# Interactive Analysis Needs Theories of Inference[*]

Jessica Hullman[†] Andrew Gelman[‡]

6 Nove 2020

**Abstract**

Computer science research has produced increasingly sophisticated software interfaces for interactive and exploratory analysis, optimized for easy pattern finding and data exposure. But assuming that identifying what's in the data is the end goal of analysis misrepresents strong connections between exploratory and confirmatory analysis and contributes to shallow analyses. We discuss how the concept of a model check unites exploratory and confirmatory analysis, and review proposed Bayesian and classical statistical theories of inference for visual analysis in light of this view. Viewing interactive analysis as driven by model checks suggests new directions for software, such as features for specifying one's intuitive reference model, including built-in reference distributions and graphical elicitation of parameters and priors, during exploratory analysis, as well as potential lessons to be learned from attempting to build fully automated, human-like statistical workflows.

# 1 Introduction

## 1.1 Data visualization in hypothesis-driven science

Data analysis is a decidedly human task. As Tukey and Wilk wrote a half century ago [128], "Nothing—not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers—nothing can substitute here for the flexibility of the informed human mind." Data analysis is a microcosm of the scientific method. We conjecture, conduct experiments, find approximations, are surprised, or not surprised, and arrive at perspectives on data that help us make decisions in the world.

The inevitability of human judgment in analysis has led to the development of interactive interfaces to help analysts more easily conduct ad hoc data exploration and analysis. These range from programmatic environments for exploratory analysis like computational notebooks, to modern business intelligence tools that can create dashboards or trellis plots without the user having to manually specify visual encodings, to visualization recommenders that serve up data summaries optimized for perception and exposure of patterns.

If we look to the state of the art in exploratory and visual analysis software, a simple theme seems to unite the motivations behind research threads: the value of the interface is to get out of the way of the data, so the human analyst can find the patterns or "insights" it holds. Innovations in research and commercial systems aim to create a responsive environment where queries are met at the "speed of thought," to enable more flexible inputs by which users can query and analyze data and to efficiently summarize data despite the scalability problems that arise as datasets grow larger.

The presumption behind prioritizing data exposure in building these tools is that exploratory and confirmatory stages of an analysis workflow are easily distinguished. According to many accounts of how knowledge is created during data analysis, so-called exploratory analysis is "model-free" and consists of preparing and familiarizing oneself with data, searching for useful representations or transformations, and noting interesting observations. Confirmatory analysis, on the other hand, involves verifying that data support a hypothesis or generating new hypotheses [71, 101, 105, 124]. However, in practice this distinction proves murky. Model-driven inference plays a role even in canonically exploratory activities; after all, what is surprising is defined

[*] We thank Jeff Heer, Alex Kale, Hanspeter Pfister, Matthew Kay, and Jennie Rogers for helpful comments.

[†] Department of Computer Science & Engineering and Medill School of Journalism, Northwestern University.

[‡] Department of Statistics and Department of Political Science, Columbia University, New York.

by the implicit model of our expectations. With the help of our visual system we engage in processes comparable to fitting implicit models to data when we examine visualizations for distribution and trend, and we judge fit when we notice outliers and other deviations from symmetries inherent in graphical forms like histograms or scatterplots.

If people were perfect statistical processors, we wouldn't worry about implicit model fitting during interactive analysis. But there is evidence that bias can affect outcomes of data analysis. In the social sciences, a high rate of false discoveries is synonymous with the replication crisis, referring to many failed attempts to replicate what were believed to be high quality experiments in psychology and other fields that have relied heavily on null hypothesis significance testing. Writing decades earlier, Diaconis described people's proclivities to look for patterns and their stubbornness to revise their beliefs later as "magical thinking" in data analysis [25]. Gelman and Loken [50] analogize the flexibility that arises when one has full choice over how to filter or otherwise transform data, specify models and tests, and make other analysis decisions to a garden of forking paths that exists implicitly even in analyses in which only one path was chosen. That motivated reasoning can influence the results an analyst reaches in exploratory analysis is the premise of recent work in computer science that pursues algorithms and interfaces for mitigating the symptoms of too much flexibility, such as by tracking and adjusting for comparisons being made.

However, without an underlying theoretical basis to drive their work, computer scientists and statisticians can easily end up designing software that exacerbates magical thinking and conflicts with real world analysis stakes and goals. A good theory of inference can help account for how people do analysis and prescribe how they should. This allows researchers to do their own "model checks" in evaluating software and trying to understand what types of activities and representations analysis interfaces should support.

## 1.2 Scope of this paper

The message of this article is that despite their sophistication, interactive analysis and graphics tools that emphasize exposing patterns in data have a limited ability to improve data analysis without underlying theories of inference. We first consider the origins of interactive data analysis, and how they might have led to a fixation in system design on exposure, the "laying open of the data to display the unanticipated" [128]. We provide examples of negative implications of data exposure as a primary goal in designing systems from interactive visualization and databases research.

We propose that to design effective interfaces for interactive analysis, researchers need theory. A good theoretical framework can guide design, provide normative targets for evaluation, and be used to develop more rigorous understanding of how people do analysis. At a high level, we think of analysts using interactive GUI analysis tools as developing and updating "pseudo-statistical models" that capture their evolving knowledge about the phenomena data are intended to represent. In our view these intuitive statistical models are used to generate predictions, in the form of reference structures, that endow graphics and other data summaries with meaning. As one compares the predictions of their implicit model to data, they realize where their assumptions about a data generating process are off, they update their beliefs in various ways, and they decide how to proceed in their analysis. We argue that rather than assuming that human tendencies in forming beliefs and making decisions are outside the scope of what can be represented computationally, we should be developing and evaluating formalizations of these implicit models that provide testable implications to drive understanding.

We review some recent approaches to formalizing the role of statistical graphics in inference, based in Bayesian or classical statistical philosophies, noting where they overlap versus diverge. We discuss how models like these can help us reason more concretely about how mental (implicit) statistical models are applied in interactive exploratory analysis, how canonically exploratory activities are instrumental to confirmatory analysis, and how the interactive tools we build can enforce these connections. We argue that a Bayesian understanding of exploratory analysis as driven by model checks provides a generalizable framework for developing interactive analysis tools, which aim to naturally integrate exploratory and confirmatory activities.

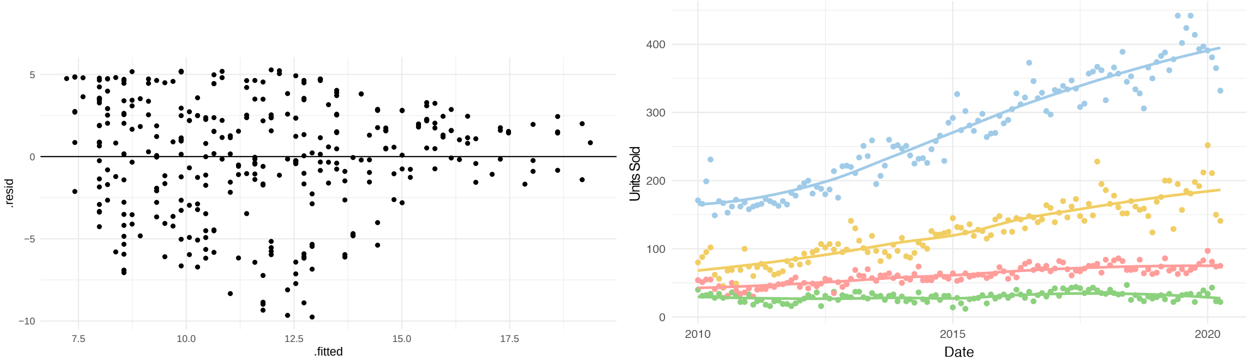

Figure 1: *A residual plot (left) is optimized to show deviations from the reference distribution via the position of the horizontal line at 0 and the degree to which y-positions of points deviate from being randomly distributed about 0. For a line chart showing trends by region (right), the reference distribution is less obvious. It might represent predictions, for example, from an implicit model that assumes all regions have the same growth trend and don't interact with one another.*

# 2 Origins of interactive data analysis

## 2.1 Tukey on exploratory data analysis

It may seem quite obvious that if you are doing data analysis, the interface you use should prioritize representation and easy access to the data. This way of thinking is partially motivated by the exploratory data analysis movement pioneered by John Tukey in the 1960's. Tukey popularized the idea of exploratory data analysis (EDA) as a natural complement to confirmatory data analysis (CDA), writing that "[t]he simple graph has brought more information to the data analyst's mind than any other device. It specializes in providing indications of unexpected phenomena" [125]. Classic goals of EDA of detecting underlying structure or patterns in data are often framed as "model-free." For example, while CDA has been likened to a jury deciding whether a defendant is guilty or not, exploratory analysis is akin to the detective developing hunches [10, 126, 140].

However, many of Tukey's writings also imply that model fitting and graphics go hand in hand. Some of the graphics he espoused can be interpreted in terms of a model; "hanging rootograms," for example, would be difficult to motivate without reference to a Poisson count model. He spoke of the value of the "iterative character of the relationship of exposing and summarizing" in exploratory analysis. By attempting to fit models to data, one learns about what doesn't fit, which "can often be more effectively approached and structured because there has been some fit, even a poor one" [128].

At a high level, if EDA is understood to be discovery of the unexpected, then this is defined relative to the expected. We note two practical implications of this duality:

1. Any exploratory graph should be interpretable as a model check, a comparison to "the expected," which implies that when constructing such graphs we should be able to figure out what is the model being used as a basis of comparison. Sometimes, as with a residual plot (Figure 1 left), this comparison is obvious; other times we can gain insight by carefully considering what sort of model is being implicitly checked by a graph. For example, a graph showing parallel time series for subgroups in a population—say, single family home sales in different regional markets—can be interpreted as a check of, or exploration of discrepancies from, a zero-interactions model in which changes are the same across all regions (Figure 1 right).

2. Exploratory analysis can be made more effective by comparing to more sophisticated models. EDA is often thought of as an alternative to model-based statistical analysis, but once we think of graphs as comparisons to models, it makes sense that the amount we've learned increases with the complexity of the model being compared to. Indeed, this is implied by various principles of effective graphics, to use line plots, small multiples, and other tools to foreground comparisons of interest, many of which

appear throughout Tukey's work.

There is a corresponding argument in classical hypothesis testing or confirmatory data analysis, that more is learned from rejection of a complex model than from rejection of a trivial null model such as a hypothesis that all effects are exactly zero. In some ways, EDA is like an omnibus test in that we are open to all sorts of violations of the model, but with the difference that in exploratory analysis we are interested not so much in rejection as in the particularities of the discrepancies between model and data: rather than tailoring tests to particular alternatives, we rely on human pattern-finding abilities to motivate the development of future hypotheses. We describe this view, which aligns with many of Tukey's statements about graphs, in more detail in Section 4.

While Tukey's advocating for using graphics to identify hypotheses and expose raw data might have seemed to be an argument for throwing out the models completely, from what we know about the context that he was working in, we suspect this level of enthusiasm for graphics may have been his knee-jerk reaction to a general lack of respect for exploratory methods. The statistics research community of the time was dominated by formal proofs of confirmatory methods aimed at "legitimizing variation by confining it by assumption to random sampling" and "restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with 'known' probabilities of error" [128]. Tukey's writings on graphics and exploratory analysis represented his pushing back against the dominance of CDA in the field, his trying to "move the center of gravity away from an (over)emphasis on mathematical theory to a greater balance between methodology, theory, and applications" [42]. Moreover, unifying exploratory and confirmatory analysis is no small task. Consider for example Tukey's statement that "[f]ormal statistics has given almost no guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed" [128], which conveys that he grasped the challenge of incorporating both in a single framework. We shouldn't assume he cared only about exploratory analysis and graphics simply because he didn't have the answer ready at hand.

## 2.2 Innovations in user interfaces

Modern interactive data analysis also owes much to developments in computer science, in the same way that earlier advances in statistical modeling by Laplace, Gauss, and so forth accompanied progress in mathematics. As Tukey began writing about exploratory data analysis, computer scientists such as Engelbart, Kay, Sutherland, and others made pioneering efforts in the development of software interfaces for "intelligence augmentation." As promoted by Engelbart, intelligence augmentation is associated with "increasing the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems" [33]. Increased capability could come as better efficiency (perhaps framed as "more rapid comprehension" or "speedier solutions") as well as improved perception of possible solutions to problems that before seemed unsolvable. The broad framing of IA by these early pioneers outlined a vision for transforming interactions with computers in which graphical user interfaces for data analysis were a natural step.

Tukey's and colleagues' system PRIM-9 augmented the capabilities of a human by enabling perception of higher dimensional data without the usual restriction to two dimensions [40]. A user could "dissect" multivariate data through point cloud rotation, use masking to select subregions of a space, and isolate particular subsamples. Because an analyst will rarely be able to specify the "optimal" projection, finding an appropriate one requires moving about in a multi-dimensional space, which PRIM-9 enabled through controlled continuous rotation [42]. Tukey's work on PRIM-9 led to further developments through projection pursuit, the incorporation of automation into interactive visualization by optimizing a projection index to detect interesting directions of study [43].

