A Formative Study of Interactive Bias Metrics in Visual Analytics Using Anchoring Bias

Emily Wall, Leslie Blaha, Celeste Paul, Alex Endert

A Formative Study of Interactive Bias Metrics in Visual Analytics Using Anchoring Bias

Emily Wall[1], Leslie Blaha[2], Celeste Paul[3], and Alex Endert[1]

[1] Georgia Tech, Atlanta, GA
{emilywall,endert}@gatech.edu
[2] Air Force Research Laboratory, Pittsburgh, PA
leslie.blaha@us.af.mil
[3] U.S. Department of Defense, Washington, D.C.
clpaul@tycho.ncsc.mil

Abstract. Interaction is the cornerstone of how people perform tasks and gain insight in visual analytics. However, people's inherent cognitive biases impact their behavior and decision making *during* their interactive visual analytic process. Understanding how bias impacts the visual analytic process, how it can be measured, and how its negative effects can be mitigated is a complex problem space. Nonetheless, recent work has begun to approach this problem by proposing theoretical computational metrics that are applied to user interaction sequences to measure bias in real-time. In this paper, we implement and apply these computational metrics in the context of anchoring bias. We present the results of a formative study examining how the metrics can capture anchoring bias in real-time during a visual analytic task. We present lessons learned in the form of considerations for applying the metrics in a visual analytic tool. Our findings suggest that these computational metrics are a promising approach for characterizing bias in users' interactive behaviors.

Keywords: Cognitive bias · Anchoring bias · Visual analytics.

1 Introduction

Human-in-the-loop approaches to data analysis combine complementary strengths of humans and computers. In visual data analysis, people leverage cognitive and perceptual systems to think about data by analyzing the views created. However, cognitive science tells us that people are inherently biased [31]. At times, biases act as mental shortcuts and help people analyze data quickly [15]. Yet there are situations where biases may lead to suboptimal analysis processes or decisions. **Anchoring bias**, for example, describes the tendency for people to rely too heavily on initial information when making a decision [12]. In the analytic process, this tendency leads people to preferentially weight some information and neglect other information, often leading to poorly informed decisions.

The impact of bias on decision making can be further compounded in mixed-initiative visual analytic approaches. Mixed-initiative visual analytics systems

leverage adaptive computational models that learn from and adjust to user feedback [19]. These models incorporate latent knowledge about the data or the domain from users through interactions. However, what if mixed-initiative systems learn from biased behaviors or even amplify the users' biases [13]?

In cognitive science, bias is typically measured by analyzing decisions people make during controlled laboratory experiments (e.g., [12,24,32]). It is understood that bias can influence perceptual judgments, memory recall, and deliberative choice making, each of which are involved in visual analytics [25]. In the context of data visualization and visual analytics, researchers have begun to characterize bias from analysis of perceptual judgments [6,7,35] or interaction data [3], where a user's behavior with an interactive tool is treated as a proxy for their cognitive state. All of these works, however, rely on post-hoc analysis of user data. While informative to the ways visualization design can influence the severity of bias, waiting until a task is completed does not allow for online intervention by systems prior to a potentially erroneous decision.

Enabling mixed-initiative systems to adapt to or mitigate cognitive biases requires an understanding of bias in real-time, during the analysis process [16]. Recently, we introduced theoretical metrics to quantify bias from user interactions in real-time during the visual data analysis process [36]. The metrics focus on characterizing *human* bias rather than other forms of bias that may be present in the analysis process (i.e., bias in analytic models, data sampling, etc.). The metrics track user interactions with the visualization, data, and analytic models in the system to create a quantitative representation of analytic provenance. The theoretical formulation in [36], however, relies on assumptions untested on actual user data, leaving many open questions regarding how to implement and apply the theory in a visual analytic tool.

In this paper, we explore how to bring the theoretical metrics into practice; specifically: **how to incorporate the interactive bias metrics into a visual analytic tool**. To do so, we implemented the metrics in a tool and conducted a formative study to examine how bias can be observed in users' interactions through the lens of the bias metrics. Our goal is to leverage a well-known and highly studied form of bias (anchoring [12,14]) to influence participants' analysis processes in a controlled way, to study the metrics under predictably biased behavior patterns. Our analysis suggests anchoring bias can be observed in users' interactive behavior through the lens of the bias metrics. The primary contributions of this paper include (1) guidelines for applying the bias metrics in visual analytic systems (Section 6), and (2) results of a formative study showing how the metrics can be used to capture anchoring bias (Section 5).

2 Related Work

Bias in Cognitive Sciences. Bias is a concept that has been widely studied in cognitive science. Cognitive bias refers to subconscious errors or inefficiencies resulting from the use of heuristics for decision making [20, 21, 31]. There are dozens of these types of errors that commonly impact decision making, and

specifically data analysis and sensemaking [18]. A prominent example is confirmation bias, the tendency to search for and rely on evidence that confirms an existing belief [24,38]. In this paper, we focus on anchoring bias, defined earlier.

Framing describes the manner in which a choice is presented to people, including the language used, the context provided, and the nature of the information displayed [32,33]. For example, a positive framing of a medical treatment risk would present probability of lives saved; a negative framing presents the same information in terms of lives lost. Framing has been found to strongly shape decision making [30]. The way that information or task goals are introduced to people has a strong impact on how they will conduct their analyses. Thus in our formative study, described later, we leverage task framing to induce anchoring bias in participants. Doing so allows us to evaluate how anchoring bias manifests in user interaction patterns for a visual data exploration and classification task.

Bias in Visual Analytics. The topic of bias in visual analytics has recently garnered increasing attention. Gotz et al. [16] addressed the issue of selection bias in examining healthcare data. They proposed a way to quantify how subsets of data may be unintentionally biased due to correlated attributes in a filtered dataset. Dimara et al. [7] examined the attraction effect in information visualization, the phenomenon where a person's decision between two alternatives is altered by an irrelevant third option. They observed that this bias is present in the use of data visualizations [7] and can be mitigated by altering the framing of the task [6]. Other recent work has begun to organize and formalize the types of bias relevant in the visualization and visual analytic domains [5,8,34,37].

Perhaps most similar to our work is Cho et al. [3] who replicated effects of anchoring bias in a visual analytic tool. In their study, participants were tasked with predicting protest events by analyzing Twitter data. They elicited anchoring bias in participants through priming, then measured reliance on particular views in a multi-view system through post-experiment metrics like total proportion of time in each view. We similarly aim to show over-reliance on some visual information sources, but we will instead quantify the behavioral effects of anchoring bias through the bias metrics [36].

