
DATA HUNCHES: INCORPORATING PERSONAL KNOWLEDGE INTO VISUALIZATIONS

Haihan Lin*
University of Utah

Derya Akbaba*
University of Utah

Miriah Meyer
Linköping University

Alexander Lex
University of Utah
alex@sci.utah.edu

ABSTRACT

The trouble with data is that often it provides only an imperfect representation of the phenomenon of interest. When reading and interpreting data, personal knowledge about the data plays an important role. Data visualization, however, has neither a concept defining personal knowledge about datasets, nor the methods or tools to robustly integrate them into an analysis process, thus hampering analysts' ability to express their personal knowledge about datasets, and others to learn from such knowledge. In this work, we define such personal knowledge about datasets as *data hunches* and elevate this knowledge to another form of data that can be externalized, visualized, and used for collaboration. We establish the implications of data hunches and provide a design space for externalizing and communicating data hunches through visualization techniques. We envision such a design space will empower users to externalize their personal knowledge and support the ability to learn from others' data hunches.

Keywords Data Visualization · Design Space · Situated Knowledge

1 Introduction

Data-driven decision-making and reasoning is now considered the gold standard for businesses [1], sports [2], and scientists [3]. In these contexts, data is often considered complete, objective, neutral, and transparent. In practice, however, uncritical reliance on data alone can lead to poor decisions and outcomes. Good decision makers, including business leaders, team managers, and scientists, have a deep understanding of the data they are analyzing and know its limitations. Critical scholars argue, however, that visualization researchers often assume that the data is perfect and fail to consider any nuance that may exist in the data [4, 5, 6, 7, 8], leaving the task of evaluating and judging datasets and visualizations to the user alone.

We encountered the challenge of visualizing imperfect data in one of our own design studies that we conducted in col-

laboration with clinicians, who analyze data about blood transfusions to improve patient outcomes and limit the use of valuable resources [ref removed]. When we collected feedback on a prototype visualization tool, our collaborators expressed concern about the data: that the amount of recycled blood — a patient's own blood that is re-used during surgery — captured in the data was much lower than they expected. Based on their experience, almost all surgeries make extensive use of blood recycling, and they had a hunch that the low blood recycling values were due to the data frequently not being recorded appropriately in the electronic health record system. Our collaborators worried that when other clinicians saw the same discrepancies, they would lose trust in the visualization and the data, and thus curb their willingness to take a more data-driven approach in their work. In this example, an expert had specific knowledge about imperfections in the dataset, and the expert was able to provide an estimate of what the data could be. This knowledge, however, was implicit and specific to an individual expert and not available to others, a phenomenon reported in other design studies [9, 10] as well as studies of tools for casual users [11].

We consider these hunches to be critical for, and central to, data analysis conducted by teams or multiple stakeholders. They provide a richer and fuller view of the world than data can provide alone. Our view is grounded in a critical theory perspective that considers data to be an imperfect

*These two authors contributed equally to this work.

This is the authors' preprint version of this paper. License: CC-BY Attribution 4.0 International. Please cite the following reference:

Haihan Lin, Derya Akbaba, Miriah Meyer, Alexander Lex. Data Hunches: Incorporating Personal Knowledge into Visualizations. 2021.

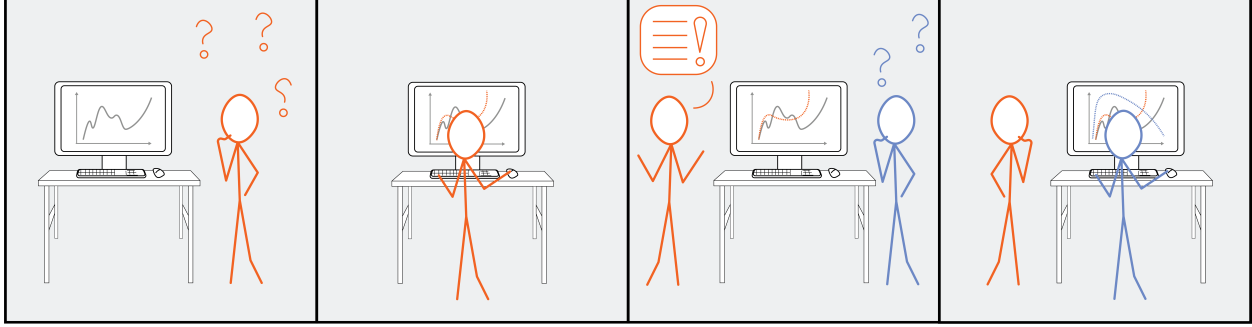


Figure 1: A comic-style representation of data hunches. When a viewer looks at a data visualization and has a hunch about the data that is not represented, they are able to externalize the data hunch through visual methods that we recommend in this paper. Following the externalization of their data hunch, they can communicate with others using the same visual space as the original data. Data hunches can be externalized by anyone based on their personal knowledge and communicated to anyone through visualization interfaces and other collaborative mediums.

and incomplete representation of the world [5, 7, 8, 6, 12]. This critical perspective also considers the knowledge that people have to be situated and pluralistic, with a more complete and objective view of reality coming from the combination of many perspectives [13]. We hence argue that it is through explicitly combining data with externalized hunches about the data that we can enable more productive, precise, and accurate data analysis.

Naming, acknowledging, and valuing these hunches — which we refer to as *data hunches* — opens a wealth of new visualization design opportunities that are fundamentally different from other approaches of considering imperfect data. Uncertainty quantification and visualization [14, 15, 16] assumes that many imperfections can be accounted for, modeled, and quantified to make data better represent reality, which results in datasets that a visualization designer encodes, and a reader passively consumes. Implicit error [9], on the other hand, does not assume that imperfections can necessarily be quantified and modeled, but does approach personal knowledge as a way to detail necessary corrections to data. Our conception of data hunches instead views data as a fixed perspective of the world, and the hunches people have about the data as equally valuable, important, and enlightening. This conception offers new ways to think about how we might use visualizations to both externalize hunches and communicate them to others.

In this paper, we conceptualize data hunches to describe the knowledge that users bring to the data analysis process that augments and complements the data, and in turn becomes part of the interpretation. We consider data visualizations to be a productive mechanism for supporting both the externalization and communication of hunches with the underlying data. To this end, our work includes two core contributions:

- **A conceptualization of data hunches:** We define data hunches and discuss their epistemological foundation, types, and relationship to other existing approaches; and

- **A design space for data hunches:** Through graphical techniques, we demonstrate how data hunches can be externalized and communicated within a visualization context to facilitate collaboration.

We offer a discussion of various considerations for implementing data hunches, such as the potential for malicious intent, the scenarios best suited to hunches, and how hunches might impact trust. Despite open questions about these considerations, we consider this work the first step toward a wealth of productive opportunities for valuing and including personal knowledge in visual data analysis.

2 Data is Imperfect

As a substrate meant to tell us something about the world, data is an imperfect representation. Measurement errors, modeling assumptions, and missing context all make the values stored in a dataset neither a perfect nor complete view of the world. Data imperfections from measurement, modeling, and forecasting phases of analysis are often quantified as uncertainty, with a host of visualization techniques for communicating uncertainty to support transparency and decision-making [16]. Other imperfections stem from the limited view of the world that data values provide, requiring contextual knowledge about who created the data, how, and under what conditions in order to productively interpret them [9]. In this section, we review the literature about the imperfections of data in order to situate our arguments for data hunches as an overlooked opportunity to improve what we can know during visual analysis.

2.1 Data as Imperfect Measurements

As data is captured, processed, and analyzed, it acquires discrepancies from the exact values as they might exist in the world. These discrepancies result from our inability to perfectly measure and collect information about the world

due to a variety of challenges [17, 18]: limited resources, such as not being able to sample every person in a population of interest; limited measurement capabilities, such as the numerical precision of an instrument; or limited knowledge about the future, such as the unpredictability of forecasting weather. In the literature, these discrepancies are referred to as *uncertainty*, and they are often quantified into metrics that describe “the possibility that the observed data or model predictions could take on a set of possible values” [19]. Although the specific definition of uncertainty varies across the literature [20] and fields of study [21], the following are considered common sources for uncertainty [22, 23]: imprecision in measurement apparatuses, modeling assumptions and differences, errors in data collection, incompleteness, and variations in data.

The visualization community has devoted significant effort to understanding how people perceive and understand quantified uncertainty [20, 22, 15, 24, 25, 26, 27]. Building on this knowledge, the community has developed a range of techniques for visually communicating data with its associated uncertainty [17, 28, 16]. Some approaches attempt to intuitively encode uncertainty through modifications of a data item’s graphical mark using blurring [14], sketchiness [29], or value-suppressed color schemes [30]. Other approaches have instead explored visual representations that directly display summary statistics [31, 24] or hypothetical outcomes [32].

Despite extensive research on the visualization of quantified uncertainty, many visualization practitioners and data workers hesitate to include quantified uncertainty in their visualizations and workflows for reasons of comprehension and messaging [19, 20]. Instead, many include qualitative expressions of uncertainty. For example, in an interview study with visualization practitioners, the majority of participants reported using text to warn viewers of the potential uncertainty in a visualized dataset [19]. These qualitative expressions of uncertainty are themselves a source of cognitive uncertainty [20], providing the visualization designer’s own subjective view of what is, and is not, imperfect in the data. The concept of data hunches that we describe in Section 4 encompasses these types of qualitative uncertainty.

2.2 Data as an Imperfect Representation

Data measurements are often a proxy for the thing we really care about in the world, and they are an imperfect proxy at best. For example, if we wanted to know something about how the HCI community has grown over the last decade – a phenomenon without a clearly associated metric – we could look at numbers of attendees at the CHI conference as a proxy. These numbers, while themselves not precisely reflecting community size, have some caveats; 2020 attendance is 0 because the conference was canceled due to the COVID-19 pandemic, and 2021 shows a significant increase over previous years, likely due to the virtual format of the conference. CHI attendance is thus an

imperfect representation of the size HCI community, and one of many possible representations we could choose.

