Noisy Work: A Review of The Uncertainty Visualisation Literature

Harriet Mason

3/20/24

Table of contents

1 Current Todo List

This is a list of tasks that need to be completed before I consider the paper complete. Please keep this in mind when reading.

2 Background

2.1 Introduction

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 benefit from the power of data visualisation, however this does not mean it is always done with success. 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 is highlighted by data sets such as Anscombe's quartet (Anscombe 1973) or 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 when visualised. Anscombe quartet is shown in Figure 1, because describing the data is never the same as seeing it. Instead of having to repeatedly check endless hypothesis to find interesting numerical features, visualisations **tell** us what is important about our data and **how** it might diverge from what we expect.

Uncertainty visualisation is a somewhat new subfield of visualisation, with early papers that specifically reference an "uncertainty visualisation" popping up in the late 90s and early 2000s (*Cite Gap 1*). Despite the youth of the field, commonly used uncertainty visualisations such as a box plot or histogram have been around for almost as long as statistical graphics themselves

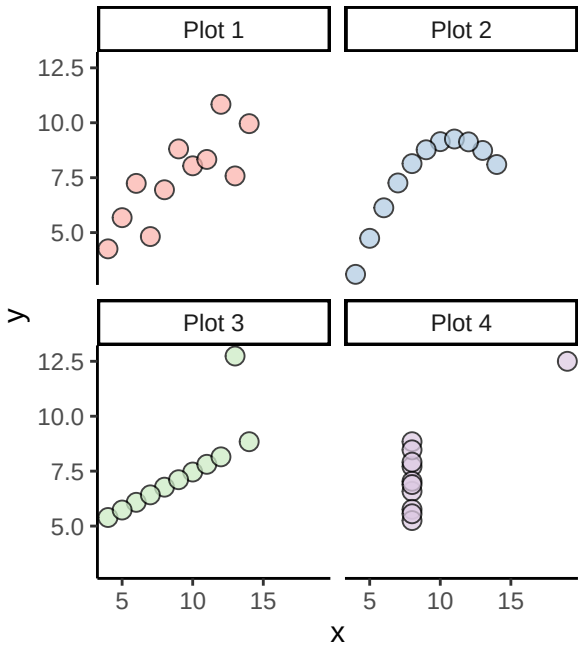


Figure 1: The four scatter plots that make up Anscombe's quartet. The four scatter plots are visually distinct but have the same mean, standard deviation, and correlation. The visualisation highlights the importance of plotting your data to identify interesting features that are hidden by other summary statistics.

(*Cite Gap 2*). Early mentions of "uncertainty visualisation" appear in computer science papers discussing uncertainty visualisation as a subfield of geospatial information visualisation, specifically for the many uncertainties that come with satelite images (*Cite Gap 1*). In the modern day, uncertainty visualisation has exploded into its own field that applies to a wide range of contexts and is no longer bound by geospatial information (*Cite Gap 1*).

2.2 Similar work

Any field that experiences an explosion of new research will also invite a series of literature reviews that seek to combine and summaries that information. Given that uncertainty visualisation has been around for a few decades now, there is already a wealth of uncertainty visualisation literature reviews.

Interestingly, these reviews often do not offer overarching rules for tried and tested uncertainty visualisation, but rather comment on the *difficulties* faced when trying combine the papers from this field. Kinkeldey, MacEachren, and Schiewe (2014) found most experiments the methods for uncertainty visualisation evaluation to be adhoc, 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 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 each uncertainty visualisation paper.

There are a handful of reasons that are used to explain why it is so difficult to generalise the findings in the uncertainty visualisation literature. Some papers suggested or agree with the idea that visualisation typologies should move away from data types, uncertainty categories, and representation types and towards "task-centered typologies" as this will help generalise results (Kinkeldey, MacEachren, and Schiewe 2014; Hullman 2016). Some authors may not attribute the task to the *reason* the literature cannot be easily summarised, but do mention that the choice of best visualisation is highly dependent on the specific goal of that visualisation (Griethe and Schumann 2006; Spiegelhalter 2017). Others comment 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 difficulties in noisy uncertainty visualisation findings arises reguardless of the direction the analysis was approached, be it looking at geospatial data (Kinkeldey, MacEachren, and Schiewe 2014), or uncertainty visualisation as a whole (Hullman 2016; Spiegelhalter 2017; Griethe and Schumann 2006).

This review believe all these sources of noise have a single root cause, a lack of a cohesive and encompassing definition of uncertainty. The absence of an econompasing definition of uncer-

tainty is mentioned by every uncertainty visualisation directly (Spiegelhalter 2017; Griethe and Schumann 2006) or indirectly by describing a miriad of ways it can be considered in the literature (Kinkeldey, MacEachren, and Schiewe 2014; Hullman 2016), although it is never commented on as a source of the noise in the field.

In understanding the definition problem, we will understand 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 improvement. In understanding this purpose, we will also consider whether or not "uncertainty visualisation" is a valid sub-field of visualisation itself, or if that distinction has only arisen as a by product of the confused definition of uncertainty. Ultimately, this paper should serve as a guide for future uncertainty visualisation authors to prevent their work from becoming noise in what is already an incredibly noisy field.

3 What is Uncertainty?

3.1 Colloquial definitions

Uncertainty visualisation is made particularly difficult by the term "uncertainty" lacking a commonly accepted definition in the literature. This mishmash of terminology leads to a large body of work, all claiming to finding the best visualisation or expression of of "uncertainty" but most don't even seem to agree on what uncertainty is.

Definitions presented in papers that discuss uncertainty visualisation want it to be encompassing and include everything a layperson might consider when trying to consider "uncertainty". Some consider it to be synonymous with specific terms defined in mathematics, such as probability, variance, error, or precision. Others consider it to be an encompassing umbrella term of which for these mathematically defined objects are only examples, and include concepts that are somewhat related, such as missing values (*Cite Gap 3*). This wide array of uncertainty definitions becomes a snake eating its own tail, as many authors opt to redefine uncertainty themselves[], ignore defining uncertainty entirely [], or pick a definition from a seemingly randomly previously published papers [], all of which make variance in uncertainty definitions even worse (*Cite Gap 3.5*).

While these vague definitions might be OK for polite conversation, they are *not* foundation on which to build an entire sub-field of visualisation. In order to visualise uncertainty, it needs to be quantifiable [Griethe and Schumann (2006); Leland2005], in order to quantify uncertainty, it needs to be mathematically defined. The uncertainty visualisation literature is completely awash with papers that define uncertainty using a vague encompassing phrase, (e.g. Walker et al. (2003) defined it as any deviation from complete determinism) but go on to *quantify* uncertainty as a PDF, error, or some other easily quantified mathematical object when it comes time to do the visualisation (*Cite Gap 4*). This definition swap happens so subtly that it seems to go unnoticed by authors writing the papers, and the fact that "uncertainty visualisation" methods dont even notice they are exclusively focus on depictions

of easily quantifiable uncertainty, such as error [], mass [], or variance [] (*Cite Gap 5*). This has left the expressions of uncertainty that are hard to quantify or visually differentiate untouched, despite many papers calling for their invention (*Cite Gap 6*). There is likely only confusion and obfuscation if the vague definitions of "uncertainty" continue to be used and authors should try to be more specific with the mathematical object they are visualising, *or* a strict definition of uncertainty needs to be defined. The current method of titling papers and defining uncertainty forces readers to comb through the method of a particular visualisation experiment to understand if the uncertainty visualised is amenable to their needs.

Some papers have recognised these colloquial definitions are lacking and have made attempts to formalise the space with taxonomies, however these taxonomies come with their own issues.

3.2 Taxonomy definitions

Much in the same way that there are almost as many definitions of uncertainty as there are papers on the topic, the field is also over run with 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), if the uncertainty is in the the attribute, spatial elements, or temporal element of the data [], whether the uncertainty is scientific (e.g. error) or human (e.g. disagreement among parties) [], if the uncertainty is random or systematic [], statistical or bounded [], accuracy or precision [], if the uncertainty is about the past or the future [], the stage of the data analysis pipeline it comes from [], the severity of the uncertainty [], how quantifiable the uncertainty is (Spiegelhalter 2017), etc., etc., etc. (*Cite Gap 7*). In their own way, each of these taxonomies show an apsect 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, depicted in Figure 2. This definition organises uncertainty based on three axis, location (where the uncertianty is coming from), level (how quantifiable it is), and nature (if it comes from true randomness or incomplete knowledge).

When considering the use of taxonomies to define something inherrently vague, such as the uncertainty in an uncertainty visualisation, it is helpful to consider this quote from *The Grammar of Graphics*:

> "Taxonomies are useful to scientists when they lead to new theory or stimulate insights into a problem that previous theorizing might conceal. Classification for its own sake, however, is as unproductive in design as it is in science. (Wilkinson 2005)

Therefore, 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 definition and do little to help statisticians untangle and estimate the uncertainty in their projects. This is obvious in practice, as

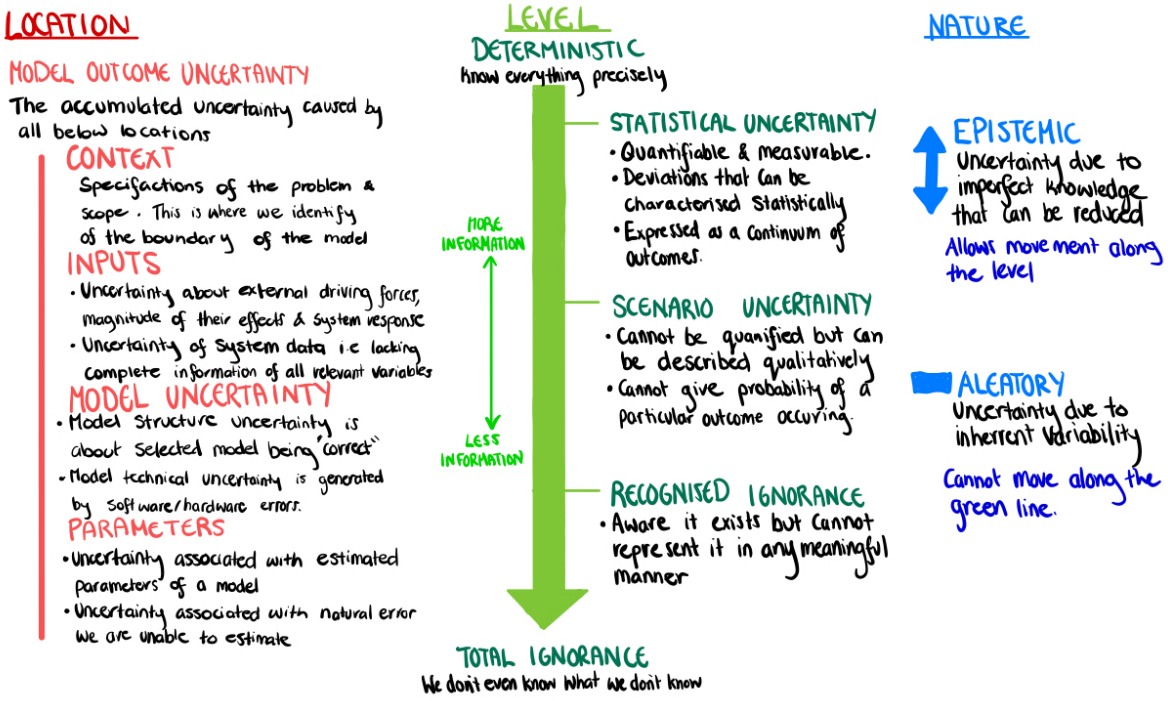


Figure 2: *I need to fix the taxonomy illustration because inputs and model uncertainty need to be swapped.* Depicts an illustration of the taxonomy described in Walker et al. (2003). From right to left the drawing shows the location, level and nature of uncertainty with examples of that category underneath. A specific source of uncertainty from the location can be mapped to a level of ignorance that can increase or decrease (i.e. moving up or down the green line) depending on the nature of the uncertainty. Identifying the location, level and nature of your uncertainty allows you to better understand it.

