DETECTING AND MITIGATING HUMAN BIAS IN VISUAL ANALYTICS

A Dissertation
Presented to
The Academic Faculty

By

Emily Wall

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology

August 2020

Copyright © Emily Wall 2020

DETECTING AND MITIGATING HUMAN BIAS IN VISUAL ANALYTICS

Approved by:

Dr. Alex Endert, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. John Stasko
School of Interactive Computing
Georgia Institute of Technology

Dr. Polo Chau
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Brian Fisher
School of Interactive Arts and Technology
Simon Fraser University

Dr. Wenwen Dou
Department of Computer Science
University of North Carolina at Charlotte

Date Approved: April 14, 2020

True knowledge exists in knowing that you know nothing.

Socrates

For MA, SM, and PM, who supported me throughout this conquest.

ACKNOWLEDGEMENTS

The completion of my dissertation was made possible by an army of friends, family, and colleagues who supported me in various ways.

Thank you to my advisor, Alex, whose guidance and patience helped me to grow into a confident and independent researcher. Thank you to my committee members, John, Polo, Brian, and Wenwen, whose critical feedback strengthened and shaped my research over the last several years. A special thanks to all of my collaborators who have worked and written tirelessly toward publication deadlines: B.C., Hannah, Jaegul, Haesun, Alex, Karthik, Zehua, Niklas, Leslie, Lyndsey, Celeste, Kris, Subhajit, Ravish, Bharath, Eli, Meeshu, Laura, Kristin, Michael, John, Souroush, Gonzalo, Arup, Kuhu, Andrew, Arpit, and Jamal.

I am grateful for several mentors who were particularly inspirational to me – who took chances on me, who guided me through scary and new opportunities, and who provided a constant source of support whenever I needed them. Thank you Mark, Lyndsey, Leslie, and Gonzalo. Thank you to my labmates Chad, Ramik, Yi, Alex, Arjun, John, Terrance, Hayeong, Tim, Fred, Julia, Bahador, Hannah, Subhajit, Arpit, Grace, Shenyu, Meeshu, Sakshi, and Matt for the paper reviews, user study participation, brainstorming, VISGivings, and the occasional beer.

Thank you to Fred and Ian for all of the Which Wich memories. Thank you to my best friends and roommates, Meeshu and Shambhavi, for listening to my practice talks, reading my papers, helping me design interfaces and business cards, and using every success and failure as an excuse to buy cheesecake. Thank you to Purr Monster, who sat faithfully by my side, and sometimes on my laptop itself, for all the years of companionship and cuddles.

Thank you to several inspirational educators, who made me love academia and never question that I would continue my education: Ms. Collins; Ms. MacKinnon; Ms. Smith; Dr. Tanenbaum; Coach Taylor; the Torbetts; and Ms. Adams, who believed in my ability to succeed early on.

Thank you to David and Noah, who provided balance in my life through traveling, pizza nights, and days at the park. Thank you to Zack, Jeff, and Sharon for the care and support that got me through my undergraduate and early graduate years. And of course, thank you to my family for their love and support; my parents Dennis and Sue; my brother and sisters: Dennis, Jr., April, Crystal, and Tarra; my aunt and uncle Ginger and Bubber; my nieces: Audrey, Ava, Emma, and Nora; my grandfather George, and my other grandparents who were unable to see me reach this goal: Audrey, Louise, and Earl.

Thank you all. You have helped to make this dissertation not only possible, but an amazing, enjoyable, and unforgettable journey.

Research for this thesis was funded in part by the National Science Foundation under Grant IIS-1813281; the Analysis in Motion Initiative at Pacific Northwest National Laboratory; the Siemens FutureMaker Fellowship; the GVU Foley Scholarship; the D.E. Shaw Exploration Fellowship; and the Graduate Fellowship for STEM Diversity (GFSD).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

People are susceptible to a multitude of biases, including perceptual biases and illusions; cognitive biases like confirmation bias or anchoring bias; and social biases like racial or gender bias that are borne of cultural experiences and stereotypes. As humans are an integral part of data analysis and decision making in many domains, their biases can be injected into and even amplified by models and algorithms. This dissertation focuses on developing a better understanding of the role of human biases in visual data analysis. It is comprised of three high-level goals:

1. **Define bias**: We present four common perspectives on the term "bias" and describe how they are relevant in the context of visual data analysis.

2. **Detect bias**: We introduce a set of computational *bias metrics* that, applied to user interaction sequences in real-time, can be used to approximate bias in the user's analysis process.

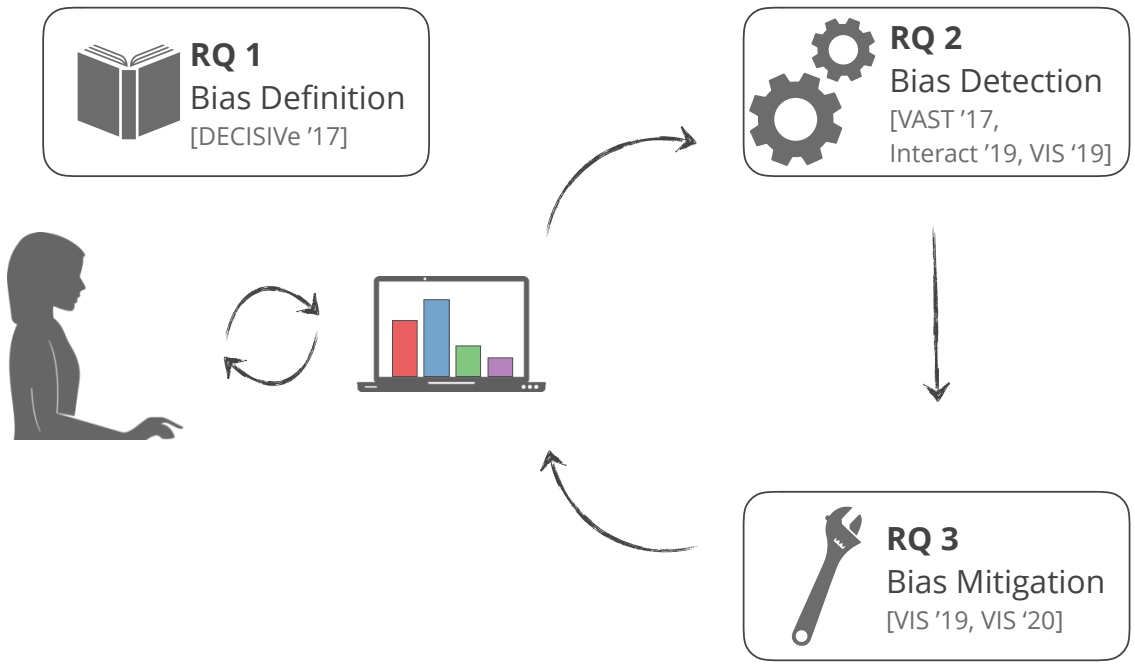
3. **Mitigate bias**: We describe a design space of ways in which visualizations might be modified to increase awareness of bias. We implement a system which integrates and visualizes the bias metrics and show how it can increase awareness of bias.

CHAPTER 1
INTRODUCTION

1.1 Motivation

Machine learning is the tool of choice for solving many large-scale computational problems, including ranking, clustering, and predictive forecasting. While many problems can benefit from machine learning, many still ultimately rely on human input, especially where the consequences of a decision are high (e.g., criminal intelligence analysis) or where domain expertise is required (e.g., personal or corporate purchasing decisions). In such cases, human-in-the-loop (HIL) [50] approaches to data analysis are required. One such solution is visual analytics. Visual analytic systems combine the complementary strengths of humans and machines in what is often presumed to be an ideal combination [93, 175]. Machines offer superior computational power and working memory, while humans have skilled perceptual capabilities and adaptive analytics [72]. Combining these sources of skill and information, the human often bears to burden of making the ultimate decision.

Existing research in visualization has largely focused on understanding the power and limitations of human perception as a primary motivator for the field as a whole [30, 98]; and more recently, there has been an increasing focus on building systems to address data exploration and decision making needs [175]. However, in contrast to the work on perception and system development, relatively little consideration has been given to the role of cognition in visualization. Specifically, these HIL approaches have not often considered inherent limitations suffered by both parties (i.e., humans are biased and error-prone; and machines require substantial training data and are susceptible to biased algorithms and techniques). Yet, when appropriately balanced, visual analytic systems have the potential to mitigate these shortcomings and indeed produce an ideal combination for many decision



k

Figure 1.1: The work in this dissertation, injected into HIL data analysis processes, can enable better decisions. As the user interacts with a visual analytic system during the data analysis process, their interactions are recorded and used as a proxy for understanding their cognitive state (including biases). This information then informs mitigation strategies that alter the visualization to make the user aware of their biases and ultimately support better decision making.

making contexts. Thus, it is the goal of this dissertation to **consider the role of human bias in visual data analysis**.

My approach is three-fold (Figure 1.1): first, to **define** the overloaded term "bias" in the context of visual analytic systems; second, to develop techniques that can characterize or **detect** a person's biases in real-time during the process of visual data analysis; and third, to design systems that can **mitigate** bias by increasing users' awareness to facilitate improved decision making.

1.2 Dissertation Overview

Given the mulitiplicity of the term "bias", I first **define** in which contexts the term is relevant in the domain of visual analytics (Chapter 3). The term is frequently used in cognitive

Table 1.1: Dissertation outline and publication summary.

PART I: DEFINING BIAS IN VISUALIZATION (CHAPTER 3)

§Four Perspectives on Human Bias in Visual Analytics

Emily Wall, Leslie Blaha, Celeste Paul, Kris Cook, and Alex Endert. *DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations (at InfoVis'17)*, 2017.

§Four Perspectives on Human Bias in Visual Analytics

Emily Wall, Leslie Blaha, Celeste Paul, Kris Cook, and Alex Endert. *Cognitive Biases in Visualizations*, Springer, 2018, pp. 29–42.

PART II: DETECTING BIAS IN VISUALIZATION (CHAPTER 4)

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

Emily Wall, Leslie Blaha, Lyndsey Franklin, and Alex Endert. *IEEE Visual Analytics Science and Technology (VAST)*, 2017.

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

Emily Wall, Leslie Blaha, Celeste Paul, and Alex Endert. *Proceedings of the 17th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT'19)*, 2019.

§A Markov Model of Users' Interactive Behavior in Scatterplots

Emily Wall, Arup Arcalgud, Kuhu Gupta, and Andrew Jo. *IEEE Information Visualization (VIS) Short Papers*, 2019.

PART III: MITIGATING BIAS IN VISUALIZATION (CHAPTER 5)

§Toward a Design Space for Mitigating Cognitive Bias in Vis

Emily Wall, John Stasko, and Alex Endert. *IEEE Information Visualization (VIS) Short Papers*, 2019.

§Left, Right, and Gender: Visualizing Interaction Traces to Mitigate Social Biases

Emily Wall, Arpit Narechania, Jamal Paden, and Alex Endert. *IEEE Information Visualization (InfoVis), 2020 (under review)*

science to describe cognitive processing errors that result from the use of heuristics to make decisions [90, 91, 182]. A highly publicized example of the potential impact of cognitive bias in decision making is the Madrid Train Bombing Case, wherein an innocent person was arrested due to latent fingerprint mis-identification by forensic analysts subject to confirmation bias [45, 164]. Another common use of the term is in social settings, in which "bias" describes discriminatory stereotypes or preconceptions that impact people's judgment. For example, a laboratory experiment showed that employers perceived mothers to have lower competence and recommended lower starting salary compared to women with no children [34], while men were not penalized based on their parental status. To address this multiplicity of definitions, I present four perspectives on human bias that are particularly relevant in the context of visual analytics. This work was published at the DECISIVe workshop at IEEE VIS in 2017 [192] and later adapted as a book chapter in a book title Cognitive Biases in Visualization [193].

Given an understanding of which types of bias are relevant in visual analytics, my approach is to next computationally characterize or **detect** bias in real-time during the analysis process (Chapter 4). Combining computational approaches with human perception and sensemaking, visual analytic systems are increasingly used to perform data analysis in the digital world. These systems afford a new opportunity with respect to bias detection, by providing a new way to track and measure a person's decision making process: via *user interaction*. Interactions mark the paths of exploratory data analysis, providing insight into a person's reasoning and decision making processes [127, 145]. This information can, in turn, be used to characterize an analyst's decision making process from the perspective of bias toward specific parts of the data or system. As bias steers users' cognitive processes, it also steers users' behavior through interactions in visual analytic systems and thus the underlying models as well. In particular, I argue that *when data analysis is supported by visual analytic tools, analysts' 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 computationally detect and assess mental pitfalls in real-time during the analysis process. As a result, I present the formulation of two types of computational bias metrics (coverage and distribution), each applied to three components of data and visualization (data points, attributes, and model parameters). This work was published at VAST in 2017 [191]. I validated the metrics through a formative study assessing anchoring bias in a paper published at INTERACT in 2019 [190]. I also demonstrated how the metrics can be further refined by accounting for how the proximity of nearby data points will influence unbiased user behavior in a short paper published at IEEE VIS in 2019 [189].

Lastly, given the ability to characterize how an analyst is biased in real-time, I show how to leverage that information to **mitigate** bias in the decision making process (Chapter 5). I present a design space of considerations for building visualizations that put cognition on the forefront, published as a short paper at IEEE VIS in 2019 [196]. Systems can elevate the importance of supporting effective cognition starting with the early ideation of visualization interfaces. I implemented one such system that mitigates bias by showing users traces of their previous interactions and demonstrated its effectiveness toward increasing awareness of bias in a user study about political decision making using the visualization. This work is currently under review [195].

The outline of this dissertation and summary of publications can be found in Table 1.1.

1.3 Research Impact

Given the increasing popularity of HIL solutions for data analysis and decision making, it is imperative to assess the impact of human bias in the process. Human biases (including cognitive errors, social stereotypes, etc.) can be propagated to or even amplified by underlying computational models. A recent example that showcases the potential consequences of human bias in systems is the AI chatbot, Tay [4, 106]. The artificial intelligence was intended to be a friendly chatbot that appealed to young adults. The underlying model was continually trained by incoming tweets, causing Tay to tweet increasingly racist and misog-

ynistic messages shortly after going live. While a vulnerability in Tay was exploited, the chatbot nonetheless conveys what can happen when human bias is introduced unchecked into a system.

Hence, human bias can be injected into and amplified by algorithms. When people use such algorithms to make important decisions, this can carry heavy consequences, reinforcing biases and stereotypes and ultimately producing an echo chamber. For instance, many courts in the criminal justice system utilize automated algorithms to help judges make decisions about pre-trial release, parole, sentencing, and so on by predicting criminals' likelihood of recidivism, or recommitting a crime [9]. However, recent analysis has shown that these algorithms exhibit racial bias: often incorrectly predicting black defendants as high-risk, while incorrectly predicting white defendants as low-risk.

An awareness of these potential risks will help us develop better systems, and ultimately foster better data-driven decisions. The work presented in this dissertation will enable a new class of visualization tools that are designed not just to combine humans and machines, but to do so in thoughtful consideration of supporting effective cognition and decision making. Future systems will continue to leverage the strengths of both humans and machines, while incorporating additional measures to help guard against the limitations and biases of each.

1.4 Thesis Statement and Research Questions

Thesis Statement. Human bias has a prevalent impact on data analysis and decision making, including the way our visual system biases our perception, the way we utilize cognitive "shortcuts" to make quicker judgments, and the way our judgment is colored by stereotypes and prejudices ingrained in us through our social experiences. However, the process of visual data analysis affords a new opportunity in the detection and mitigation of human bias, namely through the use of user interaction. User interaction can serve as a rough approximation of a user's cognitive state during the process of visual data analysis. Hence, the goal of this dissertation is to (1) define bias in the context of visual data analysis, (2) for-

mulate computational metrics (herein referred to as *bias metrics*) that can be applied to sequences of user interactions to characterize bias during visual data analysis, and (3) design and evaluate interventions that can increase bias awareness in users of visualization systems. *By increasing real-time awareness of bias, people can reflect on their behavior and decision making and ultimately engage in a less-biased decision making process.*

RQ 1 (Define). How do we define human bias in the context of visual analytics?

RQ 2 (Detect). How can human bias be characterized in real-time during the analysis process?

> **RQ 2.1 (Bias Metric Formulation).** By what metrics can user interaction characterize bias in a person's visual analysis process?
>
> **RQ 2.2 (Bias Metric Evaluation).** Can bias metrics be used specifically to capture anchoring bias?
>
> **RQ 2.3 (Bias Metric Refinement).** How can the bias metrics be refined to more accurately account for unbiased interactive behavior?

RQ 3 (Mitigate). Can bias metrics be used in visual analytic systems to mitigate bias?

> **RQ 3.1 (Mitigation Design Space).** How can an interface visually communicate the characterization of a user's bias?
>
> **RQ 3.2 (Mitigation Evaluation).** How effective is the visual representation of *interaction traces* in an interface toward mitigating biased decision making?

CHAPTER 2

RELATED WORK

2.1 Studying the Analytic Process

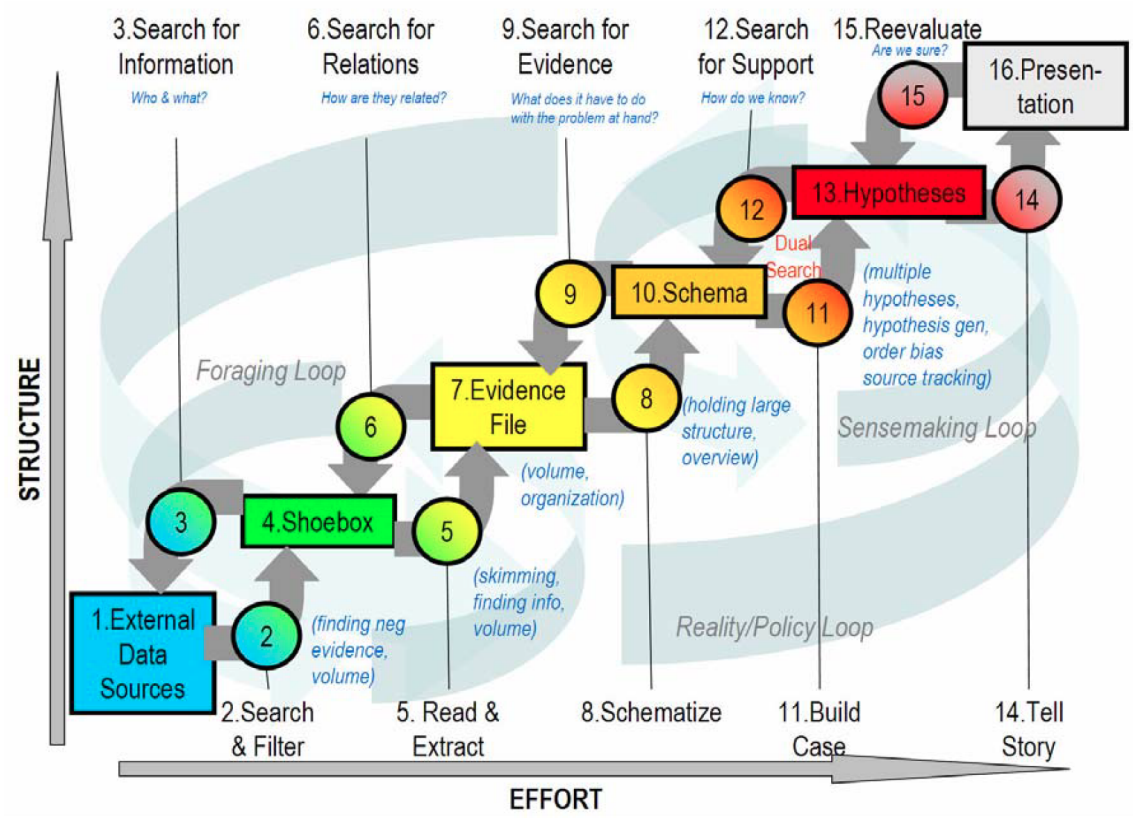


Figure 2.1: The sensemaking loop, as realized by Pirolli and Card [139].

While there is a long history of studying perception in visualization, only recently have researchers begun to more tightly integrate visualization with bodies of work in psychology [1], cognitive modeling [132], and decision making [133]. Within this relatively young space, several prevalent theories exist in the decision making and information processing literature for describing aspects of the analytic process [16]. For example, the sensemaking process was studied by Pirolli and Card [139]. Sensemaking is a term used in visual analytics to describe the process of learning about data through a visual interface; however, the

term more generally refers to the 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.

Pirolli and Card [139] studied this process 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 (Figure 2.1). Each of these higher-level processes is then decomposed into a series of cognitive actions (e.g., the foraging loop involves iteratively finding evidence from external data sources, compiling the evidence, and then skimming it to look for relevant information). On the other hand, Klein et al. [97]

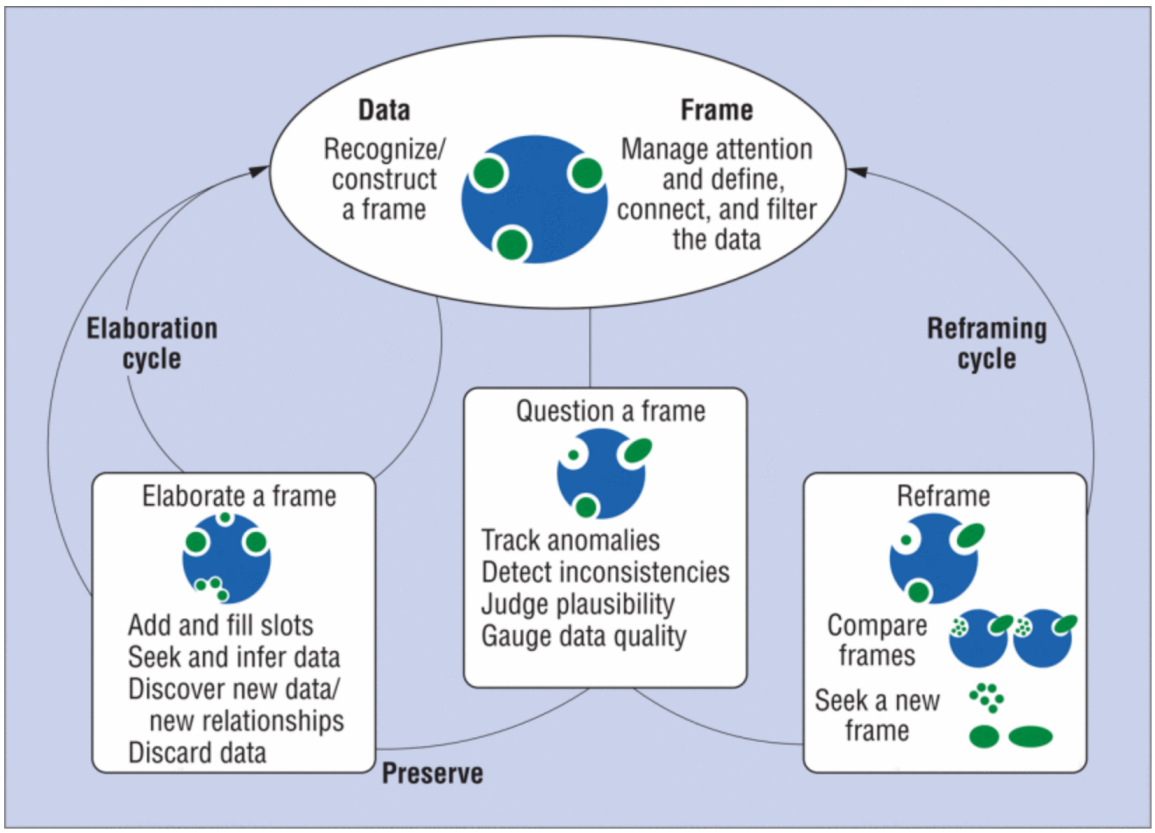


Figure 2.2: The data-frame model of sensemaking, as described by Klein et al [97].

studied the sensemaking process as an iterative framing and re-framing of information. They postulate that analysts begin with some frame of reference when examining data, then continuously compare, refine, and create new frames throughout analysis to refine their understanding of the data (Figure 2.2).

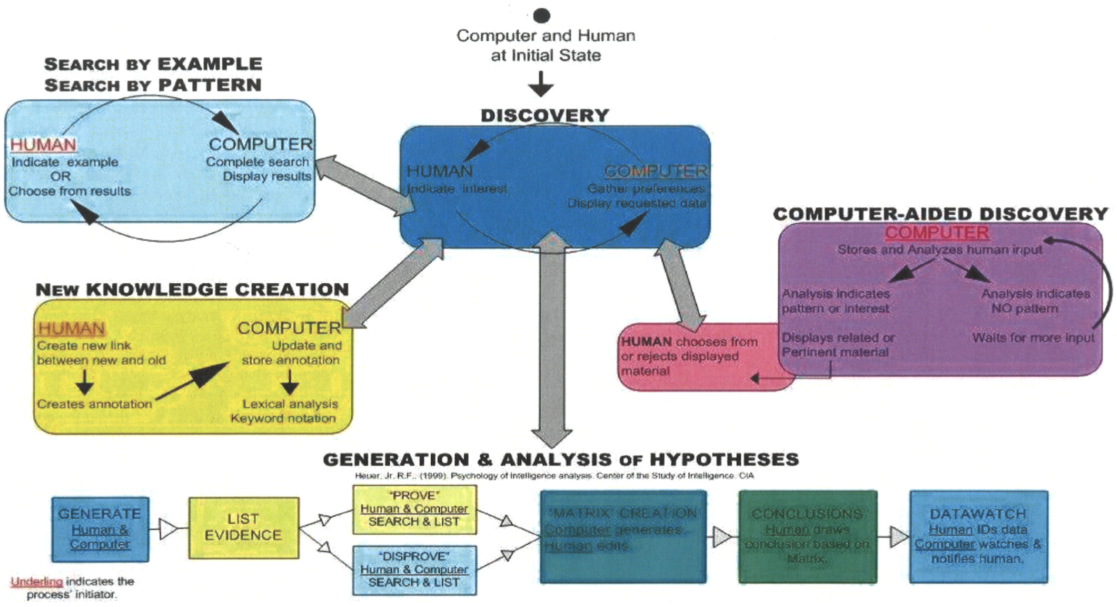


Figure 2.3: The human cognition model, as described by Green et al [72].

Other aspects of the analytic process have also been explored within the visualization and visual analytics communities. Green et al., for example, describe how a human cognition model can be utilized in visual analytics [72]. Their model details the often-complex relationship between human and computer for things like information discovery, creation of new knowledge, and generation and analysis of hypotheses. They describe how tasks and information should be distributed across humans and computers to leverage their complementary strengths in visual analytics (Figure 2.3). Similarly, Sacha et al. [152] 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 (Figure 2.4).

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,

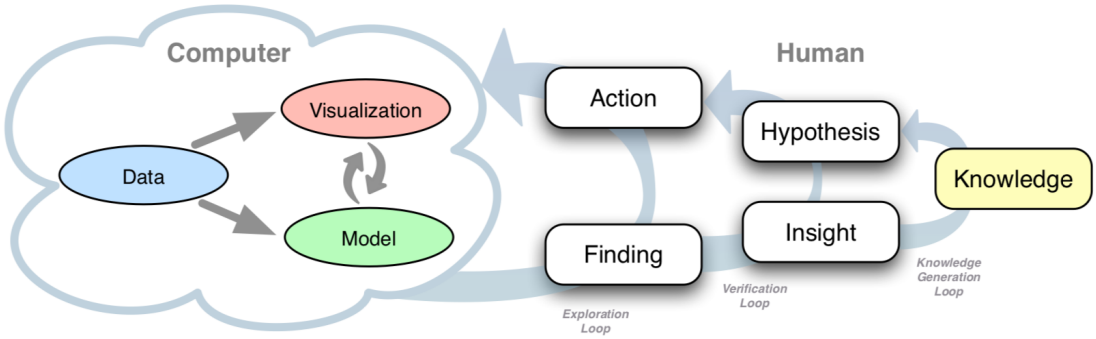


Figure 2.4: The knowledge generation model, as described by Sacha et al [152].

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 analysis cycles.

2.2 Interaction in Visual Analytics

Interaction is paramount in visual analytics [138]. 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 visual data analysis and is a vital part of the reasoning process in visual analytics [141]. Through interaction, users get acquainted with the data, form and revise hypotheses, and generate questions [3]. It allows users to focus their attention in the presence of potentially overwhelming information throughout their analysis [72].

As a key facilitator for human reasoning in visual analytics, interactions can be used to better understand more than just analytic results. They also illuminate the process that led to those results [127]. 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 record and analyze interaction data. A prominent example is GlassBox [36], which captures keyboard and mouse interactions in an interface. More recently, Nguyen et al. developed a

tool to visually group and analyze event sequences to identify common and unusual patterns across many potential users [125].

Interaction data can be a rich source of information about the user. For example, Pusara and Brodley [144] showed the uniqueness of the way users move the mouse by utilizing a supervised learning approach to identify and authenticate which user was interacting. Similarly, Brown et al. [20] showed that by analyzing the way users zoom and pan in a visual search task for "finding Waldo," they could learn about users' task performance and even some personality traits (e.g., locus of control, extraversion, and neuroticism).

Another common use for interaction data is analytic provenance. Analytic provenance is a term used to describe the trajectory of a user's analysis process, beyond the resulting choice or decision. Recent work has shown that user interaction can provide a powerful means for understanding a user's analytic provenance. For example, Dou et al. [44] showed that by analyzing user interaction logs, they could infer a user's reasoning process, including recovering the findings (decisions made after discovery), strategies (means employed to arrive at a finding), and methods (steps taken to implement a strategy and make a finding) during the analytic process. Similarly, Gotz and Zhou [69] combined manual annotations with automatically collected interaction data to identify a set of semantically meaningful actions that can be used to infer about a user's insight provenance.

Furthermore, interactive model-steering is a prevalent mixed-initiative [83] application of user interaction. Systems "take initiative" and act on behalf of users by inferring analytic model constraints from user interactions with a visualization. For example, Endert et al. [49] enabled analysts to drag documents on a canvas to re-weight underlying analytic models of document similarity. Similarly, Brown et al. [19] developed a means of learning a distance function based on the way users re-position data points on a scatterplot. Kim et al. [94] allowed users to drag points from a scatterplot into bins along either side of an axis. In response, the system re-computed weights in a linear dimension reduction algorithm. This work was later extended to non-linear dimension reduction using

sketch interactions [103]. Further, Wall et al. [194] computed an SVM model by inferring constraints from users' interactions dragging rows in a table for mixed-initiative ranking.

Thus, given prior work showing the power of interaction data for making inferences about users' cognitive state and intent, we hypothesize that user interactions can capture behaviors which correspond to human bias during visual data analysis.

2.3 Bias in Cognitive, Perceptual, and Social Sciences

Decision making may be impacted by a multitude of human biases, including cognitive, perceptual, and social biases. This dissertation focuses primarily on cognitive and social biases, described in greater detail below; however, there is also an abundance of work detailing visual perception (e.g., Gestalt principles [98], preatttentive processing theory [177]) and perceptual bias (e.g., selective perception [160], the Stroop effect [109]).

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 [107]. 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 [90].