3 Bias Metrics

In this section, we review our prior work defining the theoretical bias metrics, which is the emphasis of the analysis for the present study. The bias metrics utilize user interactions in a visual analytic tool as input. User interaction is the means by which users express their intent to the system [26,28]. User interaction has been shown to have the power to support steering analytic models [1, 10, 11, 22], inferring a user's personality traits [2], reasoning about their analytic methods and strategies [9], and understanding the generation of insights [17]. Thus, interaction can be thought of as a proxy, although lossy and approximate, for capturing a user's cognitive state. While the design of interactive behaviors in visual analytic tools may not precisely capture a user's state of mind, it can nonetheless provide coarse information about a user's sensemaking process.

Metric	Description	Example Behavior
Data Point Coverage (DPC)	measures *how much* of the dataset the user has interacted with	user interacted with only 3 of 100 players
Data Point Distribution (DPD)	measures *how evenly* the user is focusing their interactions across the dataset	user interacted with some data points dozens of times while ignoring others
Attribute Coverage (AC)	measures the *range of an attribute's values* explored by the user's interactions	user interacted with only players over 84 inches tall, when height ranges from 67 to 88 inches
Attribute Distribution (AD)	measures the *difference in the distribution* of the user's interactions to the distribution of a particular attribute in the dataset	user interacted with a uniform sample of data while the attribute follows a normal curve
Attribute Weight Coverage (AWC)	measures the *range of weights* for a particular attribute explored by the user's interactions	user sets weight values between 0–0.2, ignoring weight values less than 0 and greater than 0.2
Attribute Weight Distribution (AWD)	measures the *difference in the distribution* of the user-defined weights for an attribute to a baseline of unbiased information weighting	user weights follow an exponential distribution, with higher probability for low weight values than high attribute weights

Table 1: Metrics used in this study. Each metric computes a specific behavior which can be analyzed to detect bias.

We operationally define bias as patterns of interaction that reflect a systematic deviation from unbiased behavior consistent with a cognitive bias. The metrics are computed on logged interactions with a visualization to determine levels of bias with respect to different facets of the data. Each metric computation results in a value between 0 and 1, where 0 represents low bias and 1 represents high bias. Over time, we obtain a sequence of [0,1] metric values for each facet representing the user's level of bias throughout their analytic process. Rather than analyzing the accuracy or appropriateness of a decision *after* the decision is made, the metrics provide an interaction-by-interaction bias measurement.

The metrics compute bias with respect to *data points*, *attributes*, and *attribute weights* within the dataset and visual analytic model. For example, if a user is examining a dataset of basketball players, the bias metrics are designed to quantify a user's focus on specific players (data points), stats about players like height or free-throw percentage (attributes), and the way that the user places relative importance of those stats in analytic models (attribute weights). For our purposes, attribute weights fall in the range $[-1, 1]$ and are used to quantify the relative importance of each data attribute [22]. The attribute and attribute weight metrics are computed separately for each attribute in the dataset.

For each concept of data points, attributes, and attribute weights, there are metrics representing *coverage* and *distribution*. Coverage quantifies the proportion of elements that have been interacted with. Distribution, on the other hand,

compares the user's (potentially repeated) interactions to the underlying distribution of the data. For example, if the user performs many interactions with only a handful of basketball players, the data point coverage bias value will be closer to 1. This indicates an incomplete sampling of the data points. Similarly, if the user focuses primarily on Point Guards, for example, the distribution of interactions may significantly differ from the distribution of the player positions in the full dataset; the computed attribute distribution bias will be higher, indicating a sampling of the set dissimilar to the underlying data.

Table 1 summarizes the bias metrics. Each metric compares the user's sequence of interactions to a baseline of "unbiased" behavior. Our current baseline for unbiased behavior makes a simple assumption that all data points, attributes, or attribute weights will be interacted with in a uniform pattern. Hence, in the current formulation, we utilize a uniform distribution as the baseline for the data point and attribute weight metrics. We utilize the true underlying distribution of the attributes of the data in the attribute distribution metrics, assuming unbiased interactions will closely match the underlying distributions.

Our initial formulation of the metrics [36] was theoretical and relied on untested assumptions (e.g., about which interactions to compute on). In this work, we conduct a formative study to inform the implementation of the metrics in real visual analytic systems and to demonstrate how the metrics can be used to quantify instances of anchoring bias.

4 Methodology

We conducted a formative study to explore the implementation of the bias metrics and the ways they capture anchoring bias in real-time through users' interactive behavior. The purpose of this study is two-fold: (1) serve as a formative approach to implementing and applying the bias metrics, and (2) understand if the bias metrics can characterize participants who are exhibiting anchoring bias toward different data attributes. To test the hypothesis that the metrics can capture bias in real-time, we manipulated task framing to elicit predictably biased behaviors from participants and examined the ability of the metrics to detect patterns consistent with anchoring bias. Participants in the study were tasked with categorizing a dataset of basketball players. Using the visual analytics tool InterAxis [22] (Fig. 1), users were instructed to examine all of the available data to label 100 anonymized basketball players according to one of five positions. We deliberately encouraged participants to anchor on different data attributes (see Table 2) by randomly assigning them to a framing condition; each condition described the five positions using different attributes.

InterAxis. Participants used a scatterplot-based visual analytics tool to categorize basketball players by their position (Fig. 1). Pilot studies led us to modify the InterAxis interface from its presentation in [22] and [36] for the present study. Changes include: the y-axis custom axis options were removed; data point colors were changed to reflect participants' labels; options for saving the plot settings were removed; and experiment control options (e.g., position labels, Continue

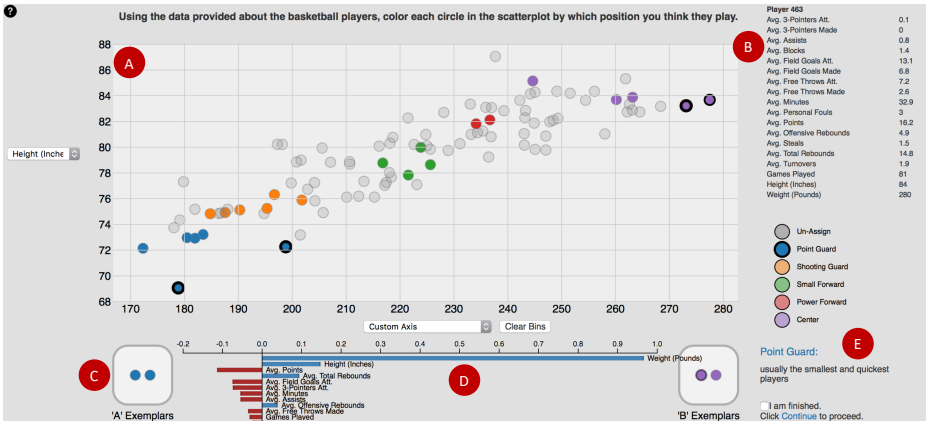

Fig. 1: A modified version of the system InterAxis [22], the interface used by participants to complete the task of categorizing basketball players.

button) were added. The data from the pilot was only for testing and feedback on our protocol and are not included in the results.