In the decade or so that followed, other statisticians made graphics contributions. Asimov [6] introduced the grand tour, which used animation to stitch together projections on high dimensional data for visual analysis in a seemingly continuous manner; the projection pursuit guided tour combined the benefits of both methods for better results when identifying low-dimensional structures in sparse high dimensional data [22]. Becker and Cleveland [8] explored brushing as a way to interactively select data in a visualization one is analyzing, in order to see the same data in other linked views, such as when viewing a scatterplot matrix. XGobi [117], followed by GGobi [118], made these state-of-the-art dynamic statistical graphic methods available in a single environment. Wilkinson's scagnostics [142, 143] explored a graph-theoretic set of measures for grouping bivariate scatterplots of high dimensional data, while his grammar of graphics provided a formal description of statistical graphics [141].

Computer scientists also began to take more interest in the new interactive capabilities for data analysis afforded by more powerful computation. Ben Shneiderman coined the term "direct manipulation" in the early 1980s to refer to systems in which objects of interest such as data points were continuously represented and could be acted on through physical manipulation or button presses [110, 111]. In contrast to the inflexible and hard to learn syntax of conventional query languages, direct manipulation was easy and produced immediately visible and reversible results [63]. One could call direct manipulation interfaces for data analysis an early step towards "democratizing data analysis," as these tools reduced the amount of specialized knowledge required to interact with data; one no longer needed to memorize rigid syntax, for example.

The late 1980's saw the emergence of visualization as a subfield of computer science [87], focused on amplifying cognition through visual methods drawn from computer graphics, vision, signal processing, human computer interaction, and others, and addressing domain applications like medical imaging, planetary sciences, and molecular modeling. Information visualization, which is closer to our focus here, concerns visualizing abstract data for which spatial mappings can be chosen more arbitrarily (e.g., statistical graphics) and was distinguished in the 1990's [17], and drew cognitive scientists and psychologists, statisticians, and cartographers. While many early advances sought to enhance data analysis among experts, the last few decades of research in the field has seen a surge of interest in making visual data analysis accessible to more novice users. Today, widely used systems like Tableau Software employ innovations that grew out of visualization research, by encoding state-of-the-art knowledge on effective visualization [84] and reducing the efforts required to manually specify views through drag-and-drop interfaces like Tableau's shelf model [115, 119] or button-driven chart type transformations [85], which interpret these user interactions as database queries.

More recently, visualization recommender systems have become an active area of research [132]. Recommenders aim to be even more hands-off than popular visualization tools like Tableau or PowerBI by suggesting views to analysts based on perceptual properties [145,146], statistical analyses [24,73,133] and/or contextual or behavioral properties [13,53,73,82], requiring minimal to no input from the user after a dataset has been loaded. Other tools literally make analysis hands-off or at least "mouse off" by supporting new input modalities like natural language [44,108,113] or touch [91,120]. Such forms of "behavior optimization" comprise the state of the art in interactive data analysis system design [103].

At the same time, the rise of "Big Data" as a fascination and challenge faced by industry has also driven increased interest in interactive analytics in database research, referring to approaches for optimizing query results for real time analysis by a human. These applications bring their own challenges [4,38], such as minimizing latency while retaining acceptable accuracy. User interfaces have not always been central to these efforts, but how to deliver visualizations and interactions in these paradigms is gaining interest [3,36,74,92,97].

# 3 Pitfalls of pattern (over)exposure

Modern interactive data analysis tools that employ information visualizations would seem to have imbibed part of Tukey's vision, by embracing graphs and transformations. Research emphasizes the importance of exposing patterns in data in ways that acknowledge the needs of a human user seeking to understand them. However, these needs are often premised on an assumption that the data itself is the analyst's end goal. The focus is directed toward making the system performant by optimizing low-level properties of how data are served up, including how they are visually encoded in graphics, how quickly query results appear, and what gestures or other primitives the analyst can use to interact with them. The governing assumption is that the analyst knows best about what to do with the data, and the system should be optimized to facilitate the searches and comparisons that users want to do.

We argue that easy access to graphs and ability to transform and manipulate data are not the best things to optimize, because data are frequently *not* what is ultimately at stake in data analysis. Rarely does a person sit down to visually analyze some data without the goal of inferring something about the world. As Tukey himself described, phenomena—referring to potentially interesting things that we can describe in non numerical terms—are what we typically want to learn about when we deal with data [127]. To put it more formally, our goal is often inference: inferring some unknown parameters of a statistical model after observing some data we think was generated by the process approximated by the model.

We are not the first to describe modeling as implicit in exploratory and interactive analysis, or exploratory and confirmatory analysis as hard to separate at times; see, e.g., [5, 7, 61, 98, 99, 102, 149]. However, we find the relative lack of theory aimed at reconciling software design with these realities unsettling. Accounts of exploratory visual analysis (EVA) in the literature tend to describe what sorts of interactions analysts do when they use interactive analysis systems, rather than prescribing what they *should* do. Surveying research on interactive visual analysis, Battle and Heer [7] define EVA as a subset of EDA that is often seen as a high level goal of analysis, but the understanding of EVA includes activities seen as precise and determined a priori (e.g., seeking evidence for a hypothesis) and those that remain vague (e.g., looking for something interesting). They describe themes that emerge across different accounts of EVA, like the identification of specific subtasks (e.g., data diagnostics, characterizing distributions and relationships), and the fact that EVA alternates between open-ended tasks (e.g., flipping through filters looking for something interesting) and more focused exploration (e.g., trying to formulate and validate a hypothesis). However, the breadth of these accounts gives the impression that researchers have reached little consensus on how interactive analysis occurs, let alone how it should occur.

Of course, there are subtasks within an inferential analysis, as well as entire use cases, where inference isn't the goal. For example, one might simply want to retrieve some "facts" from a dataset, like, who scored more points in a basketball game. Or one might want to find a new movie to watch [2] or home to buy [144], to cite classic examples, or to search for anomalous data points to identify errors or bugs, with or without a larger inferential goal. But we argue that non-inference tasks should be the exception, rather than the rule in designing for interactive analysis of abstract data.

The fact that exploratory examinations of graphs are often acts of inference does not alone make data exposure bad for inference. However, these inferences often remain informal and unchecked by confirmatory procedures. For example, one recent interview study around professional analysts' "insights" found that only a small handful mentioned that identifying what an insight is depends on one's confidence in it [80]. Other recent studies examining inference using interactive visualization systems found that most people looking at histograms of Census data in an analysis task treated patterns they saw as if they were reliable ("significant") and didn't consider how the number of comparisons they did inflated their chance of finding something interesting [151], and estimated that more than half of a group of analysts' generalizations about a population from a sample were false positives [149]. Naturally, the ways that tasks are framed, the incentives, and the participant pools in these studies make it difficult to say how prevalent false positives are in general. It seems reasonable to assume though that a lack of caution in making inferences, or an analyst's motivations to find something regardless, can lead an analysis astray.

By prioritizing data exposure over inference, researchers focused on interactive analysis and visualization lack theories that can make concrete predictions, to help them evaluate their systems or guard against invalid inferences. Common mantras, to "let the data speak" or provide an "overview first, then zoom and filter, then details on demand" [112]—aren't scoped to inference. We summarize some design implications of pattern exposure, organized by goals, scope, and evaluation, that can threaten the validity of inferences.

## 3.1 Questionable design goals

### Insights as vague targets

When it comes to evaluating the outputs of exploratory visual analysis, a common operationalization of knowledge has been the "insight." This nebulous term has been defined in various ways, with one common definition being a "complex, deep, qualitative, unexpected, and relevant revelation" [95]; cited in [80]. Chen et al. [19] describe an insight as "a fact extracted from data under analysis, such as an outlier, a pattern, or a relationship, a mental model upon which the fact is evaluated, and objective and subjective evaluations of the fact," but while better describing a possible process the definition remains extremely broad.

Insights are often framed as being closer to confirmatory processes; for example, Sacha et al. [105] distinguish between exploratory "findings" and more formalized verification loops that involve "hypothesis" and "insight," similar to canonical models of analytical sensemaking [101]. If insights are outputs of confirmatory phases, one might expect them to be considered as probabilistic statements, but relatively little research that uses insights as evaluative criteria includes degrees of belief as part of the definition. Instead, they tend to be treated more as human constructs to facilitate communication.

Moreover, like EDA itself, if the definition of insights implies unexpected, than the prior knowledge and expectations of the analyst are at play. Visual analytics researchers have proposed knowledge generation models to describe how analysts discover trends and patterns, generate knowledge, and use this knowledge to inform decision making while using visualization systems [5, 19, 70, 71, 105]. However, while some researchers have explored the role of prior knowledge and need for explicit support in visual analytics systems [18, 35, 79, 89], much of this work assumes that analysts' knowledge can at best be captured with written notes that are connected to, but not integrated with data.

How can one design an interactive analysis system to produce more insights if knowledge is at best represented in unprocessed text notes? Given the diversity of forms insights may take, the unclear connection to confirmatory analysis, and the fact that as examples of "findings" they will depend on the analyst's typically implicit and unmodeled prior knowledge, it would seem nearly impossible to design for insights short of simply making sure that the data are encoded in effective ways. A reliance on insights seems congruent with an overall design ethos of simply getting out of the way of the data, since its hard to predict what else would help.

### Unparalleled, yet unchecked, support for visual comparisons

Interactive analysis systems optimized for rapid data exposure and manipulation increase the number of queries an analyst can run on a dataset. When using standard approaches to null hypothesis significance testing (NHST), a "multiple comparisons problem" arises because NHST admits a certain percentage of false positives by definition. Hence the more tests one does, the more false positive conclusions one might expect to arrive at. If visual comparisons are analogous to significance testing, where a $p$-value is used to judge whether an effect can be ruled unlikely to be due to chance, as some statisticians have proposed [14, 15, 140], then we should control their potential to produce false discoveries.

Modern interactive analysis systems enable the user to query more and faster. In an attempt to investigate the possible severity of the multiple comparisons problem that results from flexible visual analytics, Zgraggen et al. [149] had 28 people with analysis background do exploratory visual analysis using data samples they generated from a known ground truth population, asking them to report any reliable observations or recommendations pertaining to the population from which the sample was drawn. They tracked each analyst's total number of visual comparisons, using a combination of question asking and eye tracking, and used statistical tests against the ground truth to determine the accuracy of each type of observation they saw (e.g., a comparison between two groups, a statement about the shape of a distribution, etc.). This led to an estimate that over 60% of the analysts' conclusions were spurious. Other papers similarly report on people feeling confident about conclusions drawn from visual analysis without any follow-up testing [94, 151], though these results undoubtedly depend on the experimental context and participants [7].

The "active construction" of visual comparison results by query refinement–whether adjusting an interactive visualization by zooming, filtering, or rescaling, or simply revising a SQL-style query–can be analogized more generally to the flexibility in analysis choices that gives rise to a garden of forking paths [102]. Systems that serve up easy access to data can be useful for generating hypotheses or identifying patterns that might be interesting, but analysts' a priori goals, hypotheses, or biases in interpreting the value of information may lead them to make conclusions about a statistical phenomena that aren't valid. Pu and Kay [102] relate the flexibility inherent in modern visual analytics systems to the garden of forking paths, providing a more general problem framing that subsumes the multiple comparisons problem but also allows for other "bad fits" from visual analysis, like an overfit model with high predictive error.

## 3.2 Questionable scope

### Defaults in interactive visualization

When analysts use patterns they see in visual exploration of data to make predictions about the world, they need to acknowledge variation and uncertainty. However, if we consider the plotting defaults in interactive visualization and business intelligence tools like Tableau Software or Power BI, which make it easy to make charts without specifying every aspect of their design, often the defaults would seem to discourage this awareness by aggregating data. Such systems are often optimized to support very large datasets which, if plotted as individual points, would be rendered useless by overplotting and latency.

