




Visualization Guardrails: Designing Interventions Against Cherry-Picking in Interactive Data Explorers

Maxim Lisnic , Zach Cutler , Marina Kogan , and Alexander Lex 

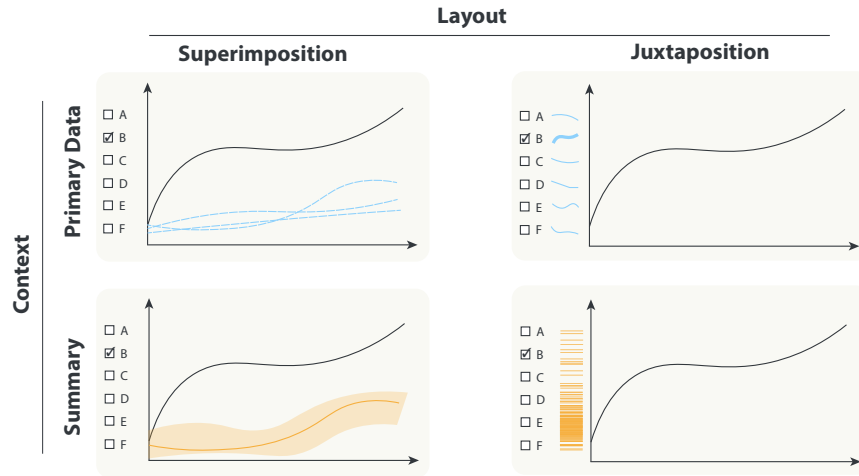


Fig. 1: The design space of visualization guardrails against cherry-picking along two dimensions: what *context* is shown (primary data or a summary) and *layout*, and where it is shown (superimposed on or juxtaposed with the main chart). The Figure shows the main data in black and the context in color: primary data in blue and summary in orange.

Abstract—The growing popularity of interactive time series exploration platforms has made visualizing data of public interest more accessible to general audiences. At the same time, the democratized access to professional-looking explorers with preloaded data enables the creation of convincing visualizations with carefully cherry-picked items. Prior research shows that people use data explorers to create and share charts that support their potentially biased or misleading views on public health or economic policy and that such charts have, for example, contributed to the spread of COVID-19 misinformation. Interventions against misinformation have focused on post hoc approaches such as fact-checking or removing misleading content, which are known to be challenging to execute. In this work, we explore whether we can use visualization design to impede cherry-picking—one of the most common methods employed by deceptive charts created on data exploration platforms. We describe a design space of *guardrails*—interventions against cherry-picking in time series explorers. Using our design space, we create a prototype data explorer with four types of guardrails and conduct two crowd-sourced experiments. In the first experiment, we challenge participants to create cherry-picked charts. We then use these charts in a second experiment to evaluate the guardrails’ impact on the perception of cherry-picking. We find evidence that guardrails—particularly superimposing relevant primary data—are successful at encouraging skepticism in a subset of experimental conditions but come with limitations. Based on our findings, we propose recommendations for developing effective guardrails for visualizations.

Index Terms—Visualization, cherry-picking, general public visualizations, misinformation interventions.

1 INTRODUCTION

Open data exploration platforms democratize access to data and visualizations of public importance. Examples include COVID-19 case dashboards on OurWorldInData [34], stock or cryptocurrency performance charts on Yahoo! Finance [53], or graphing platforms for various economic and policy indicators, such as Gapminder [42] or FRED [19]. Users not only create charts on these platforms for their own use, but also commonly share them on social media. While the popularity of data exploration platforms is a testament to their utility, the unconstrained and often unguided selection of data subsets and time periods that is commonly featured in such tools can lead to conclusions made

based on *cherry-picked data*. Such data visualizations, when shared on social media, are often misrepresented in a way that supports wrong and, in the worst case, harmful arguments. We have previously shown that cherry-picking of items and time frames is an issue that contributed to misinformation arguments in over 40% of COVID-skeptic charts shared on Twitter, most of which were screenshots of data explorers [32].

It is important to acknowledge that the problem of cherry-picking in public-facing data explorers is a *wicked problem* [41]. Biased interpretations of data, and information in general, are entangled with the viewers’ data and visualization literacy, data production quality, social and political movements, or the sense of belonging to a group with a strong ideology. This problem is hence wicked in the sense that, due to its complexity, it is resistant to any single solution. Rather, it requires continuous adaptation and innovation of approaches and interventions. In this paper, we set out to explore the possibility of interventions tackling cherry-picking in time series explorers from the perspective of visualization design. We asked ourselves: **is there anything we can do when designing and implementing data exploration platforms**

- Maxim Lisnic is with the University of Utah. Email: maxim.lisnic@utah.edu
- Zach Cutler is with the University of Utah. Email: zcutler@sci.utah.edu
- Marina Kogan is with the University of Utah. Email: kogan@cs.utah.edu
- Alexander Lex is with the University of Utah. Email: alex@sci.utah.edu

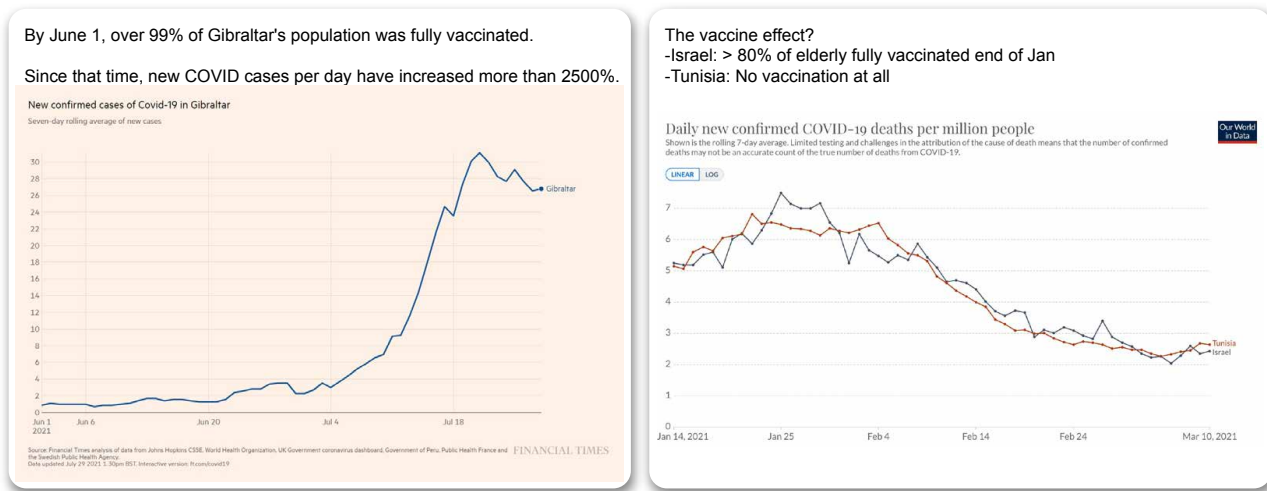


Fig. 2: Examples of tweets that spread vaccine hesitancy using screenshots of time series explorers with cherry-picked data. The [tweet on the left](#) used the Financial Times explorer to show an increase in cases in a single highly-vaccinated country (Gibraltar), implying that vaccines are harmful. The [tweet on the right](#) used the OurWorldInData explorer [34] to show two countries of different vaccination levels with similar trends of deaths (Israel and Tunisia), implying that vaccines are ineffective.

to minimize mis-use, mis-understandings, and mis-interpretations? To answer this question, we propose looking at the problem of misinterpretation and misuse of data explorers through the lens of **threat modeling** [44], and explore the possible **design space of interventions against cherry-picking** in interactive visualizations, which we term *visualization guardrails*.

We break down the issue of misinformative charts into two distinct yet equally important problems. Firstly, the *production problem*: data exploration interfaces make it very easy to create (perhaps, even nudge users towards creating) cherry-picked views. The circumstances that may lead an individual to creating a cherry-picked chart can vary from a desire to maliciously misinform, to innocent ignorance, or even to a genuine reason to be interested in zooming in on a certain subset of data. In any case, however, it should be more difficult to end up with a potentially misleading view.

The second problem is the *reaction problem*: when the resulting view is then shared with others through social media, it may end up spreading the incorrect insight and convincing others. Moreover, charts created with data exploration platforms are typically adorned with a veneer of impartiality and reputability offered by the data exploration platform’s logo and recognizable design, and hence may seem authoritative [31]. For an example, see the cherry-picked charts in the style of the Financial Times and Our World in Data in Figure 2, both reputable sources. Therefore, another major design goal for the guardrails is to introduce more nuance to views created with such reputable platforms.

To restate, our research questions are:

- What is the **design space** of guardrails against cherry-picking in data visualizations?
- Can we design guardrails that make cherry-picked charts **harder to produce**?
- Can we design guardrails that make cherry-picked charts less convincing and lead to a **more skeptical reaction** from the audience?

In this work, we describe the design space of visualization guardrails and implement a prototype data explorer with four distinct guardrails. We limit our designs to line charts showing time-series data because we previously found this to be the dominant form of charts and data used for cherry-picking in a social media context [32]. In order to investigate both the production and the reaction problems, we conducted two crowd-sourced experiments using our prototype of guardrails that target item cherry-picking. In the first experiment, we challenged participants to create cherry-picked views using control and guardrail interfaces. We then showed these charts to another set of participants in the second experiment and asked them to make a monetary decision based on the chart. Our findings show that the guardrails achieved the desired effect

of encouraging skepticism in a subset of scenarios. At the same time, a large part of the audience ignored or misunderstood the guardrails and instead focused on the main features of the visualization. Based on our results, we outline recommendations for designing guardrails against cherry-picking in data explorers.

To summarize, our paper makes several contributions:

- A conceptual framework for tackling issues of misuse, misinterpretations, and misinformation using data visualizations.
- A definition of the design space of visualization guardrails.
- A crowd-sourced user study exploring the effects of guardrails on the production of visualizations of cherry-picked data.
- A crowd-sourced user study evaluating the effects of guardrails on the audience’s reaction to cherry-picked visualizations.
- Recommendations for designing visualization guardrails, based on the results of our two user studies.

2 A THREAT MODELING FRAMEWORK FOR VISUALIZATIONS

In this section, we propose adopting the concept of *threat modeling* from computer security as a way of approaching the problem of data visualizations misused in support of misinformation.

2.1 What is Threat Modeling?

The Threat Modeling Manifesto, put forward by a group of security researchers aiming to promote security and privacy during software development, defines threat modeling as “analyzing representations of a system to highlight concerns about security and privacy characteristics” [7]. Threat modeling allows the researcher to identify things that can go wrong, pinpoint design issues, and inform mitigation measures.

In his book *Threat Modeling: Designing for Security* [44], Adam Shostack proposes a four-step framework for threat modeling, which involves answering four questions: (1) What are you building?, (2) What can go wrong?, (3) What should you do about it?, and (4) Did you do a decent job?

Example outcomes of threat modeling in the security context could be identifying opportunities for hackers to access confidential information in a data base, or designing interventions against denial-of-service attacks. But while there is a number of specific computer security-related checklists or domain-specific heuristics, threat modeling itself is a value- and principle-driven approach. Therefore, it is highly adaptable to a wide variety of security and privacy issues, as well as (as we will argue below) issues of data and visualization misuse.

2.2 Applying Threat Modeling to Visualization Threats

Next, we go over the four steps of the framework with the goal of outlining the issue of misuse of interactive visualizations used by the general public to support misinformation, as well as motivating the application of threat modeling to this issue.

What are You Building? For the purposes of this work, we focus on general public-facing interactive time series exploration portals created by local government agencies such as state and county epidemiology tracking dashboards, news organizations like Yahoo! Finance [53], or specialized data exploration platforms such as OurWorldInData [34]. In this step, the data explorer platform governance should outline the core functionality and values of the platform that should be protected and not compromised on during the threat modeling exercise. We assume that examples of such core functionality could include instant access to data for everyone, freedom of exploration without major restrictions, ability to export and share views and data, among others.

What Can Go Wrong? To answer this question, we can turn to our previous work that outlined reasoning errors in social media users' interpretations of data visualizations [32]: cherry-picking favorable subsets of data, assigning causality to salient features of charts, or not accounting for common statistical fallacies. In this work, we focus on *cherry-picking*—one of the most often used tactics—as an illustrative example. Figure 2 shows two instances of tweets using cherry-picked data explorer charts to spread misinformation and to promote vaccine hesitancy.

In his testimony before the US House of Representatives, climate scientist Richard Somerville described cherry-picking as “[making] selective choices among competing evidence, so as to emphasize those results that support a given position, while ignoring or dismissing any findings that do not support it” [45]. Some of the core goals described in the previous section, such as access to data and freedom of exploration, result in danger of making selective choices and emphasizing those results, pointing to a potential issue.