The primary view in InterAxis is a scatterplot, where each of 100 basketball players is represented by a circle (Fig. 1A). Hovering on a circle reveals details about that player (Fig. 1B). Data points can be dragged from the scatterplot into the Exemplar bins on either side of the x-axis (Fig. 1C). The system, in response, will compute a custom axis using a linear dimension reduction technique. The result is a set of attribute weights that represents the differences between the A and B Exemplar bins. The attribute weights are visualized as bars along the axis (Fig. 1D). The bars can also be interacted with by click-and-drag to directly manipulate the weights that make up the custom axis. Participants can read a description of each position by clicking on the colored circles on the right side (Fig. 1E). If they select a position on the right, the user can then label players as the selected position by clicking on the points in the scatterplot.

We selected InterAxis due to the system's highly interactive nature—to encourage users to explore and interact with the data, because the bias metrics ultimately rely on user interactions. InterAxis allows users to browse data points and attributes, and leverage an analytic model consisting of weighted attributes to project the data. This allows us to use the full set of bias metrics.

Analytic Task & Framing Conditions. Studies of anchoring bias within the cognitive science community rely on highly controlled experiments to isolate a cognitive phenomenon. However, in visual data analysis, cognitive processes are often much more complex than can be captured from such experiments. Pirolli and Card describe the sensemaking process as a series of iterative tasks involving searching for information, schematizing, presentation, and so on [27]. Hence we sought a task with enough complexity to simulate decision making within a realistic analysis scenario while maintaining tractable experimental conditions.

Position	**Size Condition**	**Role Condition**
Center (C)	Typically the *largest* players on the team	Responsible for protecting the basket, resulting in lots of *blocks*
Power Forward (PF)	Typically of *medium-large* size and stature	Typically spends most time near the basket, resulting in lots of *rebounds*
Small Forward (SF)	Typically of *medium* size and stature	Typically a strong defender with lots of *steals*
Shooting Guard (SG)	Typically of *small-medium* size and stature	Typically attempts many shots, especially long-ranged shots (i.e., *3-pointers*)
Point Guard (PG)	Usually the *smallest* and quickest players	Skilled at passing and dribbling; primarily responsible for distributing the ball to other players resulting in many *assists*

Table 2: Position descriptions used in the two framing conditions. We expected *Size* condition participants to rely more heavily on size-related attributes (i.e., Height and Weight). We expected *Role* condition participants to rely more heavily on the role-related attributes called out in the descriptions.

There are many tasks associated with performing data analysis in a visual analytic tool, such as ranking, clustering, or categorizing data [11,29]. What bias looks like can be quite different across these tasks; for this study we narrowed our scope to focus on categorization-based analysis. We found through pilot studies that categorizing basketball players was a sufficiently challenging task that led users to interact with the visual analytics tool for approximately 30 minutes. This provided a balance of task complexity and study tractability.

Participants were instructed to categorize a set of 100 basketball players by their positions by analyzing all of their stats using the InterAxis visual analytic tool [22] in Fig. 1. We used a dataset of professional (NBA) basketball player[4] statistics with names and team affiliations removed. After filtering out less active players (whose statistical attributes were too small to be informative), we randomly selected 20 players for each of five positions: Center (C), Power Forward (PF), Small Forward (SF), Shooting Guard (SG), and Point Guard (PG). Each player had data for the following stats: 3-Pointers Attempted, 3-Pointers Made, Assists, Blocks, Field Goals Attempted, Field Goals Made, Free Throws Attempted, Free Throws Made, Minutes, Personal Fouls, Points, Offensive Rebounds, Steals, Total Rebounds, Turnovers, Games Played, Height, and Weight.

Participants were assigned to one of two conditions. The two conditions differed in the descriptions provided for the five positions. In the *Size* condition, the descriptions are based on physical attributes (Height and Weight). In the *Role* condition, positions were described with respect to their typical role on the court and performance statistics. These descriptions were based on analysis of the distributions of attributes for each position as well as position descriptions

[4] `http://stats.nba.com/`

recognized by the NBA[5]. Table 2 shows the text used to describe the positions in each condition, which was available throughout the task (Fig. 1E). Similar to other experiments utilizing task framing, we described the positions from two different perspectives (sets of attributes) between the two conditions. Participants in each condition should then anchor on the attributes used in the framing to which they were assigned. We emphasize that, while the player position descriptions were framed differently, participants in both conditions were instructed to utilize all of the data to make their decisions.

Generally, anchoring bias describes an over-reliance on some information, often to the neglect of other relevant information about a decision. We operationally define interaction-based biases as increased interaction with limited subsets of data, attributes, or attribute weights over a more evenly or uniformly distributed pattern of interactions. Anchoring specifically, then, will be observed if there is biased interactions with information to which the participant has been cued and is relying on more than other information to make analytic decisions.

Verifying the Task Framing Effects. To see how the bias metrics quantify anchoring bias, we first analyzed how framing impacted user behaviors. We compared the frequencies of attributes selected for the scatterplot axes between the two framing conditions. We predicted that participants in the Size framing condition would select the Height or Weight attributes on the axes more than participants in the Role framing condition. Likewise, we predicted that participants in the Role framing condition would select the other attributes used in the position descriptions (Blocks, Rebounds, Steals, 3-Pointers, or Assists; see Table 2) on the axes more than participants in the Size framing condition.

Fig. 2 shows the results of this analysis. Each boxplot shows the number of times the given attribute was selected on the axis for participants in the Role condition (left) and the Size condition (right). Larger separation of mean and

[5] `http://www.nba.com/canada/Basketball_U_Players_and_Posi-Canada_Generic_Article-18037.html`

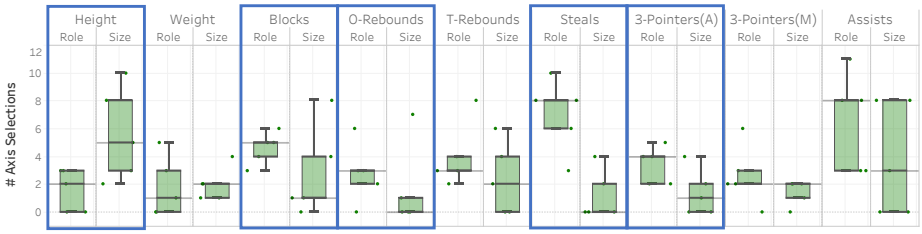

Fig. 2: Boxplots for number of attribute interaction via InterAxis axis manipulation. The thick middle line indicates the median; the box delineates the inner quartiles, and the whisker bars give the outer quartiles. Green dots indicate the sum of observations for each participant (not outliers). The blue boxes indicate attributes for which a substantial difference is seen between the two conditions.

quartile values suggests that the framing condition impacted the frequency of interaction with a given attribute, while highly overlapping boxplots suggest little or no difference for that attribute between the framing conditions. The boxplots reveal that some attribute axis selections show clear differences supporting our predictions (outlined in blue), while others exhibit little difference between conditions (e.g., Weight, Total Rebounds, 3-Pointers Made, and Assists). Participants in the Size condition interacted more frequently with Height than the Role condition participants. And participants in the Role condition interacted more frequently with performance-related attributes (Blocks, Offensive Rebounds, Steals, 3-Pointers Attempted) than participants in the Size condition. These results suggest that the participants from the two conditions anchored on the attributes described in the respective framing conditions, as predicted. These results confirm that the Role and Size conditions influenced the overall categorization behaviors in ways consistent with our intended manipulations.