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). That being said, these taxonomies can hint towards what is important when we think about uncertainty. The location axis of the Walker et al. (2003) taxonomy lines up neatly with the typical data analysis pipeline and 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) including stages as early as conceptualising the problem (Otsuka 2023) or collecting the data (Meng 2021). The level axis from the Walker et al. (2003) taxonomy also hints towards an important consideration in uncertainty, because establishing how quantifiable uncertainty is informs us how it can be communicated, something that Spiegelhalter (2017) identified in the form of "precision". It is commonly noted that each source of uncertainty in an analysis must be discussed in isolation, but combining the uncertainty from every stage is near impossible (Spiegelhalter 2017) (*Cite Gap 8*). It is clear that elements of these taxonomies are identified in many other comments on uncertainty and represent *real* considerations that need to be made when defining uncertainty, even if that definition is yet to be published.

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.

3.3 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, *posivitism* 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 useage. 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 disagreee with that statement. Variance and probability are typically considered stand ins for "uncertainty", it is often how we choose to measure it, and since probability and variance exist in descriptive statistics, descriptive statistics *must* have uncertainty. This is not necessarily true, and related distinction was made by Spiegelhalter (2017) commented on a differentiation between precise random events (such as the probability of a coin flip), and uncertainty (such as the estimated probability associated with a coin that might be biased). A sample variance is not unknown, and therefore it is not uncertain, rather it is 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, but in isolation it is not uncertain.

The idea that inference is related to uncertainty pops up in most discussions of uncertainty, however, due to the now *implicit* understanding we are performing inferential statistics, it is often brought up as a *task* or *goal* dependence. This is mentioned both by authors at a specific stage of an analysis, and by authors looking at an entire field or at all stages of the analysis pipeline. 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. 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. Munzner (2009) created a nested model for visualisation that highlighted how the first mistake that can be made in a visualisation is in the problem characterisation, and failling to do it well can cause downstream effects and damage the effectivness of a visualisation. 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. These examples show that authors that discuss uncertainty believe it is there is an important relationship between uncertainty and what we decide to keep in our analysis and throw away, i.e. the task or goal. The task or goal is, in essence, the *statistic you are drawing inference on.* Were this not true, we would have no reason to differentiate between a prediction interval and a sampling distribution because they could both be considered "uncertainty about the mean", however they are different because their inferential goals are different.

These examples highlight two key concepts that seem to be true about the relationship between uncertainty and inference:

1) Uncertainty is the by-product of inference, as we seek to draw conclusions about something unknowable.
2) Uncertainty must be defined with respect to a *specific* statistic we wish to draw inference on, there is no such thing as a *general* uncertainty that is not linked to a motivating question.

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*. These things are not "uncertainty" in of themselves, but rather contribute to the distance between the final estimate and the statistic we want to perform inference on. The field of uncertainty communication is in desperate need of a unifying mathematical definition that firmly identifies exactly how each of these things contribute to uncertainty. This may be a considerable task, but as of right now, the list of things that are "uncertainty" continues to grow, and the best methods of understanding and quantifying them are ad-hoc at best. 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, information theory tries to quantify uncertainty using the idea of entropy. None of these existing methods are thorough enough for an analyst to understand what causes uncertainty, and quantify it for communication. The need for a mathematical definition of this field is put simply by Freeman Dyson in his fameous Birds and Frogs speech:

> "Rigorous theorems are the best way to give a subject intellectual depth and precision. Until you can prove rigorous theorems, you do not fully understand the meaning of your concepts." Dyson (2010)

In order to visualise something, it needs to be quantifiable. In order to quanitfy something, it needs to be mathematically defined. The broad classes of what is consdiered "uncertainty" is not currently quantifiable in any way that is not ad-hoc. The only apsects of uncertaitny that are currently quantifiable are confidence intervals, prediction intervals, and related terms. Even then, these quantifications often only capture the uncertainty in our model or its assumptions, it ignores bias often introduced in earlier stages of the analysis. Broader and more complicated concepts such as the effects of assumptions, imputed missing variables, and model choices remain difficult to meaningfully quantify beyond ad-hoc methods.

3.4 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 visualisaiton authors directly stare at this concept, they almost unanimously agree with these concepts, however when these details are

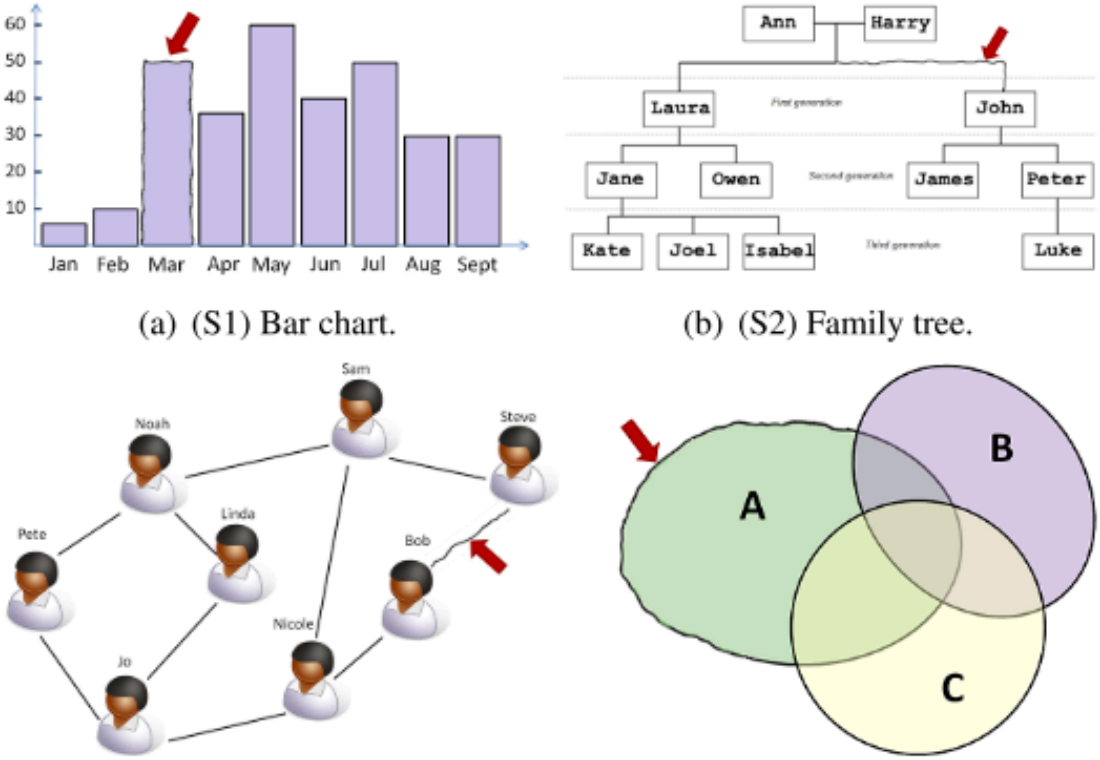
pushed the the periphery of a problem, we start to get research that confuses itself. In these cases, the authors are more confused about what they are doing than the participants.

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 effectivness 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* statistics. 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 3 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.

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



(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 3: The graphics displayed by Boukhelifa et al. (2012) to identify if there is an intuitive connection between sketchiness and uncertainty.

an expression of suspected bias in the model) both caused participants to have decreased confidence in their results and suggested modelers 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 visualisaiton, 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 visualisaiton, but simply depicts different visual representations of a PDF, however it takes reading the methodology to find this out (*Cite Gap 9*). 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.

4 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.

4.1 Establishing misunderstood concepts

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. The second is a conflation of the contextual information with the visual information. This problem is not, in isolation an issue, however it does seem to be a *key* contributor to the misunderstanding that uncertainty is an attribute of data, and not a by product of inference. Both of these issues need to be understood in order for us to fully understand the issues surrounding uncertainty visualisation.

4.1.1 Definitions of "uncertainty visualisation"

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 most "uncertainty visualisation" experiments (Ibrekk and Morgan 1987) (*Cite Gap 10*). 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 mass and would already be considered an uncertainty visualisation were our statistic of interest the population mean. 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 conceptualisation 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 10*)

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 ect) 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 aline 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.

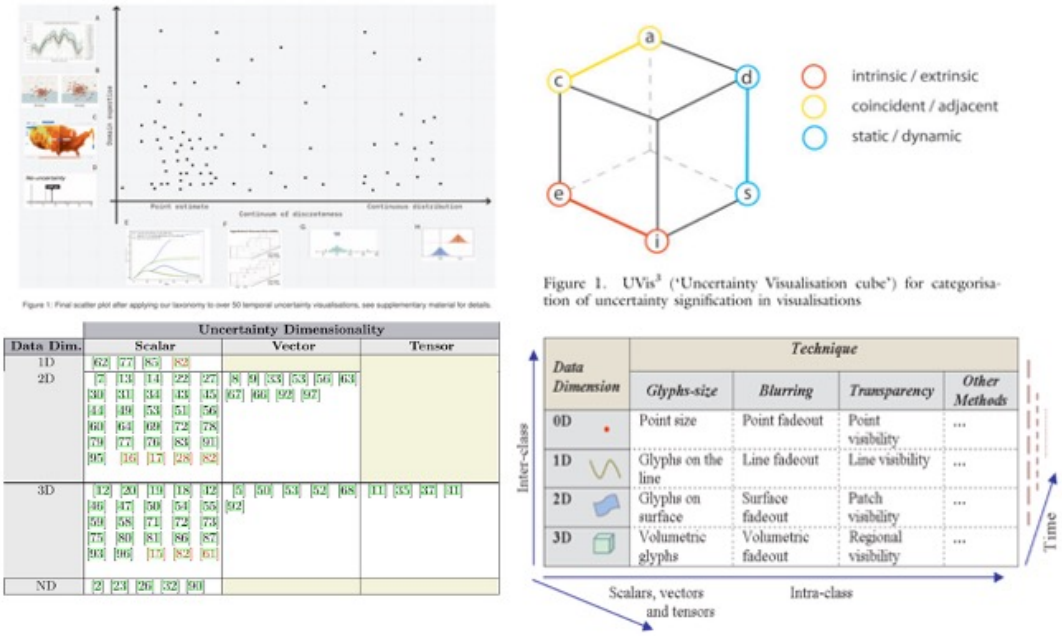


Figure 1. UVis[3] ('Uncertainty Visualisation cube') for categorisation of uncertainty signification in visualisations

Figure 4: This visualisation is here as a placeholder. I am going to replace it with an R coded visualisation that shows the 5 considerations for showing uncertainty in a plot according to Kinkeldey, MacEachren, and Schiewe (2014)

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 visualsiation 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.