Some common examples of specific types of cognitive bias include things like confirmation bias, anchoring bias, and the availability heuristic. Confirmation bias describes the tendency for people to search for evidence that confirms pre-existng hypotheses [126]. For

example, an individual may look only for articles describing current events from a political perspective that aligns with their own, while dismissing articles that disagree with their political worldview. Anchoring bias, on the other hand, describes the mental error that results from people's tendency to rely too heavily on the initial information presented to them [51]. For example, an individual looking to purchase a car will make an offer that is largely affected by the initial price presented. They are unlikely to stray far from the initial list price, even if the car is overpriced. Lastly, the availability heuristic describes the way humans tend to rely more heavily on information that is most easily remembered or most recent [181]. For example, if asked about the importance of safe driving education, an individual who had a recent car accident will likely rate the importance higher. These are just a few examples; however, there are more than a hundred types of cognitive bias that have been described in the literature [42].

In Social Sciences, on the other hand, bias often refers to prejudices or stereotypes that are socially relevant (e.g., racial bias, gender bias, age bias, etc). Social biases may be influenced by cultural norms, individual experiences or personality variations, and they can shape our decision making in a conscious or an implicit manner [73]. These biases can have severe implications in a variety of decision making domains. For example, consider the impact of racial bias in hiring. Researchers have found discrimination, either conscious or implicit, based on racial name trends [14], showing that equivalent resumes with White names receive 50% more callbacks from job applications than resumes with African American names. As a result, companies may lack a diverse workforce, which can have implications on employee turnover, group isolation or cohesion, workplace stress, and so on [148]. In the digital world, these biases can have far-reaching impacts. Combined with Machine Learning, algorithms can learn and propagate things like racial or gender bias [64, 113].

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 [70, 71, 110] or the similarity choice axiom [108, 140], 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 [24, 147]. 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 human bias exist, they are limited. Specifically, they are not suited for making people aware of their potential biases *during* the analysis process. Thus, distinct from methods for detecting bias from the *product* of decision making, 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. Based on prior work described in Section 2.2, 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 [179, 182]. 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 [99, 162, 163].

2.4 Bias in Visual Analytics

The topic of bias in visual analytics has recently garnered increasing attention. Several recent works have begun to organize and formalize the types of bias relevant in the visualization and visual analytic domains [35, 184, 192]. Perhaps most extensively, Dimara et al. [42] categorized 154 types of cognitive bias into 7 categories of a task-based taxonomy for information visualization. The categories included biases associated with estimation tasks, decision tasks, hypothesis assessment tasks, causal attribution tasks, recall tasks, opinion reporting tasks, and a miscellaneous "other" category.

Several researchers have addressed modeling of visualization history [78] and process, including the development of metrics associated with depth and breadth of analysis [88], as well as exploration uniqueness and pacing [54]. Others have explored the manifestation of a specific type of bias in the context of information visualization or visual analytics. For example, Gotz et al. [68] addressed the issue of selection bias in examining healthcare data. Because many attributes in high-dimensional datasets are often correlated (e.g., height and weight in certain datasets), selection bias in healthcare data can be a prevalent issue. Analysts may be unaware of the ways they have unintentionally biased the selection of data they are examining. Hence, Gotz et al. proposed an approach to quantify and visualize unseen selection bias.

Dimara et al. [41] examined a different bias, the attraction effect, in information visualization. The attraction effect describes the phenomenon where a person's decision between two alternatives is altered by the introduction of an irrelevant third option. By introducing decoy options to visualizations, including tables and scatterplots, the attraction effect was observed in the visualization domain. They further showed how this effect can be mitigated using interaction by providing the capability to locally delete data points from interactive visualizations before making a decision [40]. Valdez et al. [185] similarly showed the presence of bias in information visualization. In their sequence of experiments, they showed

the presence of priming and anchoring effects. Given scatterplots as stimuli, they asked participants to judge whether two classes of points were separable or not.

Perhaps most similar to the work described in this dissertation is work done by Cho et al. [29], who replicated effects of anchoring bias in a visual analytic tool. As mentioned above, anchoring bias is an over-reliance on a particular piece of information [51]; Cho et al. demonstrated that people exhibit anchoring bias when utilizing visual information sources. In their study, participants were tasked with predicting protest events by analyzing Twitter data. They elicited anchoring bias through priming, then analyzed user interaction data to measure participants' reliance on particular views in a multi-view system through post-experiment metrics (e.g., total proportion of time in each view). While we similarly propose the assessment of bias through analysis of user interaction, we focus on real-time assessment rather than post-experiment analyses.

2.5 Bias Mitigation Strategies

Prior work broadly categorized bias mitigation strategies as either *training interventions* or *procedural interventions* [100]. We incorporate these strategies and others in two high-level categories of mitigation strategies: *a priori* and *real-time*, according to when the technique is employed with respect to the analysis process. Developing a successful strategy for mitigating human bias in mixed-initiative visual analytic systems depends on identifying when and how each of the following strategies might be employed with positive outcomes [63]. There have been varying degrees of past success addressing bias in the analytics process, which we describe in greater detail below.

2.5.1 A Priori Bias Mitigation

A priori bias mitigation strategies occur before the analysis process, often in the form of educational training that may examine past errors to inform future decision making. A number of attempts have been made to mitigate bias in the domain of intelligence analysis,

including training courses, videos, and reading material. While the goal is to promote informed decision making by the analyst leading to a shift in user behavior, 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"* [79].

Serious games provided a more effective alternative to traditional means of bias training [15, 46, 57, 122, 169]. These techniques educated analysts about cognitive biases, but nonetheless did little to mitigate negative effects when biases inevitably occurred in the analytic process. They reinforce that an analyst must be proactive using feedback to adjust their behaviors to mitigate the negative effects of bias.

2.5.2 Real-Time Bias Mitigation

If biased decision making processes can be assessed and measured in real-time, bias mitigation strategies can do more than simply educate analysts beforehand. Real-time bias mitigation strategies have the potential to intervene at a more effective time. Though many of the techniques described below have not previously been used explicitly for bias mitigation, we draw inspiration from techniques we believe will encourage thoughtful reflection on people's decisions and analytic processes.

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 ways of intervening in biased decision making processes.

Non-Technological Strategies. One approach for mitigating bias in real-time is the use of non-technological strategies, including structured externalized processes, or structured

analytic techniques [80]. Perhaps the most known and accepted is Analysis of Competing Hypotheses (ACH) [79]. ACH 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.

Similarly, "consider the opposite" decision making strategies make a thoughtful reflection of evidence for the alternative hypotheses a structured part of the decision making process. This has shown promise to reduce some biases, including overconfidence, hindsight, and anchoring [10, 123]. However, these procedural thinking strategies come at the cost of potential cognitive overload, which can ultimately amplify some biases [155]. Herein, we focus on machine-assisted strategies that can lighten the cognitive burden of bias mitigation.

Increasing User Awareness. We posit that there are many ways visualization and visual analytic tools can mitigate bias in real-time by simply increasing the user's awareness of their process. For example, Dimara et al. [40] observed that highlighting points in a scatterplot that fall on the Pareto front led to a lower susceptibility to the attraction effect. While Dimara et al. [40] increased awareness in static scatterplots, we are interested in mitigation strategies in interactive settings. One approach could be to characterize the analysis process and alert the user of detected biases in real-time in the form of notifications. There is a rich space of prior work around notifications, including the tradeoffs of push v. pull [21, 114], the effects of timing and interruptions [2, 37], and so on. Such work can inform the design of notification systems for increasing user awareness of biased analytic processes.

Other researchers have tried to raise users' awareness of their analytic process by visualizing analytic provenance or coverage of the possible exploration space [11, 43, 89, 197]. With such feedback, users tended to explore more data [53], make more unique discover-

ies [197], and show greater search breadth without sacrificing depth [156]. Thus, visual characterization of the analytic process has potential to mitigate bias by altering a user's exploration.

Collaborative Mitigation. Feedback about biased behaviors can be given to a third party agent (e.g., a human teammate or a supervisor) to leverage "wisdom of crowds" [112, 168] to cancel out the noise of potentially sub-optimal individual decisions [82, 143]. This strategy could prove useful in collaborative analytic 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. In this case, prior work on fostering awareness in collaborative settings can be informative [11, 12, 13, 75, 76, 171].

Machine Initiative. Mixed-initiative [83] visual analytic tools provide a unique opportunity for bias mitigation. That is, the machine could operate as an unbiased collaborator that can act on behalf of the user, or *take initiative*, to mitigate biased analysis processes. Machine feedback supports adaptive systems or other machine-based cognitive augmentations that are responsive to the user's state. Some mixed-initiative efforts have already begun to integrate visual analytic recommendations based on user interest or semantic interactions [49]. Gladisch and colleagues [67] 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.

Mitigation through Interface and Interaction Design. Law and Basole [104] describe interface design considerations for encouraging more broad data exploration as a means for mitigating some biases. Design considerations include what constitutes a unit of exploration (i.e., to track breadth of exploration across data points or dimensions), the distinction between user- and system-driven exploration (i.e., to focus on showing users meaningful information about their own analytic process or the system's computational assumptions),

and the alternatives of related and systemic exploration (i.e., whether the expansion of information is driven by the user or not). They demonstrate a prototype tool for social network analysis that considers these dimensions in the design of the interface for encouraging broad exploration of the data. Similarly, Sukumar and Metoyer [166] present design considerations for bias mitigation in the context of the college admissions process. Their guidelines include a broad focus on tasks like easing cognitive load, supporting sensemaking, decorrelating error (i.e., not viewing students' applications as a whole), mobilizing system 2 (i.e., encouraging thoughtful consideration of application content over quick intuitive judgments), and combining formulas with intuition (i.e., to limit the drawbacks of either in isolation).

A recent study by Dimara et al. [40] showed that interaction design can itself serve to mitigate bias (namely, the attraction effect). The attraction effect describes the phenomenon where a person's decision between two alternatives is altered by the introduction of an irrelevant third option. In a decision making task using scatterplots, users were required to utilize a process of elimination, clicking to remove individual points on the scatterplot, until a single point remained as their final decision. With this interaction design, researchers observed a reduction in the attraction effect compared to participants who simply clicked on their final decision point (without first eliminating others). Hence, considering alternative interaction designs for a task could serve as an effective intervention for mitigating biased decision making.

2.6 Expertise and Uncertainty

Visual data analysis involves a number of complex cognitive processes, influenced by factors such as bias, expertise, and uncertainty. This dissertation focuses on bias; however, expertise and uncertainty are inextricably connected concepts, relating to and feeding into specific types of bias. Hence, a brief review of expertise and uncertainty provides additional context for the assessment of a user's cognitive state.

Expertise can be considered from many perspectives, including visualization literacy or expertise, domain expertise, and so on, all of which can impact the way people approach a task. For example, Xiong et al. showed that prior knowledge and beliefs about the data influence the way people interpret data and communicate with visualizations (i.e., the curse of knowledge) [199]. On the other hand, consider basic visualization literacy. Variations of literacy in visualization have been examined under many different names, including visualization literacy [18], visual information literacy [172], data literacy [149], graph comprehension [60], and graphicacy [188]. Boy et al., for instance, define visualization literacy as "the ability to confidently use a given data visualization to translate questions specified in the data domain into visual queries in the visual domain, as well as interpreting visual patterns in the visual domain as properties in the data domain" [18]. Visualization literacy, and data literacy more broadly, can have a big impact on the way people analyze data and make decisions using visualizations, influencing the views they rely on, the interactions they perform, and as a result, the distribution of their attention across the data.

Literacy and expertise, or lack thereof, ultimately influences individuals' susceptibility to specific types of bias (e.g., Dunning-Kruger Effect [101], wherein people who are unskilled in an area will overestimate their competency). Furthermore, when visualization literacy or expertise is low, people may be particularly susceptible to deceptive visualization techniques [134]. While deception typically has a negative connotation, it is often optimal to distort the presentation of data for some knowledge communication goals [32]. However, the vast majority of work on deception in visualization focuses on its malevolent uses. Correll and Heer describe techniques including data manipulation, obfuscation, and nudging, that may be used to deceive with visualizations, some of which even experts appear ill-equipped to identify [33]. O'Brien and Lauer showed that people are susceptible to deceptive visualizations, even when paired with accurate explanatory text – highlighting the importance of visualization literacy given people's high reliance on visual representations of data [131]. However, these malevolent deceptions are not without hope. Recent

work has shown that metamorphic testing [157] of visualizations could be used to automatically identify such deceptive visualizations, or "mirages" [117]. Szafir also proposed design guidelines for visualization designers and developers to avoid common misleading representations [170].

Another important influence on decision making is uncertainty. Work by Kim, Hullman, and colleagues has modeled the way people make decisions based on prior beliefs (or biases) using a Bayesian cognitive model and studied the way different representations of uncertainty [85, 92] influence people's "rational" choices [96]. Furthermore, they demonstrate that people's uncertainty about data trends leads to heavy reliance on social influence [95]. Thus, creating visualizations that better support cognition will require a more complete understanding of the concepts of bias, expertise, and uncertainty, and the intricate ways they relate to or feed into one another.

CHAPTER 3
DEFINING BIAS IN VISUALIZATION

To reach the ultimate goal of developing methods to detect and mitigate bias in visual analytics, we first must understand what is meant by the term "bias." Hence, this section describes work that has been done in response to **RQ 1** and has been published as a workshop paper [192] as well as a book chapter [193].

RQ 1: *How do we define human bias in the context of visual analytics?*

Cognitive, behavioral, and social sciences have described many ways bias can occur in people's analytic processes [97, 139, 152], decision-making strategies [20, 44], and other behaviors. Motivated by the overloaded use of the term "bias" to describe different models and concepts, we describe four different ways people tend to think about or refer to human bias that are relevant in the context of visual analytics. These perspectives include (1) bias as a cognitive processing error, (2) bias as a filter for information, (3) bias as a preconception, and (4) bias as a model mechanism. These four perspectives are not mutually exclusive; rather, they present different, potentially overlapping perspectives on bias relevant in the context of visual analytics. Furthermore, as described in Chapter 2.3, the measurement of bias according to each of these perspectives typically relies on the *products* of cognition (e.g., final choice or decision) rather than measurable parts of the *process*.

To more concretely discuss how bias can affect visual analytics, consider the following example. Suppose Susan is using a visual analytic tool to explore possibilities for purchasing a new home. She uses the tool to browse photos, explore different areas of the city, and refine her understanding of what features of a home are important to her. From her exploration, she intends to go view the homes in person and ultimately make a purchasing

decision. Throughout the following sections, we will describe how each perspective on bias can impact Susan's process and visual analytics in general. For each perspective, we provide a brief description, present an example scenario, and discuss how these perspectives inform and influence visual analytics.

3.1 Bias as a Cognitive Processing Error

3.1.1 Description

From heuristics and bias research, bias is an error resulting from an unconscious deviation from rational behavior. Cognition is frequently conceptualized as a dual-process [27]. The two processes are often termed "intuition" and "reason" [91], the former being responsible for making quick, automatic decisions, and the latter being responsible for making deliberate, reflective decisions. It is one's quick judgments that are subject to errors.

Stanovich and West referred to the two cognitive processes as system 1 (intuition) and system 2 (reason) [165]. In this analogy, system 1 is largely subconscious and prone to making errors (bias), while system 2 is responsible for recognizing and correcting errors through intentional deliberation. These types of errors result from shortcuts in cognition, broadly referred to as heuristics [91]. Bias then is described as the method or mechanism by which the error occurs. However, the process of heuristic decision making does not always lead to errors; it usually facilitates fast decision making.

3.1.2 Example

From this perspective, there are dozens of types of bias. One such example is anchoring bias [182], which refers to the tendency to be heavily reliant on an initial value or anchor. It is analogous to a center of mass: people are unlikely to strongly deviate from their center. In Susan's home-buying scenario, she will likely be subject to anchoring bias during the price negotiation of her purchase. That is, the home's initial list price forms an anchor point and will thus subconsciously impact the amount she is willing to offer. Susan's offer

for the home might have been very different had she made an offer given a different initial list price. She might even pay more money for the same home due to the tendency not to strongly deviate from the anchor point. Systems apprised of probable cognitive errors like anchoring bias have the potential to help users make better decisions by guarding against such errors, providing appropriate counterexamples, or by suggesting other ranges of data values that a user might consider.

3.1.3 Relevance to Visual Analytics

Common heuristic errors include confirmation bias [126] which describes the way people tend to accept confirmatory evidence of a pre-existing hypothesis and dismiss contrary information. Another common error is availability bias [181], where people tend to rely more heavily on information that is easily remembered (e.g., most recent). Similarly, the attraction effect [84] describes the tendency for a decision to be influenced by an inferior alternative. Collectively, these errors shape the way people search for and interpret information. Recently, Dimara et al. [41] showed that the attraction effect is present in users of information visualizations. Similarly, researchers have shown that priming and anchoring effects can be replicated in visualizations and visual analytics [29, 185]. Hence, bias impacts users outside of laboratory decision making studies and can lead to incorrect decisions and inefficiencies in visual data analysis.

3.2 Bias as a Filter for Information

3.2.1 Description

Bias acts as a filter through which we manage and perceive information. The challenge of information overload [119] motivates this analogy. Information overload, now commonly leveraged in consumer research to influence purchasing behavior, refers to a point beyond people's cognitive and perceptual limits where performance and decision making suffer [111]. Under overload conditions, people selectively allocate attention and other

mental resources to the tasks or information of highest priority. One's filter or bias thus determines what information is gathered and how sensory information is distinguished and interpreted [61].

The literature on goal-directed attention and resource allocation posits that all perception is guided by top-down influences, such as the allocation of endogenous attention [47, 142, 176]. Top-down perception governs which sensory information is identified in a scene based on goals. Bias does not make for a purely objective filter for information, however. Heuer refers to perception as an "active" process that "constructs" reality [79]; this is in contrast to a passive process that simply records reality. Similarly, obvious or important information is sometimes filtered out. For example, in one classic selective perception task, participants were shown video footage of people wearing either white or black shirts passing a basketball. Participants were asked to count how many times white-shirt basketball players on a team passed the ball to each other [160]. Most participants count the appropriate number of passes but about half fail to perceive a glaringly misfit player walk across the court. The misfit player is in a black outfit, and is consequently treated as part of the task that is selectively ignored while attention is focused on the white-shirt players. In contrast to top-down perception, bottom-up perception refers to the way external factors influence attention [150]. When there is a loud noise or someone says your name across the room, you notice despite top-down attentional and perceptual focus. Visual attention can be similarly grabbed by flashing, movement, or other visual cuing in a display.

3.2.2 Example

In our home-buying scenario, Susan may experience information overload [119] as she explores homes on the market in a visual analytic tool. She might see hundreds of homes available in the area, each with dozens of attributes. Thus, her filter or bias will govern which information she perceives and which she dismisses. For example, she may only select to view single-family homes, removing condominiums, town homes, and apartments

from the visualization. If removed, some options that may be relevant to Susan's other search criteria will not be visibly available, though still in the underlying data and system. The system may want to make some of that information known at an appropriate point in the analytic process. By leveraging knowledge about people's perceptual strengths and limitations, mixed-initiative system could present information in ways that are easy for users to understand and at a time when mental resources are available.

3.2.3 Relevance to Visual Analytics

A great deal of research in visual perception [183] has been leveraged by researchers in information visualization and visual analytics to present information in ways that are most perceptually accessible [52]. Preattentive processing theory [177], for example, describes the nature and limits of visual information processing. In creating visual representations of data, this is often used by designers as a guide to prevent overwhelming a user's perceptual limitations. Similarly, Gestalt principles [98] refer to the relationships inferred by the visual system based on proximity, groupings, symmetry, etc. between visual elements. Thus, understanding how people's filters work can inform things like which visual widgets or elements to place in close proximity to one another or which graph layout algorithm is most appropriate. Indeed, Patterson et al. [136] listed supporting attention and user mental models as some of the key visualization leverage points for design grounded in human cognition.

3.3 Bias as a Preconception

3.3.1 Description

Analysts approach mixed-initiative systems bringing all their experiences and internal influences that unconsciously shape their approaches to the analysis process. This, in turn, influences the ways they interact with systems. The consequence is that the user model within the system, the analytic products, and provenance may be shaped by each individ-

ual's unconscious biases. These types of bias may seem to have little to do directly with the task at hand. Yet, because they shape the person, there is a high likelihood they can influence mixed-initiative sensemaking.

Unconscious biases arise in a number of ways. They derive from a person's cultural beliefs and traditions, which include their implicit assumptions and expectations regarding stereotypes. Unconscious biases result from general self-confidence or self-esteem, as well as comfort or familiarity level with the capabilities of a machine's analytics and interface functions. Related personality traits render some people more risk seeking or risk averse, shaping how they push boundaries exploring a space of hypotheses or push the capabilities of the computational system. These characteristics are thus seen as a source of individual variability between people.

3.3.2 Example

Susan is avoiding listings for houses downtown in the city. Having lived in the suburbs for many years, Susan assumes that neighborhoods near downtown have higher crime rates and lower economic stability. She believes she should not make a housing investment there. The availability of recent census results and police reports within the real estate analytic tools enable Susan to explore her assumptions and refine her thinking. A mixed-initiative system may detect her avoidance of downtown properties and could prompt her to challenge her assumptions with the related data.

3.3.3 Relevance to Visual Analytics

Unconscious biases shape analysts' assumptions and stereotypes about analytical tools and mixed-initiative aids, and they shape assumptions and stereotypes about the data / analytical subjects (e.g., presumed reliability or trustworthiness of certain sources). Implicit attitudes shape the formulation of hypotheses and the questions about the assumptions and the consequences of those hypotheses. Klein and colleagues posited that the entire sensemaking

process begins with a practitioner framing the problem, and the selected framework, however minimal, then shapes what an analyst thinks about and what structure they think with [97]. Frames reflect a perspective an analyst takes to make sense of data or to solve a problem. As implicit attitudes shape an analyst's perspective, they shape the analyst's frames, thereby shaping the sensemaking process.

Use of the system is also influenced by the level of trust the user places in computational systems, which is shaped by the degree of machine autonomy the system has together with its transparency about its capabilities and uncertainty [105]. Some people are more pre-disposed to trust computational systems. This would manifest in differences in the degree of reliance an analyst places on the machine's results or recommendations. Generally, the strategy for addressing differences in reliance and trust is to find a means of trust calibration, or helping the user to adjust expectations about machine capabilities [81]. It is possible that the preconceived biases that might play into the analytic process could influence trust and reliance on the visual analytic system. Consequently, the mixed-initiative interface should be providing cues to enable the user to calibrate their trust in the machine as well.

Expertise, derived from general experience as well as explicit training, further shapes the analytical process and is shaped by implicit biases. Expertise can impact expectations and perceptions of a mixed-initiative system and the interpretations of the information visualizations under consideration. Expertise in forensic analytics, for example, may make analysts more conservative in their judgments, shaped in part by their expert understanding of the consequences of their decisions. Expertise often also provides the user with a better understanding of the limitations of the analytical tools or data collection practices, which can shape more nuanced interpretations during the analysis process.

Because they are built to record a number of different types of user behaviors throughout the analysis process, mixed-initiative systems may be particularly well-positioned to aid in the assessment of unconscious biases of analysts. We argue that it is possible for a

mixed-initiative system to capture and integrate unconscious, preconception biases into analytics through the user model and track those biases through user interactions and changes in the mental model over time.

3.4 Bias as a Model Mechanism

3.4.1 Description

Bias is the term often used in mathematical psychology to describe a decision boundary or a tendency toward one response option over another. Cognitive architectures or models are mathematical and computational approaches to formally describe mechanisms supporting perception, memory, decision making, and other cognitive functions [23]. A number of these models include a mechanism explicitly called bias, or they use a combination of mechanisms to capture the ways the aforementioned types of bias manifest in measurable behaviors, like response choice and speed. Models with explicit bias mechanisms often contain a bias parameter or measure bias as a relationship between parameters. Here, we will review two major perspectives on bias as a model mechanism, one which formalizes bias within models of mental organization and another which formalizes bias in models of decision making dynamics. Both types of behavior are necessary in visual analytics, as analysts work through their sensemaking processes of organizing information and weighing evidence against potential hypotheses and interpretations. As interactive visual analytic systems aid in the externalization of analysts' mental models, model mechanisms can help us interpret how bias is reflected in the patterns and dynamics of their interactions.

One approach to modeling bias addresses the question: where do people mentally "draw the line" between one response option and another when performing an analytic task? Many models of perceptual choice or organization describe information representation with two mechanisms [98, 183]. One mechanism is spatial organization that groups pieces of information by similarity/proximity; like objects are close in space or clustered together. The second mechanism is at least one boundary that divides the space into response

regions; object labels or choices are made according to the response regions defined by the boundary. Examples of these models include the theory of signal detection for finding signals in noise [70, 110] or categorization models [108, 129] for clustering and labeling groups of objects. Bias in these models is based on comparison to unbiased sensory input. Bias is described by a weighting of boundary regions; if regions are not equally weighted, the model represents bias toward certain responses. Other models might capture bias as a feature weighting, representing how much the respondent emphasized certain features over others.

Another major use of bias parameters is found in models of information processing dynamics behind the time to make a decision. These dynamic decision models characterize the choice between two options as a stochastic process whereby information about the options is incrementally sampled and accumulated, often in a random walk fashion, until some threshold is reached for one of the response options [24]. The evidence accumulation process governs a person's response speed and is influenced by the salience and complexity of the choice options. Bias in these models is captured by the relationship between the starting value of the evidence accumulators and the response thresholds. If the accumulator starts at zero, then the process is not biased; all responses are equally likely. If the bias parameter is non-zero, then the process is biased toward the response threshold closer to the bias value. This bias mechanism captures behaviors wherein some responses, correct or erroneous, are selected more frequently or more quickly than others.

3.4.2 Example

Homes for sale are comprised of a large number of attributes drawn from real estate descriptions. Susan is likely to have certain features along which she is organizing the options available on the market, such as number of bedrooms, number of bathrooms, basement square footage, and proximity to schools. This forms a four-dimensional mental representation space into which the houses can be organized. If she is weighing numbers of

bedrooms and bathrooms equally, we can describe her decision bias as equidistant from the category centroids or close to zero. However, Susan has strong opinions about basement square footage and proximity to schools. Based on how she organizes houses into desirable and undesirable categories, we might use models to infer that she is biased toward liking houses that are within a 10 minute walk to schools but have small basements less than 400 square feet. A system aware of these preferences might help quickly reorganize large amounts of data into a representation consistent with the user's mental representation.

3.4.3 Relevance to Visual Analytics

Visual analytic systems designed to support data exploration capture an externalization of the analyst's mental organization in the form of interaction [86, 87]. By leveraging analytic provenance [127], researchers can better understand users' strategies [44], processes that led to insights [69], and ultimately better support the sensemaking process [200]. Different spatial layouts and data encodings (including colors, shapes, etc.) reflect mental organization patterns, including perception of similarity between data points. Characterizing the biases in this mental organization process provides a quantifiable way to describe the information representation space and decision boundaries. For example, we can use the perceptual organization models to infer if the analyst is biased toward some data attributes or certain clusters/labels. We could use the sequential sampling model to identify biases in how analysts are weighing the relative utility or value of a piece of evidence.

From the perspective that bias is a model mechanism, we can also formally characterize bias from the other three perspectives described in Sections 3.1– 3.3. Although these models are implemented in a way that is rather agnostic to errors in reasoning, the bias parameters enable inferences about how errors from decision heuristics occur. For example, anchoring bias would be captured as a bias toward one of the response thresholds close to the anchor value in a model of information accumulation or decision dynamics. Bias as a filter can be formalized as a bias node or parameter in a neural network or hierarchical

model of vision [178]. This would reflect the way information might be differently sampled by an analyst based on the goal-related task they are performing. Preconception bias can be included in models as latent factors or correlates of measurable behaviors. As latent factors, biases such as gender or race stereotypes can modulate other mechanisms in the mental models, such as the organization of similar objects or response preferences [186].

3.5 Discussion

These four perspectives of bias illustrate the diversity in how people process information and form a model of the world: (1) bias as a cognitive processing error, (2) bias as a filter for information, (3) bias as a preconception, and (4) bias as a model mechanism. Each are valid perspectives that greatly shape how bias is framed in visual analytics research. However, the multiplicity in definitions sometimes leads to challenges in sharing and collaboration due to a lack of common ground. One goal of this work is to present these definitions, so that we as a community have a starting point for discussing how these perspectives fit within the visual analytics research agenda. Additionally, when considering all of these perspectives, the space in which to study bias in visual analytics increases dramatically. This leads to several open challenges and opportunities for the visual analytics community.

3.5.1 Does bias endanger mixed-initiative visual analytics?

Visual analytic applications continue to model users and adapt interfaces, visualizations, and analytic models based on their interactions. However, how do such systems differentiate between valuable subject matter expertise (which should be incorporated), and biased input? Without such techniques for identifying and guarding against biased input, applications run the risk of showing users biased views of their data that correspond to what they want to use, rather than truthful representations of the information.

For example, in model-steering situations, user input guides analytic models to focus on salient aspects of the domain being studied [48]. Without guarding against potentially

biased user input, the system may overfit the model to the biased input. The result may be a system that shows users the views they want to see, but is essentially an "echo chamber" for their own biases.

One approach for making the distinction between valuable domain expertise and biased input might be to consider the consistency or inconsistency of a user's interaction sequences. More sophisticated approaches could be derived by studying the differences in interaction sequences of domain experts and novices who are biased. It may also be useful to study large groups of users, expert or novice, modeling their processes and biases, to provide additional context to the machine intelligence about ranges of typical and outlier behaviors.

3.5.2 How to keep the machine "above the bias"?

Designing mixed-initiative visual analytic systems to reduce negative effects of biased user input is an interesting and important line of research leveraging our bias classifications. As noted by Friedman and colleagues, there are three types of bias that can influence computer systems: pre-existing, technical, and emergent biases [58, 59]. Pre-existing bias arises from the attitudes or societal norms/practices that the software designers might impart into system designs. This is akin to our bias as a preconception perspective. Concerted efforts can be made to address pre-existing bias throughout the visual analytics design process, such as using the recent GenderMag method to address gender biases in interface designs [22].

Technical biases are a consequence of technical considerations, such as choice of hardware or algorithm. Computational technical biases are unique from the various definitions of human bias we summarized herein. But because they will contribute to biases in mixed-initiative system performance, careful technical choices should be made and appropriate details should be made available to the user to facilitate informed interpretation of system behaviors.

Emergent biases arise from the use of a system, resulting from changing context or knowledge in which a system is being used. Friedman argues that these are more difficult to know in advance or even identify in practice [59]. Emergent biases are highly likely to occur in mixed-initiative systems, particularly as the interface or algorithms are shaped by any of the aforementioned biases that are influencing the user's interactions. Theoretically, the role of the machine is to be unbiased and to present a rational result based on clear rules. However, there are limitations to this approach, namely the lack of tacit knowledge and analytic context that cannot be easily modeled. This has led to the rise of user-driven machine learning that goes beyond a "supervisory" role in training [7]. Yet, as soon as the human is re-introduced into the system, the rationality of the machine is affected. How can we judge when this human-machine teaming is succeeding or failing?

We propose that mixed-initiative systems are uniquely suited to aid in the identification and mitigation of emergent biases, exactly because mixed-initiative systems reflect the user's analytic process. To do this then, we must be able to correctly interpret the user's biases as they are captured by the computational system. The four perspectives we have outlined will help the bias interpretation process. Each provides a way to identify how that source of bias plays out in the analytic process. To the degree that formal models are available for each bias perspective, those can be integrated into the system for more automated interpretations.

