The Noisy Work of Uncertainty Visualisation Research: A Review

Harriet Mason

4/29/24

Table of contents

1 Background

From entertainment choices to news articles to insurance plans, the modern citizen is so over run with information in every aspect of their life it can be overwhelming. In this overflow of information, tools that can effectively summarise information down into simple and clear ideas become more valuable. Information visualisations remain one of the most powerful tools for fast and reliable science communication.

There are many stages in our analysis that might benefit from the power of data visualisation, however this does not mean it is appropriately utilized or done with with effectiveness. (Vanderplas, Cook, and Hofmann (2020) has helpful advice here.) Visualization is an important step in exploratory data analysis and it is often utilised to **learn** what is important about a data set. The importance of data-driven discovery has been highlighted through the data sets Anscombe's quartet (Anscombe 1973) and the Datasaurus Dozen (Locke and D'Agostino McGowan 2018). Each of the pairwise plots in these data sets have the same summary statistics but strikingly different information, as revealed when visualised. Visualisations can **tell** us what is important about our data and **how** it might diverge from pre-conceived notions or what we expect.

Uncertainty visualisation is a relatively new field in research. Early papers that specifically reference "uncertainty visualisation" appear in the late 80s (Ibrekk and Morgan 1987), with geospatial information visualisation literature in the early 90s declaring this to be essential aspect of information display (MacEachren 1992; Carr, Olsen, and White 1992). Despite the new terminology visualisation of uncertainty has been present since the earliest times. For example, box plots or histograms can be considered to be displaying uncertainty in the sense of variability in observations sampled from a population distribution.

However, uncertainty visualisation takes many forms today, It can be considered to expand the application of data visualisation towards making better decisions, and research is concerned about how the construction of plots and information contained might affect the resulting decisions. With the abundance of publications it is timely to consider a review of the state-of-the-art. In fact, there have already been several reviews published.

Interestingly, these reviews often do not offer overarching rules for tried and tested uncertainty visualisation, but rather comment on the *difficulties* faced when trying to summarise the papers from this field. Kinkeldey, MacEachren, and Schiewe (2014) found most experiments on the methods for uncertainty visualisation evaluation to be ad hoc, with no commonly agreed upon methodology or formalisation and no greater goal of describing general principals. Hullman (2016) commented on the difficulty in taking overarching themes from uncertainty visualisation, as several conflated issues make it unclear if subjects did poorly in an experiment because they misunderstood a visualisation, because the question was misinterpreted, or because they used a specific heuristic. Spiegelhalter (2017) commented that different plots are good for different things, and disagreed with the goal of identifying a universal "best" plot for all people and

circumstances. Griethe and Schumann (2006) was unable to find common themes, but instead listed the findings and opinions of a collection of papers.

There are several reasons provided to explain the difficulty in generalising the uncertainty visualisation literature. Some suggested or agree with the idea that visualisation typologies should move away from data types, uncertainty categories, and representation types and towards "task-centred typologies" (Kinkeldey, MacEachren, and Schiewe 2014; Hullman 2016). Other papers indirectly hint at this by arguing that the choice of best visualisation is highly dependent on the specific goal (Griethe and Schumann 2006; Spiegelhalter 2017). Others contend that visualisations are highly dependent on the audience and there is no such thing as a "best" visualisation that will be accessible to all members (Kinkeldey, MacEachren, and Schiewe 2014; Spiegelhalter 2017). These concerns appear across all application areas, and an underlying thread is that the problem arises from a lack of a cohesive and encompassing **definition of uncertainty**.

The review provided here attempts to address these issues. We begin with understanding the definition problem, to obtain and understanding of the *purpose* of visualising uncertainty, and understand if the current visualisation tools are sufficient to achieve that purpose and if not, where there is need for new methodology. We also posit that "uncertainty visualisation" is too diffuse to be considered a field in itself, but rather should be dissected into smaller elements. The ideal role of this review to help synthesize this noisy field, by collecting together essential components, summarising different viewpoints the current literature, and provide a guide for potential research topics.

2 What is Uncertainty?

Uncertainty visualisation is directed towards communication - to the general public when disasters may hit, to decision makers in control centers, and a whole host of other public-facing tasks - needing an understanding of what it is when the term "uncertainty" is used.

Definitions in the literature appear seemingly view it to be encompassing and include everything the general public might consider to be uncertain. Some works (*XXX eg*) focus more narrowly on specific terms defined mathematically, such as probability, variance, error, or precision. Others (*XXX eg*) include broader loosely related elements, such as missing values. Several attempts to address this gap are made in the papers describing taxonomies.

2.1 Definitions in taxonomies

Taxonomies split uncertainty based on an endless stream of ever changing boundaries, such as whether the uncertainty is due to true randomness or a lack of knowledge (Spiegelhalter 2017; Hullman 2016; Walker et al. 2003), if the uncertainty is in the attribute, spatial elements, or temporal element of the data (Kinkeldey, MacEachren, and Schiewe 2014), whether the

uncertainty is scientific (e.g. error) or human (e.g. disagreement among parties) (Benjamin and Budescu 2018), if the uncertainty is random or systematic (Sanyal et al. 2009), statistical or bounded (Gschwandtnei et al. 2016; Olston and Mackinlay 2002), recorded as accuracy or precision (Griethe and Schumann 2006; Benjamin and Budescu 2018), which stage of the data analysis pipeline the uncertainty comes from (Walker et al. 2003), how quantifiable the uncertainty is (Spiegelhalter 2017; Walker et al. 2003), etc. In their own way, each of these taxonomies show an aspect of uncertainty that an author felt was important to differentiate. Walker et al. (2003) identified many common themes among this wide array of taxonomies and used them to design an encapsulating taxonomy that was boldly referred to as a "definition of uncertainty". While taxonomies can be helpful to map out the space of "things that we consider to be related to uncertainty", a taxonomy is *not* a precise definition that would allow statisticians to estimate the uncertainty in their projects. The ramifications of this is seen throughout the literature, as not understanding how to calculate uncertainty is one of the leading reasons cited by visualisation authors when explaining why they don't include it in their visualisation (Hullman 2020).

The task of providing an encapsulating mathematical definition of uncertainty is far beyond the scope of this work, however we will discuss an important but commonly misunderstood feature of uncertainty; its relationship to inference.

2.2 Defining uncertainty with respect to inference

What exactly is uncertainty, then? If we were to consider making this overarching definition, what would it need in order to be "encapsulating"? Well, let us consider what might *not* be considered uncertain in order to understand this concept a little better.

Otsuka (2023) spends the first chapter of his book discussing the place of descriptive statistics in the philosophy of the field and in doing so, highlights an interesting connection between inference and uncertainty. Descriptive statistics simply describe our sample as it is and summarises large data down into an easy to swallow format. Descriptive statistics are not seen as the primary goal of modern statistics, however this was not always the case. Around the 19th century in England, *positivism* was the popular philosophical approach to science (positivists included famous statisticians such as Francis Galton and Karl Pearson) and practitioners of the approach believed statistics ended with descriptive statistics as science must be based on actual experience and observations, therefore anything that refers to the unobservable (such as new observations or population statistics) is not true science (Otsuka 2023). By its very nature, descriptive statistics *cannot* be used to make inferences about the data because it simply exists to summaries the data, to use it to make statements about *new* data is incorrect usage. In order to make statements about population statistics, future values, or new observations we need to perform inference, which requires the assumption of the "uniformity of nature" (i.e. that unobserved phenomena should be similar to observed phenomena) (Otsuka 2023). This subtle shift, from descriptive statistics to inferential statistics was historically shunned *due to the fact it introduced the unknowable*, or in other words, uncertainty.

This philosophical understanding of statistics highlights that descriptive statistics do not have uncertainty, however some readers may disagree with that statement. Variance and probability are typically considered stand ins for "uncertainty", it is often how we choose to measure it. Additionally because probability and variance exist in descriptive statistics, descriptive statistics *must* have uncertainty. This is not necessarily true. Begg, Welsh, and Bratvold (2014) highlight that uncertainty is related to not knowing a specific value, while variability refers to the range of values a quantity can take at different locations, times or instances. Similar distinctions have been made by other authors, including Spiegelhalter (2017) who did not directly discuss variance vs uncertainty, but did differentiate between precise random events (such as the probability associated with a coin flip), and uncertainty (such as the estimated probability associated with a coin that might be biased). The variance of a sample variance can be calculated and know, therefore it is not uncertain but rather it a precise description of dispersion. If we were to discuss drawing a new observation, or estimating the true mean of a population *then* the the variance would become relevant in our discussions of uncertainty.

The idea that inference is related to uncertainty is not uncommon in the uncertainty literature, however it is often mentioned as a *task* or *goal* dependence. As the "goal" shapes every stage of an analysis, multiple authors in the uncertainty literature have commented on the need to consider quantifying and expressing uncertainty at every stage of a project (Kinkeldey, MacEachren, and Schiewe 2014; Hullman 2016; Refsgaard et al. 2007) while others highlight the importance of specific stages. Otsuka (2023) suggested that the process of observing data to perform statistics is largely dependent on our goals, because the process of boiling real world entities down into probabilistic objects (or "probabilistic kind" as he puts it) depends on the relationship we seek to identify with our data. Meng (2014) commented what is kept as data and what is tossed away is determined by the motivation of an analysis and what was previously noise can be shown to become signal depending on the the question we seek to answer. Kale, Kay, and Hullman (2019) discussed how the choices we make in our analysis impact our outcomes and introduce uncertainty. Carlin and Moreno-Betancur (2023) mentions that each research question can be can be categorised as descriptive, predictive, or causal, each of which has its own appropriate statistical methods and motivation agnostic model selection leads to statistical analysis that is devoid of meaning. Wallsten et al. (1997) argue that the best method for evaluating or combining subjective probabilities depends on the uncertainty the decision maker wants to represent and why it matters. Fischhoff and Davis (2014) looks at uncertainty visualisation for decision making decides that we should have different ways of communicating uncertainty based off what the user is supposed to do with it. It is clear that the distinction between noise and signal is important at every stage in an analysis, however combining the uncertainty from every stage is near impossible (Spiegelhalter 2017). This means that uncertainty is often discussed in isolation at multiple stages of a project, which further contributes to our fractured understanding of uncertainty.

With this understanding it becomes clear to see why uncertainty is tied to an endless string of examples in the data analysis pipeline. Uncertainty examples include imputed data, model selection, inherent randomness, biased sampling, etc, not because these things *are* uncertainty, but because they *create* uncertainty when we perform inference. Existing mathematical def-

inition of uncertainty (or related concepts) have been unable to unify this range of diverse concepts. For example, Thomson et al. (2005) suggests a mathematical formula for *examples* of uncertainty, Meng (2014) mathematically defined the variance introduced to a model by the array of model choices, and information theory tries to quantify uncertainty using the idea of entropy. These existing methods are not thorough enough for an analyst to understand what causes uncertainty, and quantify it for communication.

2.3 Mistakes made when we misunderstand uncertainty

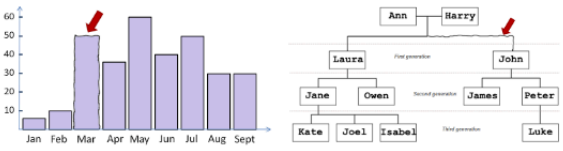
The previous section established that uncertainty is a by product of the assumptions we make when we perform inference, additionally, our quantification of uncertainty is largely task dependent, therefore the uncertainty associated with *one* distribution (which we can also view as a task or goal dependence). When uncertainty visualisation authors directly stare at this concept, they almost unanimously agree with these concepts, however when these details are pushed the the periphery of a problem, we start to get research that confuses itself. In these cases, the authors are more confused about what they are doing than the participants.

There are enough papers that fail to understand these basic principals of uncertainty that this entire paper could simply be a list with explanations. Instead, we have a handful of highlights.

Hofman, Goldstein, and Hullman (2020) performed an experiment where some participants were shown a sampling distribution of the mean, and others were shown a prediction interval around the mean, for the effectiveness of a particular treatment. The authors reported that the group that was shown the sampling distribution were willing to pay more and ascribed a higher likelihood to the belief that the treatment was more effective than a control, than the group that was shown a prediction interval. A prediction interval indicates the uncertainty associated with treatment *for a particular person* and the sampling interval shown the uncertainty associated with the *average effectiveness of the treatment*. It is already established practice in medicine to use a wide range of metrics to communicate different risks to best communicate the relevant information to a patient for a particular decision (Spiegelhalter 2017). The perceived importance of this work by Hofman, Goldstein, and Hullman (2020) hinges on a misunderstanding that the prediction and confidence interval *are* interchangeable despite performing inference on *very different* latent variables. This misunderstanding is so widespread that this finding, that shows the confidence and prediction intervals are *not* interchangeable, is considered groundbreaking enough to be published in CHI.

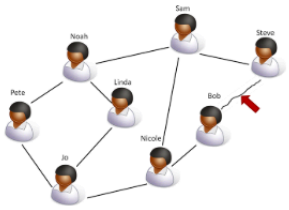
Boukhelifa et al. (2012) tried to quantify the strength of the intuitive connection between a line attribute called "sketchiness" and uncertainty. The authors of this study were aware that there was some kind of task dependence with respect to displaying uncertainty, so they tested the connection for multiple "tasks". The authors did not seem to understand the "task" dependence is likely a dependence on the inference for a particular statistic, so the "task" dependence was interpreted as *context* dependence. Figure 1 depicts the six scenarios that

participants were shown and asked to interpret what they believe the squiggly line indicates. The authors misunderstanding led to "sketchiness" being used to depict uncertainty in the categories of bar charts, in graphic networks, and in train lines, with little consideration or explanation as to *how* these things would be uncertainty. It seems a tad ridiculous for the creators of a rail network map to be "uncertain" about the existence of a train line, however we are unable to guess at what the authors of the paper believed the uncertainty would be. The most likely scenario is that the authors did not consider the specific estimate that the uncertainty was representing, which is what led to visualisations that do not make conceptual sense. Participants rightfully assumed the sketchiness therefore represented something else, such alternative options or simply ignored it.