What Should You Do About It? Shostack outlines four possible paths of action: (1) *accept* that there is an issue and do nothing, (2) *eliminate* the feature causing an issue, (3) *transfer* the responsibility to the user, or (4) *mitigate* the issue [44]. All four strategies are feasible for data explorer platforms. An example of acceptance is simple: one could do nothing. Eliminating a feature could take the form of restricting problematic interactions by, for instance, not allowing plotting any two time series on the same chart. Transferring responsibility to the user may involve requiring data or visualization literacy evaluations, but may be unrealistic to realize in platforms accessible to the general public.

Although we urge platform designers to consider all of the above strategies when performing threat modeling of their tools, in this paper we will pursue the goal of *mitigation* of the cherry-picking threat and design *guardrails* against it. In our context, due to the high possibility of bad-faith actors purposefully seeking out cherry-picked views, we posit that any transfer of responsibility strategy would be futile. We also believe that the benefits of the “problematic” features (i.e., unrestricted freedom of exploration and ability to take screenshots and share) are high, and therefore we opt to explicitly maintain such features and instead design mitigation strategies.

Did You Do a Decent Job? The goal of this step is to evaluate the success of the chosen intervention strategy. In order to answer this question, we conducted two rounds of evaluation studies, described in Sections 6 and 7. Firstly, we conducted a study challenging participants to produce cherry-picked views with and without our guardrails in an approach similar to using a Red Team in cybersecurity. This enabled us to conduct a second experiment in which a different set of participants were asked to review the charts produced in the first study.

3 RELATED WORK

In this section, we first discuss the role of cherry-picking of information for the purposes of spreading misinformation. Then, we present an

overview of existing work in data visualization on designing interventions against fallacies and cognitive biases.

3.1 Cherry-Picking and Questionable Research Practices

Cherry-picking favorable data or results has long been acknowledged as a questionable research practice across the scientific community, alongside HARKing (hypothesizing after the results are known) and p-hacking [1, 8]. Cherry-picking, or selectively showing only information that supports a given argument, is an outcome that could be unintentional and point to ignorant or inattentive practices, or be intentional and reflect malicious intent to misinform. **Unintentional cherry-picking** is a common behavior in judgement under uncertainty and could stem from a variety of cognitive biases, such as availability bias (focusing only on information that is readily available) [47], confirmation bias (focusing on information that supports prior beliefs) [38], and anchoring bias (focusing on information presented with first) [48]. **Intentional cherry-picking** is a strategy shown to support misinformation arguments about climate change denial [10], vaccine hesitancy [17], and is, more broadly, one of the hallmark tactics of denialism [26].

Cherry-picking is a practice that is not limited to data and visualizations. Quote-mining refers to cherry-picking a quote or a statement out of its original context [26]. Quoting others out of context has been practiced for a variety of purposes over the course of history, ranging from political propaganda [35] to misusing critics' quotes in misleading advertisements [40]. Both cherry-picking data and quote-mining statements can be described as sharing half-truths, meaning these tactics are difficult to debunk since they are based in truth [3]. It then follows that in order to combat the issue of focusing on a subset of information, an intervention must either raise awareness about or explicitly show or summarize the missing context.

3.2 Interventions Against Fallacies in Data Visualizations

Previous work on interventions against biases in visualization largely focuses on professional and scientific visual analytics, and specifically the forking paths problem [39], the multiple comparisons problem [54], and biases in exploratory data analysis that lead to an imbalance across many variables [49]. The interventions that have been proposed to tackle the exploratory problems primarily depend on the analyst's good faith, including strategies that automatically score their biasedness [49] similar to algorithmic approaches to detect cherry-picking in big data from the database literature [2, 29, 30], or including visualizations of the analyst's process [50]. In the context of the spread of online misinformation, we cannot typically rely on users to track their own biasedness and reliability—approaches that would fall under the strategy of transferring the responsibility described in Section 2.2.

Interventions that target biases and fallacies in narrative visualizations or at the audience level include using textual warnings against assuming that correlation equals causation [28], attaching multiple views to combat visualization mirages [56], adding interactive linking between text and data [55], as well as design alternatives for highlighting the truncation of the vertical axis [12]. Although the visualization community has raised concerns about the role of cherry-picked charts in the spread of misinformation across numerous studies [20, 21, 32, 33], to the best of our knowledge, this is the first work specifically attempting to design interventions against cherry-picking.

4 DESIGNING GUARDRAILS

In this section we describe our approach to characterizing the design space of visualization guardrails and outline the specifics and variations of the resulting design space.

4.1 Design Process

To develop a broad set of ideas, we engaged in a parallel prototyping process, where each of the four authors independently developed multiple designs [16]. Before the start of the design process, the authors put together the design brief. We agreed that the main goal of the designs would be to provide missing context and promote skepticism in the viewer [28]. Additionally, the designs should be:

- **Non-obtrusive:** there should be no restrictions on exploration and selections, and every commonly-allowed selection must be as visually salient with a guardrail as without. This goal stems from the fact that we target mitigation, rather than elimination of a feature (Step 3 of threat modeling).
- **Undemanding:** the guardrail should not directly slow down the user with pop-ups, questionnaires, or assessments. Making a user complete evaluations would more closely resemble transfer of responsibility (Step 3 of threat modeling).
- **Tamper-evident:** it should be difficult to get rid of the guardrail, for instance by cropping a screenshot. This goal is motivated by our previous work [32] that showed that most of misleading charts shared on Twitter were screenshots using various levels of cropping.

In order to come up with initial guardrail designs, all four authors independently created sketches based on the requirements. The sketches were either free-form or on top of examples of cherry-picked views, as identified by our previous work [32]. We provide all of our sketches in Appendix A. The first author then reviewed and organized the sketches into common themes and ideas, and all authors discussed the results and used them to describe the possible design space. Figure 1 presents an overview of the resulting design space along two dimensions: context, or *what* is shown, and layout, *where* it is placed.

4.2 Design Space

Given our design requirements and our designs, the task of designing a guardrail against cherry-picking in a data exploration platform is related to presenting a helpful *visual comparison*. Gleicher et al. described three ways objects could be visually compared: by superimposition, by juxtaposition, or by explicit encoding of differences [22, 23]. Explicitly encoding difference would, however, involve using a different representation of the selected items. For example, when comparing COVID cases in two countries, explicitly encoding differences could entail creating a derived dimension that subtracts the cases of the countries, and visualizing this derived dimension instead of the original data. As a result, this approach limits the saliency of selections (the original data is no longer shown) and violates our goal of non-obtrusiveness. However, the other two visual comparison strategies—*superimposition* and *juxtaposition*—fit our design requirements and describe where the guardrail could be placed.

While the visual comparison strategies describe they **layout** of the guardrails, or where the guardrail is shown, the other dimension of the resulting design space is **context**, or what is being shown. We categorize our design sketches into two types of context: the guardrail can either show *primary data* in the same units, level of aggregation, and visual language as the main data, but potentially sampled to a small set of items or *visual summaries*—transformations and aggregations of the data, or additional data that provides a summary context (e.g., a market index for stock data).

4.2.1 Layout

We describe two main types of guardrail placement: superimposition and juxtaposition. **Superimposed** guardrails exist on the same canvas and scales as the main chart data. A designer of a Superimposed guardrail has two main decisions to make. The first one is defining what contextual data should be shown as the guardrail for the given scenario

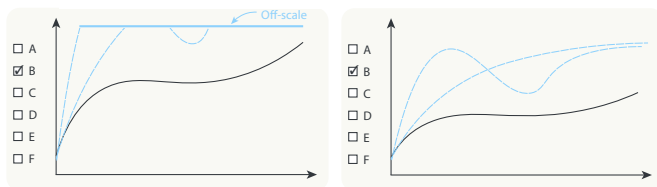


Fig. 3: Two design variations of vertical scale treatment in Superimposed guardrails. The example on the left preserves the scale of the main chart data, emphasizing the original selection but truncating the off-scale context. The design on the right adjusts the scale to fit the guardrails.

and domain. For instance, to tackle an instance of item cherry-picking, a COVID-19 data explorer showing Sweden’s cases would probably plot other Nordic countries, such as Denmark, Norway, and Finland as well. The second design decision to consider is the treatment of the vertical scale: specifically, whether the axis should be adjusted or not in cases when the contextual data would go above or below the original frame. Examples of this design variation can be seen in Figure 3. The axis could be zoomed out to include all contextual detail, thereby sacrificing saliency or detail of the main selection. Alternatively, the scale could be kept as-is, and instead would include a visual indication that there is out-of-frame context.

One of the main advantages of the Superimposed guardrails is that, by virtue of being plotted together with the main data, they are not just tamper-evident but virtually tamper-proof: it would be very difficult to remove the guardrails from the view. Additionally, Superimposed guardrails exist on the same scale and units as the main data and offer an opportunity to directly compare the values of the main chart data and the contextual data.

The disadvantages of Superimposed guardrails include the danger of overplotting—a lot of contextual information in the main frame of the visualization may not scale with many items selected. This problem could be alleviated by dynamically adjusting the size of the comparison set as a user chooses more items. Another problem is that it might not always be obvious which data items or time frames are useful “important context.” A system could leverage metadata (such as regions of the world for country data, or sectors for financial data) to make such a determination. A generic implementation that doesn’t require additional data could leverage statistical information, i.e., by including a set of representative examples of the data set.

The idea of the **Juxtaposed** layout of guardrails is to leave the main canvas of the visualization unchanged. Instead, we provide contextual information in a separate, juxtaposed view to the side, above, or below the chart. When designing Juxtaposed guardrails, the decision of their placement depends on the underlying data and target issue. For instance, if the goal of the guardrail is to give an indication that a climate change-denying author chose a small fraction of the time frame of ocean temperature data, the guardrail could run along the “problematic” dimension—in this case, the horizontal axis right below the chart (as seen in the example on the left in Figure 4).

In Juxtaposed guardrails, the decision about what constitutes important context data is less central than in Superimposed guardrails. Specifically, juxtaposition allows the designer to show *all* of the data points that would fit into the frame. Similarly, a Juxtaposed guardrail is easy to combine with other guardrails, since, unlike with Superimposed guardrails, there is no issue of overplotting the main chart.

However, Juxtaposed approaches are croppable and thus not tamper-evident (a design goal). Implementations could use strategies to reduce the croppability, for example, by embedding a Summary directly into the axis, such that cropping out the guardrail would also involve cropping out part of the axis. Alternatively, the designer could wrap the chart and the guardrails with a frame that could be indicative of parts left out. In general, however, even croppable designs could provide a

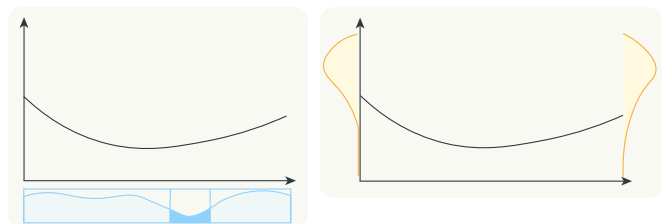


Fig. 4: Examples of Juxtaposed guardrail variations for time frame cherry-picking. On the left is an example of Juxtaposed Primary Data: the miniature view below shows the entire time frame and highlights the cherry-picked period of a dip in value. On the right is a periphery plot [37]—an instance of Juxtaposed Summary. The peripheries show that the value of interest is much higher in the periods before and after the selection.

degree of protection, as it would be much easier for an online audience to highlight cherry-picking if they could point to the full chart in replies or community notes.

4.2.2 Context

We distinguish between two types of guardrails' contextual data: they could show Primary Data of the same type as the main chart data, or they could use aggregated or transformed data in the form of a Summary.

Primary Data guardrails provide contextual data by directly plotting the primary data—meaning data at the same level of granularity and of the same units as the main data in the chart. Primary Data guardrails are shown in blue in Figure 1. The main advantage of Primary Data guardrails is their simplicity (both to implement and to understand): while they should be visually distinct from the main data, they still use the units as the chosen items and need virtually no explanation to be understood. This advantage is especially strong in a Superimposed Primary Data guardrail: it utilizes the same visual encoding and exists on the same axes as the main data, and thus its meaning is self-explanatory.

Summary guardrails, on the other hand, condense the primary data into a summary form for the purposes of providing context. They are shown in orange in Figure 1. There are many possibilities for the exact implementation of Summary guardrails, and, consequently, the designer should deliberate over what type of summary is the most meaningful, given data and domain. For instance, a simpler Summary could provide the average, inter-quartile range, or extrema of the data set. A more complicated Summary could show information about the variance or the shape of the distribution of the data.