From this perspective, data is an artifact of decisions and situated contexts that reflect the specific phenomena of an individual’s capturing of reality: “*Data are capta*, taken not given, constructed as an interpretation of the phenomenal world, not inherent in it” [33]. Data then, as an object of decision-making practices, is one representation of many possibilities. In recent publications, researchers have raised the issue of the non-neutrality of data, interrogating the epistemological underpinnings that reinforce data as an objective perspective of reality. Kitchin & Lauriault [34] argued from a critical theory standpoint that data does not exist before its creation, and, hence, data does not naturally exist in the world. Instead, data is created by people with intentions to represent some phenomenon in the world. As stated by Gitelman [4]: “raw data is an oxymoron”, the data is “situated, contingent, relational, and framed” [34].

Visualization researchers have made specific recommendations for how to expose the imperfect, representational nature of data in design practices. Dörk and colleagues [8] recommended that visualization authors disclose what they know about the underlying data, explore the provenance of the data they are using, and empower users by giving them access to interrogate the data through interactivity. Correll brought attention to the political power of data [7] and recommends ethical considerations that visualization designers should consider as they work with and visualize data. D’Ignazio & Klein [35] pointed out the hidden labor of data and suggest that the designer also visualizes the data’s provenance in order to call attention to those who collected, curated, and cleaned the dataset. These critical views of visualization design recommend increasing the transparency between visualization, designers, and viewers because data is neither objective nor perfect, and there are consequences to visualizing it as such. Acknowledging data as an imperfect representation is an acknowledgement that data is but one of many perspectives of reality. Our formulation of data hunches is directly inspired by the situated, pluralistic views of critical data and visualization scholars.

3 Methodology

Our methodology for theorizing about data hunches and developing a design space for externalizing and communicating them was based on reflective practices [36, 37]. We began by reflecting on our experiences working with a variety of domain experts who have rich knowledge about their data, knowledge that was not captured in their datasets. Through group discussions about our experiences, we recognized the missing formalization of personal knowledge and its impact in data analysis. We began mapping out the scope of data hunches, the relationship between data hunches and existing visualization concepts, and how hunches have been reported in the existing literature. This process included a literature search into data feminism,

critical data studies, and uncertainty, as well as searching works on design studies and reviewing any reported data hunches in previous design studies.

After investigating the landscape of data hunches, we iteratively developed our proposed design space. The iterations critically reflected illustrative examples from our prior experiences, and design spaces proposed for interactive visualization interfaces, uncertainty visualization, and collaborative sensemaking. We additionally received feedback on our proposed design space from our research lab and made adjustments accordingly. Finally, we used the design space to re-imagine visualization systems presented in several design studies [11, 9, 38], two of which we developed into the case studies described in Section 7.

4 Data Hunches

Situating data as one of many, but limited, perspectives of reality, we introduce personal knowledge as another perspective that deserves representation in visualizations. Personal knowledge about data can influence the interpretation of existing data [39, 40], shape how knowledge is produced [41, 42, 43, 44], and affect how decisions are made [45, 46, 16]. In this section, we frame this personal knowledge formally as *data hunches*, arguing for their potential to produce richer depictions of reality, making connections and distinctions between data hunches and existing concepts surrounding data, and providing a characterization about the common types of data hunches.

4.1 What is a data hunch?

By acknowledging that data is only one, imperfect perspective of reality, we elevate the role that personal knowledge of the data plays in the process of understanding and analyzing it. We define such knowledge as *data hunches*. More precisely, **a data hunch is a person's knowledge about how representative data is of a phenomenon of interest.** The scope of a data hunch can be individual data points, or a complete dataset, or anything in between. Shaped by a person's tacit knowledge about a particular discipline, domain knowledge, life experience, and much more, a data hunch can emerge when a person views (a visualization of) the data and deduces that the data does not completely represent the phenomenon with which they are familiar. A data hunch can be based on the missing context necessary to fully comprehend the phenomenon, discrepancies between a mental model and data, opinions on the quality of the data generation process, and so on. As one is analyzing data, data hunches influence the interpretation of the data, derived knowledge, and decisions made.

A data hunch may take the form of an *assessment* regarding the credibility of a dataset, or may reflect that a specific data item should be *included (or excluded)*. Data hunches can also propose a *directionality* of the data to indicate whether values should be lower or higher. Alternatively,

a data hunch can be expressed as a specific *value* for a data item or a *range (or distribution)* of values that better reflects the phenomenon of interest. These different types of data hunches present alternative perspectives of the phenomenon, augmenting the given perspective of the original data.

We argue that data hunches are prevalent, but often implicit, in data analysis, and potentially as important as the data itself. Using data and data hunches in tandem supports a richer representation of a phenomenon, and potentially better analysis. By acknowledging and naming data hunches, we aim to elevate the potential for personal knowledge to actively and explicitly contribute to data analysis.

4.2 Why do data hunches matter?

Data hunches are personal and can vary significantly between individuals based on a person's unique experiences and knowledge. Combining data hunches from multiple people has the potential to expose a broader, richer, more complete view of a phenomenon. This idea that combining perspectives leads to a fuller and more objective view of reality is grounded in feminist epistemology, and specifically the theory of *situated knowledges* [13].

This theory posits that knowledge cannot be obtained from a single source, but rather is best derived through a collection and collaboration across partial and overlapping perspectives. Data, similarly, cannot fully represent the natural world. From the situated knowledges perspective, data and data hunches capture different perspectives. Thus, we argue that only through overlapping data hunches with other hunches and with data can we expect to visualize a more complete representation of reality.

Visualizations can facilitate both the externalization of data hunches and their communication to others. Walny et al. [47] studied the use of data visualizations on whiteboards in corporate offices and found that visualizations as sketches promote team discussions. Similarly, a visualization designer can incorporate commenting and discussion features to promote externalization of the data hunch [9, 11], apply provenance tracking to record the influence of data hunches on the data source and vice versa [48, 49], use visualization techniques to show a collection of data hunches [50, 51, 52], and much more. In Sections 5 and 6, we propose a design space for externalizing and communicating data hunches within visualization systems to facilitate conversations, exchange opinions, and support better, more nuanced decision-making.

4.3 How do data hunches relate to existing concepts?

Data hunches are not an entirely new concept: they have existed in many forms and been treated in varying ways within the literature. By unifying existing, complementary concepts, data hunches open new design opportunities that focus on externalization and communication methods

that value multiple situated perspectives of reality. In the following paragraphs, we discuss related concepts to distinguish the boundaries of what is and what is not a data hunch.

Data hunches and quantifiable uncertainty. Data hunches, which are based on the knowledge that people bring to data analysis, are not present in the data. Therefore, we first make a clear distinction between quantifiable uncertainty and data hunches, where uncertainty can be systematically recorded, quantified, expressed, and potentially even corrected, whereas data hunches cannot. Although uncertainty visualization frequently is based on a notion of confidence intervals, data hunches can encompass expressions of uncertainty, but can also suggest alternative values, or qualitative assessments of individual data points.

Data hunches and qualitative uncertainty. Qualitative uncertainty, or indirect uncertainty, has been used to describe the quality of the underlying knowledge of data [53]. In contrast to quantifiable uncertainty, qualitative uncertainty is not easily quantified or adjusted in data analysis, and is generally conveyed through “caveats about data” [53]. Franke et al. [54] use the term confidence instead of uncertainty, as the data in humanities projects are judged and rated by experts, and the confidence that experts have in the data will impact the steps in the analysis. Ambiguity is another term used to describe qualitative uncertainty, as introduced in Nowak et al.’s [10] study, where multiple interpretations were possible based on the same data. Additionally, some works include cognitive aspects as a source of uncertainty. For example, Boukhelifa et al. [20] and Schunn & Trafon [22] described how the human-reasoning process can lead to uncertainty in the data and analytical process. Some of these terms, such as ambiguity and confidence, describe personal knowledge that influences interpretations of data, and hence overlap with our notion of data hunches. In contrast, cognitive uncertainty describes the uncertainty caused when reasoning processes lead to different interpretations of the data by different people [22, 20]. Hence, cognitive uncertainty describes the effect of personal knowledge in data analysis more generally.

Data hunches and metadata. Metadata — data about data — helps people navigate a dataset and maintains the meaningfulness of the data [55]. A common example for metadata is a library system, where a reader can easily find a book of their interest (the data) based on indexing information the library provides (the metadata). Therefore, metadata focuses on information that structures the data [56] and provides critical information about the data. In comparison, a data hunch is personal knowledge about data. Rather than providing instructions on how to interpret the data (metadata’s role), data hunches themselves influence the interpretation of data.

Data hunches and implicit error. After observing the hesitancy of public health experts to use visualizations to

analyze Zika-outbreak data, McCurdy et al. learned of the discrepancies between the data measurements and what the experts knew to be true about the spread of Zika [9]. They coined the term *implicit error* to characterize these discrepancies, and defined it as “a type of measurement error that is inherent to a dataset but not explicitly recorded, yet is accounted for qualitatively by experts during analysis, based on their implicit domain knowledge”. Although we consider implicit errors to be data hunches, the formalization of implicit error stems from an epistemological commitment to the idea that data, not people’s knowledge, objectively represents reality; if we can account for implicit error, as the paper claims, we can model systematic errors and fix data generation pipelines. Data hunches, on the other hand, speak to the value and importance of individuals’ knowledge as a perspective separate from, but complementary to, the data. We argue that this situated perspective offers a breadth of new opportunities for capturing and visualizing data hunches in support of richer data analysis.

4.4 Types of Data Hunches

We identify five types of data hunches in support of determining suitable methods for expressing and communicating data hunches from our proposed design space, discussed in Sections 5 and 6.

- **Assessment:** Assessment data hunches speak to the trustworthiness or perceived lack of quality of a dataset, or individual data items. These assessments can be combined with more specific data hunches to indicate, in the case of untrustworthy data, what the data could be.
- **Exclusion & Inclusion:** Exclusion data hunches state that certain data points should not be included in the dataset, possibly in combination with an assessment hunch. Inclusion data hunches state that a data item is missing and could be combined with a value data hunch to state their assumed value.
- **Directionality:** Directional data hunches express that values should be higher or lower. They are a middle ground between assessment hunches, which make no statement about directionality, and value hunches, which give estimates for actual values.
- **Value:** A value data hunch expresses how values in a dataset should be different. Value hunches can be about specific data items, or more holistically, can be based on functions that apply across the dataset.
- **Range & Distribution:** Range and distribution data hunches are similar to value hunches but are less specific, acknowledging uncertainty about a precise value. Instead, they express a value range, or an expected distribution of values.