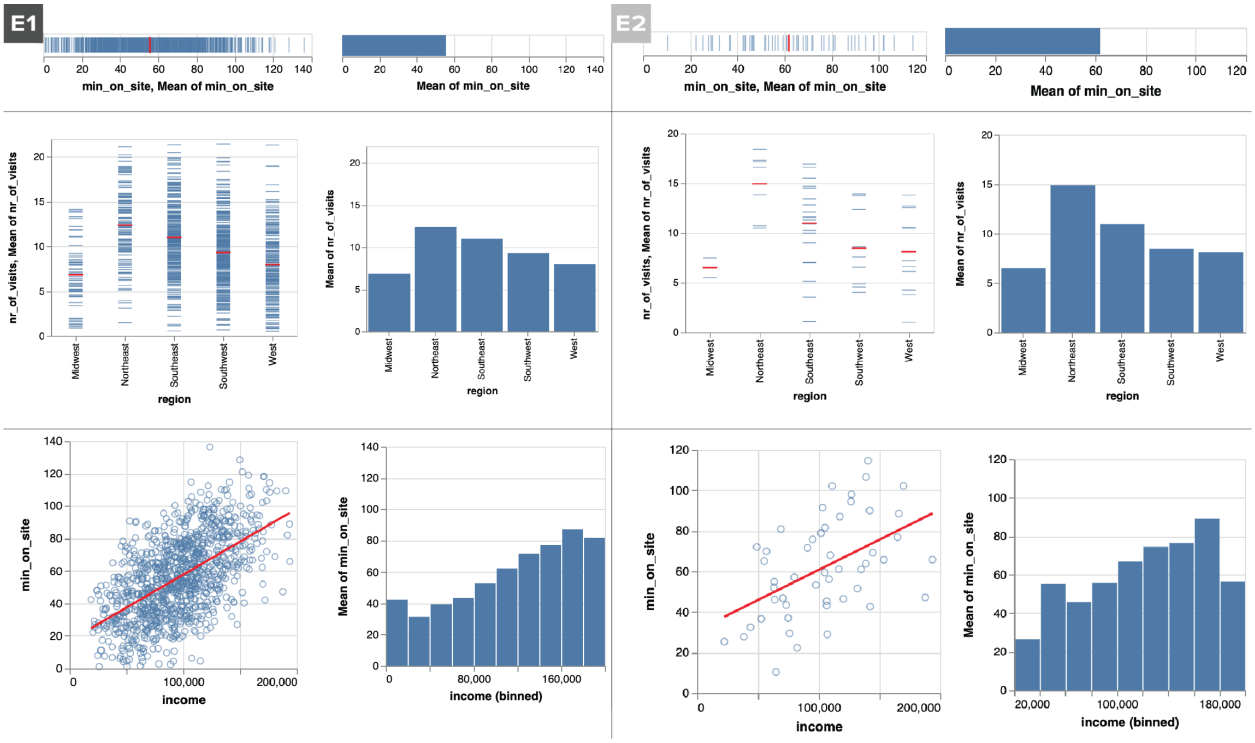

Figure 2: *Datasets with 1000 (left) and 50 (right) records, visualized using two possible default plotting approaches: showing disagreggated data by default but annotating the mean (left side of each panel), or using only mean aggregation (right side of each panel) [94]. Novice analysts might fail to adjust perceptions from aggregated views for the lower sample size on the right.*

However, many datasets analyzed with tools like Tableau do not contain millions or even thousands of rows; consider, for example, how journalists, which are one of Tableau's target user groups [122], frequently report on small datasets like polls (e.g., [107]). System designers may shrug off the heterogeneity in users' needs, assuming they will customize views to support their target inferences regardless of the defaults used by the system, but default settings may have a significant influence on certain users. For example, novice analysts may not know to change from a default or how to do so [109].

In a recent study, one of us investigated how aggregation defaults might impact novice analysts' conclusions about data [94]. In two online experiments,[1] we showed people data samples using either disaggregated views, mean aggregated views, or disaggregated views with an overlaid mark showing the mean and asked what they might conclude, if anything, about a population. People who used disaggregated views were less than one-fifth as likely to talk about effects without mentioning how big they are (e.g., "There's no difference in sales between campaigns," "Visitors from the midwest bought more"). They reported lower confidence values by an average of 6 points on a 100 point scale, and showed more sensitivity in terms of how many conclusions they drew to whether they were looking at 50 records or 1000 records. While we can't know whether in a real analysis setting, those who saw aggregated views wouldn't have tried to disaggregate them first, our experiences teaching students as they use exploratory analysis tools like Tableau or R to do analyses suggests that many won't recognize where aggregation can be harmful.

### Data-centric approximation

While research in databases has moved away from early assumptions that all data are "correct" [56], the effects of sampling and approximation error on query results are often discussed only with respect to the support of the dataset being queried. In other words, datasets are far more likely to be treated as though

[1] https://osf.io/v87wd

they are themselves objects of inherent interest than modeled as representing samples from some unobserved data generating process.

For example, consider the challenge of creating a database system that can provide an analyst with a quick enough response during exploratory analysis to make interactivity worthwhile. The need to run queries at responsive speeds that aggregate information across large databases in industry applications of business intelligence or data analytics has motivated database research in techniques like precomputation and approximate query processing (AQP), both of which aim to quickly return results when it's not possible to arrive at the non-approximate result in real time [56]. With AQP, query results are approximated, typically by taking uniform random samples from a dataset that minimize error while satisfying latency constraints, or meet some user-specified error bounds. The approximate result is returned to the user with estimated error, such as in a bar chart with error bars. Error is data-centric in the sense that it is defined relative to the dataset at hand, without considering that that data may be a non-representative sample of the underlying population (for examples of how even large large datasets can lead to poor error estimates in AQP, see [148]).

As another example, consider differential privacy (DP), which defines a database mechanism that limits the probability that a person in the dataset can be re-identified [30]. Though motivated by the problem of defining an individual privacy-preserving database that contains a representative sample for inference about a population (e.g., [31] as cited in [26]), in its most popular form, DP drops the notion of inference to focus on cases where the dataset itself is of intrinsic interest [26]. Work that assumes data will be used for inference, and consequently accounts for measurement error in formulating differentially private data releases, is relatively hard to find in research or practice despite the overwhelming interest in DP (for some exceptions see, e.g., [12, 34, 37, 69]). While some computer scientists have worked actively at the intersection of computer science and statistics (e.g., [32, 139]), Wasserman [138] comments on the seemingly large divide between the computer science view on privacy, which seeks to apply privacy-preserving mechanisms only after one has obtained some statistics, and the statistical view, which tends to prefer a sanitized dataset in light of the difficulty of predicting what future analyses they might want to use data for.

Beyond assuming a dataset itself is of intrinsic interest, how to present results from algorithms like AQP and DP has generally been brushed aside, though there are many reasons to think that people might not grasp the implications of these techniques. For DP to achieve its privacy guarantees requires a good choice of $\epsilon$, the parameter that controls the amount of noise added to the query result and consequently the risk of de-identification. But little guidance is available on how to set this parameter [29, 59], nor have many researchers pursued interfaces or other abstractions to help people reason about the trade-offs.

On the other hand, researchers in progressive computation have been said to "first develop the system and then add the user interface as an afterthought" [4], leading to clunky interactions at best . For example, many user interfaces for AQP and progressive computation like sampling-based, incremental or progressive visualization [97, 104], or "optimistic visualization" [92] visualize approximate results quickly under the assumption that an analyst will know to be cautious. Moritz et al.'s [92] Pangloss system continues to compute precise results in the background so the analyst can later check their work. However, a small study on professionals found that they did not necessarily revise their initial beliefs based on the approximation; in fact some users reported never doing this. Others similarly report that not all users of progressive computation techniques understand approximation error [39]. If, as studies of how people make decisions from uncertainty visualizations have found, people generally ignore the error bounds [62, 66] or overestimate effect sizes because they don't understand them [58], then approaches like progressive computation and AQP might exacerbate bias. One approach has been to instead construct algorithms that don't produce visualizations that deviate beyond some allowable perceptual error rate [3, 74], but this trade-off could be better approached if we first understood how susceptible people are to revising their initial beliefs in such contexts. Like others [78], we think that without a better understanding of human inference reached through more rigorous evaluation, it becomes difficult to reason about how much such systems help analysts do inference over less responsive but more precise results.

## 3.3 Questionable evaluation

### Accuracy and satisfaction as evidence of good visualization

The interactive visualization and human computer interaction research communities have devoted many user studies to checking how quickly a system or new techniques enables users to respond to questions with

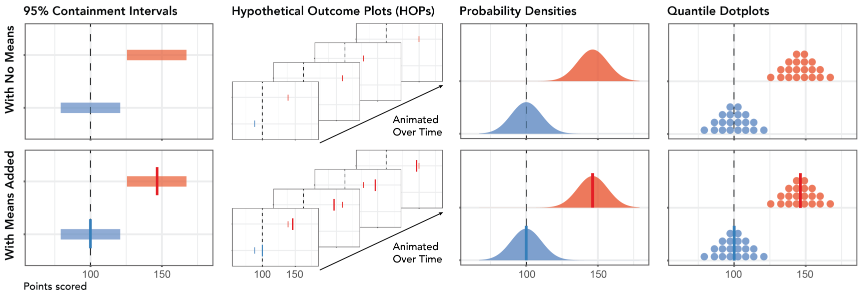

Figure 3: *Four ways of visualizing uncertainty—95% containment intervals, animated hypothetical outcomes, probability density plots, and quantile dotplots—with and without means. A recent study found that the visualizations that best supported accurate reading of probabilities (quantile dotplots) were not necessarily the visualizations that led to the least biased decisions (densities or intervals without means) [66]).*

verifiable answers and how accurately users respond to questions that largely test reading the data [64, 104]. Self-reported satisfaction and qualitative user feedback are also common [104]. There is even a long running workshop entitled "Beyond Time and Error" [1].

Perhaps the last place we would expect to see heavy reliance on time to respond and accuracy in reading data is when it comes to evaluations of static and interactive representations of uncertainty. After all, visualizing uncertainty aims to help users acknowledge limits to inference. But in a recent survey of 90 evaluative studies in this area, most focused on measures of users' accuracy reading data (e.g., how well could people estimate means or probabilities from visualizations) or self-reported satisfaction (e.g., how high did they rate a visualization when asked via Likert scale how clear or useful it was) [60].

There are clear problems with the assumptions behind a reliance primarily on accuracy and satisfaction. For one, reading data accurately when asked doesn't necessarily translate to using that information in a decision, though better decision making is often the reason authors state for pursuing better uncertainty visualizations. People also may not know what's good for them, in the sense of leading to more sound judgments and decisions, when it comes to representing uncertainty; see, e.g., the vast literature on heuristics and biases that arise from a seeming desire to ignore uncertainty [129].

As an example, a recent study showed people visualizations of two distributions representing predicted points scored in a fantasy sports game, and asked them to judge the probability of getting at least a certain number of points if they add a new player, and to make a decision about whether to invest in the new player (Figure 3). Participants made slightly more accurate estimates of the probability of winning a prize with the new player with quantile dotplots, but made less biased decisions of whether to invest in the player with other visualizations (e.g., intervals without means when variance was high, with which they made noticeably worse probability estimates [66]). It's also reasonable to expect that in many situations, when people are asked which visualizations were clearest or most enjoyable to use, they might point to the more familiar ones (e.g., intervals or density plots). Consider for example, the backlash against the New York Times animated election needle visualization in 2016 [88], a visualization that showed uncertainty effectively by making it impossible to ignore, but which some people found highly disturbing when it was introduced without warning on election night after they'd gotten to use to very different static graphics.

### Distributional assumptions about "good" analysis

To evaluate interactive analysis tools or techniques and identify cognitive bias in analysis, as recent work attempts [136, 149], requires defining what good versus bad analysis looks like. Here researchers have also tended to resort to using lower time spent on a task, or lower time required to answer a question, as desirable criteria, along with reported satisfaction through qualitative user feedback, and number of insights (see Rahman et al. [103] for a review).

Another common criteria is coverage, how much of a dataset is explored in a session (e.g., [130, 136, 145, 146]). In some cases, coverage can make sense, such as when someone has limited amount of time to analyze multiple variables in a large dataset and query latency may prevent them from getting to all of them [83]. However, in many other cases, its not the right definition of correct. Consider a case where an analyst sits down to analyze a dataset on donors to a nonprofit, which includes multiple years of donation amounts, donor gifts received, website interactions, and donor demographic information. If an analyst truly has no background at all on what to expect but must deliver a comprehensive report on all donors and their behaviors, then maybe it's fair to hope for inspection of each variable. However, such cases seem more like exceptions than the status quo in interactive analysis. For example, an analyst may have some high level question to answer, such as, Which gift should we send to the top donors to our nonprofit this year?, that leads to seemingly poor coverage.

Even if an analyst does sit down to simply learn from the data, imagine that the nex step after constructing some exploratory graphics is to concentrate on the relationship between where the donors live and how much they give. For example, the analyst might generate multiple plots looking at this relationship in different ways as well as how it interacts with the marketing campaigns in different regions. While we could call this bias because it looks like fixation on a subset of data, it's quite rational from the analyst's perspective to focus on relationships that have the potential to lead to actions (e.g., changes to marketing campaigns), over say, assessing demographic information that can't be targeted for privacy reasons.

In the absence of a guiding theory, metrics intended to capture the “correctness” of an analysis, or the fitness of an interactive analysis tool, can conflict. For example, while some studies use coverage as a measure of good analysis, a tool designed to better support the depth-first exploration identified in empirical accounts of EDA [7, 145] should lead to worse coverage. Competing measures of correctness are not surprising given that no single theoretical basis for EDA has been adopted by researchers.

# 4 Theories of inference for interactive analysis

If examining visualizations is often about making inferences about the world, how should we conceptualize this process? And what formalizations can we use to help us design better interfaces?

Though some have proposed models of knowledge generation in visual analysis [5, 71, 101, 105, 137], these have generally been conceptual models with limited ability to make concrete predictions of how to improve systems or to provide evaluation criteria for inferences. Below, we describe a high level view of interactive analysis involving intuitive fitting of models to compare to data, and then review some proposed models.

## 4.1 Implicit model checking in interactions with data

We like to think of analysts using interactive analysis software to work with data developing and updating “pseudo-statistical” models that help them make inferences about the real-world phenomena data are intended to represent, like several others [5, 102]. In contrast to statistical models an analyst might explicitly specify during a data analysis session, we call these models pseudo-statistical because while they may be approximated statistically, they may deviate from what is generally defined as rational inference.

In our view, these intuitive statistical models provide the reference structures that endow graphics and other data summaries, as well as users' interactions with them, with meaning. Predictions of these models provide the basis for implicit or explicit comparisons to data through graphics, whether the graphics are intended to evoke a strong comparison to a reference model or not. For example, imagine you are analyzing economic data and generate line charts as in Figure 4. Even if you don't have much prior knowledge on different demographics' fiscal trends, how do you judge whether you see anything interesting, and how interesting it is? Can you answer these questions without making some assumptions about what you expect to see under different assumptions about the trends?