Compared to Primary Data guardrails, the Summary guardrails are more compact and help mitigate the problem of overplotting by consolidating all of the contextual data into a single visual representation. On the downside, however, Summary guardrails may be more complex to understand as they represent a departure from the visual encoding of the main chart and may involve an uncommon visualization type. Therefore, it would be advisable to train the viewer to read the chart using annotations or training modules.

4.2.3 Implementation Alternatives

As alluded to in the previous sections, there is a number of design alternatives and decisions that a guardrail designer should consider. In this section, we describe specific implementation variations.

The **Superimposed Primary Data** guardrail is arguably the simplest, as it involves automatically plotting a subset of contextual data, there are few design choices to be made aside from defining contextual data. We observe two simple implementations of this technique in data explorers online. First is the scatter plot on Gapminder [42] that automatically includes translucent data points for the entire available data set, however the opacity of the context is adjustable and could be completely removed. The second example is Google Search: when looking up a macroeconomic or demographic metric for a given country, the resulting chart typically includes two other regions for reference. For instance, at the time of writing this paper, looking up “population of the US” returns a chart with the populations of the US over the past century, as well as those of Russia and Mexico in fainter lines.

Designers of **Superimposed Summary** guardrails have more liberty in implementation, as there are many types of visual summaries to choose from. When tackling time frame cherry-picking, the Superimposed Summary could take form of lagged variance information. Similar designs have long been popular in financial trading with the purpose of encoding past volatility of a financial instrument alongside its value, and include Bollinger Bands [6] and Standard Deviation Channels, or Envelope Channels. In the context of item cherry-picking, as discussed previously, Superimposed Summaries could take form of simple average, inter-quartile range, or extrema information of the contextual data using a representation similar to that of contour boxplots proposed by Whitaker et al. [51]. Aside from statistical summaries, a Superimposed Summary could aggregate the contextual data into a background heatmap using techniques such as DenseLines [36]. Superimposed

Summaries could also simply draw on common domain-specific benchmarks, such as stock market indices in the financial sector.

The main goal of **Juxtaposed Primary Data** guardrails is to show omitted items or time periods of the main data next to the main chart. An example of previously proposed visualization technique that could serve as a guardrail for time cherry-picking is stack zooming [27]. In the context of data explorers, Juxtaposed Primary Data could benefit from being miniature as to not be too large to gist or end up distracting from the main view. To achieve this, the specific implementations could be chosen from the set of charts designed to take up little space, such as sparklines [46] or horizon charts [25] in instances where the sign of the data provides meaningful context. Additionally, while this type of guardrail does not require a precise definition of “context data set” as Superimposed Primary Data, space constraints may necessitate decisions about what data is seen immediately beside the visualization. Juxtaposed Primary Data guardrails often can play a dual role and take the form of scented widgets [52]. For instance, a sparkline can serve as the preview of an item's data next to each item's selection checkbox—a technique commonly used in financial data explorers. In guardrails against time cherry-picking, the horizontal axis selection slider could instead be a miniature chart of the entire period which the user could brush (as seen in the example on the left in Figure 4). An example of an existing implementation of this is the exchange rate explorer on CoinMarketCap.com [5]: a user could zoom in onto a short period of, for instance, Bitcoin price chart by brushing over a miniature view below the main chart.

The guardrail type with the largest set of implementation alternatives is **Juxtaposed Summary**, as there are virtually no restrictions on what could be shown, as long as it provides useful context. One subset of design alternatives are variations of focus-and-context visualizations, or designs that allow the user to see both detail and overview simultaneously [9]. This could include distorted views, such as Multistream from Cuenca et al. [13]. Morrow et al.'s Periphery Plots [37] provide examples of non-distorted variations of this guardrail that could be helpful in the context of time cherry-picking (seen in the example on the right of Figure 4). Designs similar to periphery plots could similarly be used to tackle item cherry-picking—in this case, however, the Summary in the periphery would be calculated based on the omitted items rather than omitted time frame. The space of ways to show the distribution of contextual data is large and the choice of a specific visual encoding should depend on the features of the underlying distribution, as well as the expected audience's visual literacy, as these types of charts are typically less common in visualizations for general public. Correll provides a helpful analysis of advantages and disadvantages of distribution visualizations as well as their combinations as raincloud plots [11]. The example in Figure 1 as well as our prototype employ a vertical strip plot, but other designs we have considered include a box plot, a violin chart, or an inverted histogram (shown in the example on right in Figure 4).

5 PROTOTYPE DESIGN

To evaluate the utility of guardrails, we created prototype designs of each of the four design space quadrants: Superimposed Primary Data, Superimposed Summary, Juxtaposed Primary Data, and Juxtaposed Summary. Since visualization design decisions have to be driven by the data to be visualized, and our goal was to use the prototypes in the evaluation, we had to make decisions related to the data, features, and design. In the designing of the prototype, we prioritized *ecological validity*, therefore we used real-world data, and aimed to mimic the functionality and look of existing data explorer platforms, such as OurWorldInData's COVID-19 explorer [34]. Screenshots of our prototype implementation can be seen in Figures 5 and 8, as well as in Appendix B. A sandbox version of our prototype is available [online](#).

We chose to design for (and evaluate with) two different **data sets** that are commonly used in public-facing data exploration tools: COVID-19 death count data by country retrieved from OurWorldInData [34] and individual stock performance data retrieved from Yahoo! Finance [53]. Besides being popular data sets widely shared on social media, epidemiology and finance are also sufficiently different in topic and scale:

Select a view that best shows (and convinces your client) that the airline industry fund is the best investment.



Fig. 5: Screenshot of our prototype implementation of data explorer in the Study 1 experimental setup. Shown is the Stocks scenario with a Juxtaposed Primary Data guardrail.

epidemiology data is represented by ratio, non-zero values, while stock performance is typically shown as an interval value—percent change from a given point in time that can dip below zero. We anonymized the country and stock names to limit the influence of preconceived ideas about COVID-19 or particular stocks. We limited the number of items available to select to 15 or less to simplify the tasks in evaluation. We also chose to limit our **guardrails prototypes to item-cherry-picking**, excluding time-cherry-picking from our design and our study. We made this choice primarily to simplify our study design, as we already have four conditions and one control for two data sets to test.

To create these prototypes, we needed to make specific design decisions for each guardrail. Specifically, in the **Superimposed Primary Data** condition we defined contextual items to be countries from the same region in the Viral scenario and stocks from the same industry in the Stocks scenario. For **Superimposed Summary**, we showed the average of all items as a line and the interquartile range as a shaded area (Figure 8). For a more realistic presentation, we labeled the average line as “Marked Index” in the Stocks scenario.

In the **Juxtaposed Primary Data** condition, we chose to provide a sparkline of each item next to its label (Figure 5). We filled in the area under the line chart in light color, which made it easier to compare magnitudes among the small multiples that are stacked vertically relative to just lines. Lastly, for the **Juxtaposed Summary** guardrail we implemented a stripplot showing data of all items across all time points, with each tick representing the value of one item on a given day. The ticks are shown with slight transparency to account for overplotting, and we use color-coding to distinguish positive from negative values—a distinction relevant for the Stocks scenarios. A shaded funnel denotes which part of the global scale is currently shown on the main canvas, and also makes it more difficult to crop the plot without leaving evidence of tampering. For both of the Summary guardrails, we added a sentence explaining what the shaded area or the stripplot denote.

6 STUDY 1: PRODUCTION

In order to evaluate whether the guardrails make it *more difficult to cherry-pick data*, we conducted a crowd-sourced experiment challenging our participants to use our prototype data exploration platform with and without guardrails in place. One of the main goals of this study was to produce a data set of cherry-picked visualizations to be used in the reaction study. The other goal of this study was to evaluate whether participants find it more difficult to create such visualizations with guardrails on, as measured by their log of clicks, NASA Task Load Survey [24], as well as open text answers. This section describes our methodology and results.

6.1 Methods

The primary goal of the experiment was to get participants to use our data explorer with and without guardrails to create cherry-picked visualizations. In our experimental setup, we largely prioritized *ecological*

validity and therefore opted to create realistic scenarios and use real data. We devised four scenarios in which the participants were given a prompt, and then asked to select a view and write a caption that supports the prompt. Two of the scenarios—termed Viral A & B—were based on an anonymized data set of COVID-19 fatalities presented as a fictional viral disease in a fictional world, and the other two—Stock A & B—used a data set of individual stock performance over time, also drawn from real data but anonymized.

Although we purposefully did not use the term “cherry-picking” in the study description, in order to elicit cherry-picking, the participants were instructed to select data that best supports a given prompt. In the Viral scenarios, participants were asked to imagine that they work for a public health agency and were in charge of promoting a health policy campaign against a fictional viral disease that, judging by the data alone, does not appear to be the most effective policy. In the Stocks scenarios, participants had to imagine that they are a financial advisor who needs to promote funds that do not have strong historical performance as instructed by their boss, who claims to have deeper knowledge about the stock, without explicitly revealing so. In each scenario, the participants were required to write a short caption accompanying their visualization, to mimic the act of sharing a cherry-picked visualization on social media. Figure 5 shows a screenshot of the experimental setup.

Besides covering two different data sets and two topics, the scenarios also represented two different types of cherry-picking: in Viral A and Stock B the participants would need to “hide the competition,” or promote a group of items that have clearly better alternatives, while in scenarios Viral B and Stock A following the prompts would involve “hiding the bad apples,” or purposefully not revealing the poorer-performing subset of a group. Additionally, we attempted to design policy and stock recommendations in Viral A and Stock A to be relatively unreasonable, while prompts in Viral B and Stock B were more reasonable.

To conduct the study, we first conducted a first, in-person pilot experiment with 2 students in our school. The goal of the pilot was to establish that the scenario descriptions are straightforward to understand and that the prototype does not contain usability problems. As a result of the first pilot, we clarified the scenario description language, added a help button for participants to reference the scenarios during the task, and fixed one bug in the guardrail implementation. Next, we conducted a second, 10-participant pilot using the crowd-sourcing platform Prolific. The goal of the second pilot was to confirm that the tasks are understandable and could be reasonably completed by remote participants. The second pilot was successful and we proceeded to the main study without any changes.

For the main study, we recruited 130 participants on the crowd-sourcing platform Prolific who indicated to be fluent in English. The University of Utah IRB deemed the study exempt from full board review (IRB number 00173128). Each participant was randomly placed into one of our four guardrail conditions, and saw each of the four scenarios in a random order, two with one of the guardrails in place, and two without any guardrails, as controls. During each task, we kept a log of the participants’ selections as well as their final selection using the TTrack library [14]. After each task, the participants were asked to complete a NASA TLX [24] questionnaire to assess their subjective workload. At the end of the experiment, we revealed to participants that there has been a guardrail placed in half of their tasks, and asked them whether they noticed it and whether it had an effect on completing the task. The median completion time was 15 and a half minutes, and the participants were compensated \$5.00 for a median hourly rate of just over \$19/hr. The study was designed and implemented with the reVISit study framework [15]. Detailed instructions and prompts, as well as the whole study with all stimuli are available in [Appendix C](#) and [online](#).

After the experiment was concluded, the first and second authors independently reviewed the 520 submissions to evaluate the quality of the item selection and the text caption, and tagged them for potential problems. The authors matched in their evaluations on 395 occasions but disagreed about 125 submissions and then met to discuss the disagreements. After all the disagreements were resolved, 18% of submissions were removed for one of three reasons: (1) the participant

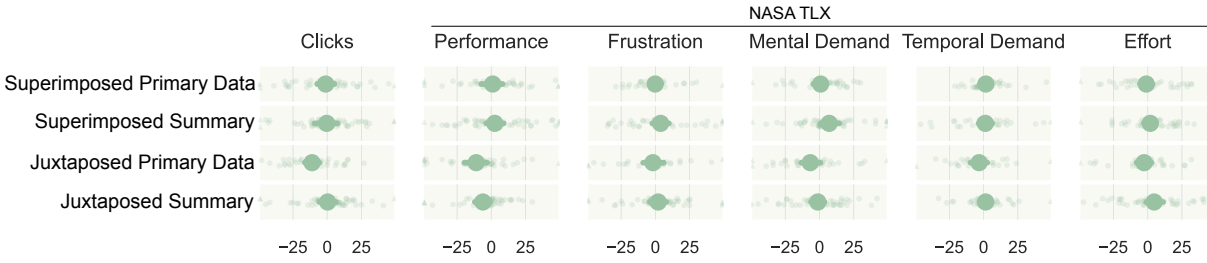


Fig. 6: Effects of guardrails on participants' task load in Study 1. Shown are the mean difference between the metrics for guardrail tasks and metrics for control tasks for each participant (0 means no effect of guardrail), as well as 95% bootstrapped confidence intervals. For clarity, large outliers (<5% of data) are shown as triangle markers on the edges. The interpretation of the NASA TLX Performance metric scale is: lower means better [24].

