COGNITIVE BIASES IN DECISION-MAKING UNDER UNCERTAINTY WITH INTERACTIVE DATA VISUALIZATIONS

by

Ryan Wesslen

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2021

Approved by:

_____
Dr. Wenwen Dou

_____
Dr. Samira Shaikh

_____
Dr. Isaac Cho

_____
Dr. Douglas Markant

_____
Dr. Jean-Claude Thill

ABSTRACT

RYAN WESSLEN. Cognitive biases in decision-making under uncertainty with interactive data visualizations. (Under the direction of DR. WENWEN DOU)

In this thesis, we hypothesize that data visualization users are subject to systematic errors, or cognitive biases, in decision-making under uncertainty. Based on research from psychology, behavioral economics, and cognitive science, we design five experiments to measure the role of anchoring bias, confirmation bias, belief bias, and myopic loss aversion under different uncertain decision tasks like social media event detection, misinformation identification, and financial portfolio allocation. This thesis makes three major contributions. First, we find evidence of cognitive biases in data visualization through multiple behavioral trace data including user decisions, user hover and click interactions, qualitative feedback, and belief elicitation techniques. Second, we design multiple experiments with interactive data visualization systems across different design complexities (coordinated multiple views to single plot), data types (social network, linguistic, geospatial, temporal, statistical), and evaluate them on user populations that range from novice to expert (crowdsourced, undergraduate, data scientist, domain expert). Third, we evaluate the experiments using multiple statistical and probabilistic techniques to measure the effects of cognitive biases including classical statistical tests, (Bayesian) mixed effects modeling, hierarchical clustering, natural language processing, and Bayesian cognitive modeling. These experiments show the promising role data visualizations and human-computer techniques could mitigate such biases and lead to better decision-making under uncertainty.

ACKNOWLEDGEMENTS

The distributional hypothesis of linguistics states that words' meanings can be derived through the "company they keep" [1]. Similarly, my success is a product of the company I have been fortunate to work with for the research within my research including Wenwen Dou, Jean-Claude Thill, Samira Shaikh, Isaac Cho, Doug Markant, Alireza Karduni, Sashank Santhanam, Svitlana Volkova, Dustin Arendt, George Banks, Haley Woznyj, Roxanne Ross, Tiffany Gallicano, Sara Levens, Min Jiang, Sagar Nandu, Emily Wall, Arpit Narechania, Mohamad Aboufoul, Gabriel Fair, Omar El-Tayeby, Bill Ribarsky, and many others.

I am grateful for funding to support my dissertation research from UNC Charlotte's School of Data Science and Project Mosaic as well as the Pacific Northwest National Laboratory.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI An acronym for Artificial Intelligence

ANOVA An acronym for Analysis of Variance

API An acronym for Confidence Interval or Credible Interval

CGF An acronym for Computer Graphics Forum

CMV An acronym for Coordinated Multiple Views

CRRA An acronym for Coefficients of Relative Risk Aversion

EuroVis An acronym for Eurographics Conference on Visualization

HCAI An acronym for Human-Centered Artificial Intelligence

HCI An acronym for Human-Computer Interaction

HOP An acronym for Hypothetical Outcome Plot

HSD An acronym for Tukey's Honest Significant Difference Test

ICWSM An acronym for International Conference on Web and Social Media

IEEE An acronym for Institute of Electrical and Electronics Engineers

InfoVis An acronym for Information Visualization or IEEE Conference on Information Visualization

LDA An acronym for Latent Dirichlet Allocation

MCMC An acronym for Markov Chain Monte Carlo

MCMC-P An acronym for Markov Chain Monte Carlo for People

MPT An acronym for Modern Portfolio Theory

MTurk An acronym for Amazon Mechanical Turk

NPMI An acronym for Normalized Pointwise Mutual Information

pLSI An acronym for Probabilistic Latent Semantic Indexing

RCE An acronym for Randomized Controlled Experiments

S&P 500 An acronym for Standard and Poors 500 Stock Index

STM An acronym for Structural Topic Modeling

VA An acronym for Visual Analytics

VaR An acronym for Value-at-Risk

VAST An acronym for IEEE Conference on Visual Analytics Science and Technology

CHAPTER 1: INTRODUCTION

Recent breakthroughs in machine learning, deep learning, and artificial intelligence has provided ample gains in the emerging field of data science. However, these advances have also led to many emerging issues surrounding the use of artificial intelligence systems like algorithmic bias [6, 7], a need for explainability [8], a lack of causality [9], privacy concerns on extraction attacks [10], a lack of human control [11], environmental costs [12], and negative societal impacts of news feed algorithms like misinformation on social media [13, 14]. To address these concerns, Ben Shneiderman has introduced a new human-centered artificial intelligence (HCAI) framework to provide reliable, safe, and trustworthy systems [15, 16]. Whereas much research in artificial intelligence (AI) systems have focused on ways to replace human decision-making, a HCAI approach focuses on how to "augment, amplify, empower, and enhance humans rather than replace them". HCAI framework has the primary objective to achieve both high levels of human control as well as high levels of automation [15]. However, one challenge to integrating more human-control into future AI systems is that past research in psychology, cognitive science, and behavioral economics has identified that humans are susceptible to cognitive biases, or systematic errors in judgement [17, 18, 19, 20, 21]. Recently, research in the fields of visual analytics, information visualization and human-computer interaction have investigated in how cognitive biases can affect the decision-making process within interactive data visualization systems [22, 23, 24]. Our purpose in this dissertation is that by expanding the research in cognitive biases in such systems, HCAI designers can better measure, avoid, and possibly mitigate such cognitive biases in order to develop more robust systems in the future.

This thesis focuses within the fields of visual analytics, information visualization and human-computer interaction, which forms the basis of the study of interactive data visualizations. Figure 1.1 outlines the research workflow explored in this thesis. First, we use research from fields of psychology, cognitive science, and behavioral economics have identified cognitive biases that are systematic errors in decision-making. We use such theories to motivate the platforms we develop to identify and measure the effects of such cognitive biases within decision-support tasks using interactive data visualizations. The next step is system development which includes user interface (web) development, machine learning training, and back end engineering (e.g., database, API) to develop end-to-end custom interfaces and experiment apparatus. Third, we carefully design randomized controlled experiments within laboratory or through crowdsourced platforms (e.g., Amazon's Mechanical Turk) to identify treatment effects from the experiments. In addition, we use a several techniques in statistics and causal inference to appropriately isolate treatment effects including classical statistical tests (e.g., ANOVA, t-tests, linear regression) and mixed effects modeling. Further, to provide context and user strategies for user behaviors, we also analyze users' open-ended feedback using traditional qualitative text analyses as well as computer-assisted text approaches like topic modeling. In the long term, such experiments would feedback to modify and develop new theories. This thesis focuses more on the first three steps while newer frameworks [25] provide future work to refine and develop theories on graphical inference which we discuss in the discussion section.

1.1 Dissertation Outline

Chapter 2 begins with an outline of background and related work. The remainder of the dissertation is organized into two parts: investigations on the transference of cognitive biases within visual analytic systems followed by experiments in information visualization that study simplified tasks using newer uncertainty visualization representations.

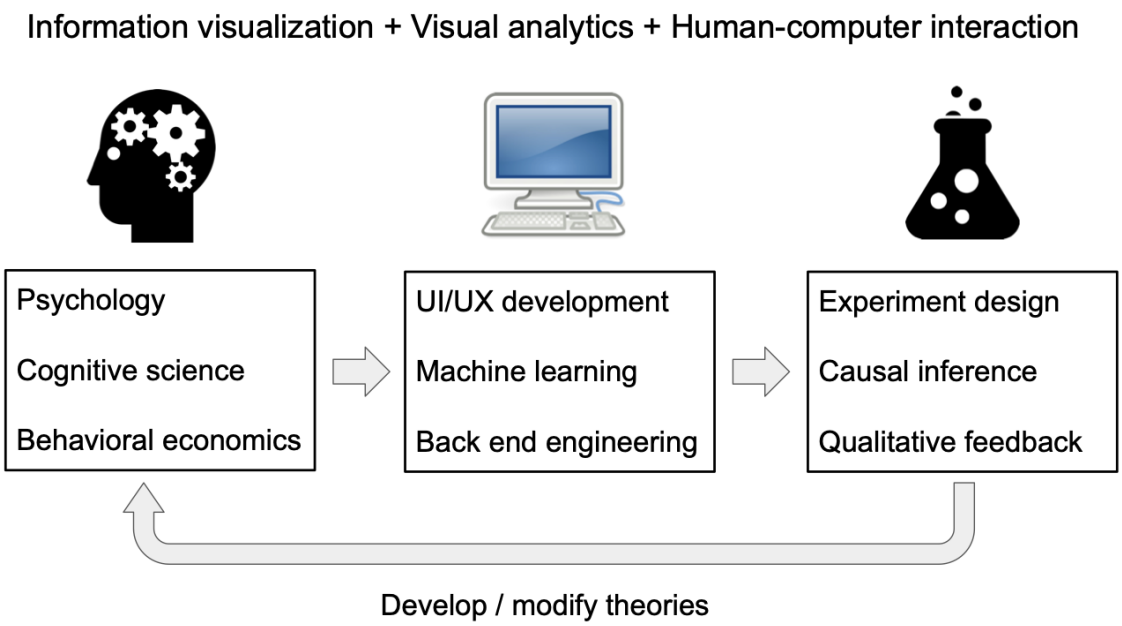


Figure 1.1: Thesis research workflow

1.1.1 Cognitive biases in visual analytics systems

Chapters 3-5 consider three different experiments that investigate the role of two cognitive biases (anchoring [26, 27] and confirmation bias [28, 19]) within two visual analytics systems. Chapter 3 investigates the role of anchoring bias within a social media event detection system, Crystal Ball [29], and explores if similar notions of anchoring can effect users decisions and interactions patterns in a controlled in-laboratory experiment. Chapter 4 provides an in-laboratory experiment to measure the effects of confirmation bias on data visualization users decisions regarding misinformation detection on social media using the Verifi system [30]. Chapter 5 combines elements of Chapter 3 and 4 by studying the role of anchoring effects using the Verifi system for the task of misinformation detection; however, it expands the studies from Chapter 3 and 4 by examining the role strategy cues provided to participants may interact with anchoring effects participants view during training.

1.1.2 Uncertainty visualizations in information visualization representations

Chapter 6 and 7 extend work from Chapters 3-5 but consider simplified user interfaces and test different but related cognitive biases on crowdsourced experiments. Chapter 6 considers the role of belief bias [20] on participants correlation beliefs through a controlled experiment and using Bayesian cognitive modeling as an approximate normative judgement to compare how participants should update their beliefs for given hypothetical correlation variable pairs. Chapter 7 then explores the effect of uncertainty visualizations [31] on myopic loss aversion for retirement investment decision-making by replicating a past behavioral economics study [3] and extending it with newer uncertainty representations to investigate if such representations can mitigate myopic loss aversion.

We then conclude the dissertation with a discussion on major takeaways from this thesis as well as ideas of future extension of this research.

1.2 Thesis Statement and Contributions

In this thesis, we hypothesize that data visualization users are subject to systematic errors, or cognitive biases, in decision-making under uncertainty. Based on research from psychology, behavioral economics, and cognitive science, we design five experiments to measure the role of anchoring bias, confirmation bias, and myopic loss aversion under different uncertain decision tasks like social media event detection, misinformation identification, and financial portfolio allocation. This thesis makes three major contributions. First, we find evidence of cognitive biases in data visualization through multiple behavioral trace data including user decisions, interactions logs (hovers, clicks), qualitative feedback, and belief elicitation techniques. Second, we design multiple experiments with interactive data visualization systems across different design complexities (coordinated multiple views to single plot), data types (social network, linguistic, geospatial, temporal, statistical), and evaluate them on user

populations that range from novice to expert (crowdsourced, undergraduate, data scientist, domain expert). Third, we evaluate the experiments using multiple statistical and probabilistic techniques to measure the effects of cognitive biases including classical statistical tests, (Bayesian) mixed effects modeling, hierarchical clustering, natural language processing, and Bayesian cognitive modeling. These experiments show the promising role data visualizations and human-computer techniques could mitigate such biases and lead to better decision-making under uncertainty.

1.3 Prior Publications and Authorship

Although I made significant contributions of the research in this dissertation, much of the work was done in collaboration with my advisor Wenwen Dou and my dissertation committee members Samira Shaikh, Isaac Cho, and Douglas Markant. Significant contributions were made as well by my co-author Alireza Karduni along with additional collaborators on the work including Sashank Santhanam, Svitlana Volkova, and Dustin Arendt. Chapter 3 was published at the IEEE VAST 2017 conference [32] along with a related co-authored paper [29] published in the same venue. Chapter 4 was published in AAAI ICWSM 2018 [30]. Chapter 5 was published in IEEE EuroVis 2019 [33] as well as in the Journal of Computer Graphics Forum (CGF). Chapter 6 was published in IEEE InfoVis 2020 [34] with an accompanying publication in the 2021 Special Edition of IEEE Transactions on Visualization and Computer Graphics. Last, Chapter 7 will be published in IEEE Vis 2021 along with a follow up in the 2022 Special Edition of IEEE Transactions on Visualization and Computer Graphics [35]. In this thesis, I use the first person plural to reflect my collaborators' contributions.

CHAPTER 2: RELATED WORK

2.1 Related Work

2.1.1 Introduction

The cognitive revolution of the 1950s enabled the mathematical formalism of cognitive processes to develop testable hypotheses of human behavior. Unlike most attempts in computer science to predict behavior based solely on past actions (i.e., behaviorism), computation-based approaches to understanding decision-making, and thus behavior, offer a possible second revolution [36]. One area where such a computation revolution may provide many gains is decision-making [37] with interactive visualizations for the Information Visualization (InfoVis) [38] and Visual Analytics (VA) [39] research communities. Developed as an extension of human-computer interaction, interactive data visualizations are tools that amplify human cognition through the use of abstract visual representations of data [40, 41, 42]. While data visualization research has a long history of studying low-level perception, InfoVis and VA research rarely study cognition (i.e., high-level, complex decision-making) or use cognitive modeling to understand how individuals make decisions [43]. Nevertheless, adopting cognitive modeling for data visualization research can provide many opportunities to accelerate innovation, improve validity, and facilitate replication efforts [44]. Early examples to understand cognitive processes while using data visualizations consider insight-based approaches [45], top-down modeling [46], or visual attention coupled with decision-making [47].