At a high level, there are a few reasons we find it natural to think about any examination of a data summary as model-driven inference. One is that many exploratory graphics themselves get their meaning from implicit reference distributions. For example, a histogram often invites a comparison to a familiar reference shapes like symmetric bell curves. Depending on what sort of data one is working with, a priori expectations may help define the reference: one expects power law-like relationships for degree distributions in many networks, and in various other online interaction data, while in credit fraud detection, plots of leading digit frequencies might be compared to the expectations of Benford's law. As an example where the reference distribution defines the spatial mappings within the chart, consider a residual plot from a linear

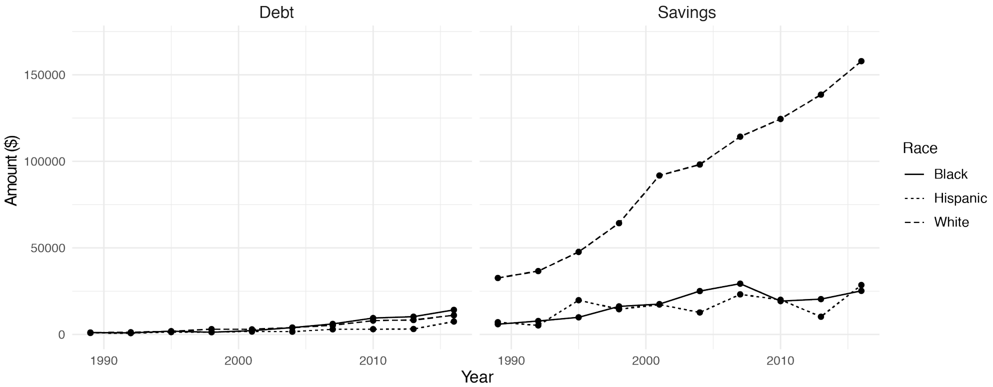

Figure 4: *Line charts comparing average student loan debt among those aged 25–55 (left) and average family liquid retirement savings by demographic (right). Depending on how much sampling error a user assumes, they might conclude that average loan debt is approximately equal by race, while noting a difference between the top line and bottom two lines on the right. Depending on their prior knowledge, both domain general and specific, as well as the goals of their analysis, they might imagine different reference distributions to judge the strength of the "signal" in the chart on the right. Data from [90].*

model (Figure 1 left), with residuals (observed minus predicted) plotted to $y$-position, a horizontal line annotated at zero, and the predictor plotted to $x$-position. The plot structure makes it easy to identify when the observed residuals depart from the implied reference distribution: when the $y$-axis range is automatically scaled to the data, we can use the degree to which the zero line is off-center to judge biases, as well as how much they depart from being equally distributed around the center line. It's difficult to imagine a chart that doesn't imply some sort of reference through the spatial frame it puts the data in. The human visual system's ability to detect deviations from symmetry makes graphics an excellent means for spotting discrepancies from a model, regardless of where one is in an analysis process.

Another reason we like the intuitive model comparison perspective view is that if we define these implicit models as Bayesian, as two of the frameworks described below do, we can make more concrete statements about the role of prior knowledge in exploratory analysis. That prior knowledge about a domain or a specific process plays a role in what someone concludes from using graphs has long been acknowledged in high level models of graph comprehension (e.g., [41, 96, 100]), but these are too conceptual to make many concrete predictions. In most interactive analysis systems today, prior knowledge is restricted to being implicit, given that few features for eliciting or representing prior knowledge exist (outside of standard options like the ability to save text notes). However, differences in prior knowledge, and consequently differences in reference distributions, are likely to explain most differences in what analysts conclude and how their interactive analysis sessions differ. For example an experienced economist who who generates the pair of line charts in Figure 4 who is interested in individual level effects might realize that they should switch to visualizing medians. Upon realizing this, they might shift the reference distributions they imagine for lines closer together, with more shift down for the noticeably distinct White line in the right chart down, which they expect to be affected to a greater extent by income inequality in that group.

Like statistical models in practice, we expect that implicit mental models can vary in their sophistication and can produce fits that range from good to poor, depending on how experienced the analyst is, how motivated they are, and other aspects of the context. For example, imagine someone examining Figure 4 to judge how likely it seems that a model fit to the Black and Hispanic lines could have produced the line for White, based on the apparent grouping. While an experienced analyst might bring in some prior knowledge, as described above, and then imagine simulating draws from reference distributions for each line, an inexperienced analysts might instead simply imagine an outline around the bottom two lines created by treating the maximum and minimum savings amounts observed over both series as the upper and lower bounds on a confidence envelope, then judging how likely to the other line is to fit in this shape. Recent empirical work on how people make effect size judgments and decisions from plots of distributions [66] describes how different users may vary considerably in the sophistication for their visual routines, from using simple visual ratios, like the ratio of a difference in means relative to the total axis length in a bar chart, to more sophisticated sampling-oriented strategies. However, regardless of the user and their level of experience, we expect that analysis benefits the more explicitly these models are considered. Less sophisticated analysts, if prompted to articulate reference distributions, might observe the issues with using shallow heuristics based on visual cues only, while more experienced analysts might be better able to improve upon their implicit models by thinking about them more explicitly. In both cases, learning about the deviation between an intuitive model and the data drives further analysis.

## 4.2 Some proposed conceptual frameworks

### Graphical inference as model check

Our proposal above is informed by a formulation of graphical inference and exploratory analysis as analogous to a "model check": a comparison of data to replicated data under a model, that one of us previously proposed [45, 46]. Like previous generalizations in the history of Bayesian statistics (e.g., generalizing from the likelihood to include a prior, such that classical models can be seen as Bayesian models with uniform priors, or hierarchical modeling for which simpler models are a special case), the generalization of the Bayesian paradigm to model checking aims to explain two seemingly opposed phenomena—exploratory versus confirmatory statistics—via a single model.

Assuming a parameter(s) of interest $\theta$, the model check formulation expands the notion of a posterior distribution in Bayesian inference from $p(y|\theta)p(\theta)$ to $p(y|\theta)p(\theta)p(y^{rep}|\theta)$ [51]. In its original formulation, $y^{rep}$ is a replicated dataset with the same size and shape of the observed dataset $y$, but produced by a hypothesized model that accounts for what is known about $\theta$. All model checks, whether exploratory (driven by graphical comparisons) or confirmatory (driven by $p$-values), represent comparisons between $y$ and $y^{rep}$. Visualizations of these, whether real or imagined, can be thought of as visual test statistics ($T(y)$ and $T(y^{rep})$); in other words they play the role of summaries that capture the amount of signal in the data (observed or imagined). The visual test statistics might even be subjected to some standard via a discrepancy function, producing something like an implicit $p$-value. To put this more simply, the model check formulation essentially says that in viewing graphics, the user imagines data produced by a process that seems reasonable to them, and compares these imagined data to the observed data plotted in the graph.

In a Bayesian framework, the hypothesized model that produces $y^{rep}$ is the posterior predictive distribution $p(y^{rep}|y)$. To understand this distribution, first consider that a Bayesian statistical model assumes that one has prior information, in the form of a distribution over possible values $p(\theta)$; one then observes new data that convey the likelihood of different values of $\theta$, also captured by a distribution $p(y|\theta)$; then Bayes' rule updates prior beliefs according to the likelihood, arriving at posterior beliefs $p(\theta|y) \propto p(y|\theta)p(\theta)$. The posterior predictive distribution can be viewed as a transformation of the posterior distribution $p(\theta|y)$ from parameter space (i.e., in terms of $\theta$) to data space (i.e., in terms of the underlying measurements). It's calculated by marginalizing the distribution of $y_{rep}$, which we can think of as a newly drawn version of $y$ given $\theta$, over the posterior distribution of $\theta$ given $y$. As an example, say $\theta$ is the rate at which new customers of a telecom company who have multiple phone lines are expected to "churn", or drop service, in a six month period. Our posterior distribution describes our beliefs about the rate (perhaps using a beta distribution), but the posterior predictive distribution tells us how many customers to expect to stop service in some finite set (which is binomially distributed assuming independent churn with a common probability), given what we've learned about $\theta$ from the new data we observed and accounting for our remaining uncertainty.

By formalizing the process that produces $y_{rep}$ as a posterior predictive distribution, the Bayesian model check analogy highlights what factors can influence how analysts judge patterns in data. Beyond describing the implicit model as simply "a process that seems reasonable to the user," as we do above, we can expect the form it takes (comprised of a model specification and fitted model) to be subject to influences of the observed data, and by extension, the type of chart used via which the observed data is perceived, as well as the prior knowledge of the analyst. To make this concrete, imagine a user of an interactive visualization system plotted all quantitative variables of interest to histograms, and then inspects these to judge distribution. An analyst who a priori expects normally distributed data based on what they know about the underlying measurements might compare what they see in the histogram to imagined histograms of normally distributed datasets of the same size from a process with the same mean and variance. Perhaps the location or scale of the replicate distribution would be adjusted based on their prior, capturing evidence they've seen in the past.

Sometimes the shape of the data itself may influence what the analyst assumes as the implicit reference. For example, if the data were very skewed and with very weak prior information about what to expect, the comparison might be to draws from some exponential distribution parameterized according to the data. How the data are plotted also matters. If instead of a histogram, the data were plotted as points to a one=dimensional scatterplot (i.e., strip plot), the analyst with a priori expectations could still use their knowledge of the variable to look for the visual density signature of a bell curve, made easier if the points were semi-transparent. However, the reduced expressiveness of the plot for showing density would make this harder than other comparisons, like judging deviation from a uniform distribution. The more we can optimize the visualization for the reference distribution, as in the residual plot, the more efficient in terms of knowledge gain we expect the analysis to be.

Most visualizations enable estimating more than one parameter, so $\theta$ is often a vector. For example, for a choropleth map of predicted vote share in an election, $\theta$ is a vector of state-specific predictions. There may also be flexibility in what aspects of the visualization the user's simulated reference distribution is intended to compare to. E.g., someone might focus their attention on just a subset of plotted data where they have some testable assumptions. As an example, imagine someone analyzing housing data who creates the line chart in Figure 1 right. They might judge the evidence it provides for discrepancies in growth trends in single family home sales by region by comparing where each region's line falls in an imagined prediction distribution from a linear model fit to all the regions' lines, assuming they don't interact. However, if they're especially interested in a more specific comparison, such as just the past couple years or so, their models and model checks might focus on just the right portion of the chart. Or as described above, someone analyzing Figure 4 right might imagine a reference distribution capturing a model fit to the observations summarized by the two lower lines in order to compare to the upper line.

The model check formulation is productive as a theory for inspiring better software because it implies that software should encourage users to consider their reference distributions regardless of the analysis stage. We speculate that reference-less charts are more likely to lead to the sort of spurious insights various empirical studies have pointed to. For example, a novice analyst who makes an U.S. map of cancer rates by county presented without any adjustment based on reference distribution (e.g., [23]) might think they see a pattern that is due to sampling error based on different county populations. We have also used the model-check formulation to reason more formally about how design strategies like visualizing uncertainty relate to goals that authors have in communication settings, like wanting their viewers to recognize a certain pattern in a chart [61]. If we think about visualizing uncertainty as reducing some arbitrariness in the viewer's implicit choice of model and fitting process, then it seems more rational to convey uncertainty despite fears viewers won't appreciate it. Hence a good model of intuitive inference can also lead to ideas that help authors anticipate and design for interpretations in communication settings.

Making reference distributions explicit in interactive analysis software could be approached in a few ways. For many types of graphics, it may be possible to build some set of reasonable reference models into software and allow the analyst to customize their graphics to make the comparison more obvious. Standard graphics for univariate and bivariate distributions like histograms could be viewed with optional reference distributions, representing different likelihoods fit to the observed data. This approach would assume a uniform prior, but if one has a prior on the location or scale of parameters these could be elicited and then draws from the corresponding posterior predictive distribution shown. If statisticians and computer scientists worked together, we imagine it could lead to a new class of interfaces that aim for the same "naturalness" currently pursued in interactive analysis but toward the goal of helping people externalize their expectations by building models. We suspect that if users become more accustomed to conceiving of graphs as model checks, this may encourage them to be more careful to design data summaries that take advantage of symmetries in reference distributions, similar to the way a residual plot does.

Given a framework like the model check, there are many questions that one could try to answer empirically toward better understanding human graphical statistical inference. For instance, is it generally true that people assume implicit distributions when they judge "signal" in graphics, or do simple heuristics based only on the data shown operate more often? In this case, the research question may be how software can shift people toward probabilistic thinking. Moreover, the Bayesian formulation assumes a posterior predictive distribution $p(y^{rep}|y)$, but as our example in Figure 4 showed, in some cases $y$ might be a subset of the data shown in the visualization, e.g. when the judgment is about how different one group or series is from others. Can we predict how a viewer will divide up plotted data to judge the significance of some pattern? We think more can be done to try to answer some of these questions, though we also expect that certain aspects of a given visual judgment process will be hard to empirically validate without asking a user for some input. For example, given someone is using a probabilistic model like a posterior predictive distribution, how well can we know how much of that distribution is influenced by their prior $p(\theta)$ versus an imagined likelihood function $p(y|\theta)$ without trying to elicit their prior? To actually faithfully fit such experimental data could require more modeling structure of correlations than would be reasonable to assume in viewers' conscious reasoning processes.