3.5.3 Is bias good or bad?

The term bias tends to carry a negative connotation. It is perceived as something that we should strive to eradicate. However, bias is not always bad. Each of the four perspectives on bias differs in how it impacts cognitive and perceptual processes.

From the perspective that bias is an error, we should work to minimize it; however, it should not be confused with the heuristic decision making processes that lead to such biases. We emphasize that heuristic decision making is not inherently bad. It usually results

in more efficient decision making. Thus, it is imperative that in attempting to mitigate bias as an error, we do not unduly limit heuristic decision making processes in general.

From the perspective that bias is a model mechanism, it is neither good nor bad. In this case, it is an objective characterization of the decision making process. While the decision making process itself may be suboptimal or erroneous (as is the case of bias as an error), here bias just describes the boundary between response options.

From the perspective that bias is a filter and the perspective that bias is a preconception, it can be both beneficial and detrimental depending on circumstances. Perceptual filters prevent us from experiencing information overload. However, they can also cause us to inadvertently filter out information relevant to a given decision. Unconscious biases like innate risk-aversion tendencies can help us to make deliberate, mindful decisions, but on the other side of the spectrum can lead to impulsive high-risk decisions. Thus, because different perspectives on bias vary widely in their potential benefits or risks, it is imperative to thoughtfully define the perspective and scope considered for bias detection or mitigation efforts.

3.6 Summary

In this section, we have addressed **RQ 1** by describing four common perspectives on bias that are relevant in the context of visual analytics, including (1) bias as a cognitive processing error, (2) bias as a filter for information, (3) bias as a preconception, and (4) bias as a model mechanism.

CHAPTER 4
DETECTING BIAS IN VISUALIZATION

After defining bias in the context of visual analytics, the next high-level goal, **RQ 2**, involves developing methods to characterize bias during the analysis process.

> **RQ 2**: *How can human bias be characterized in real-time during the analysis process?*

This question is divided into three parts: defining bias metrics (Chapter 4.1), a formative study to implement and validate the bias metrics (Chapter 4.2), and a study to refine the bias metrics by better understanding what constitutes unbiased behavior (Chapter 4.3).

4.1 Characterizing Bias with Interactive Bias Metrics

This section focuses on the first sub-question of **RQ 2**. It describes work that has been done in response to **RQ 2.1** and has been published as a conference paper [191].

> **RQ 2.1**: *By what metrics can user interaction characterize bias in a person's visual analysis process?*

Visual analytics affords a novel mechanism for measuring and characterizing bias as a result of its interactive nature. 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 [136]. Interactions mark the paths of exploratory data analysis, providing an opportunity to glean insight into a person's reasoning and decision making processes [127, 145].

Hence, we hypothesize that when data analysis is supported by visual analytic tools, analysts' 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.

In this section, we present theoretical foundations for quantifying indicators of human bias in interactive visual analytic systems and propose six preliminary metrics. Here, we adopt the perspective of bias as a model mechanism, as described in Chapter 3.4. 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. In this section, we conceptually apply these metrics to the detection of cognitive biases (i.e., as described in Chapter 3.1); however, future sections describe how the metrics can be used to detect other types of human bias.

4.1.1 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.

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 [91, 182]. 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 [66]. Heuristics frequently ignore or subconsciously

Table 4.1: Cognitive biases relevant to intelligence analysis [79] that produce the measurable behavioral indicators we focus on in this section.

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

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 [79]. 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 [66, 90]. 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

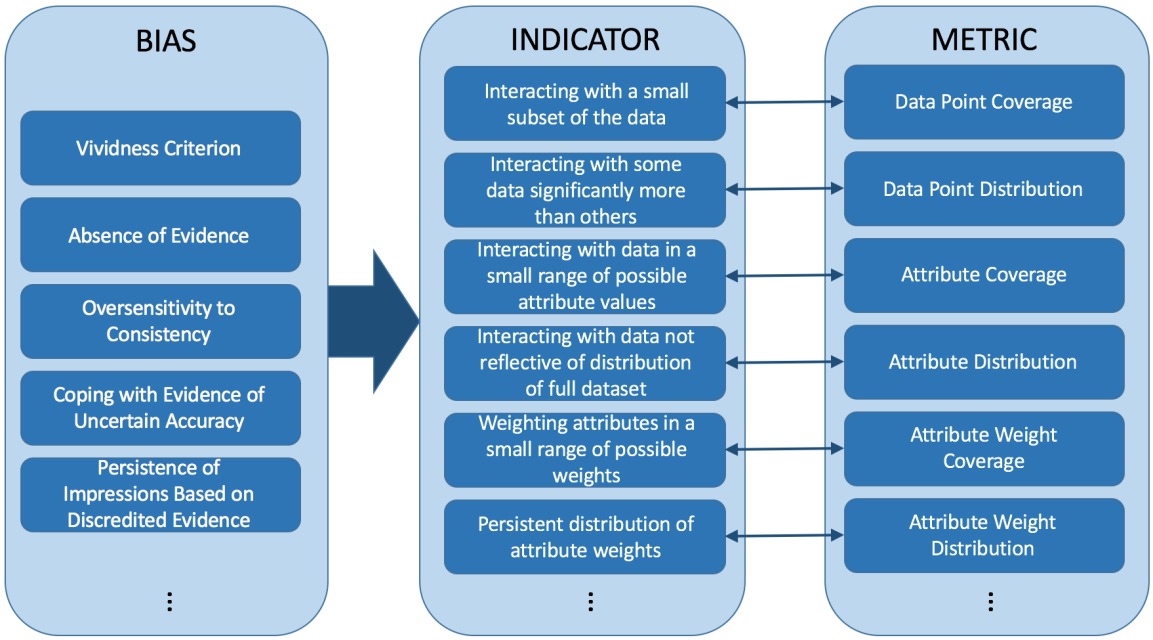


Figure 4.1: Cognitive biases result in behavioral indicators that are measurable by the proposed metrics. We scope this proposal to those indicators and metrics depicted above, but there are numerous other biases, behavioral indicators, and ways to measure those indicators.

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 [79] describes five types of cognitive biases particularly relevant for evaluating evidence, defined in Table 4.1: *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 4.1, 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 section and the set of proposed metrics for measuring indicators of bias in Figure 4.1. The block 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.

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 multi-view interface. Further, the metrics can only account for the dataset loaded in the system. For example, if an analyst is examining a dataset of criminal suspects, the metrics would not be able to infer about a bias toward a person not represented in the dataset.

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 {interaction, data point} 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 dataset 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 4.2. We define $D = \{d_1, \ldots, d_N\}$ to be a dataset 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 dataset, 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.[1]

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

[1]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 dataset sizes, but we leave consideration of those sets to later work.

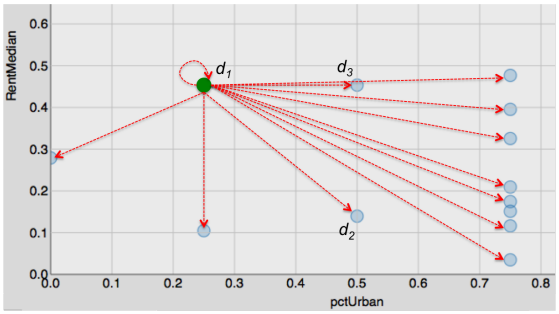
a user has made on a dataset. Coverage is related to the notion of a cover for a set. The cover for $T \cup D$ is a collection of sets whose union contains $T \cup D$ as a subset. In terms of interactions, the cover for $T \cup D$ is the union of all sets of interactions $I(D)$ possible in the analytic process. In information visualization, the concept of coverage has been studied as a means to encourage users to explore more data [43, 53, 67, 156, 197] as well as inform users of collaborators' explorations [11, 89]. 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.

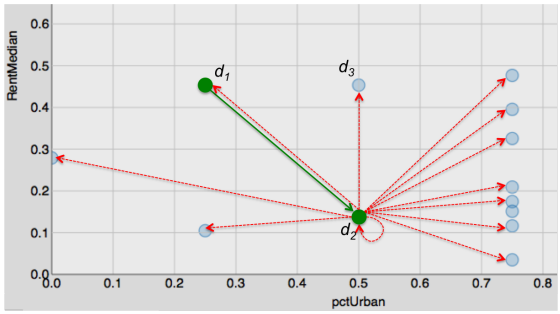
Key to our present interest in modeling evolving behavior as people interact with systems is that we can track the events in $I(D)$ that are created by the user over the course of an analytics session. We propose that by tracking these events as a Markov chain over the state space $T \cup D$, we can 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 $b \in [0, 1]$, 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

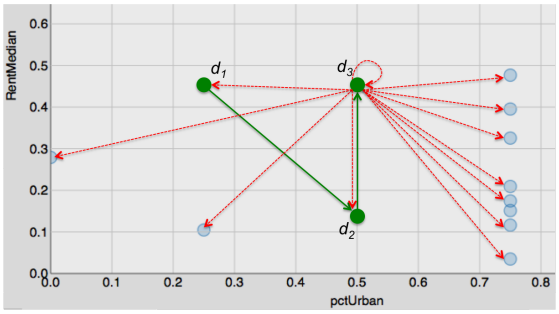
with data point d_i on step $k+1$ is $P[d_{i,k+1}|d_{j,k}] = 1/N$, meaning that each next interaction does not depend on the current interaction or the interaction history. A sequence of interactions in $I(D)$ thus forms a regular Markov chain, with the data points representing the states in the chain with transition probability matrix $P = \left[\frac{1}{N}\right]$. Figure 4.2 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 Figures 4.2a – 4.2d illustrate 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



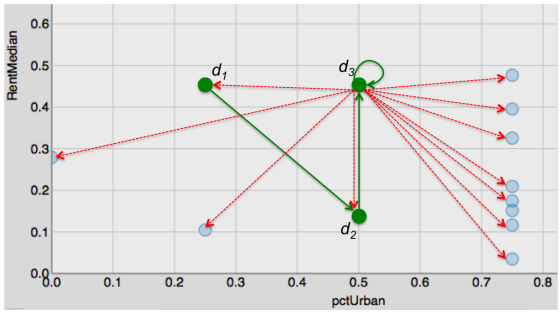
(a) Interaction 1: hover over point d_1; resulting Markov chain $\{\{hover, d_1\}\}$



(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 4.2: 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.

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 in Chapter 4.3.

4.1.2 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 4.1.

Data Point Coverage

Description. The data point coverage metric is an ordinal measure of the user's attention to the data points in the dataset. In particular, it measures the amount of the dataset 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 dataset. 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

Table 4.2: Notation used to describe the bias metrics

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
$D = \{d_1, ..., d_N\}$	dataset of size N
$A = \{a_1, \ldots, a_M\}$	set of M attributes describing dataset 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

the data point coverage is much smaller than would be predicted by the unbiased baseline model?

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 dataset. $\kappa(D_U) \leq \kappa(D)$, and $\kappa(D_U)$ approaches $\kappa(D)$ as the user explores more of the dataset.

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 dataset, 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

$$\hat{\kappa}(D_U) = \frac{N^k - (N-1)^k}{N^{k-1}}. \tag{4.1}$$

We then define the data point coverage metric b_{DPc} according to Eq. 4.2.

$$b_{DPc} = 1 - \min\left(\frac{\kappa(D_U)}{\hat{\kappa}(D_U)}, 1\right) \tag{4.2}$$

Example. To understand how this metric might be useful in capturing behavioral indicators of bias, consider the following. An analyst may propagate their 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* [79], they subconsciously rely more heavily on evidence that is vivid or personal than on evidence that is dull or impersonal. Thus, bias would be propagated through the system by interacting with only a small, vivid subset of the full set of evidence.

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 dataset.

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. 4.3.

$$\chi^2 = \sum_{n=1}^{N} \frac{(\kappa(I(d_n)) - \hat{\kappa}(I(d_n)))^2}{\hat{\kappa}(I(d_n))} \tag{4.3}$$

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. 4.4.

$$b_{DPd} = 1 - p \tag{4.4}$$

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* [79], 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.

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 dataset. 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. 4.5.

$$b_{Ac}(a_m) = 1 - \min\left(\frac{\kappa(D_{U,Q_{a_m}})}{\hat{\kappa}(D_{U,Q_{a_m}})}, 1\right) \tag{4.5}$$

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

$$\hat{\kappa}(D_U, Q_{a_m}) = \frac{Q_{a_m}^k - (Q_{a_m} - 1)^k}{Q_{a_m}^{k-1}}. \tag{4.6}$$

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* [79]. 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 them to only interact with a portion of the range of possible attribute values in the dataset.

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 dataset.

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 dataset 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 dataset 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. 4.7.

$$\chi^2 = \sum_q \frac{(\kappa(a_{m,q}) - \hat{\kappa}(a_{m,q}))^2}{\hat{\kappa}(a_{m,q})} \tag{4.7}$$

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. 4.8.

$$b_{Ad}(a_m) = 1 - p \tag{4.8}$$

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* [79]. 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.

Attribute Weight Coverage

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.

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. 4.9.

$$b_{AWc}(a_m) = 1 - \min\left(\frac{\kappa(W_{U,Q_{a_m}})}{\hat{\kappa}(W_{U,Q_{a_m}})}, 1\right) \tag{4.9}$$

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

$$\hat{\kappa}(W_{U,Q_{a_m}}) = \frac{Q^k_{w(a_m)} - (Q_{w(a_m)} - 1)^k}{Q^{k-1}_{w(a_m)}}. \tag{4.10}$$

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* [79] 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.

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. 4.11.

$$b_{AWd}(a_m) = 1 - p \tag{4.11}$$

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* [79]. 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.

4.1.3 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 thus far theoretical 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.

Generalizing the Metrics

In this section, we discuss some of the factors that were considered in defining the proposed bias metrics and how they may be adjusted to generalize the 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 their 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.

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 (e.g., [28]).

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 their 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. 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.

4.1.4 Summary

In this section, we have addressed **RQ 2.1** by describing six metrics for characterizing bias in visual analytics, including {*coverage* and *distribution*} each computed for {*data points*, *attributes*, and *attribute weights*}.

4.2 Capturing Anchoring Bias with Interactive Bias Metrics

After defining theoretical foundations for bias metrics in visual analytics, we seek evidence that the metrics are able to pick up on analysts' biased behavior. Specifically, this section focuses on the second sub-question of **RQ 2**. It describes work that has been done in response to **RQ 2.2** and has been published as a conference paper [190].

RQ 2.2: *Can bias metrics be used specifically to capture anchoring bias?*

In the previous section, we introduced metrics to quantify bias from user interactions in real-time during the process of visual data analysis. The bias metrics track the interactive process of users with respect to the visualization, data, and analytic model in the system to create a quantitative representation of analytic provenance. The theoretical formulation of the metrics, however, must be validated. Do the bias metrics give a signal when users are engaged in analytic processes known to be influenced by bias? Further, the theoretical formulation of the metrics leaves many open questions regarding the implementation details of the bias metrics in a visual analytic tool, including: (1) which interactions are used in each computation?, (2) how often to compute the metrics?, and so on. We explore these questions in the context of anchoring bias, which describes the tendency for people to rely too heavily on initial information when making a decision [51].

In this section, we present the results of a formative study to examine how bias can be observed in users' interactive behaviors through the lens of the bias metrics, while simultaneously refining our understanding of how to apply the interactive bias metrics in a real scenario. Our goal is to leverage a well-known and highly studied form of bias (anchoring [51, 62]) to influence participants' analysis process in a controlled and predictable way. Specifically, we presented users with one of two alternative task framings, each intended to encourage users to anchor on different attributes of the data. Users were tasked with categorizing basketball players' positions using an interactive visual analytic tool. We analyzed the bias metrics over the duration of each participant session. 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 put the metrics into practice. Our analysis suggests anchoring bias elicited by task framing can be observed in users' interactive behavior through the lens of the computational bias metrics.

4.2.1 Methodology

We conducted a formative study to explore the ways that anchoring bias manifests in interactive behavior during visual data analysis, specifically through the lens of the bias metrics, described in Chapter 4.1. Anchoring bias describes the tendency for people to rely too heavily on initial information when making a decision [51]. In the analytic process, this tendency leads people to preferentially weight some information and systematically neglect other information, often leading to poorly informed decisions. In order to induce such anchoring bias on participants in this study, we utilized framing effects. Framing describes the manner in which a choice is presented to people, including the language used, the context, and the nature of the information displayed [179, 180]. Framing has been found to strongly shape decision-making [174], as people tend to take more risks under negative framing conditions. The way that information, task goals, or context are introduced to people has a strong impact on how they will conduct their analysis. By using framing effects to induce anchoring bias, we are able to evaluate how a well known bias manifests in user interaction patterns for a visual data exploration and classification task.

Participants in the study were tasked with categorizing a dataset of basketball players. Using the visual analytics tool InterAxis [94], shown in Figure 4.3, users were instructed to explore the dataset and label 100 anonymized basketball players according to the position they play. Participants were randomly assigned to one of two different framing conditions, each of which described the five different positions in basketball using different attributes. The goal of using two different conditions was to anchor participants on those specific

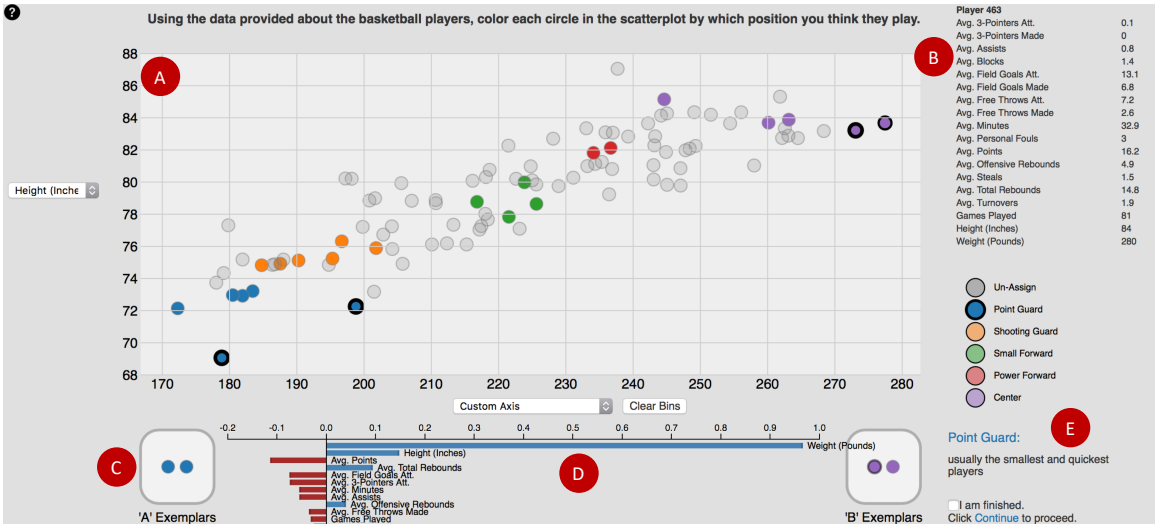


Figure 4.3: A modified version of the system InterAxis[94], the interface used by participants to complete the task of categorizing basketball players. See text for more details.

attributes during their analysis (see Table 4.3). Given the predictable ways anchoring bias influences decision making, this study explores the research question *can the bias metrics be used to characterize interactive behavioral differences when people are anchored on different attributes of the data?*

InterAxis

Participants used a scatterplot-based visual analytics tool to categorize basketball players by their position (Figure 4.3). Pilot studies led us to modify the InterAxis user interface from its presentation in [94] for ease of use in the study. Changes include: the y-axis custom axis options were removed; the color scheme was changed, data point colors were changed to reflect participants' labels; options for saving the plot settings were removed; experiment control options (e.g., Un-Assign label option, Continue button) were added. The data from the pilot was only for testing and feedback on our protocol and not included in the results.

The primary view in InterAxis is a scatterplot, where each of 100 basketball players is represented by a circle (Figure 4.3A). Hovering on a circle reveals details about that

player (Figure 4.3B). Data points can be dragged from the scatterplot into the bins on either side of the x-axis (Figure 4.3C). 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 points in the bin on the high end of the axis and the bin on the low end of the axis. The attribute weights are visualized as bars along the axis (Figure 4.3D). 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 below the detail panel (Figure 4.3E). With one of the positions selected, the user can then label players as the selected position by clicking on the points in the scatterplot.

We chose to use InterAxis for the study due to the system's highly interactive nature – to encourage users to explore and interact with the data, since the bias metrics ultimately rely on user interactions. Further, InterAxis allows users to browse data points and attributes, in addition to using 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 interactive visual data analysis, cognitive processes are often much more complex than can be captured from such experiments. Hence we sought a task with enough complexity to simulate decision making within a realistic analysis scenario while maintaining tractable experimental conditions.

There are many tasks associated with performing data analysis in a visual analytic tool, such as ranking, clustering, or categorizing data [48, 159]. What bias looks like can be quite different across these tasks; hence, for this study we narrowed our scope to focus on

Table 4.3: Position descriptions used in the two framing conditions. *Size* condition participants were expected to rely more heavily on size-related attributes (Height and Weight). *Role* condition participants were expected to rely more heavily on the role-related attributes called out in the description.

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*

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.

We asked participants to categorize a set of 100 basketball players by their positions based on their stats using the InterAxis visual analytic tool [94], shown in Figure 4.3. The study dataset was a subset derived from a dataset of professional (NBA) basketball players[2] whose names and team affiliations were 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) for a total of 100 players. Each player had

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

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 (Inches), and Weight (Pounds).

Participants were assigned to one of two conditions. The two conditions differed in the descriptions provided for the five positions (C, PF, SF, SG, and PG). In the *Size* condition, the descriptions are based on physical attributes (Height and Weight) of players. 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 descriptions of the positions recognized by NBA.[3] Table 4.3 shows the text used to describe the positions in each condition.

Participants

Ten participants (4 female, mean age 25.5 ± 2.7 years) were recruited from a large university. All but one participant had experience playing basketball, and six participants watched at least a few (NCAA, NBA, WNBA) games per season. The one participant who never played basketball watches it regularly. All participants were at least moderately familiar with information visualization. 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 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

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

completing the task, participants were administered a post-study questionnaire about their experience. Participants were compensated with a $10 Starbucks gift card.

Throughout the task, an investigator observed the participant's interactions, taking notes of each participant's strategies (e.g., labeling points the user was most confident about, relying on a particular attribute for their categorizations, using nearby data points to visually categorize an uncertain point, etc). Participants were encouraged to ask questions as needed regarding the interface, the underlying algorithmic transformations, or the meaning of an attribute. As needed, the investigator would prompt the participant to clarify their strategy to understand when shifts in strategy occur. The investigator 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 capture the input data for the bias metrics.

4.2.2 Verifying Anchoring Effects

To see how the bias metrics characterize anchoring bias, we first analyze how framing impacted user behaviors. We verified that the task framings induced an anchoring bias by analyzing behavioral differences between participants in the two conditions. This is an important first step to ensure that the signal detected by the bias metrics during the analysis process can be attributed to anchoring bias. The results of this analysis indicate that participants relied heavily on the attributes used in their respective condition groups.

The two framing conditions, Size and Role, were designed to bias participants in a con-

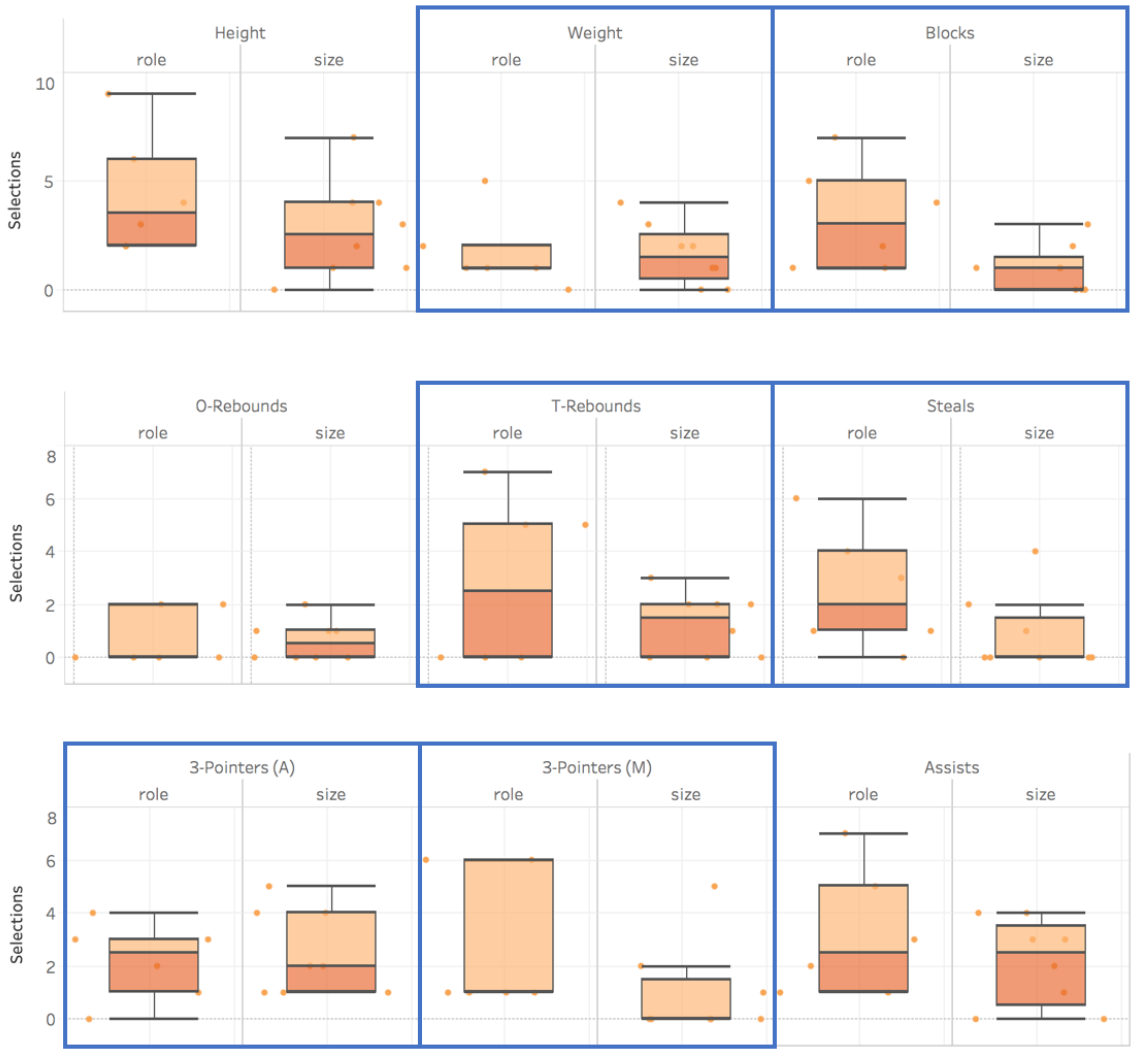


Figure 4.4: Boxplots of number of attribute interactions via axis manipulation in InterAxis. The median is indicated by the thick middle line, the inner quartiles within the box, and the outer quartiles the whisker bars. The red dots indicate the sum of observations for each participant (rather than outliers as in traditional boxplots.)

trolled way. Our prediction is that participants will anchor, or rely more heavily, on the attributes in the particular descriptions used in their condition (Table 4.3). We first compared the frequencies of attributes selected for the axes in the scatterplot between the two framing conditions. We predicted that participants in the Size condition would select the Height or Weight attributes on the axes more than participants in the Role condition. Likewise, we predicted that participants in the Role condition would select the other framed attributes

(Blocks, Rebounds, Steals, 3-Pointers, or Assists) on the axes more than participants in the Size condition.

Figure 4.4 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 quartile values suggests that the framing condition impacted the given attribute, while significantly overlapping boxplots suggest little or no difference between the framing conditions for that attribute. The boxplots reveal that some attribute axis selections show clear differences supporting our predictions (e.g., Height, Blocks, Offensive Rebounds, Steals, and 3-Pointers Attempted), while others are less clearly affected (e.g., Weight, Total Rebounds, and 3-Pointers Made). These results suggest that the participants from the two groups anchored on the attributes described in the respective framing conditions.

We also characterized if participants showed different patterns in their use of the position labels between conditions. That is, anchoring bias could influence their tendency to use one label more than others, in addition to selecting the attributes differently. In categorization models, using one label more than another is often measured with a parameter called bias, referring to the bias for some categories over others. We fit the Generalized Context Model (GCM) to the categorization identification-confusion matrices [128, 130]. The GCM defines the probability of assigning players S_i from group i to position label R_j as $P[R_j|S_i] = \frac{\beta_j \eta i,j}{\sum_{k \in K} \beta_k \eta i,k}$, where $0 < \beta_j < 0$ is the bias for label j, subject to $\sum_j \beta_j = 1$, and $\eta_{i,j} > 0$ is the similarity of player i to all the other members of position j. We fit GCM parameters with a Bayesian MCMC estimate of 10,000 runs, holding attention weights constant, and β_j was given a uniform(0,1) prior. The mean β_j parameter estimates for each category and group are plotted in Figure 4.5. If participants are unbiased in assigning position labels then we would expect all $\beta_j = 0.2$. Figure 4.5 shows β_j are around 0.2 for most positions. The Size condition has less variability in the β_j values, but there is evidence in that condition for a higher β_j for labeling Point Guard. The Role condition produced

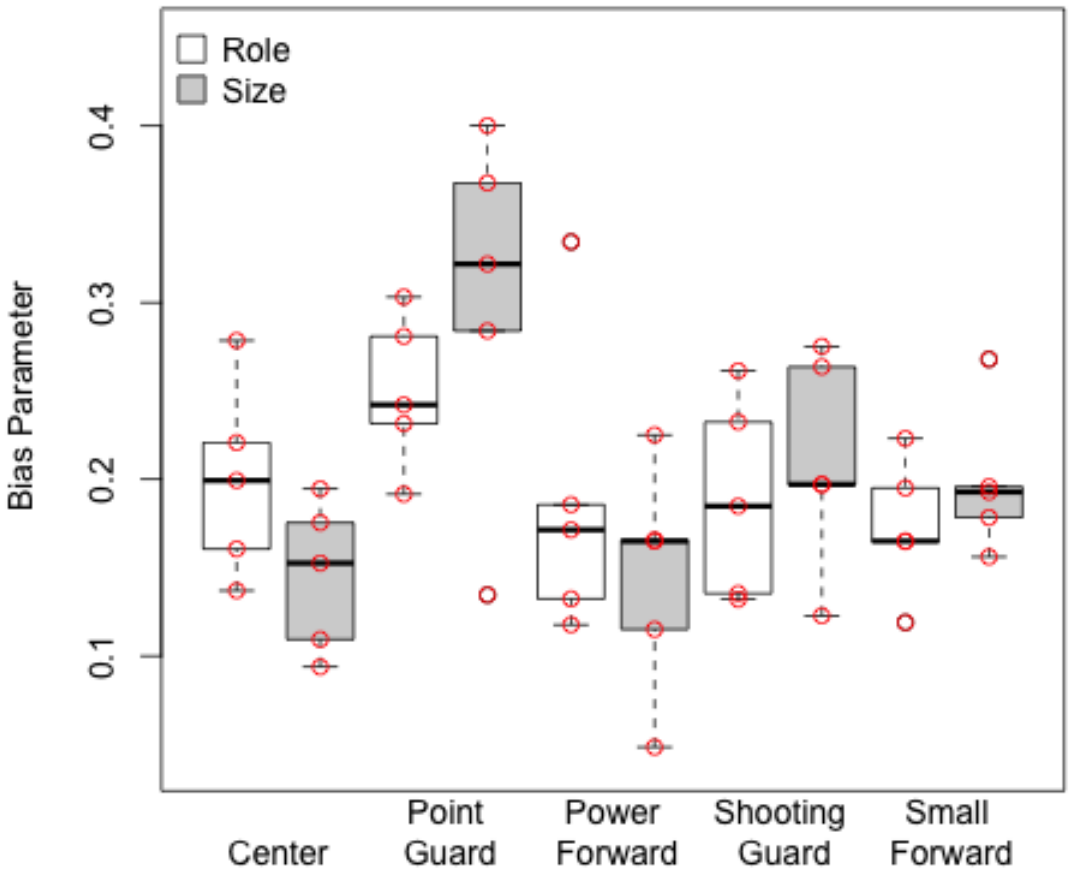


Figure 4.5: Box plots of the GCM bias parameters estimated for each position for the Role and Size conditions. Red points represent the individual participant values.

more variability in category bias, with most participants showing a bias toward assigning the label Point Guard over the other positions. Center and Power Forward show lower bias parameters, suggesting Role condition participants were less likely to assign those as labels to players.