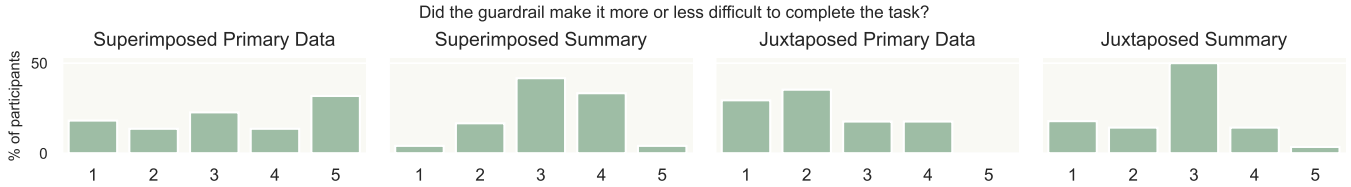


Fig. 7: Participants' perception of the guardrails' effect on the difficulty of performing the task in Study 1. Participants in the Superimposed Primary Data condition described the guardrail to make cherry-picking "much more difficult" (5 on the Likert scale) significantly more than those in other conditions. Juxtaposed Primary Data responses most often described the guardrail to make the task "much easier" (1 on the Likert scale).

did not follow the prescribed prompt and didn't select relevant items, (2) the caption submission was unintelligible, or (3) the caption submission was too short or generic (e.g., "Chart"). Following the review, the 427 remaining submissions were used for analysis.

We analyzed the resulting submissions within subjects quantitatively by comparing the number of clicks (or different views selected) and NASA TLX metrics with and without guardrail for each participant using paired T-tests. We also qualitatively analyzed the data by reviewing the caption and post-study feedback text and noting insightful themes.

6.2 Findings

As a result of Study 1, our participants generated 520 cherry-picked visualizations and captions, of which 427 passed our quality review and were used in further analysis. The submissions spanned 2 scenarios and 4 guardrails, as well as a control condition. The resulting submissions are available for view [online](#). Figure 8 shows an example submission as seen in the Study 2 interface.

Figure 6 presents an overview of the within-subjects difference of the number of clicks per task as well as the NASA TLX survey results. Although we elicited the physical demand evaluation as part of our NASA TLX questionnaire to be consistent with the original protocol [24], we do not analyze the results as they were not relevant to our study. As seen from Figure 6, the participants needed significantly fewer clicks to arrive at their final view in the Juxtaposed Primary Data condition, with an average of 21 clicks versus 32 clicks in the Control condition ($T(38) = -4.00, p < 0.0001$). Participants also needed less time to explore the data using Juxtaposed Primary Data: an average of 106 seconds compared to 143 seconds in the Control conditions ($T(38) = -2.12, p = 0.041$). The results of the NASA TLX survey show that participants reported better subjective performance using the Juxtaposed Primary Data (28.9 versus 39.7, $T(38) = -2.45, p = 0.019$) and the Juxtaposed Summary conditions (26.5 versus 31.7, $T(52) = -1.95, p = 0.055$), and lower mental demand with Juxtaposed Primary Data (48.8 versus 55.8, $T(38) = -1.84, p = 0.075$).

In their text responses, participants noted how the Juxtaposed guardrails actually made it *easier* for them to find the best cherry-picked view: "It made it easier to visualize without constantly turning off and on each data set to conceptualize which ones would work best." These results make sense, as Juxtaposed Primary Data gives a detailed yet easy-to-understand overview of the data set not only to the audience of a cherry-picked visualization, but also to the author. We note that we did not observe any text responses recounting any difficulties in the process of cherry-picking with a Juxtaposed guardrail. Figure 7 also shows that no one in the Juxtaposed Primary Data condition, and less than 5%

of Juxtaposed Summary participants reported that the guardrail made it "much more difficult" to perform the task, implying that Juxtaposed views do little to impede the production side of cherry-picking.

On the other hand, the Superimposed conditions were more frequently described as a hurdle to creating cherry-picked view. Those who saw the Superimposed Summary condition reported slightly higher mental demand (51.3 v. 43.5, $T(47) = 2.10, p = 0.041$). Their text responses elaborated that it was "harder to compare when you are telling me to convince someone [that option A is] a better choice when the line shows there is [something else] better", with another participant mentioning how they "couldn't easily show that one investment was better than others without hiding some truth." We also note that two participants declined to perform the task in the guardrail conditions, with one of the submitted captions simply reading: "I can't. I'd be a liar. [option A] is terrible."

As seen from Figure 7, those who saw the Superimposed conditions also reported that the guardrail made it "more difficult" and "much more difficult" to complete the tasks (4 and 5 on the 5-point Likert scale, respectively), with the highest value belonging to the Superimposed Primary Data view. While we carefully avoided using the term "cherry-picking" in the task description, a participant reported that the guardrail "interfered with my ability to cherry-pick the data I needed to."

A qualitative analysis of participants' captions also reveals their ability to adapt to the guardrails: in cases where the guardrails revealed context that would debunk cherry-picking, participants used text to shift the focus onto another salient feature or a redeeming quality of the data. For instance, although the guardrail clearly revealed that countries with viral containment Policy A have much more infections than others, one caption shifted the emphasis away from the magnitude and towards the trend: "Policy A's peak comes fast but comes back down just as fast." In another case, when tasked with promoting investments in Airline stocks, which the guardrail revealed to only produce average returns, one participant's caption instead promoted stock stability as a desirable attribute: "Over time, growth in stock prices in the airline industry have either been consistent with or outperformed the market average, thus being the most predictable." Another participant simply channeled optimism to explain away a dip in the investment return: "It's up and down, but it's currently down, indicating that it will increase again."

7 STUDY 2: REACTION

In order to evaluate whether the guardrails make *cherry-picked data less convincing*, we conducted a second crowd-sourced experiment asking another set of participants to review charts created by the participants of Study 1. This section describes our methodology and results.

You will need to travel to **Eldoril North (Policy A)** for work.
 You've come across this visualization and the accompanying caption.
 Please review the visualization and the caption, and then answer the questions below **based solely on this information**.
 (Please try to not rely on other visualizations you've seen.)

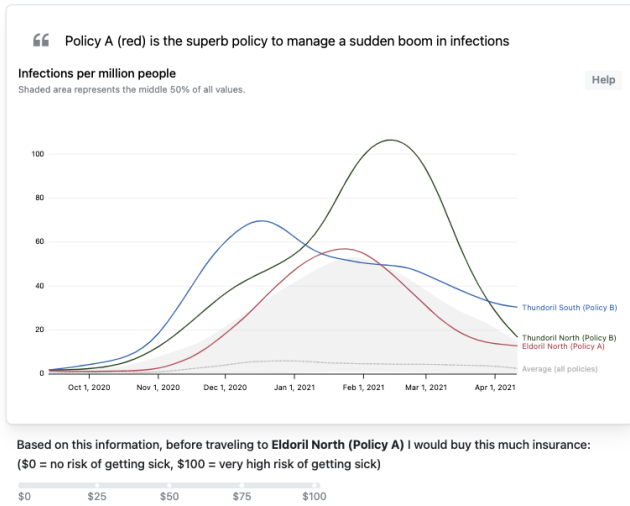


Fig. 8: Screenshot of our prototype implementation of data explorer in the Study 2 experimental setup. Shown is the Viral scenario with a Superimposed Summary guardrail.

7.1 Methods

Before running the main study, we conducted three small-scale pilots on Prolific of five participants each to verify that our description of the scenario and our questionnaire were understandable for the participants. In the first two pilots, we asked the participants to follow a think-aloud protocol and recorded their audio using the reVISit think-aloud feature [15]. After reviewing the results, we made minor changes to the UI of chart and questionnaire elements and clarified some language. The participants' audio was helpful in identifying issues with the user interface but not in prompting the participants to reflect on their responses and analyze the data. For the third pilot, we added a text box in which we asked the participants to reflect on their responses, which yielded richer results. The protocol with a text box rather than audio recording was then used for the main study.

For the main study, we recruited 160 participants on the crowdsourcing platform Prolific who indicated to be fluent in English. The study was reviewed by the University of Utah IRB and deemed exempt from full board review (IRB number 00173128). Each participant was randomly placed into one of the four scenarios from Study 1, and was shown five visualizations in random order: one with each of the guardrails and one control. All five visualizations were randomly drawn from the pool of 427 charts created by the participants of Study 1. This study was also designed and implemented with the reVISit study framework [15]. Detailed instructions and prompts, as well as the whole study with all stimuli, are available in [Appendix C](#) and [online](#).

Participants of Study 2 saw the exact same interface as participants from Study 1, except that the visualization was accompanied by the caption at the top, and the sidebar that permits selection of items was hidden for all conditions except for Juxtaposed Primary Data. The Juxtaposed Primary Data condition without a sidebar would look essentially like the control, therefore we kept the sidebar with checkboxes removed. Figure 8 shows a screenshot of the experimental setup.

In order to accurately infer how much the participants trust a given chart, they were asked a *monetary action* question about every visualization that approximates an investment game [18]. Specifically, in the Stock scenarios they were asked to decide how much they would invest in the promoted industry fund, up to \$100. In the Viral scenarios they were asked how much health insurance they would buy when traveling to the virus-affected areas, up to \$100. The participants then had to describe in 1-2 sentences the rationale behind their choice and answer a series of Likert scale questions, eliciting whether they found the visualization trustworthy, convincing, easy to understand, and whether they would share it with others. The median completion time was 10

minutes, and the participants were compensated \$2.50 for a median hourly rate of \$15/hr.

We analyzed the monetary action and Likert question results within participants and across guardrails quantitatively using repeated measures ANOVA tests, with post-hoc paired T tests. To account for multiple comparisons made, we used the Benjamini-Hochberg procedure [4], a common correction method [43]. We also qualitatively analyzed the data by reviewing the monetary action justification text and post-study feedback text and noting insightful themes.

7.2 Findings

Figure 9 presents the results of the monetary action question by guardrail and scenario. To our surprise, in most cases the guardrails did not have a statistically significant effect of nudging the audience towards skepticism. We additionally observe the relatively poorer performance of Summary guardrails compared to Data guardrails. Participants' text responses reveal that, although we added a one-sentence explanation of what the visual summaries meant, it was difficult for many to understand what was being shown. As a result, the audiences disregarded them: Superimposed Summary results matched Control in three out of four scenarios, while none of Juxtaposed Summary results had an effect. At the same time, very few of the responses to Primary Data guardrails reflected confusion, as the guardrails shared almost the same visual encoding as the main data selections.

We performed repeated measures ANOVA tests to identify whether there are differences in monetary action and Likert question responses between guardrails. To identify specifically which guardrails had an effect, we followed up with post-hoc paired T tests with Benjamini/Hochberg FDR correction comparing the Control condition with each of the guardrails. We found no statistically significant effects in the Likert question responses. In the monetary action results, we found that guardrails had the strongest effect in the Viral A scenario—the only scenario in which ANOVA tests showed a statistically significant difference between the action results by guardrail condition ($F(4, 152) = 7.79, p < 0.001$). In this scenario, participants chose to spend, on average, \$63 in insurance in the Superimposed Primary Data condition, as opposed to \$45 when seeing the control visualizations ($T(38) = -4.28, p = 0.001$). Other post-hoc tests revealed that Superimposed Summary (mean \$55, $T(38) = -1.99, p = 0.076$), and the Juxtaposed Primary Data (mean \$56, $T(38) = 2.72, p = 0.024$) also had a statistically significant effect compared to the Control condition, where the average insurance purchase was \$45.

In their text responses, many participants explicitly reference the guardrail and discuss that seeing it had an effect on their decision. One participant in the Superimposed Primary Data condition noted: “[The chart] shows [option A] to be the worst one out of the lines shown.” Similarly, a participant who saw the Viral A scenario with a Superimposed Summary said that “[the infection rate] is still considerably higher than the average and I do not think it is worth taking any risks.” A participant mentioned referencing “...the comparatives at the side to visualise how it compares to other infected places” in Juxtaposed Primary Data. We note that although some of the guardrails did not have a strong effect on the results across conditions and at the population level, they did impact the decisions of some of the participants. Specifically, we observe that a subset of participants in each of the experimental conditions similarly referred to the guardrails in their rationales.