4.1.2 Contextual Information

One of the key components of uncertainty and inference is the noise and signal dichotomy. When Otsuka (2023) discusses the information we boil away to create a "probabilistic kind" he uses the example of a coin flip and the sex of baby births as two things that are *very* different in reality, but it is only through the lense of a Bernoulli distribution that we can view these two things as the same. The fact is that this boiling down, of a real world event into data, is not something that happens independent of the statisticians, but is just as much as product of our interests (the particular relationship we hope to draw inference on) as it is a product of the event's natural properties. This is expressed simply in the text with an extension to the coin flipping example:

> "Recall that we modeled a coin toss by a Bernoulli distribution. That was because we supposed the coin will land either heads or tails and ignored the possibility of its landing on its edge. If we are to take the latter possibility into account, we should use a multinational distribution with three categories. As this example shows, the question of which statistical model we should use is not just determined by the physical nature of the object under study, but rather depends on the modeler's interests and intentions." Otsuka (2023)

With this in mind, let us turn to the concept of "contextual" information. For the purposes of this review, we will consider contexual information to be anything that makes it easy for the audience to connect the data from a probablistic kind to its real world parallel, but is largely irrelevant in the abstract probability space. A key example of this is spatial information, such as borders, road locations, terrains, etc. Spatial information is *highly relevant* when we translate from the abstract domain back into real world for communication. It can help provide reasons for *why* there may be specific characteristics in our data, however it is not required information for the model. Figure 5 shows the proces of boilling information down into data, visualising it, and including contextual information to improve our scientific reasoning. The concept that contextual information is not the same as statistical information is also highlighted in *The Grammar of Graphics* which states:

> "Geography is anchored in real space-time and statistics in abstract dimensions. This is a distinction along a continuum rather than a sharp break… but this difference in focus clearly means that a system optimized to handle geography will not be graceful when dealing with statistical graphics." Wilkinson (2005)

For the remaining of this paper, we are not going to differentiate uncertainty visualisation using contextual information, despite the fact that it is common in the field. Our reasoning for this is that organising the field according to contextual information, promotes the idea that uncertainty is a concept specific to data qualities and ignores the role of inferential statistics.

While other discussions of uncertainty may consider this distinction more relevant to their work, here we will consider this contextual information. Contextual information is important for

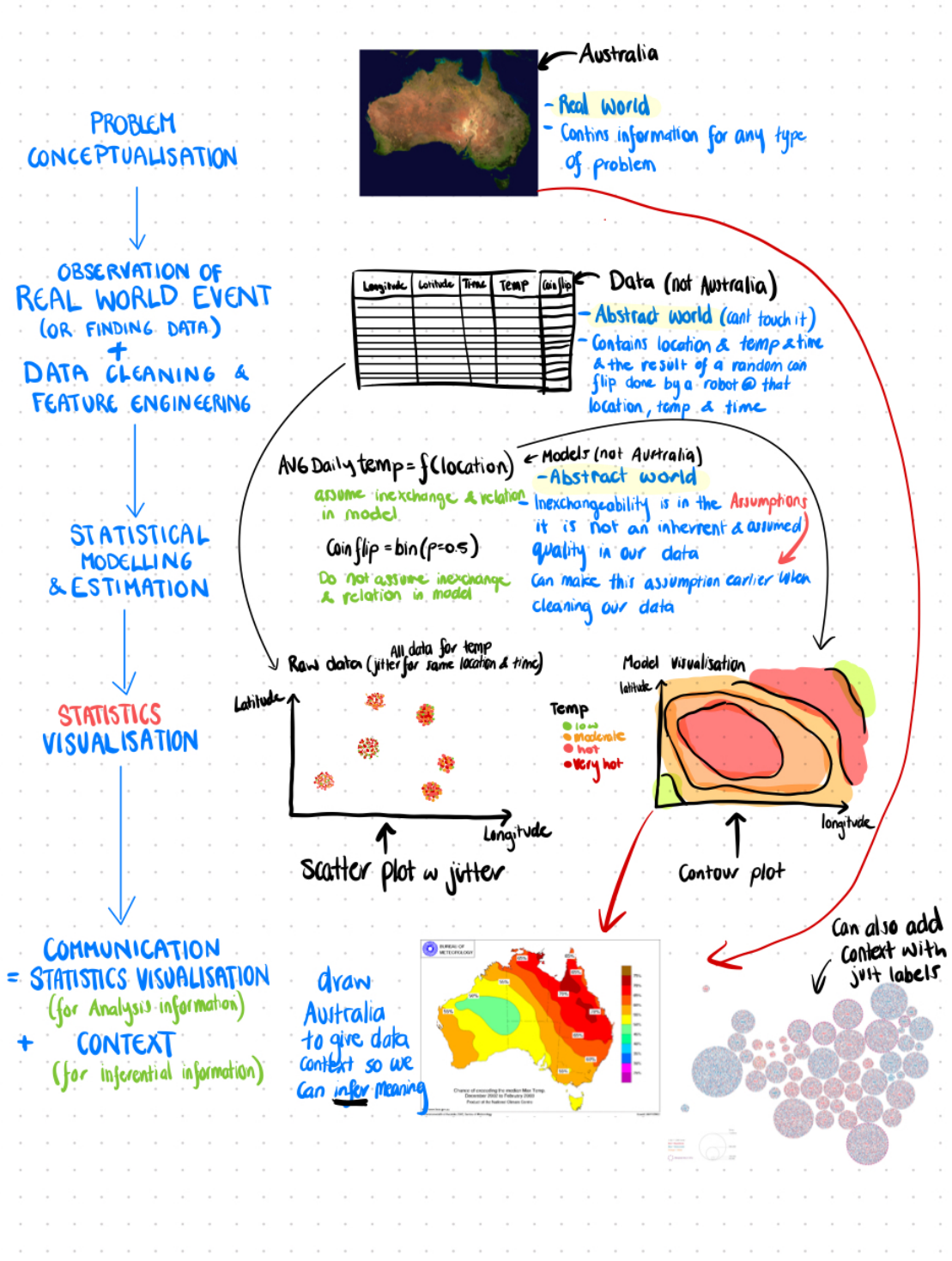


Figure 5: Data pipeline with visualisations to show why we are ignoring contextual information for the time being. The visualisations that depict the data are similar to typical statistical graphics. The purpose of the spatial information is to provide context to the information we extract from the statistical aspect of the graphic.

interpretation and understanding of graphics, but it does not in of itself generate uncertainty. Its relevance to the field of "uncertainty visualisation" is through the percieved lack of available channels. For example, a choropleth map has already used colour to represent an estimate, and position to represent spatial information, which is sometimes relevant to the probabilistic kind, and other times it is simply contextual information (such as when the modelling is done by using a state as a category rather than its geographic information). These two cases, when the spatial information is model relevant and when it is contextual, are rarely separated. It can lead authors to believe the spatial information is always of the highest importance, and thus it is always given the position axis, even when it is not, especially with respect to the uncertainty in a model. A confidence interval is a confidence interval whether it came from spatial data, temporal, or cross sectional data.

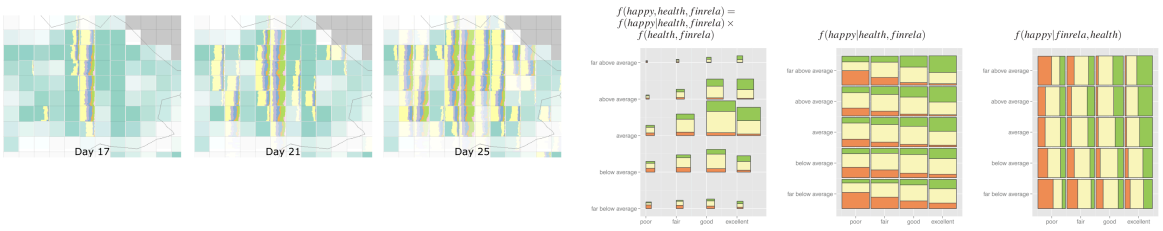
Similar to the confusing definition of uncertainty, the constant conflation of statistical and contextual information also creates a wide range of problems. Kinkeldey, MacEachren, and Schiewe (2014) discussed how uncertainty can be represented by three components attribute (what) position (where) and temporal (when) and that studied typically deal with uncertainty around attribute but rarely position and time, however it is never specified what considerations should cause attribute uncertainty to be different to position or temporal uncertainty. Kay et al. (2016) did an experiment that showed uncertainty around bus arrival times, however the visualisation used in the experiment, shown in Figure 6, is indistinguishable from most work that would be considered "attribute" uncertainty.

This paper focuses on uncertainty, specifically uncertainty that is still in the the abstract *statistics* domain. Spatial and temporal uncertainty papers limit how they communicate their uncertainty because they are considering what can be visualised inside the context provided by maps and, not what they should *try* to visualise based on the uncertainty abstraction. When we view a spatial-temporal visualisation through the lense of trying to visualise the distribution of $P(X|longitude, latitude, time)$ there is no reason this is any different from any other high dimensional visualisation problem.

In this sense we are not proposing that the current conceptualisation of data according to its physical properties is necessarily wrong, or that it leads to conclusions that are not necessarily specific to spatial or temporal data, and it may lead us to miss relevant information that is outside our domain. This is not only the case for uncertainty data. Slingsby, Reeve, and Harris (2023) discussed using a gridded glyphmap for spatial modelling, however the statistical elements are identical to a fluxuation diagram depicting $f(infection|time, age, longitude, latitude)$ presented in Wickham and Hofmann (2011). The plot by Slingsby, Reeve, and Harris (2023) is simply a special case of the graphic by Wickham and Hofmann (2011), where a separate visualisation is made for each time point, the continuous function $infection|age$ has replaced the categorical variable *happy*, a fade is added for the population size, and a map is shown in the background. Slingsby, Reeve, and Harris (2023) presents other maps in the paper, however they are also special cases of the product plots framework. This is not to say this extension is uninteresting, on the contrary we believe identifying cases where particular graphics are useful is a worthwhile endevour,

Figure 6: The uncertainty visualisation used by Kay et al. (2016) to show the uncertainty around a bus arrival time prediction.

but these plots seem to be have been developed unaware of the existing statistical framework. When we separate the field of visualisation along lines that may not be of particular relevance, we increase the likelihood of "reinventing the wheel" in each sub-dicipline.



The focus on contextual information has also created a rift in 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.
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 11*). 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.

For these reasons, the implication that uncertainty is not inference related, the barriers that cause authors to ignore existing visualisation methods, and the confusion created in the motivation of uncertainty visualization as a whole, we do not believe it is fruitful to continue separating the field of uncertainty visualisation according to its contextual information. Therefore, the rest of this review will not separate uncertainty visualisations based on this information.