An important theme to understanding decision-making with visualizations is the classic exploration-exploitation trade-off [48]. Motivated by foraging theory [49], this

problem interprets decision-making as a search process between gathering (exploration) and using (exploitation) information [50]. Interactive visualizations are an application of two such search sub-problems: visual search and information search. Sometimes linked to the idea of multi-armed bandits [50] (i.e., time allocation problem where each choice is uncertain), this problem has direct influence in interactive visualization where the user must decide how to use the interface's flexible views in making a decision. However, such flexibility comes at a cost.

A second problem results from the idea that too much flexibility (i.e., degrees of freedom) in decision-making can result in garden of forking paths. The problem of forking paths is tied to statistics' problem of multiple comparisons [51], which is a reinterpretation of the exploration-exploitation problem as a dual process between exploratory and confirmatory analysis.[1] More recently, this problem has been extended to a high rate of false positive decisions with interactive visualizations where the user has too much freedom [52, 53, 54]. The forking paths problem is a question of how users update their mental model as presented with new data in an interactive visualization. In this way, Pu and Kay [52] suggest that cognitive biases may explain user susceptibility to the forking path problem in visualizations.

Cognitive biases are systematic errors in judgment that have been long studied by cognitive psychologists and social scientists to understand how and why individuals sometimes make consistent errors in decision-making. Recently, visualization researchers have explored whether cognitive biases transfer to visualization decision-making [23, 24, 55] and, if such biases can be identified, could such findings inform the design of visualizations systems to debias or mitigate such effects [56, 57, 58, 59]. If such well-designed systems can help users to find the right explore-exploit mix, ideally such a system would safeguard against possible forking path problems and mitigate systematic biases with the ultimate goal of enabling better overall decision-making.

[1] Multiple comparisons problem is sometimes referred to as "p-hacking".

2.1.2 Cognitive Biases

Classical notions of rational behavior include adherence to expected utility theory, the laws of logic, and that uncertainty is measured through probability [60]. Under these definitions, cognitive biases are systematic deviations, or errors, from such normative behaviors. However, there are multiple perspectives towards understanding cognitive biases [61].

The *cognitive-psychological* perspective attributes biases to bounded rationality [62], or the use of heuristics due to limited time or mental processing ability. Classical examples extend from work by Tversky and Kahneman include availability bias [63], anchoring, and representativeness [26]. This approach has been extended to the dual-process framework of heuristic-deliberate that differentiate between fast, intuitive, emotional thinking (type 1) and slow, deliberate and analytic thinking (type 2) [18]. Alternatively, the *ecological-perspective* on cognitive biases views heuristics optimistically as effective tools for decision-making [64]. This approach attributes cognitive biases as the result of heuristics and the applied context. In this view, biases occur when people apply experience-based heuristics to unfamiliar situations that conflict with individuals' mental models [65]. A third framework is the *evolutionary perspective* that posits that cognitive biases are a mismatch between evolutionarily developed heuristics [66] and an agent's environment [67]. Under this framework, heuristics that may have provided advantages from an evolutionary perspective may not longer have relevance to modern environments. This approach is similar to ecological perspective but differs in attributing to the origin of cognitive biases: ecological are gained from experience while evolutionary are acquired through genetics.[2]

Nevertheless, there are many different examples of cognitive biases [24]

[2]More recently, a fourth perspective on cognitive biases, a neural network approach, has been posited by Korteling, Brouer, and Toet [61]. This approach complements and extends the other three perspectives rather than replaces them.

2.1.2.1 Anchoring Bias

Anchoring bias is the tendency for an initial piece of information, relevant or not, to affect a decision-making process. Typically, anchoring bias is associated the anchoring-and-adjustment heuristic. Anchoring-and-adjustment heuristic consists of a two step process [27]. In the first step, a person will develop their estimate, or anchor, of an open-ended question. In the second step, the person will adjust her estimate as new information is processed. Error occurs when she does fails to make a sufficient adjustment to the correct answer. In the classical studies by Tversky and Kahneman [26], participants were asked to estimate the number of African countries that are members of the United Nations. Using a wheel to generate a random number to serve as the anchor, participants final estimate was shown to be affected by the wheel's random number, even though it had no relevance to the question at hand. Extending the work by Tversky and Kahneman, psychological studies have investigated interacting conditions that can modify the effects of anchoring through modifying financial incentives [27], cognitive load [27], anchor extremity [68], and uncertainty/knowledge [69, 70]. Yet more recently, anchoring and other cognitive biases have started to be explored by visualization researchers, who are interested in whether cognitive biases like anchoring can transfer to decision-making using interactive visualizations. This work motivates the experiments in Chapters 3 and 5.

2.1.2.2 Confirmation Bias

Past research in psychology has found that individuals exhibit a tendency to treat evidence in a biased manner during their decision-making process in an effort to protect their beliefs or pre-conceived hypothesis [71], sometimes even in situations where they have no vested interest or personal stake [19]. Research has shown that this tendency, known as confirmation bias, can cause systematic errors in individual decision-making [20]. Confirmation bias can be expressed *"either as a failure to seek*

or utilize data which are inconsistent with a single hypothesis under test" or *"it may be expressed as a failure to seek or utilize evidence for alternative hypotheses"* [72]. Past research demonstrates that confirmation bias affects decision making process in contexts like policy rationalization, judicial reasoning, medicine, and science [19]. Classic laboratory experiments to study confirmation bias typically present participants with a hypothesis and evidence that either confirms or disconfirms their hypothesis, and may include cues that cause uncertainty in interpretation of that given evidence. These studies motivate the design of the studies in Chapter 4.

2.1.2.3 Belief Bias

Research in psychology shows that prior beliefs have a strong influence on people's interpretation of uncertain data [73, 74, 75, 76], especially for correlations [77, 78]. A central theory that explains why prior beliefs are important is the dual-process account of reasoning [79, 18]. This theory posits that fast heuristic processes (System 1) competes with slower analytic processes (System 2) that can affect logical decisions. Evans *et al.* [80] suggested that belief bias [79, 73] could occur as "within-participant conflict" between the two systems when participants tend to agree with an argument based on whether or not they agree with the conclusion rather than its logical conclusion. Alternatively, other research focused on theory-motivated reasoning bias based on "congruent" and "incongruent" evidence relative to an individuals' belief systems [74]. These theories motivate design aspects for the studies in Chapter 6.

2.1.2.4 Myopic Loss Aversion

Financial decision-making is central to decisions like retirement investing as typical investors make decisions about how to allocate funds across a wide range of assets that vary in risk (e.g., stocks vs. bonds). A seminal study by Mehra and Prescott [4] found a surprising reluctance to take on risk, in that standard economic models could not account for the large historical premium for riskier investments (the "equity

premium puzzle"). Benartzi and Thaler [21] theorized that individuals deviate from the predictions of neoclassical economic theory due to two factors, an oversensitivity to the possibility of losses, and evaluation of returns over short time periods, a combination they referred to as *myopic loss aversion*. Benartzi and Thaler [3] showed that myopic loss aversion emerges when making investment decisions with simple visualizations of the distribution of returns (i.e., bar charts). They found that investors allocated less in stocks when shown returns over a 1-year evaluation period due to aversion to short-term losses. Their results suggest that the method for visualizing investment performance can have a dramatic effect on individuals' willingness to take on risk. This work motivates the experiment presented in Chapter 7.

2.1.3 Cognitive Biases in Visualizations

Cognitive bias research in visualizations can be divided into frameworks and empirical studies.

2.1.3.1 Empirical

Empirical studies on cognitive biases in data visualizations are relatively new, beginning around 2015-2016. Most studies attempt to demonstrate evidence of traditional cognitive biases within data visualization user studies. One of the earliest mentions of cognitive biases within the visualization community comes from Ellis and Dix [22]. They posit possible cognitive biases relevant to data visualization decision-making and suggest possible case studies. Following this introduction, data visualization research on cognitive biases have developed studies to explore biases like attraction [81, 58], selection [82], availability [83], anchoring [32, 84, 33, 85, 86] and confirmation bias [30, 87]. However, with the exception of Dimara *et al.* [58] and Gotz *et al.* [82], these studies have largely been exploratory in nature with the attempt to measure evidence of the existence of such biases, with no experiments on intervening and demonstrating how visualizations can mitigate against such biases. Most of

these studies were motivated by past psychology studies without theoretical models on why such biases exist in visualizations. To address this problem, recent theoretical frameworks have been introduced on tasks, interaction metrics, or different types of biases (e.g., perceptual and social) with data visualizations.

2.1.3.2 Frameworks

Past explanatory frameworks for cognitive biases by cognitive psychologists divide biases on why they occur [24]. Two examples of such frameworks include Baron [88] and Pohl [89]. While beneficial for researchers across disciplines, for the data visualization community, there's no comprehensive review of cognitive biases within that community. In a first attempt to address this problem, Dimara *et al.* [24] provide a taxonomy of experimental tasks to measure cognitive biases within visualization research. Using 176 suggested cognitive biases, they classify these biases into one of seven task-based categories and qualitatively rated each by its possible impact in data visualizations.

Three alternative frameworks have been introduced to consider cognitive biases more broadly across individual biases. Wall *et al.* [23] provide a framework for measuring cognitive biases within visualizations through the development of metrics for user's interaction logs. Specifically, they introduce two broad categories of metrics: coverage and attribute metrics. Second, Calero Valdez, Ziefle, and Sedlmair [55] propose a three-tier model of perception, action, and social that corresponds to different methods in data visualization research to study biases at each tier. Last, Parsons [59] explored the role *external representations* [90] can facilitate representational biases due to differences in individual *representational fluency*, i.e., knowledge and skills to understand different representations. In this view, the best mitigate against such biases is not individual bias-level solutions (e.g., mitigation specifically for anchoring effects), but instead visualization designers to put more effort in education, training, and practice to ensure all users have sufficient representational fluency.

More recently, additional research has expanded to consider the design space of mitigating biases with additional use cases. Wall, Endert and Stasko [91] introduced eight dimensions that are part of two core components (system and context) of a visual interactive system that can be manipulated to mitigate such biases. While work in mitigating biases has been sparse, this work provides an opportunity for future studies to assess these strategies.

2.1.3.3 Cognitive Modeling in Visualizations

Cognitive modeling in visualization initially was studied as a subset of visuospatial reasoning in how individuals derive meaning from external visual representations [92]. Visualization researchers have integrated similar ideas to understand visualization cognitive processes through insight-based approaches [45] and top-down modeling [93, 46]. More recently, InfoVis researchers have used Bayesian models to understand cognitive processing of visualizations [94, 95]. Cognitive scientists have demonstrated the importance of Bayesian modeling to understanding individual decision-making [96, 97]. In this approach, an individual has some prior belief that is updated when the individual consumes additional data, resulting in their posterior beliefs. Bayesian cognition models have been used to understand deviations from optimal belief updating due to conservatism, sample-based inference (approximation) and "resource-rational" interpretations of cognitive bias [60].

Two InfoVis studies [94, 95] have combined belief elicitation with a Bayesian cognitive modeling framework. Wu *et al.* [94] examined whether people integrated prior probabilities with data in an optimal manner. They found that priors influenced predictions in a manner consistent with Bayesian inference, although to a lesser extent than predicted by the model. However, a limitation to this study was that participants were given a prior; therefore, prior beliefs cannot be examined. In contrast, Kim *et al.* [95] empirically measured participants' prior beliefs about the a target proportional quantity and used those priors to calculate the normative posterior given

the data that was presented. In aggregate, participants' judgments were consistent with predictions derived from Bayesian inference, although less so for large data sets. However, participants expressed greater uncertainty in their judgments than expected from the Bayesian model. Further, the authors connect such Bayesian modeling and belief elicitation with recent research on visualizing uncertainty through techniques like HOPs [98, 99, 100].

2.1.4 Exploration-Exploitation

While new to data visualization research, the rational modeling approach for optimal exploration-exploitation decisions within Human-Computer Interaction (user interfaces) are not new. Early versions appeared in the early 1990s and were extensions of the early rational analysis approach outlined by Anderson [101]. For example, Rehder *et al.* [102] used the Anderson decision-making framework to develop a simple model of how a user will scan through commands until she finds the most relevant command to execute. In this approach, the authors framed a user's decision-making process as a choice between (1) executing the current command (i.e., exploitation) and (2) exploring another command in the chance of finding a better command (i.e., exploration). To model this decision, the authors formalized two cost functions for either scan (i.e., explore more) or execute (i.e., act) that are a function of the user's perceived commands' relevance, cost of moving and undoing execution, and the conditional probabilities of each hypothesis given its relevance. Under several assumptions, the user will either scan (explore) or act (exploit) based on whichever function has the lowest cost.

2.1.5 Decision-Making Framework

Another important factor would be incorporating a decision-theoretic modeling framework within visualization empirical studies [103].[3] Although not new to the vi-

[3]Hullman *et al.* [103] recommendation of decision frameworks is in the context of uncertainty in visualizations however their argument can apply to evaluating any visualization system.

sualization community (e.g., [5]), decision frameworks offer two significant advantages: increased participant motivation and utility-maximization frameworks to determine optimal decisions to use as benchmarks. First, incentives (i.e., extrinsic awards) can provide higher participation effort and, thus, performance. Such extrinsic motivating factors generally are monetary awards linked to better performance; however, points through gamification can provide more subtle incentives as well. Further evidence from neuroscience has also found that incentive manipulations can improve participant motivation [104]. Second, by creating such incentives, it's easier to quantify rewards as utility, thus enabling Von Neumann-Morgenstern utility functions [105]. This lends itself to the incorporation of probabilities and infers rational choice in which behaviors can be interpreted as individual "utility-maximizing."

2.1.6 Visualizing Uncertainty

A related and important research area is communicating uncertainty to individuals through visualizations. Recently, a significant amount of research within the InfoVis and VA community has gone to better understand different ways to incorporate uncertainty within visualization systems. To organize such approaches, Hullman *et al.* [103] provide a six-level taxonomy for measuring uncertainty in data visualization literature.

There are three important reasons uncertainty representations are important for interactive data visualizations. First, visualizations can aid decision-making under uncertainty by conveying measurements of uncertainty without requiring core statistical knowledge. Tsai, Miller, and Kirlik [106] find empirical evidence for the ability of interactive visualizations to convey Bayesian reasoning for Bayes-naïve participants (i.e., individuals who are unfamiliar with Bayes theorem). Second, uncertainty can be measured as social information based on others' performance on specified task. Kim, Reinick, and Hullman [107] provide a controlled experiment to measure the effect of showing other participants' expectations on user's ability to recall the data, the extent

to which they adjust their expectations to align with the data, and their trust in data accuracy. Last, a better understanding of how individuals perceive uncertainty can yield better design guidelines for the development of interfaces that consider uncertainty. For example, Greis *et al.* [108] provide design guidelines for HCI researchers on evaluating whether and how to present uncertainty within visualizations.