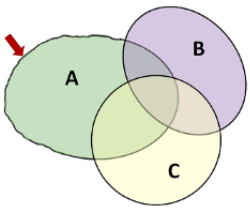


(a) (S1) Bar chart.

(b) (S2) Family tree.



(c) (S3) Social network.



(d) (S4) Venn diagram.

Fig. 6. The four abstract scenarios used in the study.

(a) (S5) Rail network map.

(b) (S6) Utility map.

Fig. 7. The two non-abstract scenarios used in the study.

Figure 1: The graphics displayed by Boukhelifa et al. (2012) to identify if there is an intuitive connection between sketchiness and uncertainty. These graphic were made without concern as to what (i.e. which statistic) in the image is supposed to be uncertain. This leads to the images being difficult to connect to uncertainty even if we assume tha that is what it is supposed to represent. For example, it is unclear how you can be uncertain about a train line (S5).

Many authors do not understand bias *is* uncertainty (something that has been proven mathematically by Meng (2014)) and whether we view uncertainty as bias or variance depends on

the resolution of our problem. L. M. K. Padilla et al. (2021) found that high uncertainty in the model estimates (calculated uncertainty) and low forecaster confidence (which is typically an expression of suspected bias in the model) both caused participants to have decreased confidence in their results and suggested modellers express both if they are relevant. Kale, Kay, and Hullman (2019) discussed the importance of communicating decisions made in the data analysis pipeline and being aware of the alternatives. While there is nothing wrong with the results of these papers, the fact that they were published shows some level of surprise by these results. This work indicates that there was a significant number of authors in the field were not already aware that choices introduced early in the data analysis pipeline create bias and therefore uncertainty in our final values. The surprise that both bias and variance contribute to final uncertainty shows that authors do not understand both will contribute to the *distance* of your estimate to the value you are drawing inference on (AKA error).

The misunderstanding that descriptive statistics also extends to visualisation. The exploration step of an analysis, which includes descriptive statistics, exploratory data visualisation, and unsupervised machine learning techniques, is performed without a prior hypothesis, however misunderstandings of this fact appears frequently in the literature. Griethe and Schumann (2006) commented that "if visualization is used as a means to explore a data volume or to communicate its contents the uncertainty has to be included". K. Potter et al. (2010) aimed to create a summary plot that "concisely presented data with uncertainty information" to create an exploratory visualisation tool that visualised uncertainty. Exploratory visualisations often differ from descriptive statistics because the explicit statistic we are drawing inference on is less explicit. A versatile visualisation such as a scatter plot allows for a viewer to consider several hypothesis at once, each of which may have its own associated uncertainty. Hullman and Gelman (2021) argue that there is no such thing as a "model-free" visualisation, therefore visualisation require robust visualisations of uncertainty as we are always performing inference. While we agree with this sentiment, it is clear that uncertainty *cannot* be defined without a specific motivating question, and therefore trying to include uncertainty in an exploratory visualisation stage (which by definition does not have a hypothesis) is not possible.

This confusion (or indifference) also creates many imprecisely named papers. Many papers will boast a title that claims to be about uncertainty visualisation, but simply depicts different visual representations of a PDF, however it takes reading the methodology to find this out (*Cite Gap 4: Organise uncertainty quantification*). While this is a minor by product of the definition issue discussed earlier, it does result in a literature that is needlessly difficult to navigate.

The reality is that a lot of uncertainty visualisation authors do not seem to have an intuitive understanding of uncertainties connection to inference, when inference is being performed, and how to design experiments that capture this relationship. This misunderstanding is not due to laziness or the fault of the authors, but rather is likely caused by the absence of a strict definition of uncertainty.The imprecise and confusing set of existing definitions are creating a field in which the authors themselves do not know what they are testing.

3 What is Uncertainty Visualisation?

If we simply built the idea of uncertainty visualisation up from "uncertainty" which we already know has a precarious definition we would understand the field of "uncertainty" visualisation must also be precarious. You cannot quantify something that is not defined, and you can't visualise something you cannot quantify. If we cannot quantify the full concept uncertainty in any meaningful way (other than with examples such as missing values or a PDF) then there is no such thing as an "uncertainty" visualisation. Unfortunately this problem is not quite that simple.

3.1 Definitions of "uncertainty visualisation"

Before discussing the issues facing uncertainty visualisation due to a poor concept *of* uncertainty visualisation, we want to establish two ideas that should be considered in conjunction with the lacking definition of uncertainty. The first is the sidestepping of the "uncertainty" definition problem by suggesting "uncertainty visualisation" is inherently different. Some authors may consider "uncertainty" to be poorly defined, but "uncertainty visualisation" to be well established. We will spend the next section explaining why that is not the case by discussing several taxonomies of uncertainty visualisation.

In order to visualise something, it needs to be quantifiable. In order to quantify something, it needs to be mathematically defined. The broad classes of what is considered "uncertainty" rarely have mathematical defintions, and therefore, cannot be meaningfully visualised. For these reasons simple quantifiable uncertainties, such as confidence intervals dominate the literature, while visualisations for more complicated sources of uncertainty, such as the effects of assumptions, imputed missing variables, and model choices remain elusive.

A large drive of the field of uncertainty visualisation, has been the idea that "uncertainty visualisation" (defined as one word) is somehow different to "uncertainty" "visualisation" (separately defined as two words). When "uncertainty visualisation" is defined as a single word we lose some nuance in our understanding of the driving force behind some issues, and problems with visualisation and problems with uncertainty become conflated. This appears in the literature as a mountain of seemingly conflicting statements about what is, or is not, an "uncertainty visualisation". For example Wilkinson (2005) mentions that popular graphics, such as pie charts and bar charts omit uncertainty, however at least one or both of these charts are used in a significant number of uncertainty visualisation experiments (Ibrekk and Morgan 1987; Olston and Mackinlay 2002; Zhao et al. 2023; Hofmann et al. 2012). Wickham and Hofmann (2011) suggests their product plot framework, which includes histograms, should have a way to measure uncertainty, but does not consider that a histogram is *already* a depiction of PMF and would already be considered an uncertainty visualisation were our statistic of interest a new oberservation from our data set. These conflicting ideas and lacking definition of what is means for a visualisation to be an "uncertainty visualisation" is a cornerstone of some clearly questionable results within this field.

Since the concept of "uncertainty visualisation" is harder to define than "uncertainty", "uncertainty visualisation" taxonomies are typically less thorough than taxonomies for "uncertainty". They also tend to blur the line between the mathematical elements of uncertainty and the visual depiction of those mathematical objects. While we will not focus on these taxonomies, it is still important to mention them and highlight what we can learn from their success' and failures.

Some taxonomies of uncertainty highlight how confusing most authors find the concept, and how difficult it is to articulate the rules of an uncertainty visualisation when uncertainty itself if not defined. Kristin Potter, Rosen, and Johnson (2012) organised several existing uncertainty visualisations into groups based on the dimensionality of the data (1D, 2D, 3D, and No Dimension) and the dimensionality of the uncertainty (Scalar, Vector, Tensor). However, because the term "PDF", a statistical object that describes a random variable that is typically a one dimensional function, is used to describe both the data and the uncertainty for all dimensions, it is hard to understand how this should work. Grewal, Goodwin, and Dwyer (2021) created a taxonomy that mapped uncertainty visualisations to some point in a 2D space defined by the "domain expertise" and "continuum of discreteness" (that scaled from "point estimate" to "continuous distribution"). This conceptualization could be considered a visual extension of the law of large numbers, because as a point estimate is a single sample, as the sample size gets larger the distribution will appear to be a continuous. However, assumes every point estimate is a sample from the same distribution (a mean is not technically a sample from a prediction interval, but it is a sample from the sampling distribution) and it conflates discreteness from a too small a number of observations, discreteness from a precision problem, discreteness because that is the true nature of the variable, and discreteness as a visualisation choice. Griethe and Schumann (2006) organised uncertainty visualisations into two cases (1) a hypothesis test was preformed to confirm the validity of the visualisation and (2) the visualisation has uncertainty depicted. This conceptualization conflates hypothesis testing and confidence intervals, and while they are related concepts, a hypothesis test is not necessarily uncertainty. Of course the authors of these papers did not intend for the taxonomy to be a complete description of uncertainty visualisations, however these distinctions that are often ignored are subtly important. (*Cite Gap 9: Other uncertainty vis taxonomies*)

There are a small handful of taxonomies that highlight some interesting features in the way uncertainty is visualised. Kinkeldey, MacEachren, and Schiewe (2014) categorised uncertainty according to whether or not it was: 1. Implicit (a sample) or explicit (a depiction of mass) 2. Intrinsic (alter existing symbols symbols to represent uncertainty) or extrinsic (add new objects to represent uncertainty) 3. visually integral or separable (the uncertainty can be separated from the data and read independently) 4. coincidence (uncertainty and data are represented in the same plot) or adjacent 5. static or dynamic (animated, interactive etc) Of the groups they identified, they actually only used (4), (2), and (5), left out (1) because most visualisations are explicit, and (3) corresponds to (1) in most cases (Kinkeldey, MacEachren, and Schiewe 2014). This taxonomy suggests interesting considerations for a visualisation of a distribution (even though it was not intended to be a visualisation of a distribution). A similar version of this taxonomy was presented by L. Padilla, Kay, and Hullman (2022) who

commented that visualisation can be organised into two categories, "graphical annotations of distributional properties" and "visual encodings of uncertainty" which functionally align with the intrinsic/extrinsic distinction. The first four categories boil down to the question "should uncertainty and signal be conveyed together as one variable or separately as two". This consideration will become more relevant later in this review.

Most of these taxonomies are created by observing existing uncertainty visualisations, which means they suffer from many of the same conceptual issues that the field has with uncertainty in general. Additionally, there does not seem to be anything these taxonomies offer that would not be better established by separate taxonomies for uncertainty and visualisation. The difference between "uncertainty visualisations" and "data visualisations" is not technically in the visual element, it is mathematical. Kinkeldey, MacEachren, and Schiewe (2014) almost acknowledges this in their own paper that discusses an uncertainty visualisation taxonomy when they claim "future typologies should take different categories of tasks into account (1) communication tasks (2) analytical tasks (3) exploratory task", a common typology for information visualisation in general. The process of understanding and estimating uncertainty requires knowledge of the data, the statistical methods used to make an estimate, and the assumptions of a model. Visualising the statistics that represent uncertainty should be no different than depicting the statistics that represent any other element of a graphic, this is something we will show in the following sections of this paper.

3.2 Why does the field of "uncertainty visualisation" exist at all?

It is unclear exactly *why* uncertainty visualisation is a field of interest. The two most common justifications are: 1) Uncertainty is fundamentally differently to other variables due to psychological heuristics involved in the interpretation of uncertainty (e.g. risk aversion). 2) Uncertainty is not of interest in of itself and is typically layered on at the end of a the visualisation pipeline, so uncertainty visualisation is the field of adding error information into already established graphics and *thus* uncertainty information is a high dimensional visualisation problem.
(*Cite Gap 10: Reason uncertainty visualisation is different*) Papers will often discuss uncertainty in relation to one of these motivating reasons, the evaluation experiments motivated by (1) often perform different visualisations of PDFs, while the papers motivated by (2) will focus on trying to impute uncertainty as error within one of the existing channels, such as colour (*Cite Gap 10*). These dual motivations that co-exist with uncertainty visualisation authors rarely mentioning the others, create confusion in the field as to its purpose. Kinkeldey, MacEachren, and Schiewe (2014) also identified this issue when they highlighted that the literature makes it unclear if uncertainty is a variable in of itself, something that should be interpreted with the main variable, or if it is metadata.

Uncertainty is noise, but most evaluation experiments measure it as signal. Uncertainty acts as a form of statistical "hedging" for signals found in an analysis. Interviews with experts in statistic back up this primary motivation, as ignoring uncertainty information of often

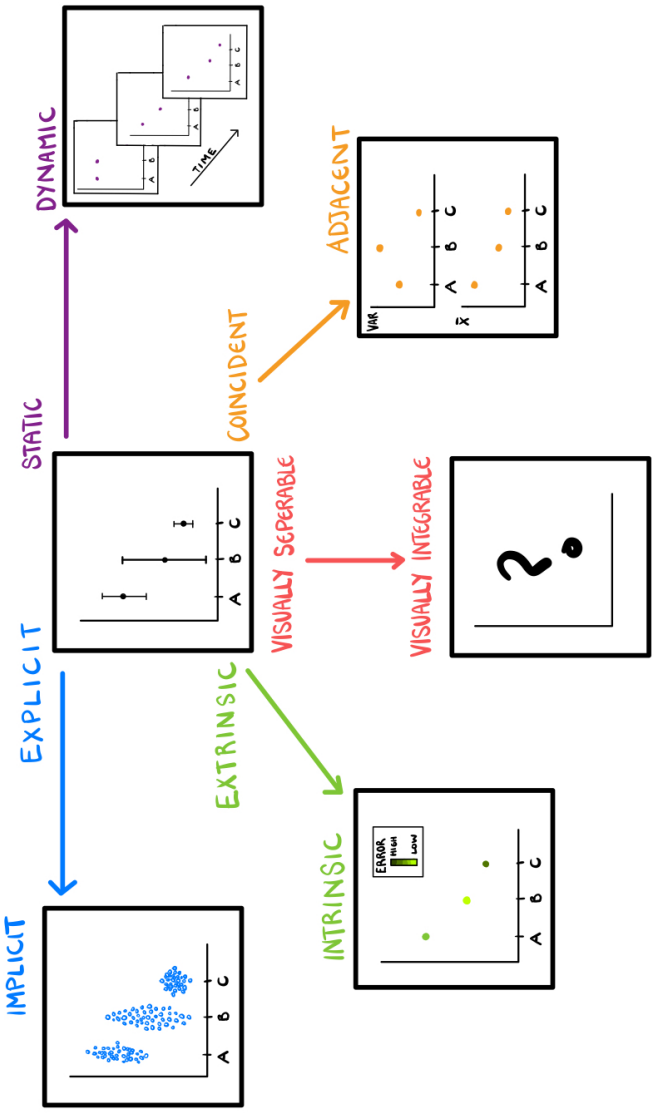


Figure 2: Visualisation of Kinkeldey, MacEachren, and Schiewe (2014) uncertainty visualisation taxonomy. It can be seen that most of the categories can be considered a change to the typical error bar chart. Many of the distinctions conflate changes of the input data and statistics calculated and with visual changes, highlighting how defining uncertainty and visualisation together conflates several elements of a graphic.

expressed as being similar to fraud or lying (Hullman 2020; Manski 2020). Noise and signal are not an inherent property of variance, or meta-data or the aspects of an analysis that are usually secondary considerations, it is a property of the data that *promotes* a message we are trying to infer and the data that *suppresses* it, if that message is invalid. When we ask the viewer of a plot to look at data and extract a value, we are asking them to perform inference on that value. There will be noise associated with that answer and that is uncertainty. If we ask direct question about some uncertainty metric, we have turned the uncertainty into signal because that is what the participants are drawing inference on.