4.2 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.

4.2.1 Information asymetry 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*). 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. *Other examples* (*Cite Gap 13*). 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 reminicent of our problems with the definition of "uncertainty".

Unlike uncertainty, the concept of the "information" in a plot is not entirely avoided and works such as *The Grammar of Graphics* attempt to outline exactly what information is contained within a plot. When creating a graphic there are several tasks that must be completed in a specific order, regardless of whether or not we are working with a uncertainty visualisation or not, this pipeline is illustrated in Figure 7, and contains all the steps of the grammar of graphics (Wilkinson 2005). First the mathematical information in a plot is established in the "varset" setps, and then that mathematical information is visually represented, which is defined in the "graph" steps. 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 asymetry 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. Additionally, the grammar of graphics was not designed to extend to "uncertainty visualisations" as Wilkinson (2005) referred to uncertainty as "semantics", so concepts that are common to uncertainty mathematically, 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 outcoms 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 would 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.

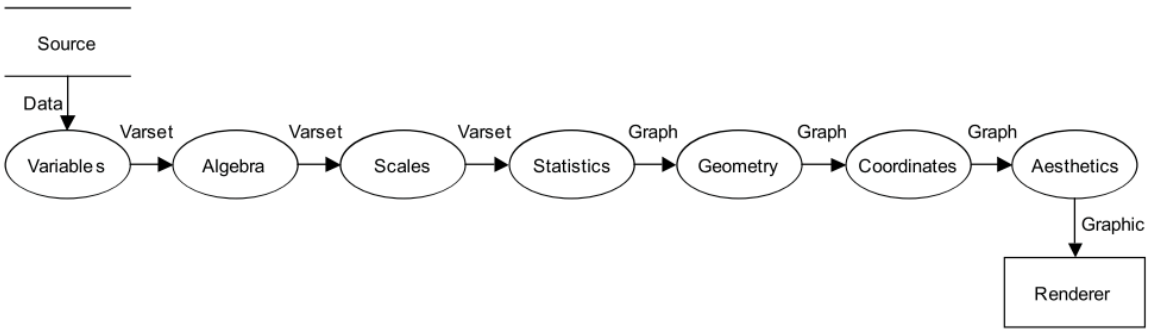


Figure 2.2 *From data to graphic*

Figure 7: The grammar of graphics data analysis pipeline from Wilkinson (2005). the "varset" stages are the mathematical components of the graphic, the mathematical components define the "information" in the graph, while the "graph" stages are the visual elements that determine the "computational" cost of the graph.

Alternative information consideration within well defined frameworks may also be worthwhile to consider. Wu et al. (2023) tried to eliminate information asymetry by esentially 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.

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. *Other examples* (*Cite Gap 14*).

This collection of examples starts to paint a pretty clear picture. Visualisations with rather shocking information asymetry 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.

4.2.2 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, color saturation, color 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 guestalt heuristics, 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, centered 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. *Other examples* (*Cite Gap 15*).

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 uncetainty 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, maxmimums, variance, and any other statistic that can be considered related to "uncertainty" is just that, a statisitc. If we want to visualise the relationship between a temperature estimate and our forcasted variance on that esitmate, 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 take aways 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?

5 Uncertainty is noise, not signal

Uncertainty is noise, but most evaluation experiments measure it as signal. Noise and signal are not an inherrent 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 *supresses* 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 visualisaiton and the results of the studies in the field. Uncertainty is poorly defined, therefore authors do not understand it is not an inherrent 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. It also highlights a unique and facinating 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 fundamentally transparency to the viewer about a statistical analysis. 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* 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.

5.1 Issues with the current methods of measuring uncertainty

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 conceptialisation of uncertainty.

5.1.1 Performance

L1: Behavioral Targets - Performance (241; 64.8%): how effectively a user can extract information, make inferences, or make decisions with a visualization. - Interpretation & Semantics (64; 17.2%): the ease with which a user associates uncertainty with an encoding. - Quality of User Experience (67; 1.8%): the user's valuation of the visualization, such as their preference or satisfaction.

L2: Value extraction - Accuracy (134; 36%): Difference from a ground truth response (e.g., [37, 40]). - Confidence/Accuracy Alignment (5; 1.3%): Correlation between confidence and accuracy (e.g., [11]).

Decision and risk avoidance - Decision Quality (14; 3.8%): Difference from a rational deci- sion standard (e.g., [23, 68]). - Risk Avoidance (23; 6.2%): The degree to which a user's response attempts to avoid risk (e.g., [35]).

Trust - Confidence (24; 6.5%): The degree of belief in the validity or truth of a judgment, data set, visualization, etc. (e.g., [16]).

Qualitative - Impact on Decision-Making (3; 0.8%): Effects on decisions where no ground truth is implied (e.g., [15]). - Awareness of Judgment Process (8; 2.1%): The user's ability to recognize aspects of their judgment (e.g., [15]). - Understand internal model (48; 12.9%): Seeking to understand how a judgment is made (e.g. [13, 87]). - Efficiency (37; 9.9%): How well a visualization supports quick judgments (e.g., [59]).

Semantics - Intuitiveness/Reading ease (31; 8.3%): How "naturally" a visualization supports correct interpretations (e.g., [7]).

?? Dont care - Memorability (3; 0.8%): How well a fact is remembered (e.g., [39]). - Learnability (0): The user's ability to improve their performance over time - Satisfaction (38; 10.2%): The user's aesthetic valuation of a visualization (e.g., [22, 61]). - Interaction (4; 1.1%): How much interaction a visualization receives in time, number of clicks, etc.

L3: Evaluation goal - Compare impacts of multiple uncertainty visualizations (255; 68.5%): Results of one or more tasks are used to rank two or more uncertainty visualizations (e.g., [11, 40, 53]). - Determine the impact of presenting uncertainty (98; 26.3%): At least one visualization that does not contain uncertainty information is evaluated (e.g., [68]). - Validate effectiveness of an uncertainty vis (5; 1.3%): A user study is designed to confirm that a visualization improves some response(s). (e.g. [80]). - Understand why/how a visualization works (54; 14.5%): Re- sponses to a visualization(s) are analyzed to identify or confirm some judgment mechanism (e.g., [87]). - Understand interactions with user characteristics (53; 14.2%): The effect of properties of the user (expertise, location, etc.) on responses is analyzed (e.g., [1]).

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 visualsiation in later work. Until that later work is done, it is often difficult to accept one particular representation. While justification could come in the form of discussing established concepts in visualisation, such as the hierarchy of perceptual tasks, even these may need an evaluation study to back up their claims.

The reasoning for this is obvious. There are a large number heuristics and biases that are not obvious to us when designing visualisation. Additionally, these heuristics and baises can change depending on the larger scope of the graphic and the population we are communicating with. Additionally, since there is often a miriad of ways to visualise any particular mathematical object, to adopt specific visualisation that is not already common practice needs reasonable justification, lest we run into a range of unforseen heuristic pitfalls.

The remainder of section 4 will focus on the results and methods of uncertainty visualisation evaluation experiments. There are six main types of evaluation experiments in the uncertainty visualisation literature, they are:
- Association perceptual tasks: These experiments identify perceptual tasks that are/could be associated with uncertainty, and test its association with uncertainty, check it's number of distinguishable levels, and/or check how quickly/accurately people can draw the uncertainty

information from the graphic.
- Value perceptual tasks: Participants are shown some depiction of uncertainty and are asked to extract a value, make a comparison, or do something simple to show that the graphic is readable.
- Heuristic Checks: These experiments are checking if a perceptual bias that has been identified through public use of a particular plot, other evaluation experiments, or gut feeling identified through public use of a plot or in a different.
- Plot Comparisons: These papers describe an uncertainty visualisation evaluation experiment, where they compare two or more uncertainty visualisaitons on a specific set of questions. These questions can be about specific statistics (e.g. What is the probability that A>B?) or about a secondary goal such as trust or decision making (e.g. How much do you trust this estimate?).
- Qualitative analysis: These experiments show participants an uncertainty visualisation for an open ended or specific task, and asks them to describe how the visualisation was used to come to a conclusion, or they are asked to describe what they see in the visualisation. - Method papers: These papers discuss issues with the evaluation methods of uncertainty visualisaiton and suggest changes or improvements, typically with results indicating the benefit of the change.

These experiments are not mutually exclusive within a paper, and a single study will often contain several experiments, usually from several of these categories. These distinctions are important because each have their own issues in how they relate to uncertainty more broadly. Below we will go through the five key issues in uncertainty visualisation, describe how they mainfest in the uncertainty visualisation literature, and offer some tentative solutions to these problems.

We will also comment on some of the uncertainty visualisation papers that suggest a visualisation without an evaluation experiment. Some of these plots may offer a solution to the unique problems we mention in this paper and these visualisations will be discussed in "Part 6: Great Examples". We will also consider these papers to illustrate the issues that visualisations are aware of and working to fix, and discuss whether or not these directions of research are likely to bear fruit, and why.

> Edit note: Currently these notes are just from spiegelhater because this is the only paper I have moved notes over from. Other sources will be used here lol. Working out what exactly they are communicating is a bit of a pain though.

Risk and uncertainty are not only different in how they are communicated, but also why they are communicated. Risk communication focuses on communicating probability around events. These events are typically a binary event that is known to be random, so a version of communication that does not mention risks (e.g. telling someone they will or will not get cancer instead of communicating their risks about it) inherently carries with it a lack of transparency. Additionally risk communication is often communicated to avoid or information about an unfavorable outcome, so framing can be important if we want to influence decisions

in a certainty way. These motivations do not extend to the communication of uncertainty information more broadly.

There are some elements of risk communication that seem to translate to uncertaitny communication more broadly. Using words to communicate risk is discouraged because people can misinterpret the actual risk involved, so when possible numerical estimates of risk should be communicated, however this can make information harder to understand for those with poor numeracy skills (Spiegelhalter 2017). People performing better and preferring with numerical estimates of risk translates to uncertainty communication more broadly *inc (citations for this).* Risk communication also needs to have clear objectives, use plain language, limit information to only what is necessary, and segment the audience to allow for differences in interest and knowledge (Spiegelhalter 2017). These general concepts of communication also extend to communication of uncertainty. On the other hand, it is not sensible to assume that probability specific concepts such as framing, intuitiveness of probabilities, or reference classes are still relevant when we consider uncertainty communication as signal supression, not as a synonym of risk or probability communication. In this same vein, most uncertainty communication papers very often discuss risk-aversion as something that needs to be considered despite it being unclear exactly how risk-aversion is related to uncertainty communication *(citations for the papers that do this).*

In the same sense that risk and uncertainty are different, risk-aversion and amiguity-aversion (which can be considered the uncertainty version of risk aversion) are also not the same. Risk-aversion is the tendency of people to prefer outcomes of low uncertainty to those outcomes with high uncertainty despite the outcomes with high uncertainty having a higher or equal expected outcome. In the case of communicating risk, the only reason this heuristic would be relevant was if risk-aversion was a mistake that needed to somehow be corrected for, rather than a heuristic that *reveals the utility of decreased uncertainty.* Risk-aversion is often treated as a mistake or something that should be avoided by the uncertainty visualisation literature even though that is not necessarily true, depending on what is causing the risk-aversion (e.g. is it due to poorer estimates or increased awareness of negative outcomes). If a particular communication method leads people to place a greater value on certainty or make worse estimates, that is the aspect of importance, not risk-aversion which 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. Risk-aversion has a large number of confounding factors that make it difficult to understand if it is something to be considered in risk communication, ambiguity-aversion has a similar problem. Ambiguity-aversion is the tendency of people to prefer to take on certain risks rather than unknown risks. It is essentially risk-aversion applied to risk. For this reason discussions of ambiguity-aversion have the same issues as the conversations about risk-aversion, where they do not consider ambiguity-aversion to be an estimate of the utility of decreased uncertainty, it contains a subtext that argues against transparency, and it is considered a valuable insight in of itself despite conflating many factors.