Together, these results confirm that the Role and Size conditions influenced the overall categorization behaviors in ways consistent with our intended manipulations. With this overall evidence for shifting biases and attribute selection patterns, we turn to exploring the interaction patterns to see how these biases are reflected in our novel bias metrics. We note that the critical evaluation of the metrics herein relies on within-subject analysis, because the ultimate application will be quantifying the bias of a single user while performing visual analytic tasks. Consequently, our between-condition analyses remain qualitative as

we compare the within-subject metrics patterns.

4.2.3 Analysis and Results

We analyzed the user study data with the high-level goal of understanding how anchoring bias manifests in participants' behavior through the lens of the bias metrics. 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 analyze 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 deviate from expected behavior. To analyze if the metrics can capture bias, we use the collected interaction logs to simulate the real-time computation of the bias metrics after each user's session in order to avoid influencing the analysis process. We note that the bias metrics create 74 unique time series per participant (DPC + DPD + 18 attributes x {AC, AD, AWC, AWD}). In the scope of this work, we narrow the focus of our discussion to only attributes that were referenced in the position descriptions to analyze if they picked up on the induced bias. We discuss a few selected examples of findings from the computed bias metrics. Visualizations of all metrics can be found in the supplemental materials.[4]

Participants' accuracy for categorizing players averaged 53% ($SD = 18\%$) over the course of 33.6 minutes ($SD = 14$ min). Some interactions were filtered out to reduce noise in the bias metric computations. Namely, hovers less than 100 ms were removed as likely "incidental" interactions performed unintentionally while navigating the mouse cursor to a different part of the interface. Because hovering in the interface shows a data point's details, particularly short hovers were likely not intentional interactions by the user to get information. 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,

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

see the Discussion section.

Metrics over Interaction Sequences

After every interaction, each bias metric computation results in a value between [0,1] quantifying the level of bias at that time. Computed over time, the metrics produce a sequence of values quantifying the level of bias throughout the analysis process which can be visualized as a time series. One way to test whether bias metrics capture anchoring is to look for differences in these sequences between the two conditions.

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* condition) will have higher metric values in the associated condition than in the other. For example, we expect the time series of AD values for Assists for Role condition participants to be higher than the values for Size condition participants. To address this question, we visualized the time series for each of the 74 metrics.

Figure 4.6 shows the 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. Visual examination of Figure 4.6 finds 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.

This visual trend can also be observed 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 $M_{\text{Role}} = 0.0849$ ($SD = 0.042$).

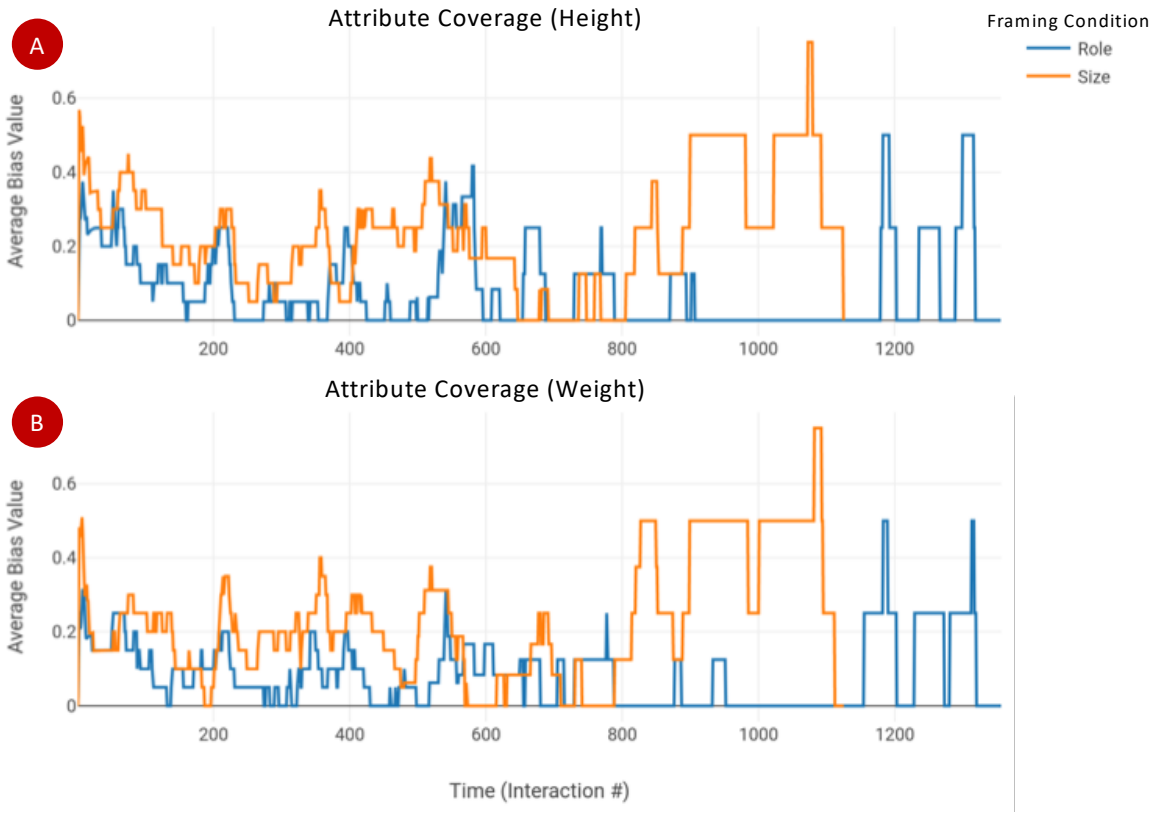


Figure 4.6: A visualization of the average Attribute Coverage (AC) metric for the attributes (A) Height and (B) Weight for participants in each condition. Size condition participants (in orange) tended to have higher AC bias for Height and Weight than Role condition participants (in blue), consistent with our predictions.

This evidence supports our hypothesis; however, not all metrics show a discernible difference 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 mostly on particularly tall players might have the same metric value as a user focusing mostly on particularly short players. That is, simply knowing the metric value captures presence of a bias in the *coverage* or *distribution* of the data, attributes, or attribute weights; however, the number itself does not differentiate the source within the data distributions. In the next section, we address this by looking not just at the $[0, 1]$ metric values, but the underlying coverage or distribution that comprises that computation.

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. One way to illuminate the framing effects on user interaction patterns is to compare the metrics broken down into components of coverage and distribution rather than just the summative $[0, 1]$ values. Thus, in this analysis we visualized this 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. That is, a black cell in the DPD metric for one participant may represent a different number of interactions than a black cell in the DPD metric for a different participant. The scales are defined in each plot.

Figure 4.7 shows what the metrics DPD, AD for Assists, and AWD for Assists look like 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 represents the type of position being labeled during that time period (blue = PG, orange = SG, green = SF, red = PF, and purple = C). 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, where darker shades represent a greater number of interactions.

In Figure 4.7A, the y-axis shows a row for each data point to illustrate DPD. The DPD metric shows more bias toward players who are PGs while attempting to label PGs (2) than while attempting to label SFs (1), consistent with correct categorizations. This can visually indicate the user's bias toward particular players based on their interactive behavior. In Figure 4.7B, 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 (1) than while labeling Centers (2), consistent with Role framing. In Figure 4.7C, the y-axis illustrates the breakdown of attribute weight ranges (AWD) into four quantiles. The AWD metric for Average Assists

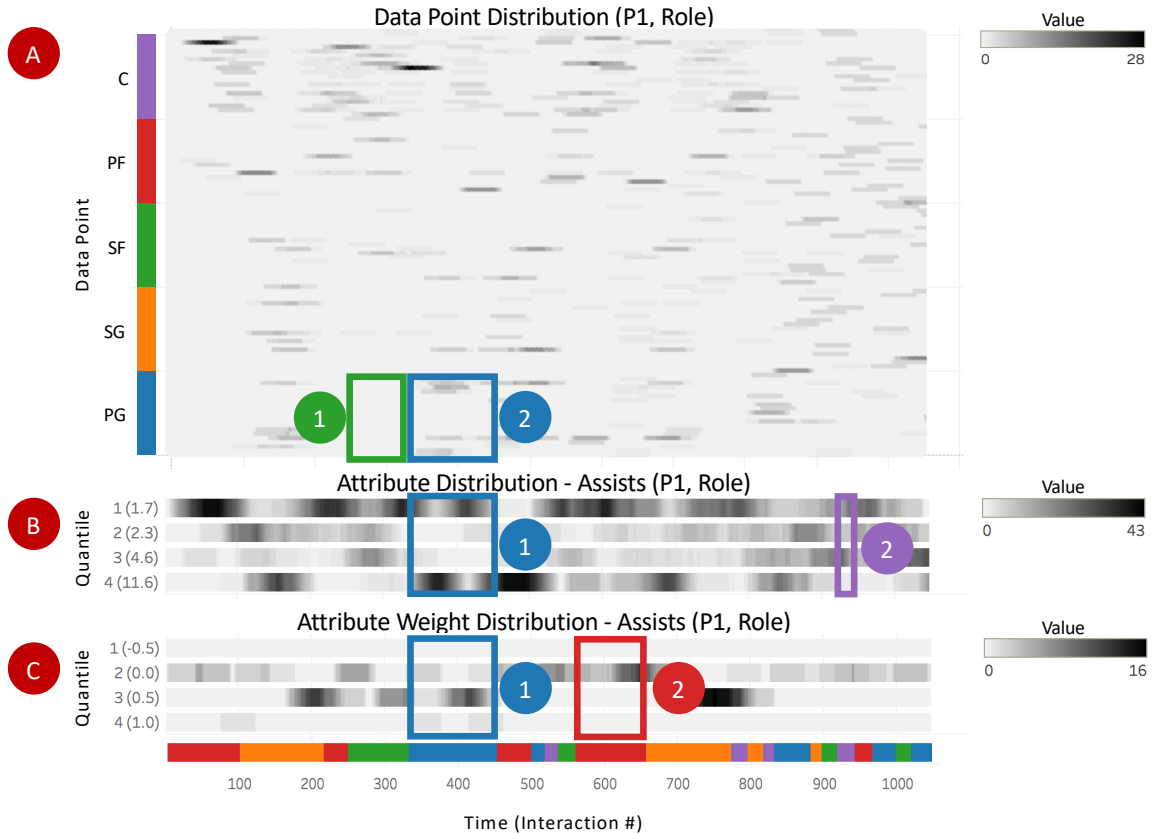


Figure 4.7: Visualizations of three of the bias metrics for a Role condition participant: (A) the DPD metric, (B) the AD metric for Average Assists, and (C) the AWD metric for Average Assists. While labeling Point Guards (PG; blue boxes), compared to labeling other positions (SF = green boxes, C = purple boxes, PF = red boxes), the participant exhibited more bias toward PG players (A) and the Assists attribute (B) and (C) from the Role condition PG description.

indicates a bias toward higher weighting of the attribute while labeling Point Guards (1) than while labeling Power Forwards (2). The Role condition Point Guard description is intended to influence participants to anchor on the Average Assists attribute. Hence, Figure 4.7B and Figure 4.7C visually capture a user's anchoring bias toward an attribute.

Figure 4.8A visually compares 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.

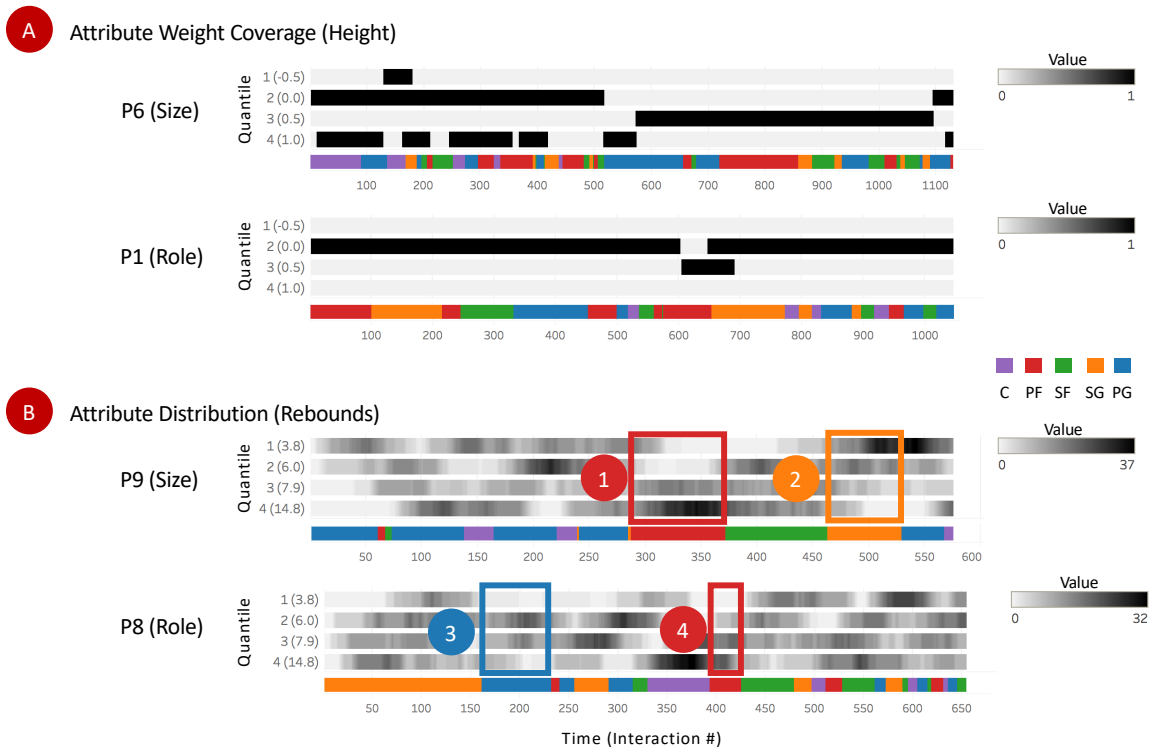


Figure 4.8: (A) Visualization of the AWC metric. The Size condition participant (top) showed more *coverage* of the range of Height attribute weights than the Role condition participant (bottom). (B) Visualization of the AD metric for Total Rebounds. Participants focused more on upper parts of the Rebounds distribution while labeling PFs (red boxes) than other positions.

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 [31]. 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 quantiles. For the Size participant (top), $L = 14.9$ indicating more switching, and $L = 229.8$ for the Role participant (bottom), reflecting a very long time in a single quantile which is seen in Figure 4.8A. Heatmaps for all metrics and all participants can be seen in the supplemental material.

Similarly, Figure 4.8B shows how AD for Average Total Rebounds compares for one Size condition participant (top) and one Role condition participant (bottom). Role con-

dition participants were told that PFs 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 size (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 quantiles 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 quantiles.

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

4.2.4 Applying the Bias Metrics

This study constitutes the first application of the bias metrics described in Chapter 4.1, and explores how to analyze them for 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, includ-

ing 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

Fisher et al. [56] describe visual analytics as a "translational cognitive science." That is, they posit that visual analytics is a cognitive science that must travel between the often-disjointed worlds or pure science and design. In the present work, these worlds collide in the design of tools intended to measure bias from user interaction. Science motivates the interaction design to best externally reflect internal cognitive processes, while design focuses on creating a seamless and enjoyable user experience.

Designing a visualization system often involves understanding potential user needs, including things like ease-of-use, learnability, or powerful 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 design goals. While incorporating bias computation and visualization in an interface has the potential to promote better analysis behaviors, it ultimately relies on interpreting user interactions as a meaningful capture of 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 [94] 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 interaction difficult from the perspective of the bias metrics. Further, participants would be less likely to interact with individual 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. Similarly, Dimara et al. [40] conducted a study in which they had users click to locally delete individual data points on a scatterplot until a single point remained (as opposed to clicking to select the single data point in the decision). While this showed promise for mitigating bias (specifically, the attraction effect), it came at the cost of a tedious user experience.

Such trade-offs must be considered when integrating bias metrics into practical tool design. And they raise important questions for future research, such as: if the interaction design in 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 or lacked meaning. 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 a user's 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: Chapter 4 presents a formulation of metrics for characterizing

bias based on prior user interactions; however, it does not inform us *when* to compute the metrics or *how many prior interactions* should be computed on. In our 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 (Figure 4.3E) 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. While an objective characterization of bias, the value b itself is not actionable. 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 a heatmap (Figures 4.7–4.8). 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.

4.2.5 Discussion

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 and zoom and pan on the scatterplot, different clusters of points or outliers might draw the user's attention. In future work, we would like to explore baselines that account for visual salience to better model unbiased behavior. 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 baseline models of unbiased behavior 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 X 791 interactions per participant, in addition to overall metrics like task accuracy. While this sample is large enough for our present analysis, ten participants

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 a cognitive phenomenon (bias) in the context of a real-world problem (using a visual analytic system for categorization and decision making). 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 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 interface interactions could be substantial. We may need to identify meaningful ways to incorporate time on task into the metric computations.

Temporal Interaction Weighting

In the previous section, we discussed the impact of different windowing techniques for computing the bias metrics. One potential improvement on these variations would be to come up with 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 interactions that were 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.

4.2.6 Summary

In this section, we have addressed **RQ 2.2** by conducting a study to assess whether anchoring bias can be detected using the bias metrics described in the previous section. We presented the results of a study where participants were assigned to one of two conditions for a categorization task using a visual analytic system. Comparing the two conditions, we found that user interactions interpreted through the bias metrics captured strategies and behaviors reflecting the manipulated anchoring bias. To produce a stronger signal of bias, we posit that baseline models of unbiased behavior may need to be refined.

4.3 Refining Interactive Bias Metrics

We defined bias metrics and validated that they can be used to characterize anchoring bias from user interactions in a scatterplot. Now, we seek to revisit one of the fundamental assumptions made in the theoretical formulation of the metrics: that at any point in the analysis, the user is equally likely to interact with any data point. We refine the metrics by trying to better understand what users' unbiased behavior looks like. Specifically, this section focuses on the third sub-question of **RQ 2**. It describes work that has been done in response to **RQ 2.3** and was published as a short paper at IEEE VIS [189].

> **RQ 2.3**: *How can the bias metrics be refined to more accurately account for unbiased interactive behavior?*

In this section, we analyze the assumptions made about how to model unbiased behavior in the metrics described in Chapter 4.1. The baseline of unbiased behavior was theorized as a Markov model, where each combination of {data point, interaction type} constitutes a unique state. However, "unbiased behavior" was initially suggested to be represented as equal probabilities between all states in the Markov model. This assumes randomness in the user's interactive behavior, which we posit is an unreasonable assumption for most tasks and interfaces. Hence, we experimentally challenge the assumption of equal probabilities of interactions by exploring people's actual interaction sequences as they analyze data.

To refine the metrics, we replicate the study conducted in Chapter 4.2, again using anchoring bias with the basketball player categorization task. Anchoring bias describes the tendency for people to rely too heavily on initial information when making a decision [51]. However, while in the previous study we characterized anchoring bias by analyzing the *differences* in user behavior between two task framing conditions, in this study, we characterize unbiased behavior by analyzing the *commonalities* in user behavior between the two task framing conditions.

From recorded interactions, we derive a Markov model representing users' observed

interactive behavior across two bias conditions. Our analysis indicates that, rather than equal probabilities of all interactions, people's interactions can be better modeled roughly based on the proximity of data points. That is, people are more likely to interact with nearby data points than those that are far away.

4.3.1 Experiment Methodology

We conducted a study to explore the assumptions of "unbiased" behavior in the bias metrics. In the original formulation of the bias metrics [191] and subsequent experiment [190], users' interaction sequences were compared to unbiased behavior defined by *equal probabilities for all interactions*. However, we believe this assumption is likely ill-fit for most tasks and interfaces. In recent work, researchers constructed a theoretical Markov model based on size of data points (pixel area on the screen) as an approximation for probability of interaction [35]. We are motivated by such work to create a more precise model of unbiased user behavior based on experimental observations.

We hypothesize that *proximity* can be used to better model user behavior. That is, people will be more likely to interact with nearby data points than far away data points, by starting with what they know (the initial anchoring information) and expanding their analysis, analogous to local exploration of graphs [137]. To test this hypothesis, we replicated the experiment conducted in Chapter 4.2, summarized below, but refocused data analysis to examine probabilities of interaction sequences. Participants were randomly assigned to one of two task framing conditions, designed to *anchor* them on specific attributes of the dataset. They were tasked to utilize all of the data to categorize 100 anonymized basketball players by position (Center, Power Forward, Small Forward, Shooting Guard, or Point Guard) using InterAxis [94] (Figure 4.3). To our knowledge, there is no known way to explore truly "unbiased" or perfectly neutral user behavior. Users will be impacted by the framing of the task, prior biases and experiences, etc. Hence, we approximate unbiased behavior by examining the commonalities between two groups of participants who are biased

in a controlled way.

InterAxis

Participants utilized a scatterplot-based visualization tool, InterAxis [94], the same version of the tool used in the experiment in [190]. In the dataset of basketball players, each player is represented in the scatterplot by a circle (Figure 4.3A), where details (statistics including Height, Weight, Rebounds, Free Throws, etc.) about a player can be seen on the right (Figure 4.3B) by hovering over a circle in the scatterplot. The axes of the scatterplot can be manipulated by selecting from a drop-down, or by dragging points into the bins on the left and right sides of the x-axis (Figure 4.3C). The system then computes a weighted combination of attributes representing the difference between the points in the bins. The weights can be further manipulated by dragging the bars beneath the x-axis (Figure 4.3D). Users can click one of the colored circles on the right (Figure 4.3E) to display a description of that position. Subsequently clicking on a point in the scatterplot will color and categorize that player accordingly.

Analytic Task & Framing Conditions

As in the previous study [190], we likewise focus on the task of data categorization. Participants were tasked to categorize 100 anonymized NBA basketball players[5], 20 players for each of the five positions: Center (C), Power Forward (PF), Small Forward (SF), Shooting Guard (SG), and Point Guard (PG). Participants were not shown the name or team of the players, but were given the following statistics: 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 (Inches), and Weight (Pounds).

Participants were randomly assigned to one of two conditions. In each condition, we

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

manipulated task framing [179] to impact users' analysis in a controlled way. The two sets of position descriptions in the task were designed to *anchor* participants on a specific set of attributes or statistics in the data (Figure 4.3E). Participants in the *Size* condition were shown descriptions of the five positions that used statistics about their physical size (i.e., Height and Weight), while participants in the *Role* condition were shown descriptions that used statistics associated with their typical role on the court. For full experimental details, including the specific language used in each framing condition, as well as additional analyses, please refer to supplemental materials [6].

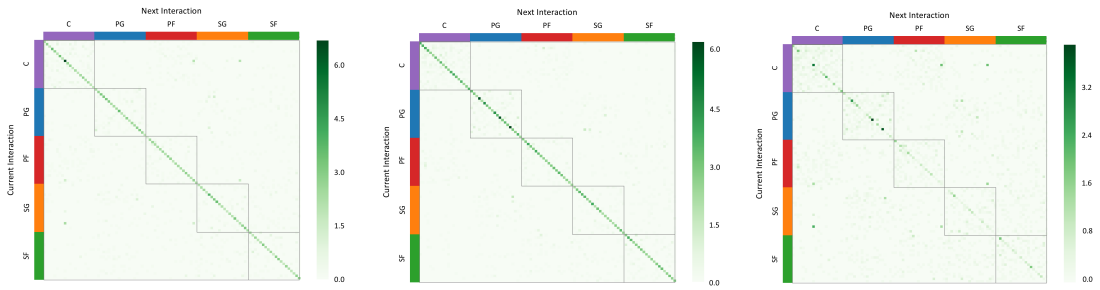
Participants

We recruited 13 participants to complete our study (7 in the Size condition). Eligible participants completed a screening questionnaire to demonstrate sufficient background knowledge about the domain (basketball) and visualization literacy (scatterplot interpretation) [18, 106]. There was no compensation to participants in the study.

Procedure

The procedure for this experiment followed the same as in [190], with differences detailed below. Participants provided informed consent and completed two questionnaires (demographic & interface usability). They were shown videos demonstrating how to use InterAxis. Different from the procedure in [190], participants in this study were given the opportunity to get accustomed to the interface for 5 minutes with a small dataset of 15 cars to be categorized by type (as either sedan, SUV, or sports car); they were also shown a refresher video on basketball positions. The main task took approximately 15-20 minutes, during which interactions in the interface were logged. Different from [190], we collected one additional piece of information in the interaction logs to aid our analysis: the locations of all data points at the time of each interaction. In total, the experiment took about 45

[6] `https://github.com/gtvalab/bias-markov`



(a) Role Condition **(b)** Size Condition **(c)** Size - Role

Figure 4.9: Aggregate probability transition matrices by condition. Rows (current interaction) and columns (next interaction) represent each of 100 basketball players, grouped by position. The highlighted squares along the diagonals indicate subsequent interactions with the same player position. Darker squares indicate higher probabilities.

minutes.

4.3.2 Data Analysis and Results

For simplicity in an initial model, we aggregated all interaction types (click, hover, drag) with a data point into a single Markov state. Next, we filtered out some interactions. Hovers and drags less than 100ms were likely accidental interactions [124], while the user passed from one intentional point to the next; so we removed those interactions. Participants performed, on average, 1043 interactions ($SD = 390$) which filtered down to an average of 527 interactions ($SD = 148$). Participants had an average categorization accuracy of 54% ($SD = 12\%$). Two participants (P12 and P13) did not label all 100 players in the scatterplot. They categorized $\frac{89}{100}$ and $\frac{97}{100}$, respectively. Next we describe and visualize the probabilities resulting from our analysis.

Comparing Conditions

Figure 4.9 shows aggregate matrices representing the probability of interacting with subsequent players in the scatterplot. Rows indicate the "current" interaction, and columns represent the "next" interaction. Hence, a cell is colored darker according to the probability of interacting first with the associated "current" player and then with the "next"

player, where players in each matrix are ordered by their position. We see similar patterns across both conditions. Namely, there is a strong trend along the diagonal. That is, there is approximately a 50% chance that from a given state (player interaction), users next transition will remain in the same state (interact with the same player again), regardless of the condition (50.04% for role condition, 54.74% for size condition). The difference matrix between the two conditions is shown in Figure 4.9c, revealing near-0 differences between most transition probabilities in the two conditions (98.5% of transition probabilities < 0.1). Collectively these results suggest similar transition probabilities between states, regardless of condition.

Proximity Analysis

In this analysis, we wanted to approximate the probability of interacting with visually nearby data points. To do so, we defined new Markov states by dividing the scatterplot into equal size grids (Figure 4.10): $2x2$ (4 states), $3x3$ (9 states), and $4x4$ (16 states), and assessed the probability of interacting with points within and between these fixed grid squares. We chose to use a fixed grid overlay for our analysis in order to examine proximity even when the position of individual points on the dynamic scatterplot may be changing. From the previous analysis, we know that multiple interactions with the same player are significantly more likely (e.g., hover on a player then click to label). Hence, in this analysis, we remove subsequent interactions with the same player to see if interactions with differ-

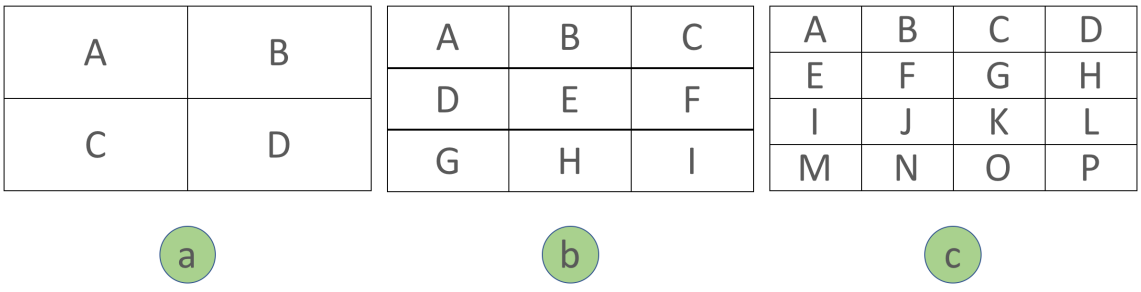
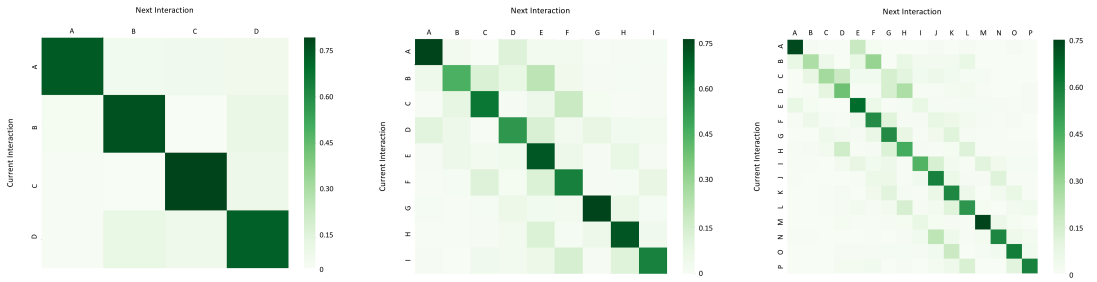


Figure 4.10: Interactions within the scatterplot were grouped into states in the Markov model by dividing the scatterplot into (A) a 2x2 grid, (B) a 3x3 grid, and (C) a 4x4 grid.



(a) $2x2$ grid (4 states) **(b)** $3x3$ grid (9 states) **(c)** $4x4$ grid (16 states)

Figure 4.11: Aggregate probability transition matrices of all participants when Markov states are defined by grouping points in the scatterplot in a $2x2$, $3x3$, and $4x4$ grid. Darker squares indicate higher probabilities.

ent basketball players tend to still follow trends of proximity. Furthermore, we observe no significant difference between conditions, so here we present results aggregated for all 13 participants. Figure 4.11 shows the results of this analysis. We observe the hypothesized pattern of proximity: users are more likely to interact with other data points within the same grid square (i.e., nearby data points) than data points in different grid squares (i.e., far away data points). This is evident by the stronger colors and hence higher probabilities along the diagonal. In $Markov_{2x2}$, we find that nearby interactions (diagonal probabilities in Figure 4.11) comprise, on average, 75.3% of subsequent interactions. Similarly, in $Markov_{3x3}$ and $Markov_{4x4}$, we find nearby interactions to comprise 64.36% and 54.29% of subsequent interactions, respectively. Apart from subsequent interactions within the same grid square (higher diagonal probabilities), we also observe a trend in $Markov_{3x3}$ and $Markov_{4x4}$ parallel to the diagonal, indicating that people often perform subsequent interactions with *adjacent* grid squares.