Participants. Ten participants (4 female, mean age 25.5 ± 2.7 years) were recruited from a large university. Nine participants had experience playing basketball, and six participants watched at least a few (NCAA, NBA, WNBA) games per season (self-reported). The one participant who never played basketball watches it regularly. All participants were moderately familiar with information visualization, based on Likert ratings provided in a background survey. Participants were randomly assigned to either the Size or Role condition.

Procedure. Participants began with informed consent, completed a demographic questionnaire, and were shown a 5-minute video describing the task and demonstrating use of the InterAxis tool to complete the task. The demonstration used different position descriptions than the study. Participants then completed the main task, using InterAxis to categorize 100 basketball players into one of five positions. There were no time limits for completing the task. After completing the task, participants completed a post-study questionnaire about their experience and were compensated with a $10 Starbucks gift card.

A moderator observed participants' interactions during the task. Participants were encouraged to ask questions as needed regarding the interface, the underlying algorithmic transformations, or the meaning of an attribute. The moderator did not reveal information about the underlying distribution of positions in the dataset or additional attributes that might be used to help categorize players.

Timestamped logs of the users' interactions were automatically recorded, including interactions with data points (labeling, hovering to reveal details, and dragging to axis bins), interactions with axes (selecting a new attribute for an axis, dragging to adjust attribute weights, and recomputing attribute weights based on interactions with the bins), and interactions with position descriptions (clicking to reveal a description and double clicking to de-select a position description). The interaction logs serve as the input data for the bias metrics.

5 Analysis and Results

We analyzed the user study data with the high-level goal of understanding how to use the bias metrics to quantify and characterize *anchoring bias*. The bias metrics provide us with the ability to characterize a user's analytic process in real-time by quantifying aspects of their interaction patterns in which they may be exhibiting bias. In particular, we analyzed the bias metrics from the granularity of (1) the sequences of $[0, 1]$ metric values over time, and (2) where in the distribution of the data user interactions deviated from expected behavior. From the perspective of the bias metrics, participants subject to anchoring bias could be observed to have (1) higher $[0, 1]$ bias metric values for the anchored attributes, and/or (2) instances during the analytic process where they interact more heavily with part of the distribution of the anchored attribute.

To analyze if the metrics can capture bias, we used the collected interaction logs to simulate the real-time computation of the bias metrics after each user's session to avoid influencing the analysis process[6]. We note that the bias metrics created 74 unique time series per participant (Data Point Coverage + Data Point Distribution + 18 attributes x {Attribute Coverage, Attribute Distribution, Attribute Weight Coverage, Attribute Weight Distribution}). In the scope of this work, we narrow the focus of our discussion to only attributes that were referenced in the framing of position descriptions (Table 2). We discuss a few selected examples of findings from the computed bias metrics. Visualizations of all metrics can be found in the supplemental materials.[7]

Participants' accuracy for categorizing players averaged 53% ($SD = 18\%$) over the mean duration 33.6 minutes ($SD = 14$ min). Some interactions were filtered out to reduce noise in the bias metric computations. According to Newell's time scale of human action [23], deliberate cognitive actions are on the order of 100 ms+. Because hovering in the interface shows a data point's details, particularly short hovers were likely not intentional interactions. Thus, hovers with duration less than 100 ms were removed as likely "incidental" interactions performed unintentionally while navigating the cursor to a different part of the interface. Participants performed an average of 1647 interactions ($SD = 710$), which filtered down to an average of 791 non-incidental interactions ($SD = 300$). For additional discussion on which interactions are included in the bias metric computations, see Section 6.

Metrics over Interaction Sequences. Computed over time, the bias metrics produce a sequence of $[0, 1]$ values quantifying the level of bias throughout the analysis process, which can be visualized as a time series. We hypothesized that the attributes explicitly described in each condition (Height and Weight for the *Size* condition; Blocks, Rebounds, Steals, 3-Pointers, and Assists for the *Role*

[6] Note that while the ultimate goal of the metrics is online interpretation and mixed-initiative adaptation, the present work collected full interaction sequences of metrics for post-hoc analysis, to ensure the metrics can capture bias and to elucidate how to effectively put the metrics into practice.

[7] `https://github.com/gtvalab/bias-framing`

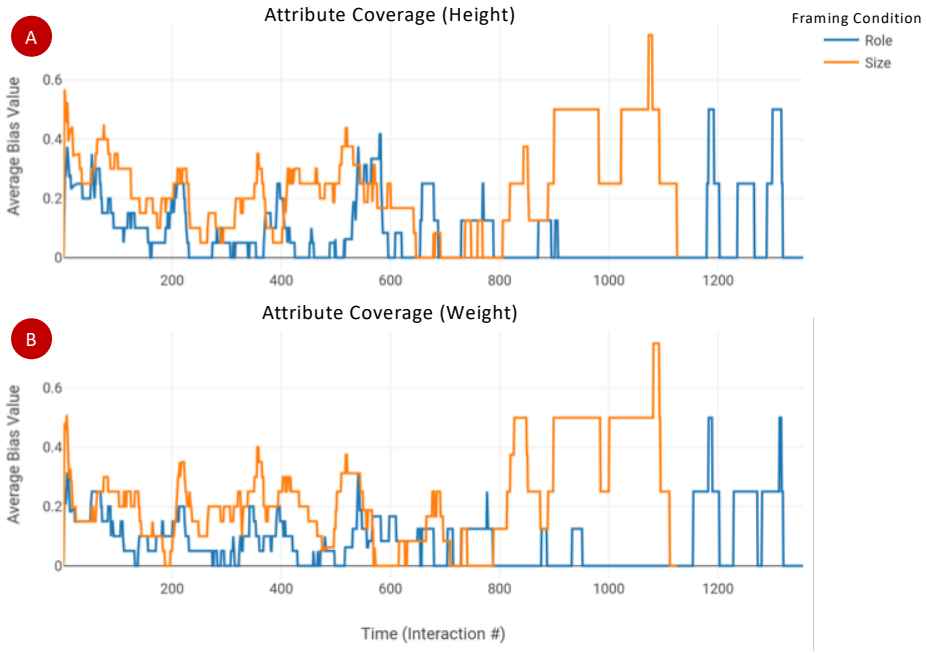

Fig. 3: A visualization of the average Attribute Coverage (AC) metric for the attributes (A) Height and (B) Weight. Size condition participants (orange) tended to have higher AC bias for both Height and Weight than Role condition participants (blue), consistent with our predictions.

condition) will have higher metric values in the associated condition than in the other. For example, we expected the time series of Attribute Distribution values for Assists to be higher for Role condition participants than for Size condition participants. To evaluate this hypothesis, we visualized all 74 metrics' time series.