The idea that risk and ambiguity aversion reveal the utility of decreased variance seems to

seldom be considered in the literature, therefore it should be mentioned here, otherwise this issue might continue. However, considering risk aversion is at its core, a trade off between bias and variance, it is bizzare to assume it to be a mistake when statisticians perform these trade offs all the time. If risk-aversion is truly "irrational" behaviour, every statistical textbook that discusses the use of a biased but consistent estimator in a finite sample case (a common example would be specific cases of the maximum likelihood estimator) should be tossed out. Discussing risk and ambiguity aversion as though it is some diversion from a perfectly rational, mathematical choice, implies there is *no value to a decrease in variance which is not even true within mathematics*. 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 concept of risk aversion treats risk as "noise", as though it is something that only exists to be ignored, but if the common place use of biased MLE's is an indication of anything, it is that *ignoring uncertainty* is the irrational choice.

While this is a focus on uncertainty visualisation, these issues in uncertainty communication need to be acknowledged because they bleed into the uncertainty visualisation literature. Uncertainty visualisation are compared on their ability to communicate explicit risk, elicit trust, and prevent risk or ambiguity avoidance, continuing these issues in uncertainty communication at large. While uncertainty visualisation can make it easier to communicate the complicated details of uncertainty information, it carries some challenges that are unique to the visualisaition, specifically with respect to the "signal supression" concept. In simple estimates or verbal communication, the signal is often easy to identify because it is what we are explicitly saying. 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. These unique and facinating challenges that are faced by viewing uncertainty visualisaiton through the lense of "signal supression" have been almost completely untouched by the literature.

What is meant by "uncertainty" may seem obvious to some, but when you attempt to quantify or visualise it you will quickly find yourself asking, "uncertainty about… what?". Do you mean uncertainty on an estimate? On a forecast? How many steps ahead is this forecast? Are we only considering the uncertainty in the estimate or in the parameters or are we considering the possibility of measurement error or biased inputs? Signal and noise can only be untangled in the presence of a motivating question.

The previous sections of this paper established several issues in uncertainty communication at large. In this section, we will discuss the current literature on uncertainty visualisation, how the uncertainty communication issues more broadly manifest in the specific case of visualisation, and what general lessons we can take from the existing work.

5.1.1.1 Decision making and risk aversion

Decision making questions that typically involve uncertainty discuss the concept of "risk-aversion" as though it is a mistake. The irony of an uncertainty visualisation pper expressing the importance of uncertainty visualisation, but setting up a ground truth or optimum strategy in which participants *should* ignore the uncertainty information is not lose on us. This perspective does not make sense because it implies participants should make decisions entirely based on the average. A possible soltion for this is to provide a utility framework for a particular experiment (Hullman 2016), however it is unclear how easily participants could adopt a utility framework that is different to their own. Cheong et al. (2016) introduced a very simple optimum strategy into their experiment on bushfire uncertainty visualisations, where participants should choose to evacuate a house as soon as its likelihood of catching fire is 50% or higher. 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. These papers also very rarely consider what the optimum decision would be in a given scenario, and a difficulty with anticipating the best strategy is conflated with the failings of the uncertainty visualisation. Ultimately, we agree with Hullman (2016) who suggested that what determines an approripate ground truth is largely a philosophical exercise.

5.1.1.2 Trust and confidence

Authors also occasionally use trust as a proxy for uncertainty. Similar to measuring uncertainty through decision quality, this method also has its own series of issues. 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. It does seem to be the case that uncertainty communication increases trust because it is a proxy for transparency *(add citations)*. Unfortunately a large proportion of the literature discusses trust as something that is directly related to uncertainty visualisation rather than as an observable by product. In viewing uncertainty communication as directly related to trust, not related through the proxy of transparency, several unobserved variables are conflated. The amount of uncertainty that is present in the information, whether or not the information is trustworthy or makes sense, prior beliefs of the participants, and the trustworthiness of the source are a few examples of variables that are conflated when authors directly consider trust to be of direct interest. Considering trust, and not transparency, as the metric of importance in uncertainty communication leads to a questionable subtext that argues against transparency. 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".Being concerned about audiences perception of trust, without first establishing if what we are communicating is *trustworthy*, leads to the implication that details that result in information not being considered trustworthy should be avoided. Authors that imply that a decrease in trust or an

increase in ambiguity or risk aversion are metrics of importance in of themselves do not understand the purpose of uncertainty communication as methods to increase transparency. 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).

Additionally, some studies use a ground truth in these trust experiments, which makes for a very confusing read on exactly how participants are supposed to use the information. - Zhao et al. (2023) displayed a model prediction with 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 uncertaitny information should somehow influence participants to use their own prediction. If the Beuro of Meterology was uncertain about their rain prediction, that would not be sufficient conditions for me to decide to make my own rain predictions. 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 infororation should infludence the choices of the participants.

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.

5.1.1.3 Deterministic questions

- You cannot make a deterministic statements about a random variable (citation: the entire field of probability). Many of these issues just boil down to a really pervasive misunderstanding of statistics. I do not expect the general population to understand this fact about random variables, however *I would expect* researchers investigating *methods of statisticl communication* would at least understand enough to ask appropriate questions that *are answerable*.
- while we are here, you cant answer a probability question with a deterministic. I cannot say yes or no to a probability question without providing a threshold.
- Basically, you need a probability threshold to answer these questions, something it does not seem like the authors are collecting, and I am unsure if the authors are aware of this themselves because they never mention it.

5.1.1.4 Other (questionable) attemps to capture uncertainty

There are other studies that try to include uncertainty information without directly asking for it but do not go down the route of asking for a decision or levels of trust. These papers often try to ask a question that should utilise both the uncertainty and signal in the response. This is an interesting approach, but these papers are often still looking for a "signal" which is specified in the form of a "ground truth" that is used to evaluate the plot. Unfortunately this usually comes in the form of slightly cryptic questions that create a large amount of noise on the interpretation side (Hullman 2016).

Cite gap ___: Weird questions to capture uncertainty - Correll, Moritz, and Heer (2018) "place your 5 ships on the safest locations on the board." (minimise danger) - Ibrekk and Morgan (1987) asked participants for the "best" estimate while displaying a population density function. The participants proided the mode instead of the mean, although there was no indication the mean was what was being asked about. - Hofmann et al. (2012) showed two distributions in 20 different visualisations (a lineup protocol) using a jittered sample, a density plot, a histogram, and a box plot. Participants were asked to report in which of the plots was "the blue group furthest to the right" and to provide reasoning from a multiple choice list and grade their certainty. The experiment set up is shown in Figure 8. The participants answers were then compared to a ground truth where the correct plot had 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.
- L. M. Padilla, Ruginski, and Creem-Regehr (2017) made several confusing assumption when determining the ground truth of their experiments that were assessing storm visualisations. Experiment 2 assumed that the participants should provide responses that indicated they perceived the intensity of the storm to be independent of the uncertainty distribution of its location. This assumption seems counterintuitive as best, however experiment 3 asked participants to "decide which oil rig will receive more damage based on the depicted forecast of the hurricane path" for which they *were* supposed to incorperate uncertainty information and not assume they were independent. The authors seemed to be unaware of these confounding mistakes in their assumptions of how the participants were supposed to utilise the uncertainty information.
- 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 extract the mean and variance from four uncertainty visualisations, bar charts with error bars, box plots, gradient plots, and violin plots (that were adjusted to make sure the mathematical information they were showing was the same) from two side by side distributions. Participants were also asked to answer a question that was supposed to incorperate both the signal and the noise, such as "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 favor of A to 7=Outcome will be most

in favor of B. The ground truth statistic for this question was a scalar multiple of Cohen's d, indicating participants were supposed to incorperate uncertainty information using a very specific formula that was likely unknown to them but assumed to be used implicitly. - 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 didnt 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%. The authors failled to recognise that this "complicated decision making task" boilled down to a simple value extraction problem, which the text made easiest to extract. The results from the experiment indicated that many of the participants also failled to recognise the obvious and simply strategy.

- Blenkinsop et al. (2000) tailored their questions according to the visualisation shown, which implies an understanding that different visualisations contain different information. Despite this the authors still compared the visualisations according to which were "useful for uncertainty visualisation".
- Blenkinsop et al. (2000) did try to include uncertainty as noise, asking participants to search for a specific outcome (the land classified as grass) in a random . However the participants failing to identify the signal (something that would be expected of an uncertainty plot) was seen as a sign the task was too complicated. While the task in this experiment *was* certainty far too complicated (and poorly communicated) this discussion indicates the authors did not consider the signal suppression to be the goal of the visualisation, and rather an annoying distraction.
- Ibrekk and Morgan (1987) asked participants for the "best estimate" which was provided in terms of most probable value (mode), the mean and the median, the "ground truth" statistic used was the mean, despite the "best estimate" largely depending on the loss function of the model and the methods used.
- Ibrekk and Morgan (1987) used several displays of a probability density functions, since "formally equivalent representations are often not psychologically equivalent".
- Ibrekk and Morgan (1987) highlights that in the face of a vague question, participants will use the plot to decide what the authors mean. The authors take this as a
- Boone, Gunalp, and Hegarty (2018) organised responses from participants by the approach they used and tried to account for participants prior beliefs (although the assessment of those beliefs used the word "likely" which is open to interpretation).

There is so little understanding of information in uncertainty graphics that participants are frequently provided with information that is insufficient to answer the question at all, an issue that is regularly missed by the experimenters. In order to test whether or not people

incorrectly interpreted the cone of uncertainty that is used to communicate cone tasks, experimenters regularly do not give participants enough information to answer the questions (L. M. Padilla, Ruginski, and Creem-Regehr 2017; Boone, Gunalp, and Hegarty 2018). The cone of uncertainty provides a 95% confidence interval for the *eye* of a hurricane, which allows us to know *where the eye of the storm will likely go*, this does not, give us enough information to answer if the storm will *hit* a particular point, because this requires information about the size and intensity of the storm, *and* an assumption about whether or not the size and intensity will change over time. Since these experiments are often trying to *test* for common misreadings of the plot, such as only thinking the area inside the cone will be damaged or believing the cone represents size and the storm is getting bigger. The authors often fail to recognize that the exact information that has been witheld for testing is necessary for answering their questions, and considering that information is necessary to mark participants as correct or incorrect, it is hard to understand exactly how they evaluated the experiment participants. Ibrekk and Morgan (1987) expected participants to calculate the mean using a pie chart that had numerical bins combined (e.g. segment one was 0-2 inches of snow, and segment two was 2-4 inches of snow) so it could not be done by calculation. The authors also expected participants to use a CDF to calculate the mean, which needs to be converted *back* into a PDF to do so mathematically, so it was unclear how participants were supposed to do this calculation visually unless they were to just make a blind guess. The authors interpreted this as the participants misinterpreting how the CDF works, not seeming to realise they gave them an impossible task. In light of this, the main take away from these papers seems to be "if you do not give people the necessary information to answer a question, they will be unable to answer it", which does not seem interesting enough to justify several papers.