The study most relevant for this context was Micallef, Dragicevic, and Fekete [109] in which the authors conducted a controlled crowdsourced experiment to evaluate participants' ability to engage in Bayesian inference through different visualizations. In their experiment, the authors create a series of variations of Euler diagrams [110, 111] to test participants' ability in Bayesian reasoning to the class mammography problem [112]. In their study, they found that participants' Bayesian reasoning was lower past experiments and that visualizations exhibited no measurable benefit. Moreover, in a second round they found even providing additional text for context to the visualization did not provide any additional value. However, they did find some evidence that visualizations may provide value when text is provided but without explicit numerical values. Ultimately, the authors' argued for the need of much more experiments to better understand the context of this problem, especially when these problems are applied to non-experts with diverse populations (like crowdsourcing sites like MTurk).

CHAPTER 3: THE ANCHORING EFFECT IN DECISION-MAKING WITH VISUAL ANALYTICS.

3.1 Introduction

Researchers in multiple fields, including psychology, economics and medicine have extensively studied the effect of cognitive biases on decision making [113, 114, 115]. Cognitive biases are rules of thumb or heuristics that help us make sense of the world and reach decisions with relative speed [116]. Decision making, the process of identifying solutions to complex problems by evaluating multiple alternatives [117] has been increasingly exacerbated due to explosion of big data [118]. To facilitate human decision-making processes on large and complex datasets, Visual Analytics (VA) combines automated analysis techniques with interactive visualizations to increase the amount of data users can effectively work with [119]. Evidently, the effectiveness of VA to support decision making is an area that warrants study. Our goal in this work is therefore to conduct a study which incorporates three complementary strands of research, given the premises that VA supports decision making, and that decision making is impacted by cognitive biases. Specifically, we investigate how users' decision making processes are impacted by cognitive biases when using VA systems to analyze large and complex datasets. Moreover, we explore if and how cognitive biases are reflected in the way that users interact with visual analytic interfaces.

In the context of VA research, many recent VA systems [120, 121, 122, 123, 124] designed to facilitate the decision making on large and complex datasets contain coordinated and multiple views (CMV). By presenting different visual representations that show various aspects of the underlying data and automatically coordinating operations between views, multiple coordinated views support exploratory analysis

to enable insight and knowledge discovery [125]. In visual interfaces that employ CMV design, users often have choices on which views serve as primary vs. supporting views for their analysis and on the strategies to switch between different views.

The flexibility of visual interfaces with coordinated and multiple views make cognitive biases such as anchoring bias particularly relevant to study. People find cognitive biases to be useful heuristics when sorting through large amounts of information, when task constraints or instructions prime them to focus on specific types of information, or when asked to make quick decisions and analyses. This has been demonstrated for several biases and shown that biases affect decision-making processes in predictably faulty ways that can result in decision-making failures when information is discounted, misinterpreted, or ignored [18]. Additionally, the biases affect not only regular users, but also expert users, when thinking intuitively [18]. One type of bias, the anchoring effect describes the human tendency to rely too heavily on one/the first piece of information offered (the "anchor") when making decisions [115]. Research has demonstrated that individuals anchor on a readily accessible value and adjust from it to estimate the true value, often with insufficient adjustments. For instance, if a person is asked to estimate the length of the Mississippi River, following a question on whether the length is longer or shorter than 500 miles, their answer will be adjusted from the 'anchor' value of 500 miles and will underestimate the true length of the Mississippi River. The effect of such anchors have been extensively studied in multiple tasks in the laboratory and in the field (for a detailed review see [126]). However, the effect of anchoring in Visual Analytics interfaces have not been systematically studied. More importantly, the effect of anchoring bias on the strategies that users deploy to interact with the visual interface and their analysis outcomes remains an open question.

In this chapter, we study the effect of anchoring on users' exploration processes and outcomes. When interacting with visual interfaces employing CMV design, there

is a possibility that users rely too heavily on one particular view. The reasons for such reliance include but are not limited to prior experience, familiarity with certain visualizations, and different ways they were trained to use the visual interface. The significance and impact of such anchoring is the subject of our study.

Prior work in the VA community provides empirical data on cognitive costs of visual comparisons and context switching in coordinated-multiple-view visual interfaces [127, 128]. Findings from these experiments inform design guidelines of CMVs. However, there is little research on how cognitive biases transfer to visualizations, in particular to visual interfaces with coordinated multiple views. MacEachren [129] argues that prior efforts in visualization of uncertainty deal with representation of data uncertainty, but do not address the reasoning that takes place under these conditions. We therefore aim to investigate the impact of anchoring effect on human decision-making processes when using VA systems, because it has been shown to be overwhelmingly affect decision-making [115]. Our experiment design addresses several challenging requirements that are necessary to derive meaningful implications: first, the experiments need to be conducted using a VA system with tasks relevant to decision-making based on large and complex datasets; second, measures and experiment data that reflect users' decision making processes (beyond task completion time and accuracy) need to be collected; third, novel analyses methods need to be developed to tease out the effect of anchoring bias on decision making with VA systems. Accordingly, our work makes the following original contributions:

- To situate our study in complex decision making tasks with visual interfaces, the experiments are conducted with a sophisticated visual analytics system [130] with multiple coordinated views. The design of the visual analytics system enables the visual anchor on either geo or time related representation through tutorial/training.

- In order to study the effect of anchoring bias on the decision-making processes

with greater nuance and granularity, we collect not only quantitative measures about users' performance, including questionnaire responses, but we also collect detailed interaction logs within the visual interface. The interaction logs capture the decision-making process at a action level. Significant differences in actions were found between subjects assigned to different visual anchors.

- In addition to running statistical tests on the quantitative measures collected through pre- and post-questionnaires, we apply two novel methods of analysis - graph analysis and structural topic modeling - to analyze the paths and patterns of users interactions and identify the effect of anchoring bias. Our analysis revealed that visual anchors impact users' decision-making processes while numerical anchors affect the analysis outcomes.

3.2 Background and Related Work

In this section, we describe literature in the areas relevant to our study.

3.2.1 Background on Anchoring Effect

Humans have the tendency to rely on heuristics to make judgments, which can lead to efficient and accurate decisions [131], however these heuristics may also lead to systematic errors known as cognitive biases [18]. Psychologists have long studied the presence of cognitive biases in human decision making process [132, 18]. The anchoring and adjustment bias, defined as *the inability of people to make sufficient adjustments starting from an initial value to yield a final answer* [132], is one of the most studied cognitive biases that can lead individuals to make sub-optimal decisions. In the classic study by Tversky and Kahneman [132], the authors found evidence that when individuals are asked to form estimates, they typically start with an easily accessible value or reference point and make adjustments from this value. While such an approach may not always lead to sub-optimal decisions, research has demonstrated that individuals typically fail to adjust their estimates away from their initial start-

ing point the *anchor*. Research has shown that anchoring affects decision making in various contexts, including judicial sentencing [133], negotiations [134] and medical diagnoses [113]. Given this documented prevalence of anchoring bias in various contexts of decisionmaking activities, we hypothesize that such effects may also be present when individuals interact with data while using visual analytics.

3.2.2 Visual Analytics and Cognitive Biases

Sacha *et al.* [135] investigate how uncertainties can propagate through visual analytics systems and examine the role of cognitive biases in understanding uncertainties, and also suggest guidelines for the design of VA systems that may further facilitate human decisionmaking. Similarly, research in the detection of biased decision making with VA software is in the early stages [136]. Harrison *et al.* found through a crowd-sourcing experiment that affective priming can influence accuracy in common graphical perception tasks [137]. George *et al.* [138] examined robustness of anchoring and adjustment effect in the context of decision support systems. Although their study revealed the presence of anchoring bias in the user's decision making task of estimating the price of house, their decision support system did not contain a highly complex visual interface consisting of coordinated multiple view. Researchers have also investigated the role of various other biases such as confirmation bias [139] and attraction effect [81] in the context of visual analytics. Dimara *et al.* [81] studied attraction effect using crowdsourcing experiments to determine that attraction bias did in fact generalize to information visualization and that irrelevant alternatives may influence users' choice in scatterplots. Their findings provide implications for future research on how to possibly alleviate attraction effect when designing information visualization plots but no study to date has explored the anchoring bias in visual interfaces. Additionally, no research to date has examined the interaction patterns and activities of users in decisionmaking while these users are explicitly anchored under controlled experimental conditions.

In the next section, we describe a novel approach to analyzing the users' interaction patterns which is grounded in the analysis of web log data.

3.2.3 Use of Topic Models for Analyzing Web Logs

For our analysis of the interaction logs, we employ a variant of topic models, structural topic modeling (STM), that facilitates testing the effect of document-level variables on topic proportions. By characterizing the temporal sequence of actions taken by the user during their interactions with the interface as a 'text document' and characterizing actions as 'topics', we are able to test the effects of several factors, which include not only demographic variables such as age and gender, but also the effects of anchoring bias on the user's actions, and hence their decision-making processes. Although topic models have been used to analyze web logs previously, our application of STM to user interaction logs is novel by providing a mechanism to test the effect of independent variables on actions (topic proportions). Early applications of topics models [140, 141] to analyze web log behavior used probabilistic latent semantic indexing (pLSI) [142], a predecessor model to LDA-based topic models [143]. In the case of analyzing web log data, the pLSI model has been helpful in capturing meaningful clusters of users' actions, and found to surpass state-of-the-art methods in generating user recommendations for Google News [141].

One limitation of this method was that it did not consider time or user-level attributes (independent variables) within the model. To address the issue of time, Iwata *et al.* [144] created a LDA-based topic model (Topic Tracking Model) to identify trends of individuals' web logs on two large consumer purchase behavior datasets. In their model, they created a time component to identify the dynamic and temporal nature of topics. As we will discuss in section 5, we address the same concern by creating a time component in our STM model. Further, we employ STM's flexible causal inference framework as a mechanism to test anchor bias by treating each anchor group as additional independent variables.

3.3 User Experiment

In this section, we first describe our research questions and provide a detailed description of the visual analytic system used in the experiment. We then describe the experiment design rationale and tasks designed to elicit and test anchoring bias, and provide details about the experimental procedures and participants next.

3.3.1 Research Questions

Given that our research lies at the intersection of anchoring bias, decision making processes, and visual analytics systems, we designed two types of anchors, namely visual and numerical to evaluate their effects in the context of visual analytics systems. **The numerical anchor** is based on many psychology studies to test whether the participants can adjust away from the numerical anchor in their final answers. **The visual anchor** is designed specifically to prime people with different views in visual analytics interfaces with CMV design. The design of the numerical anchors is to evaluate if users are subject to anchoring bias when using visual analytics interfaces to aid decision making in a way similar to what's found by previous experiments conducted without the use of a visual analytics interface; while the design of the visual anchors is to test specifically whether users can be anchored visually and how that affects the analysis process and outcome. More specifically, we seek to answer the following research questions with respect to the impact of anchoring on decision-making activities using visual analytics systems:

- RQ1 - Visual Anchor: Can individuals be anchored on a specific view in a CMV?
- RQ2 - Numerical Anchor: Are the effects of numerical priming transferable to VA?
- RQ3 - Interaction Patterns I: How does anchoring influence the *paths* of interactions?

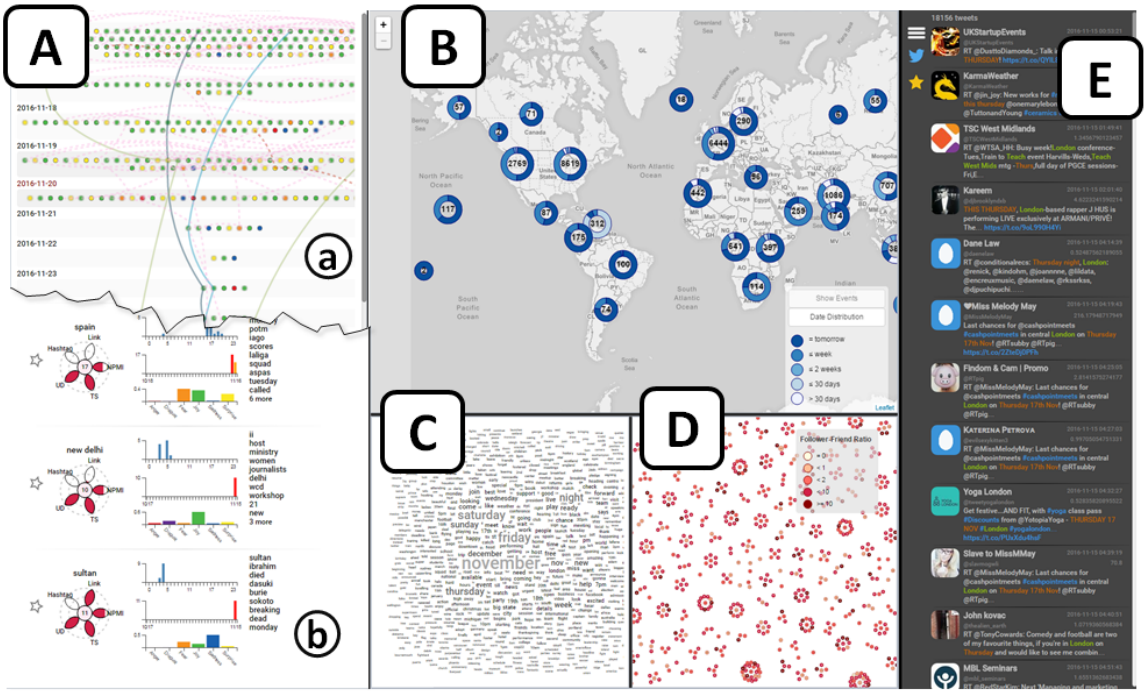


Figure 3.1: Crystal Ball interface: the interface has 4 main views: (A) calender view, (B) map view, (C) word cloud view, and (D) social network view. The calendar view shows the future event overview (a) by default. The event list (b) is shown when the user selects a subset of future events. The tweet panel (E) is shown when the user clicks the Twitter icon.

- RQ4 - Interaction Patterns II: Are there *systematic differences* in the interaction patterns and information seeking activities of individuals primed by different anchors?

To answer these questions, we designed and conducted a controlled experiment using a custom VA system, which is described next.