This is likely the cause of the contradiction between uncertainty visualisation papers following identical results to normal information visualisation papers, and authors believing uncertainty is inherently *different* to normal visualisation. Each of the sections in this paper likely contribute in some way to this contradiction in conceptualisation of uncertainty visualisation and the results of the studies in the field. Uncertainty is poorly defined, therefore authors do not understand it is not an inherent property of the data and is actually related to a particular inference question, this incorrect view that uncertainty is always variance or probabilities leads to the class of "uncertainty visualisations" that depict different mathematical objects or visual features with little in the sense of consistent rules, this complicated class has caused visualisation authors to miss that they are simply repeating the established evaluation results from normal information visualisations. This situation is a bit of a mess.

If uncertainty is inherently different, we would expect evaluation studies of uncertainty visualisations to produce different results to similar evaluation studies on "normal" variables. This does not seem to be the case.

3.3 Mistakes made when we misunderstand uncertainty visualisation

In the previous section, we highlighted several problematic experimental designs or pieces of research that were the result of misunderstandings around the definition of uncertainty. Here, we will discuss the issues that arise due to a misunderstanding of "uncertainty visualisation". The two primary problems are the regular comparisons of mathematically incomparable plots, and reproducing results that are already widely understood in the larger information visualisation literature.

3.3.1 Information asymmetry in evaluated plots

Visualisation authors are almost unanimous in commenting that the "information" in two plots must be the same in order for the visual techniques to be compared (Cleveland and McGill 1984; Kinkeldey, MacEachren, and Schiewe 2014) (*Cite gap 12: Equal information*). However, what makes two visualisations equal in "information" is not consistent and is regularly mentioned without it being clear exactly what this information is. This is easily illustrated by the wide array of comments made by visualisation authors when they explain why two representations

are being compared in their evaluation study. Ibrekk and Morgan (1987) displayed 6 PDFs in their experiment because "formally equivalent representations are often not psychologically equivalent", but do not explain why what makes a representation "formally equivalent" or why only one representation was presented for the 95% CI, box plot, and CDF. Hofman, Goldstein, and Hullman (2020) comments that "theoretically" the sampling distribution of the mean and the prediction interval of a new observation are equal "so long as one knows the sample size", but does not recognise the several assumptions and statistical knowledge that is required to compare the two. (*Cite Gap 13: examples of what is defined as "information"*). Kinkeldey, MacEachren, and Schiewe (2014) adopts an existing definition also that suggests two graphics are informationally equivalent if all the information in one plot is inferable from the other and vice-versa, but adds that two plots are computationally equivalent if that information can be extracted from both plots with similar easy and speed. It is clear this lacking definition of "information" in a visual is reminiscent of our problems with the definition of "uncertainty".

Works such as *The Grammar of Graphics* attempt to outline exactly what information is contained within a plot, and in doing so allow plots to be considered a statistic (Vanderplas, Cook, and Hofmann 2020). This statistic that *The Grammar of Graphics* defines fundamentally a descriptive statistic. Figure 3, and contains the set of necessary and sufficient steps that are required to make a graphic, as defined in *The Grammar of Graphics* (Wilkinson 2005). It can be seen that the grammar assumes we have our data, performs a set of functions on that data, and produces a statistic. The grammar of graphics fundamentally describes the graphic as it is, not as it *would* be were our sample size to increase to infinity. As a matter of fact, the graphics assumption that we start from a data set that has an unchanging sample size is at odds with inference that typically involves assuming an infinitely large sample size. **?@fig-infintegraph** shows that as n approaches infinity, two graphics that are very similar according to the *The Grammar of Graphics* can diverge into different graphics, while two graphics that are very different, can converge into the same graphic. *The Grammar of Graphics* itself does not recognize this issue, as uncertainty is referred to as "semantics" and is ultimately hand waved away. This does not mean describing a graphic using inferential statistics is completely unheard of. Some graphical descriptors, such as Product Plots (Wickham and Hofmann 2011) describe graphics using the underlying distribution, however the framework is not a substitute for *The Grammar of Graphics*. Other authors have noticed difficulties with inferential statistics and and created an addition to the grammar that should make depicting a distribution easier (Pu and Kay 2020; **Kay2023?**). Unfortunately, this addition falls short as it does not recognize *why* latent distributions are difficult to describe using *The Grammar of Graphics* and carries over many of the same problems.

This does not mean *The Grammar of Graphics* should be abandoned when we visualize uncertainty, rather, these difficulties are something we should keep in mind when

Tierney and Cook (2023) generates a visualisation for missing values, however because it fits nicely into the grammar of graphics framework, it is inherrently a descriptive statistic.

If uncertainty is fundamentally a by-product of inference, then the visualisations as described by *The Grammar of Graphics*

Most uncertainty visualisation experiments are written with a seeming indifference to the grammar of Graphics, despite it being necessary to untangle experimental results. The primary reason for this seems to be the absent definition of uncertainty. Uncertainty is fundamentally related to inference and descriptive statistics cannot have uncertainty.

Since two plots cannot be compared on their visual elements if they contain different information (because you will never be able to identify if the difference is because of informational asymmetry or visual processing) we could understand two plots as being the same if their "varset" steps are identical. However because uncertainty is not defined mathematically, we cannot reliably add it in at the "statistic" stage. , such as resampling do not quite fit within the framework. For example, if we have one set of data and two graphics, where one graphic represents the data as a series of points, and the the other depicts a smoothed density using a line. These two plots differ in *both* the statistic and geometry stages of the grammar of graphics, however as n (the number of observations in our sample) gets larger, these two plots will become visually indistinguishable (depending on the collision modifier used). This result aligns nicely with large sample theory, because as n increases the data and mass should become mathematically indistinguishable, and many uncertainty visualisation experiments notice an interaction between sample size and graphic effectiveness so this result is likely also practically true (Kale et al. 2018; Newburger, Correll, and Elmqvist 2022; Hofmann et al. 2012) but despite the increasing similarity in what is actually depicted in the graphic, the sample and mass visualisations would remain different visualisations within the grammar of graphics framework. Additionally, authors rarely attribute this visual version of large sample theory to results where it is clearly applicable. Kale et al. (2018) found that users were more sensitive to the underlying trend when shown the HOPs plot over the static outcomes plot, this effect went away when the speed of the animation increased, indicating that performance difference might be due to overplotting in the static case and a smaller sample size in the static condition might produce the same result. The issue of blurred elements of the grammar of graphics in the case of distribution visualisations does not end here. Visualisations that depict mass such as a confidence interval, a boxplot, a PDF, a letterbox plot, a violin plot, a histogram, a quantile dot plot, etc, all show some version of the mass of a variable at different resolutions, but whether these graphics differ in something as low down in the grammar pipeline as the variable stage, or later at the geometry stage is unclear.

Alternative information consideration within well defined frameworks may also be worthwhile to consider. Wu et al. (2023) tried to eliminate information asymmetry by essentially looking at a visualisation and identifying what information can be extracted and then using a "rational agent benchmark" to determine how a rational person would use that information for a decision making task. While this solution is interesting, it may have different results depending on who is applying it as they may use different methods or notice different elements when extracting information from a graphic. The mathematically defined concept of sufficient statistics is also scarcely mentioned likely because it is specific to a particular ground truth statistic, however uncertainty *is* specific to a particular statistic, so it could be a useful avenue for understanding when two plots differ in their ability to answer a particular question.

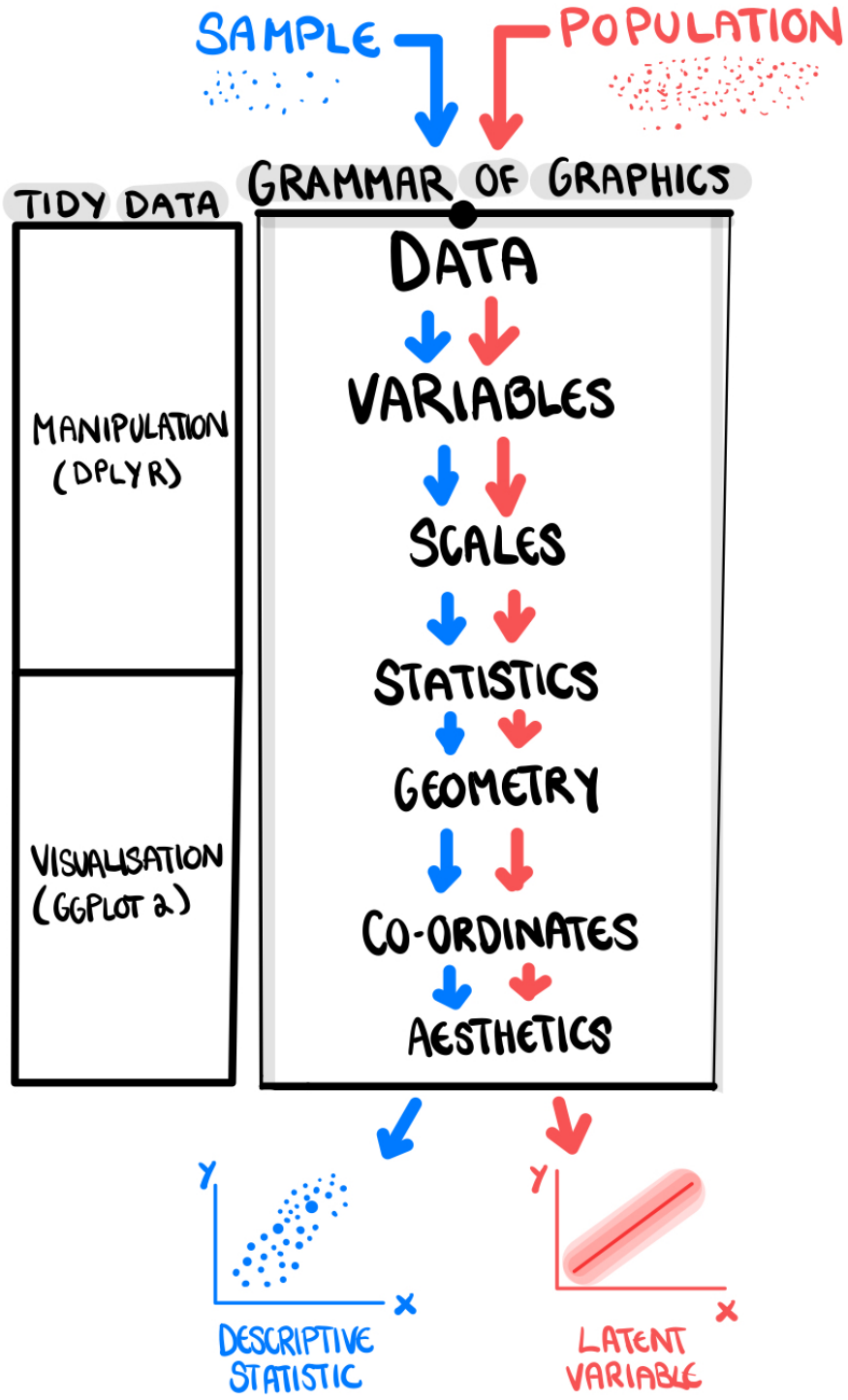


Figure 3: (The grammar of graphics data analysis pipeline from Wilkinson (2005) written in black, with the tidy data equivalent written in blue. The product plot framework and the grammar of probabilistic graphics frameworks are in green. It can be seen that the grammar of graphics provides a set of instructions that create a graphic, however the inferential power of this graphic is not considered and relies on many additional assumptions that are not defined by the grammar.

