Warning, Bias May Occur: A Proposed Approach to Detecting Cognitive Bias in Interactive Visual Analytics

Emily Wall* Georgia Tech

Leslie M. Blaha† Pacific Northwest National Laboratory

Lyndsey Franklin‡ Pacific Northwest National Laboratory

Alex Endert§ Georgia Tech

ABSTRACT

Visual analytic tools combine the complementary strengths of humans and machines in human-in-the-loop systems. Humans provide invaluable domain expertise and sensemaking capabilities to this discourse with analytic models; however, little consideration has yet been given to the ways inherent human biases might shape the visual analytic process. In this paper, we establish a conceptual framework for considering bias assessment through human-in-the-loop systems and lay the theoretical foundations for bias measurement. We propose six preliminary metrics to systematically detect and quantify bias from user interactions and demonstrate how the metrics might be implemented in an existing visual analytic system, InterAxis. We discuss how our proposed metrics could be used by visual analytic systems to mitigate the negative effects of cognitive biases by making users aware of biased processes throughout their analyses.

Keywords: cognitive bias; visual analytics; human-in-the-loop; mixed initiative; user interaction;

Index Terms: H.5.0 [Information Systems]: Human-Computer Interaction—General

1 INTRODUCTION

Visual analytic systems gracefully blend sophisticated data analytics with interactive visualizations to provide usable interfaces through which people explore data [43, 71]. User interaction is central to the effectiveness of visual analytic systems [21, 56, 80]. It is the mechanism by which people and systems communicate about the data, allowing people to become an integral part of the data exploration process. Analysts leverage their domain expertise and reasoning skills to explore the data via the user interface. They communicate their intents and questions to the system, realized as guiding analytic models or changing the visualization's parameters.

We argue that user interactions play a powerful second role in addition to shaping analytical models: interactions form an externalized record of users' thought processes. Interactive visual analytics supports guiding endogenous attention, creating and organizing declarative memory cues, parsing and chunking information, aiding analogical reasoning, and encouraging implicit learning [55]. Interactions mark the paths of exploratory data analysis, providing an opportunity to glean insight into a person's reasoning and decision making processes [54, 62].

The balance between human interaction and machine automation is the primary focus of mixed-initiative visual analytics [39]; however, the trade-offs of incorporating humans into analytics are not well understood. In mixed-initiative tools, people interact with applications to steer computational models, to explore alternative representations, and to augment models with valuable subject matter expertise. These human-in-the-loop (HIL) approaches enable insights in many domains, especially where uncertainty is high and human reasoning is a valuable addition to data-intensive computation [20].

However, incorporating human reasoning and analysis into computational models may have unwanted side effects. Prior work in cognitive psychology informs us that there are inherent limitations to cognitive processes, such as working memory capacity limits [11,51]. One limitation relevant to analytic processes and visual data analysis is cognitive bias, errors resulting from the use of fallible decision making heuristics [29, 42]. Evidence that cognitive biases impact users' decision making abounds; recent work has shown that information visualization users are not immune to cognitive biases [13]. While bias might exist and be propagated through a system via data collection (e.g., convenience sampling bias), data processing (e.g., algorithm bias), visual mappings (e.g., visual perception bias), etc. [27, 64], here we focus on cognitive bias injected by analysts.

Several cognitive biases have been previously identified as particularly relevant to data analysis and the intelligence process [38] (see Table 1). Such biases can have far-reaching effects, influencing the evidence upon which analysts rely and the hypotheses they form. Further, when user interaction in visual analytic tools is intended to guide analytic models, cognitive biases might be propagated to and amplified by the underlying computational models. The resulting biased analytic models may ultimately prompt analysts to make incorrect or inferior decisions, or simply echo the users' biases back to them. This constitutes an emergent bias in computational systems [27]. We note that cognitive bias is not all bad nor does use of heuristics always produce errors in reasoning; on the contrary, use of heuristics is often positive, producing quicker and more effective decision making. Such efficiencies may be useful within an HIL system. Thus, we seek ways to understand how bias arises in HIL analytics, to harness positive effects when useful and mitigate negative effects when they might be damaging.

We hypothesize that when data analysis is supported by visual analytic tools, analysts' cognitive biases influence their data exploration in ways that are measurable through their interactions with the data. This presents an opportunity to leverage user interactions to detect and assess mental pitfalls in real time during the analysis process. While models exist that incorporate measures of human bias, they rely on the final *products* of cognition (e.g., a final choice decision). This does not allow for the real-time measurement of bias in the decision making process. Instead, we propose that cognitive bias can be detected earlier in an analysis *process*, using metrics applied to the user's interactions. Real-time assessment of cognitive performance can be leveraged for adaptive interfaces, responsive to individualized user needs [5, 25]. However, the critical first step in developing systems for cognitive augmentation or mitigation is construction of conceptual frameworks for detecting and assessing a user's cognitive state [22, 28]. This provides the theoretical basis for interpreting behaviors to provide the right machine-based bias interventions.

In this paper, we present theoretical foundations for quantifying indicators of cognitive bias in interactive visual analytic systems

*e-mail: emilywall@gatech.edu
†e-mail:leslie.blaha@pnnl.gov
‡e-mail:lyndsey.franklin@pnnl.gov
§e-mail: endert@gatech.edu

and propose six preliminary metrics. These metrics are based on the notions of *coverage* and *distribution*, targeting assessment of the process by which users sample the data space. We propose a way to quantify interactions and a naïve baseline model for an unbiased analysis against which the metrics can be interpreted. We emphasize that our proposed metrics do not map one-to-one onto any particular biases. Rather, they describe behavioral indicators that might result from any number of underlying cognitive biases. We discuss how future refinement of the baseline models will serve to shape the interpretation of the metrics for bias assessment, and illustrate the metrics in action with InterAxis [44].[1]

2 Why Study Bias in Visual Analytics?

Many research efforts have been dedicated to developing or identifying how visualizations support human cognition in effective ways. For example, Green et al. [36] introduced a human cognition model in response to increasing complexity of data visualizations, resulting in design guidelines grounded in cognitive science. Wright et al. [79] also noted that designing interfaces to facilitate external visual thinking can minimize the risk for some cognitive biases. Fisher et al. [24] added additional structure to the ways cognition had been previously considered by applying a translational cognitive science model to visual analytics. Patterson et al. [55] identified a set of visualization leverage points, together with suggested metrics, where knowledge of attention and memory processes can guide visualization design. We build on this understanding of the cognitive mechanisms supporting the analysis process by contributing a novel set of metrics for quantifying behavioral indicators of cognitive bias based on user interactions.

Human cognition is particularly relevant to HIL systems. The design of these systems combines the complementary strengths of humans (adaptation, accommodation, and perceptual expertise) and machines (working memory and large-scale information processing) [36] and are widely considered superior to human-only or machine-only alternatives for specific tasks and domains [39]. Human-only approaches can result in too heavy a cognitive load and do not scale with increased processing requirements. Machine-only approaches can result in a lack of user trust and are infeasible if appropriate training data is not readily available. Thus, the field of HIL visual analytics focuses on finding the appropriate balance of human and machine effort [39]. Numerous systems implement such approaches for dimension reduction in scatterplots [6, 19, 44, 47], distance function learning [7], ranking [76], and sensemaking recommendations [10]. However, the trade-offs of mixed-initiative systems have yet to be fully explored. In particular, while humans bring intuition and domain expertise into analytics, they also bring *bias* into analytics via interaction.

Many HIL systems utilize interaction for model steering. For example, ForceSPIRE [18] uses semantic interaction to incrementally update parameters based on the user's interactions while shielding users from model complexities. However, the way humans interact with data, in this case text documents, is subject to their cognitive biases. They may subconsciously pay particular attention to documents that confirm a pre-existing hypothesis (confirmation bias) [53, 77] or rely heavily on documents which are most recent (availability heuristic) [72]. As bias steers users' cognitive processes, bias also steers users' behavior through interactions in visual analytic systems and thus the underlying models as well. Consequently, changes in model parameters and data statistics systematically reflect the analyst's bias. Recently, Gotz et al. [32] used this logic to address selection bias in healthcare data using a quantitative distance measure to compare variable distributions in the analyst's selected data subset to that of the whole data set. Similar quantitative approaches could be leveraged to capture multiple types of cognitive bias, shaping model evolution through interactive visual analytics.

[1] Live demonstration can be found in the supplemental video.

In the case of intelligence analysis, cognitive biases can result in dire real world consequences. One highly publicized example is the Madrid Train Bombing Case, where the confirmation bias of forensic analysts contributed to the misidentification and arrest of an innocent man [16, 69]. Such cases motivate the need to better understand the role of bias in HIL visual analytics to enable people to make better decisions about the desired balance of human and machine control. Further, by understanding when bias may be present, we can potentially integrate ways to mitigate the negative effects and ultimately produce better analytic results.

3 Related Work

In the following sections, we discuss work relevant to the challenge of cognitive bias in visual analytics. Those areas include related work on bias from cognitive science (Section 3.1), understanding how people perform analyses (Section 3.2), and describing prior work on capturing and inferring about user interactions (Section 3.3).