All these types of data hunches can be expressed for an individual data item, groups of data items, or whole datasets. For example, a value data hunch could apply to a single point (this should be twice as much), to a few data points (all of the items of that type should be twice as much), or to the whole dataset (all data points should be twice as much).

4.5 Context of Data Hunches

What a data hunch says about the data is different from *why* a person has the hunch. In the previous section, we described different types of hunches, focusing on what a data hunch can articulate about what someone perceives as a more true representation of reality. Equally important is *why* someone has the hunch; we refer to the *why* as the *context* of the data hunch.

The context of a data hunch is as critical for its interpretation as the context of data. For the past decade at least, scholars in the field of critical data studies have asserted that data cannot be considered out of context. D’Ignazio & Klein observed that the failure to consider data outside of its context could run the risk of “analytic misstep[s]” [6]. In their critique of big data, boyd & Crawford described how data analytics inherently strip data of their context in pursuit of an objective representation of the world, but that this move is an error: “taken out of context, data lose meaning and value” [12]. In this paper, we argue that like data, a data hunch without context is meaningless, aligning with Seaver’s provocative reminder that “taken out of context, *everything* loses its meaning” [57].

We propose that data hunches *require* context for their meaning and trustworthiness to become clear. Thus, we consider it vital to design mechanisms that capture a data hunch’s context when externalizing a hunch, as well as ways to provide access to that context when communicating a hunch to others. As designers work with stake holders to determine how data hunches are externalized, they should also explore how context can be recorded and shared, as understanding the context of a data hunch is essential to evaluate its trustworthiness.

5 Design Space: Externalizing Data Hunches

Given the various types of data hunches, what are the possibilities to externalize (record) them in a visualization? An externalization technique should allow people to record and express their personal knowledge regarding the original data that is not reflected in the original visualization, nor can it be systematically captured by other processes. Data hunches of different types can often be externalized in different ways. We consider three approaches here: in abstract space, in visualization space, and in data space.

Although the different externalization approaches all serve the purpose of describing a data hunch, each approach has

unique benefits and drawbacks. We consider the following criteria to evaluate each technique:

- **Expressiveness** Can a complex data hunch be expressed, with underlying reasoning?
- **Immediacy**: Is a data hunch expressed in the space the data is visualized, and hence can it be read easily?
- **Consistency**: Can a technique for externalization be used for all other visualization techniques?
- **Discernibility**: Is an externalization easily recognizable as a data hunch or is there a risk of confusing a data hunch with the original data?
- **Technical complexity**: Is an externalization/communication method easy to implement, or are sophisticated UI elements required? Can existing systems be easily retrofitted?
- **Scalability**: Can many data hunches by different individuals be communicated efficiently?

5.1 Externalization in Abstract Space

Externalization in abstract space refers to methods for recording data hunches that are not directly mapped to data, or to the visual representation of data.

5.1.1 Structured elicitation

Structured elicitation, shown in Figure 2a, is an externalization method that captures data hunches through structured UI elements, such as forms, ratings, and votes. Structured elicitation could take the form of “star”-ratings (*assessment*), or by asking analysts whether they think a value should be higher or lower (*directionality*).

Rating, for example, has been used widely in visualization research as a way to judge some quality of the visualization by a group. Quispel & Maes [58] used ratings to investigate preferences of visualization types between different groups of people. McCurdy et al. [9] used structured forms to elicit data hunches from domain experts.

Structured elicitation could be used for all types of data hunches, although its lack of *immediacy* (forms cannot be sensibly embedded in a data visualization) implies that it is less suited for specific expressions of alternative data *values* or *ranges & distributions*. Data hunches generated via structured elicitation can be analyzed and presented in a *scalable* way, as summaries can be easily generated from the responses. Instead of implementing a complicated system that allows sketching or manipulations, designers can use forms to gain some basic information before opting to implement more complex methods. Structured elicitation is also *consistent* for different chart types, making it much easier for the designer to reuse the implementation in different projects. Structured elicitation lacks *expressiveness*, as only predetermined questions can be answered.

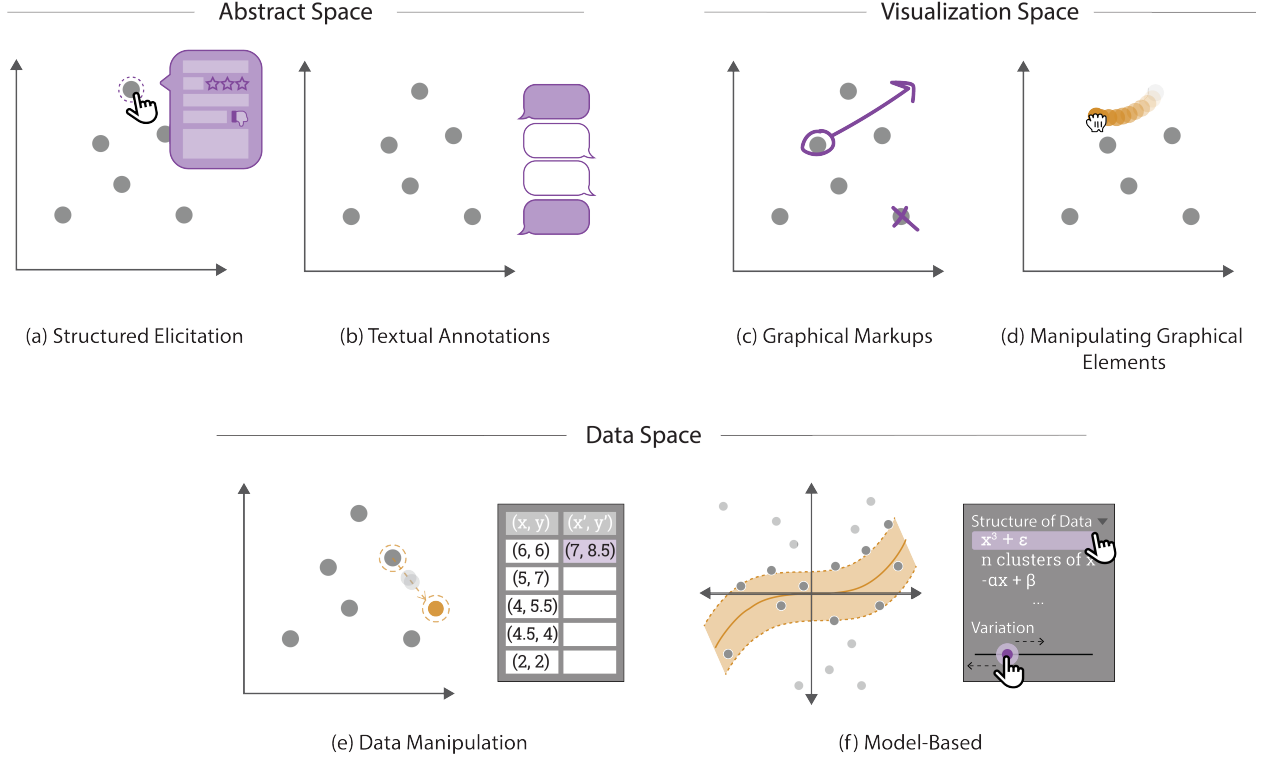




Figure 2: Design space for externalizing (recording) data hunches. We distinguish three categories: externalization in **abstract space**, such as through (a) forms and (b) annotations; externalization in **visualization space**, such as (c) markups or (d) direct manipulation; and externalization in the **data space**, such as (e) data manipulation or (f) the expression of expected patterns via models. We color code expressions of data hunches by whether they are immediately available in data space (orange ) , or are qualitative (violet ).

5.1.2 Textual annotations

Textual annotations, illustrated in Figure 2b, provide the ability to express a data hunch and to describe the context of a data hunch. Textual annotations are extremely *expressive* and can be used to describe all types of data hunches. However, textual annotations lack *immediacy* to help users see their hunches in visualization space. Data hunches explained in text (e.g., “these values should be twice as high”) cannot be easily translated into data or visualization space, and hence are difficult to aggregate or summarize (low *scalability*).

Although textual annotations can be used to express data hunches, they can also be used to provide context about a data hunch, which we argue in Section 4.5 should be externalized with a hunch to support understanding and trust. Such context includes reasoning, concerns, or other comments about a data hunch. Therefore, annotations should be easy to combine with other forms of expression for data hunches.

Textual annotations have low *technical complexity* and are *consistent* throughout various chart types. Previous works have explored various forms of annotations. Liu et al. [59] proposed an annotation system that uses structured format strictly according to the domain context. As another

example implemented differently, Goyal et al. [60] offered more freedom to users by allowing them to use a notepad for free-form notes during their experiment.

5.2 Externalization in Visualization Space

Data hunches can be expressed in the same space as the original visualization. The advantage of this approach is that it provides high *immediacy*, since, for example, a value data hunch can use the same marks and channels as the original visualization, making it the most intuitive dimension for users to express the data hunch.

5.2.1 Graphical markups

Graphical markup (see Figure 2c) refers to adding visual elements directly to a visualization, using approaches such as pen/mouse-based sketching, or adding elements to a visualization using functionality similar to a drawing program. Such markups can use the same marks and channels as the original visualization, but are not limited to it: Users have the freedom to express their data hunches with the encodings they prefer, as shown in Figure 3. Visual markups can be a powerful tool for users to express data hunches of various types, including *exclusion/inclusion* (e.g., by crossing out data points, as shown in Figure 2c), *directionality*

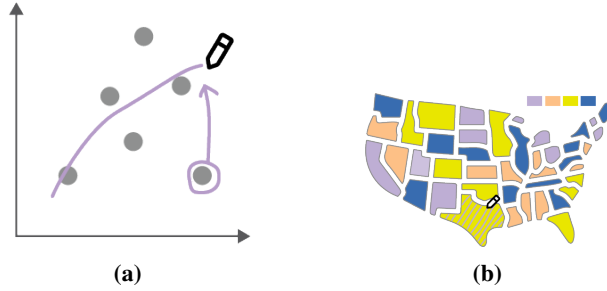


Figure 3: Using graphical markups to externalize data hunches in a choropleth map and a scatter plot. (a) In the scatter plot, we see a markup of an expected trend-line, and an arrow for a *directionality* data hunch. (b) In the choropleth map, the markup uses the same channel (color) as the original visualization to indicate a categorical *value* hunch, yet uses hatching instead of a constant area color.