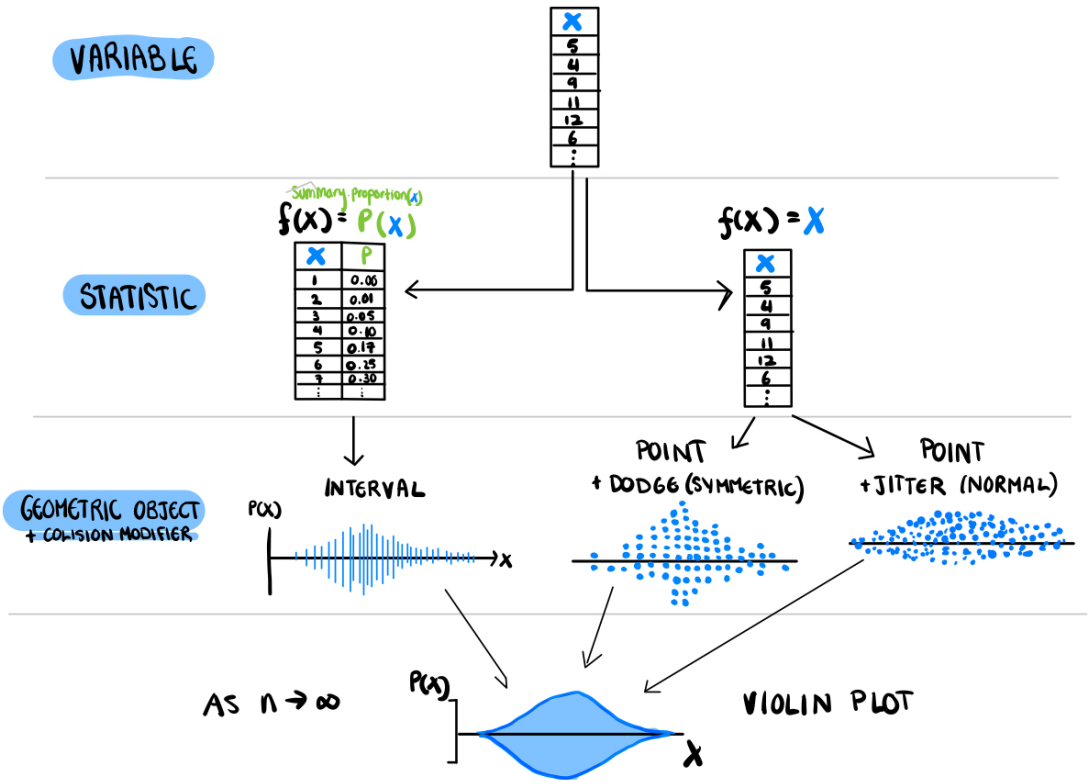


Figure 4: This graphic shows three simple graphics made from the same data set that diverge according to the grammar of graphics. Despite being distinct graphics from a stage as low down as the statistic (or data manipulation in the tidy data framework), as our sample size increases to infinity, the grahics converge to the same distribution and become visually indistinguishable.

Ultimately, this all points to confusion as to when two visualisation contain information asymmetry, a problem that becomes more pronounced in uncertainty visualisation. Ibrekk and Morgan (1987) compared a 6 visualisations of a PDF, a box plot, a CDF and a mean with a confidence interval on the basis that all of them were "uncertainty visualisations". They found that people are better at extracting the mean from a plot when they are shown a plot that contains a mean with a confidence interval than when they are shown a PDF, a box plot, or a CDF (where the mean could not be read off the plot). Hofman, Goldstein, and Hullman (2020) and Zhang et al. (2022) compared prediction and sampling distributions because they are both "uncertainty" that is typically depicted around the mean. They found that people are better at answering questions about a prediction interval when shown a prediction interval instead of a sampling distribution. Hullman, Resnick, and Adar (2015) compared the static error bars and violin plots of the marginal distributions of two variables (A and B) to an animated plot that depicted a single outcome of the joint distribution of A and B in each frame of an animation, a comparison that was justified because all plots were "uncertainty visualisations". They found that the plot that depicted outcomes from $P(A, B)$ where the distance between the two outcomes was equivalent to $a - b$ was better at answering the question "What is $P(B < A)$" than violin or error bar plots depicting $P(A)$ and $P(B)$ separately. (*Cite Gap 14: Examples of mathematically obvious results*).

This collection of examples starts to paint a pretty clear picture. Visualisations with rather shocking information asymmetry are regularly compared because they are both "uncertainty visualisations". This results in a series of experiments where the visual aspects of the graphic are not even required to anticipate the experimental results. This issue likely results in the lacking mathematical definition preventing us from defining and quantifying "uncertainty" mathematically and then including this quantification in a well defined framework such as the grammar of graphics.

3.3.2 Repeating perceptual task experiments

Given the previous section, one might consider problems in uncertainty visualisation to be fixed if we could restrict evaluation experiments to cases where the information in each graphic is almost identical. Unfortunately this still often results in uninteresting or obvious results.

If two graphics are visually different but identical in the information they contain, they must differ in how that information is depicted. Once we have a mathematical expression of uncertainty, the visualisation of that uncertainty is theoretically identical to the visualisation process of any other variable. For simple tasks such as value extraction, there is a hierarchy to perceptual tasks where extracting visual information in some forms is easier than others. The hierarchy was originally established 40 years ago by Cleveland and McGill (1984), below is an updated version summarised by Vanderplas, Cook, and Hofmann (2020):

1) Position along a common scale.
2) Position along a non-aligned scale.

3) Length, direction, angle, slope
4) Area
5) Volume, density, curvature
6) Shading, colour saturation, colour hue
7) Discriminable shape
8) Indiscriminable shape

This hierarchy is a good general rule, however it can change from person to person (Davis et al. 2022) Additionally, there are other graphical rules to consider such as gestalt principles, broader methods of extraction, and attention principles (Vanderplas, Cook, and Hofmann 2020). These established visualisation concepts allow us to anticipate the ease with which certain pieces of information will be extracted from a plot. We can use these concepts to understand the computational complexity of a graphic. A bizarre feature of the uncertainty visualisation literature is that it does not work to build upon these existing principles or identify the ways in which uncertainty visualisations may diverge from these rules. These building block concepts of visualisation are seldom mentioned.

It is difficult to find examples of uncertainty visualisation experiments where the plots do contain the same information, however when they do, the results align with existing information visualisation research. Technically, a PDF and a mean with confidence intervals both have enough information to extract the mean of the distribution, however they both have a very different computational cost. To extract the mean using a PDF, a participant would need to identify the point along the x axis that splits the area under the curve in half. If a participant is provided with a mean with a confidence interval, extracting the average is a simple task of reading the position on an aligned scale. Ibrekk and Morgan (1987) found that when asking participants for the "best estimate" (which they thought should be interpreted as the mean of the distribution) of a skewed distribution, participants provided the mean when given a mean with confidence intervals and the mode when given a PDF (Ibrekk and Morgan 1987). Similar results to this occur over and over again in the uncertainty visualisation literature. Gschwandtnei et al. (2016) found that visualisations where the start time of an interval could literally be read off the plot (error bars, centred error bars, and ambiguation) performed better than the plots (accumulated probability, gradient, and violin) where the start time involved some guesswork because the drop off was gradual. Cheong et al. (2016) found that participants were better at answering questions when they were explicitly given the relevant probability in text rather than when they needed to read it off a map. (*Cite Gap 15: Examples of replicated perceptual task experiments*).

These results show that uncertainty is not technically different to any other variable. When trying to anticipate the results of these studies, we can use the same principles of information equivalence and difficulty of relevant mental tasks to understand which plots will outperform others. This leads us to wonder… why is uncertainty visualisation a field at all?

It is clear that authors intuitively *feel* that uncertainty is different even though this difference has not been captured in existing uncertainty visualisation experiments. Some authors cited

uncertainty was special because of its psychological qualities, but we have yet to see any evidence of that here. Kinkeldey, MacEachren, and Schiewe (2014) also noticed that it is not clear whether or not uncertainty is just another variable, as many value extraction papers treats it as such in their experimental design and uncertainty may need to be in a variable class of its own. Hullman (2016) commented that it is straightforward to show a value but it is much more complex to show uncertainty, but did not explain *why* uncertainty information faces more difficulty. After all, confidence intervals, minimums, maximums, variance, and any other statistic that can be considered related to "uncertainty" is just that, a statistic. If we want to visualise the relationship between a temperature estimate and our forecasted variance on that estimate, can we not just use a scatter plot, as we would any other pair of variables? There is nothing special about these statistics that should differentiate them from a mean or median and imply their visualisation should be "special" in a way that would warrant its own field. Especially when the core takeaways from uncertainty evaluation experiments do not seem to differ at all from the established results in information visualisation. So, what is it that makes authors believe uncertainty is different?

4 Issues with measuring uncertainty as noise, not signal

The previous section highlights a unique and fascinating problem faced by uncertainty visualisation. If asking direct questions about uncertainty causes us to treat it as a signal, how do we evaluate uncertainty as *noise*?

Uncertainty, acting as it is intended, is transparency. There are many secondary benefits that come with this improved transparency, such as better decisions, more trust in the results and more confidence in the authors. These secondary benefits are, however, *not* the immediate goal of uncertainty. The following sections will discuss the issues and limitations in measuring uncertainty through these secondary metrics and provide suggestions as to how future studies should consider measuring uncertainty.

4.1 Issues with the current methods of measuring uncertainty

Uncertainty visualisation papers can be organised according to the *goal* of the experiment. Evaluation experiments are the standard rule for visualisations because the human brain is not as reliable as mathematical calculation. Therefore, user studies often aim to assess the limitations, biases, and heuristics of our mental calculator so that we can better understand the problems we may encounter when we plot our data. This is not to say any paper that suggests a visualisation without an evaluation experiment is completely lacking in justification, and there are many papers that suggest a novel visualisation without an evaluation study. Sometimes these papers are a preliminary step in finding a solution for common problems and intend to evaluate the visualisation in later work. The reasoning for this is obvious. There are often heuristics and biases that are not obvious to us when designing visualisation. Additionally,

these heuristics and biases can change depending on the larger scope of the graphic and the population we are communicating with (Spiegelhalter 2017; Kinkeldey, MacEachren, and Schiewe 2014).

The overarching goal of most uncertainty visualisation evaluation papers belong to one of: performance (how effectively a participant can extract information from a plot), interpretation and semantics (the ease with which a particular visualisation is associated with uncertainty), and quality of user experience (if the participants liked the plot or not)(Hullman et al. 2019). Hullman et al. (2019) found that the majority of papers evaluate visualisations on performance (approximately 65% of the papers they surveyed) or interpretation and semantics (approximately 17%) and both of these evaluation goals will run into problems because of their conceptualisation of uncertainty.

4.1.1 Performance

There are many metrics experimenters use to evaluate the performance of a visualisation. Each metric, such as can use accuracy, decision quality, confidence, trust and others each have their own issues which prevent them from accurately capturing the effects of uncertainty in a visualisation.

By far the most common metric used is accuracy, which is used by approximately 36% of evaluation studies (Hullman et al. 2019). In previous sections, we pointed out issues that arise when accuracy (of extracting probability or variance) is used as the primary metric in evaluating uncertainty visualisations, the work simply replicates known concepts in information visualisation. Subjective metrics, such as how much participants liked the aesthetic representation of a visualisation, or metrics with an unclear motivation or use, such as how memorable the information in a plot is or how much a user interacted with a plot, are ignored in this section. Below we will explain in detail why experiments that try to measure uncertainty using methods such as decision making, trust, or other types of questions, also fail to capture the effects of visualising uncertainty in a plot.

4.1.1.1 Decision making and risk aversion

Decision making tasks are often described by authors as a more realistic version how plots are used in practice. While we do not disagree with this statement, exactly *why* these tasks are different to value extraction tasks is never explained. While the task may be more similar to how plots are actually used, that does not necessarily mean it is a better *experimental* environment for evaluating plots. Let us consider this in more detail. A normal value extraction task involves:

1) correctly interpret the question
2) extract the specified value
3) report the value

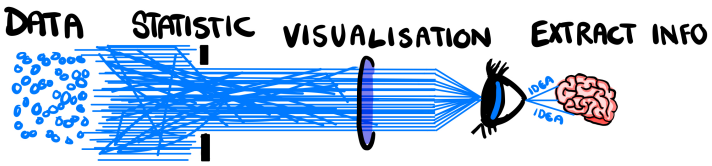


Figure 5: Visualisation of the process of transforming raw data to a visualisation insight. Summarising the data using common statistics removes information from our data. Once the data is in a format that represents the varibales of interest, organising that information in a visualisation makes it easier to extract insights from these statistics and convert them to insights. If information is removed at the statistics stage, it cannot be added back in at the visualisation stage, as it is meerly an efficient technique to organise and present information to enable a large range of insights.

On the other hand, a decision making task using an uncertainty visualisation involves:

1) Correctly interpret the question
2) Use some risk utility function to set a threshold for an "acceptable level of risk"
3) Extract the value for from the plot that will be compared to the threshold
4) Make a judgement based off that value

The key aspect that separates a decision making task from a value extraction task, for uncertainty visualisations, is the inclusion of a utility function. While this might seem like a small change, it actually completely warps the experiment in unexpected.

The first issue is that individuals will all have their own individual risk utility function, that is, how much they want to avoid or engage with risks or uncertainty. Therefore decision making experiments that have an additional layer of noise that value extraction experiments do not. We cannot be sure if participants answered differently from the ground truth because a visualisation was difficult to read *or* because the participants risk utility function did not align with the one set by the authors. Several authors have offered solutions to this issue, however the problem is deeper than any of them seem to realise.

Hullman (2016) suggested providing a utility framework for each experiment, to instruct the participants in how to account for the uncertainty information. This is a method that seems to have been adopted by Fernandes et al. (2018) who describe the following scenario to participants who are trying to maximise the coins in their experiment: > Subjects gain coins for every minute they are able to continue an activity that is valuable to them (e.g., watching TV at home) before going to the bus stop, and gain a bonus for arriving at their intended

destination early. Subjects incur a coin penalty for time spent waiting at a stop for a bus to arrive.

Any payment scheme that is incorporated into an experiment will also implicitly set up a utility framework, as it is assumed participants will try to maximise their payments. In tasks such as this, the ground truth of each question is typically selected to be the value a ration agent would select and visualisations are evaluated on the basis of how far the participants responses are from that ground truth. The problem with this, is that all the complexity of the question is in *working out* the best response, and it has little to do with the visualisation itself. The visualisation aspect of these studies becomes a value extraction experiment.