Fig. 3 shows the Attribute Coverage (AC) metric for (A) the Height attribute and for (B) the Weight attribute. The blue line represents the AC metric time series averaged over all Role condition participants. The orange line represents the AC metric time series averaged over all Size condition participants. Fig. 3 shows that Size condition participants tended to have higher peaks (metric values closer to 1) and longer peaks (over greater spans of time) in the AC bias metric for the Height and Weight attributes than Role condition participants, consistent with the framing condition predictions.

We confirm this trend by comparing bias values averaged over the full interaction sequence for participants in each condition. Size condition participants had an average value of $M_{\text{Size}} = 0.2211$ ($SD = 0.066$) for the *Height* AC metric compared to $M_{\text{Role}} = 0.0952$ ($SD = 0.016$). Similarly, for the Weight AC metric, the Size condition participants had an average value of $M_{\text{Size}} = 0.2120$ ($SD = 0.098$) compared to the Role condition participants $M_{\text{Role}} = 0.0849$ ($SD = 0.042$).

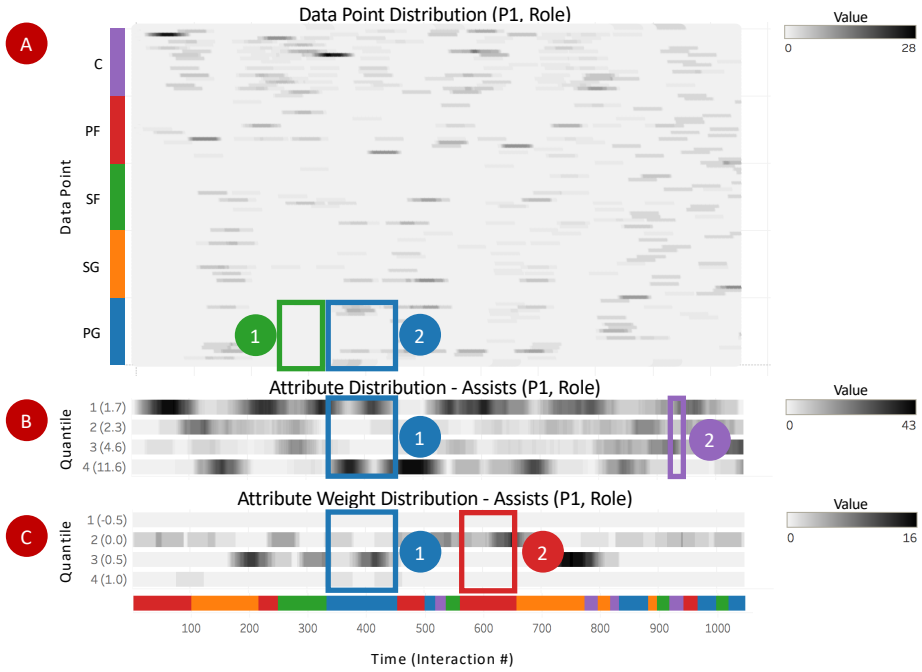

Fig. 4: Visualizations of three of the bias metrics for a Role condition participant: (A) the DPD metric, (B) the AD metric for Assists, and (C) the AWD metric for Assists. While labeling PGs (blue boxes), (A) the participant showed more bias toward PGs than while labeling other positions (SF; green). (B) The participant also showed greater bias toward the high end of the distribution for Assists while labeling PGs than other positions (C; purple), and (C) weighted Assists more heavily while labeling PGs than while labeling other positions (PF; red).

This evidence supports our hypothesis; however, not all metrics show a discernible difference in $[0, 1]$ values between the two conditions. One potential explanation for inconsistent effects is the level of granularity in the analysis. The bias metric values indicate the degree of bias; however, they do not indicate the source of the bias. For example, a user focusing on particularly tall players might have the same metric value as a user focusing mostly on short players. That is, simply knowing the metric value informs us of a bias; however, the number itself does not differentiate the source within the data distributions. Next, we address this by examining the underlying coverage or distribution.

Coverage and Distribution of Bias Metric Values. To compute the metric values, an intermediate step is to break down the user's interactions with data points, attributes, and attribute weights into quantiles and distributions to see how they deviate from unbiased interactive behavior. One way to show the framing effects on user interaction patterns is to compare the metrics broken down into components of coverage and distribution rather than just summative $[0, 1]$

values. In this analysis we visualized the breakdown of coverage and distribution metrics using a heatmap. Note that because the bias metrics are computed independently for each participant, the color scale used to shade the cells is likewise normalized for each participant. The scales are defined in each plot.

Fig. 4 illustrates the metrics Data Point Distribution (DPD), Attribute Distribution (AD) for Assists, and Attribute Weight Distribution (AWD) for Assists for one Role condition participant. All of the metrics share a common x-axis of time, captured as the interaction number. The colored bars beneath the time represent the type of position being labeled during that time period (blue = Point Guard, orange = Shooting Guard, green = Small Forward, red = Power Forward, and purple = Center). The shading in a particular (x, y) position represents the count of interactions that fall within the given bin at the given point in time; darker shades represent a greater number of interactions.

In Fig. 4(A), the y-axis shows a row for each data point to illustrate DPD. This type of plot can visually indicate the user's bias toward (interaction with) particular players based on their interactive behavior during different time periods. For example, the DPD metric shows more bias toward players who are Point Guards (PGs) while attempting to label PGs (Fig. 4(A,2)) than while attempting to label Small Forwards (Fig. 4(A,1)), consistent with correct categorizations.

In Fig. 4(B), the y-axis illustrates the distribution of attribute values (AD) broken down into four quantiles. The AD metric for Average Assists shows a stronger bias toward players with a high number of Average Assists while labeling PGs (Fig. 4(B,1)) than while labeling Centers (Fig. 4(B,2)), consistent with Role framing. In Fig. 4(C), the y-axis illustrates the breakdown of attribute weight ranges (AWD) into four quantiles. The AWD metric for Average Assists indicates a bias toward higher weighting of the attribute while labeling PGs (Fig. 4(C,1)) than while labeling Power Forwards (Fig. 4(C,2)). The Role condition PG description is intended to influence participants to anchor on the Average Assists attribute. Hence, Figures 4(B) and 4(C) visually capture a user's anchoring bias toward an attribute.