(e.g., by supplementing an arrow, as shown in Figure 3a), *value* (e.g., by shading an area in a color, as shown in Figure 3b), and *range/distribution* (e.g., by drawing an area where a value is expected to be). The process of graphical externalization also helps with the understanding of and reasoning about visual information [61]. Markups can also efficiently add summary information, such as trend-lines or clusters, and hunches about such summaries (see Figure 3a). Graphical markups provide high *discernibility*, as sketches are commonly easy to distinguish from UI elements.

Previous works have studied sketching for annotations of visualizations. In Kim et al.’s study [62], the participants preferred graphical markups over textual annotations alone. Marasoiu et al. [63] implemented a prototype that uses graphical markups to facilitate communications and to clarify hypotheses in data analysis. Romat et al. [64] presented graphical markup systems to facilitate sensemaking.

However, graphical markups can lead to visual clutter [62], making it much less *scalable* compared to other methods, even though innovative methods exist to show summaries of many people’s sketches [65, 66]. Also, not all visualization methods are equally amenable to being marked-up: whereas, for example, scatter plots are well suited for annotations, space-filling techniques, such as maps, heatmaps, and treemaps, leave little white space and frequently use color encoding, making it harder for annotations to stand out. When developing a markup interface for data hunches, designers must also choose the degrees of freedom and control about the annotation techniques. Although free-form sketches, for example, are most expressive and easy, controlled objects, such as arrows or “strike-out” Xs, could be mapped back to data, potentially improving *scalability*. Implementing graphical markups also can be *technically complex*, requiring designers to add features usually found in drawing tools. Free-form sketching also comes with the disadvantage that users might use ambiguous encodings that might be difficult to resolve.

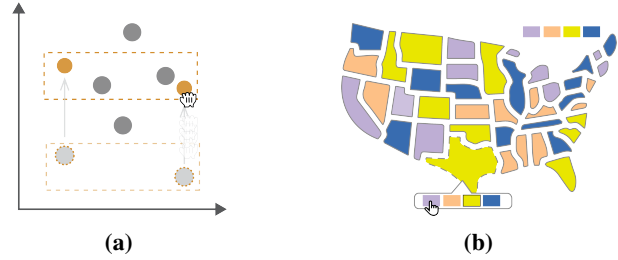


Figure 4: Illustrations of graphical elements manipulations. (a) In the scatter plot example, a user can adjust the location of the data points by making a selection and dragging the points to a position that matches their data hunch. (b) In the choropleth map example, users can select the color based on predetermined color scheme to match the value according to their data hunch.

5.2.2 Manipulating graphical elements

Manipulating graphical elements involves moving, removing, adding, or otherwise changing (e.g., changing color) parts of the visualization that encode data. For example, designers could support externalizing data hunches of *value* type in a scatter plot chart using direct manipulation: users could drag a point up or down according to their data hunches (see Figure 2d). Similarly, a user could pick an alternative color from a color-palette drawing from the visualization’s original color palette for an element on a map (Figure 4b). More sophisticated interfaces could be used to indicate possible ranges of values. Manipulation could also be implemented for aggregated selection, allowing users to make adjustments to groups of elements, such as moving a brushed group of points in scatter plots together in the same direction (Figure 4a). In contrast to markups, graphical manipulations are a direct manipulation of the marks and channels presented by the original visualization (resulting in high *immediacy*), and hence can be easily translated into data space.

Previous works have suggested direct manipulation on visual encodings is a viable way to edit data and provide visual demonstrations of thought processes. Baudel [67] presented editing single or groups of data items in a dataset using graphical manipulations in data visualizations. Saket et al. [68] used graphical manipulations (through repositioning, resizing, and recoloring marks in visualizations) to help users express their expected visualization with increments in direct manipulations, and in turn, the system suggests visual transformations. Saket et al. [69] also provided empirical guidelines and strategies on how people use direct manipulations on graphical encodings to achieve tasks.

One risk with manipulation is that the externalized data hunch is indistinguishable from the original data visualized (resulting in low *discernibility*). Designers must take precautions against this, e.g., by coloring manipulated data points and showing their original location, although the specific approach depends on the underlying visualization.

Overall, we argue that manipulating graphical elements is an intuitive and systematic way to externalize data hunches and record them in the data space. Users have the freedom to adjust the chart according to their data hunches, and designers can systemically record these expressions, making the technique *scalable*. The biggest drawbacks of manipulation are high *technical complexity*, and low *consistency*: Implementations for each chart type require innovative UI solutions and have to be adjusted for every visualization technique. The level of *expressiveness* also depends on the quality of the implementation.

5.3 Externalization in Data Space

Data hunches can also be expressed directly in data space, i.e., by manipulating the data directly, instead of going through a data visualization. For externalization through data space, we consider *data space manipulation* and *expressing model-based hunches* as the two main techniques for this dimension. The former is suitable for data hunches for specific data items, whereas the latter can be especially useful for data hunches about whole datasets.

5.3.1 Data space manipulations

To directly express data hunches of a *value* or *exclusion & inclusion* hunch in data space, users can edit data to express a data hunch (see Figure 2e). Alternatively, users could also use expressions to manipulate data items in bulk. This approach does not directly involve a data visualization and is a relatively easy approach to implement for designers, yet it is considerably less *expressive* than other methods. Also, data space manipulations suffer from low *discernibility*: at an absolute minimum, a visualization has to track different “versions” of a dataset, corresponding to individual user’s hunches, instead of manipulating the original data. In practice, it is advisable that visualizations also use unique encodings for data hunches to differentiate from the original data. To support other types of data hunches (*expressiveness*), such as ranges or directionality, designers would have to provide more sophisticated editing interfaces for data sources.

Romat et al. [64] included data editing in their digital ink externalization system, a functionality requested by participants. Although this functionality was added post facto, it illustrates a preference for editing data directly in externalization systems. Data space manipulation affords excellent *consistency*, because it does not depend on the design of the visualization. However, we argue that it is less *immediate* than externalizations on top of a data visualization. As all techniques that can translate a data hunch into data space, it offers great *scalability*.

5.3.2 Expressing model-based hunches

Although the data hunches we have discussed so far focus on explicitly manipulating or annotating data items, users can also have a preconceived notion of how a dataset should behave (Figure 2f). For example, an analyst would

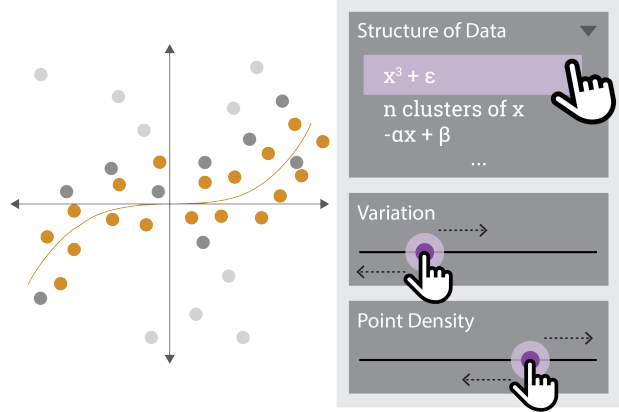


Figure 5: Example for expressing model-based hunches. Users can select a model from a selection of predefined models or specify their own. Subsequently, Users could adjust the variance of the model as well as the number of visualized points. The interface here shows the points that are included as part of the data hunch model selection and the original data points that would not fit into the model.

expect that a certain dataset follows an exponential function. In this case, rather than focusing on a particular data point, the designer can implement an externalization method that allows users to express a data hunch of *range/distribution* type that is described by models and formulas, as illustrated in Figure 5.

Marasoiu et al. [63], for example, presented an interface that allows users to sketch models, which then generates data points based on the sketch, as a way to facilitate communication between customers and analysts. A model-generating technique would allow users to create a model of their data hunch, and the system then generates data points based on the model. For example, a system could recognize a group of points in a scatter plot as a cluster [70], and let analysts move that cluster around. Alternatively, the designer can also provide a selection of common mathematical models for the user to choose from based on their data hunch, and then make the data point adjustment after the selection and generation.

If a model is used to generate data points, many of the attributes of data space manipulations apply, including the danger of low *discernibility*, communicating hunches as real data, and overlooking the need to communicate a model-based hunch. An additional opportunity for communicating the data hunch, and for increasing *immediacy*, is visualizing the model together with the original and modeled data in the visualization.