Cheong et al. (2016) tested multiple different visual representations of uncertainty for representing the likelihood of a house being burned down based on its location. Their payment scheme, which paid out \$0.10 for a correct choice (i.e. staying when the house was not burned down or leaving when the house was burned down) and 0 for an incorrect choice (i.e.leaving when the house didn't burn down or staying when the house burned down), meant participants were incentivised to base their entire leave/stay decision on whether or not the likelihood of a fire at their house is above or below 50%. If the participants *did* correctly identify the optimum strategy and answer accordingly, the "decision making experiment" was actually *just a value extraction experiment*. The decision making aspect ceased to matter, all that mattered was identifying if the probability of a fire was greater than 50%. Interestingly, despite this very obvious and simple tactic to maximize payout from the experiment, it seems like many participants did not adopt it, instead acting much more warily as though they were considering whether or not they would *actually* evacuate in the event of a fire. This means it is likely that introducing a payment scheme or a utility function creates too much mental labour for the participants to answer "correctly", and they would rather stick with their own risk utility function that feels *natural* than go through the mental labour of adopting another. Setting up an experiment that requires participants to do laborious calculations or strenuous mental effort for a small marginal benefit is unlikely to result in successful participant responses because it is not in the "spirit" of visualisation. (*Cite Gap 16: Examples of value extraction decision making tasks*)

It is clear that audiences primarily interact with visualisations through their "System 1" brain (Daniel 2017). This aligns with other authors comments that visualisation is primarily about "gists" [Spiegelhalter2017]. In this sense, asking participants a question that requires them to shift to their "System 2" is entirely contrary to *why* we use visualisation. The exact reasons we describe visualisation to be so powerful, that they are able to communicate complicated information fast and efficiently *is a product of the system 1 brain* (Daniel 2017), however this means displaying information using visualisations have the same weakness' as system 1 thinking, and visualisation authors need to be aware of this. Asking participants to answer questions that require complicated calculations does not reflect how visualisations are used, and will likely be ignored by the participants. While we agree with Hullman (2016) who suggested that what determines an appropriate ground truth is largely a philosophical exercise, we differ

in the sense that we do not think it should be done, lest you create a needlessly complicated value extraction task.

Additionally, several papers mention "risk-aversion" in participant decisions as a negative by-product of a particular visualisation design, however that displays a failure to understand that risk aversion is *not* a mistake, but rather it *reveals the utility of decreased uncertainty*. It may be interesting to investigate why a particular visualisation elicits increased risk-aversion, for example is it due to poorer estimates or increased awareness of negative outcomes, however simply reporting this as "risk aversion" conflates these factors. The work that advocates for transparency in risk-communication gives you whiplash when it's subtext argues that transparency leads to "incorrect" conclusions. There have been other discussions on the appropriateness of discussing risk-aversion as a bias (Vranas 2000), but few have made the point that if uncertainty holds some *value* there is not technically a "right" decision at all.

The final issue with decision making tasks is that calculating "optimum" choices is not a task we use uncertainty for, it is a task that utilises *risk*. Risk and uncertainty are slightly different, as risk is known probabilities and uncertainty is unknown probabilities, so uncertainty is what you get when you cannot accurately define risk (Spiegelhalter 2017). Communicating risk allows people to weigh up options and make an optimum decision, uncertainty hedges information to let people know there may not even be an optimum strategy. In this sense, many "uncertainty" decision making experiments are actually just risk decision making experiments. Not only this, but the inclusion of uncertainty information should *not* impact the choices of a rational agent. A well behaved rational agent *should* ignore uncertainty information, so long as the estimate provided is still the best case. This point seems to be lost on many uncertainty visualisation authors. Zhao et al. (2023) displayed a model's prediction and its estimated uncertainty and asked participants if they wanted to submit the model estimate as their answer, or make their own prediction. At no point did the authors seem to realise a rational agent would accept the model estimate every time, no matter how uncertain the information was, and participants were actually disqualified from the study for taking this approach to answering the questions. (*Cite Gap 17: Examples of studies where the correct choice was to ignore the uncertainty information*)

This leaves us with several issues in decision making experiments. They are noisy because of participants individual utility functions and when that is not the case they are thinly veiled and over complicated value extraction experiments. Additionally, authors don't even seem to agree how participants should incorporate uncertainty information, with some authors designing complicated utility functions, and others not realising the best response to their experiment is to ignore uncertainty information entirely. For these reasons, uncertainty visualisation experiments should either be designed with these caveats in mind and be designed to untangle why a particular visualisation will elicit a specific decision, or the field should steer away using decision quality and related metrics as a basis for evaluating visualisation performance.

4.1.1.2 Trust and confidence

Authors also often measure the impact of visualising uncertainty on participants trust in an estimate. Trust is a by product of displaying uncertainty rather than the goal of it, and viewing the relationship in the converse direction can lead to misguided research.

If the purpose of displaying uncertainty information is to appropriately hedge a signal with noise, then it should be assumed that trust is only related to uncertainty communication through increased transparency and honesty. Considering trust, and not transparency, as the metric of importance in uncertainty communication can lead to a questionable subtext that argues against transparency, something that has been noticed by several other authors [Spiegelhalter (2017); ONeill2018]. Hullman (2020) found that author simultaneously argued that failing to visualise uncertainty was akin to fraud, but also many avoided uncertainty visualisation because they didn't want their work to come across as "untrustworthy". These authors are optimising *trust* rather than *transparency*, which means they opt to leave out uncertainty information when it does exactly what it is supposed to, decrease certainty in conclusions.

Science communication should be primarily concerned with accuracy. Setting trust and risk-aversion as the variables of interest implicitly encourages statisticians to set trust and risk-aversion as the primary goals of communication. The issue of trust being divorced from trustworthiness has been commented on by other authors (O'Neill 2018), however the issue still persists in the uncertainty visualisation literature. Zhao et al. (2023) displayed a several variations of a visualisation of a model prediction and its uncertainty and took participants using the model prediction as a sign of trust. They reported that visualising uncertainty information caused participants to trust the model in the low variance case, but the results in the high variance case were inconclusive. The discussion made it clear the authors thought the uncertainty information should make the visualisation more trustworthy, but conflating trust and the use of a prediction implied uncertainty information should somehow influence participants to use their own prediction, even though a prediction being uncertainty does not necessarily mean it is incorrect. Despite this, the authors seemed to assume that the uncertainty information *should* have an influence on that, showing they had not deeply considered *how* uncertainty information should influence the choices of the participants. (*Cite Gap 18: examples of studies where authors measure trust*)

A similar measure to trust is using "confidence" in an extracted value or a decision. Interestingly, "confidence" is also used to try and capture the clarity of a message in a normal visualisation. Confidence cannot simultaneously be a measure of clarity of visualisation *and* a way to capture the uncertainty expressed in a visualisation.

4.1.1.3 Other (questionable) attempts to capture uncertainty

There is also a swath of studies that are aware a question that boils down to a value extraction experiment, or a question that should be answered by literally ignoring uncertainty information is not what we want when we consider uncertainty visualisations. These papers often try to ask a question that should utilise both the uncertainty and signal in the response, however this is

rarely what actually occurs. This method typically results in in cryptic or confusing questions that create a large amount of noise on the interpretation side of the analysis (Hullman 2016).

Some authors opt for asking slightly vague questions that imply a use of uncertainty, but compare it to a ground truth that is very specific. Ibrekk and Morgan (1987) asked participants for the "best estimate" which was evaluated in accuracy by comparing it to the mean, however the "best estimate" depends on the loss function we are using, and a loss function of minimised error was not implied by the question. Hofmann et al. (2012) showed two distributions in 20 different visualisations (a line-up protocol) using a jittered sample, a density plot, a histogram, and a box plot and asked participants. Participants were asked to report in which of the plots was "the blue group furthest to the right" The experiment set up is shown in Figure 6. The participants answers were then compared to a ground truth where the correct plot had a blue distribution with a right shifted mean. By comparing the results to a ground truth statistic and marking participants as "wrong" or "right", the error from the participants that had an alternative interpretation to the concept of "furthest right" was conflated with the error from a the visualisation choice. These papers make it unclear if the participants got the answers wrong because they misunderstood the question or because of something related to the plot. Therefore, this method leads to inconclusive results about the plot design, and is not advised.

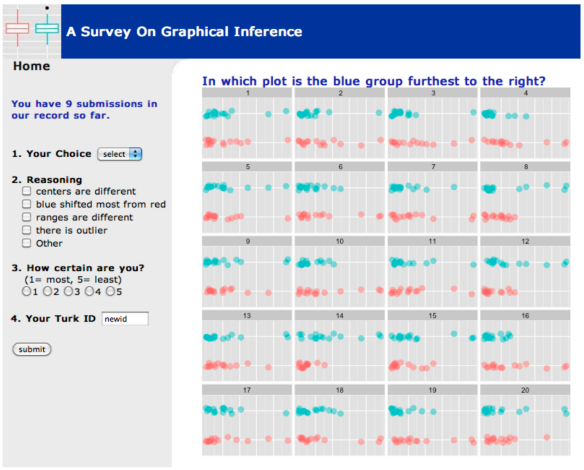


Figure 6: This shows the user interface for the experiment performed by Hofmann et al. (2012). The question of "furthest to the right" is open to interpretation.

Another method used by authors is to ask a deterministic question about a random event. L. M. Padilla, Ruginski, and Creem-Regehr (2017) provided participants with a visualisation of the cone of uncertainty and asked then to "decide which oil rig will receive more damage based on the depicted forecast of the hurricane path". The cone of uncertainty provides a 60% confidence interval for the location of the eye of a hurricane, which allows us to know the area where the eye of the storm will go, it does not given any information about the intensity of a storm, the size of a storm, or even if a location will be hit. This inclusion of

determinism seems to cause the authors to stumble themselves, as they are not consistent with their assumptions. In their first experiment L. M. Padilla, Ruginski, and Creem-Regehr (2017) indicated the correct answer was to assume that the storm was equally intense no matter how far from the centre of the distribution an oil rig was, however answering their third experiment correctly hinged on assuming the intensity of the storm at a particular point (which in this experiment they phrased as damage) *does* change in intensity as you move away from the centre of the distribution. Given these conflicting assumptions, it is unclear how the participants were supposed to adjust the probabilistic path information to answer a deterministic question about which oil rig would receive the most damage. Other authors have commented on the complexity of communicating hurricane risk because the path, storm surge and wind speed are all important and cannot be ignored (Spiegelhalter 2017). The flip side of this is asking participants for a deterministic answer to a probabilistic question. Correll and Gleicher (2014) asked participants "how likely is candidate B to win the election?" when the two distributions indicated voter preference. Participants were not able to answer the question about likelihood in term of probability, but were instead given seven options from 1=Outcome will be most in favour of A to 7=Outcome will be most in favour of B. The ground truth statistic for this question was a scalar multiple of Cohen's d, indicating participants were supposed to incorporate uncertainty information using a very specific formula that was likely unknown to them but assumed to be used implicitly.

The final method used by authors is to just explicitly ask about uncertainty and signal information separately. Sanyal et al. (2009) mapped uncertainty to dots and signal to a 3D surface and asked participants to identify areas of high and low signal and high and low uncertainty. Participants were not asked to combine that information in any way, and the signal and the noise were treated as separate variables. Correll and Gleicher (2014) asked participants to separately extract the mean and variance from four uncertainty visualisations. These methods explicitly view the uncertainty and signal as two separate variables that should be extracted from a plot, and not two variables that should be interpreted together. Even viewing these questions as a routine check to make sure the signal information isn't impacted by the uncertainty is counter intuitive, because the whole point *of* the uncertainty is to impact the signal information.

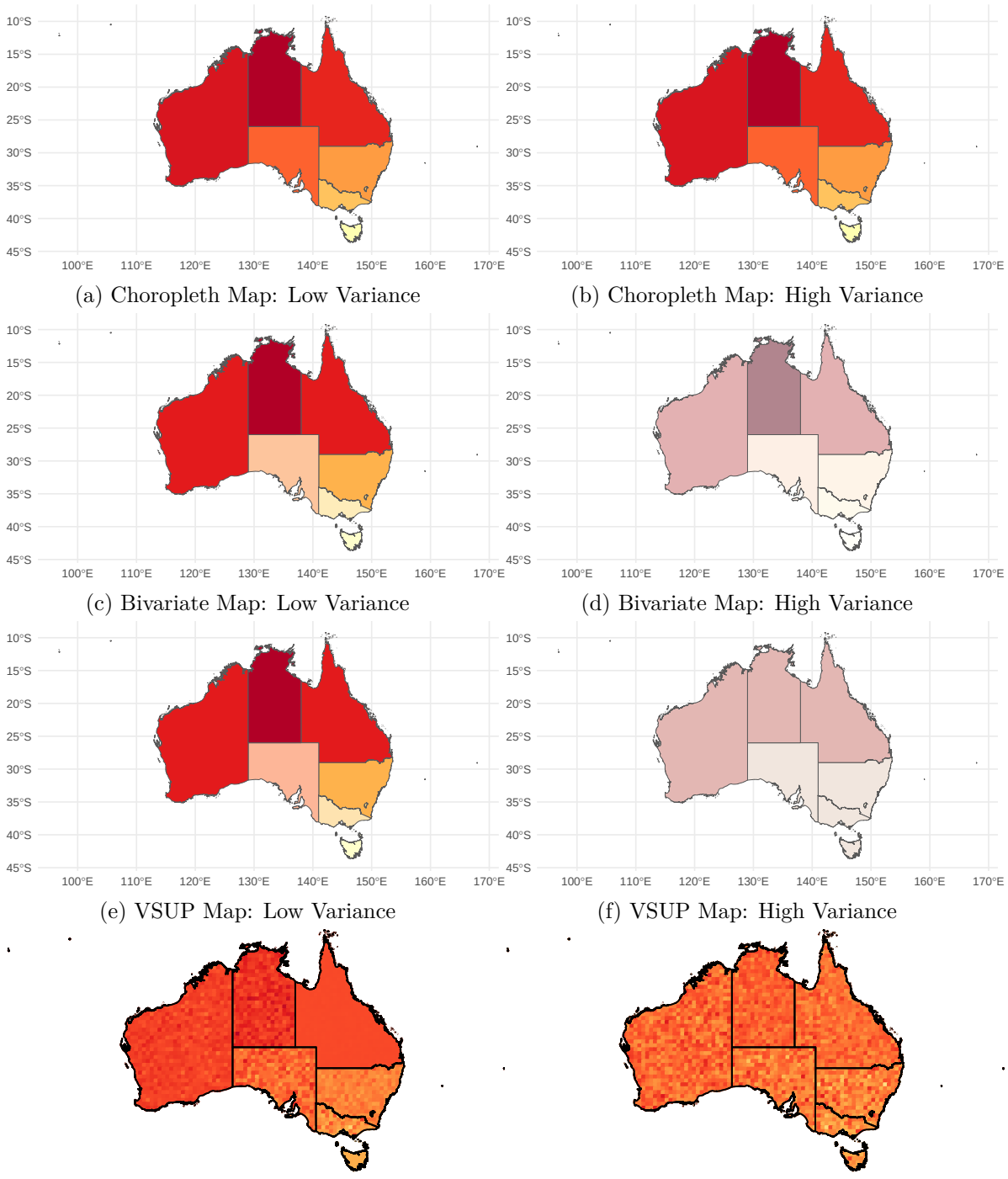
These examples are a bit complete mishmash of methods, however they point to a larger issue that goes beyond decision making, trust and confidence experiments. Authors *have no idea* how to evaluate the effects of uncertainty in an uncertainty visualisation.

