Reasoning about uncertainty—broadly construed as considering a distribution of possible events or states of the world—plays a crucial yet understudied role in how people make sense of data. For example, in the past two years laypeople and analysts alike used visualizations of covid-19 data to grapple with questions about when and where cases were increasing, to inform decisions about personal risk and public health. *Often the way interfaces present data can invite failure modes of human reasoning with uncertainty*, especially the tendency to ignore or downplay possible interpretations of data. Design choices in covid-19 visualizations, such as comparisons of case numbers between selected geographic regions—emphasize specific interpretations of data and make it easier to ignore others, contributing to divergent perceptions of risk. Visualization authoring tools and design guidelines largely fail to account for the cognitive processes people use to interpret a given chart and how they may interpret data differently in light of their prior knowledge and beliefs. More broadly, data interfaces tend to completely omit latent uncertainties—e.g., reasonable alternative ways of analyzing data which might produce different results—inviting inferences and decisions based on limited information. Given the increasing scope and complexity of modern data science, people who work with data require tools that elevate uncertainty and help them reason with it more explicitly.

In my research, **I create visualizations and analysis software to help people reason with uncertainty in data**. I do this in two ways: By measuring, modeling, and theorizing about human behavior with data interfaces, I develop a more rigorous and flexible science of visualization design. Drawing inspiration from theories in psychology, economics, and statistics, I prototype data analysis tools to test design hypotheses about how to promote careful judgements with uncertainty in data. My research won multiple paper awards at IEEE VIS in recent years for contributing new ways to measure and model user behavior with uncertainty visualizations. My ongoing and future work puts these empirical findings and the theories that inspire them into practice, contributing interfaces that can change how we build data analysis software. I aim to develop tools that both align the design of data visualizations with natural human capacities for reasoning about uncertainty and support scientists in surfacing uncertainties about their data that may otherwise be ignored.

RETHINKING EVALUATIONS OF PEOPLE'S BEHAVIOR WITH VISUALIZATIONS

The dominant paradigm in visualization research assumes that visualization *effectiveness* can be described in terms of the ability of the average user to discern the data values encoded in a chart [1]. For example, studies suggest encoding data values as position rather than color or area because the average user can compare positions faster or more accurately than colors or areas. Visualization recommender systems behind popular commercial software like Tableau rely on these performance averages to rank and suggest possible data encodings [12]. This notion of effectiveness makes assumptions about human behavior which I problematize in my work: (1) that effectiveness is about atomic tasks such as reading and comparing values from charts, rather than the kinds of applied inferences and decisions which are especially common when reasoning with uncertainty; and (2) that evaluations can support generalizable visualization design recommendations by averaging over variations in the way people think, disregarding the influence of cognitive mechanisms such as strategies. I create and impel visualization evaluations that **measure effectiveness in ways that are theoretically meaningful for applied tasks** such as inferences and decisions under uncertainty and **model heterogeneity in people's reasoning with visualizations**.

Prevailing notions of visualization effectiveness fail to account for how lay audiences make inferences about underlying trends in noisy data. Such inferences are prevalent in data journalism, for example, when members of the public attempt to infer whether there is a growth trend in noisy time series data such as covid-19 cases or the monthly jobs report [6]. I ran a series of experiments where I showed Mechanical Turk workers reference visualizations of growth versus no growth in the monthly jobs report with sampling error (Fig. 1, top), and I asked them to judge whether jobs were likely to be increasing in example charts (Fig. 1, bottom). Whereas a typical uncertainty visualization evaluation would measure people's speed or accuracy [4], I adopted

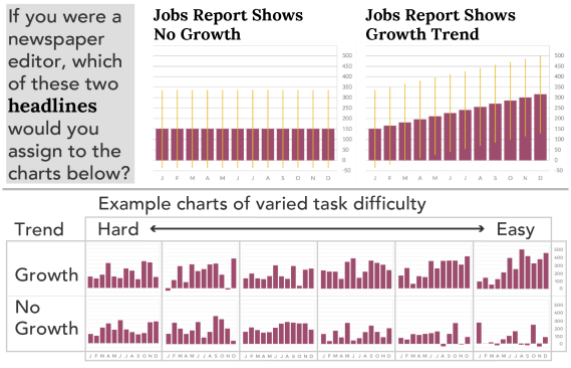

Figure 1. Task for my experiment on detecting trends the jobs report. Notice how difficult this task can be when using error bars as a reference for sampling error.