The differences in effect sizes between scenarios additionally suggest to us that the interplay between guardrails, the egregiousness of the cherry-picking attempt, and the specific shape and magnitude of the underlying data plays an important role. As previously discussed in Section 6.2, when confronted with a revealing guardrail, the producers of cherry-picked visualizations often shifted the focus of their argument from the magnitude of values to a visualization shape- or domain-specific “redeeming factors”: a recent change in trend or the overall variance of the trend. This is an artifact of the complexity of real-world data used in our experimental setup and we hypothesize that a study with artificial data that controls for the nuances of the trends could result in stronger effects of guardrails.

It is also important to highlight that, despite a number of participants

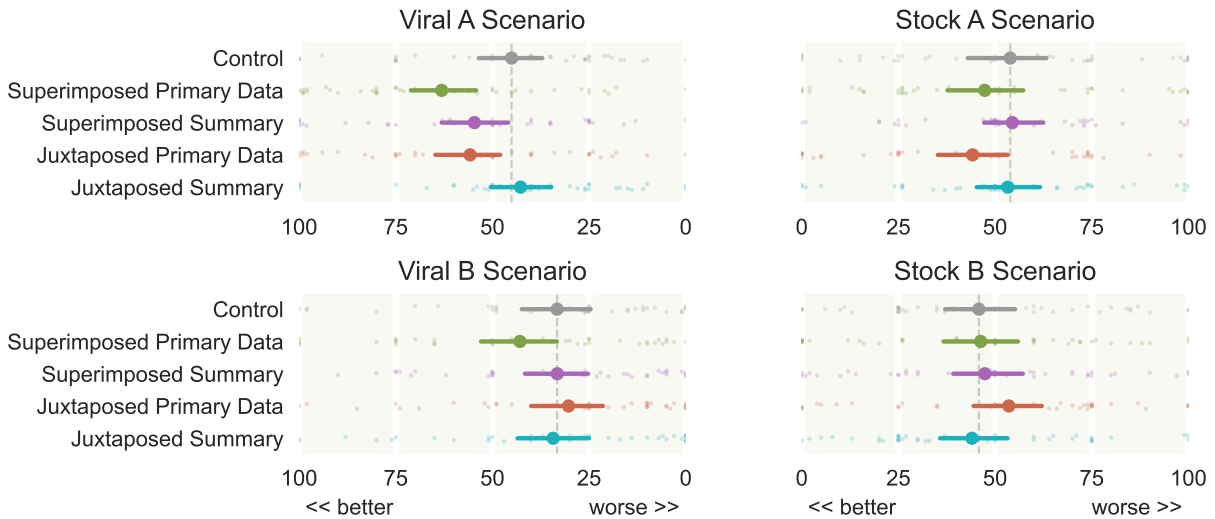


Fig. 9: Results of the monetary action question in Study 2. Shown are average values and bootstrapped 95% CI ($n = 1000$) and individual data points in the background. Note the inverse scale between the Viral and Stock scenarios: the monetary action in the Viral scenarios involves making a decision about insurance purchase (less insurance = trusting cherry-picking), while in the Stock scenarios it involves making an investment (more investment = trusting cherry-picking).

directly citing guardrails in their decision to resist the cherry-picked argument, after we revealed the guardrails at the end of the experiment, 38% of participants reported having not noticed any guardrails in place, with another 23% having noticed but not understood what they were supposed to mean. Taken together, these results suggest that one of the pitfalls of adding guardrails is that the audience may simply not notice or understand them. And even of the participants who did, some still preferred to not pay attention to the guardrails and focus on the main salient features of the visualization. One participant explained: “I just followed the figures and how they were either rising or falling”, with others echoing that they tried to understand the main data at “face value” instead of focusing on the guardrails.

8 DISCUSSION & DESIGN RECOMMENDATIONS

Based on our work, we find evidence that our guardrail designs have the potential to lessen the problem of cherry-picking in data explorers, however they come with limitations and important considerations. In this section, we present a discussion of our findings and provide design recommendations for effective use of guardrails.

Most of the guardrail designs against cherry-picking involve providing a definition of “contextual information”: whether its demographically-comparable subsets or countries with similar climate. Correctly defining context would thus require a careful examination of the domain at hand and consulting a domain expert. An **evaluation of the domain as well as existing misinformation** on the topic would also help understand which type of cherry-picking should be targeted. For instance, while misinformation about climate change typically involves time frame cherry-picking [10], COVID-19 conspiracy theories are most often based on item cherry-picking of countries [32].

In selecting the context to use as guardrail, we find that Primary Data guardrails were much easier for the audiences to notice and understand. Therefore, we recommend to **prioritize guardrails that are as close in visual encoding to the main data as possible**, and utilize Primary Data guardrails as default. While certain domains may benefit from providing statistical summaries to deter cherry-picking, **Summary guardrails should always be accompanied by a tutorial or detailed annotations** to help users understand them and take them into account.

Our results also indicate that, in terms of layout, Superimposed guardrails both make it more difficult to create cherry-picked views, and that such views are more often not as convincing to the audiences. At the same time, the Juxtaposed guardrails provide limited protection against the production of cherry-picked views, and sometimes can even be helpful in exploring the data. However, we did not find visualizations created with Juxtaposed guardrails to be more misleading. Juxtaposed and Superimposed guardrails are also easily combined, and therefore

we would recommend **introducing Superimposed guardrails to prevent cherry-picking, and using Juxtaposed guardrails primarily as a way to make data explorers more easy to use.**

Importantly, we also found evidence that cherry-picking is an *adversarial process*. Users adjust to successful guardrails and seek out other ways to successfully cherry-pick: our previous work shows that people often crop visualization screenshots [32], and experiments in this paper show that authors shift focus onto other features of data. Consequently, it is crucial to **continuously adapt to ever-evolving cherry-picking strategies** and the types of misinformation that are currently popular, and update the decisions on what information can provide relevant context. An analysis of the data explorer usage log could provide useful insights about what subsets of data users focus on and additionally guide guardrail implementation.

9 CONCLUSION & FUTURE WORK

In this paper, we describe an approach to designing technical interventions against the misuse of data visualizations in support of misinformation. We examine cherry-picking in visualization through the lens of threat modeling and describe the design space of *guardrails*: design interventions against such deceptive tactics. We are hopeful that data exploration platforms adopt similar interventions in their designs. This would allow future work to examine the role that guardrails would play in complex real world contexts. As we have discussed, the misinformation using data visualizations is adversarial and a wicked problem. Because of this, studying real world adoption of guardrails would be especially important: while the results of our experiments show moderate effects of guardrails encouraging skepticism, it is challenging to predict the exact effects of guardrails on online data discourse.

We hypothesize that although guardrails may not always influence an individual in isolation, they could have indirect effects in a world where the general public is familiar with guardrails. For instance, guardrails could provide evidence that triggers a fact-checking discussion on social media, while guardrails cropped out of a screenshot could alert the audience of tampering attempts. The use of guardrails against misinformation could eventually be associated with reputable sources and serve as a trustworthiness indicator in and of itself [31]. Aside from adapting and evolving their tactics, it is also likely that malicious actors would migrate to other, non-guardrailed platforms, and create a demand for “alternative” data exploration sites.

In addition to studying the effects of guardrail adoption, future work should also examine strategies to make guardrails—in particular Juxtaposed designs—less amenable to cropping, as well as the effects of combining multiple forms of guardrails.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation (IIS 2041136, IIS 1751238, and CNS 2213756).

SUPPLEMENTAL MATERIALS

All supplemental materials are available at <https://osf.io/fcwkv/>.

REFERENCES

- [1] C. Andrade. HARKing, Cherry-Picking, P-Hacking, Fishing Expeditions, and Data Dredging and Mining as Questionable Research Practices. *The Journal of Clinical Psychiatry*, 82(1):25941, Feb. 2021. doi: 10.4088/JCP.20f13804 3
- [2] A. Asudeh, H. V. Jagadish, Y. W. Wu, and C. Yu. On detecting cherry-picked trendlines. *Proceedings of the VLDB Endowment*, 13(6):939–952, Feb. 2020. doi: 10.14778/3380750.3380762 3
- [3] A. Barchetti, E. Neybert, S. P. Mantel, and F. R. Kardes. The Half-Truth Effect and Its Implications for Sustainability. *Sustainability*, 14(11):6943, June 2022. doi: 10.3390/su14116943 3
- [4] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x 8
- [5] Binance. CoinMarketCap. Binance, 2013. 5
- [6] J. Bollinger. *Bollinger on Bollinger Bands*. McGraw-Hill, New York, 2002. 5
- [7] Z. Braiterman, A. Shostack, J. Marcil, S. de Vries, I. Michlin, K. Wuyts, R. Hurlbut, B. S. Schoenfeld, F. Scott, M. Coles, C. Romeo, A. Miller, I. Tarandach, A. Douglan, and M. French. Threat Modeling Manifesto. <https://www.threatmodelingmanifesto.org/>, 2020. 2
- [8] F. Büttner, E. Toomey, S. McClean, M. Roe, and E. Delahunt. Are questionable research practices facilitating new discoveries in sport and exercise medicine? The proportion of supported hypotheses is implausibly high. *British Journal of Sports Medicine*, 54(22):1365–1371, Nov. 2020. doi: 10.1136/bjsports-2019-101863 3
- [9] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers Inc., 1999. 5
- [10] J. Cook, P. Ellerton, and D. Kinkead. Deconstructing climate misinformation to identify reasoning errors. *Environmental Research Letters*, 13(2):024018, Feb. 2018. doi: 10.1088/1748-9326/aaa49f 3, 9
- [11] M. Correll. Teru Teru Bōzu: Defensive Raincloud Plots. *Computer Graphics Forum*, 42(3):235–246, 2023. doi: 10.1111/cgf.14826 5
- [12] M. Correll, E. Bertini, and S. Franconeri. Truncating the Y-Axis: Threat or Menace? In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–12. ACM, Honolulu HI USA, Apr. 2020. doi: 10.1145/3313831.3376222 3
- [13] E. Cuenca, A. Sallaberry, F. Y. Wang, and P. Poncet. MultiStream: A Multiresolution Streamgraph Approach to Explore Hierarchical Time Series. *IEEE Transactions on Visualization and Computer Graphics*, 24(12):3160–3173, Dec. 2018. doi: 10.1109/TVCG.2018.2796591 5
- [14] Z. T. Cutler, K. Gadhave, and A. Lex. Ttrack: A Library for Provenance Tracking in Web-Based Visualizations. In *IEEE Visualization Conference (VIS)*, pp. 116–120, 2020. doi: 10.1109/VIS47514.2020.00030 6
- [15] Y. Ding, J. Wilburn, H. Shrestha, A. Ndlovu, K. Gadhave, C. Nobre, A. Lex, and L. Harrison. reVISit: Supporting Scalable Evaluation of Interactive Visualizations. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 31–35, Oct. 2023. doi: 10.1109/VIS54172.2023.00015 6, 8
- [16] S. P. Dow, A. Glassco, J. Kass, M. Schwarz, D. L. Schwartz, and S. R. Klemmer. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction*, 17(4):18:1–18:24, Dec. 2011. doi: 10.1145/1879831.1879836 3
- [17] È. Dubé, J. K. Ward, P. Verger, and N. E. MacDonald. Vaccine Hesitancy, Acceptance, and Anti-Vaccination: Trends and Future Prospects for Public Health. *Annual Review of Public Health*, 42(1):175–191, Apr. 2021. doi: 10.1146/annurev-publhealth-090419-102240 3
- [18] H. Elhamedi, A. Gaba, Y.-S. Kim, and C. Xiong. How Do We Measure Trust in Visual Data Communication? In *2022 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 85–92. IEEE, Oklahoma City, OK, USA, Oct. 2022. doi: 10.1109/BELIV57783.2022.00014 8
- [19] Federal Reserve Bank of St. Louis. FRED, Federal reserve economic data. Federal Reserve Bank Of St. Louis., 1997. 1
- [20] L. W. Ge, Y. Cui, and M. Kay. CALVI: Critical Thinking Assessment for Literacy in Visualizations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18. ACM, Hamburg Germany, Apr. 2023. doi: 10.1145/3544548.3581406 3
- [21] L. W. Ge, M. Easterday, M. Kay, E. Dimara, P. Cheng, and S. L. Franconeri. V-FRAMER: Visualization Framework for Mitigating Reasoning Errors in Public Policy. *Proceedings of CHI 2024, ACM Conference on Human Factors in Computing Systems*, 2024. doi: 10.1145/3613904.3642750 3
- [22] M. Gleicher. Considerations for Visualizing Comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, 2018. doi: 10.1109/TVCG.2017.2744199 4
- [23] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011. doi: 10.1177/1473871611416549 4
- [24] S. G. Hart and L. E. Staveland. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock and N. Meshkati, eds., *Advances in Psychology*, vol. 52 of *Human Mental Workload*, pp. 139–183. North-Holland, Jan. 1988. doi: 10.1016/S0166-4115(08)62386-9 6, 7
- [25] J. Heer, N. Kong, and M. Agrawala. Sizing the Horizon: The Effects of Chart Size and Layering on the Graphical Perception of Time Series Visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*, pp. 1303–1312. ACM, 2009. doi: 10.1145/1518701.1518897 5
- [26] M. Hoofnagle and C. J. Hoofnagle. What is Denialism? *SSRN Electronic Journal*, 2007. doi: 10.2139/ssrn.4002823 3
- [27] W. Javed and N. Elmquist. Stack zooming for multi-focus interaction in time-series data visualization. In *Proceedings of the IEEE Symposium on Pacific Visualization (PacificVis '10)*, pp. 33–40. IEEE, 2010. doi: 10.1109/PACIFICVIS.2010.5429613 5
- [28] P.-M. Law, L. Y.-H. Lo, A. Ender, J. Stasko, and H. Qu. Causal Perception in Question-Answering Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–15. ACM, Yokohama Japan, May 2021. doi: 10.1145/3411764.3445444 3
- [29] Y. Lin, B. Youngmann, Y. Moskovitch, H. V. Jagadish, and T. Milo. On detecting cherry-picked generalizations. *Proceedings of the VLDB Endowment*, 15(1):59–71, Sept. 2021. doi: 10.14778/3485450.3485457 3
- [30] Y. Lin, B. Youngmann, Y. Moskovitch, H. V. Jagadish, and T. Milo. OREO: Detection of cherry-picked generalizations. *Proceedings of the VLDB Endowment*, 15(12):3570–3573, Aug. 2022. doi: 10.14778/3554821.3554846 3
- [31] M. Lisnic, A. Lex, and M. Kogan. "Yeah, this graph doesn't show that": Analysis of Online Engagement with Misleading Data Visualizations. Preprint, Open Science Framework, Sept. 2023. doi: 10.31219/osf.io/q6j3u 2, 9
- [32] M. Lisnic, C. Polychronis, A. Lex, and M. Kogan. Misleading Beyond Visual Tricks: How People Actually Lie with Charts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21. ACM, Hamburg Germany, 2023. doi: 10.1145/3544548.3580910 1, 2, 3, 4, 9
- [33] L. Y.-H. Lo, A. Gupta, K. Shigyo, A. Wu, E. Bertini, and H. Qu. Misinformed by Visualization: What Do We Learn From Misinformative Visualizations? *Computer Graphics Forum*, 41(3):515–525, 2022. doi: 10.1111/cgf.14559 3
- [34] E. Mathieu, H. Ritchie, L. Rodés-Guirao, C. Appel, C. Giatino, J. Hasell, B. Macdonald, S. Dattani, D. Beltekian, E. Ortiz-Ospina, and M. Roser. Coronavirus Pandemic (COVID-19). <https://ourworldindata.org/coronavirus>, 2020. 1, 2, 3, 5
- [35] M. Mayer and R. J. Evans. *They Thought They Were Free: The Germans, 1933-45*. The University of Chicago Press, Chicago, 2017. 3
- [36] D. Moritz and D. Fisher. Visualizing a Million Time Series with the Density Line Chart. *arXiv:1808.06019 [cs]*, Aug. 2018. 5
- [37] B. Morrow, T. Manz, A. E. Chung, N. Gehlenborg, and D. Gotz. Periphery Plots for Contextualizing Heterogeneous Time-Based Charts. In *2019 IEEE Visualization Conference (VIS)*, pp. 1–5, Oct. 2019. doi: 10.1109/VISUAL.2019.8933582 4, 5
- [38] S. Plous. *The Psychology of Judgment and Decision Making*. McGraw-Hill Series in Social Psychology. McGraw-Hill Higher Education, New York, 2007. 3
- [39] X. Pu and M. Kay. The Garden of Forking Paths in Visualization: A Design Space for Reliable Exploratory Visual Analytics : Position Pa-