3.3.2 Crystal Ball - a visual analytics system used for the experiment

To study the anchoring effect during complex decision making tasks performed in a visual analytics system, we conduct the experiment with Crystal Ball, a visual analytics system that facilitates users in identifying future events from Twitter streams [130]. Detecting future events from tweets is a challenging problem as the signals of future events are often overwhelmed by the discussion of on-going events. Crystal

Ball is designed to detect possible future events from streaming tweets by extracting multiple features and enables users to identify potentially impactful events.

3.3.2.1 Analyzing large and noisy Twitter data

On average, around 500 million tweets are posted on Twitter per day by more than 300 million Twitter users [145]. However, many of them discuss past and ongoing events, and news headlines. To find, identify and characterize possible future events, the Crystal Ball system pipeline contains multiple components, including entity extraction, event identification and a visual interface.

The pipeline first extracts location and date from tweets. If the extracted date refers a future time and the extracted location is valid, then the tweet goes to the event identification component. Even if a tweet may mention a future time and valid location, it is possible that the tweet does not contain any informative content. Thus, in order to determine the quality of tweets as indicators of future events, we employ 7 measures: Normalized Pointwise Mutual Information (NPMI) of time-location pairs, link ratio, hashtag ratio, user credibility, user diversity, degree centrality and tweet similarity.

3.3.2.2 Multiple Coordinated views in the Crystal Ball Interface and User Interactions

Figure 3.1 shows the Crystal Ball interface. The interface has four main views: a calendar view, map view, word cloud view and social network view. **The calendar view** displays a list of future events (Figure 3.1A). By default, it shows overview of future events (event overview, Figure 3.1a). The event overview shows all identified events and connections among them. Circles represent identified future events. The circles are grouped by dates. Events that have a same location are connected with a solid line and events that have same keywords are connected with a dotted line.

The event view shows detailed event information (event list) when the user selects

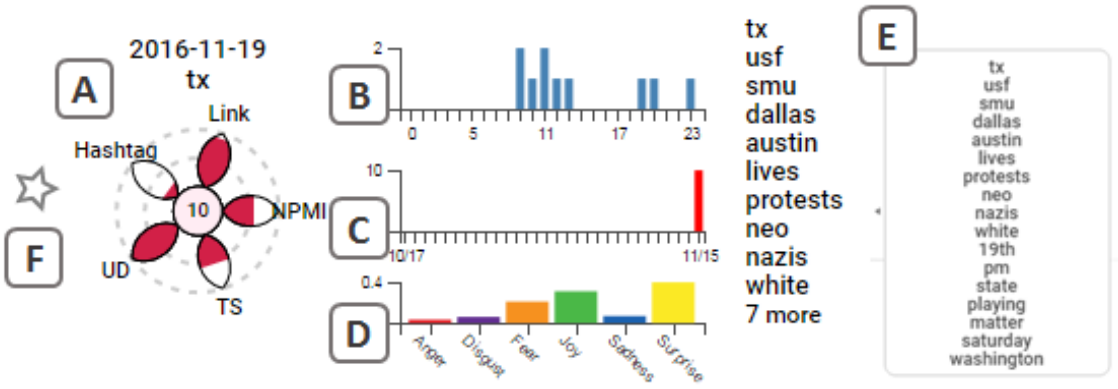


Figure 3.2: Event list. The flower glyph shows 5 measures of the future event and the number of tweets in the center (A). The three bar charts in the center show hourly distribution of tweet positing time (B), the number of tweets pointing to the event in last 30 days (C), and averages of emotion scores of the tweets (D). A list of keywords that summarize the tweets are displayed next to the bar charts (E). The user can bookmark the event by clicking the star icon (F).

a subset of the future events as shown in Figure 3.1b. Figure 3.2 shows enlarged image of Figure 3.1b. A flower glyph visualizes five of seven measures of a future events with the number of tweets in the center (Figure 3.2A). The five measures are link and hashtag ratios, NPMI, user diversity and tweet similarity. Two timeline bar charts visualize distribution of tweet positing time (Figure 3.2B) and the number of tweets in last 30 days (Figure 3.2C). The bottom bar chart shows average emotion scores of the tweets (Figure 3.2D). Keywords that summarize tweets of the event is displayed on the right side of the view (Figure 3.2E). The event can be bookmarked as favorite by clicking the star icon (Figure 3.2F). The bookmarked events are stored in database so that the user can review them anytime.

The map view shows identified events on the map to indicate where they will occur (Figure 3.1B). Events are aggregated based on the zoom level and are shown as rings. The color of ring proportions represent event dates ranging from tomorrow (dark blue) to more than a month (light blue). Clicking a ring will show its tweets as circles. Clicking a circle will show a tooltip showing the tweet.

There are two facilitating views to help users explore and further analyze the future

events: **word cloud view** and **social network view** (Figure 3.1 C and D). The word cloud view shows keywords extracted from tweets of the identified events. The size of keywords represent frequencies of the keywords. The word cloud view is updated when selected events are changed. The social network view represents relationships between future events and Twitter users. Clusters in the view represent future events in same locations. In many cases, a cluster has several future events in a same location.

User Interactions: The highly interactive and exploratory nature of the Crystal Ball interface enables users to start their exploration and analysis of future events from any of the four main views present in the interface.

A user can start with the calendar view in order to know when the event will occur. Hovering the mouse over a circle in the event overview will highlight the corresponding events on the map, word cloud and social network view. The user can find events that share a same location or keywords by examining links. The user can select a particular date then the event list will be shown in the calendar view that shows all events of the date with detailed information. Other views will be automatically updated to show corresponding events in the views.

Alternatively, the user can start the analysis from the map view to make sense of where the event will occur first. The map view shows detailed evenets when zooming into a region of interest. When the zoom factor is lower than a zoom threshold, the calendar, word cloud and social network views are updated to show the events in the current map extent. The user can open the event list to show all the events in the map extent by clicking the "show events" button on the bottom right of the map view.

The interactions implemented in Crystal Ball allows users to perform exploratory analysis to support decision-making tasks. Consequently, the decision making process is reflected by the actions participants take within Crystal Ball. In our experiment, in order to analyze the effect of anchoring bias on a decision making task conducted in Crystal Ball, we defined and logged 39 unique user interactions. Figure 3.3 lists 36

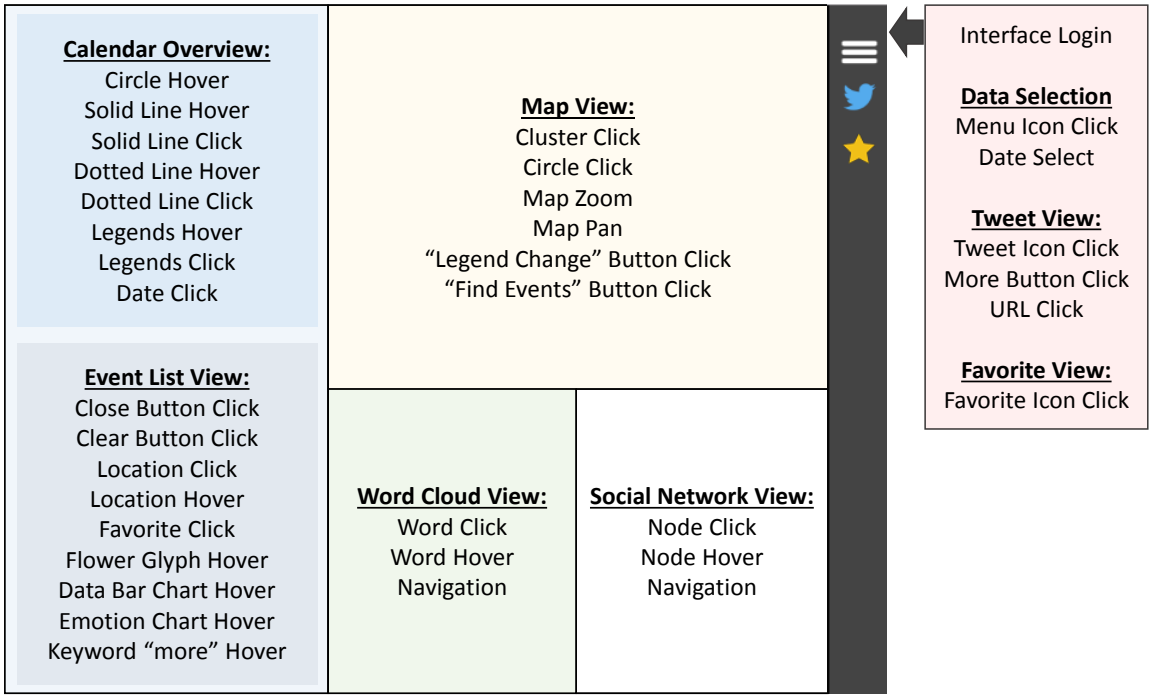


Figure 3.3: User interaction logs: the figure displays main user interaction logs of the Crystal Ball interface. Each view has different interaction logs based on its visual elements.

user interactions, situated in their corresponding views. The rest of the 3 interactions were used rarely but our participants during their interactions with Crystal Ball. The interface records all user interaction logs with a timestamp and a user's name to database which are analyzed in Section 3.5 in order to show users' decision making process.

3.3.3 Design Rationale

The anchoring effect has been replicated in numerous studies in the laboratory and in the field [115]. Our experiment design is thoroughly grounded in these best practices of controlled experimental studies in that we use priming to elicit the anchoring bias. First, we focused the participant's experience around a well-defined cognitively engaging decision-making task - we asked participants to estimate the number of protest events in a given period of time and in a given location. We conducted our experiment with the Crystal Ball interface to predict and detect protest events from

Twitter data. The *Calendar View* and the *Map View* as described in Section 3.3.2 serve as the **time** and **geo (visual)** anchor. In order to test our hypotheses, we followed a 2×2 between-subjects factorial design with two factors (numerical and visual) and each factor had two levels as described below.

3.3.4 Experimental Stimuli

The visual and numerical anchors for the experiments were devised in order to prime the participants in two ways. The numerical anchor primed participants on a number (High or Low) and the visual anchor primed participants on a specific view in the Crystal Ball interface (map view, representing geo anchor or calendar view, which represented the time anchor). The decision-making task presented to the participants is one of the four choices presented below:

Geo + high/low anchor: Do you think that the number of protest events in the state of California <geo anchor first> was higher or lower than 152 (or 8) <high (or low) numerical anchor> between November 10, 2016 and November 24, 2016 <time anchor>?

Time + high/low anchor: Do you think that the number of protest events between November 10, 2016 and November 24, 2016 <time anchor first> was higher or lower than 152 (or 8) <high (or low) numerical anchor> in the state of California <geo anchor>?

As can be noted, the magnitude of the numerical anchor, either high or low, is subject to the experimental condition. These high and low numerical anchors were chosen based on the actual number of protest events present in the data (as determined by trained annotators). Additionally, the **order** of presentation of the visual anchors varies in the two questions. The visual anchors were further **reinforced** through custom training videos orienting the participants to the use of the Crystal

Ball interface.[1] The two training videos reinforce the visual anchors by starting and driving the analysis from either the map view (geo) or the calendar view (time).

3.3.5 Experiment Design

3.3.5.1 Procedures

The data collection for this study involved in-person laboratory participation. Participants were recruited via in class recruitment, email to listservs and the psychology research pool at our university. Sessions were conducted between February 10th, 2017 and March 15th, 2017. After signing up for the study, participants were assigned a unique code for secure identification. Associated with this code, was the random assignment to one of four experiment conditions (High/Geo, Low/Geo, High/Time, Low/Time). Participants were asked to come to the lab for the duration of one hour. The experimenter would first elicit their responses to informed consent. Next, the participants would view two training videos specifically designed for this experiment. The first video was a general training video (duration 5 minutes) which oriented them to the use of the Crystal Ball interface and its basic functionality (e.g., primary and supporting interactions). This video was shown to all the participants, regardless of experimental condition. Next, the participants were shown a priming video (duration of 3 minutes) based on their visual anchoring group. The priming video was designed to guide the users through a case scenario through either Geo or Time Visual Anchors.

Following the training, the participants were asked to complete a pre-test questionnaire. The pre-test questionnaire consisted of questions related to participant's demographics (age, gender, education), their familiarity with visual analytics systems and social media, and Big-5 personality questions [146]. The informed consent, training video and pre-test questionnaire typically took around 20 minutes to complete. The participants were then assigned the task, and asked to interact with Crystal Ball for about 25 minutes. We designed and implemented interaction logging with

[1]Please refer to supplemental materials for the two training videos.

the Crystal Ball interface to capture their timestamped actions as they proceeded through the task. The interaction logging is transparent to the participants. At the end of their interaction, participants were asked to estimate the number of protest events based on their analyses within Crystal Ball.

Next participants were asked to complete a post-test questionnaire. The completion of the post-test questionnaire ended their participation in the study. The post-test contained questions regarding the usability of the system (ease, attention, stimulation, likability), level of engagement during the task and questions to gauge their susceptibility to bias. The bias questions consisted of eight questions designed to measure the level of bias. Participants were compensated by either a $5 gift card or class credit assigned at the discretion of the class instructor willing to assign extra credit.

3.3.5.2 Participants

A total of 85 participants completed the study. We discarded the data for four participants due to usage of incorrect identification codes during the experiment. Distribution of participants across experiment conditions was relatively even and is represented in Table 1. Figure 4 shows a summary of the demographic characteristics of participants across factors including age, gender, education and major. We note that there is an even balance of participants across various demographic characteristics such as gender (male vs. female), age (different age ranges) and education background (computing vs. other majors), although there is some skewness in the data towards students pursuing Masters degrees. Participant demographic characteristics were also balanced across the four experiment conditions due to random assignment of participant to experiment condition. Males and females participants were uniformly distributed across experiment conditions (25% in each condition, $SD = 10\%$ for males, $SD = 6\%$ for females). Average ratio for males to females was 1.02 in each experiment condition. Average proportion of undergraduate, masters and PhD

Table 3.1: Distribution of participants in 4 conditions. Row-Numerical anchor; column-Visual anchor.