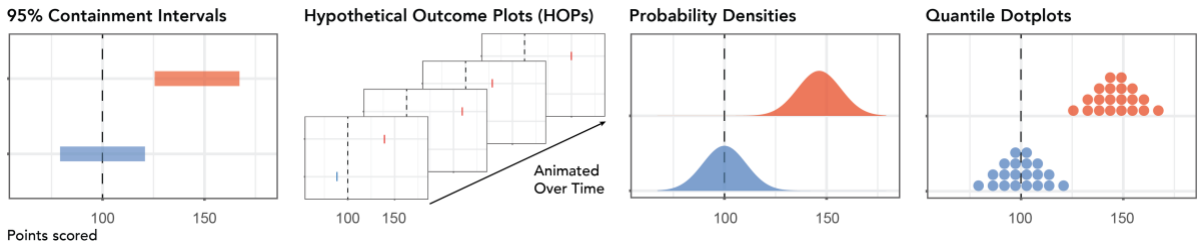

Figure 2. Uncertainty visualizations evaluated in my study on visual reasoning strategies for decision-making.

psychometric models from perceptual psychology to compare people's *sensitivity to evidence* of a growth trend in the data when showing them different kinds of reference visualizations. Specifically, I hypothesized that people would be able to infer the correct trend in more ambiguous examples when they referenced hypothetical outcome plots (HOPs, animated sequences of possible outcomes [5]), rather than static depictions of sampling variability such as error bars. My results verified this hypothesis, suggesting that experience-able visualizations—showing possible realizations from a data generating process—calibrate people's expectations about what a pattern might look like better than an aggregated depiction of those same realizations. In contrast, an evaluation of mere accuracy would have struggled to detect a difference between visualizations and would not have addressed perceptual calibration.

Surprisingly little visualization research to date applies decision theory to evaluate visualizations or considers the role of people's visual reasoning strategies—what they attend to in visualizations and how they formulate heuristic judgments with a chart. Perhaps this is because visualization evaluations tend to focus a-theoretically on metrics like accuracy, response time, or user satisfaction [4], reflecting the view that effectiveness is about straightforward perceptual optimization rather than *what chart users actually do* with information they extract from charts. I investigated how crowdworkers use uncertainty visualizations (e.g., Fig. 2) to make incentivized decisions, drawing on methods from behavioral economics to model how chart users balance monetary costs and payoffs with probabilities of events [8]. I found that people satisfice with visualizations, often adopting oversimplified strategies which lead to biased decisions that fail to optimize monetary payoffs. For example, most people tend to compare probability distributions by focusing on the distance between them on a chart and ignoring their variance. I also found that people are prone to switch between different strategies they use to decode a visualization. Both satisficing and strategy switching were novel findings, overlooked by previous visualization evaluations which averaged over heterogeneous user behaviors to derive misleadingly oversimplified rankings of visualization effectiveness. These findings led me to propose strategy-aware models of visualization effectiveness as the basis for a new generation of visualization recommenders, which will more carefully align likely data interpretations with the intent of visualization designers. This work won the InfoVis Best Paper Award at IEEE VIS 2020.

SUPPORTING MODEL-BASED REASONING IN VISUAL ANALYTICS

Theories of statistical inference suggest that analysts test interpretations of data by comparing counterfactual patterns with observed data [3]. However, with limited support for statistical modeling in many visual analytics (VA) tools, analysts must imagine these counterfactual patterns and make such comparisons in their heads (Fig. 3). I pursue a vision of VA tools that support model-based reasoning by **innovating evaluation methods for VA** and **prototyping visualization software that puts statistical theory into practice**.

How well do typical VA applications, optimized for exposing patterns in data, support the kind of counterfactual comparisons required for causal inference? Measuring the quality of causal inferences with visualizations requires a "normative" benchmark for users' inferences, however, in most applied settings the data generating process is unknown, making it

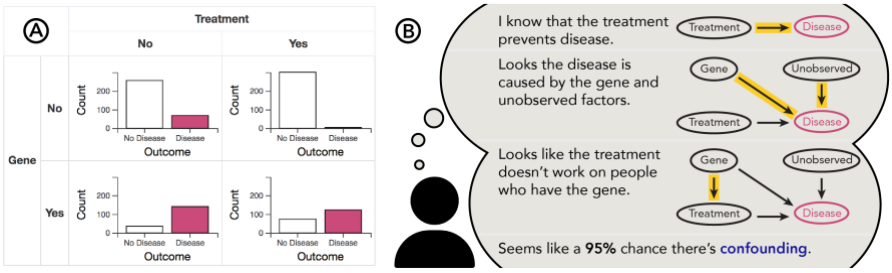

Figure 3. Imagined counterfactuals: (A) the user inspects a faceted bar chart; (B) the user builds up a causal explanation by reasoning about how well the data matches a series of counterfactual predictions.