6 Design Space: Communication and Collaboration

Once data hunches are expressed, the next challenge is to appropriately visualize them. The method of communi-

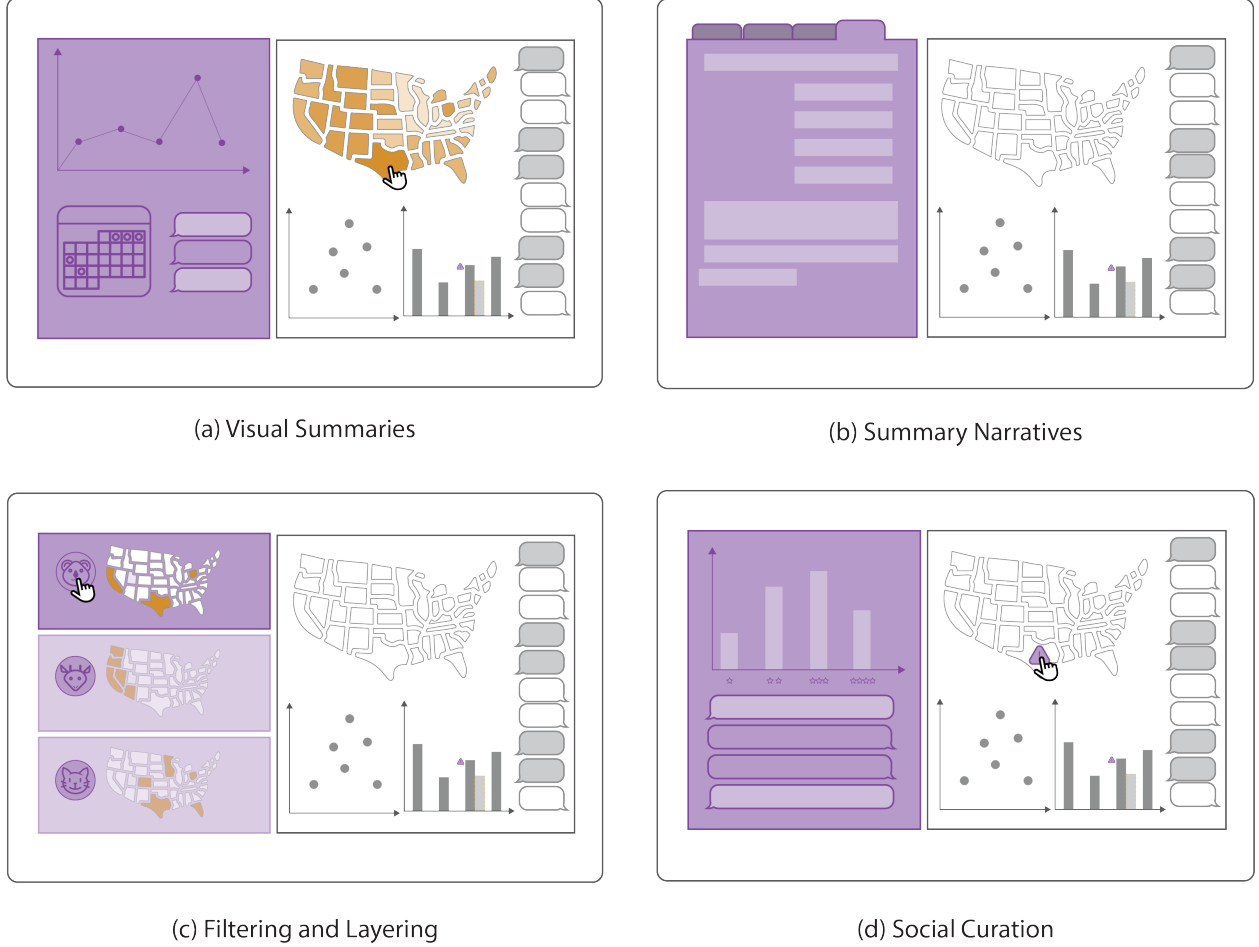




Figure 6: Design space to support data hunches for communication and collaborations. We present four techniques that focus on the social aspects of data hunches, and are all suitable for exploring multiple data hunches within the same interface. Violet spaces  indicate qualitative information related to  indicate data hunches externalized in or transcribed into the data space.

cating data hunches also obviously depends on the externalization method used, especially when data hunches are not available in the data space. Some basic tactics such as using tooltips and showing data hunches on-demand can be used to organize the data hunches in the visualization interface. One rule that should always be followed is that data hunches should always be shown as distinct from the original data encoding (see the *discernibility* criterion), clearly indicating that what is shown is a person’s data hunch. Techniques such as using sketchiness to discern data hunches from the original visualization can be effective for the task [71].

Although externalizing individual data hunches is important as a memory aid and for recording personal knowledge, the main benefit of expressing and reading data hunches is in collaborative scenarios. A data hunch represents an externalization of specific situated knowledge [13], and through the collection of data hunches across multiple people over time, we can use visual methods to generate a dialog about where situated knowledges overlap and where

they differ. The encoding of many people’s data hunches, however, can be challenging, as individual encodings of data hunches can quickly clutter the visualization. As part of the collaborative process, it may be useful to see trends, clusters, or higher level reasoning across individuals and their respective data hunches. In this section, we present a design space that focuses on using data hunches for collaboration and communication, orthogonal to the design space for externalization.

6.1 Visual Summaries

Creating a summary visual of all the hunches can offer a quick overview of areas on the original visualization where hunches have been created. Techniques such as heatmaps can provide an overview of where in the visualization data hunches occur, signaling hot spots that will likely require closer inspection of the data or an in-depth discussion about the content behind the individual data hunches, shown in Figure 6a. The designer also has the op-

portunity to more precisely aggregate data hunches when they are recorded in data space. For example, it is possible to group together the data hunches that reflect an increase in value in one group and the ones representing a decrease in another, in order to display the general trends in how data hunches have been externalized. Additionally, a designer can use a visual summary to present the values of data hunches for specific data items, possibly showing a group consensus of how specific data items may better reflect reality. Information from structured elicitation, such as low trust in specific data points, could also be quantified and summarized.

Data hunches externalized through *graphical markups* are hard to use in visual summaries, without transcribing the markups manually or through the use of algorithms, although approaches such as Forma Fluens [65] or the *New York Times*’ “you draw it” [66] of overlaying many user’s sketches could prove interesting. Similarly, textual comments could be mined for re-occurring insights.

6.2 Summary Narratives

Beyond communicating the data hunches themselves, reasoning and justification are what make data hunches valuable; after all, if the data hunch is only a visual expression, other users can learn very little about the meaning and context of a data hunch, as previously discussed in Section 4.5. Therefore, we recommend implementing textual annotations that serve as a space to capture *summary narratives*, important characteristics about the context, and reasoning for externalizing data hunches. The designer can choose to implement summary narratives through a structured form (as shown in Figure 6b), or leave the summary narrative open so that it does not preemptively scope the information that people might provide. We relate this implementation to work done in the fairness in machine learning community, where it has been recommended that datasheets accompany datasets, so that appropriate context is recorded about the data [72].

Instead of specific elements or magnitude, summary narratives place the emphasis on expressing the overall sentiment a user takes regarding their data hunch(es): such as specific beliefs surrounding the credibility of a data source, personal experiences that informed their data hunches, or general background knowledge that they are bringing to their interpretation and externalization. Showing such context is important for communicating higher level reasoning and influential factors in externalization. This type of information may be captured in *textual annotations*, as was seen in the sense.us [11], but is challenging to extract across multiple users and annotations.

6.3 Filtering & Layering

In some instances, an abundance of data hunches may hinder collaboration, especially when *who* created the data hunch affects *how* it is interpreted. We envision systems for expressing data hunches to be especially useful in en-

vironments where participants know each other and know each other’s roles, such as in a scientific community, or in a hospital system. One way of filtering data hunches would hence be to show data hunches of only users one knows, or users they trust. Data hunches could also be filtered by metadata, such as the role of the person expressing the data hunch in an organization (e.g., include only data hunches by technical staff), or the time a data hunch was logged (e.g., include only data hunches expressed after an important event that changed their opinions regarding the data). Finally, data hunches could be filtered by some quality metric, such as reputation scores of its author, or whether the data hunch also provides context (discarding hunches that provide no rationale). Instead of filtering, such information could also be used to weigh data hunches, e.g., giving data hunches that also provide contextual annotations a higher weight in a summary visualization. Ultimately, if these methods are used for analysis or decision-making, we suggest that the decision maker should justify why they placed more weight on certain data hunches over others.

For data hunches that are externalized through methods such as *graphical markups*, this can be a particular useful technique to display the data hunches without overwhelming the reader with too many visual elements on the screen. Filtering and layering can also be an important way to elicit data hunches in collaborative settings where it is important to remove social influence from the externalization. In instances where stakeholders think that the process of externalizing data hunches may be easily influenced by social factors, e.g. seeing what others did and thus changing one’s own externalization to match, filtering systems can display only the original visualization prior to externalization. Discussing the likelihood of other people’s opinions impacting their own externalization with stakeholders is important.

6.4 Social Curation

With several data hunches externalized on the same data visualization, other members in the group can have varying opinions on them. Previous works [11, 50] have shown that forums for data visualizations can facilitate further understanding and promote discussions on the data visualization. Similar to online forums, designers can implement social aspects of data hunches: rankings, voting mechanisms, and commenting systems, as shown in Figure 6d. Social curations through these methods can lead to two valuable outcomes: gathering sentiment on these data hunches and promoting discussions about data hunches. With multiple data hunches expressed within a group, voting and ranking systems can let members in the group easily express their agreement or disagreement with a data hunch, and they can also elaborate more with comments and discuss the data hunches with other members. New readers to the visualization can have an easier time navigating through different data hunches, if they are sorted by discussion and ranks. Furthermore, data hunches can be built on top of other data hunches or influenced by other data hunches,

and the designer can implement provenance tracking to see how data hunches intertwine, in which the structure can be much clearer when put in forums.

Another approach would be to enable experts to curate a report based on data hunches, similar to how scientists with expertise in a particular area analyze publications and summarize them in review publications. Such curated data hunches could summarize the data hunches of large crowds, highlight common themes or areas of disagreements, and link to the original data hunches for reference.

7 Case Studies

Here we present two case studies to apply our design space to prior work. We chose these prior works because they include designs that leverage personal knowledge, cover a range of domain expertise, and include illustrative examples of how individuals externalized and collaborated through visualization tools. In the first example, the authors look at personal knowledge in a specific domain, whereas in our second example, the authors work with public information. For each case study, we review the original work and highlight how data hunches exist under different names. From there, we present new design opportunities afforded by our reconceptualization of personal knowledge through the lens of data hunches.