5.1.2 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 bluriness 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 guarentee that the uncertainty will act as an approriate signal supressor. 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-supressing uncertainty pallet (Correll, Moritz, and Heer 2018).

Value-supressing 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

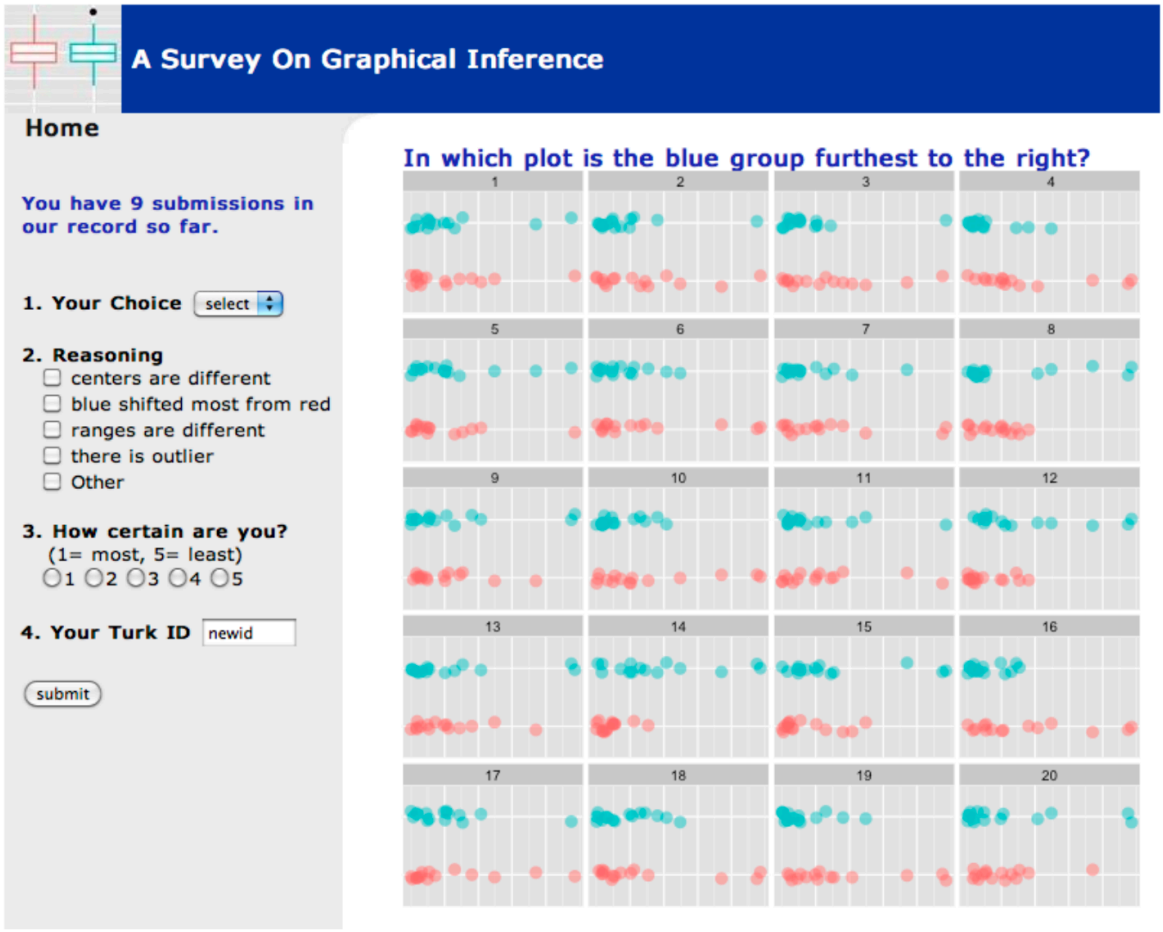
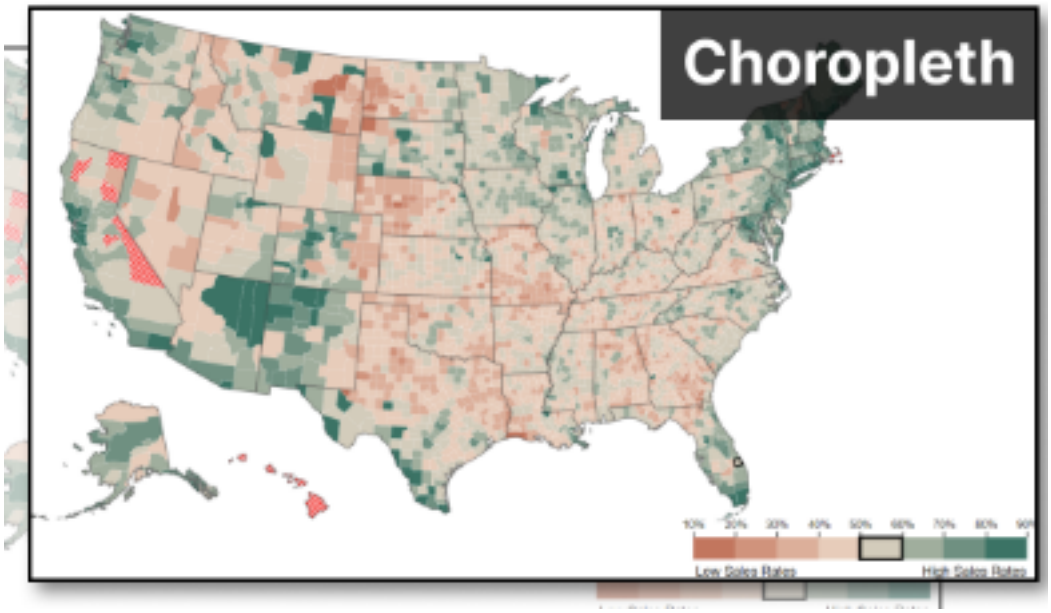
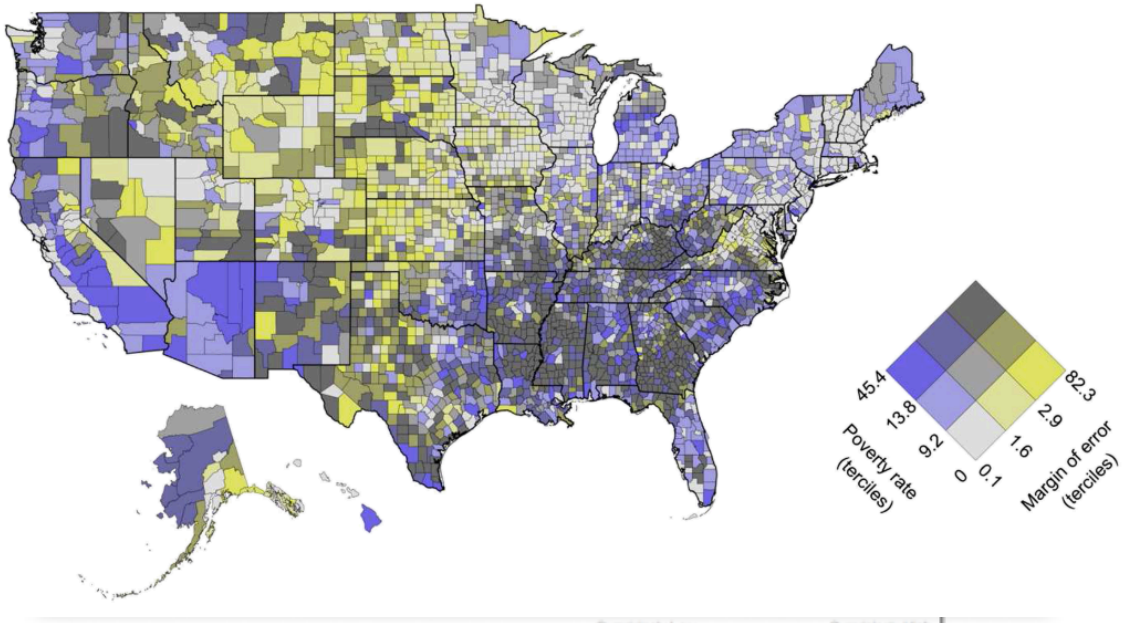


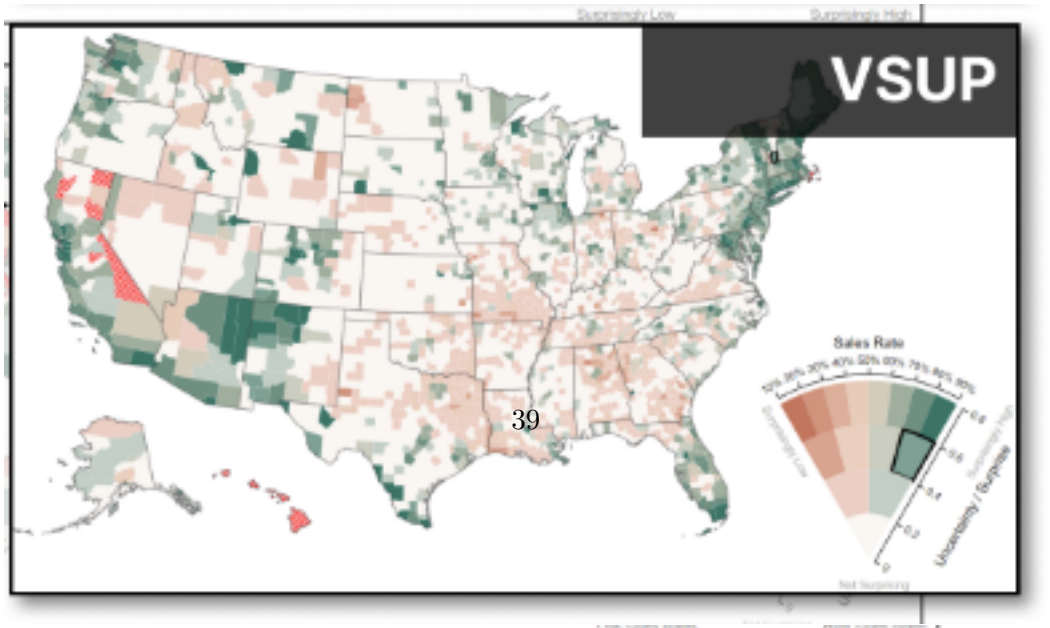
Figure 8: This shows the user interface for the experiment performed by Hofmann et al. (2012). The question of "furthest to the right" is open to interpretation.

are appropriately *supressed* by the uncertainty. Hence the name of the pallet. Figure 9 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 supression 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 doesnt 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 supressing 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 "supressed signal" is represented by a single variable. This statistic can then be expressed in a one dimensional colour space, which is a method addopted by the Baysian surprise metric map (Ndlovu, Shrestha, and Harrison 2023) and the excedance probability map [Lucchesi2017]. .