### Graphical inference as Bayesian cognition

The utility of Bayesian models of cognition for understanding a range of human behaviors also informs our conception of intuitive modeling above. In cognitive science, Bayesian models of cognition [54, 55] have gained traction for modeling various forms of human cognition, including object perception [72], causal reasoning [114], and knowledge generalization [123]. These models assume individual cognition relies on Bayesian inference: an individual's implicit beliefs about the world are captured by a "prior"; when exposed to new information they update their prior according to Bayes' rule, arriving at posterior beliefs. One of us has applied Bayesian models of cognition to how people draw inferences when shown visualized data, either eliciting their prior beliefs about a parameter (e.g., [75, 76]) or endowing priors, showing them new data, and then eliciting their posterior beliefs to compare to normative Bayesian posterior beliefs from one or more models reflecting different ways that Bayesian updating could occur.

As an example of what a Bayesian cognition model emphasizes, assume someone views a plot like Figure 1 right during exploratory analysis, and implicitly want to compare the trends they see. We assume that they may bring some prior beliefs, however weak, to the task. If they care about slopes, they might have a prior $p(\theta)$ where $\theta$ represents some vector of priors $\beta_{1,i}$ for each line's slope, specifying, for example, that the slope is slightly more likely to be positive for each group and falls somewhere between a negligible amount (e.g., the rate predicted by population growth) and a steady increase of up to say 20% per year. Or maybe the prior comes directly from the estimates of another smaller data sample on the same topic (housing markets in this case). We assume that the user views each new series in terms of the likelihood it implies of different slope values and implicitly arrives at posterior beliefs about the slope of each line by updating their prior beliefs to reflect the newly observed series using Bayes rule. Comparing their posterior beliefs to those of an agent who perfectly updated according to Bayes' rule provides some insight into how "rationally" the user learned from the new data. While they may make additional comparisons atop these posterior beliefs about each slope, such as comparing them to judge how different they are, but the Bayesian cognition perspective focuses more on the updating.

An important distinction concerns whether Bayesian cognition is used in a normative sense, where a Bayesian model is used to define "good belief updating" as a standard for comparing to people's belief updates from data, versus a descriptive sense, in which observations of people's belief updates are analyzed to gain insight into human inference, ideally using principled tools for model evaluation and model selection [121]. In other words, applying Bayesian cognitive modeling does not in itself imply that people must be perfect Bayesians (in fact much of the literature in mathematical psychology, as well as applications in computer science [68, 75, 76, 147], explores how to explain deviations from normative Bayesian updating).

How does the Bayesian cognition model relate to the Bayesian model check formulation described above? Both rely on the generalizability of a Bayesian modeling framework for describing human inference. Both can be used descriptively or normatively. Taking the latter perspective, both would seem to imply that people make of use Bayes' rule to update their beliefs and prescribe how to improve this process. The model check formulation suggests that we want to figure out our implicit models so we can better use graphics to check their fit and refine them, while Bayesian cognition implies that we want to find graphics that lead to closer-to-Bayesian belief updating. In many ways, their normative versions are complementary: Bayesian cognition emphasizes trying to achieve more rational updating in the context of a predefined model, while the model check formulation focuses more on how a Bayesian perspective enables us to think about implicit models that people use to take into account observed data as well as prior information. Improving people's behavior with graphs via Bayesian cognition should lead to more sound implicit models and reference distributions from the standpoint of information accumulation. We could in this sense consider the Bayesian cognition framework a submodel of the Bayesian model check framework, one which tends to focus on the implicit belief updating step that uses the observed data to update the prior, producing the posterior distribution and corresponding posterior predictive distribution.

In some applications, descriptive and normative perspectives on Bayesian cognition go hand in hand. Normative Bayesian updating can be used as an evaluative standard to better understand how people perceive the informativeness of data than is possible with standard evaluation approaches for visualizations like evaluating perceptual accuracy. For example, when people's priors are elicited about parameters they have some real world experience with (e.g., common disease rates, rates describing properties of groups of people), when shown larger datasets (e.g., in the thousands) many deviate considerably from doing a Bayesian update [75, 76]. Separating the deviation from Bayesian posterior beliefs into location deviation (e.g., how far is the mean of their posterior from the mean of the Bayesian's posterior) versus variance deviation (e.g., how much more or less certain are they than a Bayesian would be) indicates that variance updates are the culprit, aligning with biases described in behavioral economics and judgment and decision making like non-belief in the law of large numbers [11]. Going a step further, we can get more insight into how off people's perceptions of variance are by assuming they have no reason to distrust either source of information, and calculating an approximation of their "perceived sample size," e.g., the size of the equivalent random sample that a Bayesian would have needed to see to arrive at that posterior [76]. This approach provides a more easily interpretable metric for evaluating visualizations than other measures of the similarity of distributions, like Kullback-Leibler divergence. For example, for simple examples where people make inferences about proportion parameters from survey data, we've seen extreme insensitivity to large samples [76], perhaps because icon-based representations of very large survey results require representing more than one observation per icon. However, we've found we can up the average perceived sample size for a huge dataset ($N = 750K$) from about 400 to about $70K$ by animating the icon arrays [76].

Developing pseudo-Bayesian updating to predict people's belief updates in different settings also motivates engineering better ways to show data, those that are likely to lead to more Bayesian intuitive updates. For example, in a progressive computation or AQP setting (Section 3.2), conservatism in belief updating deriving from a bias like non-belief in the law of large numbers would suggest showing more frequent, smaller N updates on a processing query, rather than a single approximate result followed later only by the final precise result [11]. Toward more descriptive ends, one might incorporate sources of deviation from normative updating based on factors other than the statistical informativeness of the data. For example, hierarchical models in which hyperpriors describe the bias a person expects from a given information source can be used to reflect on the forms and strength of distrust in data as a reason for deviation in some settings.

In other descriptive applications, applying Bayesian models of cognition has also led to observations that people's posterior beliefs look closer to normative Bayesian posterior beliefs in aggregate [76]. This result is related to various models in economics yielding as-if-rational behavior from aggregates of individually irrational actors. One proposed reason why belief updates can look Bayesian in aggregate but not individually is that while people have a prior probability distribution which encodes their beliefs, they do not form judgments using the entire distribution at once [134]. Instead, they are "noisy" Bayesians who take a small number of samples from the distribution (which they may or may not have access to [93]), and reason with these samples instead of the full distribution. If people's natural processes for reasoning under uncertainty involve sampling, then formulations like the Bayesian posterior predictive check above are compatible, and in developing interactive interfaces for analysis, we might make use of sample-based representations. For example, we can elicit their prior or posterior beliefs by asking them to "sketch" sets of representative samples rather than full distributions (see, e.g., [76]). Correspondingly, uncertainty affecting data could be shown using samples (e.g., [62]), so that a single metaphor is used for uncertainty throughout an interactive analysis or visualization tool.

Of course there are also challenges and limitations with this approach. One is eliciting priors. Undoubtedly how this is done matters, but it can be hard to evaluate if one's gotten the right prior from someone. Another challenge is that assuming people always will or should do Bayesian updating may be unrealistic, similar to assuming that people always think in terms of reference distributions may be unrealistic. This is not because these theories are saying that people need to be consciously aware of model checking or Bayesian inference to use it; that would be misunderstanding large swaths of research in cognition that show that people are often capable of sophisticated inference without necessarily understanding the mechanisms they are using. Rather, both Bayesian cognition and Bayesian model checking, in their canonical forms, do not account for many factors that may lead to shallow processing that diverges from their predictions, like lack

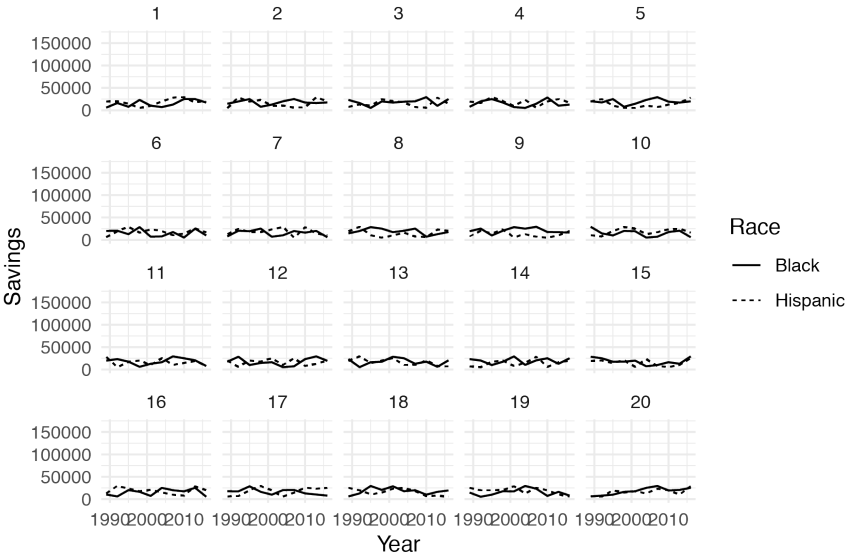

Figure 5: *Lineup that hides the real data for Black and Hispanic from Figure 4 amongs 19 "null plots" representing expected lines given no growth trend in savings (in this case generated by randomly permuting the Savings values for each group). If users can identify the actual data from the lineup, they are said to have approximated a hypothesis test with type 1 error rate of* $1/N$.

of attention or time spent on a graphical inference task. However, despite these limitations, as a means of learning about how people draw inferences and stimulating ideas about how to make software help them do better (i.e., diverge less from normative behavior), Bayesian cognition, like the model check formulation, enables precise predictions and provides new ways of measuring behavior with graphics for checking against human behavior.

### Graphical inference as implicit hypothesis testing

Some statisticians have proposed formalizing graphical statistical inference as null hypothesis significance testing [14, 140]; Buja et al. [14], for instance, argue that discovering some insight using a visualization is akin to rejecting some null hypothesis. This understanding has led to several types of graphical tools. The Rorschach method involves producing an array of "null" plots, visualizing data drawn from a null distribution that represents samples from a data generating process where no pattern exists. The idea is that looking at such plots can calibrate the eyes for sampling variation as one examines data. The "lineups" approach also relies on null plots generated in the same way, but produces an array of $N$ plots, where one of the plots is of the some observed dataset $y$ and the other $N - 1$ plots are null plots. If an analyst can identify which of the $N$ plots shows the observed data, they are said to have performed a visual test equivalent to a hypothesis test with type 1 error rate of $1/N$. Both approaches requires the analyst to specify the null generating mechanism.

Applied to the line chart in Figure 4, the goal of the implicit hypothesis testing view might be to formulate the task of judging whether there's something interesting as a comparison of the observed data to data generated by a null hypothesis like "all slopes are the same." Users could then see if they could pick out the real data, which, given the discrepancy between the slope of the White line and the Hispanic and Black lines, would be easy. However, if they had a more specific question, such as whether the data support the interpretation that the growth rates for Hispanic and Black are identical, they could create a lineup that uses a similar null distribution but is defined for only those two lines (Figure 5), for example by simple shuffling each series' $y$-values while keeping $x$ consistent.

Lineups are an innovative attempt to combine graphics and confirmatory data analysis, but they may fail to achieve equivalence with actual statistical tests as a result of unmodeled factors that affect how well someone can detect the graphical deviations that identify the observed data. For example, the user's visual acuity and how the data are encoded are likely play a role in how well calibrated the presumed null hypothesis testing is. Figure 4 uses the same $y$-scale as the original graphic, but if it were scaled instead to the range required just for the Hispanic and Black lines, the task would feel different. In fact, studying how people look at lineups to better understand graphical statistical inference has become its own line of research [9, 21, 86, 131, 150], providing some evidence of our view above that a good attempt at formalizing graphical inference can lead to better understanding of human visual inference and where it deviates from expectations.

More generally, the Bayesian perspective on graphical comparisons as model checks subsumes treating visual comparisons as null hypothesis significance tests as a special case. While the lineup and Rorschach require the analyst to think about their reference distribution by asking them to specify a null generating mechanism, we prefer the model check framework's greater emphasis on understanding deviations from a model over knowing for sure whether a difference is zero. Taken literally, this latter assumption seems to conflict with the idea of analysis as estimating unknown quantities: when would we really expect two different quantities in the world to be exactly the same? The Bayesian formulation also provides room for prior knowledge to be involved in intuitive model checks.

#### Multiple comparisons problem and correction

If examining graphs is akin to doing hypothesis tests in a classical statistical paradigm, then as we describe in Section 3.1, the more graphical comparisons one makes, the greater the risk of finding a pattern that does not replicate with new data.

In their study of the multiple comparisons problem in exploratory analysis, Zgraggen et al. [149] identified each analyst's explicit hypotheses (those stated by the participant) and implicit hypotheses (those not reported, but identified later in interviews or using eye tracking) to estimate how of many conclusions they drew were false positives. Since eye-tracking remains unrealistic to embed in real interactive analysis tools, other research by the authors proposes heuristics based on session logs to detect visual comparisons made while someone interacts with a visualization system [151]. For example, not every visualization with a filter is a hypothesis test, but every visualization with a filter condition is a test of the null hypothesis that the filter condition makes no difference compared to the distribution of the whole dataset. While it's interesting to think about detecting implicit references automatically, we suspect that completely "hands-off" approaches may be more error prone in inferring what comparisons analysts are doing compared to features that help them make their intuitive reference distributions more explicit, e.g., by letting them sketch predictions.