Fig. 5(A) visually compares Attribute Weight Coverage (AWC) for Height between two users from different conditions. The position descriptions used in the Size condition were designed to anchor participants on Height and Weight attributes. The Size condition participant (top) showed greater coverage of the range of attribute weights (as shown by the black bars in all four quartiles) and spent more time with a high, positive weight applied to the Height attribute. Comparatively, the Role condition participant (bottom) covered less of the range of possible attribute weights and spent the vast majority of their analysis with a low weight applied to the Height attribute. We can quantify this difference using the L metric from recurrence quantification analysis [4]. L gives the average length of diagonal segments in a recurrence analysis. Applied to the metric state, larger L values reflect staying in a state longer while smaller L values reflect switching more frequently between quartiles. For the Size participant (top), $L = 14.9$ indicating more switching, and $L = 229.8$ for the Role participant (bot-

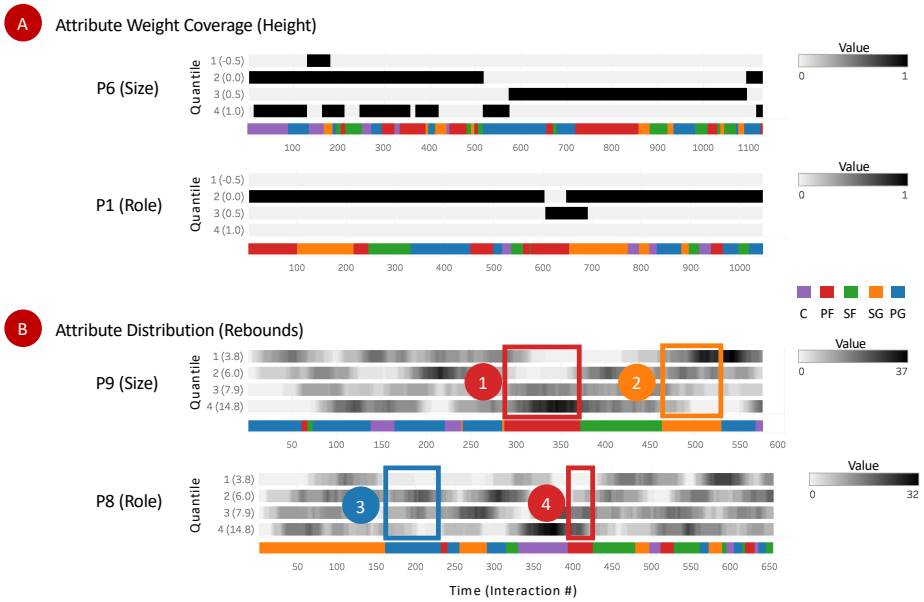

Fig. 5: (A) Visualization of the AWC metric for Height. The Size condition participant (top) showed greater *coverage* of the range of Height attribute weights than the Role condition participant (bottom). (B) Visualization of the AD metric for Total Rebounds. Participants focused more on upper parts of the Rebounds distribution while labeling PFs (red boxes) than other positions.

tom), reflecting a very long time in a single quartile which is seen in Fig. 5(A). Heatmaps for all metrics and all participants are in the supplemental material.

Similarly, Fig. 5(B) shows how Attribute Distribution (AD) for Average Total Rebounds compares for one Size condition participant (top) and one Role condition participant (bottom). Role condition participants were told that Power Forwards (PF) typically have a high number of Rebounds. While labeling PFs, both the Role condition participant (4) and the Size condition participant (1) showed interactions with greater focus toward the upper parts of the distribution (Q3 and Q4). Similarly, both the Role condition participant (3) and the Size condition participant (2) interacted with lower parts of the distribution (Q1 and Q2) while labeling other positions. While the Size condition participants were not explicitly told about the importance of Rebounds for PFs, there is a correlation between the Weight of PFs and Rebounds ($r = 0.414$, $p = 0.069$), which could explain the similar patterns across the two conditions. Looking at the distribution patterns, we see both participants spent some time in all quartiles for the AD metrics. For the participant in the Size condition (top), $L = 21.2$, and for the Role condition participant (bottom), $L = 16.8$. The participants had similar L magnitudes, but the relatively larger value for Size condition participant indicates less switching between quartiles.

In summary, the task framing impacted which attributes people rely on in their interactive analysis process. These visualizations collectively demonstrate the promise that the real-time interaction-based bias metrics can detect anchoring bias toward particular attributes of the data.

6 Applying the Bias Metrics

This study constitutes the first application of the real-time bias metrics in [36], and explores how to analyze them to capture a specific type of cognitive bias. Consequently, we identified a number of challenges to consider and extracted several lessons learned in moving from theory to implementation in measuring bias through interactions. Additional sources of variability in user activities arise in the real-world analysis process that challenge theoretical assumptions. Implementation choices made early in the design process may need to adjust or adapt on the fly to accommodate unforeseen activities by the experimental participants. In this section, we present guidelines and considerations for integrating and applying the bias metrics, including a discussion on interaction design for the bias metrics, which interactions should be included in the bias metric computations, and how to interpret the metrics.

Designing for Measurement v. Usability. Designing a visualization system often involves understanding potential user needs, including things like ease-of-use, learnability, or analytic capabilities. These goals each necessitate particular design decisions. Incorporating interaction-based bias metrics in an interface likewise entails its own design requirements which may conflict with other goals. While incorporating bias computation in a visualization has the potential to promote better analysis behaviors, it ultimately relies on interpreting user interactions as a meaningful capture of the analytic process. Hence, the design must facilitate sufficient, meaningful, recordable interactions. In other words, the analysis process must be explicit in the interaction design of the interface.

For example, in the modification of the InterAxis [22] system for the evaluation discussed in the user study methodology section, we debated the interaction design for labeling basketball players' positions. A lasso tool could be an efficient way to label players in bulk; however, providing such a tool would make the interpretation of the aggregate interaction difficult from the perspective of the bias metrics. Further, participants would be less likely to interact with specific data points, read their individual attributes, and make a decision.

Given that the bias metrics rely on abundant interaction data, we instead decided to use single click to label data points and hover to reveal details about individual data points. This decision came at the expense of a potentially less frustrating user experience, as echoed by participants after the study. Such tradeoffs must be considered when integrating bias metrics into practical tool design. When the risk of biased analysis is low or the potential consequences are low, designers and developers may opt to focus on designing for usability. An important question to consider for future research is *if the interaction design of an*

interface does not organically produce sufficient interaction data to measure, to what extent is it acceptable to violate user experience to achieve it?