3.1 Bias in Cognition

Prior work in cognitive psychology informs us that there are two key components to understanding reasoning and decision making processes: (1) how information is organized mentally (including perceptual, memory, and semantic organization); and (2) how that organization is aligned with decision boundaries or mapped to response criteria [48]. Cognitive activities in both areas are susceptible to pitfalls that can result in misinterpretations or erroneous decisions. For information organization processes, these pitfalls include perceptual illusions and false memories. For decision making processes, these pitfalls are collectively referred to as logical fallacies and cognitive biases. These various pitfalls arise naturally from our perceptual and intuitive decision making processes. Therefore they cannot be avoided or eliminated. However, we can be aware of their occurrence and use deliberate reasoning processes to scrutinize and overcome the negative consequences of biased cognition [41].

Bias can be defined in different ways for visual analytics [75]. In this paper, our definition most closely aligns with the perspective of "bias as a model mechanism." While bias typically has a negative connotation, it is not always undesirable. At its most basic level, bias can be thought of as a way to describe where in the decision process or organizational space people place their decision criteria. That is, where do people draw the line between one response option versus another when performing some cognitive task. From this perspective, there are multiple modeling approaches with a parameter quantifying bias for a given task or decision process. Models of perceptual organization, such as the theory of signal detection [34, 35, 50] or the similarity choice axiom [49, 58], use proportions of correct and incorrect responses to describe performance in terms of perceptual discriminability and decision boundary bias. Stochastic decision making models of choice behavior use proportions of response choices and response speeds to capture bias as a relationship between the speed of mental evidence accumulation and response thresholds [9, 63]. A commonality among these techniques for quantifying bias is that they rely on post-experiment analysis of the decision making process. That is, the models for bias are based on the *product* of a user's cognitive operations. This places a strong constraint on the use of these approaches to situations wherein we have complete sets of decisions.

From this body of related work, we learn that while product-based analyses for detecting bias exist, they are limited. Specifically, they are not suited for making people aware of their potential biases during analysis. Thus, we are motivated to establish methods to detect cognitive bias during the interactive exploration *process*, inferred through user interaction over the course of an analytic task. We conceptualize interaction in visual analytic systems as a direct capture of the reasoning process used during data analysis. In this way, user interactions constitute a novel set of measurable behaviors that could

be used to study and model logical fallacies and cognitive biases in the analytic process [73, 74]. Our assumptions are consistent with the recent efforts to use hand, mouse, or eye tracking trajectories to model continuous cognition, which have shown that the shapes of movement across a computer interface reflect mental organization and biases throughout the whole response process [46, 67, 68]. In this paper, we describe methods for real-time detection of potentially biased analysis behavior from user interaction sequences.

3.2 Studying the Analytic Process

The process of learning about data through a visual interface is often referred to as the visual analytic sensemaking process. Sensemaking, however, is a more general process by which information is gathered, hypotheses are formulated, evidence is extracted, and the hypotheses are evaluated. For HIL data analytics, this is a process of exploring the data attributes together with the data model predictions and attempting to explain any patterns against the conceptual models or hypotheses framing the problems of interest.

The sensemaking process was studied by Pirolli and Card [57] by performing a cognitive task analysis with intelligence analysts. They proposed that the sensemaking process could be roughly described by two loops: (1) a foraging loop to search for information, and (2) a sensemaking loop to resolve an understanding of the information. Klein et al. [45] studied the sensemaking process with the observation that analysts begin with some frame of reference when examining data, then continuously compare, refine, and create new frames throughout analysis. Similarly, Sacha et al. [65] describe the process of knowledge generation in visual analytics in terms of the related roles of the human and computer. Their model consists of loops for knowledge generation, verification, and exploration. It is clear from these models that the process of learning and making inferences about data can entail a number of cognitive and perceptual decisions, such as data identification, pattern detection, information discrimination, classification, and selection between discrete options. Multiple types of bias may be introduced into the process by each type of decision, and they may be compounded over the repeated sensemaking cycles.

3.3 Interaction in Visual Analytics

Interaction is paramount in visual analytics [56]. It advances a visualization from one state to the next, allowing users to navigate and understand increasingly complex data. Interaction facilitates human reasoning; it is the mechanism by which users go through the process of analysis and is a vital part of the reasoning process in visual analytics [59]. Through interaction, users get acquainted with the data, form and revise hypotheses, and generate questions [1]. It allows users to focus their attention in the presence of potentially overwhelming information throughout their analysis [36]. As a key facilitator for human reasoning in visual analytics, interactions are used to better understand more than just analytic results. They also illuminate the process that led to those results [54]. Typically, however, interaction is ephemeral; that is, once it has triggered the appropriate system response, the information contained in the interaction is discarded.

In response to this loss of data, log analysis tools have been developed to support analytic provenance. A prominent example is GlassBox [12], which captures keyboard and mouse interactions in an interface. Interaction data has been used for things like interactive model-steering [18], user authentication based on mouse movements [61], and even inferring personality traits [8]. Another common use for interaction data is analytic provenance, where users' analytic processes, strategies, and methods can be reconstructed based on their interaction sequences [15, 33]. Thus, given prior work showing the power of interaction data for making inferences about users, we hypothesize that user interactions can capture behaviors which may correspond to biased analysis.

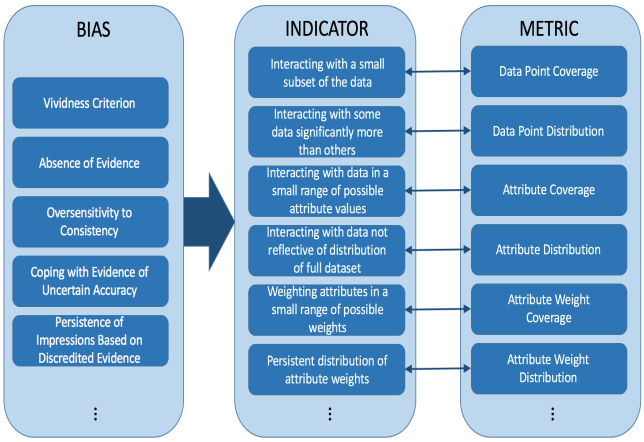


Figure 1: Cognitive biases result in behavioral indicators that are measurable by the proposed metrics. We scope this paper to those indicators and metrics depicted above, but there are numerous other biases, behavioral indicators, and ways to measure those indicators.

4 Formalizing Cognitive Bias in Visual Analytics

In this section, we outline the ways cognitive bias may manifest in the analytic process and discuss relationships between bias indicators and the proposed metrics.

4.1 Behavioral Indicators of Bias in Interaction

Cognitive bias is a consequence of heuristic decision making processes that allow people to simplify complex problems and make more efficient judgments [42, 73]. A heuristic is a "rule of thumb" for making an inference, or a strategic way in which information is ignored to get to a decision faster [30]. Heuristics frequently ignore or subconsciously weight certain types of information. As a subconscious cognitive process, heuristics also play an integral role in visual analytics. Concerted efforts have been made to delineate the cognitive biases to which analysts may be susceptible [38]. This provides a starting point for understanding biases in the inference and sensemaking process.

There are dozens of cognitive biases captured in the heuristics and biases literature [30, 41]. The cognitive biases relevant to a set of interactions are dependent on the nature of the task people are performing. We focus herein on the cognitive biases that typically make the evaluation of evidence an effective process. We refer to the evaluation of evidence as the process by which data are determined to be relevant to the analysis process at hand. Heuer [38] describes five types of cognitive biases particularly relevant for evaluating evidence, defined in Table 1: *vividness criterion, absence of evidence, oversensitivity to consistency, coping with evidence of uncertain accuracy*, and *persistence of impressions based on discredited evidence* (also known as the *continued influence effect*). Each type of bias, including those in Table 1, impacts people's behavior in predictable ways. The third column in the table gives an example of how each given type of bias might specifically influence a user's interactions. For each of these examples, we can compute on several measurable patterns of user interaction, which we refer to as **behavioral indicators of bias** or just **indicators of bias**.

We emphasize that our approach is based on the claim that there is *not* a one-to-one mapping between cognitive biases and the proposed metrics. When a user is biased, we expect to find these patterns in their interactions; however, detecting a particular indicator does not necessarily tell us which type of cognitive bias may have caused the behavioral response. We have diagrammed this relationship between the types of cognitive biases discussed in this paper and the set of proposed metrics for measuring indicators of bias in Fig. 1. The block

Bias	Description	Interaction Manifestation
Vividness Criterion	humans rely more heavily on information that is specific or personal than information that is abstract or lacking in detail	e.g., analyst frequently returns to / interacts with data points that are rich in detail
Absence of Evidence	humans tend to focus their attention on the information that is present, ignoring other significant pieces of evidence that may be missing	e.g., analyst filters out a subset of data, forgets about it, and makes future decisions without accounting for the missing data
Oversensitivity to Consistency	humans tend to choose hypotheses that encompass the largest subset of evidence	e.g., analyst interacts almost exclusively with data that supports the largest encompassing hypothesis, dismissing other data
Coping with Evidence of Uncertain Accuracy	humans tend to choose to accept or reject a piece of evidence wholly and seldom account for the probability of its accuracy	e.g., analyst filters out data that supports a seemingly unlikely hypothesis, thus fully rejecting it
Persistence of Impressions Based on Discredited Evidence	humans tend to continue to believe information even after it has been discredited (also known as the *continued influence effect*)	e.g., analyst continues to interact with data supporting a hypothesis that has been disproved

Table 1: Cognitive biases relevant to intelligence analysis [38] that produce the measurable behavioral indicators we focus on in this paper.

arrow between biases and indicators represents a many-to-many mapping, the particulars of which we defer to future work. Here we focus on developing metrics that relate to individual indicators of bias.