Presuming visual hypothesis tests can be detected as someone does exploratory analysis, the solutions that are typically proposed involve using hold-out sets. For example, Zgraggen et al. [149] compared the false positive rate when the comparisons identified in their study were followed up with by running as statistical tests on the same dataset versus testing using a new dataset of the same size from the same population. Computer science theorists have proposed more sophisticated approaches to defining hold-out sets for interactive analysis [27, 28]. These approaches won't necessarily help when datasets aren't very large, and some, such as the reusable holdout [28] require careful tuning of parameters that are difficult to reason about. Pu and Kay [102] take a higher level view, summarizing hold out and other validation approaches along with other possible corrections a visual analysis system could do for multiple comparisons, like increasing "perceptual regularization" that makes data harder to see the more comparisons one does [102].

More generally, as mentioned above, the assumption that visual comparisons are null hypothesis significance tests is a special case of the view of graphical comparisons as model checks in a Bayesian framework. In a Bayesian framework, attempts to correct for multiple comparisons can also seem like unnecessary "policing" of analysis; after all, it is often reasonable to adapt one's analysis to the data. Within a Bayesian framework, the multiple comparisons problem is often not considered a problem, as long as regularization is achieved in other ways and all relevant aspects of a model are reported with uncertainty [49].

### 4.3 Some implications for design

There are some takeaways of our argument in terms of software design requirements and opportunities. First, if model-driven inference underlies exploratory analysis, then systems should be capable of representing data generating processes (DGPs), including the model form that is believed to have generated the data (e.g., a

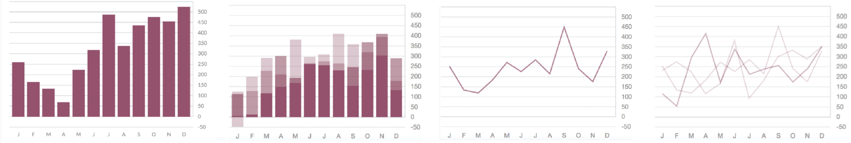

Figure 6: *Standard statistical graphics of annual jobs report data that one might encounter in an interactive visual analysis tool (first and third from left) versus depictions of probabilistic animations (second and fourth from left) means). Some research has found that people are better able at discriminating which of several models produced a noisy sample when the models in question are depicted using animation [67]).*

Gaussian process for various physical properties, or a binomial process for counts of successes or failures). While this could be done in a console or inserted in a script if the environment is already programmatic (e.g., interactive notebooks) and users are adept at statistical programming, in general users shouldn't have to leave the graphical environment they are in to switch to writing code. We like the idea of model sketching interfaces that use graphics as much as possible to elicit user input and provide feedback. For example, if the users' imagined reference distribution is generated by a model fit to only a subset of the data shown in the chart, they might select a subset by interacting with the chart, then right click to open up a window that lets them parameterize the type of model they imagine, with options customized based on the possible parameters of interest given the nature of the data.

The Bayesian models of graphical examination as posterior predictive check and Bayesian cognition motivate making it easier in tools for users to articulate prior distributions over parameters. Again in the interest of avoiding interrupting workflows to dive into code, the user could ideally "sketch" a prior graphically then generate data from it to validate as part of building models to simulate their mental models [76]. Thought should be put to how to do this, of course, given that different elicitation interfaces can lead to different strategies for formulating priors and and modeling "noise" from the interface [76, 106].

Closely aligned to representing DGPs is a need to explicitly represent uncertainty, both by plotting the reference distributions produced by externalized mental models and to express variation and uncertainty in observed data to help the user as they formulate intuitive models. While many graphical user interface tools for data analysis like Tableau or PowerBI support representing uncertainty in parameter estimates (e.g., through error bars and confidence envelopes), our own experience helping students or people we consult for represent uncertainty in common GUI systems leads us to think many users who aren't capable of doing formal modeling themselves will struggle to add uncertainty representations to their plots, even when they know they should. Above we discussed plotting observations rather than aggregations by default as one simple way software design can prioritize variation and uncertainty. One could go a step further, implementing non-parametric bootstrapping as the basis for all plotted data using animated hypothetical outcome plots (HOPs) [62] (Figure 6). We think the typical exploratory analysis tool could offer a much richer playground for analysts to explore and check their hypotheses in ways that currently remain implicit if probabilistic animation and interactive graphical approaches to eliciting predictions (e.g., drawing some representative samples, positioning intervals) were available.

One might argue that technically, many widely used graphical user interface tools for data analysis, like Tableau or MS Excel, currently do provide "modeling tools" in the form of built-in statistical tests and regression modeling functions. However, we see important differences between simply putting confirmatory and exploratory functions in the same tool and enforcing a strong connection between the two. Once users can specify their intuitive models and see distributions of predictions from them alongside observed data, some of the "noise" that currently probably gets in the way of good inferences due to having to maintain reference distributions in their heads is reduced, and the connection between exploration and confirmation is strengthened. Similarly, GUI and programmatic statistical software intended for confirmatory analysis might default to graphical model checks when presenting the typical tables of model results and fitness summaries. Probabilistic animation can be a useful tool here too, and again we should be aware of the connection between the graphical exploration and the explicit or implicit model being checked.

Seasoned tool builders might wonder what sort of new failure modes we might be creating by putting more sophisticated support for DGPs in interactive analysis tools? How might the new distributional features be misunderstood by users? One way to try to guard against this is to work toward a "grammar" for model sketching that allows the analyst to build up to more complex models from simpler components, and warns them when the models they specify are likely to be hard to fit (e.g., require a lot of data), though this would require some thought to construct. The hope is that considering implicit models as something to be gradually sketched in more detail as one works with data could influence how analysts think about model building in general, such that they become less likely to jump headfirst into complicated high dimensional models or sophisticated tests they may not understand as soon as their mindset turns confirmatory.

# 5 Comparing human to automated statistics

Beyond using theories to produce testable statements about how humans do analysis and to stimulate new design ideas, considering how we might remove humans from the analysis process entirely may also paradoxically help us find ways to improve interactive analysis. In other words, how could an Artificial Intelligence do statistics? We use this question as a thought exercise for further reflecting on the types of knowledge and strategies that come into play during interactive analysis.

In the old-fashioned view of Bayesian data analysis as inference-within-a-supermodel, it's simple enough to imagine an AI replacing a person: it simply runs some equivalent to a probabilistic program to learn from the data and make predictions as necessary. But in a modern view of statistical practice—iterating the steps of model-building, inference-within-a-model, and model-checking—it's not quite as clear how the AI works. By taking what currently seems vague and framing it computationally, we might discover useful regularities or patterns in human statistical workflows.

To fix ideas, we shall discuss Bayesian data analysis, which can be idealized by dividing it into the following three steps [48]:

1. Setting up a full probability model—a joint probability distribution for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.

2. Conditioning on observed data: calculating and interpreting the appropriate posterior distribution—the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.

3. Evaluating the fit of the model and the implications of the resulting posterior distribution: how well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step 1? In response, one can alter or expand the model and repeat the three steps.

Currently, human involvement is needed in all three steps listed above, but in different amounts:

1. Setting up the model involves a mix of look-up and creativity. We typically pick from some conventional menu of models (linear regressions, generalized linear models, survival analysis, Gaussian processes, splines, Bart [20], and so forth). Tools such as the probabilistic programming language Stan enable putting these pieces together in unlimited ways, with similar expressiveness to how we formulate paragraphs by putting together words and sentences. Right now, a lot of human effort is needed to set up models in real problems, but we could imagine an automatic process that constructs models from parts.

2. Inference given the model is the most nearly automated part of data analysis. Model-fitting programs still need a bit of hand-holding for anything but the simplest problems, but it seems reasonable to assume that the scope of the "self-driving inference program" will gradually increase. For example, for thirty years we have been able to automatically monitor the convergence of iterative simulations [52]. With the no-U-turn sampler, a recursive algorithm builds a set of likely candidate points spanning a wide swath of the target distribution and stops when it starts to retrace its steps, thus avoiding the need to tune the number of steps in Hamiltonian Monte Carlo [57].

3. The third step—identifying model misfit and, in response, figuring out how to improve the model—are likely the toughest part to automate. We often learn of model problems through open-ended exploratory data analysis, where we look at data to find unexpected patterns and compare inferences to our statistical experience and subject-matter knowledge. Indeed, a primary piece of advice we espouse to statisticians is to integrate that knowledge into statistical analysis, both in the form of formal prior distributions and in a willingness to carefully interrogate the implications of fitted models.

By considering how to fully automate all three steps, we can identify some ways to improve interactive software. The space of model parts we deem necessary to support step 1, for example, should directly guide the types of built-in options that interactive analysis tools offer an analyst to specify their implicit models. When it comes to step 2, inference within the model, we might try to build in automatic checks (for example, based on adaptive fake-data simulations) to flag problems with fitting a specified model when they appear. This could help us think about how users might note immediate problems with an implicit model as they examine graphics.

How would an AI do step 3? The AutoML approach to model evaluation typically involves choosing a preferred loss function to minimize, e.g., generalization error on held-out data, estimated using standard procedures like cross validation. But human model checking often combines model fitness measures with more qualitative assessments of how well model predictions align with domain knowledge. One approach closer to human model checking is to simulate the human in the loop by explicitly building a model-checking module that takes the fitted model, uses it to make all sorts of predictions, and then checks this against some database of subject-matter information such as a knowledge graph. This is one avenue for attempting to mimic the "Aha" process behind concepts like "insight" that drives scientific revolutions. Trying to construct this would undoubtedly require deeper inquiry into how humans check model fit, and might lead to ideas for building interactive systems, like making it easy for analysts to scan through many predictions from their models or transform them into different measures to ask "does this look right."

All that is left, then, is the idea of a separate module that identifies problems with model fit based on comparisons of model inferences to data and prior information. It's less clear how techniques from AI and ML research should be combined to do that; this may be the hardest part of the pipeline to remove humans from the loop. However, by attempting to combine existing technologies we are likely to learn more about how to think about humans doing model checks, which might also feed new interface optimizations.

# 6 Discussion

## 6.1 Integration of data exploration and modeling

Our work contributes to the statistical reform movement in a very broad sense by focusing on how software contributes to bad inference. When tools enforce a strict separation between exploratory and confirmatory analysis, we expect users to be more likely to become convinced of spurious findings and more likely to abuse confirmatory statistics in search of support. We haven't tested this assumption, but the various studies pointing to bias in exploratory visual analysis don't bode well.

Our argument about the potential consequences of prioritizing pattern exposure in creating tools for interactive EDA should not be construed as saying that exposing raw data is generally bad in analysis contexts. In contexts like communicating statistical results, showing the data or properties of the raw data can be very useful for providing information about effect size, especially in light of many readers' tendencies to overestimate effects [58]. However, as we point out above, there are various consequences of assuming non-generalization tasks as the default mode of exploratory analysis when building interactive analysis interfaces. We suspect that there is some link between prioritizing pattern exposure and a tendency in the history of interactive visualization research to consider perceptual phenomena over cognition, perhaps because at when one cares more about performance than mechanism, perception is easier to measure.

Instead, we argue for research that pursues a tighter integration between models, graphics, and data querying, motivated by a view of interactive analysis as a process of users comparing intuitive pseudo-statistical models to data via model checks. Given theories like Bayesian model checking and Bayesian cognition, it becomes clearer what types of interactive features would help with this, like built in reference distributions, and more features to push analysts toward specifying assumed models and priors while examining graphics, to name a few. As Licklider envisioned back in the 1960's, imagine having "a graphics display that allowed you to see the model's behavior—and then if you could somehow grab the model, move it, change it, and play with it interactively" [135]. While some visual analytics systems aim to make certain model results, like from deep learning, explorable (e.g., [65, 116]), the general mantra behind building tools for exploratory interactive analysis should be driven by models as well, and interface with formal model building, inference, and checking. It's encouraging that at least one recent system seems to move in this direction, by combining support for model building and checking with standard graphical exploratory data analysis support [77].

Of course, as system developers and researchers, we face a daunting challenge. Supporting interactive tools for exploratory analysis is already quite complex. For example, to make huge datasets interactive at all requires a number of database and visualization-based optimizations. Many of the interactive analysis innovations we've surveyed have an important role to play in reducing the many manual efforts required to do interactive analysis. Recommendations based on graphical features (e.g., [145, 146]) or statistical analysis of data aimed at guiding analysts toward interesting trends or anomalies (e.g., [81, 133]) are great examples of this. However, we think the field of interactive data analysis could better achieve its goals of transforming how people interact with data if such innovations were guided by theories of inference. This is not to suggest that this task will be easy, as there is much still to learn about how to gently introduce modeling capabilities without interrupting an analyst's flow, and about what users of different profiles do given more advanced modeling tools and asked to specify their expectations.

Our work also shouldn't be taken as implying that research in the subareas of visualization and databases we've surveyed are devoid of self-criticism; recent years have seen a strong critical theme emerge in research in the field of interactive visualization, for example. We think that innovation in theoretical foundations will be necessary if we ever hope to "guarantee," or even thoroughly evaluate, that the tools we build lead to good inference. The symbiosis of analyst and machine that occurs in the flow of exploratory and confirmatory statistical analysis makes it difficult to make progress on this front without considering what's going on, and what should go on, in the analyst's head.