These maps make the importance of combining uncertainty and signal in a single visual channel clear. A chloropleth 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 supressed signal as a single variable, rather than visualising uncertainty separately using two different channels.


Choropleth
Low Sales Rates
High Sales Rates


45.4
13.8
9.2
0
0.1
1.6
2.9
82.3
Poverty rate (terciles)
Margin of error (terciles)


VSUP
Sales Rate
Uncertainty / Surprise

5.2 Suggestions to measure uncertainty

5.2.1 Heuristic checks

- Have a better hypothesis? Like its more specific. They are looking for a very specific thing *wrong* with a visualisation.
- We wary of trying to control something and observe it at the same time

A lot of work that identifies heuristics or biases in plots do not hinge on the previous assumptions. 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)

5.2.2 Just noticiable signal

Connecting our evaluation of uncertainty back to the reasons uncertainty is believed to be different is likely the key to improving evaluation studies. 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 expressed as being similar to fraud or lying and the goal of "improving decisions" is only seen as a secondary outcome of an appropriately hedged signal (Hullman 2020; Manski 2020).

5.2.3 Thinking "smaller picture" with

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. Let us consider a decision making experiment, there are several steps that Simply asking "Do people make better decisions with

Experiment rant from confirmation

5.2.4 Comparisons to lineup plots and hypothesis testing

It is unclear how participants are supposed to incorperate uncertainty information into their responses when uncertainty visualisation authors themselves are unsure. Other authors interpret this difficulty as a combination of the uncertainty and task complexity (Kinkeldey, MacEachren, and Schiewe 2014), and not as a general confusion. All these examples sever to illustrate that *we dont know how to measure uncertainty* and if we don't know how to measure it, *we cannot evaluate the quality of a visualisation*. It could be argued that a well done uncertainty visualisations should have an imperceptible signal unless the signal would be identified with a hypothesis, but this idea also has a miriad of issues that come with it. There is already a series of issues with the reject/do not reject concepts in hypothesis testing, so it is already questionable to set the results of a hypothesis test as the ground truth in an uncertainty visualisation experiment, however these may be additional problems that go beyond the binary nature of the issue. (**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. This increased sensisitivity could be the result of another aspect of uncertainty that is ignored. Statstical tests are typically built upon a series of assumptions, and it is difficult to identify *which* assumption failure caused the data to fail the statistical tests. Lineup plots allow us to see *how* our data is different to the assumed distribution, and better understand *why* our data may have failled a test [1]. These experiments utilised the lineup 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. **?@fig-hypyvs** shows the conceptual difference between the lineup protocol and uncertainty visualisation. A lineup protocol displays the uncertainty about the null and identifies if the true data plot is identifiable (and therefore significiantly difference) while an uncertainty plot displays the variance of the estimated value and assesses if the null of "no signal" is within this plot (e.g. if the error bars overlap with zero). [2] This implies a visible signal in an uncertainty plot should indicate some divergence from the null.

The problem of trying to assess uncertainty without knowing *why* we are depiting uncertainty has been noticed by other authors. Spiegelhalter (2017) noted we "cannot assess the quality of risk communication unless the objectives are clear".

6 Great Examples

> Edit notes: this section seems to go off a bit on dimensionality of visual tasks and stuff that might be better put in the noise and signal section about measuring

[1] Did Patricks paper compare statistical tests to lineup protocols in the event an earlier assumption of the tests, such as, independent observation? I think he did check some. I wonder if you can quantify the information difference you get from a rejected statistical test vs a rejected line plot

[2] Does this flip have mathematical implications in how the plots should be assessed?

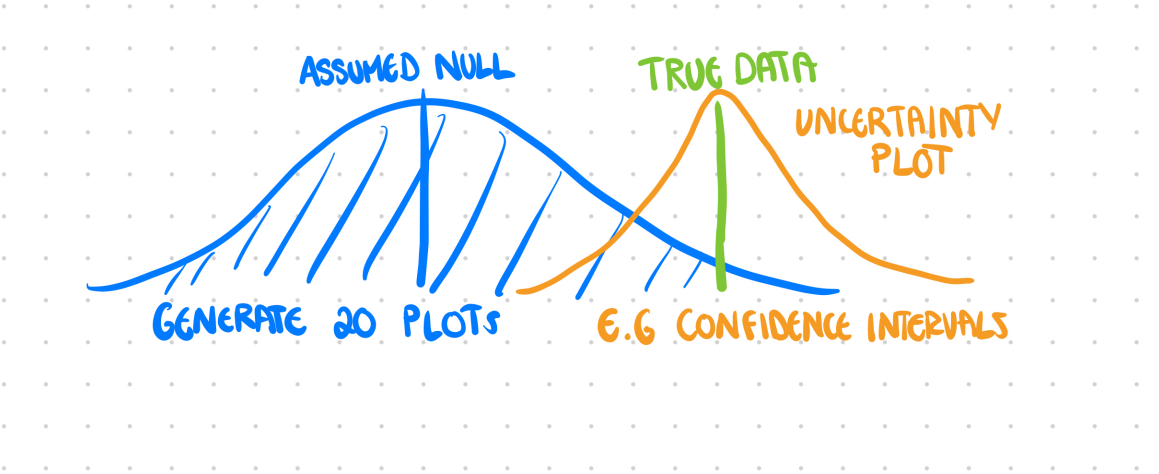


Figure 10: Difference between the null plot vs the uncertainty plot.

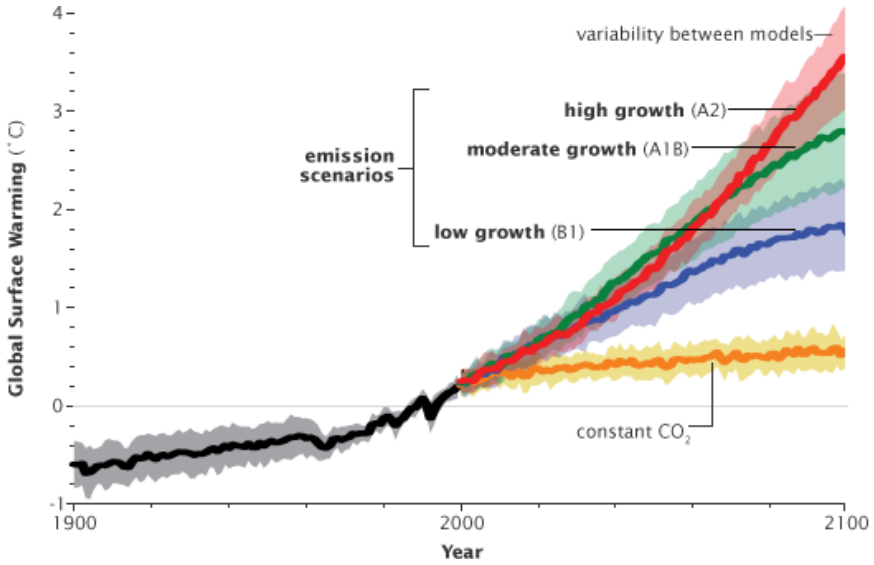
uncertainty.

Of course we do not believe every uncertainty visualisation is unusable and every experiment is build upon misguided assumptions. Despite the common problems detialled 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 identifies or fixes pitfalls in the previous sections (without identifying them) or does an experiment that identifies useful information that avoids these pitfalls.

Uncertainty that is created by steps that are further down the pipeline are often ignored by the uncertainty visualisation literature, however there are visualisations that make these considerations. If uncertainty had a more encapsulating mathematical definition, the uncertainty created at earlier stages might be able to be visualised with the model uncertainty that is typically expressed through confidence intervals. That being said, even with a definition that allows us to combine these uncertainties into a single value, it still might be desirable to see which stage of the analysis is creating uncertainty in the visualisaiton. One example of a visualisation that includes that information is visualisations of climate scenario uncertainty, as shown in **?@fig-climatescenario**. Uncertainty is introduced into this model by the choices of individuals which is difficult to model or predict and is also unlikely to be constant over time. Therefore specific scenarios are modelled and shown, each with their own model uncertainty. This highlights a method that can be used to visualise uncertainty from earlier in the pipeline that is often forgotten or ignored via assumptions.

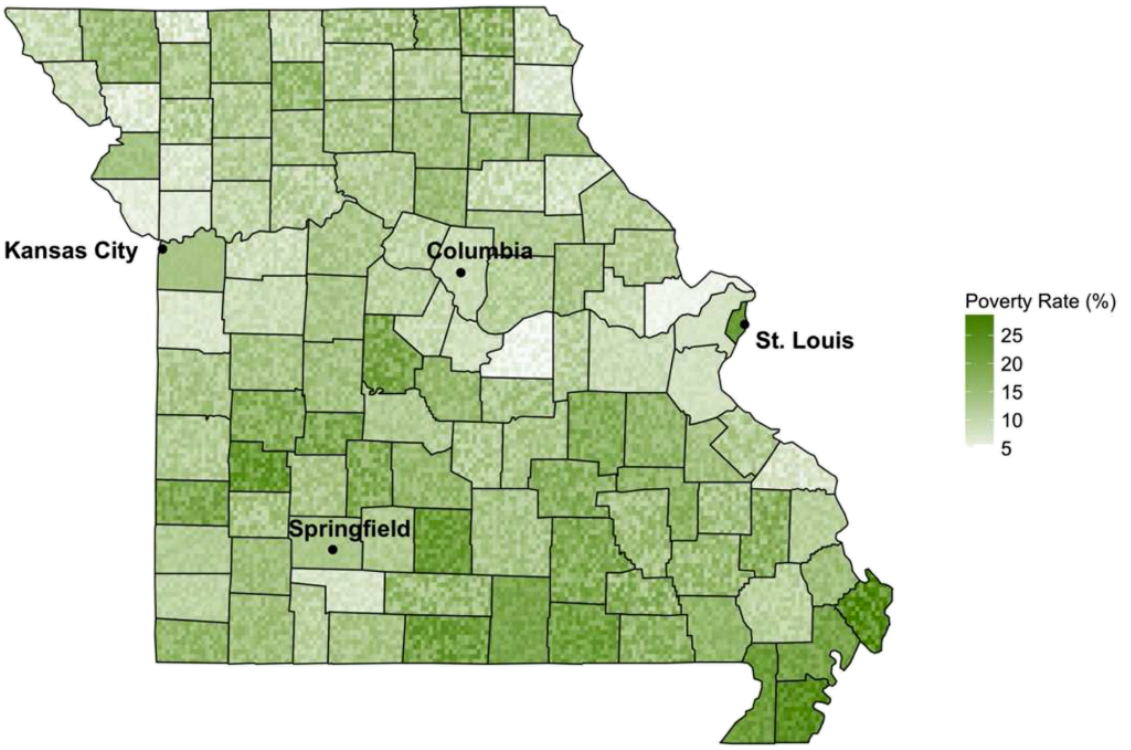
There are also some interesting qualitative studies that take an investigative stance to the use of uncertainty in decision making rather than an evaluation stance. Daradkeh (2015) presented participants with ten investment alternatives and asked participants “from among available alternatives, which alternative do you prefer the most”, and were asked to think aloud

and consider the uncertainty in their decision making. This study was qualitative and rather than setting a "correct" answer that would have forced a value of risk on the participants, 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.

