen, et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

[69] Ingram Olkin, Issa J. Dahabreh, and Thomas A. Trikalinos. Gosh - a graphical display of study heterogeneity. *Research Synthesis Methods*, 3:214–223, 9 2012.

[70] Amy Orben, Tobias Dienlin, and Andrew K Przybylski. Social media's enduring effect on adolescent life satisfaction. *Proceedings of the National Academy of Sciences*, 116(21):10226–10228, 2019.

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

[72] Amy Orben and Andrew K Przybylski. Screens, teens, and psychological well-being: evidence from three time-use-diary studies. *Psychological science*, 30(5):682–696, 2019.

[73] Arthur Pap. Theory of definition. *Philosophy of science*, 31(1):49–54, 1964.

[74] Harold Pashler and Eric–Jan Wagenmakers. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7:528–530, 11 2012.

[75] Chirag J. Patel, Belinda Burford, and John P.A. Ioannidis. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68:1046–1058, 9 2015.

[76] Charles Perin, Pierre Dragicevic, and Jean-Daniel Fekete. Revisiting bertin matrices: New interactions for crafting tabular visualizations. *IEEE transactions on visualization and computer graphics*, 20(12):2082–2091, 2014.

[77] Juho Piironen and Aki Vehtari. Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, 27:711–735, 5 2017.

[78] Gregory J. Poarch, Jan Vanhove, and Raphael Berthele. The effect of bidialectalism on executive function. *International Journal of Bilingualism*, 23:612–628, 4 2019.

[79] Julia M. Rohrer, Boris Egloff, and Stefan C. Schmukle. Probing birth-order effects on narrow traits using specification-curve analysis. *Psychological Science*, 28:1821–1832, 12 2017.

[80] Omer Sagi and Lior Rokach. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249, 2018.

[81] Abhraneel Sarma, Alex Kale, Michael Moon, Nathan Taback, Fanny Chevalier, Jessica Hullman, and Matthew Kay. multiverse: Multiplexing alternative data analyses in r notebooks (version 0.6.1.2). *OSF Preprints*, 2021.

[82] Leonard J. Savage. *The Foundations of Statistics*. Dover Publications, Inc., 2 edition, 1972.

[83] Michael Sedlmair, Christoph Heinzl, Stefan Bruckner, Harald Piringer, and Torsten Möller. Visual parameter space analysis: A conceptual framework. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2161–2170, 2014.

[84] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. 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, sep 2018.

[85] R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, and B. A. Nosek. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1:337–356, 9 2018.

[86] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science*, 22(11):1359–1366, 2011.

[87] Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143:534–547, 2014.

[88] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN Electronic Journal*, pages 1–26, 2015.

[89] Uri Simonsohn, Joseph P. Simmons, and Leif D. Nelson. Specification curve analysis. *Nature Human Behaviour*, 4:1208–1214, 11 2020.

[90] Paul E. Smaldino and Richard McElreath. The natural selection of bad science. *Royal Society Open Science*, 3:160384, 9 2016.

[91] David Spiegelhalter. Risk and uncertainty communication. *Annual Review of Statistics and Its Application*, 4:31–60, 3 2017.

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

[93] Bret Victor. Explorable explanations. Online. http://worrydream.com/ExplorableExplanations/, 2011.

[94] Kirsten G. Volz and Gerd Gigerenzer. Cognitive processes in decisions under risk are not the same as in decisions under uncertainty. *Frontiers in Neuroscience*, 6:1–6, 2012.

[95] Martin Voracek, Michael Kossmeier, and Ulrich S Tran. Which data to meta-analyze, and how? *Zeitschrift für Psychologie*, 227:64–82, 2019.

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

[97] Junpeng Wang, Subhashis Hazarika, Cheng Li, and Han-Wei Shen. Visualization and visual analysis of ensemble data: A survey. *IEEE transactions on visualization and computer graphics*, 25(9):2853–2872, 2018.

[98] Jelte M. Wicherts, Coosje L.S. Veldkamp, Hilde E.M. Augusteijn, Marjan Bakker, Robbie C.M. van Aert, and Marcel A.L.M. van Assen. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7:1–12, 2016.

[99] Cindy Xiong, Joel Shapiro, Jessica Hullman, and Steven Franconeri. Illusion of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):853–862, 2020.

[100] Cristobal Young. Model uncertainty and the crisis in science. *Socius*, 4, 2018.

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