Our argument above has focused primarily on analysis applications involving abstract data, where standard statistical graphics are the norm. In some other applications of interactive analysis and scientific visualization, users may have a harder time expressing their implicit models. For example, doctors might want to search for clinical features in large databases of medical imagery to help them in making diagnoses (e.g., [16]). When experts' implicit models are based in recognizing of visual-spatial signatures, it may be harder to elicit them, or at least require very different interfaces than we propose here. However, the fact that interactive interfaces are moving toward eliciting more input from domain experts like doctors' to facilitate their work even when their implicit models are hard to formally represent suggests some parallels despite the different assumptions that can be made about the data [16].

## 6.2 The value of a formal model of inference

We suspect that to some who are used to building systems based on intuitions might be asking, is the trouble of a formal model worth it? After all, it will never perfectly predict what people do, and it requires some significant changes to how systems are designed, e.g., by adding more sophisticated functions for making implicit models more explicit and emphasizing variation without overwhelming analysts.

We understand why one might ask these questions, but think that they are limiting. Beyond stimulating new ideas for designing interfaces and new ways of evaluating them, conceiving of interactive analysis as checks against pseudo-statistical mental models pushes us toward identifying testable implications of different formalizations of this process. This level of specificity makes it possible to recognize where our assumptions are wrong. This process is not completely absent in the status quo approach, in that current efficiency-oriented evaluations of interactive systems can also help researchers realize when their intuitions are wrong. The point it is that its likely to be less direct and error prone than if a more formal, normative model were available, similar to how it is inefficient to relying on the yes/no answers of null hypothesis significance tests.

As in many areas of human endeavor, reflecting on goals and assumptions can be valuable for many reasons, and once we formulate statistical graphics as comparisons, we can make progress by considering what are the models being implicitly compared to. A good theoretical framework of modeling feeds a process in which we learn from the ways that peoples' behavior deviate from model predictions. Just as Tukey spoke of the iterative character of the relationship of exposing and summarizing in exploratory analysis, by attempting to develop and fit statistical models to explain people's behavior during interactive analysis, we learn valuable lessons about what doesn't fit that feed new and improved modeling attempts. Speaking on a meta level, we want models of intuitive modeling in interactive analysis so that researchers, including us, who use these methods can check their models.

As we have discussed, data visualization and exploratory data analysis can be seen as a form of model checking, with the goal of revealing the unexpected beyond what is already in a model of the world, which in turn points us to an interrogation of our models and preconceptions. We also described how the Bayesian model checking formulation was useful to us for thinking about strategies for communicative visualization, like whether or how to visualize uncertainty. Similarly related to communicative visualization, graphics are often described as visual storytelling. The connection here is that stories can themselves be viewed as model checks or as explorations of anomalies, with the "twist" in a good story corresponding to a confounding of expectations [47]. Putting these together suggests that designers and readers should consider visualizations not just as artifacts on their own but also with respect to the stories they tell and the default narratives they overturn. A good model of inference can help us see the similarities between more than one pair of seemingly opposed activities.

# References

[1] BELIV '06: Proceedings of the 2006 AVI Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization, New York, 2006. Association for Computing Machinery.

[2] C. Ahlberg and B. Shneiderman. Visual information seeking using the filmfinder. In Conference Companion on Human Factors in Computing Systems, pages 433–434, 1994.

[3] D. Alabi and E. Wu. Pfunk-h: Approximate query processing using perceptual models. In Proceedings of the Workshop on Human-In-the-Loop Data Analytics, pages 1–6, 2016.

[4] G. Andrienko, N. Andrienko, S. Drucker, J.-D. Fekete, D. Fisher, S. Idreos, T. Kraska, G. Li, K.-L. Ma, J. Mackinlay, et al. Big data visualization and analytics: Future research challenges and emerging applications. In BigVis 2020: Big Data Visual Exploration and Analytics, 2020.

[5] N. Andrienko, T. Lammarsch, G. Andrienko, G. Fuchs, D. Keim, S. Miksch, and A. Rind. Viewing visual analytics as model building. In Computer Graphics Forum, volume 37, pages 275–299. Wiley Online Library, 2018.

[6] D. Asimov. The grand tour: A tool for viewing multidimensional data. SIAM Journal on Scientific and Statistical Computing, 6(1):128–143, 1985.

[7] L. Battle and J. Heer. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in Tableau. In Computer Graphics Forum, volume 38, pages 145–159. Wiley Online Library, 2019.

[8] R. A. Becker and W. S. Cleveland. Brushing scatterplots. Technometrics, 29(2):127–142, 1987.

[9] R. Beecham, J. Dykes, W. Meulemans, A. Slingsby, C. Turkay, and J. Wood. Map lineups: effects of spatial structure on graphical inference. IEEE Transactions on Visualization and Computer Graphics, 23(1):391–400, 2016.

[10] J. T. Behrens. Principles and procedures of exploratory data analysis. Psychological Methods, 2(2):131, 1997.

[11] D. J. Benjamin, M. Rabin, and C. Raymond. A model of nonbelief in the law of large numbers. Journal of the European Economic Association, 14(2):515–544, 2016.

[12] T. Brawner and J. Honaker. Bootstrap inference and differential privacy: Standard errors for free. Unpublished manuscript, 2018.

[13] D. Bromley, S. J. Rysavy, R. Su, R. D. Toofanny, T. Schmidlin, and V. Daggett. DIVE: A data intensive visualization engine. Bioinformatics, 30(4):593–595, 2014.

[14] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367(1906):4361–4383, 2009.

[15] A. Buja, D. Cook, and D. F. Swayne. Inference for data visualization. In Joint Statistical Meetings, 1999.

[16] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2019.

[17] S. Card, J. Mackinlay, and B. Shneiderman. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, 1999.

[18] M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. Van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and knowledge in visualization. IEEE Computer Graphics and Applications, 29(1):12–19, 2008.

[19] Y. Chen, J. Yang, and W. Ribarsky. Toward effective insight management in visual analytics systems. In 2009 IEEE Pacific Visualization Symposium, pages 49–56. IEEE, 2009.

[20] H. A. Chipman, E. I. George, R. E. McCulloch, et al. BART: Bayesian additive regression trees. Annals of Applied Statistics, 4(1):266–298, 2010.

[21] N. R. Chowdhury, D. Cook, H. Hofmann, M. Majumder, and Y. Zhao. Utilizing distance metrics on lineups to examine what people read from data plots. arXiv preprint arXiv:1408.1889, 2014.

[22] D. Cook, A. Buja, J. Cabrera, and C. Hurley. Grand tour and projection pursuit. Journal of Computational and Graphical Statistics, 4(3):155–172, 1995.

[23] M. Correll and J. Heer. Surprise! Bayesian weighting for de-biasing thematic maps. IEEE Transactions on Visualization and Computer Graphics, 23(1):651–660, 2016.

[24] Ç. Demiralp, P. J. Haas, S. Parthasarathy, and T. Pedapati. Foresight: Rapid data exploration through guideposts. arXiv preprint arXiv:1709.10513, 2017.

[25] P. Diaconis. Magical thinking in the analysis of scientific data. Annals of the New York Academy of Sciences, 364(1):236–244, 1981.

[26] C. Dwork. Automata, languages and programming: 33rd international colloquium, ICALP 2006, Venice, Italy. Proceedings, Part II, chapter Differential Privacy, pages 1–12, 2006.

[27] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. Generalization in adaptive data analysis and holdout reuse. In NIPS, pages 2350–2358, 2015.

[28] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth. The reusable holdout: Preserving validity in adaptive data analysis. Science, 349(6248):636–638, 2015.

[29] C. Dwork, N. Kohli, and D. Mulligan. Differential privacy in practice: Expose your epsilons! Journal of Privacy and Confidentiality, 9(2), 2019.

[30] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In TCC, pages 265–284. Springer, 2006.

[31] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In Annual International Cryptology Conference, pages 528–544. Springer, 2004.

[32] C. Dwork and A. Smith. Differential privacy for statistics: What we know and what we want to learn. Journal of Privacy and Confidentiality, 1(2), 2010.

[33] D. C. Engelbart. Conceptual Framework for the Augmentation of Man's Intellect. Spartan Books, 1963.

[34] G. Evans, G. King, M. Schwenzfeier, and A. Thakurta. Statistically valid inferences from privacy protected data, 2019.

[35] P. Federico, M. Wagner, A. Rind, A. Amor-Amorós, S. Miksch, and W. Aigner. The role of explicit knowledge: A conceptual model of knowledge-assisted visual analytics. In 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), pages 92–103. IEEE, 2017.

[36] J.-D. Fekete and R. Primet. Progressive analytics: A computation paradigm for exploratory data analysis. arXiv preprint arXiv:1607.05162, 2016.

[37] C. Ferrando, S. Wang, and D. Sheldon. General-purpose differentially-private confidence intervals. arXiv preprint arXiv:2006.07749, 2020.

[38] D. Fisher, R. DeLine, M. Czerwinski, and S. Drucker. Interactions with big data analytics. Interactions, 19(3):50–59, 2012.

[39] D. Fisher, I. Popov, S. Drucker, and M. C. Schraefel. Trust me, I'm partially right: Incremental visualization lets analysts explore large datasets faster. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1673–1682, 2012.

[40] M. A. Fisherkeller, J. H. Friedman, and J. W. Tukey. PRIM-9, an interactive multidimensional data display and analysis system. Dynamic Graphics for Statistics, pages 91–109, 1988.

[41] E. G. Freedman and P. Shah. Toward a model of knowledge-based graph comprehension. In International Conference on Theory and Application of Diagrams, pages 18–30. Springer, 2002.

[42] J. H. Friedman and W. Stuetzle. John W. Tukey's work on interactive graphics. Annals of Statistics, pages 1629–1639, 2002.

[43] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers, 100(9):881–890, 1974.

[44] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology, pages 489–500, 2015.

[45] A. Gelman. A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. International Statistical Review, 71(2):369–382, 2003.

[46] A. Gelman. Exploratory data analysis for complex models. Journal of Computational and Graphical Statistics, 13(4):755–779, 2004.

[47] A. Gelman and T. Basbøll. When do stories work? evidence and illustration in the social sciences. Sociological Methods and Research, 43:547–570, 2014.

[48] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. Bayesian Data Analysis, third edition. CRC Press, 2013.

[49] A. Gelman, J. Hill, and M. Yajima. Why we (usually) don't have to worry about multiple comparisons. Journal of Research on Educational Effectiveness, 5(2):189–211, 2012.

[50] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University, 2013.

[51] A. Gelman, X.-L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, pages 733–760, 1996.

[52] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences (with discussion). Statistical Science, 7:457–511, 1992.

[53] D. Gotz and Z. Wen. Behavior-driven visualization recommendation. In Proceedings of the 14th International Conference on Intelligent User Interfaces, pages 315–324. ACM, 2009.

[54] T. L. Griffiths, C. Kemp, and J. B. Tenenbaum. Bayesian models of cognition. In R. Sun, editor, Cambridge Handbook of Computational Cognitive Modeling, pages 59–100. Cambridge University Press, 2008.

[55] T. L. Griffiths, J. B. Tenenbaum, and C. Kemp. Bayesian inference. In K. J. Holyoak and R. G. Morrison, editors, The Oxford Handbook of Thinking and Reasoning. Oxford University Press, 2012.

[56] J. M. Hellerstein and M. Stonebraker. Readings in Database Systems. MIT Press, 2005.

[57] M. Hoffman and A. Gelman. The no-U-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. Journal of Machine Learning Research, 15:1351–1381, 2014.

[58] J. M. Hofman, D. G. Goldstein, and J. Hullman. How visualizing inferential uncertainty can mislead readers about treatment effects in scientific results. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2020.

[59] J. Hsu, M. Gaboardi, A. Haeberlen, S. Khanna, A. Narayan, B. C. Pierce, and A. Roth. Differential privacy: An economic method for choosing epsilon. In 2014 IEEE 27th Computer Security Foundations Symposium, pages 398–410. IEEE, 2014.

[60] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. IEEE Transactions on Visualization and Computer Graphics, 25(1):903–913, 2018.

[61] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay. In pursuit of error: A survey of uncertainty visualization evaluation. IEEE Transactions on Visualization and Computer Graphics, 25(1):903–913, 2019.

[62] 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, 10(11):e0142444, 2015.

[63] E. L. Hutchins, J. D. Hollan, and D. A. Norman. Direct manipulation interfaces. Human-Computer Interaction, 1(4):311–338, 1985.

[64] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. IEEE Transactions on Visualization and Computer Graphics, 19(12):2818–2827, 2013.

[65] M. Kahng, N. Thorat, D. H. P. Chau, F. B. Viégas, and M. Wattenberg. Gan lab: Understanding complex deep generative models using interactive visual experimentation. IEEE Transactions on Visualization and Computer Graphics, 25(1):1–11, 2018.

[66] A. Kale, M. Kay, and J. Hullman. Visual reasoning strategies for effect size judgments and decisions. IEEE Transactions on Visualization and Computer Graphics, 2020.

[67] A. Kale, F. Nguyen, M. Kay, and J. Hullman. Hypothetical outcome plots help untrained observers judge trends in ambiguous data. IEEE Transactions on Visualization and Computer Graphics, 25(1):892–902, 2018.