4.2 What Can We Measure?

To identify ways in which we might measure bias from interaction data, we need to develop two key pieces of theory: (1) what can be measured, and (2) a method of interpreting the measurements.

To address (1), we must identify the sets of possible things that can be measured, from which we can derive metrics. Herein we focus on combinations of {types of interaction} with {objects of interaction}. That is, types of interaction include things like clicks, hovers, and drags afforded by a system that can be explicitly captured by event listeners. Semantically similar interactions supported by other device modalities can be mapped to our proposed metrics, but ultimately need to be bound to event handlers. For our preliminary metrics, objects of interaction currently include data points, attributes, and attribute weights; however, we could conceivably measure interactions with many other objects, including analytic model parameters or interactions with particular views in a multiview interface. Further, the metrics can only account for the data set loaded in the system. For example, if an analyst is examining a data set of criminal suspects, the metrics would not be able to infer about a bias toward a person not represented in the data set.

To address (2), we must develop baseline models of behavior that would reflect performance under assumptions of non-biased information gathering or decision making to make appropriate inferences about biased behaviors. We assert that we can formulate models of interaction behavior by conceptualizing the set of data points and possible interactions with those points as a state space over which we can define Markov chains. That is, we let each interaction with a data point be a state in a state space. A user performing that {point, interaction} combination has transitioned to the associated state in the Markov chain. The transition probabilities are the likelihood of subsequent interaction options given the current state or current interaction. For example, if clicking on a point means you are likely to next click on a point in close proximity, the transition probability would be high between those two states. As we will develop further, the data set defines the points, the interface defines the possible interactions on those points, and together, the visual analytic system defines the state space. Our Markov chain provides a generalizable approach to describing any sequence of interactions with an analytic system. The model can be changed to capture different analytic behaviors by simply altering the transition matrix for the Markov chain on that state space. In this way, we can study different patterns of biased and unbiased behaviors to define relevant baselines for different domains all within a common theoretical framework. But in this work, we will use a simple Markov chain, defined later, making minimal assumptions about what constitutes unbiased behaviors.

To formalize our preliminary metrics, we first define some common notation, which is summarized in Table 2. We define $D = \{d_1, \ldots, d_N\}$ to be a data set of size N. Each data point d_i has a set of M attributes, $A = \{a_1, \ldots, a_M\}$. We define D_U to be the unique set of data points interacted with by a user. $I(D)$ is the set of interactions by a user on the data set, and $T = \{\text{click}, \text{hover}, \ldots\}$ is the set of interaction types. Within a visual analytic system, the set of possible interaction events is $T \cup D$, the union of the set of interaction types afforded by the interface and the set of data points.[2]

In a finite set of items, we define the concepts of *coverage* and *distribution*. *Coverage* refers to the degree to which $I(D)$ has sampled or covered the set $T \cup D$. We mean to use coverage in an intuitive way here, referring roughly to the amount of data exploration that a user has made on a data set. Coverage is related to the notion of a cover for a set. The cover for $T \cup D$ is a collection of sets whose union contains $T \cup D$ as a subset. In terms of interactions, the cover for $T \cup D$ is the union of all sets of interactions $I(D)$ possible in the analytic process. In information visualization, the concept of coverage has been studied as a means to encourage users to explore more data [14, 23, 31, 66, 78] as well as inform users of collaborators' explorations [2, 40]. The concept of coverage is motivated by the desire to ensure that the full extent of the data is considered, even if it represents an outlier or otherwise lesser portion of the distribution of data.

Alternatively, the concept of distribution is motivated by the desire to ensure that the user's interactions with the data are proportional to the actual dispersion of the data. *Distribution* refers to the dispersion of the set of interactions $I(D)$. Distribution differs from coverage in that it accounts for repeated interactions rather than considering only the binary notion of set membership. For a set of interactions, the probability frequency function over the dimension of interest for $I(D)$ defines the shape of the dispersion of the data with which the user has interacted.

[2] We note that in most non-streaming visual analytic systems, T and D, as well as $T \cup D$ are finite; streaming data systems have the potential for countably infinite data set sizes, but we leave consideration of those sets to later work.

Key to our present interest in modeling evolving behavior as people interact with systems is that we can track the events in $I(D)$ that are created by the user over the course of an analytics session. We propose that by tracking these events as a Markov chain over the state space $T \cup D$, we can define metrics characterizing $I(D)$ in ways that reflect information gathering and decision making processes. When compared to a baseline, these proposed metrics will enable us to assess when behavior differs from the baseline in meaningful ways. In the present work, we focus on meaningful deviations that might reflect cognitive biases. Further, for each metric, we define the bias value $0 \leq b \leq 1$, where higher values indicate more prominent indicators of bias, and lower values indicate less prominent indicators of bias.

For our preliminary metrics, we assume a simple baseline model of independent, equally likely interactions with any data point. At any given time, the probability of interacting with data point d_i on step $k+1$ is $P[d_{i,k+1}|d_{j,k}] = 1/N$, meaning that each next interaction does not depend on the current interaction or the interaction history. A sequence of interactions in $I(D)$ thus forms a regular Markov chain, with the data points representing the states in the chain with transition probability matrix $P = \left[\frac{1}{N}\right]$. Fig. 2 illustrates the Markov chain resulting from four interactions with a scatterplot. The sequence of actions taken by the user was: (1) hover over point d_1; (2) hover over point d_2; (3) hover over point d_3; and (4) click on point d_3. The resulting Markov chain, given in set notation is $\{\{hover, d_1\}, \{hover, d_2\}, \{hover, d_3\}, \{click, d_3\}\}$. The green trajectory over Fig. 2a to 2d illustrates the sequence of interaction events as a movement through a state space, with the growing list of {interaction, data point} dyads forming the set $I(D)$ for this user. The dashed red arrows show the unbiased baseline model, where a transition from the current (green) point to every other point, including self-transition, is equally likely.

While the assumption of uniformity is naïve, it is intended to be only a preliminary point of comparison. It allows us to establish the metrics while making few assumptions about what unbiased behavioral indicators look like, because they are likely domain and interface dependent. However, we note that the Markov chain approach allows us to flexibly swap out the transition probability matrix without altering the computation of the proposed metrics themselves. We further discuss the process of creating better baseline representations of unbiased behavior as future work in Section 7.1.

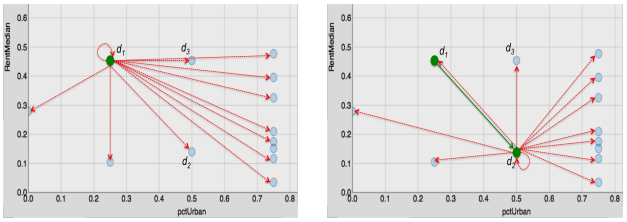
5 Preliminary Metrics for Cognitive Bias

We hypothesize that when cognitive bias is present, it should manifest in particular patterns of interaction with the data. In this section, we propose six preliminary metrics for detecting behavioral indicators of bias based on a user's interactions. We quantify behavioral indicators and define the expected values derived from the Markov chain baseline model. For each metric, we give a description, the mathematical formulation, and an example use with a type of bias from Table 1.

5.1 Data Point Metrics

5.1.1 Data Point Coverage

Description. The data point coverage metric is an ordinal measure of the user's attention to the data points in the data set. In particular, it measures the amount of the data set with which the user has interacted compared to the expected amount. In an unbiased exploration of the entire available data, the metric decreases over time as the user interacts with more of the data set. Of course, early in the analysis, fewer data points will have been interacted with than later in the analysis, so we must account for the number of possible interactions. So the question for the metric with respect to bias is: Is there a time in the process where the the data point coverage is much smaller than would be predicted by the unbiased baseline model?



(a) Interaction 1: hover over point d_1; resulting Markov chain $\{\{hover, d_1\}\}$

(b) Interaction 2: hover on point d_2; resulting Markov chain $\{\{hover, d_1\}, \{hover, d_2\}\}$

(c) Interaction 3: hover over point d_3; resulting Markov chain $\{\{hover, d_1\}, \{hover, d_2\}, \{hover, d_3\}\}$

(d) Interaction 4: click on point d_3; resulting Markov chain $\{\{hover, d_1\}, \{hover, d_2\}, \{hover, d_3\}, \{click, d_3\}\}$

Figure 2: The Markov chain formed by the first four interactions with a scatterplot, superimposed on top of a visualization for illustrative purposes. The set of {interaction, data point} combinations constitutes the states of the Markov chain. Subsequent interactions are conceptualized as the transitions between the states. A green point indicates a data point that has been interacted with. The red arrows indicate possible transitions from the current state.