	Geo	Time	Grand Total
High	20	21	41
Low	22	18	40
Total	42	39	81

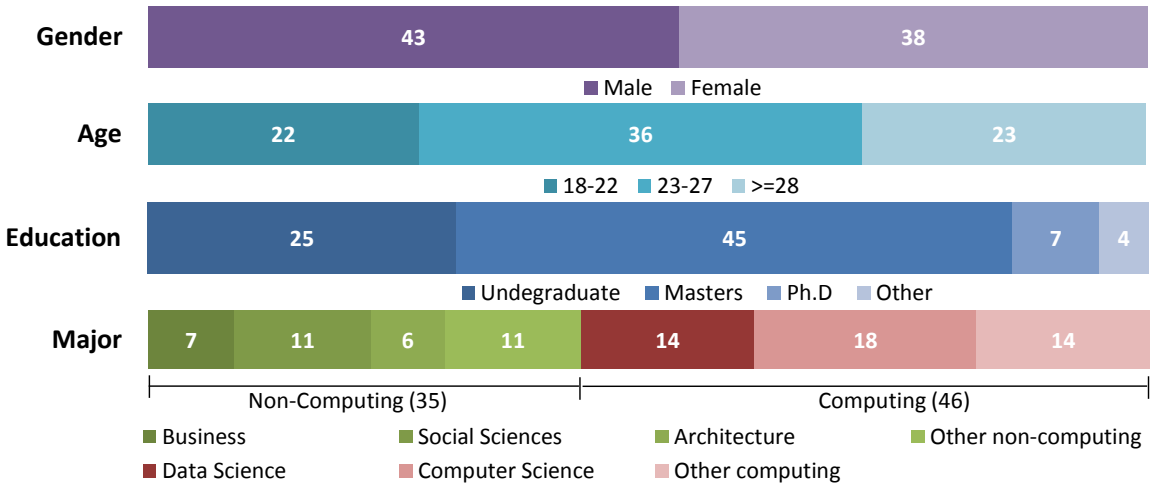


Figure 3.4: Demographic. A summary of demographic information of the participants based on gender, age, education and major.

education level in each experiment condition was also 25% (SD = 13%, 10% and 24% resp.).

3.4 Experiment Results: Analyzing Quantitative Measures

Two types of quantitative analyses are conducted to answer research questions RQ1 and RQ2 introduced in Section 3.3.1.

3.4.1 RQ1 - Visual Anchor: Can individuals be anchored on a specific view?

To quantitatively evaluate whether participants can be anchored on a view in Crystal Ball, we conducted two types of stastical analysis - two-way ANOVAs and Bonferroni-corrected pairwise t-tests. We extracted the overall time duration a participant spent in geo or time-oriented views from the interaction logs by taking into account the time stamp of each action occurred in a particular view.

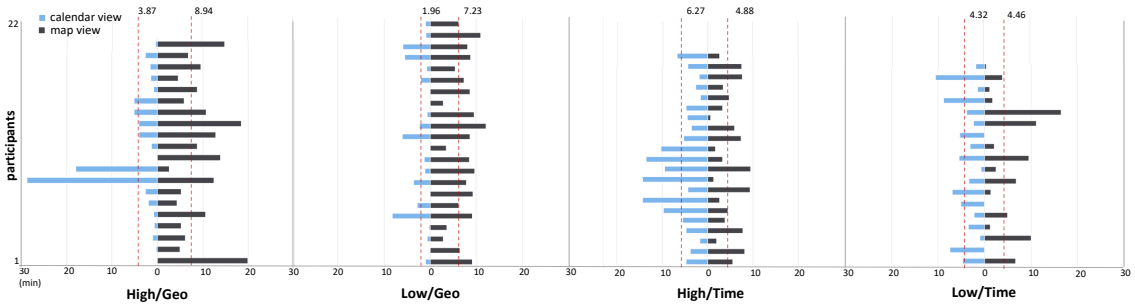


Figure 3.5: This figure provides a summary of the amount of time spent in calendar and map views on each of the four different conditions. The red dashed line is the mean of the amount of time spent in calendar and map view.

A two-way ANOVA was conducted on the influence of two independent variables (numerical and visual anchor) on the amount of time spent in different views (map vs. calendar). The main effect for visual anchor was statistically significant and had an F ratio of $F(1, 78) = 11.57, p < .001$. The main effect for numerical anchor indicated that the effect for numerical anchoring did not significantly affect the time spent in map vs. calendar view ($p > 0.05$). The interaction effect was not significant, $F(1, 77) = 0.12, p > 0.05$.

Bonferroni corrected pairwise t-tests (α=0.05/4) were conducted to compare the duration of time spent in the different conditions with α=0.0125 level of significance. We found that visual anchor had significant effect on the time spent in map view vs. calendar view across both conditions ($p < 0.01$ in both cases) whereas the numerical anchor did not ($p > 0.05$ in both cases).

In Figure 5, we show the duration of time spent in each view for each participant across all four experimental conditions. The x-axis represents the time in minutes, with the blue bars representing duration in calender view and the black bar representing duration in map view; the y-axis are the participant ids in each condition. We see from charts labeled High/Geo and Low/Geo that participants spent significantly more time in the map view vs. the calendar view in the geo anchoring conditions. The charts labeled High/Time and Low/Time reveal that in the Time priming conditions,

the time spent in each view was variable and no statistical trends can be observed. We have included four separate charts in Figure 5 to provide sufficient comparative detail across the experiment conditions.

3.4.2 RQ2 - Numerical Anchor: Are the effects of numerical priming transferable to VA?

The effect of numerical anchor on time spent within Crystal Ball. As reported in Section 4.1, two-way ANOVAs conducted in order to determine the effects of numerical anchoring indicated the main effect for numerical anchor did not significantly affect the time spent in map vs. calendar view ($p > 0.05$). The interaction effect was also not significant, $F(1, 77) = 0.12, p > 0.05$.

These findings indicate that being primed by a numerical anchor did not have an effect on the amount of time spent in map view compared to the calendar view. We discuss the implications of these findings further in the discussion section, and suggest that more investigation is needed to determine the cause of these effects.

The effect of numerical anchor on the decision-making outcome. To further asses the impact of numerical anchoring, we analyzed the responses given by participants in the pre- and post-tests. The participants were asked to estimate the number of protest events, before and after their interactions with the data and the interface. The findings are shown in Figure 3.6. On the x-axis we show the two groups in the numerical anchoring condition High and Low. On the y-axis are each participant's estimates regarding the number of protest events, before the interaction (in orange) and after the interaction (in red). Mean post-test responses are in black. Our findings indicate that participants were consistently anchored on initial number presented to them in the framing of the questions ($p < 0.05$). Our findings suggest that the effects of the classic anchoring bias elicited by priming with numerical anchors in previous laboratory studies can be replicated in VA. We did not find any effects of the visual anchoring (geo vs. time) on the final outcome ($t(40) = 2.02, p > 0.05$), suggesting

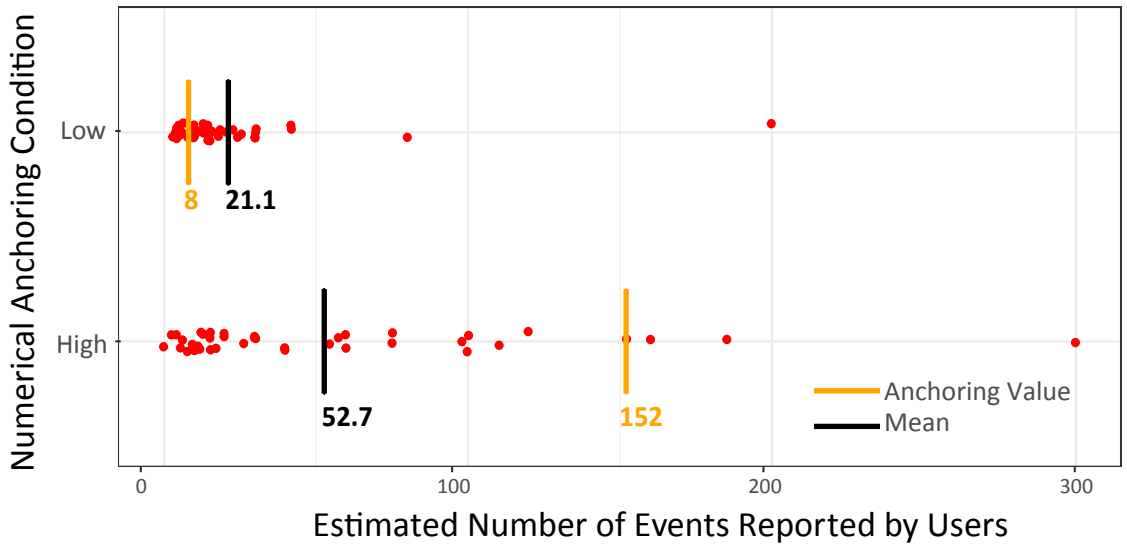


Figure 3.6: The estimated number of geo-political events reported by each participant is represented by red dots. The orange line represents the anchor value and black line is the mean of estimated number of political events.

that the effects of the visual anchor may be more subtle than can be determined via post-test questionnaire responses. We conducted detailed analyses to capture these effects, which we shall describe in the next section.

3.5 Experiment Results: Analyzing Interaction Logs for User Activity Patterns

To test the hypothesis that user interaction logs reflect the participants' decision making processes, we applied two additional types of analyses on the logs in order to evaluate the impact of anchoring effect on the patterns of user interactions. The two analyses address research questions RQ3 and RQ4 (Section 3.3.1).

3.5.1 RQ3: How does anchoring effect influence the paths of interactions?

To analyze the paths users take during their analysis with Crystal Ball and the effect of anchors on these paths, we developed a novel method to study the sequences of users' interactions as a network of interaction nodes. We constructed the interaction network as follows: each interaction is logged with five attributes: time stamp of the interaction, the view it took place in, the type of interaction, and detailed description

for each interaction (e.g., 12:56:35.56, Calendar view, Click, Zoom 89.55 36.00). As shown in Fig 3.3 there are 36 main interactions as well as 3 main secondary interactions over multiple views. Each of these interactions form the nodes in a network. The edges in the network are chronological pairs of interactions. For example, if a user has zoomed on the map and then hovered on a particular location in the calendar view, this would add an edge between the *Map Zoom* and *Calendar Location Hover* nodes. The edge weight would incrementally increase for each additional observed pair. For visualization purposes, we disregarded self-loops (i.e, repeated actions) because we are more interested in the relationship between different interactions and the paths of interactions taken by our users. This method yields a weighted directed graph which enables us to cluster interactions through community detection, rank each action by multiple centrality measures and compare aggregate user path differences controlling for each anchor. The network visualizations in this section were created with Gephi [147].

In this section, first we take actions of all users into account to get a complete picture of users' paths of interactions. We then analyze differences in users' paths to detect different user strategies controlling for the two anchors (visual and numerical). We studied the network of all user logs, as well as our four experiment anchors. We did not find significant differences in the networks created from logs of users primed on the two numerical anchors. Hence in this discussion, we will focus on the three remaining networks: a full network (AllNetwork), a Geo-anchored network (GeoNetwork) and a Time-anchored network (TimeNetwork).

3.5.1.1 Analyzing the network of all interactions

By adopting an exploratory data analysis method, we started by analyzing different features of AllNetwork (39 nodes and 640 edges) (Figure 3.7). We first utilized the community detection algorithm developed by Blondel *et al.* [148], which resulted in 5 different communities of interactions. Most of these communities are comprised of

interactions that occur within the same view or have a close semantic relationship to each other. The community detection results allowed us to categorize nodes in our network into three main groups that were in line with our initial system design strategies: preliminary interactions, primary interactions, and supporting interactions. The primary interactions include those that users have to go through in order to find the events of interest. The supporting interactions are those that users perform in order to find supporting information to confirm the previously found events. The preliminary interactions such as login and clicking on the menu bar were used infrequently as they are not critical to the analysis process. Figure 3.7 shows the network colored and annotated based on the community detection results.

In order to measure the importance of different interactions, we utilized the Pagerank algorithm [149]. Since Pagerank takes into account the weight of edges between interactions, it is much more powerful than simply calculating the frequency of each interaction. Pagerank assigns probability distributions to each node denoting the importance of the node. These probability distributions are appropriate metrics for importance of the interactions in our system as they show the likelihood of a random surfer in the network to traverse to a specific end node. The top ranked interactions in our interface are all from the primary action communities with the exception of *Word Hover*. As seen in Figure 3.7, edge weights between important nodes in the same community are higher than ones between different communities. Furthermore, we can observe mutual higher weighted paths between these higher ranking interactions. Some exceptions to this observation is when the users are moving away from a view to another conduct more in depth analysis. For example, in AllNetwork, the edge between *Events Location Hover* to *Events Location Click* is weighted very strong, but the path of opposite direction is not. We can interpret *Events Location Click* as an interaction that drives users out of this community to others such as Word Cloud and Social Network to get complementary information of an event (See Fig 3.7). The

AllNetwork and the analyses resulting from it serve as a reference for the comparisons we wish to make across the GeoNetwork and TimeNetwork.

3.5.1.2 Comparing interaction networks of Geo- vs. Time-anchored users

To answer our research questions of whether visual anchor has an effect on the way different groups of participants interact with the visual interface, we constructed two networks based on the actions of the geo and time-anchored groups. These networks consist of the same 39 action nodes but have different edges and weights allowing us to compare the interactions of participants primed on the two visual anchors through the lens of their respective networks.

Similar to our analysis of AllNetwork, we first started by detecting communities within GeoNetwork and TimeNetwork. Interestingly, the results show similar community structures to AllNetwork. However, there are subtle differences that point to the differences regarding the usage of Crystal Ball between the two groups. For example, the action of *Favorite Icon Click* (through which users can save an event to view later in the Favorites Menu) in GeoNetwork is part of the preliminary actions community, but for TimeNetwork it is part of the Time related primary actions community. This subtle change could indicate that the time primed users had more interactions between saving an event as a favorite and then viewing the list of favorite actions in comparison to our Geo primed users.

We calculated Pagerank for interaction nodes in both these networks. Comparing these values would allow us to understand important interactions within each network and how they are affected by the visual anchor. Figure 3.8 illustrates significant differences of the two networks. In the GeoNetwork, the top nodes are a mixture of interactions from the Map and Event views, with the highest ranked interaction from the Map view. This pattern is consistent with the strategies shown in the Geo priming video. In contrast, in the TimeNetwork, the top ranked nodes are interactions within the Events view and the Calendar View, which is also consistent

with the strategy shown in the time-anchor video. Other important but lower ranking actions in TimeNetwork are from the Map View. Furthermore, by observing the paths between the Events community (colored purple in Figure 3.8) in both GeoNetwork and TimeNetwork, we see that weight of the edge between *Events Location Hover* and *Events Location Click* is relatively higher in the GeoNetwork in comparison to the TimeNetwork. This could indicate that our Geo primed users use maps to explore and primarily use hovering and clicking on a location together to view more details in the map, word cloud, and social network views. Our time primed users on the other hand, utilize the hovering on locations to explore events. These differences show interesting behavioral variations in sequences of interactions between our two groups. These differences show that time primed users are more likely to use the Calendar and Events view actions as their primary exploratory tool. Figure 3.8 shows the comparisons between these two networks and two bar charts comparing the top 5 ranked nodes in each network.