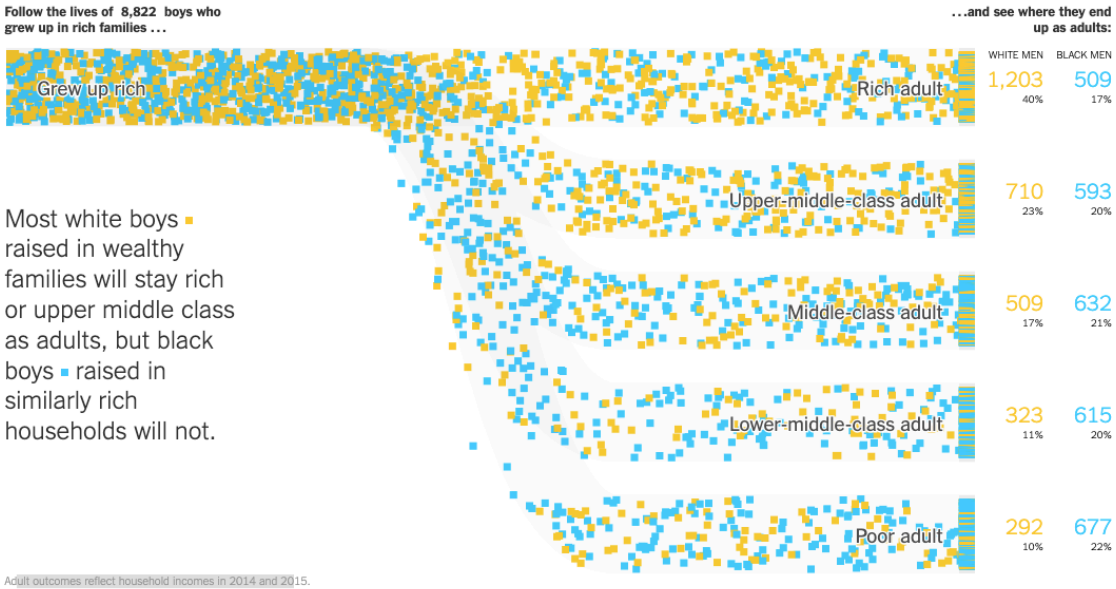


{fig-climatescenario}

Another method that has been used is to visualise samples from the distribution we are trying to draw inference on instead of estimates along with a variance or error. This has been used in maps with the pixel map [Lucchesi2017], but is more commonly used in animation with the HOPs plots (Hullman, Resnick, and Adar 2015) or similar concepts (Blenkinsop et al. 2000) . The difficulty with resampling methods is that they are typically conflated with animation even though this does not necessarily need to be true, as can be seen in the **?@fig-pixelmap** and the New York Times class mobility figure, shown in **?@fig-nyt**, an animated version of which can be found here. Therefore, so long as the sample size of the number of outcomes is kept to a level where individual samples are still visible, there is no evidence that an outcome plot *needs* to be animated. Visualisations that opt to express a signal as a sample rather than an estimate have the potential to supress 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 supression, 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 the messiness of the data that sits behind a model.



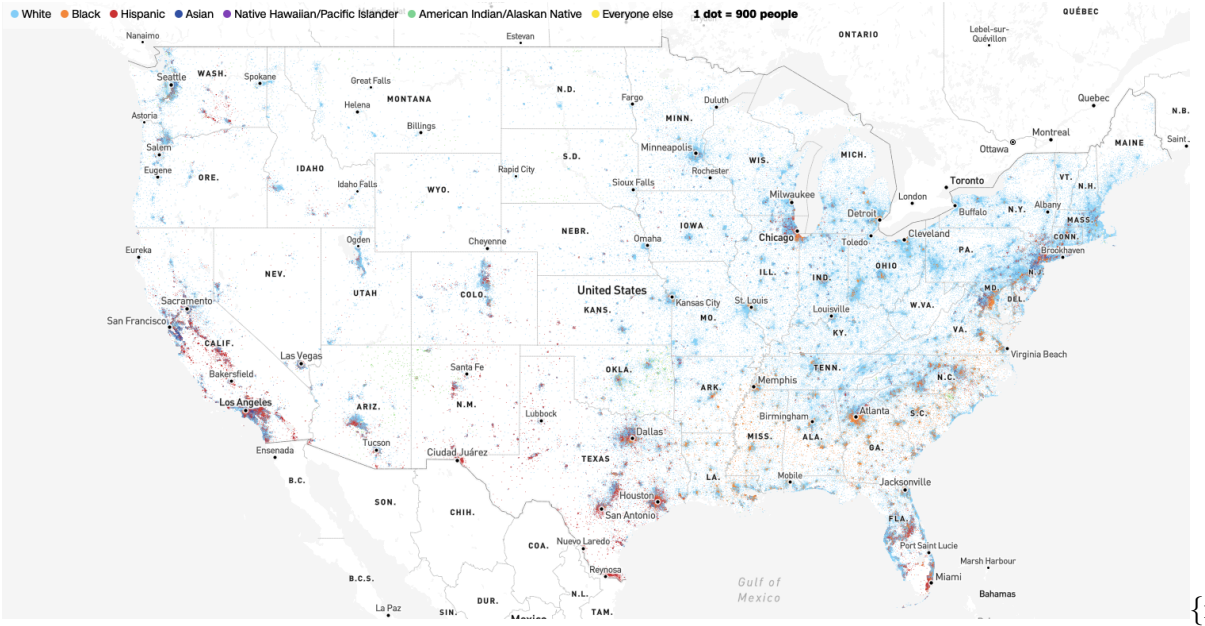
{fig-pixelmap}



{fig-nyt}

A final method to perform signal supression is simply to visualise the data if it is available. An example of this is shown in **?@fig-census** 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, i.e. 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. The raw data can also be used in a similar method to the HOPs plot, where a statistic of importance is implied by the visualisation, but the signal is supressed using the data itself instead of resampling methods. 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 lineup protocol was implemented.



{fig-census}

These great examples show there are visualisations that are being made that have a good conceptualisation of uncertainty and express it as signal supression, however this work is not being done formally within visualisation. These industry visualisation are likely made with a large amount of back and forth between different levels of an organisation, something that is important for good visual communication (Spiegelhalter 2017).

7 Future work

This paper has identified several issues in the uncertainty visualisation literature and with those issues, we have several suggestions for future work that would likely be fruitful for the field.

Our first suggestion for future work is a mathematical defintion of uncertainty. Specifically, a definition that has a place for all the concepts that currently sit under the "uncertainty" umbrella, explains how these concepts relate to uncertianty, and also allows us to quantify them in a way where they can be combined or visualised. This work would not only benefit uncertainty visualisaiton, but also statistics more broadly. There is currently a lot of work that quantifies the uncertainty through bias and variance at various stages of the data analysis pipeline separately. Elements such as imputed data, assumptions, sampling methods, and analyst decisions all have their own independent work, quantification and discussions, but methods to brand between these concepts are few and far between. A conceptual framework that allows us to combine these results would be incredibly beneficial to the field.

The concept of uncertainty should also be formalised within the grammar of graphics. Not only would a formalisation of uncertainty visualisation within the grammar of graphics allow us to iron out some of these confusions, but they would make it easier to understand when existing visualisation methods apply and can be used to explain our results.

If we are going to consider uncertainty as noise, not signal, there needs to be a way to identify this signal supression in an experimental design. It is clear that value retrival and questions with a specific ground truth are not appropriate, as uncertainty visualisation will simply follow the existing visualisation rules. Testing a secondary feature of a plot, that is, how strong a signal is becomes a little more complicated. Line up protocols have been used in adjacent work that look at the strength of the signal in a plot, and this idea of identifying if plots have some "barely noticiable differences" could be utilised. 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.

Additionally, information visualisation needs a more precise concept of the difference between two plots, and that framework should be utilised in experiments. 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. Visualisation has the grammar of graphics, however we have already discussed how this can break down in the case of distribution visualisation.

If a visualisation researcher would prefer to perform experiments rather than formalise methods, the task dependency many authors in uncertainty visualisation mention would be a useful direction for research. Unfortunately the current visualisation evaluation literature is *far* too noisy to allow us to identify what might be the cause of this task dependence. It is unclear if it is a conflation between task and the ground truth/displayed statistics concepts discussed earlier,

or an actual difference in the way we percieve graphics. A good starting place for this idea would be to identify if there is task dependency in perceptual tasks, since all perceptual task research to date has been done on value retrival. That is, an experiment that compares things like area, position, colour and other perceptual tasks on value exaction, comparisons, and other simple tasks. It is also 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 isnt 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 experiemnts. 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. Reguardless, 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.

8 Conclusion

In this paper we have highlighted a series of misunderstandings, confusions, and methodical errors in the uncertainty literature.

Many of these issues are not problems in isolation, so long as the authors themselves are aware of it and state these assumptions in their papers. Our problem with this body is work is that the authors themselves seem to be oblivious to this vast array of problems, considering not a single one of the problems discussed here are mentioned explicitly in a single visualisation paper.

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.

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.

Boone, Alexander P., Peri Gunalp, and Mary Hegarty. 2018. "Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations." *Journal of Experimental Psychology: Applied* 24 (3): 275–95. https://doi.org/10.1037/xap0000166.

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.

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.

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.

Dyson, Freeman J. 2010. "Birds and frogs in mathematics and physics." *Physics-Uspekhi* 53 (8): 825–34. https://doi.org/10.3367/ufne.0180.201008f.0859.

Fischhoff, Baruch, and Alex L. Davis. 2014. "Communicating scientific uncertainty." *Proceed-*

ings 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 Transac-*

tions on Visualization and Computer Graphics 25 (1): 892–902.

Kay, Matthew, Tara Kola, Jessica R. Hullman, and Sean A. Munson. 2016. "When (ish) is my bus? User-centered visualizations of uncertainty in everyday, mobile predictive systems." *Conference on Human Factors in Computing Systems - Proceedings*, 5092–5103. https://doi.org/10.1145/2858036.2858558.

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.

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.

———. 2021. "Enhancing (publications on) data quality: Deeper data minding and fuller data confession" 184 (4): 1161–75. https://doi.org/10.1111/rssa.12762.

Munzner, Tamara. 2009. "A nested model for visualization design and validation." *IEEE Transactions on Visualization and Computer Graphics* 15 (6): 921–28. https://doi.org/10.1109/TVCG.2009.111.

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.

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.

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.

Slingsby, Aidan, Richard Reeve, and Claire Harris. 2023. "Gridded Glyphmaps for Supporting Spatial COVID-19 Modelling." https://www.staff.

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.

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.