Formulation. For data point coverage, we consider the size of the set of interactions relative to the expected value of the baseline model. We define $I(D)$ and D_U as above. Let $\kappa(D_U)$ be the size or cardinality of the set of unique points interacted with at any point in time, and let $\kappa(D) = N$ be the cardinality of the whole data set. $\kappa(D_U) \leq \kappa(D)$, and $\kappa(D_U)$ approaches $\kappa(D)$ as the user explores more of the data set.

From the baseline Markov chain defined by the sequence of interactions in $I(D)$, we define $\hat{\kappa}(D_U)$ as the expected number of unique data points interacted with in $I(D)$. After k interactions on a data set, or k transitions in the Markov chain, we can define a set of k-multisets, which are the sequences of length k with N possible objects in which any single data point could be revisited up to k times. In k-multisets, the expected value of the number of unique data points visited in k interactions is defined by

$$\hat{\kappa}(D_U) = \frac{N^k - (N-1)^k}{N^{k-1}}. \tag{1}$$

We then define the data point coverage metric b_{DPc} according to Eq. 2.

$$b_{DPc} = 1 - \min\left(\frac{\kappa(D_U)}{\hat{\kappa}(D_U)}, 1\right) \tag{2}$$

Example. To understand how this metric might be useful in capturing behavioral indicators of bias, consider the following. An analyst may propagate her/his bias by focusing on (e.g., repeatedly interacting with) or ignoring (e.g., not interacting with) certain data points. For example, when an analyst uses the *vividness criterion* [38], s/he subconsciously relies more heavily on evidence that is vivid or personal than on evidence that is dull or impersonal.

Notation	Description
b_μ	bias metric from the set of all metrics μ, with range $b_\mu \in [0,1]$, where higher values indicate more prominent indicators of bias
$D = \{d_1, ..., d_N\}$	data set of size N
$A = \{a_1, \ldots, a_M\}$	set of M attributes describing data set D
$T = \{click, hover, ...\}$	set of interaction types
D_U	unique set of data points interacted with by the user, where $D_U \subseteq D$
$I(d_n)$	set of interactions with data point $d_n \in D$
$\kappa(X)$	cardinality of set X
$\hat{\kappa}(X)$	expected cardinality of set X, based on a Markov chain model of user interactions
$w = [w(a_1), \ldots, w(a_M)]$	attribute weight vector

Table 2: Notation used to describe the bias metrics

Thus, bias would be propagated through the system by interacting with only a small, vivid subset of the full set of evidence.

5.1.2 Data Point Distribution

Description. The data point distribution metric is a measure of bias toward repeated interactions with individual data points or subsets of the data. Here we compare the frequency function of data point interactions to a baseline uniform distribution of interactions across all D. Data point distribution aids in determining if the user is focusing their interactions unevenly across the data set.

Formulation. We can detect this by measuring the distribution of interactions with the data points. The baseline model of independent, equally-likely interactions with the data points predicts a uniform distribution of interactions. We compute the χ^2 statistic, comparing the actual number of interactions with each data point to the expected baseline uniform distribution according to Eq. 3.

$$\chi^2 = \sum_{n=1}^{N} \frac{(\kappa(I(d_n)) - \hat{\kappa}(I(d_n)))^2}{\hat{\kappa}(I(d_n))} \tag{3}$$

Here, $\kappa(I(d_n))$ denotes the observed number of interactions with data point d_n, while $\hat{\kappa}(I(d_n))$ denotes the expected number of interactions with d_n. Derived from the regular Markov chain of interactions with $P = [1/N]$, after k interactions, $\hat{\kappa}(I(d_n)) = k/N$, equivalent to the expected number of times returning to data point d_n in k steps. The p-value is obtained from the χ^2 distribution with $N-1$ degrees of freedom, then the metric value is defined according to Eq. 4.

$$b_{DPd} = 1 - p \tag{4}$$

Example. To understand how this metric might be useful in capturing behavioral indicators of bias, again consider the *vividness criterion* example. When an analyst uses the *vividness criterion* [38], they subconsciously rely more heavily on evidence that is vivid or personal than they do evidence that is dull or impersonal. Consequently, when evaluating evidence and forming hypotheses, they are likely to return to those most vivid pieces of information disproportionately to their actual value as evidence. This is measurable by considering the distribution of interactions across data points.

5.2 Attribute Metrics

5.2.1 Attribute Coverage

Description. Different from considering the way the set of interactions cover the set of data points, we can also consider the way the points in D_U cover the ranges of values for the data attributes, A. Thus, for each attribute, the attribute coverage metric measures the range of values explored by the user's interactions. It gauges whether the data interacted with by the user presents a comprehensive or narrow image of the full range of values along each dimension of the data set. If a user interacts with data in the full range of values for a given attribute, the metric will be low; alternatively, if a user only interacts with data in a small range of the possible attribute values, the metric will be high.

Formulation. Attribute coverage is computed for each attribute separately, though a single data point interaction impacts all attributes simultaneously. Attribute coverage refers to the degree to which the user interactions have sufficiently covered the range of attribute values. For categorical attributes, we define "sufficiently covered" to mean that at least one data point has been interacted with for each value $q \in Q$ that the attribute can take. For continuous attributes, we define "sufficiently covered" by quantizing the data into Q quantiles.

Let $I(D)$ and D_U be defined as above. Let Q_{a_m} be the set of Q categorical values or quantiles for attribute a_m. We then define the attribute coverage metric for attribute $a_m \in A$, according to Eq. 5.

$$b_{Ac}(a_m) = 1 - \min\left(\frac{\kappa(D_{U,Q_{a_m}})}{\hat{\kappa}(D_{U,Q_{a_m}})}, 1\right) \tag{5}$$

where $\kappa(D_{U,Q_{a_m}})$ is the cardinality of the set of values/quantiles for attribute a_m covered by the set of unique data points with which the user has interacted. Thus, b_{Ac} is greater when the user has not interacted with data over the full range of values of a_m.

Similar to the data point coverage metric, the sequence of Q_{a_m} sampled in k interactions forms a k-multiset for attribute a_m. In k-multisets, the expected value of the number of unique attribute values visited in k interactions is defined by

$$\hat{\kappa}(D_U, Q_{a_m}) = \frac{Q_{a_m}^k - (Q_{a_m} - 1)^k}{Q_{a_m}^{k-1}}. \tag{6}$$

As this is computed per attribute, there will be as many b_{Ac} scores as there are attributes of the data. It is possible for a person to have broad attribute coverage of some attributes and low attribute coverage of others.

Example. Consider an analyst subject to *oversensitivity to consistency* [38]. This bias can cause the analyst to dismiss evidence that is not part of the greatest encompassing hypothesis. It may lead to fruitless pursuit of an incorrect hypothesis if alternative evidence is not weighed and considered appropriately. Thus, an analyst subject to this bias might see consistent evidence that a suspect's vehicle is black and only examine black cars. The analyst might be dismissive of different accounts that the vehicle was blue or silver and consequently neglect to properly investigate alternatives. The bias would thus cause her to only interact with a portion of the range of possible attribute values in the data set.

5.2.2 Attribute Distribution

Description. The attribute distribution metric is a measure for detecting bias toward particular attributes of the data. For each attribute of the data, we compare the distribution of the data interacted with to the distribution of the full data set.

Formulation. Define $A = \{a_1, ..., a_M\}$ as the set of attributes describing the data. For numerical attributes (e.g., car price), we compare the distribution of data that has been interacted with D_U

to the distribution of the full data set D using a Kolmogorov-Smirnov (KS) test, a nonparametric test for comparing continuous distributions. The KS statistic for attribute a_m is defined by $S_{(N,n',a_m)} = \sup_x \left| F_{D,N,a_m}(x) - F_{D_U,n',a_m}(x) \right|$, where $F_{D,N,a_m}(x)$ and $F_{D_U,n',a_m}(x)$ are the cumulative distribution functions for attribute a_m over the whole data set and the subset of unique interaction points, respectively, $n' = \kappa(D_U)$, and sup is the supremum function. We compute the empirical p-value using the KS distribution.

When the attribute a_m is categorical (e.g., gender), we apply a χ^2 test with $\kappa(Q_{a_m})$ degrees of freedom to compare the distribution of data across the categorical values. We define the test statistic according to Eq. 7.

$$\chi^2 = \sum_q \frac{(\kappa(a_{m,q}) - \hat{\kappa}(a_{m,q}))^2}{\hat{\kappa}(a_{m,q})} \tag{7}$$

In this case, each value of q in $a_{m,q}$ represents a different value of the categorical attribute a_m. The observed value $\kappa(a_{m,q}) = \kappa(I(D))$ where $d_n[a_m] = a_{m,q}$ represents the number of data points interacted with by the analyst that have value q for attribute a_m. The expected values $\hat{\kappa}(a_{m,q})$ are derived from the actual distributions of the attribute values.

For both numerical and categorical variables, we define the attribute distribution metric b_{Ad} for attribute a_m using the p-value for the KS-test and χ^2-test, respectively, according to Eq. 8.

$$b_{Ad}(a_m) = 1 - p \tag{8}$$

Thus, the value of $b_{Ad}(a_m)$ increases when the distribution of attribute a_m values of data points in D_U significantly differs from the distribution of attribute a_m values in D.