4.1.2 Interpretation and semantics

Interpretation and semantics experiments are seeking to identify a dimension (or visual task) that uncertainty naturally maps to. These experiments inherently view uncertainty as a variable that is separate to the variable on which we have mapped our signal. For example, lets say we have a map were maximum daily temperature is presented using points where the colour (red for hot and blue for cold) of the point is associated with the temperature, and

the blurriness of that point is associated with the variance *of* that temperature. It is highly likely that our brains will not flatten that into a single variable depicting noise and signal, but rather *separately* extract the temperature (colour) and uncertainty (blur) information as two independent variables. If the variables are extracted separately, there is no guarantee that the uncertainty will act as an appropriate signal suppressor. This problem has been noticed by others in the field, that typically use this method (specifically in the spatial uncertainty context) and a desire for representations that integrate uncertainty and signal is one of the reasons for the invention of the value-suppressing uncertainty pallet (Correll, Moritz, and Heer 2018).

Value-suppressing uncertainty pallets (VSUP) were developed as a method that would allow the signal and the noise to be interpreted together such that insights gained by the viewers of a plot are appropriately *suppressed* by the uncertainty. Hence the name of the pallet. Figure 7 depicted this map colouring approach and several other extensions on the typical choropleth map that differ in where in the visualisation process they combine the noise and signal information into a single concept of "valid signal". The first and most basic map is a simple choropleth map where each value (the colouring of each local government area) has no associated "uncertainty". The next map (which embodies the approach taken by the semantics experiments) is the bivariate map which maps the signal to colour value and the uncertainty to colour hue. If visualisation was performing signal suppression then that would mean the two dimensional space defined by colour value and colour hue can be mentally "flattened" into a single dimension of "valid value" that can captures the signal the our brain would need to be able to flatted this two dimensional space into a single space of "signal validity". The idea of two perceptual tasks flattening into one variable in the mind of the viewer may be wishful thinking, but it is not impossible given we are not certain on how the perceptual tasks are mapped within the human brain. Sterzik et al. (2023) found that when a value was mapped to the textures of stippling, hatching, and triangles, and found that the difference between two points on this one dimensional texture was actually a 2D space (likely "texture business" and "light/darkness"). That being said, if we look at the visual signals presented by the bivariate map, where the contrasting light and dark areas actually has no important meaning, it is unlikely this occurs for colour value and hue (or at least it doesn't occur in a way that is useful for uncertainty visualisation). Instead of hoping that uncertainty might collapse signal values into a single dimension, we can do some of that work ourselves, by using a VSUP which collapses the colour space such that high uncertainty values cannot be extracted. It is unclear how useful VSUPs are at actually combining signal and noise and therefore suppressing plot level insights, as they have only been tested on simple value extraction tasks that require evaluating a single point (Correll, Moritz, and Heer 2018; Ndlovu, Shrestha, and Harrison 2023) rather than looking for spatial relationships (which is arguably what maps are for). Following along with this trend, the next way we might consider visualising uncertainty is to combine uncertainty and signal at the earlier stage so the "suppressed signal" is represented by a single variable. This statistic can then be expressed in a one dimensional colour space, which is a method adopted by the Bayesian surprise metric map (Ndlovu, Shrestha, and Harrison 2023) and the exceedance probability map [Lucchesi2017].



(a) Choropleth Map: Low Variance

(b) Choropleth Map: High Variance

(c) Bivariate Map: Low Variance

(d) Bivariate Map: High Variance

(e) VSUP Map: Low Variance

(f) VSUP Map: High Variance

(g) Pixel Map: Low Variance

(h) Pixel Map: High Variance

Figure 7: A visualisation of a typical choropleth map, as well as three other maps that are used to display uncertainty on the choropleth map. There is a high variance and low variance example for each type of map to show how well the technique supresses signal. Each map was created using the same data with the same base palette. At first glance, the high uncertainty bivariate and VSUP maps just look like maps with a low saturation colour palette rather than map with high uncertainty. This visualisation makes it clear that supression methods that plot uncertainty to a second axis, such as hue, make uncertainty appear as a second variable, rather than a signal supression on our estimate.

These maps make the importance of combining uncertainty and signal in a single visual channel clear. A choropleth map will show signal that is not valid inference because of high uncertainty. At the other end of the spectrum, the bivariate map will show signal that is not always interesting because it forces us to interpret uncertainty and signal separately. As we move through these methods, it seems that the validity of any overarching insight becomes more visible at the cost of our ability to extract particular values of signal or noise. Therefore, given that the primary goal of visualisation *is* insights (North 2006), visualisation authors should err on the side of representing suppressed signal as a single variable, rather than visualising uncertainty separately using two different channels.

4.2 Suggestions to measure uncertainty

So, the current methods of measuring or understanding the role of uncertainty in a visualisation is questionable at best, however this is not because visualisation authors are missing the mark, but rather uncertainty is *particularly* difficult to express in a visualisation. In simple estimates or verbal communication, the signal is often easy to identify because it is what we are explicitly saying. Unlike statistical models, visualisations are used in both data exploration and communication. This means what exactly is a *signal* in any particular visualisation is hard to identify, since we often let the visualisation *tell us* what the signal is. Additionally, you cannot add noise to *every single possible* signal one might take from a visualisation. Two people looking at the same visualisation might, just by chance, develop two entirely different insights and draw inference on two completely different statistics. These unique and fascinating challenges that are faced by uncertainty visualisation have been completely untouched by the literature. This section will cover some interesting research in uncertainty visualisation and suggestions for better ways to measure uncertainty.

4.2.1 More specific hypothesis

Heuristic checks are useful because they look at unknown pitfalls that might exist in interpretation of current plots. Since the hypothesis for these experiments are usually quite specific, e.g. “do people perceive a an outcome that is within the bar as more likely than one outside it, even if both outcomes are the same distance from the mean?”. This means they are less likely to fall into the trap of trying to answer questions that are *far* too broad to be answered with a single experiment (e.g. “is a scatter plot better at showing uncertainty than a box plot”). This work also provides useful insights for experiments by highlights pitfalls participants might fall into when they review the results of evaluation experiments (Hullman 2016). Newman and Scholl (2012) found that participants were more likely to view points within the bar as more likely than points outside of the bar in bar charts with error bars. Similar effects have been identified in other types of uncertainty displayed. L. M. Padilla, Ruginski, and Creem-Regehr (2017) found that points that were on an outcome of an ensemble display were perceived as more likely than points not on an outcome, even when the point that was not on a specific

outcome of the ensemble was closer to the mean of the uncertainty distribution. The sine illusion can cause the confidence interval of a smoothed sine curve to seem wider at the peaks than the troughs, causing us to underestimate uncertainty associated with changing values (Vanderplas and Hofmann 2015).

In a similar vein, experiments that verify smaller aspects of plot design might be more useful to the field in the long run because it helps contribute to a larger working theory of "how do we see visualisations". Many visualisation experiments try to compare two plots with several differences, but do not seem to be interested in the mechanisms by which we extract information from visualisations. Small perceptual tasks that seek to answer small but highly relevant questions (for example, if colour hue and colour value can be perceived as a single signal suppressed variable) would be useful to the field.

4.2.2 Qualitative Studies

Alternatively visualisation research could shift away from the accuracy concept all together ask questions that allow for open ended responses. This method can enlighten authors as to *how* the uncertainty information was used by the participants. Hofmann et al. (2012) tried to capture this by asking participants why they considered a particular plot to be more "right shifted", however this qualitative assessment does not seem to have made it into the final paper. Daradkeh (2015) presented participants with ten investment alternatives and asked participants "from among available alternatives, which alternative do you prefer the most", and were asked to think aloud and consider the uncertainty in their decision making. The experimenters goal was to observe and organise the methods people use when making decisions in the face of uncertainty. This study was an excellent example in a useful experimental design. They highlighted the specific aspects of uncertainty that participants typically considered, such as the range of outcomes that are above/below a certain threshold, minimum and maximum values, the risk of a loss, etc, and mapped where in the decision making process participants made these considerations. Data visualisation is commonly utilised as a tool in data exploration, so it is not uncommon for a data analyst to make a plot with only a vague goal and pull out a large number of adjacent observations. This experimental framework could replicate this process.

4.2.3 Just noticeable signal

It could be argued that a well done uncertainty visualisations should have an imperceptible signal unless the signal would be identified with a hypothesis test, almost like a reverse line up protocol, but this idea also has some issues that should be considered. The reject or do not reject concepts in hypothesis testing do not offer a complete image of uncertainty, and exploration of uncertainty visualisation largely stems from a desire to move away from this binary framework.

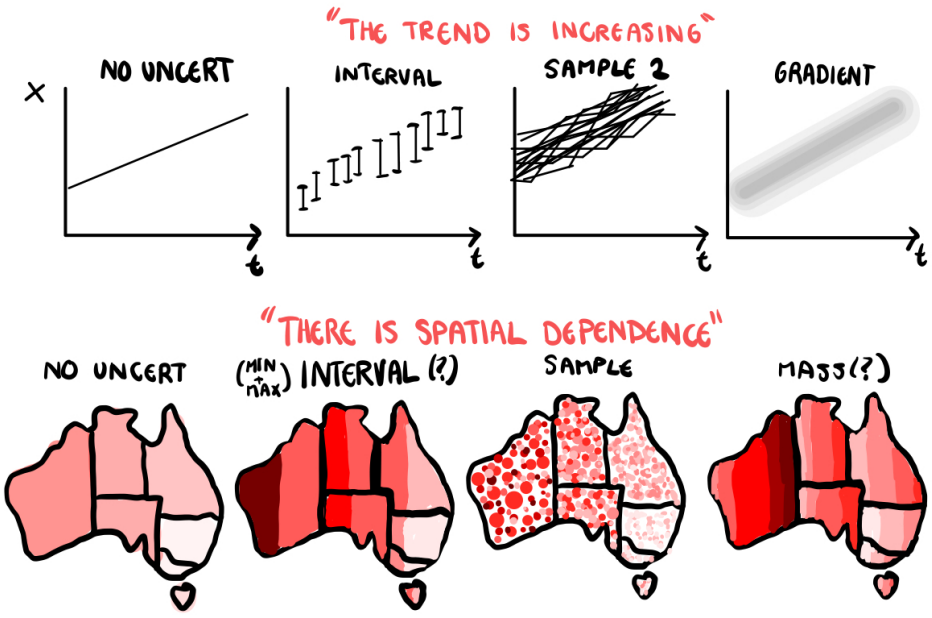


Figure 8: The visualisation showing the kind of signal that we would expect an uncertainty plot to supress

Additionally, setting up an uncertainty visualisation where the participants are expected to notice the signal once the data behind the visualisation passes a hypothesis test implies the signal *is* noticeable to a human at that level. (**Patrick2023?**) compared people ability to recognise patterns in a residual plot to typical statistical tests and found human viewers looking at a plot were less sensitive than the typical residual tests. These experiments utilised the line-up protocol which has been suggested as a method to check if perceived patterns are real or merely the result of chance (Buja et al. 2009; Wickham et al. 2010; Chowdhury et al., n.d.). This concept bears similarity to the goal of uncertainty visualisation, but it is not quite the same. Figure 9 shows the conceptual difference between the line-up protocol and uncertainty visualisation.

5 Great Examples

Despite the common problems detailed in the previous sections, there is some interesting work in the uncertainty visualisation space. This can come in the form of an uncertainty visualisation that attempts to visualise a typically ignored aspect of uncertainty, or an experiment that avoids the pitfalls detailed in the previous sections.

The literature focus on "quantifiable" uncertainty leaves the variance that occurs at early stages in our analysis that is difficult to quantify ignored and forgotten. Some authors have chosen to focus on these unquantifiable cases and look at expressing more complicated cases of uncertainty. Tierney and Cook (2023) builds upon the tidy data principles to allow users to handle missing values. This includes data plots with a missing value "shadow" that allows visualisation authors to identify if the variables used in a plot have any structure in their missing values, which would contribute to uncertainty. Another example of uncertainty that is often ignored is the uncertainty resulting from human choices. Climate scenario uncertainty, shown in Figure 10, attempts to display the range of climate change outcomes that can result from a range of best and worse case human choices.