A New Baseline

Results of our experiment suggest that users are more likely to interact with nearby data points than far away data points when performing a categorization task with an interactive scatterplot. How do we now incorporate this information into a new probability matrix that represents a baseline of unbiased behavior?

We tend to favor simple models or modifications over more complex ones, with modest changes to the equal-probability baseline. Hence, we propose that in the context of our experimental task, a more accurate baseline of unbiased behavior could adjust from the equal-probability baseline by distributing interaction probabilities such that subsequent interactions with the *same* data point comprise roughly 50% of interactions from any given state. We could likewise account for proximity by grouping points in grid squares (as in Figures 4.10-4.11) and defining probabilities of subsequent interactions within each grid square (nearby interactions) as at least 50% of interactions from any given state, according to the grid size chosen.

4.3.3 Discussion

Explaining Unbiased Interaction Sequences

This experiment provides a more accurate baseline of unbiased behavior in the context of our tool, dataset, and analytic task. However, we posit that these results may not be especially generalizable. Higher probability of interactions with a specific quadrant of the dataset could be explained by the structure of the task. For instance, because the player descriptions tended to point users to a specific part of the distribution (i.e., the tallest players, the players with the highest number of Assists, etc.), interactions with the high end of the axis likely all occurred within a given quadrant. With all else equal, a slightly different problem framing may likely have yielded a vastly different baseline model. Hence, it is important to account for the specific context of a problem when defining a baseline, including the tool, task framing, and so on. Our experiment provides a model by which more accurate baseline models can be derived through pilot studies for interfaces that may utilize the bias metrics [190, 191].

Other Notions of Proximity

In this work, we focused on understanding how proximity can be used to model users' interactive behavior. However, we only roughly estimated proximity by grouping interactions into Markov states based on a grid pattern. The purpose of this choice was the ease with which it could be computed using a Markov model. Future work could consider other notions of proximity (e.g., measure the precise pixel distance between points).

Future Models

While the current study focused on analyzing data from the perspective of proximity, there are many other variables that could impact user behavior. Future work could include an examination of how aspects of visual salience [115] impact interactive behavior (e.g., default size of data points in the scatterplot, variable encodings using hue or opacity, etc).

Overfitting

There are numerous ways to model unbiased behavior, as mentioned above. However, a common danger among them is to create models that are overfit to user data from a single experiment. Hence, we must exercise caution in how we define or alter models of unbiased behavior, keeping in mind that often the simplest approaches work best. The next step given the current work to improve the baseline is to implement and compare it against other potential baselines of unbiased behavior to see how well the resulting metrics are able to detect deviations in user behavior in real-time.

4.3.4 Summary

In this section, we have addressed **RQ 2.3** by conducting a study to observe how users actually interact with visualizations. We replicated the study conducted in Chapter 4.2, where participants performed a categorization task. We approximated unbiased behavior

as the commonalities between participants in two different framing conditions. As a result, we were able to refine the model of unbiased behavior used in the bias metrics.

CHAPTER 5
MITIGATING BIAS IN VISUALIZATION

The final question, **RQ 3**, focuses on how to leverage the bias metrics in the design of visualizations to mitigate biased decision making.

> **RQ 3**: *Can bias metrics be used in visual analytic systems to mitigate bias?*

This question is divided into two parts: defining a design space of bias mitigation for visualization (Chapter 5.1) and evaluating one strategy, visualizing interaction traces (Chapter 5.2).

5.1 Designing Bias Mitigation Strategies

Once we have metrics to characterize a user's bias during visual data analysis, the next high-level goal, **RQ 3**, involves exploring ways to utilize that characterization to ultimately mitigate biased decision making. Specifically, this section focuses on the first sub-question. It describes work that has been done in response to **RQ 3.1** and was published as a short paper at IEEE VIS [196].

> **RQ 3.1**: *How can an interface visually communicate the characterization of a user's bias?*

As interactive visualizations are increasingly used for data analysis and decision making in widespread domains, these processes can be improved by designing systems that can both leverage analysts' cognitive strengths and guard against cognitive limitations and weaknesses, including biases. In this section, we focus on **deriving a design space for visualization systems that can mitigate bias**.

Prior work detailing bias mitigation, or debiasing techniques (Chapter 2.5), has largely relied on non-technological strategies, like training courses [77, 79]. However, as data analysis increasingly takes place through technological media, particularly using visualization, we are motivated to consider ways in which visualization design can improve decision making processes. While some prior work has provided guidelines toward mitigating one type of bias in a particular context [40, 68], we take a more general approach aimed at increasing real-time awareness of bias abstracted from a specific scenario. Given the recent emergence of bias mitigation in visualization (and hence relatively little work done in this area), our design space is derived from (1) prior work describing bias mitigation strategies outside of the visualization community, as well as (2) potential areas of visualization research that may inform the design of systems that mitigate bias.

Toward this goal, we must make a key assumption: that systems have information about bias in the user's decision or analytic process. Prior work has developed techniques that make this assumption reasonable. For example, computational methods exist for quantifying bias in the analytic process [68, 190, 191]. Given this information, or other forms of de-biasing information, the goal is then to design systems that can help people make better decisions by compensating for the ways in which people are likely to make cognitive errors.

Many different types of biases can have common observed behavioral effects [100, 191]. For example, an analyst subject to vividness criterion [79] (over-reliance on information that is vivid or personal) may interact with a particularly vivid data point repeatedly. The same behavior is likely to be observed if the analyst is subject to a different type of cognitive bias, like the continued influence effect [79] (continued reliance on a piece of evidence, even after it has been discredited). As a result, some mitigation strategies can, to varying degrees, have an effect on multiple types of biases [10]. Hence, in this design space, we do not focus on any specific type of cognitive bias. Rather, we find it prudent to introduce design considerations for mitigating bias and improving analytic *processes*,

agnostic to a specific type of bias.

Within this context, the contribution of this work is the derivation of 8 dimensions of vis design that designers should consider when developing systems to mitigate biased decision making, or retrofitting such capabilities in existing tools. These dimensions represent aspects of a visualization system that can be manipulated in specific contexts to mitigate biased decision making. We concretize these dimensions through examples using a hypothetical VA system, `fetch.data`, to illustrate potential bias mitigation interventions. While prior work on bias mitigation in the context of visualization and visual analytics is limited, we find it timely to scaffold design efforts going forward when building systems that can mitigate biased decision making.

5.1.1 Driving Areas in Visualization Research

While prior work on mitigating cognitive bias in the visualization domain is sparse [40, 68], we are motivated to define a design space in this emergent area. Hence, in the context of visualization research, we derive inspiration from sub-fields of visualization research that may be leveraged to mitigate biased decision making processes.

Guidance

According to Ceneda et al., guidance can be defined as "a computer-assisted process that aims to actively resolve a knowledge gap encountered by users during an interactive VA session"[26]. In other words, systems that guide users provide some form of assistance during interactive data analysis (e.g., Scented Widgets [197] in collaborative settings or VizAssist [17] for visualization creation).

Bias mitigation in VA can be loosely thought of as a form of guidance, where the goal is to impact the user's decision making in such a way as to promote a more balanced analytic *process* and / or a more reasonable *product* or final choice decision. Within Ceneda et al.'s [26] characterization of guidance, we focus on the *output means* in the context of bias

mitigation. What can we show the user to facilitate an analytic process that is less prone to the potentially negative effects of cognitive bias?

Analytic Provenance

Analytic provenance is a description of the analytic process leading to a decision [127]. Many researchers have shown the impact of raising users' awareness of their process. Researchers have shown ways to measure or visualize the user's *coverage* of the data throughout analysis [11, 89], leading users to make more discoveries [197] and analyze the data more broadly [53, 104]. This body of research shows promise that provenance awareness can alter user behavior in the context of bias mitigation.

Mixed-Initiative VA

Mixed-initiative [83] VA tools explore the balance between human and machine effort and responsibilities. Some systems leverage users' interaction sequences to infer about their goals and intentions in an analytic model (e.g., [19, 194]). These types of mixed-initiative tools inspire potential ways of mitigating cognitive bias as people use visualizations. In particular, the machine could operate as an unbiased collaborator that can act on behalf of the user, or *take initiative*, to mitigate biased analysis processes.

5.1.2 Design Space

In this section, we describe 8 dimensions (D1-D8) important to the design of bias mitigation strategies in VA tools. These 8 dimensions are not strictly orthogonal, nor are they exhaustive. Rather, they represent our view on the aspects of visualization systems that may be manipulated for the purposes of mitigating bias given current technologies. Due to limited prior work on bias mitigation in VA, the process for deriving this design space was largely ad-hoc, guided primarily by literature review in driving areas of vis research (Section 5.1.1). Many of the dimensions are related (Figure 5.1).

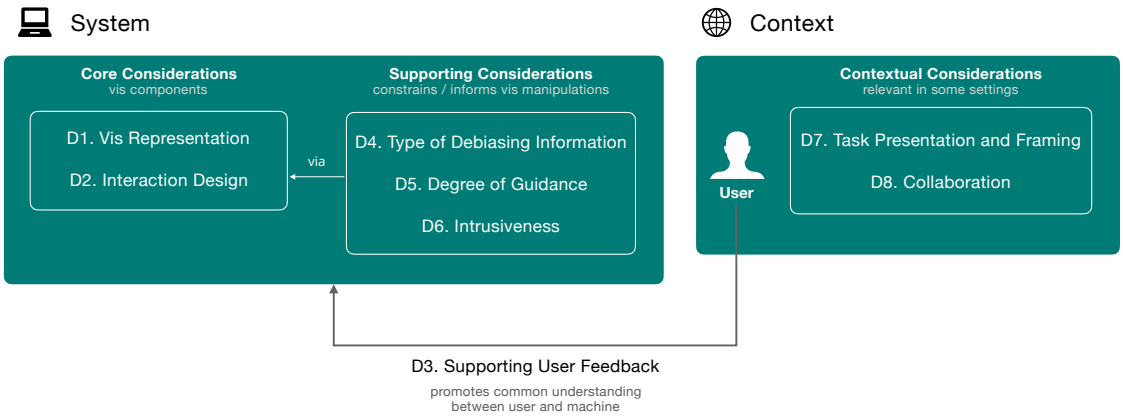


Figure 5.1: The design space is comprised of 8 dimensions, described in Section 5.1.2. D1 (VISUAL REPRESENTATION) and D2 (INTERACTION DESIGN) are the two *core* components of a visualization [55] that can be manipulated to mitigate biased decision making processes. How these components are manipulated is informed and constrained by *supporting* considerations, including D4 (TYPE OF DEBIASING INFORMATION), D5 (DEGREE OF GUIDANCE) and D6 (INTRUSIVENESS). Some *contextual* considerations may only be relevant in specific settings, including D7 (TASK PRESENTATION AND FRAMING) and D8 (COLLABORATION). Finally, D3 (SUPPORTING USER FEEDBACK) connects the user and contextual setting to the system by promoting a common understanding between user and machine.

To ground our design space, we describe applied examples using a common scenario. Suppose a hiring manager at a tech company uses a VA tool, `fetch.data` (Figure 5.2, top) to analyze tabular data about job applicants. From potentially hundreds of applications on file, the hiring manager wants to select a handful of candidates to interview. Suppose the system is comprised of three interactive views: (A) a scatterplot view, (B) a filter panel, and (C) a ranking table view. Scatterplot axes can be configured, the table sorted, and filters used to adjust the subset of data viewed. For each design dimension below, we describe the concept and revisit this example to illustrate how a visualization could be retrofitted to mitigate biased decision making.

D1: Visual Representation

Concept. There are many possibilities for representing information that may have a debiasing effect. For bias interventions intended not to impose significant disruption to the user's natural analytic process, designers may opt for *peripheral* or *ex-situ* visualizations. Periph-

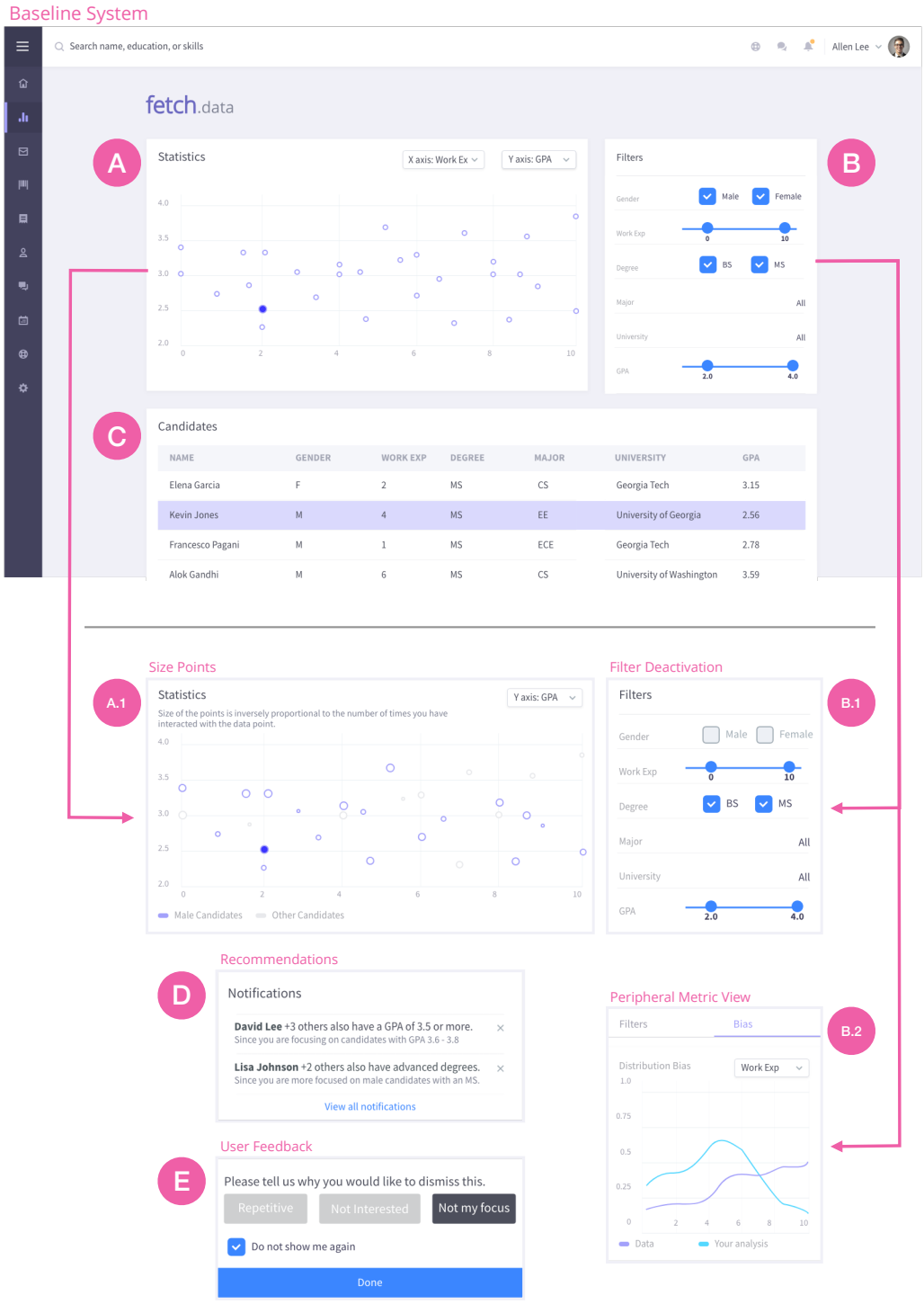


Figure 5.2: An illustration of the system, `fetch.data`, used to analyze tabular data about job applicants. The baseline system (top) consists of (A) a scatterplot view, (B) a filter panel, and (C) a ranking table view. Possible bias mitigation interventions are shown below. (A.1) sizes candidate data points based on the analyst's interactions with them. (B.1) shows the system disabling the gender filter. (B.2) shows a peripheral view of metrics quantifying bias. (D) shows recommendations for candidates the analyst has not yet examined. (E) shows a pop-up allowing the analyst to provide feedback when dismissing a notification.

eral visualizations would appear in a separate view of the interface, potentially available on demand, and hence may be less likely to call the user's attention away from the primary visualization. On the other hand, *in-situ* visualizations would appear within existing views. For example, in-situ visualizations could encode debiasing information in previously unused visual channels (e.g., opacity, color, position, etc). The choice between in-situ v. peripheral display of debiasing information should be informed by (1) type of debiasing information, and (2) intended level of user attention to that information. Furthermore, the representation of information should follow conventions described in vis research. For example, chart types [153] and optimal visual encodings [30] should be considered based on the type of data presented.

Example. Suppose the hiring manager is subject to anchoring bias, or the tendency to rely too heavily on initial "anchoring" information [51] (in this case, the first few résumés received). If the first handful of candidates happened to be males, successful bias mitigation strategies could draw the user's attention away from potential gender bias. Some metrics of bias (e.g., those described in Chapter 4.1) compare the distribution of user interactions to the underlying distributions of the data. This could be shown in a *peripheral* view showing both distributions (Figure 5.2, B.2). Alternatively, a single metric quantifying the severity of the bias could be encoded as an ambient background display where color or opacity represents level of bias. In another example, history (provenance) could be shown *in-situ* by encoding size of scatterplot points as time spent examining each candidate, drawing attention to those (female candidates) who may have been unintentionally ignored (Figure 5.2, A.1).

D2: Interaction Design

Concept. Altering the interaction design may be another impactful way to mitigate bias. For example, a designer's choice between a rectangle or lasso selection may have implications about how a user approaches a problem or task. Similarly, a system could disable

interaction with data / views when biased behavior is detected. However, altering interaction design and affordances to mitigate bias can often come at the expense of perceived user control and system usability. Designers of bias mitigation interventions should weigh the tradeoffs of these choices so usability is not unduly compromised.

Example. Consider a filtering widget designed to mitigate bias. If the hiring manager applies a filter to exclude female candidates in the data, a typical system response would be to remove female candidates from the views in the visualization. The system could instead respond by presenting a split or duplicated scatterplot view: one in which the manager's intended data is shown (male candidates), and one in which the filtered data is shown (female candidates). Alternatively, the system could disable interactions with filters for which a bias is exhibited (Figure 5.2, B.1).

D3: Supporting User Feedback

Concept. While the primary objective of bias mitigation interventions is to communicate information from the system to the user, supporting user feedback is likewise important. In real-world systems that may be able to characterize user bias with limited accuracy, it can enable the user to communicate information outside the scope to the underlying model of bias (e.g., that a presumed bias is not due to unconscious error, but rather an external task constraint). When user feedback is supported, users may be given an increased sense of mutual understanding or common ground with the system. Further, models of user bias might be improved as a result.

Example. In the hypothetical hiring scenario, suppose the system detects a strong (gender) bias in that the hiring manager has primarily interacted with male candidates. One system response could be to recommend female candidates (Figure 5.2, D). However, the hiring manager's focus could be the result of a constraint on the task unknown to the system (e.g., a division of labor between two managers). If the manager dismisses the recommendation of female candidates, the system can elicit feedback (e.g., via a pop-up dialog) to clarify

information potentially outside the system's purview (Figure 5.2, E). Reasons may include things like a repetitive recommendation, an irrelevant recommendation, or an external task constraint. According to the hiring manager's selection, the system may alter the underlying model of bias to account for these preferences or constraints.

D4: Type of Debiasing Information

Concept. A primary consideration in designing bias mitigation strategies is the type of debiasing information that the system will capture and communicate to the user. Types of debiasing information that could promote user awareness includes things like analytic provenance, summative metrics that quantify the analytic process [54, 88], and so on. We could further conceive of future systems that are able to identify specific types of bias the user may be subject to by name (e.g., confirmation bias [126], anchoring bias [51], etc). Systems should ideally communicate information about potential biases in a way that guides users to counteract them (i.e., they should be *informative* and *actionable*).

Example. Suppose the hiring manager is exhibiting signs of availability bias [181], or a heavy reliance on information that is most easily remembered or most recent (i.e., the most recent application received). When bias is detected (i.e., the hiring manager is exhibiting signs of availability bias), the system could show provenance information to the hiring manager by adding an additional view to the interface that shows a snapshot of various stages of history of the manager's analytic process (e.g., like the history shown in [78]). Alternatively, the system could show the results of summative interaction metrics, similar to the metric visualization in [191] (Figure 5.2, B.2). This could enable the hiring manager to reflect on their process and adjust.

D5: Degree of Guidance

Concept. Degree of guidance is analogous to Ceneda et al.'s *guidance degree* in VA guidance [26]. It can be thought of as a spectrum that refers to how much the system

"helps" the user. On one end of the spectrum, the system provides little intervention, while on the other end, the system more aggressively steers the user. Ceneda et al. describe three scenarios for degrees of guidance: *orienting*, *directing*, and *prescribing*, examples of which are described below. The degree of guidance adopted must be considered alongside tradeoffs of user experience. Systems that deny user control may come at the expense of perceived usability issues.

Example. An *orienting* bias mitigation strategy would promote user awareness of their biases. For example, the system could size candidates in the scatterplot according to the hiring manager's focus (where larger points represent neglected candidates; Figure 5.2, A.1). A *directing* bias mitigation strategy could suggest candidates to the hiring manager to consider from the pool of candidates who have not been analyzed (Figure 5.2, D). A *prescribing* bias mitigation strategy would involve the system assuming initiative or otherwise taking control from the user. An example of this might be disabling filters or interactions with specific candidates (Figure 5.2, B.1).

D6: Intrusiveness

Concept. Intrusiveness refers to how much the system interrupts or otherwise intrudes on the user's analysis process. On the low end of the spectrum, bias information may be presented peripherally or even on demand (i.e., user attention optional). Highly intrusive mitigation strategies may present information front and center requiring the user's attention until the perceived bias is addressed. This is akin to the distinction between reactive (e.g., system responds only when prompted by the user) v. proactive systems (e.g., system makes suggestions to the user) [198]. Proactive interventions would necessitate higher intrusiveness. The level of intrusiveness of the intervention should not outweigh the intended benefit, however. In lower-cost decisions (e.g., analyzing a dataset of food to construct a weekly menu), a highly intrusive bias mitigation strategy would likely be unwelcome to the user. On the other hand, the intrusion may be acceptable for decisions that carry greater

importance (e.g., criminal intelligence analysis). This dimension is distinct from D5 (DEGREE OF GUIDANCE). Consider the following analogy: suppose a person asks her friend for directions from point A to point B. The friend may draw a map, suggest GPS, or walk her friend there herself (i.e., DEGREE OF GUIDANCE). If she walks with her friend, she may exhibit a spectrum of INTRUSIVENESS (e.g., how closely does she stand to her friend).

Example. In our hypothetical scenario, a minimally intrusive mitigation strategy may present bias information to the hiring manager only *on-demand*. For example, there may be a tab in the interface that reveals information about the model of user bias when clicked on (Figure 5.2, B.2). A more intrusive bias mitigation strategy could be a pop-up notification that repeatedly alerts the hiring manager until a less biased analysis state is reached (Figure 5.2, D).

D7: Task Presentation and Framing

Concept. Changes to the presentation of information can have an impact on the analytic process and outcome. Framing has been found to strongly shape decision-making [174], including richness of language used and positive v. negative terminology to describe logically equivalent information [179]. For instance, in one study, researchers showed that people chose one treatment (surgery) over another (radiation therapy) when it was described as having a 90% short-term survival rate v. a 10% immediate mortality rate [116]. In addition to language, visual framing or anchoring can also shape decision making [29]. In situations where designers of bias mitigation interventions have control of the task, thoughtful consideration should be given to the often subtle-seeming aspects of task presentation.

Example. This contextual consideration is primarily limited to situations in which the designer has control over the presentation of the task (e.g., in a user study). In our hypothetical scenario, the job description can impact the analysis process. For example, the framing of criminal background criteria may alter the hiring manager's threshold for minimally viable candidates (e.g., the negative framing "does not have a criminal record" may

lead to a lower decision threshold than the positive framing "has a clean record"). Visual framing of information can also impact decision making (i.e., the relative size and spatial arrangement of multiple views, the order in which the hiring manager is trained to use them, etc).

D8: Collaboration

Concept. Collaborative contexts have potential to mitigate bias by allowing others to check an analyst's work. By leveraging "wisdom of crowds", collaboration helps to ensure that a single sub-optimal individual decision does not prevail [82, 143]. Analysts teaming on a project may be alerted to biased behaviors, to ensure they cross-validate each other's work. In this case, prior work on fostering awareness in collaborative settings can be informative [11, 13, 76]. Collaboration is contextually relevant, as it may be infeasible in many scenarios due to the nature of the decision (e.g., a personal healthcare decision) or other constraints (e.g., division of labor).

Example. To leverage collaboration to mitigate biased decision making, designers of the vis tool could show traces of other hiring managers' exploration behaviors. For example, this could entail coloring points in a scatterplot based on which have been previously examined by other hiring managers (e.g., [11]), to promote social accountability.

5.1.3 Characterizing Existing Systems

Two recent works have designed interventions within visualization systems to mitigate cognitive bias [40, 68]. For each, we describe the context of the problem, the bias intervention, and how it fits within the aforementioned design space.

Mitigating Selection Bias

In analyzing high dimensional data sets, many dimensions may exhibit correlations. Hence, when attempting to select a sample from a larger dataset, the analyst may unintentionally

filter out a representative part of a population (i.e., selection bias) [68].

To mitigate selection bias, Gotz et al. modified a visualization tool, DecisionFlow. Specifically, they modified an existing view in the visualization (D1, in-situ) by adding a color-coded bar after each subsequent data selection to depict the similarity of the subset to the original dataset. The color-coding of the bar was based on a computed value (D4, bias metric) that quantified the differences in variable distributions between the two datasets. They also added a secondary view (D1, ex-situ) that provided details about how variables of the data were constrained either via direct or unintentional filtering via correlation. These modifications represent an *orienting* degree of guidance (D5) that is relatively unintrusive (D6). They did not modify the interaction design (D2), task presentation (D7), or collaborative nature (D8) of the system, and did not enable user feedback (D3).

Mitigating the Attraction Effect

Dimara et al. designed an experiment to test two different strategies for mitigating the attraction effect (the phenomenon where a person's decision between two alternatives is altered by the introduction of an irrelevant third option) in scatterplots [40].

In one strategy, they highlighted optimal choices with a brightly colored stroke (D1, in-situ) before users clicked to select their choice point. This constitutes an *orienting* degree of guidance (D5). In another design, they altered the task framing (D7) and interaction design (D2) from "select a point" to "eliminate points until only one remains". While this was more effective than the first strategy, it could have usability implications as it represents a more intrusive (D6) design. For both strategies, they do not support user feedback (D3) or collaboration (D8). By virtue of these mitigation strategies taking place within an experiment, the debiasing information (D4) was a precondition to the study.

5.1.4 Discussion

This design space does not exhaustively include all possible contextual design considerations when building visualization systems that can mitigate biased decision making. For example, the device type may drive design choices that are compatible with varying input modalities or screen sizes (e.g., haptic feedback or other non-visual channels when screen real-estate is limited). Mitigation strategies may also be adaptive to the type of user of the system (casual user, domain expert, data analyst, etc). Furthermore, while we have focused on improving decision making *processes*, agnostic to a specific type of bias, there may be more targeted mitigation strategies that address a specific type of bias. Some of these limitations could be overcome by future systematic literature review (e.g., revisiting ad-hoc dimensions).

Choices within this design space must be balanced with potentially conflicting design considerations. For example, higher levels of INTRUSIVENESS may mitigate bias, but at the expense of user frustration in using the system. In addition, we have assumed that the TYPE OF DEBIASING INFORMATION is given a priori. However, the collection of this information within a system may necessitate its own design considerations. Systems that compute bias metrics based on user interaction sequences (e.g., [189, 190]) will have constraints on VISUAL REPRESENTATION and INTERACTION to ensure that the user's interactions adequately capture their cognitive process. Hence, this may conflict with bias mitigation strategies that involve altering that design.

5.1.5 Summary

In this section, we have addressed **RQ 3.1** by describing a design space of considerations that should be made when creating bias mitigation solutions. The design space consists of 8 dimensions, related to the *core* components of the vis that can be manipulated (VISUAL REPRESENTATION and INTERACTION DESIGN), *supporting* considerations that drive the design (TYPE OF DEBIASING INFORMATION, DEGREE OF GUIDANCE, and IN-

TRUSIVENESS), *contextual* considerations that are only relevant in some scenarios (TASK PRESENTATION AND FRAMING and COLLABORATION), and keeping the user in the loop by SUPPORTING USER FEEDBACK.

5.2 Evaluating a Bias Mitigation Strategy

After designing different ways of utilizing the bias metrics for mitigation, it is next important to evaluate the effectiveness of these strategies. Specifically, this section focuses on the second sub-question of **RQ 3**. It describes work that has been done in response to **RQ 3.2** and is in preparation for submission to IEEE VIS [195].

> **RQ 3.2**: *How effective is the visual representation of interaction traces in an interface toward mitigating biased decision making?*

As the sheer volume and ubiquity of data increases, data analysis and decision making are increasingly taking place within digital environments, specifically facilitated by interactive visual representations of data. These environments provide a new way to measure and characterize cognitive processes: by analyzing users' interactions with data. Analyzing user interactions can illuminate many aspects about the user and their process, including identifying personality traits [20], recovering a user's reasoning process [44], and most relevant to the present work, quantifying human biases [191]. This work utilizes the power of user interaction with the goal of **increasing users' awareness of potential biases** that may be driving their data analysis and decision making. In particular, we examine the effectiveness of **visualizing traces of users' interactions** toward increasing awareness of bias.

To assess the impact of these interaction traces, we designed an interactive scatterplot-based visualization system (Figure 5.3). We conducted two formative studies and one empirical study in which users utilized this interface in a political decision making scenario. We curated a dataset of fictitious politicians in the U.S. state of Georgia and asked participants to select a committee of 10 responsible for reviewing public opinion about the recently passed Heartbeat Bill, banning abortion in the state after 6 weeks. In this scenario, several types of bias may impact analysis, including gender bias (e.g., bias favoring one gender over another), party bias (e.g., voting along political party lines, regardless of po-

tential ideological alignment from candidates in another party), age bias (e.g., preferential treatment of candidates based on age), and so on. Note that we do not aim to address overt biases or discrimination in this work; rather, we believe visualization *can* have an impact on increasing user awareness of potentially unconscious biases that may impact decision making in critical ways.

The primary contributions of this work are results of two formative studies and one empirical study in which we analyze the effectiveness of visualizing interaction traces toward increasing user awareness of social bias. We analyze two interventions: visualizing interaction traces in *real-time* while analyzing the data and in a *summative* view after completing the task. Our findings suggest that visualizing interaction traces, particularly in a summative format after the analysis process, is a promising way to increase users' awareness of bias. Furthermore, when decisions are complex (e.g., with high dimensional data), real-time visualization of interaction traces may lead users to choose political committees that are more proportional to the underlying dataset, specifically with respect to gender bias. In the following sections, we present a description of the dataset and interface used in the studies, specifics of the methodology, findings from two formative studies and one empirical study, and a discussion of how these results can inform the design of future systems that can mitigate potentially biased analyses.