[68] A. Karduni, D. Markant, R. Wesslen, and W. Dou. A bayesian cognition approach for belief updating of correlation judgement through uncertainty visualizations. arXiv preprint arXiv:2008.00058, 2020.

[69] V. Karwa and S. Vadhan. Finite sample differentially private confidence intervals. arXiv preprint arXiv:1711.03908, 2017.

[70] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. Mastering the Information Age: Solving Problems with Visual Analytics. Goslar: Eurographics Association, 2010.

[71] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In Visual Data Mining, pages 76–90. Springer, 2008.

[72] D. Kersten and A. Yuille. Bayesian models of object perception. Current Opinion in Neurobiology, 13(2):150–158, 2003.

[73] A. Key, B. Howe, D. Perry, and C. Aragon. Vizdeck: Self-organizing dashboards for visual analytics. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, pages 681–684, 2012.

[74] A. Kim, E. Blais, A. Parameswaran, P. Indyk, S. Madden, and R. Rubinfeld. Rapid sampling for visualizations with ordering guarantees. In Proceedings of the VLDB Endowment International Conference on Very Large Data Bases, volume 8, page 521. NIH Public Access, 2015.

[75] Y.-S. Kim, P. Kayongo, M. Grunde-McLaughlin, and J. Hullman. Bayesian-assisted inference from visualized data. IEEE Transactions on Visualization and Computer Graphics, 2021.

[76] Y.-S. Kim, L. Walls, P. Krafft, and J. Hullman. A bayesian cognition approach to improve data visualization. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 2019.

[77] T. Kraska. Northstar: An interactive data science system. Proceedings of the VLDB Endowment, 11(12):2150–2164, 2018.

[78] B. C. Kwon, J. Verma, P. J. Haas, and C. Demiralp. Sampling for scalable visual analytics. IEEE Computer Graphics and Applications, 37(1):100–108, 2017.

[79] T. Lammarsch, W. Aigner, A. Bertone, S. Miksch, and A. Rind. Towards a concept how the structure of time can support the visual analytics process. In Proceedings of the International Workshop on Visual Analytics (EuroVA) in Conjunction with EuroVis, pages 9–12, 2011.

[80] P.-M. Law, A. Endert, and J. Stasko. What are data insights to professional visualization users? arXiv preprint arXiv:2008.13057, 2020.

[81] D. J. L. Lee and A. G. Parameswaran. The case for a visual discovery assistant: A holistic solution for accelerating visual data exploration. IEEE Data Engineering Bulletin, 41(3):3–14, 2018.

[82] H. Lin, D. Moritz, and J. Heer. Dziban: Balancing agency & automation in visualization design via anchored recommendations. In Proceedings of the 38th Annual ACM Conference on Human Factors in Computing Systems - CHI '20, page 12, 2020.

[83] Z. Liu and J. Heer. The effects of interactive latency on exploratory visual analysis. IEEE Transactions on Visualization and Computer Graphics, 20(12):2122–2131, 2014.

[84] J. Mackinlay. Automating the design of graphical presentations of relational information. ACM Transactions On Graphics (TOG), 5(2):110–141, 1986.

[85] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. IEEE Transactions on Visualization and Computer Graphics, 13(6):1137–1144, 2007.

[86] M. Majumder, H. Hofmann, and D. Cook. Validation of visual statistical inference, applied to linear models. Journal of the American Statistical Association, 108(503):942–956, 2013.

[87] B. H. McCormick. Visualization in scientific computing. Computer Graphics, 21(6), 1987.

[88] R. McCormick. The NYT's election forecast needle is stressing people out with fake jitter. The Verge, 2016.

[89] N. McCurdy, J. Gerdes, and M. Meyer. A framework for externalizing implicit error using visualization. IEEE Transactions on Visualization and Computer Graphics, 25(1):925–935, 2018.

[90] S.-M. McKernan, C. Ratcliffe, C. E. Steuerle, C. Quakenbush, E. Kalish, T. Meko, B. Chartoff, F. Blackshaw, and S. Lei. Nine charts about wealth inequality in America (updated). Urban Institute, 2017-10-05.

[91] Microsoft. Power BI.

[92] D. Moritz, D. Fisher, B. Ding, and C. Wang. Trust, but verify: Optimistic visualizations of approximate queries for exploring big data. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 2904–2915, 2017.

[93] M. C. Mozer, H. Pashler, and H. Homaei. Optimal predictions in everyday cognition: The wisdom of individuals or crowds? Cognitive Science, 32(7):1133–1147, 2008.

[94] F. Nguyen, X. Qiao, J. Heer, and J. Hullman. Exploring the effects of aggregation choices on untrained visualization users' generalizations from data. In Computer Graphics Forum. Wiley Online Library, 2020.

[95] C. North. Toward measuring visualization insight. IEEE Computer Graphics and Applications, 26(3):6–9, 2006.

[96] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. Decision making with visualizations: A cognitive framework across disciplines. Cognitive Research: Principles and Implications, 3(1):29, 2018.

[97] Y. Park, M. Cafarella, and B. Mozafari. Visualization-aware sampling for very large databases. In 2016 IEEE 32nd International Conference on Data Engineering (ICDE), pages 755–766. IEEE, 2016.

[98] A. Perer and B. Shneiderman. Balancing systematic and flexible exploration of social networks. IEEE Transactions on Visualization and Computer Graphics, 12(5):693–700, 2006.

[99] A. Perer and B. Shneiderman. Systematic yet flexible discovery: Guiding domain experts through exploratory data analysis. In Proceedings of the 13th International Conference on Intelligent User Interfaces, pages 109–118, 2008.

[100] S. Pinker. A theory of graph comprehension. Artificial Intelligence and the Future of Testing, pages 73–126, 1990.

[101] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In Proceedings of International Conference on Intelligence Analysis, volume 5, pages 2–4. McLean, VA, 2005.

[102] X. Pu and M. Kay. The garden of forking paths in visualization: A design space for reliable exploratory visual analytics: Position paper. In 2018 IEEE Evaluation and Beyond-Methodological Approaches for Visualization (BELIV), pages 37–45. IEEE, 2018.

[103] P. Rahman, L. Jiang, and A. Nandi. Evaluating interactive data systems. VLDB Journal, 29:119–146, 2020.

[104] S. Rahman, M. Aliakbarpour, H. K. Kong, E. Blais, K. Karahalios, A. Parameswaran, and R. Rubinfield. I've seen enough incrementally improving visualizations to support rapid decision making. Proceedings of the VLDB Endowment, 10(11):1262–1273, 2017.

[105] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. IEEE Transactions on Visualization and Computer Graphics, 20(12):1604–1613, 2014.

[106] A. Sarma and M. Kay. Prior setting in practice: Strategies and rationales used in choosing prior distributions for bayesian analysis. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2020.

[107] S. Saul. New 2020 poll shows three-way tie among Sanders, Warren and Biden. New York Times, 2019-08-26.

[108] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. In Proceedings of the 29th Annual Symposium on User Interface Software and Technology, pages 365–377, 2016.

[109] R. C. Shah and J. P. Kesan. Policy through software defaults. In Proceedings of the 2006 International Conference on Digital Government Research, pages 265–272. Citeseer, 2006.

[110] B. Shneiderman. A computer graphics system for polynomials. Mathematics Teacher, 67(2):111–113, 1974.

[111] B. Shneiderman. The future of interactive systems and the emergence of direct manipulation. Behaviour & Information Technology, 1(3):237–256, 1982.

[112] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In Proceedings of the 1996 IEEE Symposium on Visual Languages, pages 336–343. IEEE, 1996.

[113] A. Srinivasan and J. Stasko. Orko: Facilitating multimodal interaction for visual exploration and analysis of networks. IEEE Transactions on Visualization and Computer Graphics, 24(1):511–521, 2017.

[114] M. Steyvers, J. B. Tenenbaum, E.-J. Wagenmakers, and B. Blum. Inferring causal networks from observations and interventions. Cognitive Science, 27(3):453–489, 2003.

[115] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. IEEE Transactions on Visualization and Computer Graphics, 8(1):52–65, 2002.

[116] H. Strobelt, S. Gehrmann, H. Pfister, and A. M. Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. IEEE Transactions on Visualization and Computer Graphics, 24(1):667–676, 2017.

[117] D. F. Swayne, D. Cook, and A. Buja. Xgobi: Interactive dynamic data visualization in the X window system. Journal of Computational and Graphical Statistics, 7(1):113–130, 1998.

[118] D. F. Swayne, D. T. Lang, A. Buja, and D. Cook. Ggobi: evolving from xgobi into an extensible framework for interactive data visualization. Computational Statistics & Data Analysis, 43(4):423–444, 2003.

[119] Tableau Software. Tableau Desktop.

[120] Tableau Software. Tableau launches Vizable, a breakthrough mobile app for data exploration, 2015-10-20.

[121] S. Tauber, D. J. Navarro, A. Perfors, and M. Steyvers. Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. Psychological Review, 124(4):410, 2017.

[122] S. Teal. Journalists: Now Tableau Prep is free for you, 2018.

[123] J. B. Tenenbaum, T. L. Griffiths, and C. Kemp. Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Sciences, 10(7):309–318, 2006.

[124] J. J. Thomas and K. A. Cook. Illuminating the Path: The Research and Development Agenda for Visual Analytics. National Visualization and Analytics Ctr, 2005.

[125] J. W. Tukey. The future of data analysis. Annals of Mathematical Statistics, 33(1):1–67, 1962.

[126] J. W. Tukey. Exploratory Data Analysis. Addison-Wesley, 1977.

[127] J. W. Tukey. Data-based graphics: Visual display in the decades to come. Statistical Science, 5(3):327–339, 1990.

[128] J. W. Tukey and M. B. Wilk. Data analysis and statistics: An expository overview. In Proceedings of the November 7-10, 1966, Fall Joint Computer Conference, pages 695–709, 1966.

[129] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. Science, 185(4157):1124–1131, 1974.

[130] S. van den Elzen and J. J. van Wijk. Small multiples, large singles: A new approach for visual data exploration. In Computer Graphics Forum, volume 32, pages 191–200. Wiley Online Library, 2013.

[131] S. VanderPlas and H. Hofmann. Spatial reasoning and data displays. IEEE Transactions on Visualization and Computer Graphics, 22(1):459–468, 2015.

[132] M. Vartak, S. Huang, T. Siddiqui, S. Madden, and A. Parameswaran. Towards visualization recommendation systems. ACM SIGMOD Record, 45(4):34–39, 2017.

[133] M. Vartak, S. Rahman, S. Madden, A. Parameswaran, and N. Polyzotis. See DB: Efficient data-driven visualization recommendations to support visual analytics. Proceedings of the VLDB Endowment, 8(13):2182–2193, 2015.

[134] E. Vul, N. Goodman, T. L. Griffiths, and J. B. Tenenbaum. One and done? Optimal decisions from very few samples. Cognitive Science, 38(4):599–637, 2014.

[135] M. M. Waldrop. The Dream Machine: J. C. R. Licklider and the Revolution That Made Computing Personal. Viking Penguin, 2001.

[136] E. Wall, L. M. Blaha, L. Franklin, and A. Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In 2017 IEEE Conference on Visual Analytics Science and Technology (VAST), pages 104–115. IEEE, 2017.

[137] X. Wang, D. H. Jeong, W. Dou, S. Lee, W. Ribarsky, and R. Chang. Defining and applying knowledge conversion processes to a visual analytics system. Computers & Graphics, 33(5):616–623, 2009.

[138] L. Wasserman. Minimaxity, statistical thinking and differential privacy. Journal of Privacy and Confidentiality, 4(1), 2012.

[139] L. Wasserman and S. Zhou. A statistical framework for differential privacy. Journal of the American Statistical Association, 105(489):375–389, 2010.

[140] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. IEEE Transactions on Visualization and Computer Graphics, 16(6):973–979, 2010.

[141] L. Wilkinson. The grammar of graphics. In Handbook of Computational Statistics, pages 375–414. Springer, 2012.

[142] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., pages 157–164. IEEE, 2005.

[143] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. IEEE Transactions on Visualization and Computer Graphics, 12(6):1363–1372, 2006.

[144] C. Williamson and B. Shneiderman. The dynamic homefinder: Evaluating dynamic queries in a real-estate information exploration system. In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 338–346, 1992.

[145] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. IEEE Transactions on Visualization and Computer Graphics, 22(1):649–658, 2015.

[146] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pages 2648–2659, 2017.

[147] Y. Wu, L. Xu, R. Chang, and E. Wu. Towards a Bayesian model of data visualization cognition. In IEEE Visualization Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe), 2017.

[148] Y. Yan, L. J. Chen, and Z. Zhang. Error-bounded sampling for analytics on big sparse data. Proceedings of the VLDB Endowment, 7(13):1508–1519, 2014.

[149] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2018.

[150] Y. Zhao, D. Cook, H. Hofmann, M. Majumder, and N. R. Chowdhury. Mind reading: Using an eye-tracker to see how people are looking at lineups. International Journal of Intelligent Technologies & Applied Statistics, 6(4), 2013.

[151] Z. Zhao, L. De Stefani, E. Zgraggen, C. Binnig, E. Upfal, and T. Kraska. Controlling false discoveries during interactive data exploration. In Proceedings of the 2017 ACM International Conference on Management of Data, pages 527–540, 2017.