- per. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 37–45, 2018. doi: [10.1109/BELIV.2018.8634103](https://doi.org/10.1109/BELIV.2018.8634103) 3
- [40] L. Reina. Why movie blurbs avoid newspapers.. <https://www.thefreelibrary.com/Why+movie+blurbs+avoid+newspapers-a018662749>, 1996. 3
- [41] H. W. J. Rittel and M. M. Webber. Dilemmas in a general theory of planning. *Policy Sciences*, 4(2):155–169, June 1973. doi: [10.1007/BF01405730](https://doi.org/10.1007/BF01405730) 1
- [42] H. Rosling and Z. Zhang. Health advocacy with Gapminder animated statistics. *Journal of Epidemiology and Global Health*, 1(1):11, 2011. doi: [10.1016/j.jegh.2011.07.001](https://doi.org/10.1016/j.jegh.2011.07.001) 1, 5
- [43] A. Sarma, X. Pu, Y. Cui, E. T. Brown, M. Correll, and M. Kay. Odds and Insights: Decision Quality in Exploratory Data Analysis Under Uncertainty, Jan. 2024. doi: [10.31219/osf.io/qa8dj](https://doi.org/10.31219/osf.io/qa8dj) 8
- [44] A. Shostack. *Threat Modeling: Designing for Security*. Wiley, Indianapolis, IN, 2014. 2, 3
- [45] R. Somerville. Climate Science and EPA's Greenhouse Gas Regulations. Testimony before the US House of Representatives Committee on Energy and Commerce Subcommittee on Energy and Power. <https://www.govinfo.gov/content/pkg/CHRG-112hrg66704/html/CHRG-112hrg66704.htm>, 2011. 3
- [46] E. Tuft. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 2nd ed., 2001. 5
- [47] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, Sept. 1973. doi: [10.1016/0010-0285\(73\)90033-9](https://doi.org/10.1016/0010-0285(73)90033-9) 3
- [48] A. Tversky and D. Kahneman. *Judgment under Uncertainty: Heuristics and Biases*, vol. 185. Cambridge University Press, 1974. 3
- [49] E. Wall, L. M. Blaha, L. Franklin, and A. Endert. Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 104–115. IEEE, Phoenix, AZ, Oct. 2017. doi: [10.1109/VAST.2017.8585669](https://doi.org/10.1109/VAST.2017.8585669) 3
- [50] E. Wall, A. Narechania, A. Coscia, J. Paden, and A. Endert. Left, Right, and Gender: Exploring Interaction Traces to Mitigate Human Biases. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):966–975, Jan. 2022. doi: [10.1109/TVCG.2021.3114862](https://doi.org/10.1109/TVCG.2021.3114862) 3
- [51] R. T. Whitaker, M. Mirzargar, and R. M. Kirby. Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2713–2722, Dec. 2013. doi: [10.1109/TVCG.2013.143](https://doi.org/10.1109/TVCG.2013.143) 5
- [52] W. Willett, J. Heer, and M. Agrawala. Scented Widgets: Improving Navigation Cues with Embedded Visualizations. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '07)*, 13(6):1129–1136, 2007. doi: [10.1109/TVCG.2007.70589](https://doi.org/10.1109/TVCG.2007.70589) 5
- [53] Yahoo! Yahoo! Finance. Yahoo! Inc., 1997. 1, 3, 5
- [54] E. Zraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '18)*, p. 479. ACM, 2018. doi: [10.1145/3173574.3174053](https://doi.org/10.1145/3173574.3174053) 3
- [55] C. Zheng and X. Ma. Evaluating the Effect of Enhanced Text-Visualization Integration on Combating Misinformation in Data Story. In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*, pp. 141–150. IEEE, Tsukuba, Japan, Apr. 2022. doi: [10.1109/PacificVis53943.2022.00023](https://doi.org/10.1109/PacificVis53943.2022.00023) 3
- [56] Q. Zhu, L. Y.-H. Lo, M. Xia, Z. Chen, and X. Ma. Bias-Aware Design for Informed Decisions: Raising Awareness of Self-Selection Bias in User Ratings and Reviews. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–31, Nov. 2022. doi: [10.1145/3555597](https://doi.org/10.1145/3555597) 3

A DESIGN SKETCHES

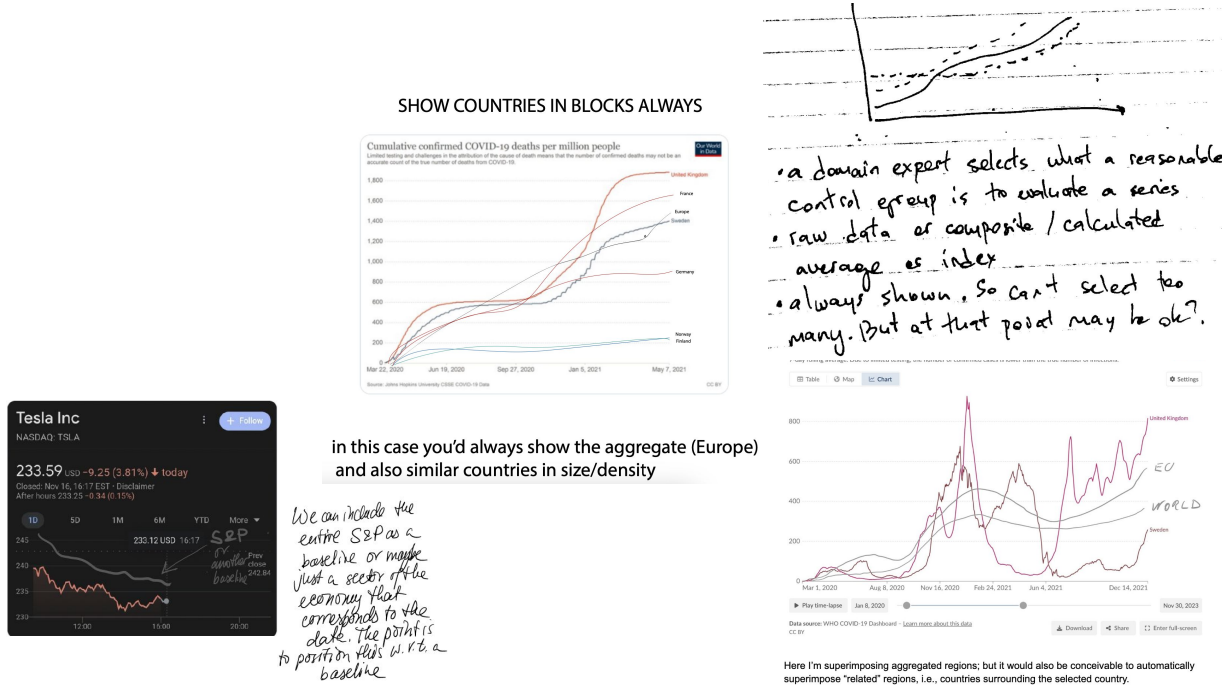


Fig. 10: Sketches of initial designs. Later categorized as Superimposed variations.

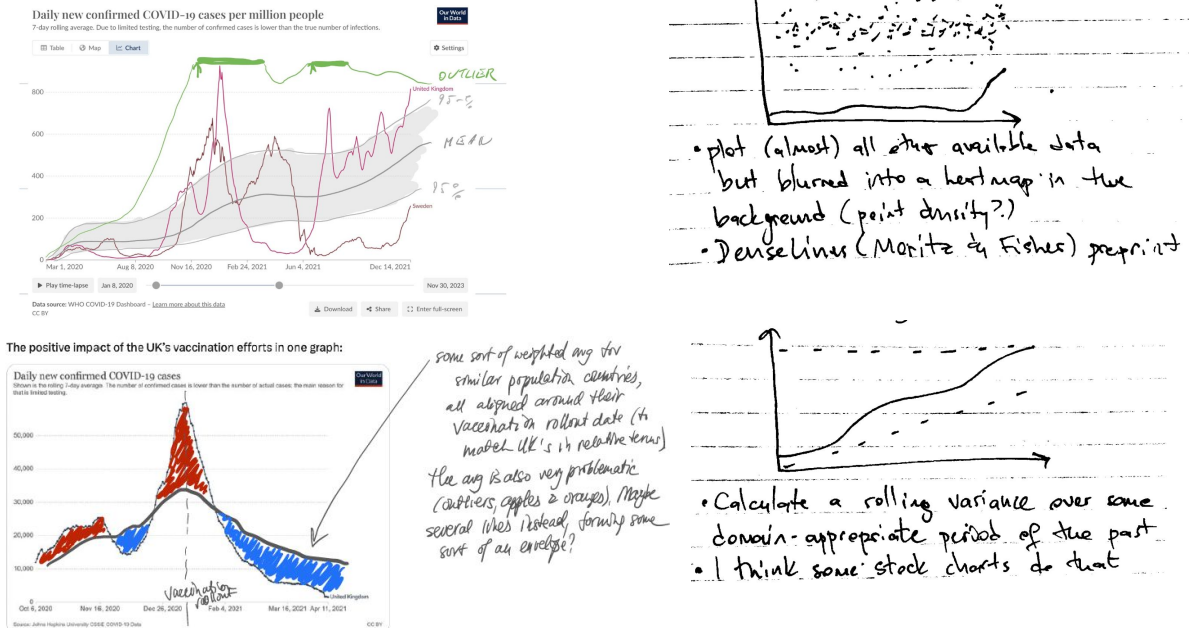


Fig. 11: Sketches of initial designs. Later categorized as Superimposed variations.