5.2.1 Design Motivation

To increase user awareness of potential biases driving their decision making, we are motivated by literature in cognitive science on nudging [173] and boosting [74], that can influence people's behavior and decision making by altering the choice architecture (i.e., the way that choices are presented). We apply this analogy in the context of visualization with the goal of "nudging" users toward a less biased analysis process. In visualization research, prior work has shown some ability to impact user behavior, resulting in more broad exploration of the data (e.g., by coloring visited data points differently [53] or by

adding widgets that encode prior interactions [197]). Furthermore, we are inspired by work on reflective design [158], wherein our purpose is not to prescribe an optimal decision to users, but rather to encourage thoughtful reflection on motivating factors of those decisions while users maintain full agency.

5.2.2 Methodology

To study the effect of visualization of interaction traces toward mitigating social biases, we selected a political decision making task that was recently relevant and might elicit multiple types of social biases (e.g., gender bias, political party bias, etc). We conducted two formative studies (whose analysis was exploratory in nature) and an empirical study that had a common task, visualization system, dataset, and procedure.

Task

The USA has a two-party political system: Democrats and Republicans [5]. In these studies, we focus on a political decision making scenario in the state of Georgia. In Georgia congress, committees may be formed to explore complex issues, draft legislation, and make recommendations [65]. Many such committees, particularly subcommittees around specific issues, may be formed by top-down appointment [65]. With membership in committees often decided by an individual or by few, the decision can be subject to an individual's biases.

In May 2019, Georgia's incumbent Governor Brian Kemp signed a bill banning abortion after 6 weeks (earlier than the previous state law of 20 weeks) [151]. Scheduled to take effect in January 2020, the bill was received with significant controversy [1]. Supporters of the abortion bill hoped its effect would culminate in an overturning of Roe v. Wade (US federal court decision protecting a woman's right to an abortion, 1973), while opponents hoped to challenge the bill before it became law. This series of studies leverages this controversial context to study bias intervention strategies via increasing user awareness of bias.

[1] A judge ruled in favor of an injunction to block enforcement of the bill in October 2019 [39]

Specifically, given a dataset of fictitious politicians, participants were instructed to *select a committee of 10 politicians responsible for reviewing public opinion in the state of Georgia on the recent controversial Heartbeat Bill*. We selected this task to simulate a realistic decision making scenario common in American politics. Furthermore, this topic and dataset can potentially elicit numerous types of social biases (e.g., gender bias, political party bias, age bias, etc), both explicit and implicit. The instruction for forming the committee was intentionally vague to avoid suggesting any particular criteria to participants.

System

Overview. We developed a visualization system intended to increase user awareness of potential social biases in decision making. To assess the effectiveness of the interface, we produced two versions: a Control (C) version of the interface, and an Intervention (I) version of the interface, in which the Control interface was augmented to visualize traces

Figure 5.3: The interface used in these studies. The primary view is an interactive scatterplot (A). Hovering on a data point populates a detail view below (B). Participants can add a data point (politician) to their list of committee members on the right (C). Data can be filtered according to categorical (D) and ordinal & numerical attributes (E). As the user interacts with the data, their interaction traces are visualized in the top right in real-time, comparing the distribution of the user's interactions to the underlying dataset (F).

of the user's interactions with the data in real-time. The Intervention version is depicted in Figure 5.3; components A-E are common across the Control and Intervention interfaces. The primary view is an interactive scatterplot (A), where the x- and y-axes can be set to represent attributes of the data via selection in a drop-down menu. Hovering on a point (politician) in the scatterplot populates the detail view (B), which shows all of the attributes of that politician. Clicking on the point in the scatterplot or on the star icon in the detail view adds the politician to the list of selected committee members (C). Selected politicians are shown in the scatterplot with a thick red border. Categorical attributes (e.g., gender, occupation, etc) can be filtered in the panel on the left (D) with drop-downs, and ordinal & numerical attributes (e.g., age, policy views, etc) can be filtered on the bottom left (E) using range sliders.

Interaction Traces. In the Intervention interface, the user's interaction traces are shown in the interface in two ways: with respect to *data points* and with respect to *attributes*. First, the points in the scatterplot are given a blue fill color once the user has interacted with the politician, with darker shades representing a greater number of interactions (Data Point Distribution metric, Chapter 4.1; Figure 5.3A). The Control interface, by comparison, uses no fill color on the points in the scatterplot. Second, the top right view (Figure 5.3F) compares the user's interactions to the underlying distributions of the data for each attribute. The attribute tags are colored with a darker orange background when the user's interactions deviate more from the underlying data and with a lighter orange or white background when the user's interactions more closely match the underlying distribution of data (Attribute Distribution metric, Chapter 4.1). Numerical attributes (age pictured) show a gray curve representing the underlying distribution of data and a superimposed blue curve representing the distribution of the user's interactions (primarily with younger politicians). Beneath the curve, the distribution of numerical attributes is broken down into four quartiles (Attribute Coverage metric, Chapter 4.1) and colored according to whether the user has interacted with data in each quartile. Categorical attributes compare user interactions

to the underlying dataset using bar charts.

Technologies. We developed the tool using the Angular 7 framework [8] for the web interface and D3.js [38] and Vega-Lite [187] to render the visualizations. We developed the server in Python 3 and leveraged Socket.IO [161] for real-time, bidirectional communication with the web interface (user interactions sent to the server to compute bias metrics, and the computed bias metrics sent back to update the visualization in real-time).

Dataset

We generated an artificial dataset of fictitious politicians. Each row in the dataset represents a politician, where each is described by attributes such as GENDER, POLITICAL PARTY, OCCUPATION, etc. Variations of the dataset were used in each study, shown in Table 5.1 and described in the relevant sections. The Python script to generate the datasets is included in supplemental materials[2].

[2] `https://github.com/gtvalab/bias-mitigation-supplemental`

Table 5.1: Attributes describing the fictitious politicians in each of three studies. The names are sampled from U.S. census data [146]. The distributions of biographical attributes in the Formative Study 2 and Main Study columns are based on those found in the 115th U.S. House of Representatives [118].

	Attribute	Formative Study 1 (X)	Formative Study 2 (Y)	Main Study (Z)
Biographical Attributes	Name	Sampled randomly by gender		
	Party	50% Democrat; 50% Republican	46% Democrat ; 54% Republican	
	Gender	50% Female; 50% Male	Female (28% if Democrat; 12% if Republican); Male (72% if Democrat; 88% if Republican)	
	Occupation	25% each: Career Politician, Doctor, Lawyer, Business	26% Career Politician; 24% Business Person; 17% Lawyer; 11% Educator; 7% Judge; 3% Financier; 3% Doctor; 3% Farmer; 2% Military; 2% Engineer; 1% Minister; 1% Scientist	38% Lawyer; 23% Career Politician; 21% Business Person; 9% Educator; 5% Scientist; 4% Doctor
	Education	-	4% High School; 2% Associate's; 25% Bachelor's; 22% Master's; 5% PhD; 38% Law; 4% Medical, constrained by Occupation	-
	Religion	-	88% Christian; 6% Jewish; 2% Mormon; 1% Muslim; 1% Hindu; 2% Unaffiliated	
	Age (Years)	-	Sampled from normal distr. with μ = 58 years, σ = 10 years	
	Experience (Years)	33% each: Low, Medium, High	Sampled from normal distr. with μ = 9 years, σ = 3 years	
Policy Attributes	Ban Abortion After 6 Weeks	33% each: In Favor, Neutral, Opposed	-/+ 3, constrained by party: D (-) R (+)	33% each: In Favor, Neutral, Opposed
	Legalize Medical Marijuana	-	+/- 3, constrained by party: D (+) R (-)	-
	Budget for Free School Lunch	-	+/- 3, constrained by party: D (+) R (-)	-
	Increase Gun Control Legislation	-	+/- 3, constrained by party: D (+) R (-)	-
	Ban Alcohol Sales on Sundays	-	-/+ 3, constrained by party: D (-) R (+)	-
	Increase Budget for Medicare	-	+/- 3, constrained by party: D (+) R (-)	-
	Increase Budget for VA	-	+/- 3, constrained by party: D (+) R (-)	-

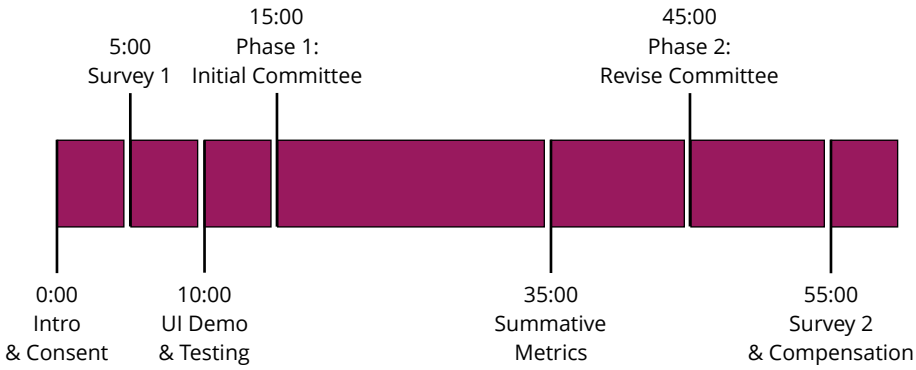
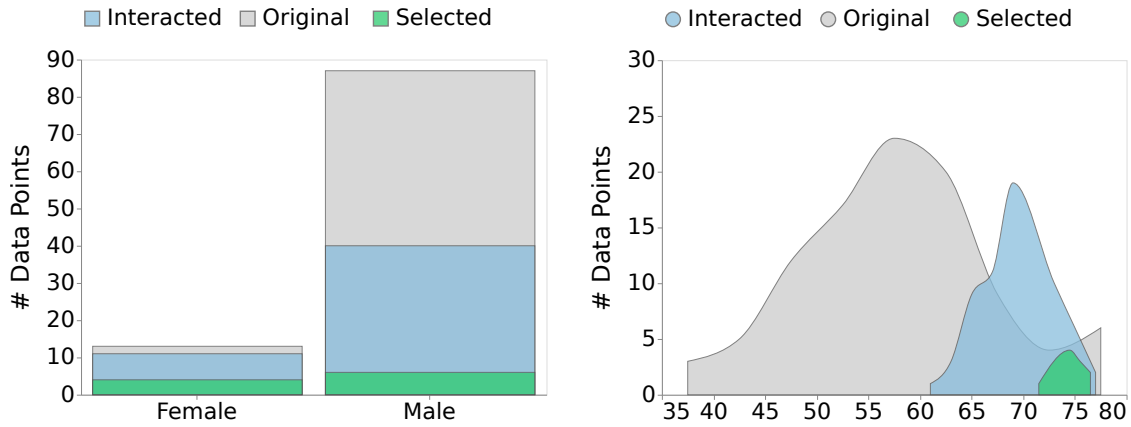


Figure 5.4: Typical timeline for both formative studies and the main study.

Procedure

Participants in the user studies were randomly assigned to use one of two versions of the tool: Control (baseline) or Intervention (visualizing interaction traces in real-time). Each study took approximately 45 minutes to 1 hour, divided as shown in Figure 5.4. The study administrator first obtained informed consent, then participants completed a background questionnaire. Participants were shown a demonstration video of the interface using a cars dataset and then given the opportunity to practice by *choosing a shortlist of 5 cars they would be interested to test drive*. Participants then performed the main task of the study.

There were two high-level phases of the main task. In the first phase, participants chose an initial group of 10 politicians to form their committee. Meanwhile, their interactions with the data were logged (axis configurations, filter interactions, click and hover interactions with data points, and so on). Next, users were shown a summative visualization intervention (e.g., Figure 5.5) that depicted, for each attribute of the dataset: the underlying distribution (gray), the distribution of the user's interactions (blue), and the distribution of their committee (green). Then, based on any imbalances the participant observed, they were given the opportunity to reflect and revise their committee if desired.



(a) Categorical attribute (GENDER)

(b) Numerical attribute (AGE)

Figure 5.5: An example of the summative metric view shown to participants after choosing their initial committee. The distribution of the dataset is shown in gray, the user's interactions in blue, and their selected committee members in green.

5.2.3 Formative Study 1

The goal of this formative study was to understand the bias baseline – that is, to what degree do things like gender, political party, and so on, impact people's decision making in this scenario? In this study, all participants used the **Control** version of the interface, described in the previous section, to complete the task. Our hypothesis was that people would focus *explicitly* on 1-2 attributes of the data, while other attributes may be sources of unintentional *implicit* bias.

Dataset

We created a dataset of 144 politicians containing one politician with each unique combination of GENDER (Male, Female), PARTY (Republican, Democrat), OCCUPATION (Doctor, Lawyer, Business, Career Politician), EXPERIENCE (Low, Medium, High), and the policy view BAN ABORTION AFTER 6 WEEKS (Opposed, Neutral, In Favor). Names for each politician were generated based on U.S. census data [146].

Participants

We recruited 6 student participants from a large university (3 female, 3 male). 5 of the participants self-reported that they most identified with the Democratic party, while 1 most identified with the Republican party. All 6 indicated they were opposed to banning abortion after 6 weeks. We discuss the limitation of participant sampling bias (specifically for political party affiliation) in Section 5.2.6. Participants self-reported an average of 3.7 out of a 5-point likert scale for familiarity analyzing data using visualizations (1 = least familiar, 5 = most familiar). We refer to participants from this study as X01-X06.

Results

Participants rated the importance of attributes that influenced their committee choices on a scale from 1 (least influential) to 7 (most influential). Participants indicated that they most explicitly relied on attributes such as the policy view BAN ABORTION AFTER 6 WEEKS ($\mu = 6.3$), OCCUPATION ($\mu = 6.2$), GENDER ($\mu = 5.8$), EXPERIENCE ($\mu = 5.7$), and PARTY ($\mu = 5.2$), while ignoring attributes like FIRST NAME ($\mu = 1.3$) and LAST NAME ($\mu = 1$).

Many participants intentionally balanced the committee along several attributes (seeking "balanced representation" – X02). For example, four participants balanced by GENDER (5 men and 5 women). The same four also balanced by PARTY (5 Republicans and 5 Democrats). One participant, X06, intentionally biased heavily toward GENDER (choosing a committee of 10 women), while attempting to balance other attributes to "represent different views and backgrounds of women as this is a decision for women to make."

The ways that participants biased their committee selections were explicit but nuanced. While X05 balanced across GENDER and PARTY, she chose a committee with all 10 members opposed to the bill, explicitly prioritizing "members (who) were very opposed to the bill." Similarly, X01 chose a committee of 4 politicians who were opposed to the bill, while only 3 each who were neutral or in favor of the bill (i.e., breaking ties by biasing toward

those opposed to the bill), hoping for "a vote against the abortion bill" while also seeking a committee that was "equally balanced amongst those who opposed, neutral, and in favor of the bill." The same participant, who identified as a Democrat, chose more Republicans than Democrats in their committee (6 v. 4).

When bias was present (in the form of unequal choices across options), it appeared to be the result of participants' explicit choices of give and take rather than implicit or unintentional biases. One exception became clear in the summative review: X01 said "oh, that's interesting" upon realizing he had unintentionally focused on politicians with more EXPERIENCE. However, this insight was the exception rather than the rule, as most participants seemed unsurprised by the system's accounting of their interactions and committee member selections. We hypothesize this may be the result of the relatively low dimensionality of the data (5 attributes) and relatively few choices per attribute (2-4 possible values per attribute) that enabled participants to maintain a reasonable mental bookkeeping of the attributes they cared about. Hence, in the next formative study, we examine higher dimensional data, with the goal of understanding the impact of implicit bias on people's decision making. Furthermore, the balanced dataset is unrealistic in American politics (e.g., political party is strongly correlated with people's views on specific policies), so the next study examines a more realistic distribution of data.

5.2.4 Formative Study 2

In Formative Study 1, we found that participants were able to reasonably maintain a mental accounting of their choices in the data and made explicit choices about how to balance or bias the attributes as a result. The goal in this formative study was two-fold: (1) to increase the dimensionality of the dataset from the previous study (and hence the level of realism as people have more things to consider when making selections), and (2) to test the effectiveness of an additional intervention (*real-time* interaction traces) toward mitigating potential biases. To address the latter goal, participants were divided into either a **Control** or **In-**

tervention condition, which dictated which version of the system they used to complete the task. All participants still completed two phases of the study: initial committee selection, followed by summative review of interactions and selections, then revised committee selection. The task and system are the same as those described in Section 5.2.2.

Dataset

We sought to increase the realism in this study by increasing the dimensionality of the dataset (i.e., people have more features to keep in mind during their decision), and deriving the dataset of fictitious politicians based on distributions found in the 115th US House of Representatives [118]. In this version of the dataset, each of 100 fictitious politicians is described by biographical attributes (e.g., OCCUPATION, RELIGION, EXPERIENCE, etc) and policy attributes (e.g., each politician's view on issues like LEGALIZE MEDICAL MARIJUANA, etc). Policy attributes take on an integer number $\in$ [-3, 3] representing the *strength* of the policy (1, 2, or 3; 0 is neutral) and *position* (in favor + or opposed -). For example, a politician with -1 toward the policy LEGALIZE MEDICAL MARIJUANA would be somewhat opposed to legalizing medical marijuana. Politicians are assumed to primarily vote along party lines, with a 1% chance of voting against their party and a 5% chance of a neutral (0) policy. For non-neutral policy positions, values were sampled from a distribution of 30% $\pm$1, 50% $\pm$2, and 20% $\pm$3, representing our general view that more neutral policies ($\pm$1) are somewhat more likely than more extreme policies ($\pm$3), with party-dependent policies ($\pm$2) being most likely. The dataset used in the study was produced by sampling from the distributions described in Table 5.1, Formative Study 2 column.

Participants

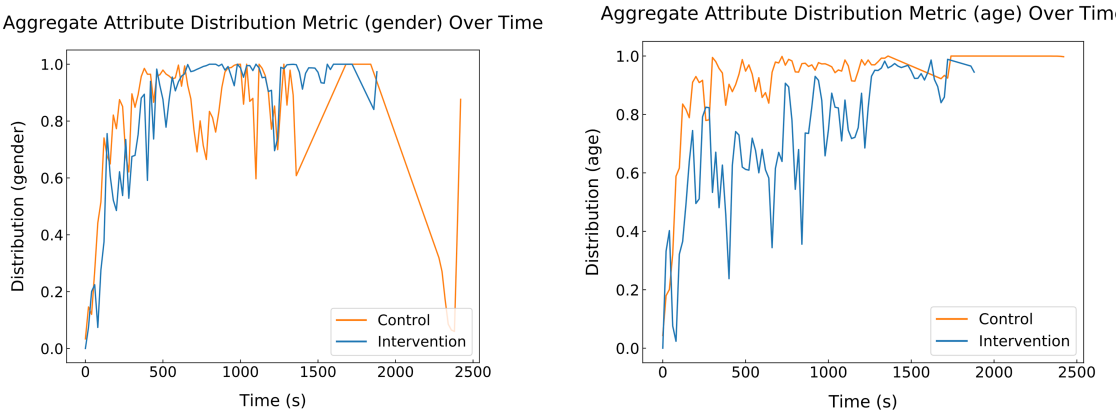
In this study, we recruited 24 student participants from a large university (3 female, 21 male). 21 of the participants self-reported that they most identified with the Democratic party, and 3 participants most identified with the Republican party. All participants self-

reported at least moderate familiarity with analyzing data using visualizations ($\geq$ 3 out of a 5-point likert scale), with an average of 4.1. References to specific participants are labeled according to condition (i.e., Y01-C – Y12-C for Control and Y01-I – Y12-I for Intervention).

Results

We can analyze the bias mitigation strategies with respect to two measures of success, which we refer to as *process* v. *decision*. That is, we can measure bias in the user's analysis *process* (using bias metrics [191]) as well as bias in their final *decision* (by looking at how balanced the political committee is with respect to various attributes of the data). Data analysis in this study was exploratory in nature. Next, we report on the most significant results; however, analysis for all attributes is included in supplemental materials.

Bias in Analysis *Process*. We analyze *bias in the analysis process* between the Control and Intervention conditions by comparing the bias metric values (Chapter 4.1). Viewed over time, the bias metric values (Attribute Distribution, in particular) can indicate how



(a) GENDER attribute **(b)** AGE attribute

Figure 5.6: Study 2: Average Attribute Distribution metric values for Control (orange) v. Intervention (blue) participants. Higher values (closer to 1) represent higher bias compared to the distribution of the attribute in the full dataset. There is no clear difference between conditions for GENDER (a), but Control participants exhibited more bias toward AGE than Intervention participants (b).

closely the user's focus on the data analysis was proportional to the underlying data. Since the data violates conditions of normality, we analyze participants' bias metric values using the Kruskal-Wallis test by ranks, a non-parametric one-way ANOVA [102].

Figure 5.6 shows the aggregate Attribute Distribution (AD) metric values over time for participants in each condition for (a) GENDER and (b) AGE. Time is shown on the x-axis, while the bias metric is shown on the y-axis (higher values = more bias). Orange curves represent the average AD bias metric value for Control participants, while blue curves represent the average AD bias metric value for Intervention participants.

Some attributes show little noticeable difference between conditions (e.g., GENDER), while others show a clear distinction (e.g., AGE). For AGE (Figure 5.6b), Control participants (orange) tended to have higher metric values over time than Intervention participants (blue), suggesting that Intervention participants interacted with politicians whose ages were more proportional to the underlying dataset. We verify this trend by comparing the mean AD bias metric value for AGE over time for Control and Intervention participants ($\mu_C = 0.857$, $\mu_I = 0.729$, $H = 3.360$, $p = 0.057$).

To better understand the potential impact of real-time bias mitigation strategies, we examine bias metric values over time in conjunction with user interactions with the interaction trace view (Figure 5.3F). For example, Figure 5.7 shows bias metric values for AD of PARTY over time for an Intervention participant. The user exhibits high bias throughout most of the task. Red vertical lines indicate moments when the user interacted with PARTY in the *real-time* interaction trace view. After interacting with the interaction trace view, the participant's bias toward PARTY decreases. One possible explanation is that the user observed bias in their interactions toward Democratic politicians in the interaction trace visualization and consequently went on to focus on Republicans to reduce the bias. This trend is not universally true across all participants and attributes; however, it does demonstrate some promise that the interaction trace view has potential to impact the way that users interact with data.

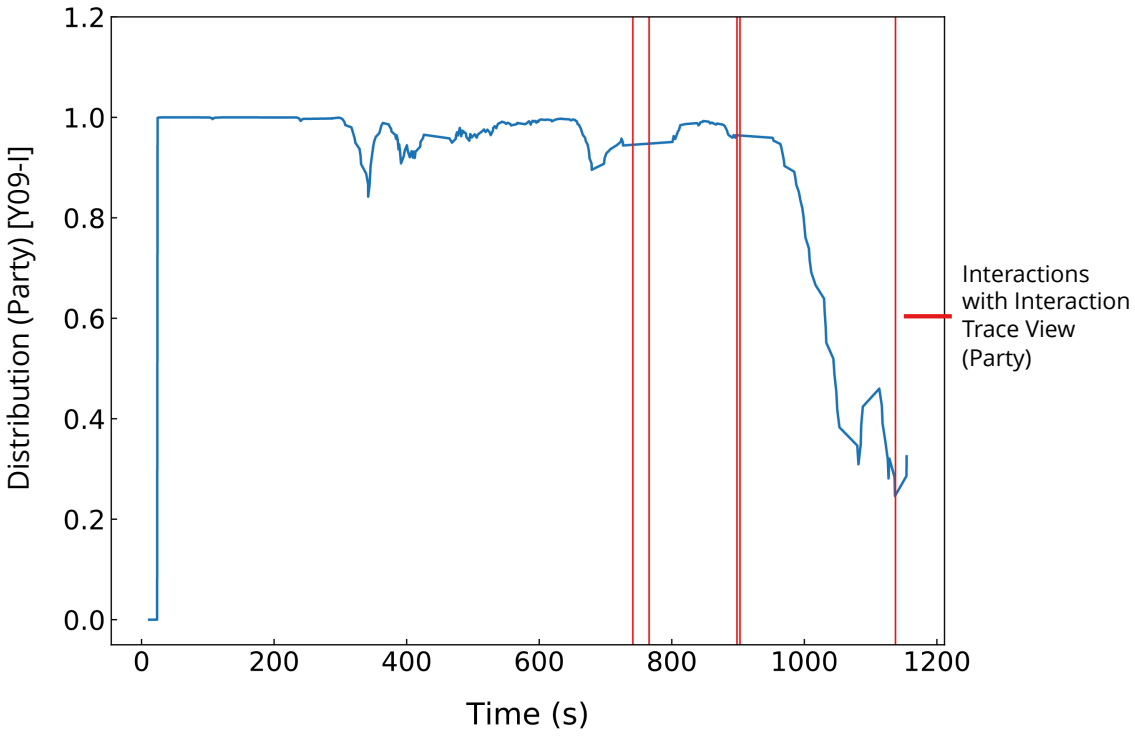


Figure 5.7: The Attribute Distribution (PARTY) metric value for one Study 2 Intervention participant. Vertical red lines indicate interactions with PARTY in the real-time interaction trace view (Figure 5.3F).

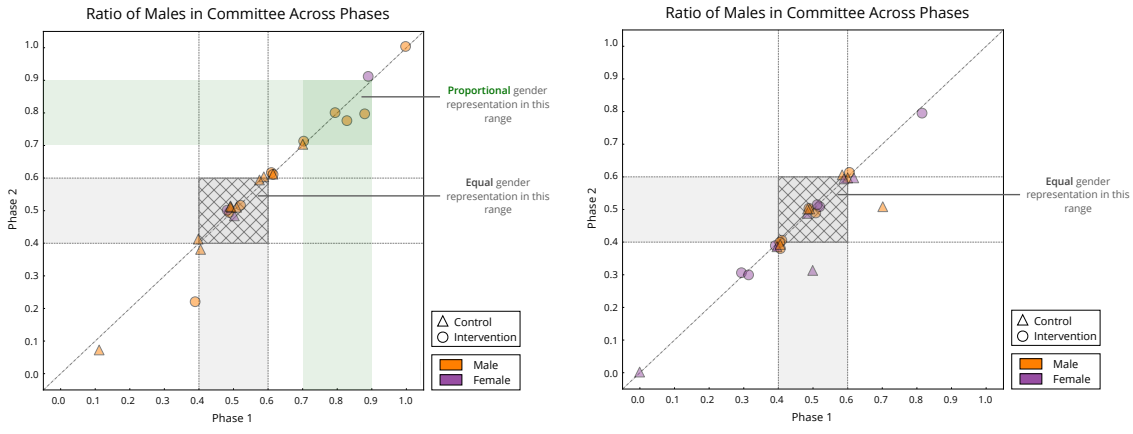
Balance: Bias in Final *Decisions*. After reviewing the *summative* interaction trace visualization, 12 participants immediately resubmitted their initial committee. Another 5 participants looked back at the data but ultimately made no revisions, and 7 participants (4 C, 3 I) changed members of their committee. We describe analyses below from both phases, and discuss this result further in Section 5.2.6.

During the study, several participants mentioned some desire to *balance* the committee they selected. However, people tended to refer to balance in different ways. Many participants thought of a balanced committee as one which has diversity and / or equal representation of values (e.g., Y15-C said, "I would have preferred to have members from all the religious groups and education levels in the committee"); whereas others thought of a balanced committee as one which was proportional to the dataset and / or electorate (e.g., Y13-C said, in reference to the criteria for selecting his committee, that he sought "an accurate representation of the opinions of the people of Georgia"). Some people seemed

to refer to both notions in conflict. For example, upon realizing his chosen committee had 5 men and 5 women, Y09-I observed "There were more men in the dataset. I chose pretty balanced." We further address this multiplicity in balance perspectives in Section 5.2.6. Supplemental materials contain visualizations and analyses for all other attributes of the data, as well as additional ways of quantifying balance (e.g., by notions of *diversity* or *proportionality*).

We use the term *balance* in the context of our analysis to quantify an objective ratio of attribute values. For instance, we can discuss the issue of gender balance in the chosen committees by examining the ratio of male politicians chosen in each participant's committee. Since the data violates conditions of normality, we analyze participants' balance in committees using the Kruskal-Wallis test by ranks, a non-parametric one-way ANOVA [102]. We focus herein on GENDER, since our exploratory analysis indicated the strongest effect.

We compare Control v. Intervention participants in Phase 1 and in Phase 2 of the study. Of all attributes of the data, 9 participants (6 C, 3 I) self-reported that GENDER was the primary consideration in their decision making. Figure 5.8a shows the ratio of male committee members chosen in Phase 1 (x-axis) and in Phase 2 (y-axis). We find that



(a) Study 2 **(b)** Main Study

Figure 5.8: The balance of GENDER in committees chosen by 24 participants in (a) Study 2 and (b) Main Study. Balance is shown as the ratio of men in each participant's chosen committee in Phase 1 (x-axis) and Phase 2 (y-axis), shape-coded by condition and color-coded by participant gender.

Intervention participants choose a higher ratio of male committee members than Control participants in Phase 1 of the task ($\mu_C = 0.492$, $\mu_I = 0.683$, $H = 4.795$, $p = 0.029$) as well as in Phase 2 ($\mu_C = 0.492$, $\mu_I = 0.658$, $H = 4.537$, $p = 0.033$).

While many participants thought of a balanced committee as one in which both genders were represented in equal numbers (e.g., Y18-C expressed his criteria included "gender equal"), the bias metrics represent the user's interactions in comparison to the underlying *distribution* of the data. In the study dataset, 13% were female politicians, while 87% were male. Hence, the use of real-time visualization of interaction traces in the Intervention condition tended to nudge participants toward a more male-dominant committee, consistent with the underlying distribution of the data (Figure 5.8a, green), compared to people's tendency to pursue equal numbers of each gender (Figure 5.8a, gray).

Qualitative Feedback. Participants found the *summative* metric visualization most useful, with a median Likert rating of 4.5 out of 5. They found varying utility in the *real-time* bias metric visualizations (interaction traces): median 4 out of 5 for coloring points by Data Point Distribution; 2 out of 5 for Attribute Coverage visualizations; 4 out of 5 for Attribute Distribution of numerical attributes; and 3 out of 5 for Attribute Distribution of categorical attributes. Of Intervention participants, two did not interact with the interaction trace view at all, two interacted only toward the end of Phase 1, two interacted early on or at random points in the analysis, and 6 interacted many times throughout the decision making process. Collectively, these ratings suggest that participants preferred the *summative* metric visualization over the *real-time* metric visualization.

With respect to the increased dimensionality of the dataset from Study 1, we found that participants focused on explicitly balancing or biasing a few attributes, but could not as easily maintain a mental accounting of *all* attributes in this dataset. Several participants indicated frustration as a result (e.g., Y07-I said "I can only do one thing at a time"). Similarly, participants expressed more surprise about how their interactions and selections mapped to the underlying dataset when considering the *summative* view, potentially indi-

cating that the view increased their awareness of bias in their analysis process (e.g., Y10-I said "I'm surprised I didn't choose a doctor"). Some attributes that participants were less explicitly focused on had high bias as a result. For instance, 0 participants indicated that AGE was the most important attribute in their decision making; yet, high bias was observed (Figure 5.6b). However, we see promise in showing interaction traces to increase users' awareness of biases that may be both explicit and implicit.