7.1 Zika-Outbreak Data Visualization

McCurdy et al. [9] presented their collaboration with epidemiologists studying Zika virus outbreak data, a dataset collated across different South American countries. After following general design study guidelines [73] for developing appropriate task abstractions, they presented their visualization prototype and then received feedback that indicated discrepancies between the data visualized in the prototype and the experts' knowledge. Probing further, they came to understand that the experts perform a series of mental modifications to the data, informed by their personal and domain knowledge. For example, they knew that Country A reported all cases regardless of investigation status, whereas Country B reported only fully investigated cases, resulting in fewer officially recorded cases for Country B in the dataset. After re-examining how the analysts were using the visualization, McCurdy et al. proposed a framework of externalizing implicit error and provided a structured method for the experts to record such knowledge. Implicit error is defined as the discrepancy in the data that is not explicitly recorded but accounted for in the analysis process, which falls under our definition of data hunches. Therefore, we will refer it as a data hunch in the following text.

In McCurdy et al.'s prototype, *structured elicitation* and *textual annotations* are the main methods used to externalize data hunches (right of Figure 7). In the annotation template, the experts are asked to identify the region and describe the data hunch with text. The template also solic-

its other attributes, such as the impact of the data hunch in the analysis stage, the potential necessary adjustments, and the expert's confidence in their hunch. These options are multiple choice questions, and the expert can choose what to answer according to the scope of their data hunch. After the data hunch is externalized, the existence of the data hunch is communicated through pins on the map, and the details of the data hunch are showed in a *summary narrative* on demand. All the data hunches are layered on top of and distinguished from the original visualization.

Leveraging our proposed design space, several changes could be made to expand the expressions of data hunches, making them more intuitive and explicit. Specifically, McCurdy et al.'s prototype did not utilize visualization space for externalization. Although text-based annotations are necessary for recording reasoning processes, they lack the immediacy to demonstrate impact and change in a cohesive visual manner. We suggest alternative designs for three areas of the original prototype: the map showing the number of cases per country, line charts that depict country specific stats, and methods for communications and collaboration across different analysts. In all these areas, the designer can use *manipulating graphical elements* and *graphical markups* to help the experts express their data hunch more explicitly.

Choropleth map. In the initial instantiation of data hunch externalization, McCurdy et al. use pins and text annotations. Instead, the designer could implement a color adjustment slider (a graphical manipulation) to help the expert directly describe their hunch on the map. This type of manipulation affords experts the opportunity to externalize their data hunch about a specific country in relation to the Zika cases in other countries, without having to pick a precise number of cases, which is an admittedly difficult task.

Line charts. In the prototype, upon clicking a country, additional charts are revealed showing more data about Zika cases over time. The designer could implement graphical markup and manipulations for these individual charts, in combination with the existing text annotations. These additions will make the externalization process more intuitive. Graphical markups and manipulations will give analysts the ability to more precisely indicate where and how they think the values differ. By shifting the work of externalization to a visual channel, the textual annotations can be reserved for reasoning and contextualizing data hunches.

Collaborations. Since the original visualization uses a choropleth map, a heatmap for data hunches is not appropriate. Instead, a designer has other options to make it easier for a team of analysts to share and communicate their data hunches. One suggestion is allocating the role of *social curation* to a team member to summarize and report on trends across hunches. Another method is enabling *filtering and layering* of hunches, so that analysts can fo-

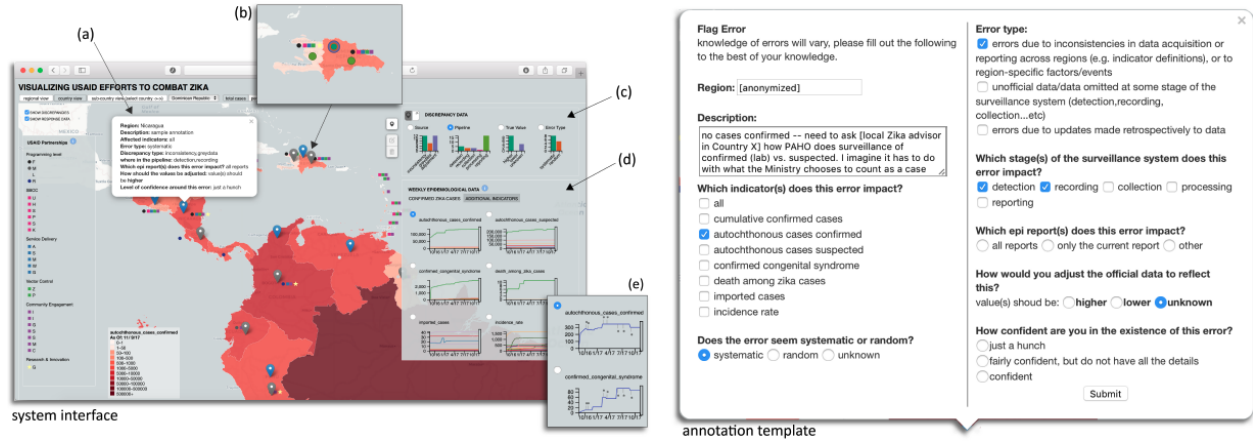


Figure 7: McCurdy et al.'s [9] prototype that implemented the framework for implicit error. We recommend changes to the choropleth map, the country specific line charts, and the ability to share and communicate multiple data hunches.

cus on either specific hunches or a collection of hunches. Since color is the main visual encoding of the map, we recommend using a fill of patterns or textures in combination with opacity to illustrate the existence of multiple and overlapping data hunches.

7.2 Online Forum sense.us

Heer et al. [11] studied asynchronous online collaborations through an interactive visualization interface, shown in Figure 8. The authors recruited 24 participants and observed how they use comments and graphical annotations analyze data in a social context. The authors chose to visualize census data about the U.S. labor force in the interface, sense.us, reasoning that the participants could easily relate to this data and consequently express their opinions about it in annotations. The authors observed that participants utilized many features implemented in sense.us for their social data analysis. Heer et al. reported that the participants had conversations linked to specific views of charts and used graphical annotations to point at specific aspects of interest on the chart. This study demonstrates that with a social platform with interactive features for data analysis, users were able to gain a deeper understanding of the data. Heer et al. described these externalizations as “contextual knowledge”, “commentary”, “historical knowledge”, and “personal anecdotes” — we consider all of them to be data hunches.

Using the terminology of our design space, the sense.us system mainly utilizes *graphical markups* (left of Figure 8) and *textual annotations* (right of Figure 8) to enable social data analysis. A user can make comments on a chart or a state of a chart, and the comments are preserved in the comments listing. The sense.us system also allows users to use graphical markups to reference certain points in the charts and make textual annotations regarding the graphical markups. Comments with references to graphical markups are marked with a special symbol in the comment listings. Collaboratively, sense.us allows users to view others' com-

ments and respond in a comment thread. In their discussion section, the authors note that while graphical markups were expressive, they wondered whether there were “methods of sketching that can be somehow data-aware?”. Addressing this question, we make specific recommendations in response to interactions that were highlighted in the paper.

Externalizing expectations Heer et al. reported where users made comments on the data and data visualizations. In their example of where a user expected a different pattern for the percentage of the U.S. work force in the military, the user used graphical markups to highlight points where they expected bigger jumps in the data. The designer could implement more intuitive features for the externalization process. For area charts, *manipulating graphical elements* enables the user to externalize their hunch in the same visual space, while keeping the externalization “data-aware” as suggested by Heer and colleagues. Alternatively, the designer can use the *data manipulation* technique to allow the user to directly input their data hunches.

Navigating data hunches. An important functionality of sense.us is that the textual annotations and graphical markups are “doubly-linked” — comments were associated with specific frames of the visualization, where navigating either would result in seeing the corresponding view. This interaction situated the data hunches with the data visualization. To extend this method of collaboration, the designer can use *social curation* and *filtering and layering*. The designer can implement more collaborative features for sharing and tracking data hunches, such as a ranking mechanism where users can agree or disagree with each other's data hunches. This method, used in conjunction with layering techniques, can help users track others' data hunches and provide a visual overview of where data hunches converge and where they diverge.

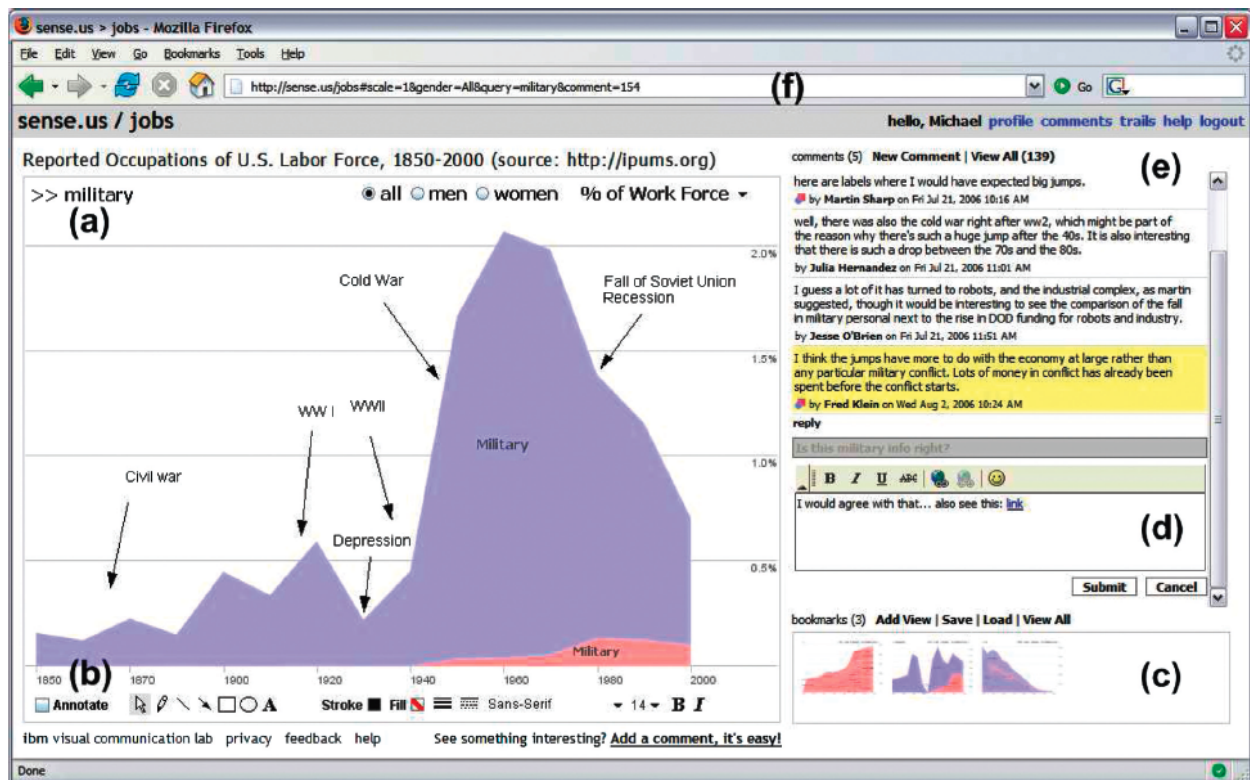


Figure 8: Sense.us [11] interface. Using our design space, we recommend that the designer implement data and graphical manipulations, in addition to the graphical markups that the interface already utilizes. For better navigation of the data hunches, we propose ranking and voting on data hunches, sharing data hunches, and layering the data hunches on the visualization for an overview of where the data hunches are externalized.