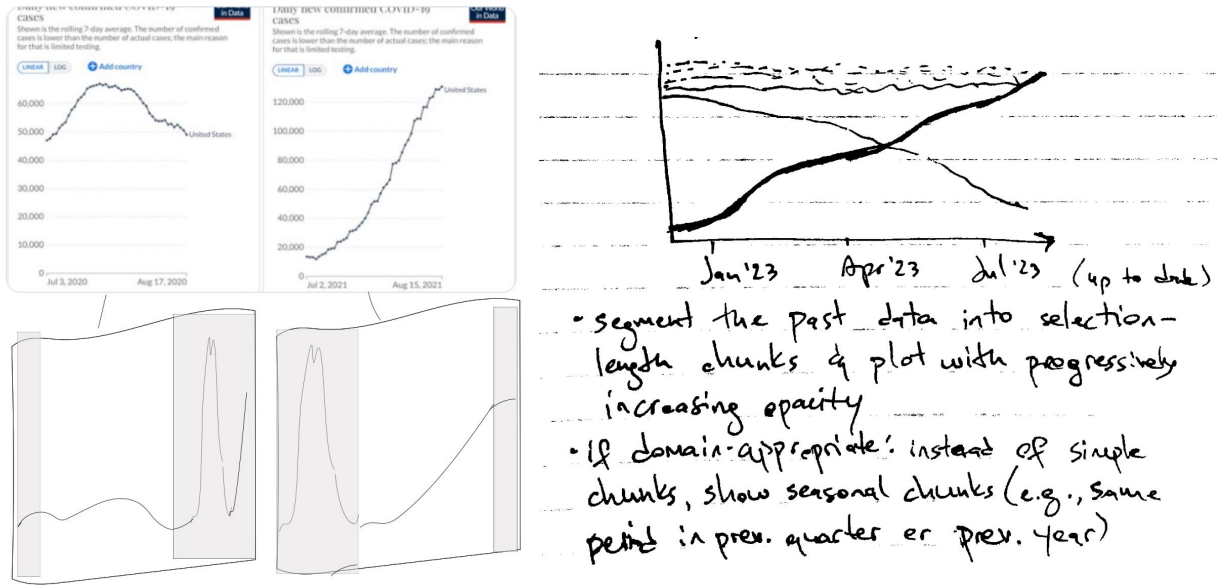


Fig. 12: Sketches of initial designs. Later categorized as Juxtaposed variations.

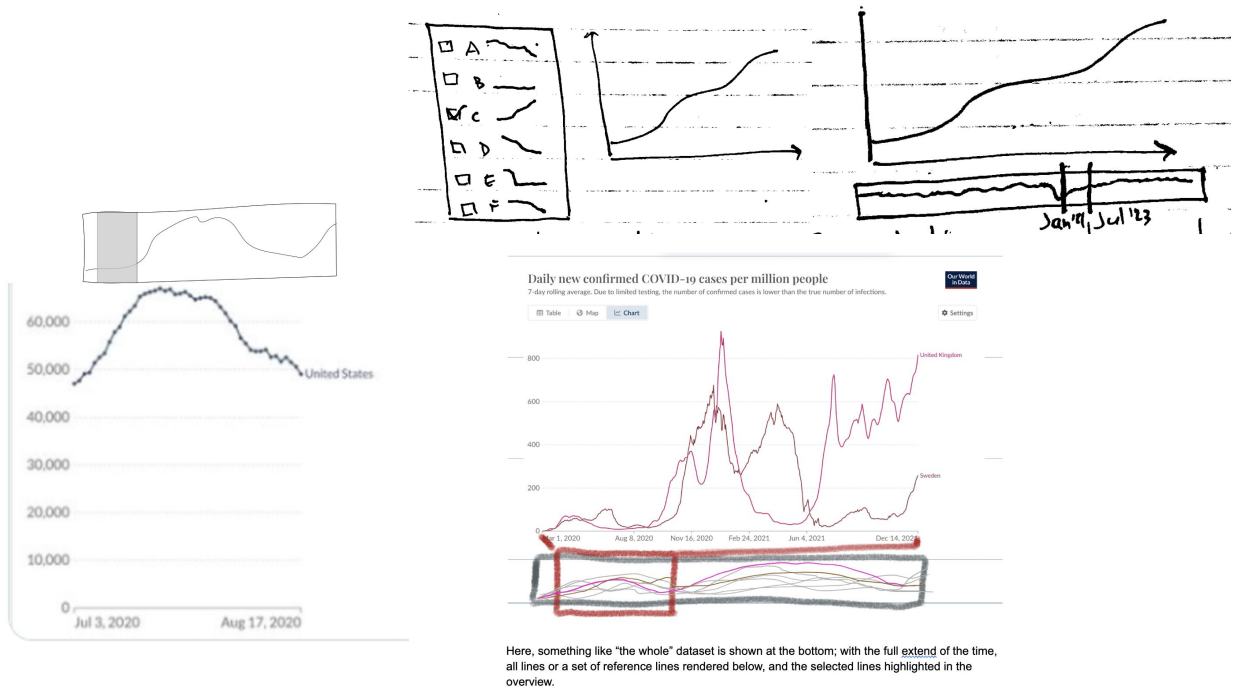
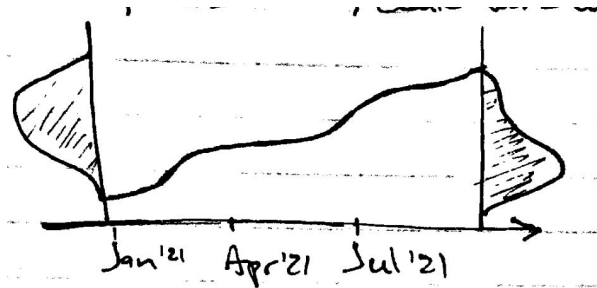
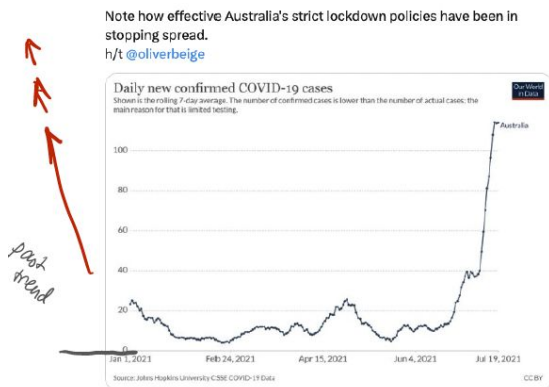
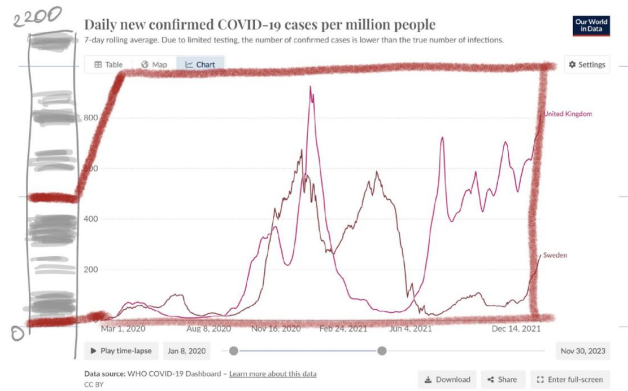


Fig. 13: Sketches of initial designs. Later categorized as Juxtaposed variations.



xtaping a heat map that samples all data items and all time points, to give a sense of where the scale we are. This deals with both item and time cherry picking.



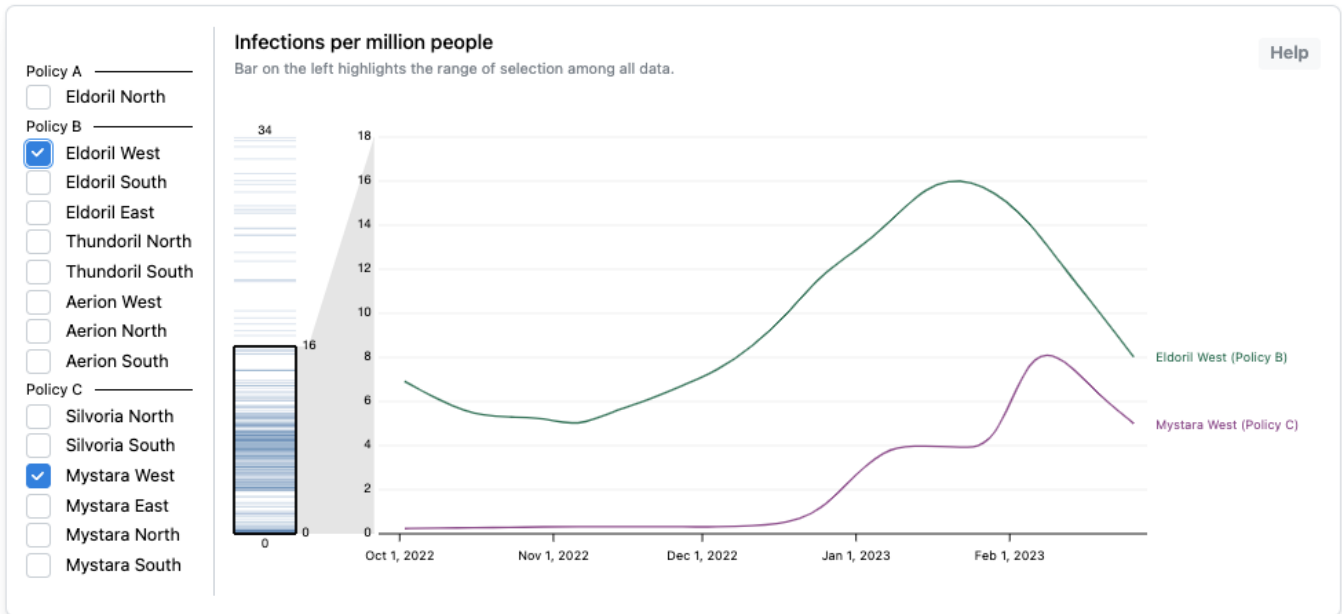
This heat map could also be a histogram, and it could be two separate histograms; for the items in the dataset, and for reference items.

These could be on the left, right, or both.

Fig. 14: Sketches of initial designs. Later categorized as Juxtaposed variations.