Example. Consider an analyst subject to *oversensitivity to consistency* [38]. If the analyst focuses on the data that is consistent with the greatest encompassing hypothesis, the distribution of the data in D_U will likely be skewed compared to the distribution D. In the case of examining suspect vehicles, 75% of the analyst's interactions may be with black cars while only 15% of the candidate vehicles are black. Thus, this metric can capture bias along particular dimensions of the data.

5.3 Attribute Weight Metrics

Attribute weights are used in visual analytic systems implicitly or explicitly to quantify the importance of each attribute in the data toward some decision. Users often specify attribute weights by interacting with interface sliders to specify each attribute's importance. The attribute weight metrics compare the coverage and distribution of weights that each attribute has been assigned by the user or system. We define an attribute weight vector $w = [w(a_1), \ldots, w(a_M)]$ comprised of numerical weights assigned to each attribute.

5.3.1 Attribute Weight Coverage

Description. We can consider the way the weights in w cover the possible ranges of values for the attribute weights. Thus, for each attribute, the attribute weight coverage metric measures the range of values explored by the user interactions. It gauges whether the attribute weights identified by the user's interactions present a comprehensive or narrow image of the full range of weights for each attribute. If a given attribute has had a wide range of weights applied, the metric will be low; however, if the weight for a given attribute has not taken on a diverse set of values, the metric will be high.

Formulation. With respect to attribute weights, the notion of coverage can be determined by comparing the weights the user has assigned to each attribute to the possible range of attribute weights. Again, this form of coverage is not about the shape of the distribution of weights for each attribute. Rather, attribute weight coverage refers to the degree to which the user interactions have sufficiently covered the range of attribute weight values. We first quantize each attribute's weight into Q quantiles. We then define "sufficiently covered" to mean that at some point, the weight for attribute a_m has taken on a value in each of the Q quantiles.

Let $Q_{w_{a_m}}$ be the set of quantiles for the weight of attribute a_m. We then define the attribute weight coverage metric for attribute $a_m \in A$, according to Eq. 9.

$$b_{AWc}(a_m) = 1 - \min\left(\frac{\kappa(W_{U,Q_{a_m}})}{\hat{\kappa}(W_{U,Q_{a_m}})}, 1\right) \tag{9}$$

where $\kappa(W_{U,Q_{a_m}})$ is the cardinality of the set of weight quantiles for attribute a_m covered by the set of unique attribute weights that the user has defined. Thus, b_{AWc} is greater when the user has not defined w_{a_m} to have a diverse range of values.

Similar to the attribute coverage metric, the sequence of $Q_{w(a_m)}$ sampled in k interactions forms a k-multiset for attribute weight $w(a_m)$. In k-multisets, the expected value of the number of unique attribute weights visited in k interactions is defined by

$$\hat{\kappa}(W_{U,Q_{a_m}}) = \frac{Q^k_{w(a_m)} - (Q_{w(a_m)} - 1)^k}{Q^{k-1}_{w(a_m)}}. \tag{10}$$

Example. After a piece of evidence has been discredited, analysts should re-weight attributes in accordance with new information. However, analysts subject to *persistence of impressions based on discredited evidence* [38] will likely continue to rely on the same weighting of attributes throughout their investigation. The bias would thus influence the analyst to examine a smaller part of the range of attribute weights.

5.3.2 Attribute Weight Distribution

Description. The attribute weight distribution metric detects bias toward particular weightings of data attributes. For each data attribute, we compare the distribution of the changes in attribute weight to a baseline exponential distribution of changes in weight.

Formulation. The attribute weight distribution metric is based on the distribution $F(\Delta w(a_m))$ of the amount of change in an attribute weight between two interaction at times τ_i and τ_j, $\Delta w(a_m) = w_{\tau_i}(a_m) - w_{\tau_j}(a_m)$. The baseline assumption is that users will be more likely to make small changes (e.g., $\Delta w(a_m)$ close to 0) to the weight of an attribute than they are to make large changes. In the present, we assume a baseline exponential distribution, $f_{\hat{\Delta}}(x) = \lambda e^{-\lambda x}$, with $\lambda = 1$. We compare the two distributions using a KS test. The KS statistic for the weight of attribute a_m is defined by $S_{(\Delta w(a_m))} = \sup_x \left| F_{\Delta w(a_m)}(x) - F_{\hat{\Delta} w(a_m)}(x) \right|$, where $F_{\hat{\Delta} w(a_m)}(x) = (1 - e^{-x})$. We then define the attribute weight distribution metric b_{AWd} for attribute a_m using the p-value for the KS test, according to Eq. 11.

$$b_{AWd}(a_m) = 1 - p \tag{11}$$

Thus, $b_{AWd}(a_m)$ increases when the distribution of weights for attribute a_m is far from the expected exponential distribution.

Example. As with the attribute weight coverage metric, consider the example of the *persistence of impressions based on discredited evidence* [38]. After a piece of evidence has been discredited, the analyst is likely to change the attribute weights very little if at all. Thus, the tail of the distribution representing large changes in attribute weights would be smaller than the expected distribution.

6 Example Application

In this section, we present an example of how the proposed bias metrics might be incorporated into a visual analytic system.

The System. InterAxis [44] is an exploratory visual analytic system that allows users to steer scatterplot axes by interacting directly with data points. The user can interact with the data by:

- hovering over a point to see details,
- dragging a point into a bin along either axis, and
- dragging an attribute bar to change its assigned weight.

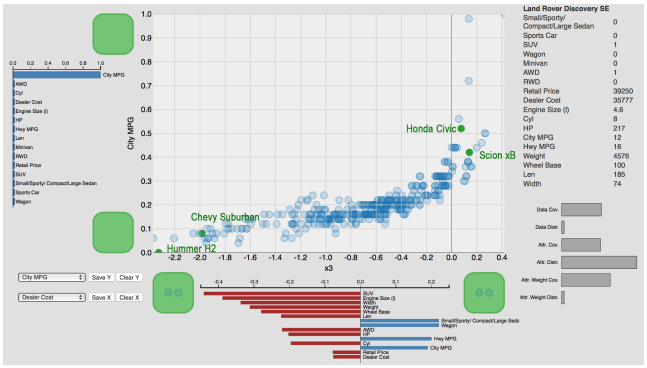
Dragging points into the bins on the high or low side of the axes represents the user's semantic distinction between the data in the bins (e.g., things they like v. things they do not like; things that are important v. things that are unimportant; etc). The system then computes an attribute weight vector based on the difference between the two sets of examples for each axis. The attribute weight vector is bound to the axis, and the data points on the scatterplot are moved to their respective locations along the axis, with items semantically similar to the respective examples appearing on either side of the axis. Fig. 3 shows the InterAxis interface with the addition of the bias metrics, described below. The original paper by Kim et al. [44] contains further details on the interface and underlying model. Herein, we use the Cars data set [37].

Interactions. We modified the system by incorporating a custom Javascript library used to log interactions and compute the proposed metrics based on those interactions. For InterAxis, we track the interactions with data points supported by the system: hovers and drags. We also track interactions with the attribute weight vectors for each axis: when users select an attribute to bind to the axis, directly modify the weights by clicking and dragging, or when the axis is recomputed based on dragging points into the bins. The custom library uses Jerzy [60], a statistics library for Javascript, to derive probabilities from the KS test.

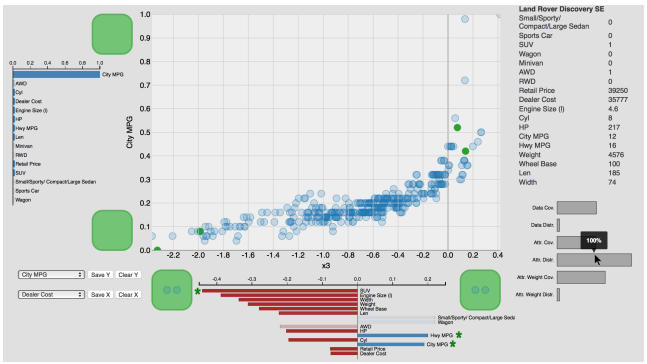
Bias Metrics. The six bias metrics are recomputed after each interaction. The level of bias computed for each metric is a number between 0 and 1, encoded as the width of the bar in the metric visualization (seen on the lower right of each image in Fig. 3). For the attribute and attribute weight metrics, the bar encodes the maximum metric value over all the attributes. Hovering on the bars reveals a tooltip showing the percentage value for each metric, shown in Fig. 3b and 3c. Additional details are encoded in visual channels not otherwise used by the system to facilitate easier interpretation of the metrics. Hovering over the bar for *data point coverage* changes the sizes of the data points. Data points in D_U are given a larger radius, while data points not in D_U are given a smaller radius. Hovering over the bar for *data point distribution* similarly changes the sizes of the data points. Points with the most interactions are given the largest radius, while points with the fewest interactions are given the smallest radius. Hovering over the bars for the attribute and attribute weight metrics (*attribute coverage*, *attribute distribution*, *attribute weight coverage*, and *attribute weight distribution*) encodes the metric value for individual attributes as the opacity of the bars along the axes. High opacity represents a higher bias metric result. Thus, darker bars correspond to dimensions along which the bias is greater. Some of these views can be seen in Fig. 3.