Summary of Findings. In this study, we found qualitative evidence that increased dimensionality of the data and cardinality of attribute values resulted in some attributes less explicitly managed and considered by participants. We also found that showing interaction traces tended to lead participants to choose more proportional GENDER composition in selected committees. In the case of a dataset with distributions sampled from the U.S. House of Representatives, this means that we encouraged participants in the Intervention condition to ultimately choose more men in their committees. We discuss the implications of this result further in Section 5.2.6.

5.2.5 Primary Study

The two formative studies (Sections 5.2.3-5.2.4) informed us that (1) with relatively low dimensional data, people are effective at maintaining a mental accounting of the data and making explicit choices, and (2) showing interaction traces appears to be a promising way of nudging participants to make choices of GENDER representation more proportional to the underlying dataset. However, Study 2 had a few key limitations. Namely, there was gender bias in our sampling of participants (21 men, 3 women), and the simulated dataset differed from the intended population (U.S. House of Representatives) in a few major ways. In this study, we address these two concerns.

System

We made a few changes to the system for this study. In Study 2, beneath the distribution curves, numerical attributes also showed coverage (Attribute Coverage metric) by visualizing four quartiles of the attribute range (Figure 5.3F, bottom), colored according to which have been interacted with (blue) or not (gray). In Study 2, participants tended not to notice or use this method of showing interaction traces and found it to be the least useful based on Likert ratings (median 2 out of 5). Thus, we removed this feature for the main Study. We also removed the default selection of the visible attribute in the interaction trace view (so that the distribution comparison is only visible when the user selects an attribute). Now, if a participant never interacts with the view, we can be sure they did not gain some unmeasurable insight by looking at the default view. We also allowed both categorical and numerical attributes to be assigned to axes in the scatterplot (compared to numerical attributes only in Study 2), with data points jittered to prevent overplotting.

Dataset

We sought a balance between the control in the Study 1 dataset and the realism in the Study 2 dataset. While the distributions in the Study 2 dataset were sampled from the U.S. House of Representatives, there were key differences in the simulated dataset instance we used. Minor variations included 52 Republicans and 48 Democrats, compared to expected values of 54 and 46. More notable differences in the distributions sampled from v. the dataset used in Study 2 included (1) only one female Republican (expected value of 6), (2) no ministers (expected value of one), and no scientists (expected value of one). Furthermore, we found that the high dimensionality led users to entirely ignore many attributes of the data. In this study, we rectified these differences.

In this version of the dataset, each of 144 fictitious politicians is described by the attributes in Table 5.1, Main Study column. Compared to the Study 2 dataset, we reduced the cardinality of OCCUPATION to 6 options (from 12); we removed EDUCATION (given that

OCCUPATION is often highly correlated); we removed all policy-related attributes, except for BAN ABORTION AFTER 6 WEEKS. Furthermore, we ensured that the sampled dataset included the appropriate distributions per attribute, including the intersection of GENDER and PARTY.

Participants

In this study, we recruited 24 student participants from a large university (12 female, 12 male). 23 participants self-reported that they most identified with the Democratic party, and 1 participant most identified with the Republican party. All participants self-reported at least moderate familiarity with analyzing data using visualizations ($\geq$ 3 out of a 5-point likert scale), with an average of 4.1. References to specific participants are labeled according to condition (i.e., Z01-C – Z12-C for Control participants and Z01-I – Z12-I for Intervention participants).

Hypotheses

Based on observations and findings from the previous formative studies, we develop the following hypotheses for the present study:

- **H1** Intervention participants will exhibit less bias in the *analysis process* than Control participants w.r.t. AGE.
- **H2** Intervention participants will choose more proportional *committee selections* than Control participants w.r.t. GENDER.
- **H3** Participants will use the *real-time* interaction trace view throughout the analysis to keep track of their process.
- **H4** *Real-time* review of interaction traces view will impact *interactive behavior*.
- **H5** *Summative* review of interactions and committee selections will increase *awareness* of bias.
- **H6** *Summative* review of interactions and committee selections will impact *interactive*

behavior.

H7 Participants will find the *summative* interaction trace view more useful than the *real-time* view.

Results

Bias in Analysis Process. Based on results in Study 2, we hypothesized that Intervention participants would exhibit less bias in the analysis process for AGE than Control participants. We compare the average Attribute Distribution bias metric value for AGE over time for Control and Intervention participants. We find that Intervention participants have a lower average bias metric value over time; however, the result is not statistically significant ($\mu_C = 0.854$, $\mu_I = 0.784$, $H = 0.013$, $p = 0.908$). Hence, our results provide little support for **H1**.

Balance: Bias in Final Decisions. While Study 2 results showed significant differences in the final decisions participants made for GENDER composition of their committees, we do not find a statistical difference in this study with respect to the ratio of men in committees in Phase 1 ($mu_C = 0.492$, $mu_I = 0.467$, $H = 1.565$, $p = 0.211$) or in Phase 2 ($mu_C = 0.458$, $mu_I = 0.467$, $H = 0.430$, $p = 0.512$). The ratio of male committee members selected by participants is shown in Figure 5.8b. Thus, our results indicate no support for **H2**. 8 participants (4 C, 4 I) ultimately chose a balanced committee with respect to GENDER (5 men, 5 women). 11 others got close to a 50%-50% split with a balance of 4 men, 6 women (2 C, 4 I) or 6 men, 4 women (4 C, 1 I).

Effects of *Real-Time* Intervention. Only participants in the Intervention condition experienced the *real-time* visualization of interaction traces. Based on observations in Study 2, we hypothesized that the majority of participants in this study would utilize the interaction trace view (Figure 5.3F) throughout their analysis process (similar to Figure 5.9A). However, three participants never interacted with the interaction trace view, and three interacted

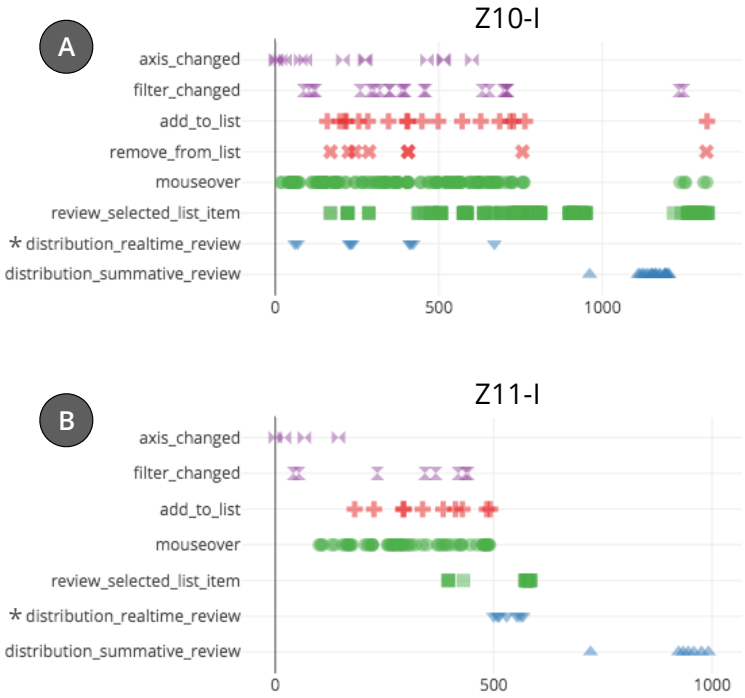


Figure 5.9: Interactions performed by two participants in the Intervention condition of the main Study. The x-axis represents time (in discrete interactions). (A) Participant Z10-I interacted with the interaction trace view (distribution_realtime_review) throughout their analysis; (B) Participant Z11-I interacted with the interaction trace view only toward the end of Phase 1 of the study, before reaching the *summative* phase (distribution_summative_review).

only at the end of Phase 1 to check their work before submitting (similar to Figure 5.9B). Only two participants utilized the *real-time* interaction trace view throughout their analysis as we anticipated. Thus, our results indicate no support for **H3**.

After interacting with the *real-time* interaction trace view, participants often made changes to their selected committee during Phase 1. Because the interaction trace view compares a user's interactions to the underlying distribution of the data, we hypothesized this would lead to changes in user behavior to *make the committee more proportionally representative of the underlying dataset*. By qualitative examination, we find some instances where this is true. For example, Figure 5.10A shows how one participant's PARTY balance shifts significantly after examination of the *real-time* interaction trace view for PARTY (blue triangle). The participant adjusted the committee from 10 Democrats to 4 Republicans and 6 Democrats. In fact, examination of the interaction trace view made the participant aware

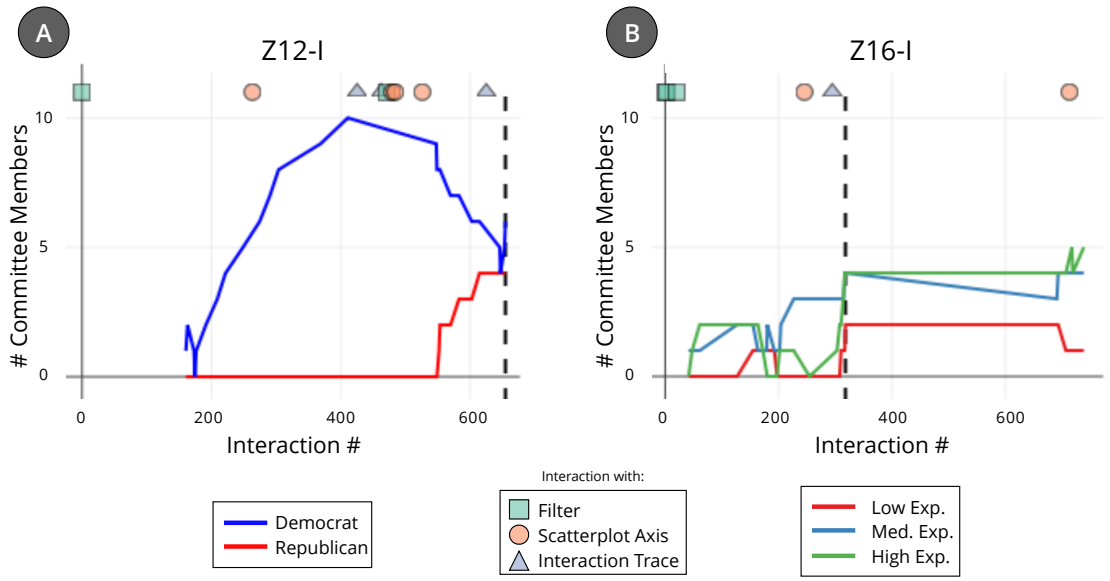


Figure 5.10: The evolving balance of committee choices for (A) PARTY and (B) EXPERIENCE. After participants interacted with the *real-time* interaction trace view (blue triangles), the balance of their committee shifted.

of a mistake in her analysis: "I forgot I had only filtered by Democrats."

While we do see some instances where committees were adjusted to be more proportionally representative, we also see instances where participants adjusted their choices to be more diverse or to be more consistent with pre-existing biases. For example, Figure 5.10B shows one participant's balance of EXPERIENCE prior to submitting the initial committee in Phase 1 (dashed vertical line). Prior to examining the EXPERIENCE interaction trace view (blue triangle), the participant had only selected politicians with a medium amount of EXPERIENCE ($6 \leq x \leq 10$). After examination, she added committee members who had high and low EXPERIENCE as well. While the majority of politicians fell in the medium EXPERIENCE bin ($88/144 \approx 61\%$), this participant adjusted toward creating a more diverse committee (4 high, 4 medium, 2 low) rather than one that was proportionally representative. Thus, while we do observe participants make adjustments to their committees after using the interaction trace view, there is no clear trend about the *direction* of those adjustments.

In addition to the interaction trace view (Figure 5.3F), we also showed interaction traces by coloring points in the scatterplot based on the frequency of user interactions with those

points, inspired by the technique described in [53]. However, different from the results presented in [53], we ultimately observed very little effect on participants' breadth of exploration. Since the data is normal, we performed a simple t-test to confirm. On average, participants in the Intervention condition interacted with more unique data points ($\mu_C = 38.75$, $\mu_I = 41.67$, $t = -0.433$, $p = 0.670$) and interacted more times per data point ($\mu_C = 7.25$, $\mu_I = 8.58$, $t = -0.877$, $p = 0.390$), although the results are not statistically significant. Hence, analysis for **H4** is inconclusive.

Effects of *Summative* Interaction Trace View. All participants, regardless of condition, were exposed to the *summative* visualization of their interaction traces and committee choices. We hypothesized that reviewing this *summative* visualization would lead participants to both increased awareness of potential biases, as well as influence their behavior in some cases. Before proceeding to the Phase 2 revision, participants were asked to reflect on their observations from the *summative* phase. Using an open-coding approach, we analyzed participants' videos during the *summative* review phase for indications of *awareness*: statements indicating the participant learned something about their process or selections. All four authors coded one video and discussed to develop an initial coding. Using that coding, all authors again applied that coding to a new video and revised the codes. After 3 iterations, we achieved theoretical saturation, and one author proceeded to code the remaining videos according to the **type of statement** (awareness, reflection, clarification) and the **object of the statement** (data, interactions, committee, interface-interaction-trace, interface-general, unknown). We distinguish *awareness* from *reflection* (statements discussing something the participant already knew about) and *clarification* (questions about the data or interface). We coded a total of 170 statements, 75 of which were statements about awareness ($\mu = 3.125$ per person). 22/24 participants spoke at least one statement indicating new awareness about their interactions or selections. The two participants who did not make any awareness statements were Intervention participants who already saw the interaction traces in *real-time* prior to the summative review phase. This confirms **H5**.

However, despite 22/24 participants expressing increased awareness or surprise during the *summative* review phase, only 8 participants (4 C, 4 I) revised their choice of committee in Phase 2 of the study. 11 participants (7 C, 4 I) interacted more but ultimately chose not to revise, and 5 participants (1 C, 4 I) resubmitted their initial committee without any further interaction. One reason participants expressed for not revising their committee was concern that making changes along one attribute might unintentionally have an effect on another (Z09-I said "I didn't realize I focused so much on business people ... But I'll leave it for now because it might mess up balance elsewhere"). More Intervention participants immediately resubmitted than Control participants (4 v. 1, respectively), which could suggest that Intervention participants already had increased awareness of how their interactions and selections mapped to the underlying data due to the *real-time* interaction trace views. Hence, they may have incorporated that information into their decision making in Phase 1, whereas Control participants only saw this comparison in the *summative* view. This provides some support for **H6**.

Of the revisions participants made in Phase 2, only two people changed GENDER composition (perhaps because GENDER was often an explicit focus in Phase 1). Comparatively, all 8 participants who revised shifted the distribution of OCCUPATION in their selected committee, 5 participants changed the composition of the policy BAN ABORTION AFTER 6 WEEKS, and 4 participants changed the composition of PARTY and RELIGION. However, as with interactions with the *real-time* interaction trace view, there was no clear trend about the direction in which people revised their committees after looking at the *summative* interaction trace view. Along different dimensions, participants made revisions that (1) increased the diversity or equal representation of the committee, (2) increased the committee's representativeness of the underlying population in the dataset, or (3) were consistent with the participant's pre-existing biases.

Subjective Feedback. Overall, participants found the *summative* metric visualization to be quite useful (median Likert rating of 4 out of 5). Real-time interaction trace visualiza-

tions were rated a median of 4 out of 5 for Data Point Distribution coloring of scatterplot points and a median of 3.5 out of 5 for Attribute Distribution in the interaction trace view. In line with our findings from Study 2, these results suggest that the *summative* metric visualizations may be more useful to participants than the *real-time* metric visualizations, confirming **H7**. We posit this could be due to participants' preference to focus on the task at hand, then reflect and perform revisions to reduce cognitive load.

Summary of Findings. In this study, we addressed two of the primary limitations of Study 2: sampling bias of participants with respect to gender and issues with representation in the dataset. Our analysis indicates that the strongest effect we observed in Study 2 on the impact of interaction traces on GENDER composition of committee selections has disappeared in this study. This loss of effect could be the result of many potential factors, which we discuss further in Section 5.2.6. Nonetheless, qualitative analysis suggests that participants positively viewed the *summative* interaction trace view, which increased participants' awareness of bias and in some cases, impacted their behavior as well.

5.2.6 Discussion

Where did the bias go? In Study 2, we observed a significant difference in Control v. Intervention participants regarding GENDER composition of selected committees. However, in the main Study, this effect disappeared, in addition to a reduction in the magnitude of effect on AGE for the bias metrics. One explanation could be that the complexity of the choice was reduced compared to Study 2 (lower dimensionality of data and lower cardinality of attribute values); hence, participants were able to better maintain a mental accounting of attributes they wanted to balance (as in Study 1). It could also be due to the change in interface to allow categorical attributes to be assigned to the scatterplot axes. With categorical attributes assigned to axes, the points in the scatterplot formed distinct clusters, which could turn a complex cognitive balancing task into a simple visual decision making task (i.e., to choose a point from each cluster). Figure 5.11 demonstrates the distinction.

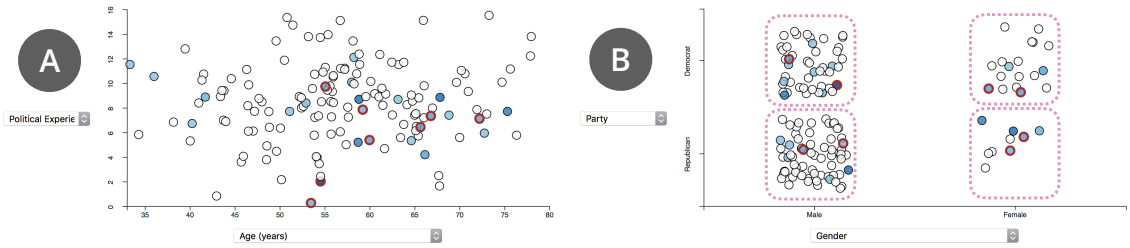


Figure 5.11: Points on the scatterplot are spread out when only numerical attributes can be assigned to axes (A). When categorical attributes can be assigned to axes, clusters of points form, offloading a cognitive task to a perceptual one (B).

Z04-C said, about the benefit of clusters: "Since I have this [Political Experience, Party] on the [Y, X] axis already, the filter kind of becomes redundant." It is also possible that the intervention in Study 2 had a stronger effect for participants of one gender, and when the sampling of participants in the main study addressed gender bias by recruiting equal numbers of male and female participants, the effect disappeared. We find some evidence in support of this explanation: when comparing male v. female participants in the main Study, we find some significant differences in committee balance for Intervention participants but not for Control participants. It could be that one gender is more susceptible to being "nudged" with the intervention. This analysis is included in supplemental materials. However, future studies are needed to isolate the cause.

Explicit v. Implicit Biases. In this series of studies, we have observed both explicit and implicit biases. We hypothesized that participants would choose committee members based on a few explicit criteria, and when dimensionality of data is high, that may result in implicit bias when people lose sight of other attributes. These implicit biases may simply be the result of lack of attention and unknown correlations in the data, or they could be the result of more dangerous implicit attitudes and stereotypes that drive decision making. From a behavioral perspective, the interactions users perform related to explicit or implicit bias may look similar. Eliciting user feedback can then be helpful to refine models of bias by users directly indicating if their focus was intentional or not (Chapter 5.1, dimension D3). Nonetheless, some interactions may be more indicative of explicit decision making

criteria (e.g., users assigning attributes to axes or filtering out subsets of the data), while others may be a better signal of implicit biases (e.g., hovering on points for details). In the current model, only interactions with data points (clicks and hovers) impacted bias computations (axis and filter interactions were not used). This distinction could improve future bias models.

Utility of Interaction Traces. Interventions (*summative* and *real-time* interaction trace visualizations) in Study 2 and the main Study were designed to make users more aware of potential biases in their decision making process. Participants tended to see utility in both *summative* and *real-time* visualization of their interaction traces. Y12-I said "(I want to) make a diverse group... for that, coverage and distribution tool was very helpful." A few participants expressed initial confusion about the interaction trace visualizations, which were for the most part resolved by further inspection (Y22-I said "The blue bars overlapping the gray bars was not intuitive at first go"). Many were made aware of previously unconscious biases, while others simply found it a source of interesting information (Y18-C said "oh, I actually interacted with a farmer. Look at that"). However, this source of information comes at a cognitive cost: it is yet another source of information that participants need to monitor and incorporate in their decision making. We believe this is the primary reason users preferred the *summative* interaction trace view over the *real-time* interaction trace view.

Another possible explanation for this preference is that the *summative* view showed the distribution of the participant's *choice of committee members* (in addition to the interaction distributions shown in both *summative* and *real-time* views). For some decision making scenarios (i.e., incremental decisions), the distribution of intermediate choices could be shown as another form of *real-time* intervention. This was echoed by some participants who suggested this be incorporated as part of the *real-time* intervention (e.g., Y04-I said "you should show the green bars in the task as well"). However, this approach may not work for non-incremental decision making scenarios in which only a single choice is made

(e.g., voting for a single presidential candidate v. selecting a committee of politicians). In such cases, the only *real-time* information that can be shown to the user that might make them more aware of biases in their process is their interaction traces.

Design Implications. With or without *real-time* interaction trace visualizations, some participants found indirect ways to assess balance in their committee choices (e.g., by applying filters and cycling through combinations of scatterplot axes to see the distribution of selected points). Hence, the affordances within the interface design can itself serve as a potentially powerful bias mitigation approach, promoting user awareness and enabling self-editing. Another example is enabling categorical attributes to be assigned to axes to see categorical distributions of selected points, offloading cognitive decision making to a perceptual task (Figure 5.11). Particularly in situations where cognitive overload may prevent users from managing secondary views, designing the interface to afford indirect assessment of their choices may be a better alternative.

Study Limitations. Study 2 suffered from a biased sampling of participants (mostly male Democrats). In the main Study, we were able to correct for gender bias; however, due to sampling within our university, we were unable to balance participants' political party affiliations. Furthermore, there was no condition in which participants did not see either of the interventions (Intervention participants saw interaction traces in *real-time*, and participants in both conditions saw interaction traces in the *summative* view after Phase 1). We treated Control participants' Phase 1 selections as our Control group. Alternatively, we could have added another condition in which participants used the Control interface, but rather than seeing the *summative* interaction trace view, they were shown only a list of selected committee members. It could be the case that the act of asking participants if they want to revise, regardless of any insights gained during the *summative* analysis, prompted behavioral changes. Additional studies are required to isolate this effect. Lastly, our analysis of awareness involved open-coding of participant utterances during the study. We avoided asking participants to fully think aloud throughout the study and only specifically prompted

participants in a few instances. Hence, particularly in the case of *real-time* awareness, the amount of verbalization was very specific to the individual participant. Furthermore, when participants were prompted, the act of posing a question could be a confounding factor that influenced awareness and reflection.

5.2.7 Summary

In this section, we have addressed **RQ 3.2** by presenting results of a sequence of three studies indicating (1) as data dimensionality increases, people are unable to maintain explicit mental decision making criteria for all attributes of the data, (2) when dimensionality of data is high, interaction traces may be a promising way to increase user awareness of bias and encourage users to make selections more proportional to the underlying dataset, and (3) while people found *real-time* traces of their interactions to be useful, they ultimately found the most utility in *summative* visualizations of their interactions and decisions, likely due to cognitive overload during the task itself. This suggests that showing interaction traces to the user is a promising way to increase awareness of potentially implicit biases in a political decision making scenario.

CHAPTER 6
REFLECTIONS

6.1 Perspectives on Bias

As described in Chapter 3, the term "bias" is overloaded, describing a multitude of concepts with a single term. Hence, the goal of Chapter 3 was to provide a broader context to the use of the term, including cognitive, perceptual, and social biases, as well as the objective use of the term "bias" in cognitive models. The latter perspective tends to be the most often overlooked. That is, people tend to be aware of cognitive, perceptual, and social biases; and they all tend to carry a heavy negative connotation. These biases often represent sources of human error. They likewise tend to overshadow more neutral or even positive uses of the term "bias".

For example, exogenous attention is captured by sensory input when someone speaks your name in a crowded room [142]; however, rather than a negative error, biased attention toward your own name compared to other names can be thought of as an evolutionary advantage. Similarly, people often loosely discuss "good biases" v. "bad biases", referring in the former to often explicit criteria learned from experience (i.e., bias toward student applicants with a higher GPA), compared to unconscious social biases (i.e., gender bias) in the latter sense. Clearly, bias is not always a weakness or detriment.

Nonetheless, as a result of these heavy negative connotations surrounding the terminology, we must ensure to carefully contextualize what we mean by bias, or even use alternative terminology to describe it in some cases for the sake of communicative clarity. For example, in the neutral case of *bias as a model mechanism* [193], we might instead refer to bias from this perspective as simply a *model parameter*. In this sense, the negative connotation of bias does not overshadow the intended use of the word.

6.2 Implications of Balance Definition

The bias metrics (Chapter 4.1) are formulated based on the assumption that "unbiased behavior" is proportional to the distribution of the data. Visualizing these metrics in *real-time* in Chapter 5.2 resulted in some significant changes in behavior, nudging participants to choose political committees with GENDER balanced proportionally to the distribution of the dataset. However, it begs the question if this was the **right** way to nudge participants. An alternative and widely held perspective on bias and balance is that of equal representation across diverse values. Future work can explore to what extent different visualization strategies are capable of nudging participants toward more *proportional* v. *diverse* choices. In different decision making scenarios, one definition of balance may be more appropriate than another (e.g., selecting a representative sample for a study v. deciding which job candidates to interview). Nonetheless, designers have a social responsibility to choose visualization designs and bias computation mechanisms that reflect social values without unduly influencing participant behavior.

6.3 Bias in the Data Life-Cycle

This dissertation has focused on sources of human bias in decision making that impact visual data analysis. To some degree, this assumes that the data being visualized is complete, error-free, and unbiased, and that the visual representation is not misleading people to make incorrect inferences. However, before a person sets eyes on a visualization system to explore their data, there is a large portion of the life-cycle of data (Figure 6.1), models (Figure 6.2), and visualization (Figure 6.3) that is unseen. For example, humans often decide by which strategies to sample data; humans label large amounts of data and define features to feed into ML models; ML models may be subject to particular algorithmic biases, favoring some parts of distributions over others; the way that data and models are visually presented can be biased by specific design choices in an interface; and the end

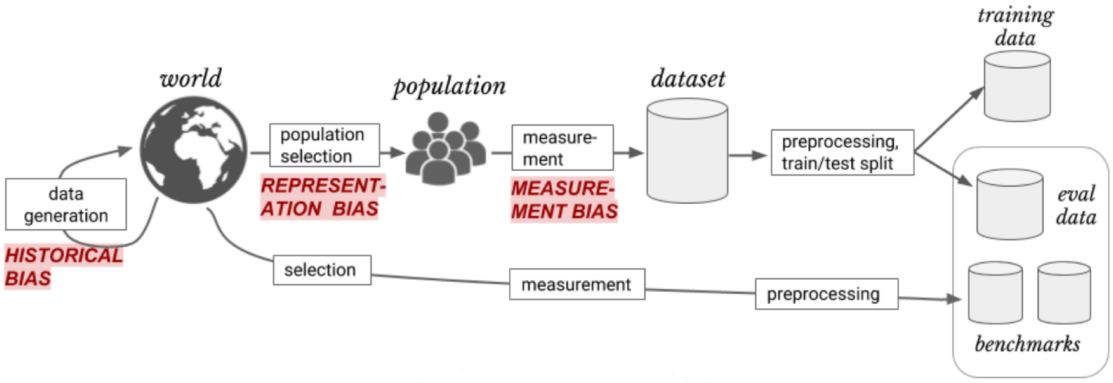


Figure 6.1: The data pipeline [167].

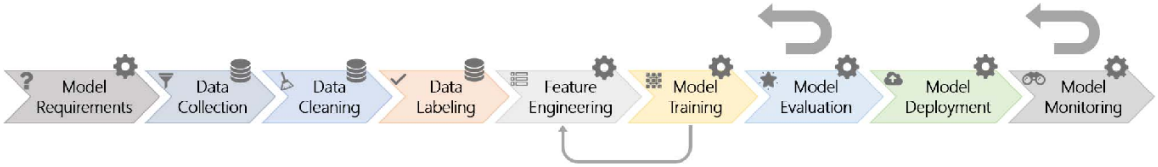


Figure 6.2: The ML pipeline [6].

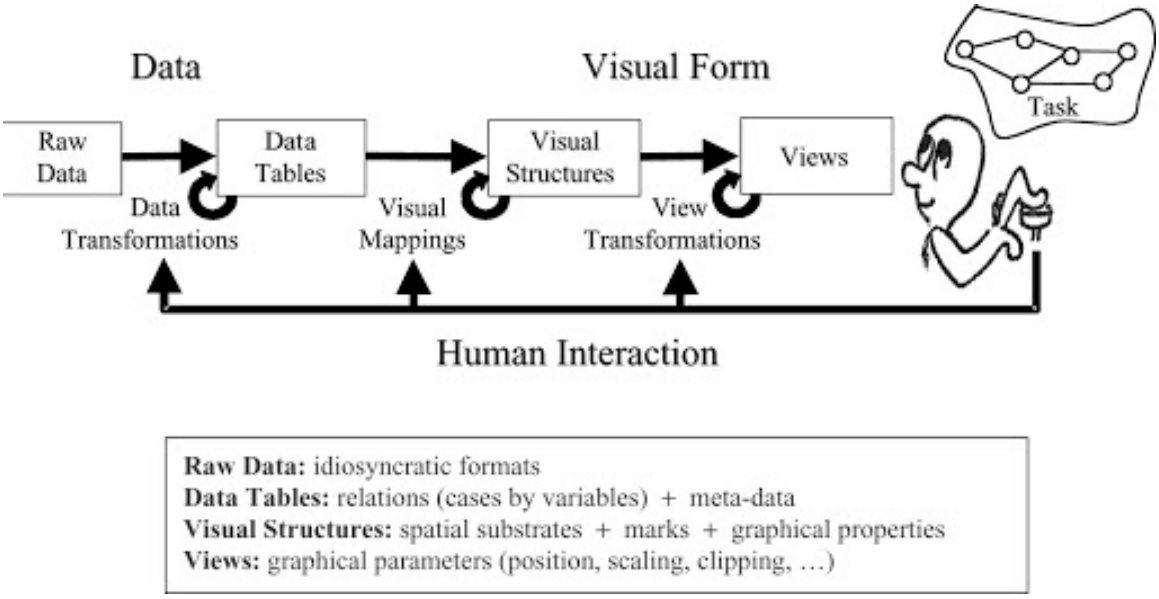


Figure 6.3: The visualization process model [25].

users of the data and model will carry their own cognitive, perceptual, and social biases that influence their decision making.

At each phase of this life-cycle of data, bias can potentially be injected into the process, by means of human or machine. There is a broad, open space that needs to be further

explored: the intricate relationship between various sources of human and machine bias. How can bias at one phase in the life-cycle impact the next? How do these biases propagate to the final decisions humans make using data and models in visualization interfaces?