8 Discussion

In our advocacy for data hunches, we have made the assumption that data hunches provide a more pluralistic and richly honest view of the world. But, what if a hunch is wrong? Or worse, what if a hunch is maliciously intended as misinformation? Externalizing and communicating the context of a data hunch is an important step toward supporting others in deciding the value and trustworthiness of a hunch. Going further, including information about who the data hunch comes from, could provide another layer of credibility and accountability. Associating a data hunch with a specific person, however, comes with a different set of issues. In some settings, the politics of an organization or field could lead vulnerable people to remain silent about contradictory hunches, depriving others of important perspectives. But anonymity can be equally caustic by invoking negative behavior toward others with opposite views [74]. How to ensure the value, credibility, trustworthiness, and transparency of data hunches is an important, yet open, question.

Another important consideration is what types of visualization systems and scenarios are most appropriate for implementing mechanisms that support data hunches. We believe that most visual data analysis involves hunches, but

designing and developing tools that support externalization and communication of multiple data hunches is likely to require significant effort. Therefore, we envision systems that support recording and communicating data hunches to be implemented primarily in scenarios where the topic of the data is of shared interest among larger communities or society in general. For example, a recent project elicited feedback from the scientific community on an animation of the SARS-CoV-2 protein structure [75]. Unlike visualization tools designed for an individual research lab, such applications target a wider audience with shared interests, where visualizing data hunches can lead to deeper impact compared to casual visualizations.

Another interesting, open question is: What happens to someone's trust in a visualization and the underlying data when data hunches are communicated in a tool? Previous work [76] has reported that *social information* can affect a user's trust and memorability about the data visualization. We anticipate similar effects with the inclusion of data hunches. We argue in this paper that data is an imperfect representation of reality, and making that imperfection visible is one goal of our work. However, if data hunches make people less trusting, will designers avoid including them, as they sometimes do with uncertainty [19]?

A reader may trust the visualization more when data hunches are provided by experts or are highly rated. On the other hand, if too many data hunches disagree with the original data, the reader may trust the source of the visualization less. In the end, the goal of conceptualizing data hunches and proposing a design space for them is to formally recognize the role of personal knowledge in understanding data and empower users to express their views. Designers should fully consider the possible impacts of data hunches before committing to including or excluding them. The work we present in this paper is only the first step in exploring a rich space of opportunities about how, why, and when to include personal knowledge about data in visualizations.

9 Conclusion

In this work, we framed the personal knowledge about how representative data is, defining such knowledge as data hunch, analyzed the implication of supporting data hunches in data analysis, and proposed a design space for externalizing and communicating data hunches in data visualizations. We mapped out the differences between data hunches and existing concepts. The proposed design space provides designers recommendations for how to integrate data hunches in their works. The ultimate goal of this work is to formalize and recognize the significant role that personal knowledge has in understanding data, which many works overlook, and elevate this personal knowledge into another form of information that can be explicitly externalized and utilized. Through this work, we seek to question the notion of data being the gold standard of representing phenomena in the world, and open up the potential to grow visualization research beyond constrained notions of data.

We recognize that our proposed design space for data hunches is based on previous research and visualization community expertise, and that we did not implement and validate demonstrations to showcase the potential use of data hunches, which we consider to be exciting future work. Based on such implementations, we plan to study how users express data hunches, the impact of data hunches on interpretation, and the influence of data hunches in analytic processes. Furthermore, we only touched the surface of the potential applications of data hunches. We plan to explore more use cases for data hunches and analyze the implications of data hunches even further.

References

- [1] Erik Brynjolfsson and Kristina McElheran. The Rapid Adoption of Data-Driven Decision-Making. *American Economic Review*, 106(5):133–139, May 2016.
- [2] Michael Troilo, Adrien Bouchet, Timothy L. Urban, and William A. Sutton. Perception, reality, and the adoption of business analytics: Evidence from North American professional sport organizations. *Omega*, 59:72–83, March 2016.
- [3] National Science Foundation. NSF’s 10 Big Ideas - Special Report. https://www.nsf.gov/news/special_reports/big_ideas/harnessing.jsp.
- [4] Lisa Gitelman. *Raw Data Is an Oxymoron*. MIT Press, Cambridge, MA, January 2013.
- [5] Yanni Alexander Loukissas. *All Data Are Local: Thinking Critically in a Data-Driven Society*. MIT Press, April 2019.
- [6] Lauren F. Klein and Catherine D’Ignazio. *Data Feminism*. The MIT Press, 2020.
- [7] Michael Correll. Ethical Dimensions of Visualization Research. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, pages 188:1–188:13, New York, NY, USA, 2019. ACM.
- [8] Marian Dörk, Patrick Feng, Christopher Collins, and Sheelagh Carpendale. Critical InfoVis: Exploring the politics of visualization. In *CHI ’13 Extended Abstracts on Human Factors in Computing Systems on - CHI EA ’13*, page 2189, Paris, France, 2013. ACM Press.
- [9] Nina Mccurdy, Julie Gerdes, and Miriah Meyer. A Framework for Externalizing Implicit Error Using Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):925–935, January 2019.
- [10] Stan Nowak, Lyn Bartram, and Pascal Haegeli. Designing for Ambiguity: Visual Analytics in Avalanche Forecasting. In *2020 IEEE Visualization Conference (VIS)*, volume 1, pages 81–85, Salt Lake City, UT, USA, September 2020. IEEE.
- [11] Jeffrey Heer, Fernanda B Viégas, and Martin Wattenberg. Voyagers and voyeurs: Supporting asynchronous collaborative visualization. *Commun. ACM*, 52(1):87–97, 2009.
- [12] danah boyd and Kate Crawford. Critical Questions for Big Data. *Information, Communication & Society*, 15(5):662–679, June 2012.
- [13] Donna Haraway. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3):575, 1988.
- [14] A. M MacEachren. Visualizing uncertain information. *Cartographic Perspective*, 13(13):10–19, 1992.
- [15] Hua Guo, Jeff Huang, and David H. Laidlaw. Representing Uncertainty in Graph Edges: An Evaluation of Paired Visual Variables. *IEEE Transactions on Visualization and Computer Graphics*, 21(10):1173–1186, October 2015.
- [16] Lace M. K. Padilla, Maia Powell, Matthew Kay, and Jessica Hullman. Uncertain About Uncertainty: How Qualitative Expressions of Forecaster Confidence Impact Decision-Making With Uncertainty Visualizations. *Frontiers in Psychology*, 11:579267, January 2021.

- [17] 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 Charles D. Hansen, Min Chen, Christopher R. Johnson, Arie E. Kaufman, and Hans Hagen, editors, *Scientific Visualization*, pages 3–27. Springer, London, 2014.
- [18] Kristin Potter, Paul Rosen, and Chris R. Johnson. From Quantification to Visualization: A Taxonomy of Uncertainty Visualization Approaches. In *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer, London, 2012.
- [19] Jessica Hullman. Why Authors Don’t Visualize Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):130–139, January 2020.
- [20] Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. How Data Workers Cope with Uncertainty: A Task Characterisation Study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3645–3656, Denver Colorado USA, May 2017. ACM.
- [21] Judi Thomson, Elizabeth Hetzler, Alan MacEachren, Mark Gahegan, and Misha Pavel. A typology for visualizing uncertainty. In *Visualization and Data Analysis 2005*, volume 5669, pages 146–157, San Jose, California, USA, March 2005. International Society for Optics and Photonics.
- [22] Christian Schunn and J. Trafon. The psychology of uncertainty in scientific data analysis. In *Handbook of the Psychology of Science*, pages 461–483. Springer Publishing Company, New York, NY, November 2012.
- [23] Henning Griethe and Heidrun Schumann. The Visualization of Uncertain Data: Methods and Problems. In *Proceedings of SimVis ’06*, pages 143–156, Magdeburg, Germany, 2006. SCS Publishing House.
- [24] Michael Correll and Michael Gleicher. Error Bars Considered Harmful: Exploring Alternate Encodings for Mean and Error. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2142–2151, December 2014.
- [25] J. Sanyal, Song Zhang, G. Bhattacharya, P. Amburn, and R. Moorhead. A User Study to Compare Four Uncertainty Visualization Methods for 1D and 2D Datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1209–1218, November 2009.
- [26] Susanne Tak, Alexander Toet, and Jan van Erp. The Perception of Visual Uncertainty Representation by Non-Experts. *IEEE Transactions on Visualization and Computer Graphics*, 20(6):935–943, June 2014.
- [27] Meredith Skeels, Bongshin Lee, Greg Smith, and George Robertson. Revealing uncertainty for information visualization. *Information Visualization*, 9(1):70–81, 2010.
- [28] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):903–913, January 2019.
- [29] Nadia Boukhelifa, Anastasia Bezerianos, Tobias Isenberg, and Jean-Daniel Fekete. Evaluating Sketchiness as a Visual Variable for the Depiction of Qualitative Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2769–2778, December 2012.
- [30] Michael Correll, Dominik Moritz, and Jeffrey Heer. Value-Suppressing Uncertainty Palettes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, Montreal QC Canada, April 2018. ACM.
- [31] K. Potter, J. Kniss, R. Riesenfeld, and C.R. Johnson. Visualizing Summary Statistics and Uncertainty. *Computer Graphics Forum*, 29(3):823–832, August 2010.
- [32] Jessica Hullman, Paul Resnick, and Eytan Adar. Hypothetical Outcome Plots Outperform Error Bars and Violin Plots for Inferences about Reliability of Variable Ordering. *PLOS ONE*, 10(11):e0142444, November 2015.
- [33] Johanna Drucker. DHQ: Digital Humanities Quarterly: Humanities Approaches to Graphical Display. <http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html#p1>, 2011.
- [34] Rob Kitchin and Tracey Lauriault. Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. SSRN Scholarly Paper, Social Science Research Network, Rochester, NY, July 2014.
- [35] Catherine D’Ignazio and Lauren F Klein. Feminist Data Visualization. *Workshop on Visualization for the Digital Humanities (VIS4DH)*, page 5, 2016.
- [36] Rowanne Fleck and Geraldine Fitzpatrick. Reflecting on reflection: Framing a design landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction - OZCHI ’10*, page 216, Brisbane, Australia, 2010. ACM Press.
- [37] Donald A. Schön. *The Reflective Practitioner: How Professionals Think in Action*. Routledge, Oxfordshire, England, March 2017.
- [38] Haihan Lin, Ryan A Metcalf, Jack Wilburn, and Alexander Lex. Sanguine: Visual analysis for patient blood management. *Information Visualization*, 20(2-3):14738716211028565, 2021.
- [39] Jimmy Moore, Pascal Goffin, Miriah Meyer, Philip Lundrigan, Neal Patwari, Katherine Sward, and Jason Wiese. Managing In-home Environments through Sensing, Annotating, and Visualizing Air Quality Data. *Proceedings of the ACM on Interactive, Mobile,*