B PROTOTYPE

Select a view that best shows (and convinces the population) that **Policy C is the best policy to combat the disease.**



Add a short text caption that will go along with your visualization: *

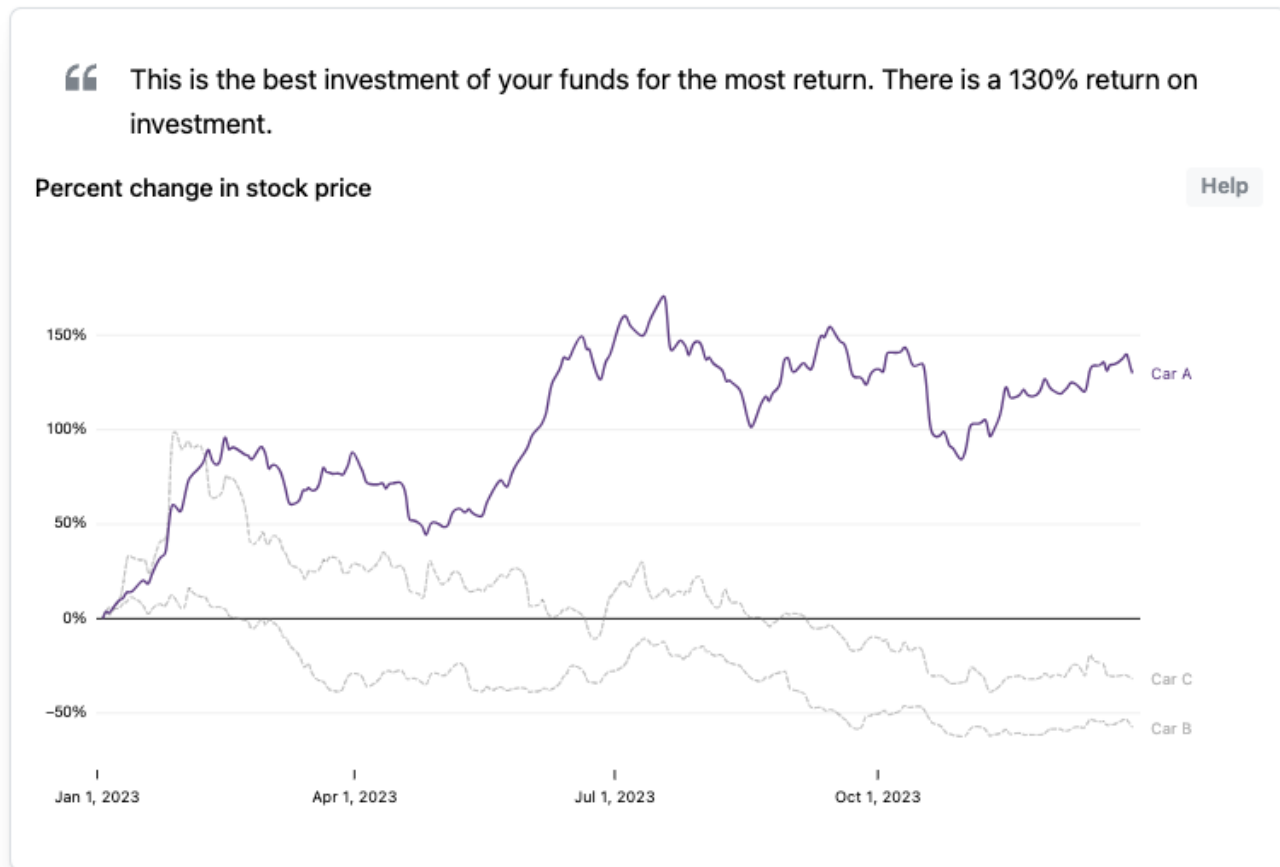
Next

Fig. 15: Screenshot of our prototype implementation of data explorer in the Study 1 experimental setup. Shown is the Viral scenario with a Juxtaposed Summary guardrail.

You've come across this visualization and the accompanying caption.

Please review the visualization and the caption, and then answer the questions below **based solely on this information**.

(Please try to not rely on other visualizations you've seen.)



Based on this information, I would invest this much in the combined **Car** industry fund:

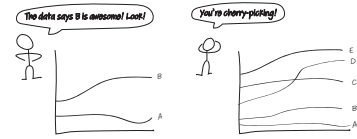
(\$0 = no chance of growth, \$100 = very high chance of growth) *



Fig. 16: Screenshot of our prototype implementation of data explorer in the Study 2 experimental setup. Shown is the Stocks scenario with a Superimposed Primary Data guardrail.

C EXPERIMENTAL SETUP

I MOTIVATION

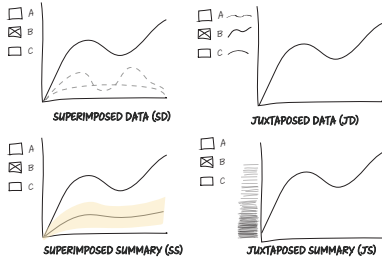


THE QUESTIONS ARE...

IS IT MORE DIFFICULT TO CREATE MISLEADING VISUALIZATIONS VIA CHERRY-PICKING WITH GUARDRAILS?
 ARE PEOPLE MORE SKEPTICAL OF CHERRY PICKED VISUALIZATIONS WITH GUARDRAILS?

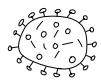
II TASKS AND CONDITIONS

- GUARDRAILS



- DATASETS

COVID



STOCKS



- SCENARIOS

VIRAL-A

UNREASONABLE
HIDE THE COMPETITION

STOCKS-A

UNREASONABLE
HIDE THE BAD APPLE

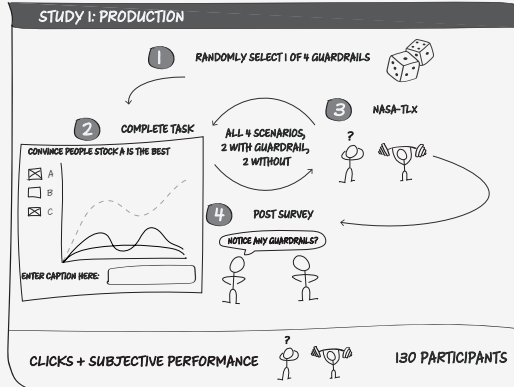
VIRAL-B

REASONABLE
HIDE THE BAD APPLE

STOCKS-B

REASONABLE
HIDE THE COMPETITION

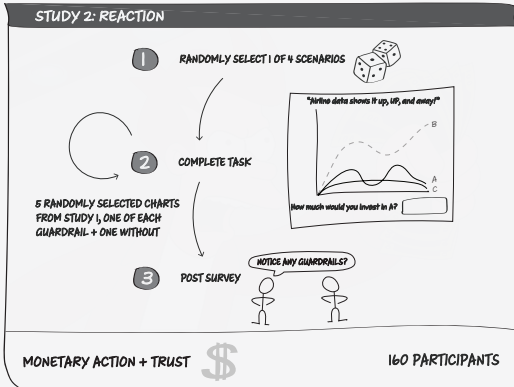
III STUDY DESIGN



AUTHORS REVIEWED 520 STUDY 1 CHARTS/CAPTIONS TO DETERMINE WHICH COULD BE USED IN STUDY 2



DETERMINED THAT 427 / 520 ARE USEFUL FOR STUDY 2.



IV ANALYSIS & RESULTS

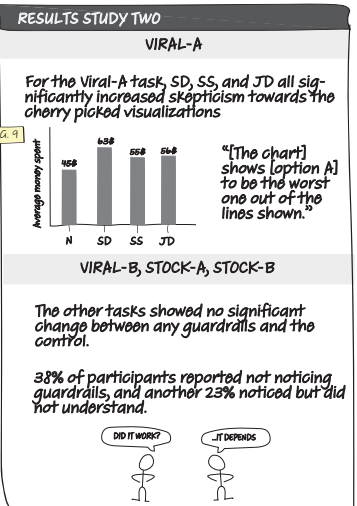
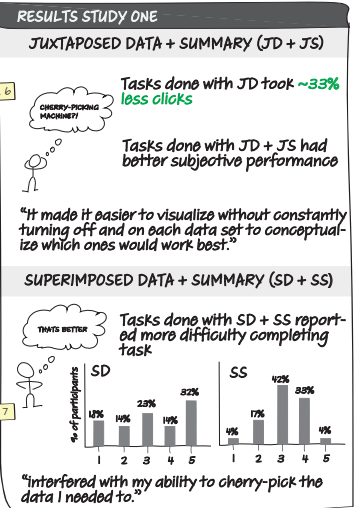


Fig. 17: Illustration of study tasks and conditions, design, analysis and results.

Instructions for the "Financial Advisor" Scenario

This scenario consists of **2 tasks**. If you need to refer to these instructions during the experiment, please click on help in the top right corner of the visualization.

Scenario: You are a financial advisor. Your client approached you asking for help in picking a new investment—an industry fund that equally invests within a single industry. The client prefers to make their decisions *solely based on the data*, and not based on any inside knowledge about the type of industry. However, your boss *does* have inside knowledge and orders you to recommend a specific industry (which is ultimately in the best interest of the client). You cannot disclose this to the client and have to use data to support your recommendations.

Task: In the experiment, you will use an interactive data explorer that shows performance of different stocks from a variety of industries. You should select a view that best shows (and convinces your client) that **the chosen industry fund would be the best investment with the highest returns**. After finalizing the view, add a caption or a slogan that will go along with your visualization.

Next

Fig. 18: Screenshot of Study 1 Stocks scenario introduction.

Instructions for the "Celestial Sniffles" Scenario

This scenario consists of **2 tasks**. If you need to refer to these instructions during the experiment, please click on help in the top right corner of the visualization.

Background: You live in a fantasy world that consists of your home country, as well as other 15 countries located across 5 continents: Eldoril, Thundoril, Aerion, Silvorla, and Mystara. Countries within each continent are similar to each other in size and demographics.

Scenario: You are an analyst for the Viral Disease Policy Center of your home country at time when there's a new viral disease called *Celestial Sniffles*. Luckily, your country has not been affected yet and has time to prepare and decide on a policy to combat it. You have access to the infection data from other countries that implemented one of the 3 available policies: A, B, or C. However, due to socioeconomic factors, the Surgeon General in your country has *already decided on a policy*.

Task: You are tasked with leading the promotion efforts that make the case for the chosen policy. In the experiment, you will use an interactive data explorer that shows infection data from other countries. You should select a view that best shows (and convinces the population) that your country's **chosen policy is the best policy to combat the disease**. After finalizing the view, add a caption or a slogan that will go along with your visualization.



Next

Fig. 19: Screenshot of Study 1 Viral scenario introduction.

Questions

Please answer the questions below about your experience completing the tasks *with* and *without* the guardrail enabled.

How familiar are you with data visualizations? *

Not familiar 1 2 3 4 5 Very familiar

Did you notice a guardrail in place in some of your tasks? *

No Yes

How clear was it what the guardrail was showing? *

Not clear 1 2 3 4 5 Very clear

Did the guardrail make it easier or more difficult to convincingly support the prompts? *

Easier 1 2 3 4 5 More difficult

Describe how the guardrail impacted your experience performing the task: *

Enter your answer here.

Did you experience any problems with our tool? *

Enter your answer here.

Did anything not render or display properly? *

Enter your answer here

Any other issues or anything you would like to tell us? *

Enter your answer here

Next

Fig. 20: Screenshot of Study 1 post-study survey.

Scenario description

Please read the scenario description below carefully, as it explains the setting for the visualizations you will see. If you need to refer to these instructions during the experiment, please click on help in the top right corner of the visualization.

Scenario: You have \$100 that you would like to invest in *industry funds* but are not sure how to best allocate that money. An industry fund is composed of all individual stocks pertaining to that industry (e.g., a Pharma fund is composed of Pharma A, Pharma B, and Pharma C stocks). Since you are not familiar with the current market, you turn to social media to read what people shared about recent stock performance.

Task: In the experiment, you will see a set of visualizations that show performance of different stocks from a variety of industries. Based on each visualization, you will be asked to decide **how much to invest in a given industry fund** and **answer a short survey**.

Next

Fig. 21: Screenshot of Study 2 Stocks scenario introduction.

Scenario description

Please read the scenario description below carefully, as it explains the setting for the visualizations you will see. If you need to refer to these instructions during the experiment, please click on help in the top right corner of the visualization.

Background: You live in a fantasy world that consists of your home country, as well as other 15 countries located across 5 continents: Eldoril, Thundoril, Aerion, Silvoria, and Mystara. Countries within each continent are similar to each other in size and demographics.

Scenario: There's a new viral disease called *Celestial Sniffles*. Most countries in the world have adopted one of 3 major disease containment policies: A, B, or C. Luckily, your country has not been affected yet but you may need to travel to one of the affected countries for work. Afraid of getting sick and stuck abroad with a huge hospital bill, you decide to purchase a traveler's health insurance policy. The price of the policy is very flexible and depends on the amount of coverage you'll get: anywhere between \$0 (no coverage) to \$100 (full coverage).

Since you are not familiar with the disease trends or containment policies, you turn to social media to read what people shared about recent infection rates.

Task: In the experiment, you will see a set of visualizations that show infection rates in different countries that adopted one of the containment policies A, B, or C. Based on each visualization, you will be asked to decide **how much to spend on your insurance policy** and **answer a short survey**.



Next

Fig. 22: Screenshot of Study 2 Viral scenario introduction.

Questions

Please answer the questions below about your experience completing the tasks:

How familiar are you with data visualizations? *

Not familiar 1 2 3 4 5 Very familiar

Did you notice a guardrail in place in some of your tasks? *

No Yes

How clear was it what the guardrails were showing? *

Not clear 1 2 3 4 5 Very clear

Did the guardrail make the visualizations more or less convincing? *

Less convincing 1 2 3 4 5 More convincing

Describe how the guardrails impacted your evaluation of visualizations. (If possible, refer to specific guardrails):

*

Enter your answer here.

Did you experience any problems with our tool? *

Enter your answer here.

Did anything not render or display properly? *

Enter your answer here

Any other issues or anything you would like to tell us? *

Enter your answer here

Next

Fig. 23: Screenshot of Study 2 post-study survey.