Another visualisation method that has a lot of potential is visualisation of samples. Visualisations that opt to express a signal as a sample rather than an estimate have the potential to suppress signal since it is not explicitly visualised, however this has yet to be shown in evaluation experiments. This is not to say that visualisations of mass would not be able to perform signal suppression, but a sample can easily be expressed using aesthetics such as colour on a map and mass visualisation often struggling with issues such as over or under smoothing. These sampling methods can show more of the messiness of the data that sits behind a model. This may not have a detrimental effect on the viewers ability to extract global statistics, as it seems they can be extracted from a visualisation of a sample with ease [Franconeri2021]. Sample visualisations have been used in maps with the pixel map, shown in Figure 11, but is more commonly used in animation with the HOPs plots (Hullman, Resnick, and Adar 2015) or similar concepts (Blenkinsop et al. 2000). This method has also been adopted by visualisation

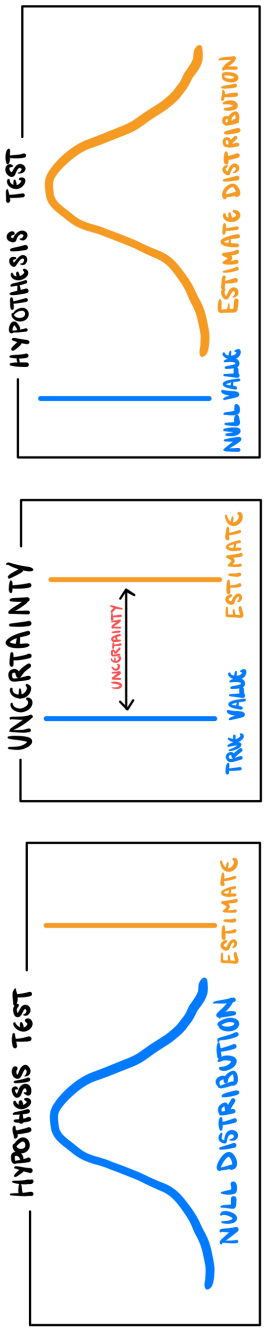


Figure 9: A line-up protocol displays the uncertainty about the null and identifies if the true data plot is identifiable (and therefore significantly difference). This can be considered the graphical equivalent of a standard hypothesis test. Uncertainty is frequently used to descibe the area around the estimate that we think might contain the true value, and we assesses if the null of "no signal" is within this plot (e.g. if the error bars overlap with zero). Uncertainty is technically the distance between the true value and the estimate.

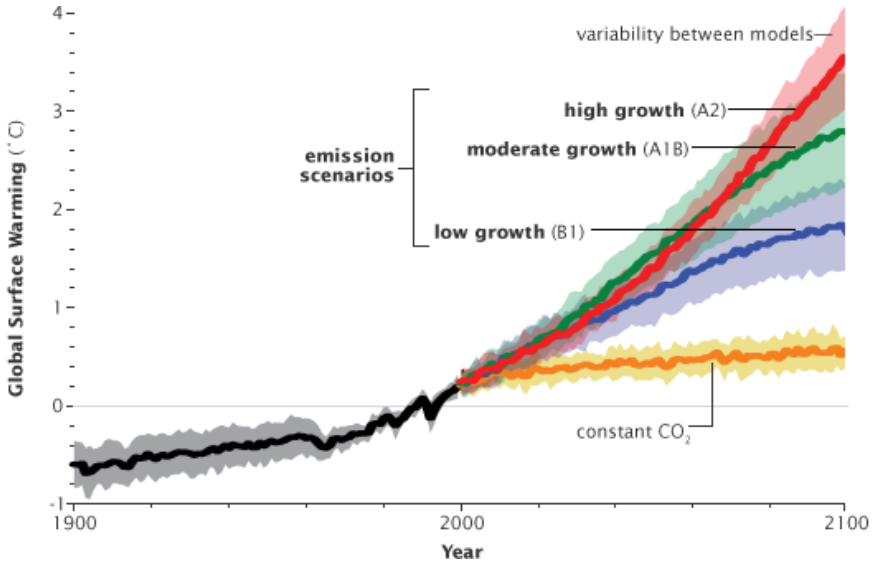
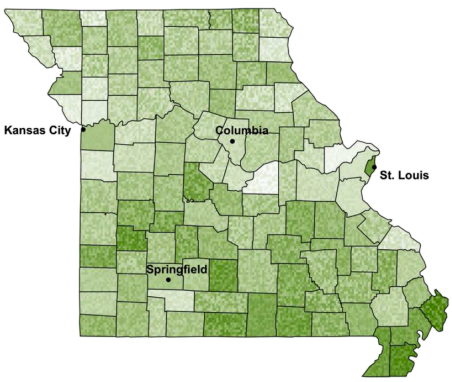
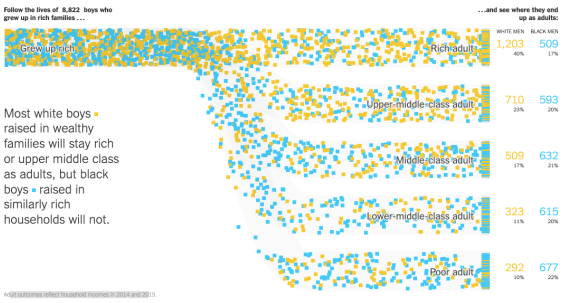


Figure 10: Notice: I will make my own Eversion of a climate scenario plot, this one is just here from a random site as a place holder. A climate scenario plot shows the forecasts of different emission scenarios depending on human choices. The model visualisation wants to show the impact of human choice on the visualisation, rather than assume it away, so five different cases are shown in the model. Each of the case studies also depicts the statistical uncertainty of the model using a confidence interval.

authors outside of academia as can be seen in the and the New York Times class mobility figure, shown in Figure 11, an animated version of which can be found here.



(a) Pixel map



(b) Outcome plot

Figure 11: Notice: Will replace both plots with my own r visualisations. The pixel map depicts a map of the different poverty rates in local government areas in the US state of Missouri. The NYT visualisation shows an screencap of an animation that depicts the economic status of white and black boys who were raised in wealthy families. By displaying a psudo-sample rather than an estimate and a variance, the visualisation does some signal suppression as the true poverty rate is masked and can only be extracted by taking the global statistic of a sample. You can also depict the signal using explicit values, as the the NYT visualisation did, but it is important that the main visual features depict the uncertainty in the estimate.

A final method to perform signal suppression is simply to visualise the data if it is available and relevant to the uncertainty distribution. An example of this is shown in Figure 12 for racial distributions spatially in America, an interactive version of the plot can be found here. This map shows the typical causes of uncertainty in a spatial model, (e.g. regions where data is sparse, ethnically diverse areas, uneven distribution of points within boundaries, etc) but it avoids the need to create a visualisation with a specific signal in mind. This is the technique typically employed by exploratory data analysis, which means it's lack of a specified signal means there is both *no* uncertainty (since we are technically not performing inference) but in the event we *do* implicitly perform inference, there is some hedging. In this sense, the best uncertainty visualisation you can get without specifying a signal you want to convey is visualising the data itself. The census dot map's addition of interactivity also allows users to zoom in and see the details that caused "uncertainty" in the form of inconsistent colours at lower resolutions when they were zoomed out. This does not mean that visualising raw data instead of implementing sampling techniques is always a valid uncertainty visualisation that will prevent insignificant signal from getting through. Buja et al. (2009) illustrated how groups that appear linearly separable in a linear discriminant analysis (LDA) visualisation of the data can actually be the result of a LDA performed on too many variables, something that

was not clear from the visualisation until the line-up protocol was implemented. However it is simple but effective option that seems to be largely overlooked by the uncertainty visualisation community.

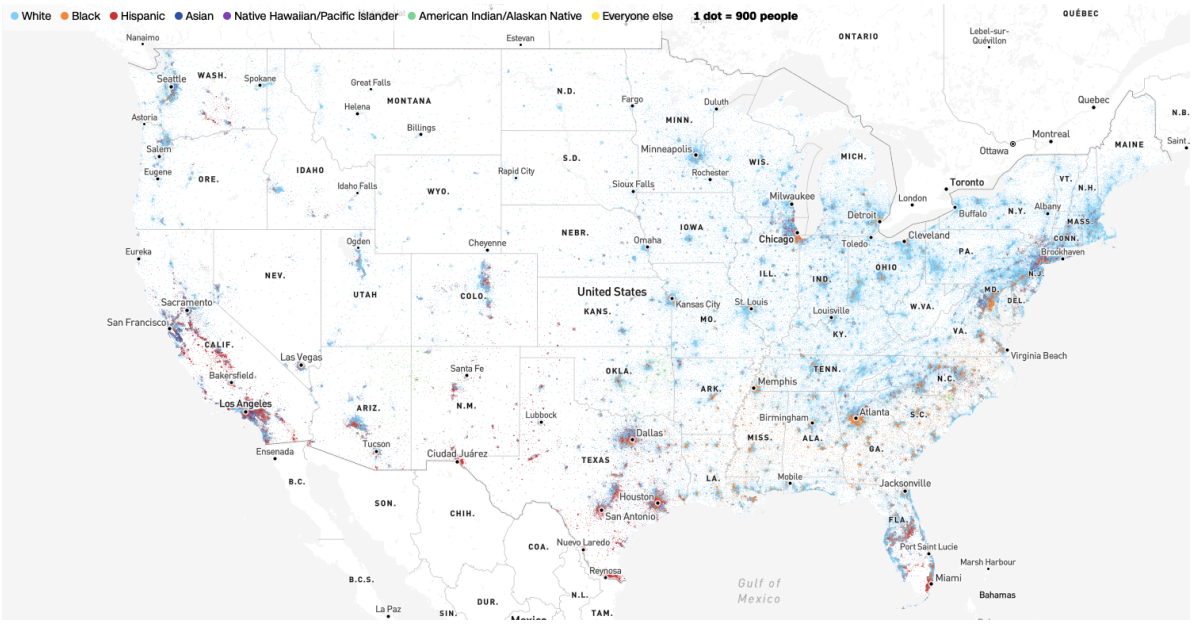


Figure 12: Census dot map

6 Future work

This paper has identified gaps in the uncertainty visualisation literature that could be filled to progress the field.

Each new development should be accompanied by a mathematical definition of the uncertainty being addressed. Ideally, a mathematically definition of uncertainty that allows us to combine these components would be developed, but in the absence of that, authors should be more specific about what aspect of "uncertainty" they are covering with their visualisation.

The concept of uncertainty should be formalised within the grammar of graphics. This formalisation would allow uncertainty visualisation authors to have a clear understanding of what is or is not an uncertainty visualisation. Additionally placing uncertainty visualisation in the framework that is used to understand existing information visualisation research would help authors understand when existing methods can be used to explain their results. incorporating uncertainty into the grammar of graphics will also give a more precise concept of the information contained within a plot. Other fields of science employ marginal changes when designing experiments to ensure it is well understood *what* aspect of their experiment is contributing to

their results, and a better sense of what "marginal" is in the case of uncertainty visualisation would greatly help the field. (*XXX Is data pipeline connected with the grammar of graphics? Should this be a recommendation?*)

Experimental practices on uncertainty visualisation need to be standardised. If we are going to consider uncertainty as noise, not signal, there needs to be a way to identify this signal suppression in an experimental design. As the literature currently exists, there is no way to combine papers to get a meaningful sense of how uncertainty information is understood by a viewer. There is also the possibility that uncertainty visualisation evaluations will need to swap to a qualitative methodology where participants are allowed to freely comment on what they notice in graphics until we establish how the existence of noise can be observed.

If an uncertainty visualisation researcher would prefer to perform experiments rather than formalise methods, there are options there too. It would be interesting to know if any perceptual tasks that can be mapped to two different visual tasks condense into a single dimension when looking for overarching signal in a plot. Alternatively, the task dependency many authors in uncertainty visualisation mention would be a useful direction to consider. It is clear that the the number of potential tasks that can be performed on a visualisation increases with with the number of observations. A single observation is limited to value extraction, two observations can be compared, multiple observations allow for shapes or global statistics to be extracted. The interaction between sample size and task is of particular interest to the uncertainty visualisation community, as uncertainty can be expressed through multiple observations using a sample, or through a single value using an error. Of course, this is limited by the fact that there also isn't a definition for what is a "task" and given the mess created by the lack of formalisation in uncertainty visualisation, it may be wise to formalise that concept before performing these experiments. Amar, Eagan, and Stasko (2005) suggested a taxonomy for information visualisation based on the types of tasks we use visualisations for and suggest 10 "analytical primitives" that we can then map to visualisations, which could be a good starting point. Regardless, these are directions of research would be fruitful to the uncertainty visualisation community even if it appears on the surface to be research that is only beneficial to the "normal" visualisation community. (*XXX Not sure what this paragraph is recommending?*)

Bibliography

Amar, Robert, James Eagan, and John Stasko. 2005. "Low-level components of analytic activity in information visualization." *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*, 111–17. https://doi.org/10.1109/INFVIS.2005.1532136.

Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. https://www.tandfonline.com/doi/abs/10.1080/00031305.1973.10478966.

Begg, Steve H., Matthew B. Welsh, and Reldar B. Bratvold. 2014. "Uncertainty vs. variability: What's the difference and why is it important?" *SPE Hydrocarbon Economics and Evaluation Symposium*, no. May: 273–93. https://doi.org/10.2118/169850-ms.

Benjamin, Daniel M., and David V. Budescu. 2018. "The role of type and source of uncertainty on the processing of climate models projections." *Frontiers in Psychology* 9 (MAR): 1–17. https://doi.org/10.3389/fpsyg.2018.00403.

Blenkinsop, Steve, Pete Fisher, Lucy Bastin, and Jo Wood. 2000. "Evaluating the perception of uncertainty in alternative visualization strategies." *Cartographica* 37 (1): 1–13. https://doi.org/10.3138/3645-4v22-0m23-3t52.

Boukhelifa, Nadia, Anastasia Bezerianos, Tobias Isenberg, and Jean Daniel Fekete. 2012. "Evaluating sketchiness as a visual variable for the depiction of qualitative uncertainty." *IEEE Transactions on Visualization and Computer Graphics* 18 (12): 2769–78. https://doi.org/10.1109/TVCG.2012.220.

Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun Kyung Lee, Deborah F. Swayne, and Hadley Wickham. 2009. "Statistical inference for exploratory data analysis and model diagnostics." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367 (1906): 4361–83. https://doi.org/10.1098/rsta.2009.0120.

Carlin, John B., and Margarita Moreno-Betancur. 2023. "On the uses and abuses of regression models: a call for reform of statistical practice and teaching." http://arxiv.org/abs/2309.06668.