Usage Scenario. Sofia is interested in exploring potential cars to purchase, using the Cars data set [37]. She first changes the scatterplot axes to Dealer Cost (X axis) and City MPG (Y axis) to reflect her two most important criteria. She hovers over several data points to understand the relationship between Dealer Cost and City MPG. She notices the top left of the scatterplot (low cost and high MPG) has hybrid and other fuel-efficient cars. She likes Scion xB and Honda Civic, so she drags the two cars to the high end of the X axis. Toward the middle and low end of the Y axis, she notices a few cars she does not like (Hummer H2 and Chevy Suburban) and drags them to the bin on the low side of the X axis. At this point, InterAxis has mapped what is important to Sofia along the X axis, shown in Fig. 3a. Cars on the right are small cars and wagons that have high city and highway MPG, and cars on the left are heavy cars with large engine sizes, with the leftmost being SUVs.

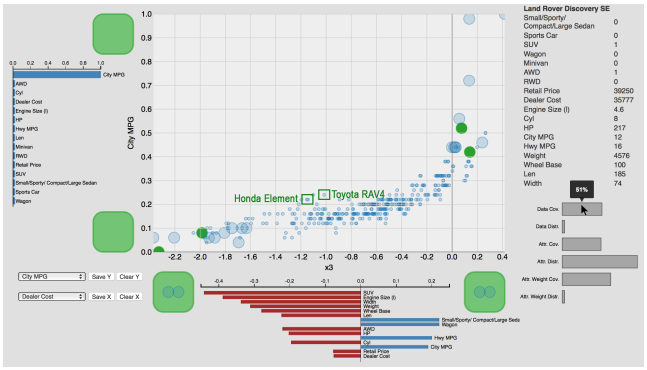
She notices that the attribute distribution bar has grown indicating a high bias, so she hovers over the bar (Fig. 3b). The tooltip shows the maximum metric value across all attributes; in this case, several



(a) Sofia's exploration of the cars in the data set. She has dragged cars she likes to the bin on the high side of the X axis and cars she does not like to the bin on the low side of the X axis.



(b) The Attribute Distribution Metric. On hover, the opacity of the bars on the axes encode the metric value for each attribute. Here, Sofia notices the dark bars indicating bias along the dimensions of SUV and city and highway MPG.



(c) The Data Point Coverage Metric. On hover, the radii of the points Sofia interacted with are increased, and the radii of the points not interacted with are decreased. Sofia notices that she has not interacted with any cars except on the extreme ends of the axes. After further exploration, she ultimately chooses Toyota RAV4 and Honda Element, unexpectedly in the middle of the X axis, to test drive.

Figure 3: A depiction of InterAxis throughout the usage scenario described in Section 6. InterAxis is a system that allows users to define custom scatterplot axes using dimension reduction by interacting with data points. The proposed bias metrics have been integrated into InterAxis in the bar visualization on the lower right portion of the interface.

of the metrics are at or near 100%. The dimensions she examines are annotated with a green asterisk. She sees that the Dealer Cost and

Retail Cost attribute bars have high opacity indicating a bias along those dimensions. Sofia acknowledges this was intentional; she is on a budget, so she interacted primarily with inexpensive cars. She also notices that the SUV attribute bar is dark. This was intentional too; she wants cars that have higher fuel economy. Next, Sofia hovers over the data point coverage bar (Fig. 3c). She notices she has interacted with many cars on the extreme ends of the X axis but not much in between. Curious what types of vehicles lie in between, Sofia hovers over several cars in the mid-range of the X axis. She sees several mid-sized cars and a few small crossover SUVs. She had previously dismissed SUVs, because she thought they had poor fuel economy.

The bias metrics computed on Sofia's interaction sequences were visualized in the interface, allowing Sofia to gain an awareness of her analytic process and biases. Her preconceived notion that SUVs have poor fuel economy led her to initially dismiss an entire class of vehicles. However, the visual characterization of her analytic process through the bias metrics shed light onto her oversight. She ultimately selects two cars of interest: Toyota RAV4 and Honda Element (annotated in Fig. 3c) to test drive.

While this scenario provides an illustrative example, the ideas generalize to other exploratory or decision making tasks. Cognitive bias impacts people's behaviors in ways that can be described and quantified from their interactions. How to best present the metric information to the user is a fascinating area of future work that we discuss further in the next section.

7 Discussion

We defined and demonstrated six bias metrics as a critical first step toward creating quantifiable models of cognitive bias in visual analytics. However, they are preliminary metrics requiring further refinement and testing. In this section, we present limitations of the current metrics as well as some of the larger open research questions.

7.1 Generalizing the Metrics

In this section, we discuss some of the factors that were considered in defining the proposed bias metrics.

Baselines. First, we define baseline distributions for the metrics that assume uniform distributions of interactions, formalized as a regular Markov chain where transitions between any two points and self-transitions are all equally likely. In many cases, this is probably not an appropriate assumption, depending on the task and context. For example, an analyst may be instructed by her supervisor to investigate only female suspects, while another analyst may be responsible for investigating male suspects. Using the current baseline comparison, the metrics would detect a bias along the gender dimension. However, if we change the baseline Markov model such that the transition probabilities make it more likely to interact with certain points over others, then the metrics can be assessed against a more appropriate baseline behavior. In general, the metrics can be refined with the context of the analyst's assigned task, opening an interesting direction of research to understand how users communicate their tasks to systems in the context of bias. Alternatively, the baseline model could be defined by interaction probabilities derived from cognitive models of decision making performance, further increasing the fidelity of the comparison of an unbiased baseline model to real human behavior.

Data Types. The metrics are agnostic to the nature of the underlying data. The notions of coverage and distribution can be applied to interactions with time-series or graph data, for example, by logging the relevant information. In the case of graphs, that might mean applying coverage and distribution concepts to the links between the data in addition to the data points themselves. For time-series data, it might be relevant to compute metrics that determine bias toward particular time windows. The key to integrating bias metrics is to use an interface enabling interactions with the data.

Log Scope. Each metric is currently computed treating all interactions equivalently, but certain types of interactions $t \in T$ might be more important or semantically meaningful in the system. Thus, the metrics could be computed and interpreted separately based on interaction type, or the interactions used to compute each metric could be weighted according to the importance of the interaction type. Similarly, the window of interactions used to compute the metrics may be an important factor for metric interpretations. We currently consider the entire history of an interaction session in the metric calculations. This approach might shed light on long-standing biases. Narrower time frames (e.g., 15 minute windows) could illuminate shorter-scale patterns of bias where the user self-adjusted or changed strategy over the session.

Interaction Types. We have primarily considered primitive interactions with data points in the proposed metrics (e.g., click, hover, drag, etc.). More complex interactions across a visual analytic system can be considered as well. The attribute weight metrics are examples that do not rely on interactions with data points, but rather consider interactions with analytic model components. We will want to account for interactions like filtering, zooming, switching between alternative visualizations, or brushing and linking between multiple coordinated views, and incidental interactions will need to be discounted. In all cases, we include the possible interactions in T so they can be included in $T \cup D$, and a Markov chain can be computed over the set of interactions $I(D) \subset T \cup D$. We can then derive appropriate baselines and relevant metrics to inform users of biases toward particular data representations.

Scalability. As the metrics are used to describe the decision making process, they can be considered a space-saving asset in the case of understanding provenance. Rather than preserving cumbersome log files for post-hoc analysis, the bias metrics might be computed during the analytic process. However, several factors might improve the scalability of the metrics themselves. For example, adjusting the window used in the metric computations could serve to improve the scalability of the proposed approach. Scalability could further be improved by computing the metrics using incremental algorithms that do not require the full interaction history to be saved and recomputed, but rather update the model based on the stream of interactions. An incremental approach would also improve the scalability of the metrics for high dimensional or sparse data.

7.2 Bias Mitigation

Our proposed bias metrics constitute an approach to real-time user state assessment, because we are tracking behaviors throughout the analytics process. There are at least three strategies for providing feedback based on the information gathered from real-time cognitive state assessment [4]: (1) provide it to the user, (2) provide it to the machine, or (3) provide it to an external agent. Developing a successful strategy for mitigating cognitive bias in mixed-initiative visual analytic systems depends on identifying when and how each of the above strategies might be employed with positive outcomes [28]. There have been varying degrees of past success addressing bias in the analytics process. We suggest how our proposed metrics may enhance HIL bias mitigation approaches.

Feedback to Users. Our bias metrics can be provided directly back to the users as feedback about their analytic processes. This leaves interpretation and any subsequent actions to the user's discretion. A number of attempts have been made to provide feedback-based bias mitigation to intelligence analysts, including training courses, videos, and reading material. These techniques have not consistently proven to be effective. As articulated by Heuer: *"Cognitive biases are similar to optical illusions in that the error remains compelling even when one is fully aware of its nature. Awareness of the bias, by itself, does not produce a more accurate perception"* [38]. Awareness can be raised by simply presenting the metrics on an interface, as in our InterAxis example in Section 6. The goal is to

promote informed decision making by the analyst, potentially leading to a shift in user behavior accordingly. Other researchers have similarly tried to raise awareness by visualizing analytic provenance or coverage of the possible exploration space [14, 40, 78]. With such feedback, users tended to explore more data [23], make more unique discoveries [78], and show greater search breadth without sacrificing depth [66]. Thus, visual characterization of the analytic process has potential to mitigate bias by altering a user's exploration.