6.4 Could the mixed-initiative system impart bias to the user?

Yes. A less-emphasized aspect of emergent bias [59] is that the structure of the user interface may influence and bias the interactions of the user. Reliance on machine automation and automated decision aids can result in automation bias. This is the heuristic use of automation instead of more vigilant information seeking and decision making [120, 121, 135]. The errors resulting from automation bias are concerning for mixed-initiative systems, wherein those errors might be integrated into the analytic results / visualizations or even the analytic processes. Of particular concern in this domain are automation commission errors. These errors are inappropriate actions resulting from over-attending to automated aids without attention to the context or other critical environmental information sources. Commission errors occur when a user accepts the recommendation of some machine analytics even when there is contrary evidence from other information sources, either internal or external to the analytics system.[1] The design of an interactive analytic interface may lend itself to overemphasizing some analytic results or mixed-initiative recommendations, such as highlighting recommendations or altering things like the size or color that might make some recommendations stand out over others. Automation bias in accepting the most strongly emphasized recommendations could lead the analyst down a biased analysis path. Does the system or the user bear the responsibility for mitigating automation bias? In this dissertation, I have demonstrated that if mixed-initiative systems can cultivate emergent biases in both the machines and the users, then mixed-initiative systems also offer new opportunities for humans and machines to team up to mitigate negative effects of bias

[1] Commission errors are contrasted with automation omission errors, which occur if the human-machine team fails to respond to system irregularities or the system fails to provide an indicator of a problematic state. In visual analytics, an omission error could occur if a system "knows" an algorithm might be mis-matched to a data type but does not alert the analyst.

through system design.

6.5 Bias Metric Accuracy

While I have demonstrated that the bias metrics (Chapter 4.1) are a promising characterization of the analytic process, a user's cognition is a complex state to model with numerous confounding factors. The accuracy of the bias metrics ultimately depends on being able to reasonably define what constitutes unbiased behavior. Then, the metrics can quantify when user behavior deviates from unbiased behavior. There are many factors, apart from inherent human biases, that influence the way people interact with data in visual interfaces, including perceptual properties that may draw the user's attention. I presented some evidence to refine the bias metrics according to perceptual distance of visual marks (Chapter 4.3); however, many other perceptual factors can drive user behavior (i.e., clusters of points, outliers, salient colors, and so on). Accounting for these perceptual factors is an important next step toward refining models of user bias.

In addition to the perceptual properties of the visualization, there are other factors that can influence user behavior. What the system infers as biased behavior may not be biased in the negative sense of the term. For example, an analyst may focus on a specific subset of data due to prior expertise, contextual knowledge not captured in the data, specific constraints of the task assigned to the analyst, and so on. Furthermore, there may be different cultural opinions of what constitutes biased behavior (e.g., with respect to gender bias in historically matriarchal v. patriarchal cultures [154]). Technologies, including bias metrics, developed by humans and embedded in the fabric of society, cannot be value-agnostic. Hence, particularly in light of the strong negative connotations of the word "bias", we must be careful to design metrics and systems that reflect societal values [58].

My goal in designing bias mitigation strategies (Chapter 5) is not to blindly assert value of the analyst's process. Instead, my goal of developing bias mitigation strategies is to encourage a more reflective decision making process. With this goal in mind, I believe

the optimal accuracy of the bias metrics to be secondary to the visual characterization of the analytic process. Even an imprecise characterization of the analytic process can nonetheless benefit users by causing them to reflect on their process before they make a decision.

CHAPTER 7
CONCLUSION

In summary, the goal of this work is to operationally define, detect, and mitigate biased analytic processes in real-time through user interaction. I hypothesize that user interactions can form a meaningful capture of users' cognitive state during the analysis process. This can be used to detect biased analysis processes. Furthermore, by increasing users' awareness of their bias in real-time through the visual interface, I hypothesize that people can ultimately make better decisions. In the course of confirming this hypothesis, I have produced the following contributions to the visualization research community:

- Operational definitions of bias in visual analytics
- Computational metrics for the detection of cognitive bias from user interaction sequences in visual analytic tools
- Formative evaluation results indicating that the computational bias metrics can be used to describe (anchoring) bias in a user's analytic process
- Empirical evidence of a refined baseline of unbiased behavior that accounts for users' tendencies to interact with nearby data points more often than far away data points
- Design alternatives for the mitigation of a biased analytic process
- Results of a sequence of user studies suggesting that showing interaction traces in an interface can increase users' awareness of potential biases

These contributions collectively advance the state-of-the-art in helping people minimize biased analysis processes.

REFERENCES

[1] A. Abdul-Rahman, R. Borgo, M. Chen, D. J. Edwards, and B. Fisher, "Juxtaposing controlled empirical studies in visualization with topic developments in psychology," *arXiv preprint arXiv:1909.03786*, 2019.

[2] P. D. Adamczyk and B. P. Bailey, "If not now, when?: The effects of interruption at different moments within task execution," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM, 2004, pp. 271–278.

[3] N. Adrienko and G. Adrienko, *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. New York: Springer-Verlag, 2005, ISBN: 9783540259947.

[4] F. Alaieri and A. Vellino, "Ethical decision making in robots: Autonomy, trust and responsibility," in *Social Robotics: 8th International Conference*, A. Agah, J.-J. Cabibihan, A. M. Howard, M. A. Salichs, and H. He, Eds., Kansas City, MO: Springer International Publishing, 2016, pp. 159–168.

[5] J. H. Aldrich *et al.*, *Why parties?: The origin and transformation of political parties in America*. University of Chicago Press, 1995.

[6] S. Amershi, A. Begel, C. Bird, R. DeLine, H. Gall, E. Kamar, N. Nagappan, B. Nushi, and T. Zimmermann, "Software engineering for machine learning: A case study," in *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, IEEE, 2019, pp. 291–300.

[7] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[8] *Angular*, n.d.

[9] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, *Machine bias*, 2016.

[10] H. R. Arkes, "Costs and benefits of judgment errors: Implications for debiasing.," *Psychological bulletin*, vol. 110, no. 3, p. 486, 1991.

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

[12] D. Baker, D. Georgakopoulos, H. Schuster, A. Cassandra, and A. Cichocki, "Providing customized process and situation awareness in the collaboration manage-

ment infrastructure," in *Cooperative Information Systems, 1999. CoopIS'99. Proceedings. 1999 IFCIS International Conference on*, IEEE, 1999, pp. 79–91.

[13] A. D. Balakrishnan, S. R. Fussell, and S. Kiesler, "Do visualizations improve synchronous remote collaboration?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2008, pp. 1227–1236.

[14] M. Bertrand and S. Mullainathan, "Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination," *American economic review*, vol. 94, no. 4, pp. 991–1013, 2004.

[15] 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*, vol. 62, pp. 452–466, 2016.

[16] P. Booth, N. Gibbins, and S. Galanis, "Towards a theory of analytical behaviour: A model of decision-making in visual analytics," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[17] F. Bouali, A. Guettala, and G. Venturini, "Vizassist: An interactive user assistant for visual data mining," *The Visual Computer*, vol. 32, no. 11, pp. 1447–1463, 2016.

[18] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete, "A principled way of assessing visualization literacy," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1963–1972, 2014.

[19] 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.

[20] 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*, vol. 20, no. 12, pp. 1663–1672, 2014.

[21] D. Burgstahler, N. Richerzhagen, F. Englert, R. Hans, and R. Steinmetz, "Switching push and pull: An energy efficient notification approach," in *Mobile Services (MS), 2014 IEEE International Conference on*, IEEE, 2014, pp. 68–75.

[22] M. Burnett, S. Stumpf, J. Macbeth, S. Makri, L. Beckwith, I. Kwan, A. Peters, and W. Jernigan, "GenderMag: A method for evaluating software's gender inclusiveness," *Interacting with Computers*, vol. 28, no. 6, pp. 760–787, 2016.

[23] J. R. Busemeyer and A. Diederich, *Cognitive Modeling*. Los Angeles, CA: Sage, 2010.

[24] J. R. Busemeyer and J. T. Townsend, "Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment," *Psychological Review*, vol. 100, no. 3, pp. 432–459, 1993.

[25] M. Card, *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[26] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski, "Characterizing guidance in visual analytics," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 111–120, 2017.

[27] S. Chaiken and Y. Trope, *Dual-Process Theories in Social Psychology*. New York: Guilford Press, 1999, ISBN: 9781572304215.

[28] W. G. Chase and H. A. Simon, "Perception in chess," *Cognitive psychology*, vol. 4, no. 1, pp. 55–81, 1973.

[29] I. Cho, R. Wesslen, A. Karduni, S. Santhanam, S. Shaikh, and W. Dou, "The anchoring effect in decision-making with visual analytics," *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

[30] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *Journal of the American statistical association*, vol. 79, no. 387, pp. 531–554, 1984.

[31] M. I. Coco and R. Dale, "Cross-recurrence quantification analysis of categorical and continuous time series: An r package," *Frontiers in Psychology*, vol. 5, p. 510, 2014.

[32] M. Correll and M. Gleicher, "Bad for data, good for the brain: Knowledge-first axioms for visualization design," in *IEEE VIS 2014*, 2014.

[33] M. Correll and J. Heer, "Black hat visualization," in *Workshop on Dealing with Cognitive Biases in Visualisations (DECISIVe), IEEE VIS*, 2017.

[34] S. J. Correll, S. Benard, and I. Paik, "Getting a job: Is there a motherhood penalty?" *American journal of sociology*, vol. 112, no. 5, pp. 1297–1338, 2007.

[35] J. A. Cottam and L. M. Blaha, "Bias by default? a means for a priori interface measurement," *DECISIVe: Workshop on Dealing with Cognitive Biases in Visualizations*, 2017.

[36] P. Cowley, L. Nowell, and J. Scholtz, "Glass box: An instrumented infrastructure for supporting human interaction with information," in *Proceedings of the 38th An-*

nual Hawaii International Conference on System Sciences, 2005. HICSS'05, IEEE, 2005, p. 296c.

[37] M. Czerwinski, E. Cutrell, and E. Horvitz, "Instant messaging: Effects of relevance and timing," in *People and computers XIV: Proceedings of HCI*, vol. 2, 2000, pp. 71–76.

[38] *D3.js*, n.d.

[39] A. Desanctis, "Judge blocks georgia's pro-life heartbeat bill," *National Review*, 2019.

[40] E. Dimara, G. Bailly, A. Bezerianos, and S. Franconeri, "Mitigating the attraction effect with visualizations," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 850–860, 2019.

[41] E. Dimara, A. Bezerianos, and P. Dragicevic, "The attraction effect in information visualization," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 471–480, 2017.

[42] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE transactions on visualization and computer graphics*, 2018.

[43] H. Doleisch, H. Hauser, M. Gasser, and R. Kosara, "Interactive focus+ context analysis of large, time-dependent flow simulation data," *Simulation*, vol. 82, no. 12, pp. 851–865, 2006.

[44] 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*, vol. May/June, no. March, pp. 52–61, 2009.

[45] I. Dror, "Combating bias: The next step in fighting cognitive and psychological contamination," *Journal of Forensic Sciences*, vol. 57, no. 1, pp. 276–277, 2012.

[46] 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)*, vol. 8462 LNCS, pp. 92–105, 2014.

[47] H. E. Egeth and S. Yantis, "Visual attention: Control, representation, and time course," *Annual Review of Psychology*, vol. 48, no. 1, pp. 269–297, 1997.

[48] 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.

[49] 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.

[50] 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*, vol. 43, no. 3, pp. 411–435, 2014.

[51] M. Englich and T Mussweiler, "Anchoring effect," *Cognitive Illusions: Intriguing Phenomena in Judgement, Thinking, and Memory*, p. 223, 2016.

[52] J.-D. Fekete, J. Van Wijk, J. Stasko, and C. North, "The value of information visualization," *Information Visualization*, pp. 1–18, 2008.

[53] 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*, vol. 23, no. 1, pp. 351–360, 2017.

[54] M. Feng, E. Peck, and L. Harrison, "Patterns and pace: Quantifying diverse exploration behavior with visualizations on the web," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 501–511, 2019.

[55] S. Few, *Now you see it: simple visualization techniques for quantitative analysis*. Analytics Press, 2009.

[56] B. Fisher, T. M. Green, and R. Arias-Hernández, "Visual analytics as a translational cognitive science," *Topics in Cognitive Science*, vol. 3, no. 3, pp. 609–625, 2011.

[57] 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*, vol. 56, no. 1, pp. 272–276, 2012.

[58] B. Friedman, "Value-sensitive design," *Interactions*, vol. 3, no. 6, pp. 16–23, 1996.

[59] B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems (TOIS)*, vol. 14, no. 3, pp. 330–347, 1996.

[60] S. N. Friel, F. R. Curcio, and G. W. Bright, "Making sense of graphs: Critical factors influencing comprehension and instructional implications," *Journal for Research in mathematics Education*, pp. 124–158, 2001.

[61] J. P. Frisby and J. V. Stone, *Seeing: The Computational Approach to Biological Vision*. Cambridge, MA: The MIT Press, 2010.

[62] A. Furnham and H. C. Boo, "A literature review of the anchoring effect," *The Journal of Socio-economics*, vol. 40, no. 1, pp. 35–42, 2011.

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

[64] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, E3635–E3644, 2018.

[65] *Georgia state senate committees*, http://www.senate.ga.gov/committees/en-US/Home.aspx, Accessed 2020-04-02.

[66] G. Gigerenzer and W. Gaissmaier, "Heuristic decision making," *Annual Review of Psychology*, vol. 62, pp. 451–482, 2011.

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

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

[69] D. Gotz and M. X. Zhou, "Characterizing users' visual analytic activity for insight provenance," *Information Visualization*, vol. 8, no. 1, pp. 42–55, 2009.

[70] 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*, vol. 29, no. 4, pp. 523–531, 1957.

[71] D. Green and J. Swets, *Signal Detection Theory and Psychophysics*. New York: Wiley & Sons, Inc., 1966.

[72] 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.

[73] A. G. Greenwald and L. H. Krieger, "Implicit bias: Scientific foundations," *California Law Review*, vol. 94, no. 4, pp. 945–967, 2006.

[74] T. Grüne-Yanoff and R. Hertwig, "Nudge versus boost: How coherent are policy and theory?" *Minds and Machines*, vol. 26, no. 1-2, pp. 149–183, 2016.

[75] C. Gutwin and S. Greenberg, "Design for individuals, design for groups: Tradeoffs between power and workspace awareness," in *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, ACM, 1998, pp. 207–216.

[76] C. Gutwin, M. Roseman, and S. Greenberg, "A usability study of awareness widgets in a shared workspace groupware system," in *Proceedings of the 1996 ACM conference on Computer supported cooperative work*, ACM, 1996, pp. 258–267.

[77] D. F. Halpern, "Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring.," *American psychologist*, vol. 53, no. 4, p. 449, 1998.

[78] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala, "Graphical histories for visualization: Supporting analysis, communication, and evaluation," *IEEE transactions on visualization and computer graphics*, vol. 14, no. 6, 2008.

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

[80] R. J. Heuer Jr, R. J. Heuer, and R. H. Pherson, *Structured analytic techniques for intelligence analysis*. Cq Press, 2010.

[81] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink, "Trust in automation," *IEEE Intelligent Systems*, vol. 28, no. 1, pp. 84–88, 2013.

[82] R. M. Hogarth, "A note on aggregating opinions," *Organizational behavior and human performance*, vol. 21, no. 1, pp. 40–46, 1978.

[83] E. Horvitz, "Principles of Mixed-Initiative User Interfaces," *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, no. May, pp. 159–166, 1999.

[84] J. Huber, J. W. Payne, and C. Puto, "Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis," *Journal of Consumer Research*, vol. 9, no. 1, pp. 90–98, 1982.

[85] J. Hullman, P. Resnick, and E. Adar, "Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering," *PloS one*, vol. 10, no. 11, 2015.

[86] E. Hutchins, "The distributed cognition perspective on human interaction," *Roots of human sociality: Culture, cognition and interaction*, vol. 1, p. 375, 2006.

[87] E. Hutchins and T. Klausen, "Distributed cognition in an airline cockpit," *Cognition and communication at work*, pp. 15–34, 1996.

[88] T. Jankun-Kelly, "Using visualization process graphs to improve visualization exploration," in *International Provenance and Annotation Workshop*, Springer, 2008, pp. 78–91.

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

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

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

[92] A. Kale, F. Nguyen, M. Kay, and J. Hullman, "Hypothetical outcome plots help untrained observers judge trends in ambiguous data," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 892–902, 2018.

[93] 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, 2008, pp. 154–175.

[94] H. Kim, J. Choo, H. Park, and A. Endert, "InterAxis: Steering Scatterplot Axes via Observation-Level Interaction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 131–140, 2016.

[95] Y.-S. Kim, K. Reinecke, and J. Hullman, "Data through others' eyes: The impact of visualizing others' expectations on visualization interpretation," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 760–769, 2017.

[96] Y.-S. Kim, L. A. Walls, P. Krafft, and J. Hullman, "A bayesian cognition approach to improve data visualization," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, 2019, p. 682.

[97] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso, "A data–frame theory of sensemaking," in *Expertise out of context*, Psychology Press, 2007, pp. 118–160.

[98] K. Koffka, *Principles of Gestalt Psychology*. Routledge, 2013, vol. 44.

[99] G. J. Koop and J. G. Johnson, "The response dynamics of preferential choice," *Cognitive Psychology*, vol. 67, no. 4, pp. 151–185, 2013.

[100] D. R. Kretz, "Experimentally evaluating bias-reducing visual analytics techniques in intelligence analysis," in *Cognitive Biases in Visualizations*, Springer, 2018, pp. 111–135.

[101] J. Kruger and D. Dunning, "Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments.," *Journal of personality and social psychology*, vol. 77, no. 6, p. 1121, 1999.

[102] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952.

[103] 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*, vol. 23, no. 1, pp. 221–230, 2017.

[104] P.-M. Law and R. C. Basole, "Designing breadth-oriented data exploration for mitigating cognitive biases," in *Cognitive Biases in Visualizations*, Springer, 2018, pp. 149–159.

[105] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[106] P. Lee, *Learning from Tay's introduction*, `https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/`, Blog, 2016.

[107] R. D. Luce, "Detection and recognition," in *Handbook of Mathematical Psychology*, R. D. Luce, R. R. Bush, and E. Galanter, Eds., vol. 1, New York: Wiley, 1963, pp. 103–190.

[108] R. D. Luce, "The choice axiom after twenty years," *Journal of mathematical psychology*, vol. 15, no. 3, pp. 215–233, 1977.

[109] C. M. MacLeod, "Half a century of research on the stroop effect: An integrative review.," *Psychological bulletin*, vol. 109, no. 2, p. 163, 1991.

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

[111] N. K. Malhotra, "Information load and consumer decision making," *Journal of Consumer Research*, vol. 8, no. 4, pp. 419–430, 1982.

[112] A. E. Mannes, R. P. Larrick, and J. B. Soll, "The social psychology of the wisdom of crowds.," 2012.

[113] T. Manzini, Y. C. Lim, Y. Tsvetkov, and A. W. Black, "Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings," *arXiv preprint arXiv:1904.04047*, 2019.

[114] J.-P. Martin-Flatin, "Push vs. pull in web-based network management," in *Integrated Network Management, 1999. Distributed Management for the Networked Millennium. Proceedings of the Sixth IFIP/IEEE International Symposium on*, IEEE, 1999, pp. 3–18.

[115] L. E. Matzen, M. J. Haass, K. M. Divis, Z. Wang, and A. T. Wilson, "Data visualization saliency model: A tool for evaluating abstract data visualizations," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 563–573, 2018.

[116] B. J. McNeil, S. G. Pauker, H. C. Sox Jr, and A. Tversky, "On the elicitation of preferences for alternative therapies," *New England Journal of Medicine*, vol. 306, no. 21, pp. 1259–1262, 1982.

[117] A. McNutt, G. Kindlmann, and M. Correll, "Surfacing visualization mirages," *arXiv preprint arXiv:2001.02316*, 2020.

[118] *Membership of the 115th congress: A profile*, https://www.senate.gov/CRSpubs/b8f6293e-c235-40fd-b895-6474d0f8e809.pdf, Accessed 2019-07-25, 2018.

[119] J. T. Milord and R. P. Perry, "A methodological study of overloadx," *The Journal of General Psychology*, vol. 97, no. 1, pp. 131–137, 1977.

[120] K. L. Mosier and L. J. Skitka, "Human decision makers and automated decision aids: Made for each other," in *Automation and Human Performance: Theory and Applications*, R. Parasuraman and M. Mouloua, Eds., Mahwah, NJ: Lawrence Erlbaum Associates, 1996, pp. 201–220.

[121] K. L. Mosier and L. J. Skitka, "Automation use and automation bias," in *Proceedings of the human factors and ergonomics society annual meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 43, 1999, pp. 344–348.

[122] 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.

[123] T. Mussweiler, F. Strack, and T. Pfeiffer, "Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility," *Personality and Social Psychology Bulletin*, vol. 26, no. 9, pp. 1142–1150, 2000.

[124] A. Newell, *Unified theories of cognition*. Harvard University Press, 1994.

[125] P. H. Nguyen, C. Turkay, G. Andrienko, N. Andrienko, O. Thonnard, and J. Zouaoui, "Understanding user behaviour through action sequences: From the usual to the unusual," *IEEE transactions on visualization and computer graphics*, 2018.

[126] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of General Psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[127] 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.

[128] R. M. Nosofsky, "Overall similarity and the identification of separable-dimension stimuli: A choice model analysis," *Perception & Psychophysics*, vol. 38, no. 5, pp. 415–432, 1985.

[129] R. M. Nosofsky, "Stimulus bias, asymmetric similarity, and classification," *Cognitive Psychology*, vol. 23, no. 1, pp. 94–140, 1991.

[130] R. M. Nosofsky, "The generalized context model: An exemplar model of classification," in *Formal Approaches in Categorization*, E. M. Pothos and A. J. Wills, Eds., Cambridge, UK: Cambridge University Press, 2011, pp. 18–39.

[131] S. O'Brien and C. Lauer, "Testing the susceptibility of users to deceptive data visualizations when paired with explanatory text," in *Proceedings of the 36th ACM International Conference on the Design of Communication*, 2018, pp. 1–8.

[132] L. Padilla, "A case for cognitive models in visualization research," 2019.

[133] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci, "Decision making with visualizations: A cognitive framework across disciplines," *Cognitive research: principles and implications*, vol. 3, no. 1, p. 29, 2018.

[134] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini, "How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, ACM, 2015, pp. 1469–1478.

[135] R. Parasuraman and D. H. Manzey, "Complacency and bias in human use of automation: An attentional integration," *Human Factors*, vol. 52, pp. 381–410, 2010.

[136] 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*, vol. 42, pp. 42–58, 2014.

[137] R. Pienta, J. Abello, M. Kahng, and D. H. Chau, "Scalable graph exploration and visualization: Sensemaking challenges and opportunities," in *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, IEEE, 2015, pp. 271–278.

[138] W. A. Pike, J. Stasko, R. Chang, and T. A. O'Connell, "The science of interaction," *Information Visualization*, vol. 8, no. 4, pp. 263–274, 2009.

[139] P. Pirolli and S. Card, "The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis," in *Proceedings of international conference on intelligence analysis*, McLean, VA, USA, vol. 5, 2005, pp. 2–4.

[140] T. J. Pleskac, "Decision and choice: Luce's choice axiom," in *International Encyclopedia of the Social & Behavioral Sciences (2nd. edition)*, J. D. Wright, Ed., Oxford: Elsevier, 2015, pp. 895–900.

[141] M. Pohl, M. Smuc, and E. Mayr, "The User Puzzle – Explaining the Interaction with Visual Analytics Systems," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2908–2916, 2012.

[142] M. I. Posner, "Orienting of attention," *Quarterly Journal of Experimental Psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[143] D. Prelec, H. S. Seung, and J. McCoy, "A solution to the single-question crowd wisdom problem," *Nature*, vol. 541, no. 7638, p. 532, 2017.

[144] 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.

[145] 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*, vol. 22, no. 1, pp. 31–40, 2016.

[146] *Random name generator*, https://github.com/treyhunner/names, Accessed 2019-07-25, 2014.

[147] R. Ratcliff and P. L. Smith, "A comparison of sequential sampling models for two-choice reaction time," *Psychological Review*, vol. 111, no. 2, pp. 333–367, 2004.

[148] B. F. Reskin, D. B. McBrier, and J. A. Kmec, "The determinants and consequences of workplace sex and race composition," *Annual review of sociology*, vol. 25, no. 1, pp. 335–361, 1999.

[149] C. Ridsdale, J. Rothwell, M. Smit, H. Ali-Hassan, M. Bliemel, D. Irvine, D. Kelley, S. Matwin, and B. Wuetherick, "Strategies and best practices for data literacy education: Knowledge synthesis report," 2015.

[150] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[151] V. Romo, "Georgia's governor signs 'fetal heartbeat' abortion law," *NPR*, 2019.

[152] 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*, vol. 20, no. 12, pp. 1604–1613, 2014.

[153] B. Saket, A. Endert, and C. Demiralp, "Task-based effectiveness of basic visualizations," *IEEE transactions on visualization and computer graphics*, 2018.

[154] P. R. Sanday, *Women at the center: Life in a modern matriarchy*. Cornell University Press, 2002.

[155] L. J. Sanna, N. Schwarz, and S. L. Stocker, "When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 28, no. 3, p. 497, 2002.

[156] A. Sarvghad, M. Tory, and N. Mahyar, "Visualizing dimension coverage to support exploratory analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 21–30, 2017.

[157] S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortés, "A survey on metamorphic testing," *IEEE Transactions on software engineering*, vol. 42, no. 9, pp. 805–824, 2016.

[158] P. Sengers, K. Boehner, S. David, and J. Kaye, "Reflective design," in *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, 2005, pp. 49–58.

[159] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *The Craft of Information Visualization*, Elsevier, 2003, pp. 364–371.

[160] D. J. Simons and C. F. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events," *Perception*, vol. 28, no. 9, pp. 1059–1074, 1999.

[161] *Socket.io*, n.d.

[162] J.-H. Song and K. Nakayama, "Hidden cognitive states revealed in choice reaching tasks," *Trends in Cognitive Sciences*, vol. 13, no. 8, pp. 360–366, 2009.

[163] M. J. Spivey and R. Dale, "Continuous dynamics in real-time cognition," *Current Directions in Psychological Science*, vol. 15, no. 5, pp. 207–211, 2006.

[164] R. B. Stacey, "A report on the erroneous fingerprint individualization in the madrid train bombing case," *Journal of Forensic Identification*, vol. 54, no. 6, pp. 706–718, 2004.

[165] K. E. Stanovich and R. F. West, "Advancing the rationality debate," *Behavioral and Brain Sciences*, vol. 23, no. 5, pp. 701–717, 2000.

[166] P. T. Sukumar and R. Metoyer, "A visualization approach to addressing reviewer bias in holistic college admissions," in *Cognitive Biases in Visualizations*, Springer, 2018, pp. 161–175.

[167] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," *arXiv preprint arXiv:1901.10002*, 2019.

[168] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.

[169] 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.

[170] D. A. Szafir, "The good, the bad, and the biased: Five ways visualizations can mislead (and how to fix them)," *interactions*, vol. 25, no. 4, pp. 26–33, 2018.

[171] A. Tang, C. Neustaedter, and S. Greenberg, "Videoarms: Embodiments for mixed presence groupware," in *People and Computers XX—Engage*, Springer, 2007, pp. 85–102.

[172] C. Taylor, "New kinds of literacy, and the world of visual information," *Literacy*, 2003.

[173] R. H. Thaler and C. R. Sunstein, *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[174] A. K. Thomas and P. R. Millar, "Reducing the framing effect in older and younger adults by encouraging analytic processing," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 2, no. 139, 2011.

[175] J. J. Thomas and K. A. Cook, "Visualization Viewpoints: A Visual Analytics Agenda," *IEEE Computer Graphics and Applications*, vol. 26, no. February, pp. 10–13, 2006.

[176] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychological Review*, vol. 113, no. 4, pp. 766–786, 2006.

[177] A. Treisman, "Preattentive processing in vision," *Computer Vision, Graphics, and Image Processing*, vol. 31, no. 2, pp. 156–177, 1985.

[178] J. K. Tsotsos, *A Computational Perspective on Visual Attention*. Cambridge, MA: MIT Press, 2011.

[179] A. Tversky and D. Kahneman, "The framing of decisions and the psychology of choice," *Science*, vol. 211, pp. 453–458, 1981.

[180] A. Tversky and D. Kahneman, "Rational choice and the framing of decisions," *Journal of Business*, S251–S278, 1986.

[181] A. Tversky and D. Kahneman, "Availability: A heuristic for judging frequency and probability," *Cognitive Psychology*, vol. 5, no. 2, pp. 207–232, 1973.

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

[183] S. Ullman, "Visual routines," in *Readings in computer vision*, Elsevier, 1987, pp. 298–328.

[184] A. C. Valdez, M. Ziefle, and M. Sedlmair, "A framework for studying biases in visualization research," in *DECISIVe 2017: Dealing with Cognitive Biases in Visualisations*, 2017.

[185] A. C. Valdez, M. Ziefle, and M. Sedlmair, "Priming and anchoring effects in visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 584–594, 2018.

[186] J. Vandekerckhove, "A cognitive latent variable model for the simultaneous analysis of behavioral and personality data," *Journal of Mathematical Psychology*, vol. 60, pp. 58–71, 2014.

[187] *Vega-lite*, n.d.

[188] H. Wainer, "A test of graphicacy in children," *Applied Psychological Measurement*, vol. 4, no. 3, pp. 331–340, 1980.

[189] E. Wall, A. Arcalgud, K. Gupta, and A. Jo, "A markov model of users' interactive behavior in scatterplots," in *2019 IEEE Visualization Conference (VIS)*, IEEE, 2019, pp. 81–85.

[190] E. Wall, L. Blaha, C. Paul, and A. Endert, "A formative study of interactive bias metrics in visual analytics using anchoring bias," in *IFIP Conference on Human-Computer Interaction*, Springer, 2019, pp. 555–575.

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

[192] 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.

[193] E. Wall, L. M. Blaha, C. L. Paul, K. Cook, and A. Endert, "Four perspectives on human bias in visual analytics," in *Cognitive biases in visualizations*, Springer, 2018, pp. 29–42.

[194] 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*, vol. 24, no. 1, pp. 288–297, 2017.

[195] E. Wall, A. Narechania, J. Paden, and A. Endert, "Left, right, and gender: Visualizing interaction traces to mitigate social biases," *IEEE Transactions on Visualization and Computer Graphics (InfoVis)*, 2020, (under review).

[196] E. Wall, J. Stasko, and A. Endert, "Toward a design space for mitigating cognitive bias in vis," in *2019 IEEE Visualization Conference (VIS)*, IEEE, 2019, pp. 111–115.

[197] W. Willett, J. Heer, and M. Agrawala, "Scented widgets: Improving navigation cues with embedded visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1129–1136, 2007.

[198] J. Xiao, R. Catrambone, and J. Stasko, "Be quiet? evaluating proactive and reactive user interface assistants," in *Proceedings of INTERACT*, vol. 3, 2003, pp. 383–390.

[199] C. Xiong, L. van Weelden, and S. Franconeri, "The curse of knowledge in visual data communication," *IEEE transactions on visualization and computer graphics*, 2019.

[200] K. Xu, S. Attfield, T. Jankun-Kelly, A. Wheat, P. H. Nguyen, and N. Selvaraj, "Analytic provenance for sensemaking: A research agenda," *IEEE Computer Graphics and Applications*, vol. 35, no. 3, pp. 56–64, 2015.