Analyzing the interactions of our participants as a network has many benefits. It allows us to take into account the sequence of interactions, as well as the paths taken by users to arrive at the conclusion. The paths taken reflect the strategies users employ during the decision-making process. Furthermore, we can take an overview of all interactions within the Crystal Ball interface and analyze what strategies need to be improved to make the interface more effective.

3.5.2 RQ4: Estimating the effects of anchoring bias on interaction patterns and information seeking activities

One drawback of the network analysis is that the estimated impact for each anchor is measured without standard errors to calculate the statistical significance of each result. To address this problem, we use structural topic modeling (STM) to measure the impact of the anchors and user-level attributes on users' actions [150]. Originally built for text summarization, STM is a generalized topic model framework

for testing the impact of document-level variables. For our model, topics are clusters of interactions measured as probability distributions over the action space. We test our hypotheses of the effect of anchoring on the topic proportions through an embedded regression-component. STM is a consolidation of three predecessor models: the correlated topic model (CTM) [151], the sparse additive generative model (SAGE) [152] and the Dirichlet multinomial regression model (DMR) [153]. The CTM model introduces a correlation structure for topic distributions while the DMR and SAGE models provide mechanisms to estimate the effect of independent variables on either topic proportions (via DMR model) or word distributions for each topic (via SAGE model). [2]

Table 3.2: Independent Variables Tested

Type	Independent Variable	Level
Condition	Visual Anchor	Time / Geo
	Numerical Anchor	High /Low
Time	Percent of Actions	Action Deciles (b-spline)
Attribute	Gender	Make / Female
	Major	Computing / Non Computing
	Age	Under 23 / Over 23
	Education	Undergraduate / Graduate
Personality	Extroversion	High / Low
	Agreeableness	High / Low
	Conscientiousness	High / Low
	Openness	High / Low
	Neuroticism	High / Low

However, as with most topic models, STM is built from the bag-of-words (BoW) assumption that provides a key advantage and disadvantage in our analysis. The

[2]We used the stm R package [154] for our analysis. This package includes additional tools for topic modeling including a spectral initialization process that aids in addressing the multi-modality problem (stability of the results).

Table 3.3: This table provides the seven actions with the highest probabilities for three sample topics: Map View, Calendar View and Event List (all tools). Action combinations (bi- or tri-grams) are denoted by the plus sign.

Rank	Map View (Topic 8)	Calendar Overview (Topic 6)	Event List: All Tools (Topic 3)
1	Map Zoom	Calendar Hover Circle	Event Keyword More Hover
2	Map Zoom + Map Zoom	Calendar Solid Line Hover	Event Keyword More Hover + Event Keyword More Hover
3	Map Pan	Calendar Dotted Line Hover	Event Keyword More Hover + Event Keyword More Hover + Event Keyword More Hover
4	Map Circle Click	Calendar Hover Circle + Calendar Hover Circle	Event Flower Glyph Hover
5	Map Zoom + Map Zoom + Map Zoom	Calendar Solid Line Hover + Calendar Hover Circle	Event Emotion Chart Hover
6	Map Cluster Click	Calendar Dotted Line Hover + Calendar Hover Circle	Event Favorite Click
7	Map Zoom + Map Pan	Calendar Solid Line Hover + Calendar Solid Line Hover	Event Flower Glyph Hover + Event Flower Glyph Hover

advantage is that it yields statistical properties (exchangability) that identifies topics as clusters of co-occurring interactions and facilitates statistical testing through the DMR (GLM regression) component. On the other hand, a disadvantage of the BoW assumption is that it ignores the order of interactions. To address this issue, we made **two modifications**: extracting bi-/tri-grams and creating a session time variable by interaction deciles. First, we extracted every bi- and tri-gram as chronological action pairs and triplets from the interaction logs. Including bi- and tri-grams and the single actions, we had 237 unique features after removing sparse features. Second, we created a time variable that divided each user's session into ten evenly distributed groups (interaction decile). Given that each user's session averaged nearly 800 individual actions, each decile maintained sufficient interactions to facilitate topic inference. Additionally, inclusion of the time variable had the advantage of increasing our sample

size (number of documents) from 81 to 810 as the document-level went from each user to a user's interaction decile (e.g. first 10% of user X's interactions).

To test the effect of anchoring bias on users' interactions, our baseline model to explain topic proportions (dependent variable) incorporates three independent variables: the visual anchor, the numerical anchor, and time as interaction deciles. After analyzing the model, we tested other demographic attributes including gender, major, age, education level, and the Big-5 personalities. Table 2 above provides a list of the independent variables tested and the categorical levels. We binned the user attributes into binary levels. Similarly, we converted the Big-5 personality results into binary levels in which users who scored above the mean were categorized as High while users who scored below the mean were categorized as Low.

3.5.2.1 The effect of visual anchor on interaction patterns estimated by topic proportions

We find the visual anchor has a significant effect on the proportion of users' interactions as topics clustered automatically in view-based groups (e.g., map, calendar, events). Figure 3.3 provides the top seven interactions for three sample topics. We observe that the interactions tend to cluster into groups related to each interaction's associated view hierarchy as shown in Figure 3.3. For example, topic 8 includes interactions related to the map view including *Map Zoom*, *Map Pan*, *Map Circle Click* and *Map Click Cluster*. Therefore, we gave topic 8 the manual label of Map View since its interactions are all related to that view. Following this approach, we created manual labels for the other seven topics. Further, we find that the topics tend to cluster in groups consistent with our network communities found in Section 5.1. For instance, the four prominent interactions of the Map View topic (by probability) have the strongest connections as well as highest PageRank in the Map View community cluster (green nodes) in Figure 3.7.

Second, we find in Figure 3.9 that the Map View and Flower Glyph topics had

the largest topic proportions. Alternatively, the social network and word cloud were the smallest topics. To test our number of topics, we followed the procedure recommended by [155] by considering multiple topic scenarios (5, 8, 10, 15, 20, 25, 30, 40) and comparing each model's held-out likelihood and average semantic coherence. We decided on an eight topic model given a high average semantic coherence and parsimony of topics (see supplemental materials).

We observe that the visual anchor had a significant effect on the Map View and Calendar View topics. Figure 3.9 provides the effect the anchors had on the topic proportions. In this plot, each dot is the estimated topic proportion difference for each topic by the two levels of each anchor. The line represents a 95% confidence interval around each estimate. From these figures, we find that the Map View and the Calendar View topic proportions have the most significant differences between the two groups. Consistent with our findings in sections 4.1 and 5.1, Geo primed users are anchored more to the view they were primed on while we see less of an effect in Time primed users. On the other hand, we found that the visual anchor had an unexpected effect with the Event List: All Tools (topic 3). Geo primed users tended to use tools like the *Keyword More* and *Emotion Bar* more than Time primed users. Alternatively, we find that the numerical anchor had only a marginal effect on two topics (Calendar View (topic 6) and Event List: All Tools (topic 3)). These results imply that the visual anchor had a more significant impact on the proportion of users' interactions than the numerical anchor. This is important as we observed opposite effect (numerical anchor was significant, visual anchor was not) in the users' estimation of the event outcome.

3.5.2.2 The effect of visual anchor on interactions used over time estimated by topic proportions

We find evidence of a temporal effect on the topic proportions. To measure this effect, we divided each user's interaction path into interaction deciles (see Section 5.1).

To aid estimation, we used a b-spline to smooth the values. Figure 3.10 provides the effect of the visual anchor (line color) and time (x-axis) for the Map View and Calendar View topic proportions. We observe a significant impact of the time of the user's session on this topic proportions. For example, Map View topic proportion is nearly twice during the user's first twenty percent of interactions than users' remaining 80 percent of interactions. Moreover, we see this distinct drop for both visual anchor groups. This observation implies that users tended to use the Map View more in the beginning of the session as they were getting acclimated to the interface. Alternatively, the Calendar View topic trended down resulting in much lower use by session end (15% Time, single digits Geo). We found marginal effects of time for the other six topics, with most nearly flat given already low topic proportions (less than 10%).

3.5.2.3 The effect of demographic variables on interaction patterns estimated by topic proportions

To test other possible variables, we ran five additional model scenarios replacing the numerical anchor variable (as it showed only marginal significance) with the demographic variables (gender, major, student level and Big-5 personality). We found that none of the variables produced significant (95%) changes in topic proportions, although some produced marginally significant effects (see supplemental materials). For example, most variation occurs in the secondary view topics (Word Cloud, Social Network and an interaction topics).

3.6 Discussion and Limitations

In this section, we provide implications of our experiment results on anchoring effect in visual analytics, and point out possible limitations related to the study design and analysis.

Experiment implications: As shown by our data analysis (section 4 & 5), our experimental results indicate that anchoring bias does transfer to visual analytics.

Most interesting finding is that the visual anchor seems to significantly impact the decision-making **process**, while the numerical anchor has a significant effect on the decision-making **outcome**. The decision-making process reflects the way that users interact with Crystal Ball; the outcome is the final answer that the participants provided at the end of the decision-making process.

Such findings have implications for user training on visual analytics systems with CMV, as well as how decision-making tasks are framed. With respect to training/tutorial, the visual analytic systems development team should provide multiple scenarios employing strategies that involve the use of different views as the primary visualization to drive the analysis. As of decision-making task framing, one should avoid accidentally anchoring the participants on an expected outcome or when possible, employ measures of cognitive bias (such as in our post-test) to evaluate the inherent cognitive bias of the users. As noted in Section §3.2.1, the tendency of humans to rely on heuristics to make judgments does often lead to efficient and accurate decisions. However, we need to determine when such heuristic decision-making is being applied, in order to ensure that the resulting decisions are optimal.

Experiment sample size limitation: As can be expected with any laboratory experiment, this research has limitations. One such limitation is the sample size of 81 participants in our experiment. However, the diversity of our sample with respect to gender, age, educational background and personality factors are steps we have taken to ensure the validity of our results. Our findings replicate the effects of anchoring that have been long studied in literature, further attesting to the validity of the experiment.

Experiment control limitation: Another limitation of our experiment is that we do not consider a control group , that is, participants who engage in the decision-making task without being primed by any anchors. While our initial study reported here was focused on determining whether the effects of anchoring are at all present

and can be elicited in such experiments, our future work will be aimed at replicating these findings in more extensive experiments with larger sample size and will include control groups for comparison.

STM analysis limitations: Fong and Grimmer [156] note that topic models are susceptible to problems in estimating marginal effects due to the zero-sum properties of topic proportions. Further, topic models cluster only based on the count and ignore interaction duration (time spent). To address such limitation, the quantitative analysis in section 4 explicitly accounted for the duration of each interaction.

3.7 Conclusion

In this chapter, we presented a systematic study and resulting analyses that investigate the effect of anchoring bias on decision-making processes and outcome using visual analytic systems. Our experimental results provide evidence on anchoring effect being transferable to visual analytics in that visual and numerical anchors affect the decision-making process and outcome respectively. The present study is a first step in an overarching research agenda of determining the use of heuristics in decision-making processes from the user interactions and if these decision-making processes can be reliably inferred then to automatically suggest ways in which to improve the process.

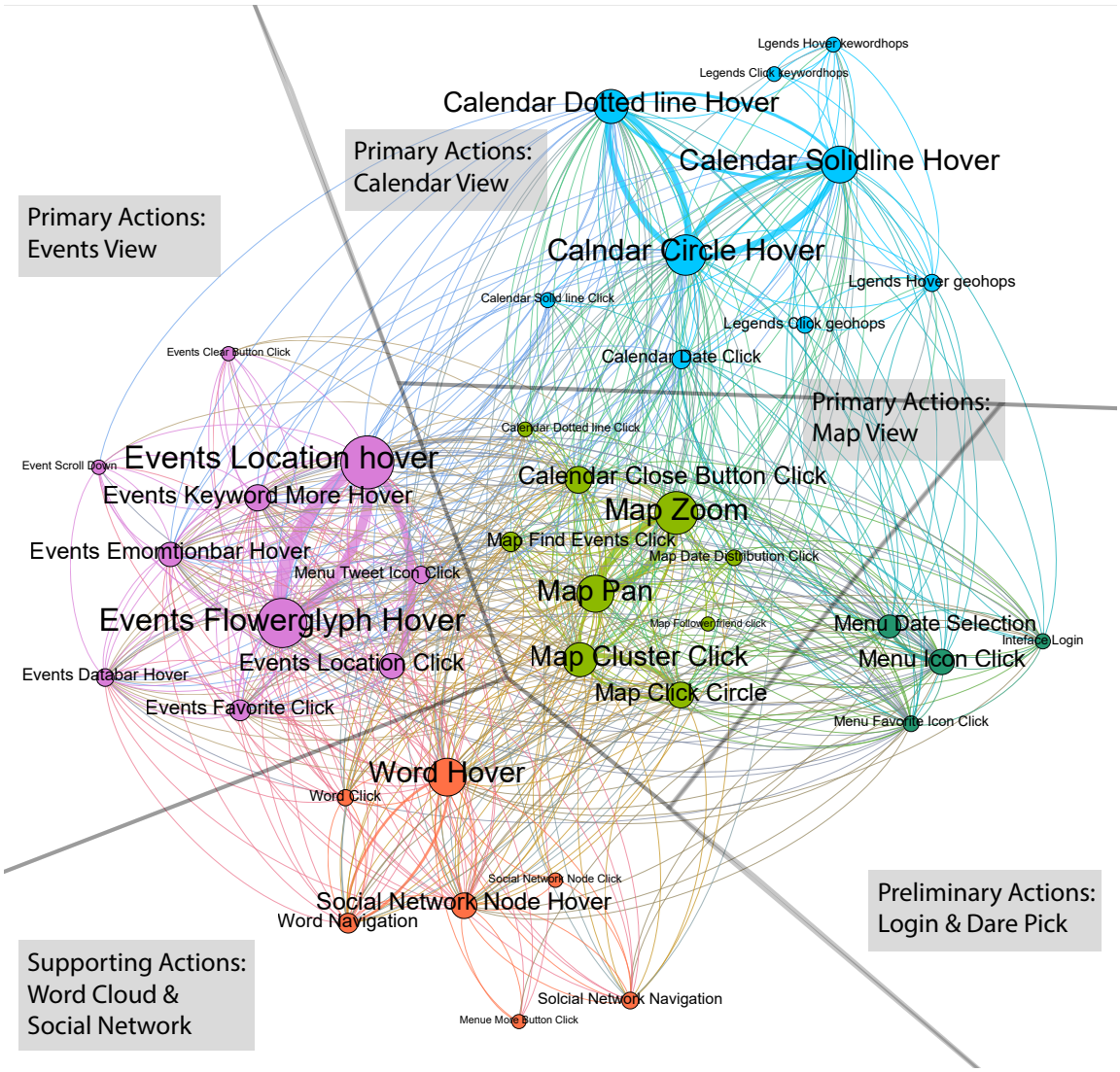


Figure 3.7: The directed network of all interactions. Nodes are interactions and edges are interactions that occur after each other. The size of nodes are proportional to Pagerank values and width of edges are proportional to the edge weights. Note, if a line is drawn between a start-node and an end-node, the outgoing edge from the start node is on the relative left side of that line.