Which Interactions to Compute On. *Incidental Interactions:* The bias metrics rely on recording and computing on sequences of user interactions. Just as we must ensure that a system's interactions are designed to explicitly capture as much of the decision making process as possible, we also need a way of knowing if some of the interactions were unintentional. For example, a user may want to hover on a particular data point in the scatterplot to get details; however, due to the particular axis configuration or zoom level, the scatterplot may be overplotted. Thus, in attempting to perform a single deliberate interaction, the user might accidentally hover on several other data points along the way. These "incidental" interactions do not reflect the user's intent in any way and should thus ideally be discarded from the bias computations to remove noise. As an initial proxy for filtering out noisy incidental interactions, we ignored all hovers less than 100 ms. Some amount of noise is to be expected when leveraging user interaction as a proxy for cognitive state. However, the fidelity of models can be improved by taking care to ensure, even with rough approximations, that the interactions computed on reflect a meaningful capture of user intent.

Interaction Windowing: Wall et al.'s [36] prior work presents a formulation of metrics for characterizing bias based on user interaction histories; however, it does not inform us *when* to compute the metrics or *how many past interactions* should be computed on. In this study, we experimented with three different techniques for scoping the metric computations. Our first approach was to compute the bias metrics after every interaction and use the *full interaction history* for every computation. Next, we tried a *rolling window* of the previous n interactions around each current interaction. The window size n then introduced another variable whose value can lead to potentially very different results. We experimented with window sizes ranging from 25 to 100 previous interactions. Lastly, we tried using key *decision points*, where the bias metrics could be computed using all of the interactions that occurred since the last decision point. We computed two variations of this: (1) using each data point label as a decision point, and (2) using the activation of a position (Fig. 1E) as a decision point. Generalizing this windowing technique, however, requires that decision points be known, which may not be the case depending on the task and interface.

Each of these windowing techniques gives a slightly different perspective on the user's bias. For example, using the *full interaction history* can shed light on long-standing biases throughout the user's analytic process, while using a *rolling window* can capture more short-lived biases. Alternatively, using only the interactions between key *decision points* can be used to characterize bias in a user's interactions associated with individual decisions. As we did not know what strategies people might use, we captured short-lived biases using a *rolling window*, size $n = 50$, computed after each interaction.

Interpreting the Bias Metrics. The bias metrics are formulated such that a value $b \in [0, 1]$ is produced, where 0 indicates no bias and 1 indicates high bias

(e.g., as shown in Fig. 3). While an objective characterization of bias, the value b itself is not actionable from a user's perspective. That is, the bias value alone does not provide sufficient detail to a user to facilitate effective reflection and correction of their behavior. For example, a user might have a high bias value for the Height AD metric. This could be due to the user focusing unevenly on short players, on tall players, or on *any* part of the distribution.

To draw actionable conclusions from the bias metric values, it is important to provide additional information to the user, specifically about where in the data or the distribution the user's interactions depart from the objective expectation. In the evaluation results, we showed one potential solution, which visualizes the *coverage* and *distribution* of interactions across data points, attributes, and attribute weights as heatmaps (Figures 4-5). Combining both the [0,1] bias values along with the *coverage* and *distribution* that comprises the bias value computation might be ideal in some situations. For example, the [0,1] bias values could be used by automated techniques to select the most concerning dimension(s) in the user's behavior. Then, using the *coverage* and *distribution* information, systems can visualize the source of bias as the imbalance between the unbiased baseline behavior and the user's interactions to encourage user reflection.

7 Limitations and Future Work

Study Limitations. One limitation of the current study was the lack of consideration for visual salience as a confounding explanation for some interactive behavior. Because users could change axis configurations, zoom, and pan on the scatterplot, different emerging clusters of points or outliers might draw the user's attention. In future work, we would like to explore redefining the unbiased baselines for the metrics that account for visual salience. Other factors can also impact users' interactive behaviors, including incidental interactions, task-switching, environmental distractions, and so on. It is of general interest to improve unbiased baseline models to account for such factors.

We have focused our analysis on an exploration of within-subjects patterns in the data, toward our goal of within-user, online use of the metrics. The present data includes, on average, 74 metrics $\times$ 791 interactions per participant, in addition to overall metrics like task accuracy. While ten participants is large enough for our present formative analysis, it is too few for strong between-subjects statistical power. Because these metrics are new, we are simultaneously developing the analyses for the metrics while testing their validity and applicability. Ultimately, our goal is to determine an effective analysis pipeline to facilitate larger data collection efforts for both within and between subjects analyses.

Generalizing Tasks and Interfaces. In this study, participants were tasked with categorizing basketball players by position in a visual analytic tool. Our goal was to study the metrics' ability to quantify a psychological concept (bias) in the context of a real-world problem (using a visual analytic system for categorization). However, the study focused on a single constrained subtask of data

analysis. In reality, data analysis can be much messier with analysts examining alternative hypotheses and switching between potentially very different subtasks in diverse analytic interfaces. In future work, we would like to examine how bias materializes in other types of interfaces and analytic subtasks (e.g., ranking, clustering, etc.) as well as how these subtasks combine into more complete sensemaking. We would also like to enable handling multiple data sources, which will challenge the current definitions of the metrics. For example, handling text documents may be challenging because clicking to open the document constitutes one interaction but the time spent reading the document without any explicit interface interactions could be significant. It is important to identify meaningful ways to incorporate time on task into the metric computations.

Temporal Interaction Weighting. We discussed above how different windowing techniques impact bias metric computations. A potential improvement on these variations would be to develop a temporal weighting scheme, where all interactions are used to compute the bias metrics, and the interactions are weighted by recency. The most recent interactions would be weighted more highly than those performed early in the user's analysis process. A rigorous evaluation of windowing and interaction weighting schemes could inform the way that we account for how current analytic processes are informed by previous ones.

8 Conclusion

The visualization and visual analytics communities are becoming increasingly aware that biases, from the way data is collected, modeled, or analyzed, may negatively impact the process of visual data analysis. Specifically for interactive data exploration, a user's cognitive biases play a role in shaping the analysis process and ultimately the analytic outcome. In this paper, we focused on implementing and applying real-time bias metrics by studying how anchoring bias materializes in user interactions. We presented the results of a formative study where participants were assigned to one of two conditions for a categorization task using a visual analytic system. We captured interaction logs from their analyses and used real-time bias metrics [36] to characterize the interactions. Comparing the two conditions, we found that user interactions interpreted through bias metrics captured strategies and behaviors reflecting the manipulated anchoring bias. These encouraging results open the potential for discovering biased behavior in real-time during the analytic process, which can have broad-reaching impact on the design and implementation of visual analytic systems.