impossible to define an *absolute* ground truth. Drawing on mathematical psychology [2], I devised a way to benchmark the quality of chart users' causal inferences *relative* to the likelihood of the data under each of a set of possible data generating models they consider [10]. I conducted a pair of experiments where I elicited crowdworkers' causal inferences from visualizations by asking them to allocate probability across a set of alternative causal explanations. I found that people struggle to make causal inferences with typical VA tools, faring no better with interactive graphics or simple bar charts than with text contingency tables. By analyzing people's sensitivity to the visual signals that differentiate between competing causal explanations, I identified numerous pain points for causal inferences with visualizations, namely that people don't know how to weigh sample size in their inferences and that they struggle to make comparisons between visualized data and counterfactual patterns they imagine in their minds' eye.

To make mental models of data generating process explicit in VA, I am leading an effort around new systems that enable analysts to express provisional statistical models and visually check their compatibility with data. The system adds operations for specifying and comparing regression models in a Tableau-like interface, and it provides automated assistance for recognizing discrepancies between data and model predictions as a stepping stone for model exploration. This research represents a proof-of-concept for novel ways of designing for model-based reasoning in exploratory data analysis, and it will seed a whole research agenda around integrating models into visual data analysis.

REPRESENTING & ENCOURAGING INTERACTION WITH LATENT UNCERTAINTY

The economist Charles Manski coined the term "incredible certitude" referring to an emphasis on point-predictions and -estimates in science communication which implies unwarranted certainty [13]. Data scientists ignore latent uncertainty when they report results from only one analysis, although alternative analyses might be sensible and yield different results. I build tools to help analysts and scientists consider latent sources of uncertainty in data analysis by **creating explicit representations of often-neglected uncertainties** and **testing design patterns that encourage interaction with these uncertainties**.

I am the student lead and primary designer on a multi-year project to build software for the Navy to help scientists conduct scientific review, synthesize the relevant literature, and make policy recommendations to decision-making officials. For the Navy, recommendations often concern the effectiveness of training programs, but other institutions use similar methods to pool evidence on interventions from medical treatments to novel technologies. I began with a formative interview study [7], published in ACM CHI 2019, where I talked with scientists who do applied research synthesis in academia, healthcare, and the Navy. My analysis showed that scientists are prone to ignore or throw up their hands at epistemic uncertainties about study quality, in part because current software does not provide representations to help users track such concerns, reason about their importance, and gauge their impact on quantitative study results.

I hypothesized that a guided process for reasoning with epistemic uncertainties about study quality would lead users to incorporate these often-overlooked uncertainties into scientific review and *meta-analysis*, a statistical procedure for combining quantitative evidence across multiple studies. To test this design hypothesis, I led the creation of a software prototype for research synthesis that provides interfaces and a workflow (Fig. 4) for: (1) defining a research question in terms of the effect of an intervention on an outcome of interest and scoping a review of relevant literature; (2) briefly reviewing of each article, extracting statistical information needed for meta-analysis as well as the kind of epistemic uncertainties that typically derail a meta-analysis; (3) triaging epistemic uncertainties and deciding what counts as solid evidence; and (4) combining study results in a meta-analysis. User feedback on the system indicates that its

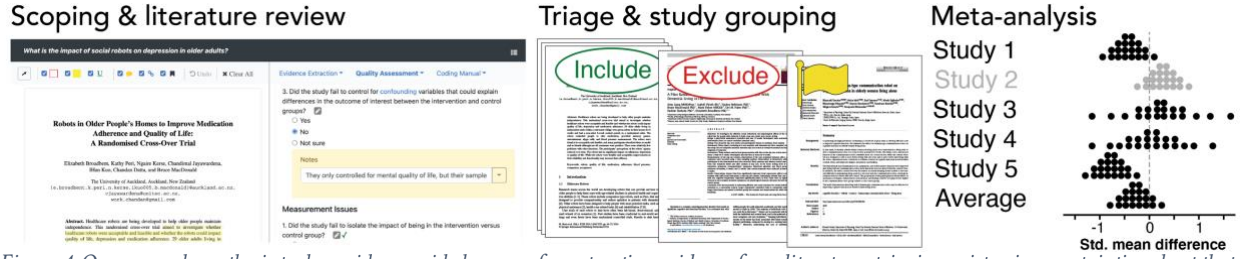

Figure 4. Our research synthesis tool provides a guided process for extracting evidence from literature, triaging epistemic uncertainties about that evidence, and combining results across studies in a meta-analysis.

design promotes transparency through documentation of analysis decisions and "changes the paradigm in terms of trying to expose what is typically neglected". This work is in submission.