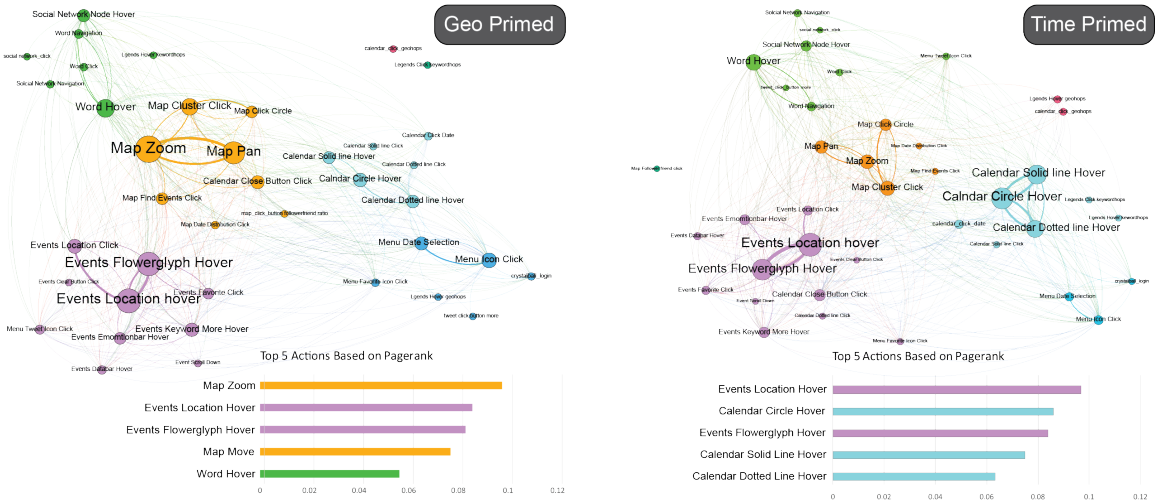


Figure 3.8: Side by side visualization of GeoNetwork and TimeNetwork. The size of nodes is proportional to Pagerank values of nodes in each graph, the color of nodes corresponds to the detected community of each node, and the width of each edges corresponds to the weight of that edges. The bar charts show the top 5 nodes based on their Pagerank value and is color coded based the community the nodes community.

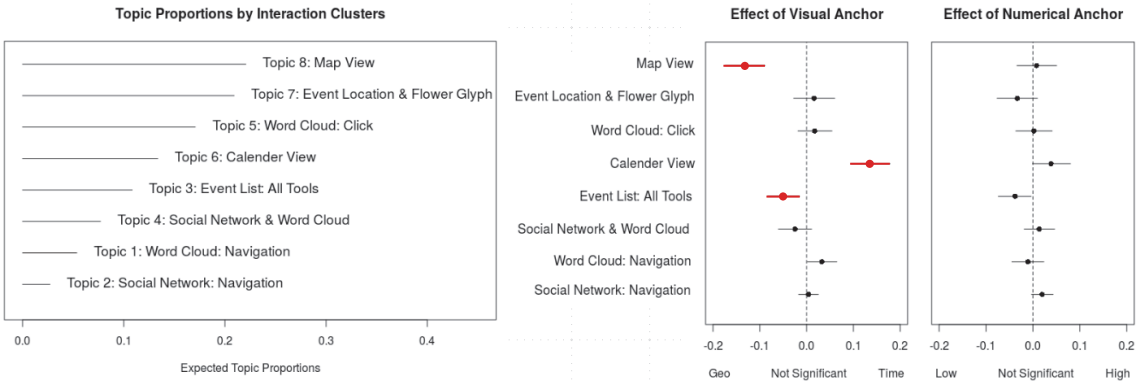


Figure 3.9: The figure on the left provides the expected topic (action-cluster) proportions with judgmental labels to aid in interpretation. The figures on the right provide the estimated effect of the visual and numerical anchors on each of the eight topics' proportions. The dot is the point estimate and the line represents a 95 percent confidence interval. The red dots/lines are topics that are significant with 95% confidence.

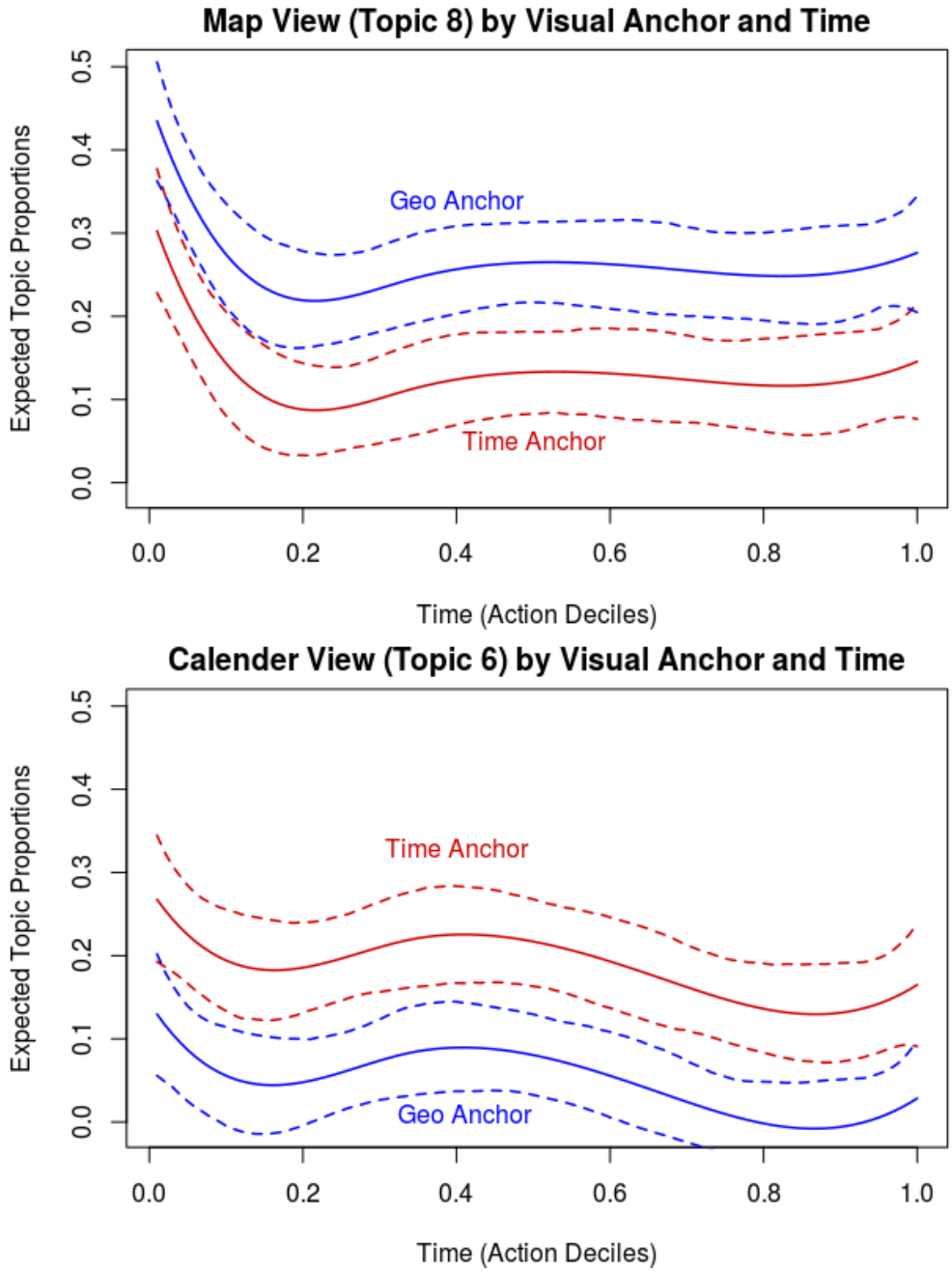


Figure 3.10: This figure provides two charts on the effect between the visual anchors (line color) and time as measured by interaction deciles (x-axis) for two topics (Map View and Calendar View). Each line is the estimated topic proportions across the session and controlling for the visual anchor. The solid line is the point estimate and the dotted line is a 95 percent confidence interval. For the interaction deciles (time), we divided users' sessions into ten evenly distributed groups. A b-spline was used to smooth the curve across the ten points.

CHAPTER 4: STUDYING UNCERTAINTY AND DECISION-MAKING ABOUT MISINFORMATION IN VISUAL ANALYTICS.

4.1 Introduction

The spread of misinformation on social media is a phenomena with global consequences, one that, according to the World Economic Forum, poses significant risks to democratic societies [157]. The online media ecosystem is now a place where false or misleading content resides on an equal footing with verified and trustworthy information [158]. In response, social media platforms are becoming "content referees," faced with the difficult task of identifying misinformation internally or even seeking users' evaluations on news credibility.[1] On the one hand, the news we consume is either wittingly or unwittingly self-curated, even self-reinforced [159]. On the other hand, due to the explosive abundance of media sources and the resulting information overload, we often need to rely on heuristics and social cues to make decisions about the credibility of information [160, 161]. One such decision-making heuristic is confirmation bias, which has been implicated in the selective exposure to and spread of misinformation [162]. This cognitive bias can manifest itself on social media as individuals tend to select claims and consume news that reflect their preconceived beliefs about the world, while ignoring dissenting information [160].

While propaganda and misinformation campaigns are not a new phenomenon [163], the ubiquity and virality of the internet has lent urgency to the need for understanding how individuals make decisions about the news they consume and how technology can aid in combating this problem [164]. Visual analytic systems that present coordinated

[1] https://www.wsj.com/articles/facebook-to-rank-news-sources-by-quality-to-battle-misinformation-1516394184

multiple views and rich heterogeneous data have been demonstrably useful in supporting human decision-making in a variety of tasks such as textual event detection, geographic decision support, malware analysis, and financial analytics [165, 166]. **Our goal is to understand how visual analytics systems can be used to support decision-making around misinformation and how uncertainty and confirmation bias affect decision-making within a visual analytics environment**.

In this work, we seek to answer the following overarching research questions: *What are the important factors that contribute to the investigation of misinformation? How to facilitate decision-making around misinformation by presenting the factors in a visual analytics system? What is the role of confirmation bias and uncertainty in such decision-making processes?*

To this aim, we first leveraged prior work on categorizing misinformation on social media (specifically Twitter) [167] and identified the dimensions that can distinguish misinformation from legitimate news. We then developed a visual analytic system, Verifi, to incorporate these dimensions into interactive visual representations. Next, we conducted a controlled experiment in which participants were asked to investigate news media accounts using Verifi. Through quantitative and qualitative analysis of the experiment results, we studied the factors in decision-making around misinformation. More specifically, we investigated how **uncertainty, conflicting signals manifested in the presented data dimensions**, affect users' ability to identify misinformation in different experiment conditions. Our work is thus uniquely situated at the intersection of the psychology of decision-making, cognitive biases, and the impact of socio-technical systems, namely visual analytic systems, that aid in such decision-making.

Our work makes the following important contributions:

- *A new visual analytic system*: We designed and developed Verifi[2], a new visual

[2]http://verifi.herokuapp.com; open source data and code provided at

analytic system that incorporates dimensions critical to characterizing and distinguishing misinformation from legitimate news. Verifi enables individuals to make informed decisions about the veracity of news accounts.

- *Experiment design to study decision-making on misinformation*: We conducted an experiment using Verifi to study how people assess the veracity of the news media accounts on Twitter and what role confirmation bias plays in this process. To our knowledge, our work is the first experimental study on the determinants of decision-making in the presence of misinformation in visual analytics.

As part of our controlled experiment, we provided cues to the participants so that they would interact with data for the various news accounts along various dimensions (e.g.,tweet content, social network). Our results revealed that conflicting information along such cues (e.g., connectivity in social network) significantly impacts the users' performance in identifying misinformation.

4.2 Related Work

We discuss two distinct lines of past work that are relevant to our research. First, we explore cognitive biases, and specifically the study of confirmation bias in the context of visual analytics. Second, we introduce prior work on characterizing and visualizing misinformation in online content.

4.2.0.1 Confirmation bias:

Humans exhibit a tendency to treat evidence in a biased manner during their decision-making process in order to protect their beliefs or pre-conceived hypothesis [71], even in situations where they have no personal interest or material stake [19]. Research has shown that this tendency, known as confirmation bias, can cause inferential error with regards to human reasoning [20]. Confirmation bias is the tendency to privilege information that confirms one's hypotheses over information that

https://github.com/wesslen/verifi-icwsm-2018

disconfirms the hypotheses. Classic laboratory experiments to study confirmation bias typically present participants with a hypothesis and evidence that either confirms or disconfirms their hypothesis, and may include cues that cause uncertainty in interpretation of that given evidence. Our research is firmly grounded in these experimental studies of confirmation biases. We adapt classic psychology experimental design, where pieces of evidence or *cues* are provided to subjects used to confirm or disconfirm a given hypothesis [28, 19].

4.2.0.2 Visualization and Cognitive Biases:

Given the pervasive effects of confirmation bias and cognitive biases in general on human decision-making, scholars studying visual analytic systems have initiated research on this important problem.

[168] categorized four perspectives to build a framework of all cognitive biases in visual analytics. [169] presented a user study and identified an approach to measure anchoring bias in visual analytics by priming users to visual and numerical anchors. They demonstrated that cognitive biases, specifically anchoring bias, affect decision-making in visual analytic systems, consistent with prior research in psychology. However, no research to date has examined the effects of confirmation bias and uncertainty in the context of distinguishing information from misinformation using visual analytic systems - we seek to fill this important gap. Next, we discuss what we mean by misinformation in the context of our work.