9 Acknowledgements

This research is sponsored in part by the U.S. the Department of Defense through the Pacific Northwest National Laboratory, the Siemens FutureMaker Fellowship, and NSF IIS-1813281. 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. Brown, E.T., Liu, J., Brodley, C.E., Chang, R.: Dis-Function: Learning Distance Functions Interactively. IEEE Conference on Visual Analytics Science and Technology (VAST) pp. 83–92 (2012)
2. Brown, E.T., Ottley, A., Zhao, H., Lin, Q., Souvenir, R., Endert, A., Chang, R.: Finding Waldo: Learning about users from their interactions. IEEE Transactions on Visualization and Computer Graphics **20**(12), 1663–1672 (2014)
3. Cho, I., Wesslen, R., Karduni, A., Santhanam, S., Shaikh, S., Dou, W.: IEEE Conference on Visual Analytics Science and Technology (VAST) (2017)
4. Coco, M.I., Dale, R.: Cross-recurrence quantification analysis of categorical and continuous time series: An r package. Frontiers in Psychology **5**, 510 (2014)
5. Cottam, J.A., Blaha, L.M.: Bias by default? a means for a priori interface measurement. DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations (2017)
6. Dimara, E., Bailly, G., Bezerianos, A., Franconeri, S.: Mitigating the attraction effect with visualizations. IEEE Trans. on Visualization and Computer Graphics (2018)
7. Dimara, E., Bezerianos, A., Dragicevic, P.: The attraction effect in information visualization. IEEE Trans. on Visualization and Computer Graphics **23**(1), 471–480 (2017)
8. Dimara, E., Franconeri, S., Plaisant, C., Bezerianos, A., Dragicevic, P.: A task-based taxonomy of cognitive biases for information visualization. IEEE Trans. on Visualization and Computer Graphics (2018)
9. Dou, W., Jeong, D.H., Stukes, F., Ribarsky, W., Lipford, H.R., Chang, R.: Recovering Reasoning Process From User Interactions. IEEE Computer Graphics & Applications **May/June**, 52–61 (2009)
10. Endert, A., Han, C., Maiti, D., House, L., Leman, S.C., North, C.: Observation-level Interaction with Statistical Models for Visual Analytics. In: IEEE VAST. pp. 121–130 (2011)
11. Endert, A., Ribarsky, W., Turkay, C., Wong, B., Nabney, I., Blanco, I.D., Rossi, F.: The state of the art in integrating machine learning into visual analytics. In: Computer Graphics Forum. Wiley Online Library (2017)
12. Englich, M., Mussweiler, T.: Anchoring effect. Cognitive Illusions: Intriguing Phenomena in Judgement, Thinking, and Memory p. 223 (2016)
13. Friedman, B., Nissenbaum, H.: Bias in computer systems. ACM Transactions on Information Systems (TOIS) **14**(3), 330–347 (1996)
14. Furnham, A., Boo, H.C.: A literature review of the anchoring effect. The Journal of Socio-economics **40**(1), 35–42 (2011)
15. Gigerenzer, G., Goldstein, D.G.: Reasoning the fast and frugal way: models of bounded rationality. Psychological review **103**(4), 650 (1996)
16. Gotz, D., Sun, S., Cao, N.: Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In: Proceedings of the 21st International Conference on Intelligent User Interfaces. pp. 85–95. ACM (2016)
17. Gotz, D., Zhou, M.X.: Characterizing users' visual analytic activity for insight provenance. Information Visualization **8**(1), 42–55 (2009)
18. Heuer Jr., R.J.: Psychology of Intelligence Analysis. Washington, DC (1999)
19. Horvitz, E.: Principles of Mixed-Initiative User Interfaces. Proceedings of the SIGCHI conference on Human Factors in Computing Systems (May), 159–166 (1999)

20. Kahneman, D.: Thinking, fast and slow. Macmillan (2011)
21. Kahneman, D., Frederick, S.: A model of heuristic judgment. The Cambridge Handbook of Thinking and Reasoning pp. 267–294 (2005)
22. Kim, H., Choo, J., Park, H., Endert, A.: InterAxis: Steering Scatterplot Axes via Observation-Level Interaction. IEEE Transactions on Visualization and Computer Graphics **22**(1), 131–140 (2016)
23. Newell, A.: Unified theories of cognition. Harvard University Press (1994)
24. Nickerson, R.S.: Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology **2**(2), 175–220 (1998)
25. Patterson, R.E., Blaha, L.M., Grinstein, G.G., Liggett, K.K., Kaveney, D.E., Sheldon, K.C., Havig, P.R., Moore, J.A.: A human cognition framework for information visualization. Computers & Graphics **42**, 42–58 (2014)
26. Pike, W.A., Stasko, J., Chang, R., O'Connell, T.A.: The science of interaction. Information Visualization **8**(4), 263–274 (2009)
27. Pirolli, P., Card, S.: Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Though Cognitive Task Analysis. Proceedings of the 2005 International Conference on Intelligence Analysis, McLean, VA p. 6 (2005)
28. Pohl, M., Smuc, M., Mayr, E.: The User Puzzle – Explaining the Interaction with Visual Analytics Systems. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2908–2916 (2012)
29. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: The Craft of Information Visualization, pp. 364–371. Elsevier (2003)
30. Thomas, A.K., Millar, P.R.: Reducing the framing effect in older and younger adults by encouraging analytic processing. The Journals of Gerontology Series B: Psychological Sciences and Social Sciences **2**(139) (2011)
31. Tversky, A., Kahneman, D.: Judgment under uncertainty: Heuristics and biases. Science **185**, 1124–1131 (1974)
32. Tversky, A., Kahneman, D.: The framing of decisions and the psychology of choice. Science **211**, 453–458 (1981)
33. Tversky, A., Kahneman, D.: Rational choice and the framing of decisions. Journal of Business pp. S251–S278 (1986)
34. Valdez, A.C., Ziefle, M., Sedlmair, M.: A framework for studying biases in visualization research. In: DECISIVe 2017: Dealing with Cognitive Biases in Visualisations (2017)
35. Valdez, A.C., Ziefle, M., Sedlmair, M.: Priming and anchoring effects in visualization. IEEE Transactions on Visualization & Computer Graphics (1), 584–594 (2018)
36. Wall, E., Blaha, L.M., Franklin, L., Endert, A.: Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. IEEE Conference on Visual Analytics Science and Technology (VAST) (2017)
37. Wall, E., Blaha, L.M., Paul, C.L., Cook, K., Endert, A.: Four perspectives on human bias in visual analytics. DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations (2017)
38. Wason, P.C.: On the failure to eliminate hypotheses in a conceptual task. Quarterly Journal of Experimental Psychology **12**(3), 129–140 (1960)