- Wearable and Ubiquitous Technologies*, 2(3):128:1–128:28, September 2018.
- [40] Peter Tolmie, Andy Crabtree, Tom Rodden, James Colley, and Ewa Luger. “This has to be the cats”: Personal Data Legibility in Networked Sensing Systems. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 491–502, San Francisco California USA, February 2016. ACM.
 - [41] P. Saraiya, C. North, Vy Lam, and K.A. Duca. An Insight-Based Longitudinal Study of Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1511–1522, November 2006.
 - [42] C. North. Toward measuring visualization insight. *IEEE Computer Graphics and Applications*, 26(3):6–9, May 2006.
 - [43] Crystal Lee, Tanya Yang, Gabrielle D Inchoco, Graham M. Jones, and Arvind Satyanarayan. Viral Visualizations: How Coronavirus Skeptics Use Orthodox Data Practices to Promote Unorthodox Science Online. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, number 607, pages 1–18. Association for Computing Machinery, New York, NY, USA, May 2021.
 - [44] Chen He, Luana Micallef, Liye He, Gopal Peddinti, Tero Aittokallio, and Giulio Jacucci. Characterizing the Quality of Insight by Interactions: A Case Study. *IEEE Transactions on Visualization and Computer Graphics*, 27(8):3410–3424, August 2021.
 - [45] Evan M. Peck, Sofia E. Ayuso, and Omar El-Etr. Data is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI ’19, pages 1–12, New York, NY, USA, May 2019. Association for Computing Machinery.
 - [46] B. Karer, H. Hagen, and D. J. Lehmann. Insight Beyond Numbers: The Impact of Qualitative Factors on Visual Data Analysis. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1011–1021, February 2021.
 - [47] J. Walny, S. Carpendale, N. Henry Riche, G. Venolia, and P. Fawcett. Visual Thinking In Action: Visualizations As Used On Whiteboards. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2508–2517, December 2011.
 - [48] Mi Feng, Cheng Deng, Evan M. Peck, and Lane Harrison. HindSight: Encouraging Exploration through Direct Encoding of Personal Interaction History. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):351–360, January 2017.
 - [49] Zachary T Cutler, Kiran Gadhave, and Alexander Lex. Trrack: A Library for Provenance Tracking in Web-Based Visualizations. In *IEEE Visualization Conference (VIS)*, pages 116–120, Salt Lake City, UT, USA, 2020. IEEE.
 - [50] Fernanda B. Viegas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. ManyEyes: A Site for Visualization at Internet Scale. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1121–1128, 2007.
 - [51] A. Mathisen, T. Horak, C. N. Klokmoose, K. Grønbaek, and N. Elmqvist. InsideInsights: Integrating Data-Driven Reporting in Collaborative Visual Analytics. *Computer Graphics Forum*, 38(3):649–661, 2019.
 - [52] Petra Isenberg, Niklas Elmqvist, Jean Scholtz, Daniel Cernea, Kwan-Liu Ma, and Hans Hagen. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization*, 10(4):310–326, October 2011.
 - [53] Anne Marthe van der Bles, Sander van der Linden, Alexandra L. J. Freeman, James Mitchell, Ana B. Galvao, Lisa Zaval, and David J. Spiegelhalter. Communicating uncertainty about facts, numbers and science. *Royal Society Open Science*, 6(5):181870, May 2019.
 - [54] Max Franke, Ralph Barczok, Steffen Koch, and Dorothea Weltecke. Confidence as First-class Attribute in Digital Humanities Data. *Proceedings of the 4th VIS4DH Workshop*, page 5, October 2019.
 - [55] Richard Gartner. What Metadata Is and Why It Matters. In Richard Gartner, editor, *Metadata: Shaping Knowledge from Antiquity to the Semantic Web*, pages 1–13. Springer International Publishing, Cham, 2016.
 - [56] Erik Duval and Wayne Hodgins. Metadata principles and practicalities. *D-Lib Magazine*, 8:2002, 2002.
 - [57] Nick Seaver. The nice thing about context is that everyone has it. *Media, Culture & Society*, 37(7):1101–1109, October 2015.
 - [58] Annemarie Quispel and Alfons Maes. Would you prefer pie or cupcakes? Preferences for data visualization designs of professionals and laypeople in graphic design. *Journal of Visual Languages & Computing*, 25(2):107–116, April 2014.
 - [59] Qi Liu, Chong Chen, Enjian Shen, Fangqing Zhao, Zhongsheng Sun, and Jinyu Wu. Detection, annotation and visualization of alternative splicing from RNA-Seq data with SplicingViewer. *Genomics*, 99(3):178–182, March 2012.
 - [60] Nitesh Goyal, Gilly Leshed, and Susan R. Fussell. Effects of Visualization and Note-taking on Sense-making and Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’13, pages 2721–2724, New York, NY, USA, 2013. ACM.
 - [61] Mary Hegarty and Kathryn Steinhoff. Individual differences in use of diagrams as external memory in mechanical reasoning. *Learning and Individual Differences*, 9(1):19–42, January 1997.
 - [62] Yea-Seul Kim, Nathalie Henry Riche, Bongshin Lee, Matthew Brehmer, Michel Pahud, Ken Hinckley, and

- Jessica Hullman. Inking Your Insights: Investigating Digital Externalization Behaviors During Data Analysis. In *Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces, ISS '19*, pages 255–267, New York, NY, USA, 2019. ACM.
- [63] Mariana Marasoiu, Alan F. Blackwell, Advait Sarkar, and Martin Spott. Clarifying Hypotheses by Sketching Data. In *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization: Short Papers*, page 5 pages, Groningen, The Netherlands, 2016. Eurographics Association.
- [64] Hugo Romat, Nathalie Henry Riche, Ken Hinckley, Bongshin Lee, Caroline Appert, Emmanuel Pietriga, and Christopher Collins. ActiveInk: (Th)Inking with Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. Association for Computing Machinery, New York, NY, USA, May 2019.
- [65] Mauro Martino, Hendrik Strobelt, Owen Corne, and Evan Phibbs. Forma Fluens - abstraction, simultaneity and symbolization in drawings. <http://formafluens.io/>, 2017.
- [66] Josh Katz. You Draw It: Just How Bad Is the Drug Overdose Epidemic? *The New York Times*, April 2017.
- [67] Thomas Baudel. From information visualization to direct manipulation: Extending a generic visualization framework for the interactive editing of large datasets. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology, UIST '06*, pages 67–76, New York, NY, USA, October 2006. Association for Computing Machinery.
- [68] Bahador Saket, Hannah Kim, Eli T. Brown, and Alex Endert. Visualization by Demonstration: An Interaction Paradigm for Visual Data Exploration. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):331–340, January 2017.
- [69] Bahador Saket, Samuel Huron, Charles Perin, and Alex Endert. Investigating Direct Manipulation of Graphical Encodings as a Method for User Interaction. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):482–491, January 2020.
- [70] Kiran Gadhave, Jochen Görtler, Zach Cutler, Carolina Nobre, Oliver Deussen, Miriah Meyer, Jeff M. Phillips, and Alexander Lex. Predicting intent behind selections in scatterplot visualizations. *Information Visualization*, 20(4):207–228, October 2021.
- [71] Aspen Hopkins, Michael Correll, and Arvind Satyanarayan. VisuLint: Sketchy In Situ Annotations of Chart Construction Errors. In *Computer Graphics Forum (Proc. EuroVis)*, volume 39 of 3, pages 219–228, 2020.
- [72] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for Datasets. page 16, 2018.
- [73] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics (InfoVis)*, 18(12):2431–2440, 2012.
- [74] Christopher P. Barlett. Anonymously hurting others online: The effect of anonymity on cyberbullying frequency. - PsycNET. *Psychology of Popular Media Culture*, 4(2):70–79, 2015.
- [75] Janet Iwasa, Miriah Meyer, Alexander Lex, Jen Rogers, Ann (Hui) Liu, and Margot Riggi. SARS-CoV-2 Visualization and Annotation Project. <https://animationlab.utah.edu/cova>, 2020.
- [76] Yea-Seul Kim, Katharina Reinecke, and Jessica Hullman. Data Through Others’ Eyes: The Impact of Visualizing Others’ Expectations on Visualization Interpretation. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):760–769, January 2018.