Carr, Daniel B., Anthony R. Olsen, and Denis White. 1992. "Hexagon Mosaic Maps for Display of Univariate and Bivariate Geographical Data." *Cartography and Geographic Information Systems* 19 (4): 228–36. https://doi.org/10.1559/152304092783721231.

Cheong, Lisa, Susanne Bleisch, Allison Kealy, Kevin Tolhurst, Tom Wilkening, and Matt Duckham. 2016. "Evaluating the impact of visualization of wildfire hazard upon decision-making under uncertainty." *International Journal of Geographical Information Science* 30 (7): 1377–1404. https://doi.org/10.1080/13658816.2015.1131829.

Chowdhury, Niladri Roy, Dianne Cook, Heike Hofmann, Mahbubul Majumder, Eun-Kyung Lee, and Amy L Toth. n.d. "Using Visual Statistical Inference to Better Understand Random Class Separations in High Dimension, Low Sample Size Data."

Cleveland, William S., and Robert McGill. 1984. "Graphical perception: Theory, experimentation, and application to the development of graphical methods." *Journal of the American Statistical Association* 79 (387): 531–54. https://doi.org/10.1080/01621459.1984.10478080.

Correll, Michael, and Michael Gleicher. 2014. "Error bars considered harmful: Exploring alternate encodings for mean and error." *IEEE Transactions on Visualization and Computer Graphics* 20 (12): 2142–51. https://doi.org/10.1109/TVCG.2014.2346298.

Correll, Michael, Dominik Moritz, and Jeffrey Heer. 2018. "Value-suppressing uncertainty Palettes." *Conference on Human Factors in Computing Systems - Proceedings* 2018-April: 1–11. https://doi.org/10.1145/3173574.3174216.

Daniel, Kahneman. 2017. *Thinking, Fast and Slow.*

Daradkeh, Mohammad. 2015. "Exploring the use of an information visualization tool for decision support under uncertainty and risk." *ACM International Conference Proceeding Series* 24-26-Sept. https://doi.org/10.1145/2832987.2833050.

Davis, Russell, Xiaoying Pu, Yiren Ding, Brian D. Hall, Karen Bonilla, Mi Feng, Matthew

Kay, and Lane Harrison. 2022. "The Risks of Ranking: Revisiting Graphical Perception to Model Individual Differences in Visualization Performance." *IEEE Transactions on Visualization and Computer Graphics* PP: 1–16. https://doi.org/10.1109/TVCG.2022.3226463.

Fernandes, Michael, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. "Uncertainty displays using quantile dotplots or CDFs improve transit decision-making." *Conference on Human Factors in Computing Systems - Proceedings* 2018-April: 1–12. https://doi.org/10.1145/3173574.3173718.

Fischhoff, Baruch, and Alex L. Davis. 2014. "Communicating scientific uncertainty." *Proceedings of the National Academy of Sciences of the United States of America* 111: 13664–71. https://doi.org/10.1073/pnas.1317504111.

Grewal, Yashvir, Sarah Goodwin, and Tim Dwyer. 2021. "Visualising Temporal Uncertainty: A Taxonomy and Call for Systematic Evaluation." *IEEE Pacific Visualization Symposium* 2021-April (April): 41–45. https://doi.org/10.1109/PACIFICVIS52677.2021.00013.

Griethe, Henning, and Heidrun Schumann. 2006. "The Visualization of Uncertain Data: Methods and Problems." *Proceedings of SimVis '06* vi (August): 143–56. http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:The+Visualization+of+Uncertain+Data:+Methods+and+Problems#0.

Gschwandtnei, Theresia, Markus Bögl, Paolo Federico, and Silvia Miksch. 2016. "Visual Encodings of Temporal Uncertainty: A Comparative User Study." *IEEE Transactions on Visualization and Computer Graphics* 22 (1): 539–48. https://doi.org/10.1109/TVCG.2015.2467752.

Hofman, Jake M., Daniel G. Goldstein, and Jessica Hullman. 2020. "How Visualizing Inferential Uncertainty Can Mislead Readers about Treatment Effects in Scientific Results." *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376454.

Hofmann, Heike, Lendie Follett, Mahbubul Majumder, and Dianne Cook. 2012. "Graphical Tests for Power Comparison of Competing Designs." http://www.public.iastate.edu/.

Hullman, Jessica. 2016. "Why evaluating uncertainty visualization is error prone." *ACM International Conference Proceeding Series* 24-October: 143–51. https://doi.org/10.1145/2993901.2993919.

———. 2020. "Why Authors Don't Visualize Uncertainty." *IEEE Transactions on Visualization and Computer Graphics* 26 (1): 130–39. https://doi.org/10.1109/TVCG.2019.2934287.

Hullman, Jessica, and Andrew Gelman. 2021. "Designing for Interactive Exploratory Data Analysis Requires Theories of Graphical Inference." *Harvard Data Science Review*, 1–70. https://doi.org/10.1162/99608f92.3ab8a587.

Hullman, Jessica, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2019. "In Pursuit of Error: A Survey of Uncertainty Visualization Evaluation." *IEEE Transactions on Visualization and Computer Graphics* 25 (1): 903–13. https://doi.org/10.1109/TVCG.2018.2864889.

Hullman, Jessica, Paul Resnick, and Eytan Adar. 2015. "Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering." *PLoS ONE* 10 (11). https://doi.org/10.1371/journal.pone.0142444.

Ibrekk, Harald, and M. Granger Morgan. 1987. "Graphical Communication of Uncertain Quantities to Nontechnical People." *Risk Analysis* 7 (4): 519–29. https://doi.org/10.1111/j.1539-6924.1987.tb00488.x.

Kale, Alex, Matthew Kay, and Jessica Hullman. 2019. "Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths." *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290605.3300432.

Kale, Alex, Francis Nguyen, Matthew Kay, and Jessica Hullman. 2018. "Hypothetical Outcome Plots Help Untrained Observers Judge Trends in Ambiguous Data." *IEEE Transactions on Visualization and Computer Graphics* 25 (1): 892–902.

Kinkeldey, Christoph, Alan M. MacEachren, and Jochen Schiewe. 2014. "How to assess visual communication of uncertainty? a systematic review of geospatial uncertainty visualisation user studies." *Cartographic Journal* 51 (4): 372–86. https://doi.org/10.1179/1743277414Y.0000000099.

Locke, Steph, and Lucy D'Agostino McGowan. 2018. *datasauRus: Datasets from the Datasaurus Dozen*. https://CRAN.R-project.org/package=datasauRus.

MacEachren, Alan M. 1992. "cartographic perspectives Visualizing Uncertain Information." *Cartographic Perspectives*, no. 13: 10–19.

Manski, Charles F. 2020. "The lure of incredible certitude." *Economics and Philosophy* 36 (2): 216–45. https://doi.org/10.1017/S0266267119000105.

Meng, Xiao Li. 2014. "A trio of inference problems that could win you a nobel prize in statistics (if you help fund it)." *Past, Present, and Future of Statistical Science*, 537–62. https://doi.org/10.1201/b16720-52.

Ndlovu, Akim, Hilson Shrestha, and Lane T Harrison. 2023. "Taken by Surprise? Evaluating How Bayesian Surprise & Suppression Influences Peoples' Takeaways in Map Visualizations." In *2023 IEEE Visualization and Visual Analytics (VIS)*, 136–40. IEEE.

Newburger, Eric, Michael Correll, and Niklas Elmqvist. 2022. "Fitting Bell Curves to Data Distributions Using Visualization." *IEEE Transactions on Visualization and Computer Graphics* 29 (12): 5372–83. https://doi.org/10.1109/TVCG.2022.3210763.

Newman, George E., and Brian J. Scholl. 2012. "Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias." *Psychonomic Bulletin and Review* 19 (4): 601–7. https://doi.org/10.3758/s13423-012-0247-5.

North, Chris. 2006. "Toward measuring visualization insight." *IEEE Computer Graphics and Applications* 26 (3): 6–9. https://doi.org/10.1109/MCG.2006.70.

O'Neill, Onora. 2018. "Linking Trust to Trustworthiness." *International Journal of Philosophical Studies* 26 (2): 293–300. https://doi.org/10.1080/09672559.2018.1454637.

Olston, C., and J. D. Mackinlay. 2002. "Visualizing data with bounded uncertainty." *Proceedings - IEEE Symposium on Information Visualization, INFO VIS* 2002-Janua: 37–40. https://doi.org/10.1109/INFVIS.2002.1173145.

Otsuka, Jun. 2023. *Thinking About Statistics: The Philosophical Foundations*. 1st ed. New York: Routledge. https://doi.org/10.4324/9781003319061.

Padilla, Lace M. K., Maia Powell, Matthew Kay, and Jessica Hullman. 2021. "Uncertain About Uncertainty: How Qualitative Expressions of Forecaster Confidence Impact Decision-Making With Uncertainty Visualizations." *Frontiers in Psychology* 11 (January). https:

//doi.org/10.3389/fpsyg.2020.579267.

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

Padilla, Lace, Matthew Kay, and Jessica Hullman. 2022. "Computational Statistics in Data Science." In, 405–26. John Wiley & Sons.

Potter, K., J. Kniss, R. Riesenfeld, and C. R. Johnson. 2010. "Visualizing summary statistics and uncertainty." *Computer Graphics Forum* 29 (3): 823–32. https://doi.org/10.1111/j.1467-8659.2009.01677.x.

Potter, Kristin, Paul Rosen, and Chris R. Johnson. 2012. "From quantification to visualization: A taxonomy of uncertainty visualization approaches." *IFIP Advances in Information and Communication Technology* 377 AICT: 226–47. https://doi.org/10.1007/978-3-642-32677-6_15.

Pu, Xiaoying, and Matthew Kay. 2020. "A Probabilistic Grammar of Graphics." *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376466.

Refsgaard, Jens Christian, Jeroen P. van der Sluijs, Anker Lajer Højberg, and Peter A. Vanrolleghem. 2007. "Uncertainty in the environmental modelling process - A framework and guidance." *Environmental Modelling and Software* 22 (11): 1543–56. https://doi.org/10.1016/j.envsoft.2007.02.004.

Sanyal, Jibonananda, Song Zhang, Gargi Bhattacharya, Phil Amburn, and Robert J. Moorhead. 2009. "A user study to compare four uncertainty visualization methods for 1D and 2D datasets." *IEEE Transactions on Visualization and Computer Graphics* 15 (6): 1209–18. https://doi.org/10.1109/TVCG.2009.114.

Spiegelhalter, David. 2017. "Risk and uncertainty communication." *Annual Review of Statistics and Its Application* 4: 31–60. https://doi.org/10.1146/annurev-statistics-010814-020148.

Sterzik, Anna, Monique Meuschke, Douglas W. Cunningham, and Kai Lawonn. 2023. "Perceptually Uniform Construction of Illustrative Textures." http://arxiv.org/abs/2308.03644.

Thomson, Judi, Elizabeth Hetzler, Alan MacEachren, Mark Gahegan, and Misha Pavel. 2005. "A typology for visualizing uncertainty." *Visualization and Data Analysis 2005* 5669 (March 2005): 146. https://doi.org/10.1117/12.587254.

Tierney, Nicholas, and Dianne Cook. 2023. "Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations." *Journal of Statistical Software* 105 (7): 1–31. https://doi.org/10.18637/jss.v105.i07.

Vanderplas, Susan, Dianne Cook, and Heike Hofmann. 2020. "Annual Review of Statistics and Its Application Testing Statistical Charts: What Makes a Good Graph?" https://doi.org/10.1146/annurev-statistics-031219-041252.

Vanderplas, Susan, and Heike Hofmann. 2015. "Signs of the Sine Illusion — Why We Need to Care Signs of the Sine Illusion — Why We Need to Care" 8600. https://doi.org/10.1080/10618600.2014.951547.

Vranas, Peter B. M. 2000. "Gigerenzer's normative critique of Kahneman and Tversky." *Cognition* 76 (3): 179–93. https://doi.org/10.1016/S0010-0277(99)00084-0.

Walker, W. E., P. Harremoes, J Rotmans, J. P. Van Der Sluijs, M. B. A. Van Asselt, P Janssen, and M. P. Krayer Von Krauss. 2003. "Defining Uncertainty." *Integrated Assessment* 4 (1): 5–17. https://www.narcis.nl/publication/RecordID/oai:tudelft.nl:uuid:fdc0105c-e601-402a-8f16-ca97e9963592.

Wallsten, Thomas S., David V. Budescu, Ido Erev, and Adele Diederich. 1997. "Evaluating and combining subjective probability estimates." *Journal of Behavioral Decision Making* 10 (3): 243–68. https://doi.org/10.1002/(sici)1099-0771(199709)10:3%3C243::aid-bdm268%3E3.0.co;2-m.

Wickham, Hadley, Dianne Cook, Heike Hofmann, and Andreas Buja. 2010. "Graphical inference for infovis." *IEEE Transactions on Visualization and Computer Graphics* 16: 973–79. https://doi.org/10.1109/TVCG.2010.161.

Wickham, Hadley, and Heike Hofmann. 2011. "Product plots." *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2223–30. https://doi.org/10.1109/TVCG.2011.227.

Wilkinson, Leland. 2005. *The Grammar of Graphics (Statistics and Computing)*. Berlin, Heidelberg: Springer-Verlag.

Wu, Yifan, Ziyang Guo, Michails Mamakos, Jason Hartline, and Jessica Hullman. 2023. "The Rational Agent Benchmark for Data Visualization." https://arxiv.org/abs/2304.03432.

Zhang, Sam, Patrick Ryan Heck, Michelle Meyer, Christopher F Chabris, Daniel G Goldstein, and Jake M Hofman. 2022. "An Illusion of Predictability in Scientific Results."

Zhao, Jieqiong, Yixuan Wang, Michelle V. Mancenido, Erin K. Chiou, and Ross Maciejewski. 2023. "Evaluating the Impact of Uncertainty Visualization on Model Reliance." *IEEE Transactions on Visualization and Computer Graphics* PP (X): 1–15. https://doi.org/10.1109/TVCG.2023.3251950.