Serious games provided a more effective alternative to traditional means of bias feedback [3, 17, 26, 52, 70]. These techniques educated analysts about cognitive biases, but did little to mitigate negative effects when biases inevitably occurred in the analytic process. They reinforce that an analyst must be pro-active using feedback to adjust her/his behaviors to mitigate negative bias effects. Analysis of competing hypotheses (ACH) [38] is a conscious tactic that can be used during the analytic process to evaluate the likelihood of multiple hypotheses in an unbiased way. ACH creates a framework for analysts to assess the relevance of each piece of evidence for multiple hypotheses, and systematically eliminate less compelling hypotheses until a single most likely hypothesis remains. While an effective analytic tool, ACH is a time-consuming process not always used in practice. Feedback from our bias metrics might encourage analysts to employ ACH more frequently.

Feedback to Machines. Machine feedback supports adaptive systems or other machine-based cognitive augmentations that are responsive to the user's state. Machine-driven automated bias mitigation could be incorporated into mixed-initiative systems. Using the metrics for automated detection and mitigation of behavioral indicators of bias would require little additional effort from the analyst if the mixed-initiative system is taking steps to determine appropriate mitigations. Some mixed-initiative efforts have already begun to integrate visual analytic recommendations based on user interest or semantic interactions [18]. Gladisch and colleagues [31] even suggest using the notion of interest through user interactions to penalize users or down-weight some recommendations to guide the user to other parts of the data space. This is one way in which mixed-initiative systems can steer users around bias-related pitfalls. As we gain a better understanding of how bias manifests in behavioral indicators, we can develop more techniques for mixed-initiative systems to leverage the bias metrics to promote desired unbiased interaction patterns.

Feedback to Other Agents. Feedback about biased behaviors can be given to a third party agent (e.g., a human teammate, a supervisor, a machine monitor). This strategy could prove useful in collaborative analytics settings. For example, analysts teaming on a project may be alerted to each other's biased behaviors, to ensure they cross-validate each other's work. We leave the development of such team-based bias mitigation to future efforts.

The potential for real-time bias detection opens up many questions surrounding how to most effectively mitigate the negative effects of cognitive bias: *How should the system inform the user when bias is detected? When and at what frequency should the system notify the user of bias or take initiative to intervene? To what extent should the system act on behalf of the user when bias is detected?* There is a rich space to be explored to understand the consequences of bias in visual analytics. Our theoretical foundations in this paper provide a starting point to improve HIL systems by better understanding and harnessing bias.

7.3 Confounding Expertise and Context

The word bias itself has a negative connotation. It evokes a sense of imperfection that we tend to think we can overcome with careful critical thinking and reflection. However, we emphasize that not all bias is bad. The same heuristic approach to problem-solving that produces cognitive biases is what allows us to not be bogged down by constant trivial decisions. It allows us to solve problems more quickly and to make fast perceptual judgments.

In the analytic process, humans have intuition and expertise to guide them. However, the interaction patterns of expert analysts and cognitively biased analysts might look very similar despite very different cognitive processes. Consider the case of an analyst focusing his attention on evidence surrounding a particular suspect. Such focus may result from cognitive bias, or it may result from quick deliberate decisions based on years of experience. The analyst might also have knowledge about the case not captured by the data at the time, like breaking new evidence. Thus, it is important to understand the role context and domain expertise play in structuring the visual analytic process to differentiate expertise from cognitive biases producing an inferior analytic process.

User annotations of their own interactions would be one possibility for improving the machine's ability to distinguish expert and biased behavior. This would facilitate creating a common understanding between the system and user by eliciting explicit user feedback and reflection. The metrics could then be adjusted in real time to weight subsequent interactions accordingly, so that confounding factors are not confused as negative biases. Consider how the metrics indicated that Sofia had a bias on price and SUV dimensions. The metrics accurately characterized Sofia's exploration, but her focus was intentional. If Sofia could annotate her intentions, the metrics could be adjusted accordingly. In future work, we hope to study the extent to which interaction patterns differ for cognitively biased users, expert analysts, and users with contextual information not captured in the data. Additionally, we hope to understand how this distinction impacts bias mitigation techniques.

8 Conclusion and Future Work

Humans and machines offer complementary strengths for visual data exploration in HIL visual analytics. However, humans are subject to inherent cognitive and perceptual limitations, including cognitive bias. While a great deal is known about bias, we lack techniques to measure bias in real-time during the visual analytics process. Thus in this paper, we focused on developing the underlying theory and set of metrics for detecting behavioral indicators of cognitive bias in visual analytics. These metrics can be used to better understand the patterns of interaction of biased individuals, to inform HIL systems that can begin to measure, monitor, and mitigate the negative effects of bias.

Future work includes filling in pieces of a larger research agenda to make the metrics usable and useful in real-world analysis scenarios. We frame these pieces of future work as three primary research questions. (1) What does unbiased behavior look like in user interaction patterns? Answering this question is critical to creating an accurate Markov Chain baseline model of unbiased behavior against which user interaction patterns can be compared. (2) How are the bias metrics related to existing post-decision models of cognitive bias? By use of appropriate problem framing [74], we can encourage different patterns of biased behavior and ultimately validate the metrics through theoretically-grounded experimental designs. (3) How do we present the metrics to users in such a way that the negative effects of bias are optimally mitigated? This involves an exploration of the interpretability of the metrics and interface design in the context of carefully controlled experiments.

Acknowledgments

The research described in this document was sponsored by the U.S. Department of Energy through the Analysis in Motion Initiative at Pacific Northwest National Laboratory (PNNL) and the Department of Defense through PNNL. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES

[1] N. Adrienko and G. Adrienko. *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Springer-Verlag, New York, 2005.

[2] K. Badam, Z. Zeng, E. Wall, A. Endert, and N. Elmqvist. Supporting team-first visual analytics through group activity representations. *Graphics Interface*, 2017.

[3] E. Bessarabova, C. W. Piercy, S. King, C. Vincent, N. E. Dunbar, J. K. Burgoon, C. H. Miller, M. Jensen, A. Elkins, D. W. Wilson, S. N. Wilson, and Y. H. Lee. Mitigating bias blind spot via a serious video game. *Computers in Human Behavior*, 62:452–466, 2016.

[4] L. M. Blaha, C. R. Fisher, M. M. Walsh, B. Z. Veksler, and G. Gunzelmann. Real-time fatigue monitoring with computational cognitive models. In *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*, vol. 9743 of *Lecture Notes in Computer Science*, pp. 299–310. Springer, 2016.

[5] B. J. Borghetti and C. F. Rusnock. Introduction to real-time state assessment. In *Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*, vol. 9743 of *Lecture Notes in Computer Science*, pp. 311–321. Springer, 2016.

[6] J. Broekens, T. Cocx, and W. A. Kosters. Object-centered interactive multi-dimensional scaling: Ask the expert. *Proceedings of the 18th Belgium-Netherlands Conference on Artificial Intelligence (BNAIC)*, pp. 59–66, 2006.

[7] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 83–92, 2012.

[8] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1663–1672, 2014.

[9] J. R. Busemeyer and J. T. Townsend. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3):432–459, 1993.

[10] K. Cook, N. Cramer, D. Israel, M. Wolverton, J. Bruce, R. Burtner, and A. Endert. Mixed-initiative visual analytics using task-driven recommendations. *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 9–16, 2015.

[11] N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24:87–185, 2001.

[12] P. Cowley, L. Nowell, and J. Scholtz. Glass box: An instrumented infrastructure for supporting human interaction with information. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences, 2005. HICSS'05*, p. 296c. IEEE, 2005.

[13] E. Dimara, A. Bezerianos, and P. Dragicevic. The attraction effect in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):471–480, 2017.

[14] H. Doleisch, H. Hauser, M. Gasser, and R. Kosara. Interactive focus+ context analysis of large, time-dependent flow simulation data. *Simulation*, 82(12):851–865, 2006.

[15] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang. Recovering reasoning process from user interactions. *IEEE Computer Graphics & Applications*, 29(3):52–61, 2009.

[16] I. Dror. Combating bias: The next step in fighting cognitive and psychological contamination. *Journal of Forensic Sciences*, 57(1):276–277, 2012.

[17] N. E. Dunbar, M. L. Jensen, C. H. Miller, E. Bessarabova, S. K. Straub, S. N. Wilson, J. Elizondo, J. K. Burgoon, J. S. Valacich, B. Adame, Y. H. Lee, B. Lane, C. Piercy, D. Wilson, S. King, C. Vincent, and R. Scheutzler. Mitigating cognitive bias through the use of serious games: Effects of feedback. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8462 LNCS:92–105, 2014.

[18] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, pp. 473–482, 2012.

[19] A. Endert, C. Han, D. Maiti, L. House, S. C. Leman, and C. North. Observation-level Interaction with Statistical Models for Visual Analytics. In *IEEE VAST*, pp. 121–130, 2011.