FUTURE PLANS

I will continue pursuing a research agenda focused on mutually aligning human data cognition about uncertainty with representations of data and statistical models in analysis and visualization software.

Strategy-aware visualization recommendation. My research has shown that the models underlying recommender systems in visualization software like Tableau neglect visual reasoning strategies [8]—how users extract and reason with information in charts— and that the resulting visualizations are insufficient for crucial sensemaking tasks such as causal inference [10]. I envision mixed-initiative visualization authoring interfaces that help designers take potential strategies into account both by (1) simulating possible interpretations based on a corpus on known strategies for a given task-encoding pair and (2) suggesting alternative design choices, such as fiduciary markings or changes to layout, that might bring the most likely interpretation of a chart in line with the designers' intent. How can visualization researchers coordinate to accumulate a sufficient corpus of data on the myriad task-encoding pairs that such a system might require? How do we elicit a visualization designer's intent, and does doing so fundamentally alter the creative process?

Beyond "incredible certitude" in explainable ML. Prevailing notions of explainability in ML systems suggest that users benefit from accounts of how models make predictions, however, empirical research shows that such explanations actually make users *less skeptical* of a model even when it makes mistakes [14]. How much does the frequent omission of model uncertainty from these explanations contribute to this blind acceptance of ML? When I interned at Microsoft Research with Rich Caruana and Jenn Wortman-Vaughan, I researched ways to estimate and visualize uncertainty in a class of glass-box ML models called explainable boosting machines. This work in progress anticipates a burgeoning research agenda investigating how statistical methods and interactive visualizations can promote deliberate reasoning about uncertainty in ML.

Visualization to calibrate expectations. Data-driven applications tend to engage with users' expectations, often implicitly, but when expectations are miscalibrated, it can lead to misunderstandings—for example, mass-confusion about the influence of mail-in voting in the 2020 presidential election. Prior work examines the role of expectations in visualization interpretation (e.g., [11]), however, very little work to date attempts to use visualization to calibrate these expectations. While my work suggests that HOPs can help people internalize expectations under a particular data generating process [9], many questions remain unanswered about the nature of visual expectations and their ties to declarative prior knowledge. What design patterns are necessary to promote strong ties between a visual expectation and a conceptual hypothesis about data? Are pre-existing beliefs "sticky" in the sense that expectations revert toward pre-existing beliefs after people look away from a calibrating display? Answering these questions about the fundamental nature of data cognition will clarify the affordances and limitations of visualizations for changing people's minds.

REFERENCES

[1] W.S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. of the ASA, 79(387), 1984.*

[2] T.L. Griffiths and J.B. Tenenbaum. Structure and strength in causal induction. *Cognitive Psychology, 51(4), 2005.*

[3] J. Hullman and A. Gelman. Designing for interactive exploratory data analysis requires theories of graphical inference. *HDSR, 7, 2021.*

[4] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE VIS, 2019.*

[5] J. Hullman, P. Resnick, and E. Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one, 2015.*

[6] N. Irwin and K. Quealy. How not to be misled by the jobs report. *The New York Times, 2014.*

[7] A. Kale, M. Kay, and J. Hullman. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. *CHI, 2019.*

[8] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. *IEEE VIS, 2021.* **InfoVis Best Paper Award.**

[9] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. *IEEE VIS, 2019.*

[10] A. Kale, Y. Wu, and J. Hullman. Causal support: Modeling causal inferences with visualizations. *IEEE VIS, 2022.* **Honorable Mention Award.**

[11] Y.S. Kim, K. Reinecke, and J. Hullman. Explaining the gap: Visualizing one's predictions improves recall and comprehension of data. *CHI, 2017.*

[12] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. Graphics, 5(2):110–141, 1986.*

[13] C.F. Manski. The lure of incredible certitude. *Working Paper 24905, 2018.*

[14] F. Poursabzi-Sangdeh, D.G. Goldstein, J.M. Hofman, J. Wortman Vaughan, and H. Wallach. Manipulating and measuring model interpretability. *CHI, 2021.*