[20] A. Endert, M. S. Hossain, N. Ramakrishnan, C. North, P. Fiaux, and C. Andrews. The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems*, 43(3):411–435, 2014.

[21] A. Endert, W. Ribarsky, C. Turkay, B. Wong, I. Nabney, I. D. Blanco, and F. Rossi. The state of the art in integrating machine learning into visual analytics. In *Computer Graphics Forum*. Wiley Online Library, 2017.

[22] D. C. Engelbart. Augmenting human intellect: A conceptual framework. Technical Report AFOSR-3223, Stanford Research Institute, Menlo Park, CA, October 1962.

[23] M. Feng, C. Deng, E. M. Peck, and L. Harrison. Hindsight: Encouraging exploration through direct encoding of personal interaction history. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):351–360, 2017.

[24] B. Fisher, T. M. Green, and R. Arias-Hernández. Visual analytics as a translational cognitive science. *Topics in Cognitive Science*, 3(3):609–625, 2011.

[25] C. R. Fisher, M. M. Walsh, L. M. Blaha, G. Gunzelmann, and B. Z. Veksler. Efficient parameter estimation of cognitive models for real-time performance monitoring and adaptive interfaces. In D. Reitter and F. E. Ritter, eds., *Proceedings of the 14^{th} International Conference on Cognitive Modeling (ICCM 2016)*. University Park, PA, 2016.

[26] J. M. Flach, C. R. Hale, R. R. Hoffman, G. Klein, and B. Veinott. Approaches to Cognitive Bias in Serious Games for Critical Thinking. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 56(1):272–276, 2012.

[27] B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996.

[28] S. M. Galster and E. M. Johnson. Sense-assess-augment: A taxonomy for human effectiveness. Technical Report AFRL-RH-WP-TM-2013-0002, Air Force Research Laboratory, Wright-Patterson AFB, 2013.

[29] G. Gigerenzer and H. Brighton. Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, 1(1):107–143, 2009.

[30] G. Gigerenzer and W. Gaissmaier. Heuristic decision making. *Annual Review of Psychology*, 62:451–482, 2011.

[31] S. Gladisch, H. Schumann, and C. Tominski. Navigation recommendations for exploring hierarchical graphs. In *International Symposium on Visual Computing*, pp. 36–47. Springer, 2013.

[32] D. Gotz, S. Sun, and N. Cao. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. *Proceedings of the 21st International Conference on Intelligent User Interfaces - IUI '16*, pp. 85–95, 2016.

[33] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009.

[34] D. Green and J. Swets. *Signal Detection Theory and Psychophysics*. Wiley & Sons, Inc., New York.

[35] D. M. Green, T. G. Birdsall, and W. P. Tanner Jr. Signal detection as a function of signal intensity and duration. *The Journal of the Acoustical Society of America*, 29(4):523–531, 1957.

[36] T. Green, W. Ribarsky, and B. Fisher. Visual analytics for complex concepts using a human cognition model. *2008 IEEE Symposium on Visual Analytics Science and Technology*, pp. 91–98, 2008.

[37] H. V. Henderson and P. F. Velleman. Building multiple regression models interactively. *Biometrics*, pp. 391–411, 1981.

[38] R. J. Heuer Jr. *Psychology of Intelligence Analysis*. Washington, DC, 1999.

[39] E. Horvitz. Principles of mixed-initiative user interfaces. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 159–166, 1999.

[40] T. Jankun-Kelly and K.-L. Ma. A spreadsheet interface for visualization exploration. In *Proceedings of the Conference on Visualization'00*, pp. 69–76. IEEE Computer Society Press, 2000.

[41] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[42] D. Kahneman and S. Frederick. A model of heuristic judgment. *The Cambridge Handbook of Thinking and Reasoning*, pp. 267–294, 2005.

[43] D. Keim, G. Andrienko, J. D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Lecture Notes in Computer Science (including subseries Lecture*

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 4950 LNCS, pp. 154–175, 2008.

[44] H. Kim, J. Choo, H. Park, and A. Endert. Interaxis: Steering scatterplot axes via observation-level interaction. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):131–140, 2015.

[45] G. Klein, B. Moon, and R. R. Hoffman. A macrocognitive model human-centered computing a macrocognitive model. *IEEE Intelligent Systems*, 21(5):88–92, 2006.

[46] G. J. Koop and J. G. Johnson. The response dynamics of preferential choice. *Cognitive Psychology*, 67(4):151–185, 2013.

[47] B. C. Kwon, H. Kim, E. Wall, J. Choo, H. Park, and A. Endert. Axisketcher: Interactive nonlinear axis mapping of visualizations through user drawings. *IEEE Transactions on Visualization and Computer Graphics*, 2626(c):1–1, 2016.

[48] R. D. Luce. Detection and recognition. In R. D. Luce, R. R. Bush, and E. Galanter, eds., *Handbook of Mathematical Psychology*, vol. 1, pp. 103–190. Wiley, New York, 1963.

[49] R. D. Luce. The choice axiom after twenty years. *Journal of Mathematical Psychology*, 15(3):215–233, 1977.

[50] N. A. Macmillan and C. D. Creelman. *Detection Theory: A User's Guide*. Psychology Press, 2004.

[51] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97, 1956.

[52] G. Mullinix, O. Gray, J. Colado, E. Veinott, J. Leonard, E. L. Papautsky, C. Argenta, M. Clover, S. Sickles, C. Hale, E. Whitaker, E. Castronova, P. M. Todd, T. Ross, J. Lorince, J. Hoteling, S. Mayell, R. R. Hoffman, O. Fox, and J. Flach. Heuristica: Decision a serious game for improving decision making. *2013 IEEE International Games Innovation Conference (IGIC)*, pp. 250–255, 2013.

[53] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.

[54] C. North, R. May, R. Chang, B. Pike, A. Endert, G. A. Fink, and W. Dou. Analytic provenance: Process + interaction + insight. *29th Annual CHI Conference on Human Factors in Computing Systems, CHI 2011*, pp. 33–36, 2011.

[55] R. E. Patterson, L. M. Blaha, G. G. Grinstein, K. K. Liggett, D. E. Kaveney, K. C. Sheldon, P. R. Havig, and J. A. Moore. A human cognition framework for information visualization. *Computers & Graphics*, 42:42–58, 2014.

[56] W. A. Pike, J. Stasko, R. Chang, and T. A. O'Connell. The science of interaction. *Information Visualization*, 8(4):263–274, 2009.

[57] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of International Conference on Intelligence Analysis*, 2005:2–4, 2005.

[58] T. J. Pleskac. *Decision and Choice: Luce's Choice Axiom*, pp. 895–900. Elsevier, Oxford, 2015.

[59] M. Pohl, M. Smuc, and E. Mayr. The User Puzzle – Explaining the Interaction with Visual Analytics Systems. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2908–2916, 2012.

[60] P. Provoost. Jerzy. `https://github.com/pieterprovoost/jerzy`, 2017.

[61] M. Pusara and C. E. Brodley. User re-authentication via mouse movements. *Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security VizSECDMSEC 04*, pp. 1–8, 2004.

[62] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen. Characterizing provenance in visualization and data analysis: An organizational framework of provenance types and purposes. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):31–40, 2016.

[63] R. Ratcliff and P. L. Smith. A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, 111(2):333–367, 2004.

[64] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, 2016.

[65] 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.

[66] A. Sarvghad, M. Tory, and N. Mahyar. Visualizing dimension coverage to support exploratory analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):21–30, 2017.

[67] J.-H. Song and K. Nakayama. Hidden cognitive states revealed in choice reaching tasks. *Trends in Cognitive Sciences*, 13(8):360–366, 2009.

[68] M. J. Spivey and R. Dale. Continuous dynamics in real-time cognition. *Current Directions in Psychological Science*, 15(5):207–211, 2006.

[69] R. B. Stacey. A report on the erroneous fingerprint individualization in the madrid train bombing case. *Journal of Forensic Identification*, 54(6):706–718, 2004.

[70] C. Symborski, M. Barton, M. Quinn, K. S. Kassam, C. Symborski, M. Barton, and M. Quinn. Missing: A serious game for the mitigation of cognitive biases. *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC) 2014*, pp. 1–13, 2014.

[71] J. J. Thomas and K. A. Cook. Visualization viewpoints: A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[72] A. Tversky and D. Kahneman. Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, 5(2):207–232, 1973.

[73] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185:1124–1131, 1974.

[74] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211:453–458, 1985.

[75] E. Wall, L. M. Blaha, C. L. Paul, K. Cook, and A. Endert. Four perspectives on human bias in visual analytics. *DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations*, 2017. To appear.

[76] E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert. Podium: Ranking data using mixed-initiative visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 2017. To appear.

[77] P. C. Wason. On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140, 1960.

[78] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, 2007.

[79] W. Wright, D. Schroh, P. Proulx, A. Skaburskis, and B. Cort. The sandbox for analysis: concepts and methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 801–810. ACM, 2006.

[80] J. S. Yi, Y. ah Kang, J. T. Stasko, and J. A. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1224–1231, 2007.