4.2.0.3 Characterizing Misinformation:

Misinformation can be described as information that has the camouflage of traditional news media but lacks the associated rigorous editorial processes [160]. Prior research in journalism and communication has demonstrated that news outlets may slant their news coverage based on different topics [170]. In addition, [171] show that the frequency of sharing and distribution of fake news can heavily favor different

individuals. In our work, we use the term fake news to encompass misinformation including ideologically slanted news, disinformation, propaganda, hoaxes, rumors, conspiracy theories, clickbait and fabricated content, and even satire. We chose to use "fake news" as an easily accessible term that can be presented to the users as a label for misinformation and we use the term "real news" as its antithesis to characterize legitimate information.

Several systems have been introduced to (semi-) automatically detect misinformation, disinformation, or propaganda in Twitter, including FactWatcher [172], TwitterTrails [173], RumorLens [174], and Hoaxy [175]. These systems allow users to explore and monitor detected misinformation via interactive dashboards. They focus on identifying misinformation and the dashboards are designed to present analysis results from the proposed models. Instead, Verifi aims to provide an overview of dimensions that distinguish real vs. fake news accounts for a general audience.

Our work is thus situated at the intersection of these research areas and focuses on studying users' decision making about misinformation in the context of visual analytics.

4.3 Verifi: A Visual Analytic System for Investigating Misinformation

Verifi is a visual analytic system that presents multiple dimensions related to misinformation on Twitter. Our design process is informed by both prior research in distinguishing real and fake news as well as our analysis based on the data selected for our study to identify meaningful features.

A major inspiration for Verifi's design is based on the findings of Volkova *et al.* [167], who created a predictive model to distinguish between four types of fake news accounts. They find that attributes such as *social network interactions* (e.g., mention or retweet network), *linguistic features*, and *temporal trends* are the most informative factors for predicting the veracity of Twitter news accounts. Our design of Verifi is inspired by these findings: (i) we included a *social network view* that shows a

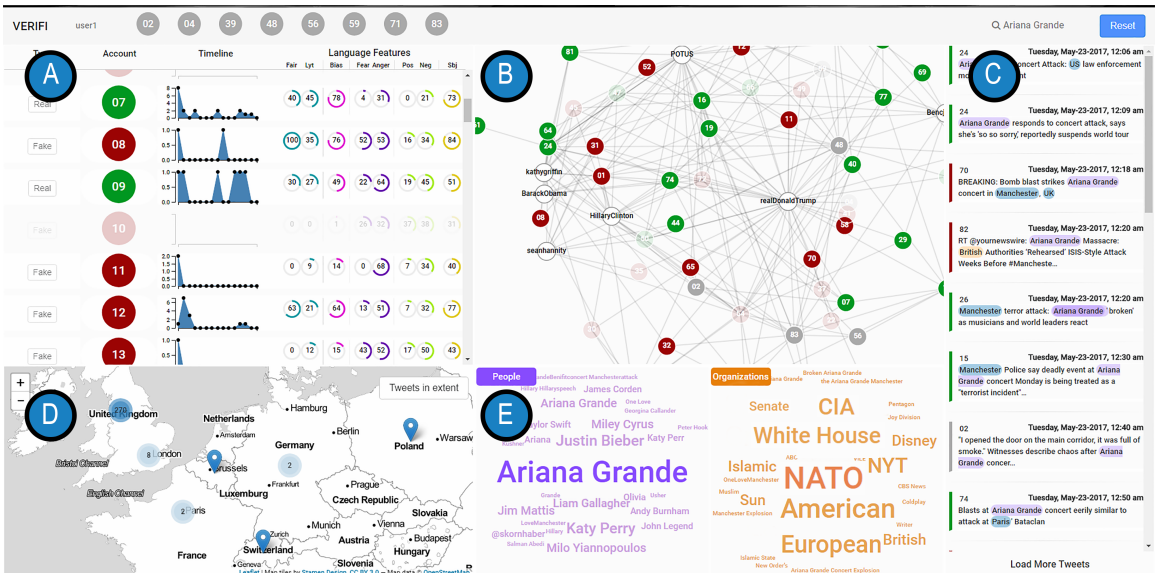


Figure 4.1: The Verifi interface: Account View (A), Social Network View (B), Tweet Panel (C), Map View (D), and Entity Word Cloud (E). The interface can be accessed at Verifi.Herokuapp.com.

visualization of account mentions (which includes retweets) as a primary view to allow users to investigate relationships between accounts; (ii) we developed an accounts view with *account-level temporal (daily) trends* as well as the most predictive linguistic features to facilitate users' account-level investigation into the rhetoric and timing of each account's tweets; and (iii) to choose the most effective *linguistic features*, we created a model to predict which linguistic features most accurately can predict the veracity of different accounts.

In addition to three different analytical cues inspired by Volkova *et al.* and our predictive model, we included visualizations and data filtering functions to allow participants to qualitatively examine and compare accounts. Based on existing research conducted on the ways news can be slanted and the diffusion of misinformation [176, 171, 177, 170], we included visual representation of three types of extracted *entities (places, people, and organizations)* to enable exploration through filtering.

Table 4.1: Distribution of types of news outlets

Type	Real	Propaganda	Clickbait	Hoax	Satire
Account	31	30	18	2	2

4.3.1 Dataset

To create our dataset, we started with a list of 147 Twitter accounts annotated as propaganda, hoax, clickbait, or satire by Volkova *et al.* [167] based on public sources. We then augmented this list with 31 mainstream news accounts [178] that are considered trustworthy by independent third-parties.[3] We collected 103,248 tweets posted by these 178 accounts along with account metadata from May 23, 2017 to June 6, 2017 using the Twitter public API.[4]

We then filtered the 178 accounts using the following criteria indicating that the account is relatively less active: (i) low tweet activity during our data collection period; (ii) recent account creation date; and (iii) low friends to follower ratio. In addition to these three criteria, we asked two trained annotators to perform a qualitative assessment of the tweets published by the accounts and exclude extreme accounts (e.g., highly satirical) or non-English accounts. After these exclusions, we had a total of 82 accounts, distributed along the categories shown in Table 4.1.

4.3.2 Data processing and analysis

To analyze our tweet data, we extracted various linguistic features, named entities, and social network structures. The role of the computational analysis in our approach is to support hypothesis testing based on social data driven by social science theories [179].

Language features: Language features can characterize the style, emotion, and

[3]https://tinyurl.com/yctvve9h and https://tinyurl.com/k3z9w2b

[4]The Verifi interface relies on a public Twitter feed collected by the University of North Carolina Charlotte.

Table 4.2: 34 candidate language features from five sources.

Source	Features	Example
Bias Language Lexicon-driven	6	Bias, Factives, Implicatives, Hedges, Assertives, Reports
Moral Foundation Lexicon-driven	11	Fairness, Loyalty, Authority, Care
Subjectivity Lexicon-driven	8	Strong Subjective, Strong Negative Subjective, Weak Neutral Subjective
Sentiment Model-driven	3	Positive, Negative, Neutral
Emotions Model-driven	5	Anger, Disgust, Fear, Joy, Sadness, Surprise

sentiment of news media posts. Informed by prior research that identified multiple language features for distinguishing real versus fake news [167], we consider five language features, including *bias language* [180], *subjectivity* [181], *emotion* [182], *sentiment* [183], and *moral foundations* [184, 185]. For example, *moral foundations* is a dictionary of words categorized along eleven dimensions, including care, fairness, and loyalty. Table 4.2 provides an overview of the features we used to characterize the language of the tweets, with each feature containing multiple dimensions.

In total, we test 68 different dimensions (i.e., 34 different language feature dimensions and each with two different normalization methods – either by number of tweets or number of words) using a supervised machine-learning algorithm (Random Forest) with a 70/30 training/validation split. We eliminated highly correlated (redundant) features (see supplemental materials). Figure 4.2 provides the ranking of the top 20 predictive language features.[5] Using this ranking, we decided to include eight language features within Verifi: *Bias, Fairness (as a virtue), Loyalty (as a virtue),*

[5]This model had a 100% validation accuracy (24 out of 24) on the 30% validation dataset.

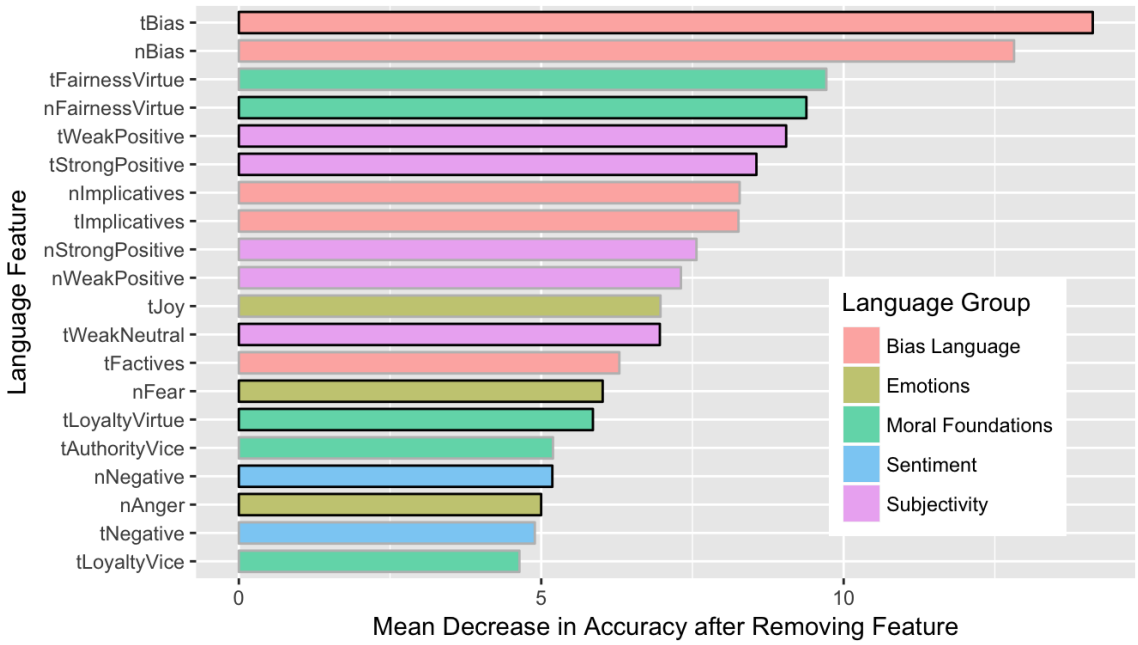


Figure 4.2: Top 20 most predictive language features of Fake and Real news outlets as measured by each feature's average effect on Accuracy. 't' prefix indicates the feature is normalized by the account's tweet count and 'n' indicates normalization by the account's word count (summed across all tweets). Features with borders are included in Verifi.

Negative sentiment, Positive sentiment, Fear, and Subjectivity to assist users in distinguishing fake and real news.[6]

Entity Extraction and Geocoding: Verifi includes a word cloud to display the top mentioned entities and enable the comparison of how different media outlets talk about entities of interest. We extract people, organization, and location entities from the tweets.

Social Network Construction: To present the interactions between the accounts on Twitter, we construct an undirected social network. Edges are mentions or retweets between accounts. Nodes represent Twitter news accounts (82 nodes) as well as the top ten most frequently mentioned Twitter accounts by our selected accounts.

[6]We averaged Strong-Weak subjectivity measures into one single measure.

4.3.3 The Verifi User Interface

The Verifi user interface is developed using D3.js, Leaflet, and Node.js. The interface consists of six fully coordinated views that allow users to explore and make decisions regarding the veracity of news accounts (Figure 4.1).

The Accounts View (Figure 4.1A) provides account-level information including tweet timeline and language features. The circular button for each account is color coded to denote whether the account is considered real (green) or fake (red). The accounts colored in gray are the ones participants are tasked to investigate in our experiment. The timeline shows the number of tweets per day. The array of donut charts shows the eight selected language features (scaled from 0-100) that characterize the linguistic content. For example, a score of 100 for fairness means that an account exhibits the highest amount of fairness in its tweets compared to the other accounts. Users can sort the accounts based on any language feature. The Account View provides an overview of real and fake accounts and enables analysis based on language features and temporal trends.

The Social Network View (Figure 4.1B) presents connections among news accounts (nodes) based on mentions and retweets (edges). The color coding of the nodes is consistent with the Accounts View (i.e., green for real, red for fake, gray for unknown). To increase the connectivity of the news accounts, we included ten additional Twitter accounts. These ten accounts (colored white) are the top-ranked Twitter accounts by mention from the 82 news accounts over the two week period. The Social Network View allows users to understand how a specific account is connected to fake or real news accounts on the social network.

Entity Views: The people and organization word clouds (Figure 4.1E) present an overview of the most frequently mentioned people and organization entities. The word clouds support the filtering of tweets mentioning certain entities of interest, thus enabling comparison across accounts. For example, by clicking on the word “Amer-

ican," accounts that mention this entity would be highlighted in both the Accounts View and the Social Network View. In addition, tweets mentioning "American" will appear in the Tweet Panel View.

Map View: The Map View provides a summary of the location entities (Figure 4.1D). When zooming in and out, the color and count of the cluster updates to show the tweets in each region. Users can click on clusters and read associated tweets. Users can also filter data based on a geographic boundary.

Tweet Panel View: (Figure 4.1C) provides drill-down capability to the tweet level. Users can use filtering to inspect aggregate patterns found in other views. Within the tweet content, detected entities are highlighted to assist users in finding information in text. This view is similar to how Twitter users typically consume tweets on mobile devices.

4.4 Experiment Design

We designed a user experiment to study how people make decisions regarding misinformation and the veracity of new accounts on Twitter with the help of the Verifi system.

4.4.1 Research Questions

Situated in the context of decision-making with visual analytics, we organized our research focus on the following research questions:

RQ1: Would individuals make decisions differently about the veracity of news media sources, when *explicitly asked to confirm or disconfirm* a given hypothesis?

RQ2: How does uncertainty (conflicting information) of cues affect performance on identifying accounts that post misinformation?

4.4.2 Experiment Stimuli

After developing the Verifi interface, we loaded data from all 82 accounts (Table 4.1) into the system. To minimize the effect of preconceived notions, all news outlet

Table 4.3: Eight accounts with masked account names. Background colors indicate real (green) and fake (red).

Mask Name	Description
@XYZ	A news division of a major broadcasting company
@GothamPost	An American newspaper with worldwide influence and readership
@MOMENT	An American weekly news magazine
@Williams	An international news agency
@ThirtyPrevent	A financial blog with aggregated news and editorial opinions
@ViralDataInc	An anti right-wing news blog and aggregator
@NationalFist	An alternative media magazine and online news aggregator
@BYZBrief	Anti corporate propaganda outlet with exclusive content and interviews

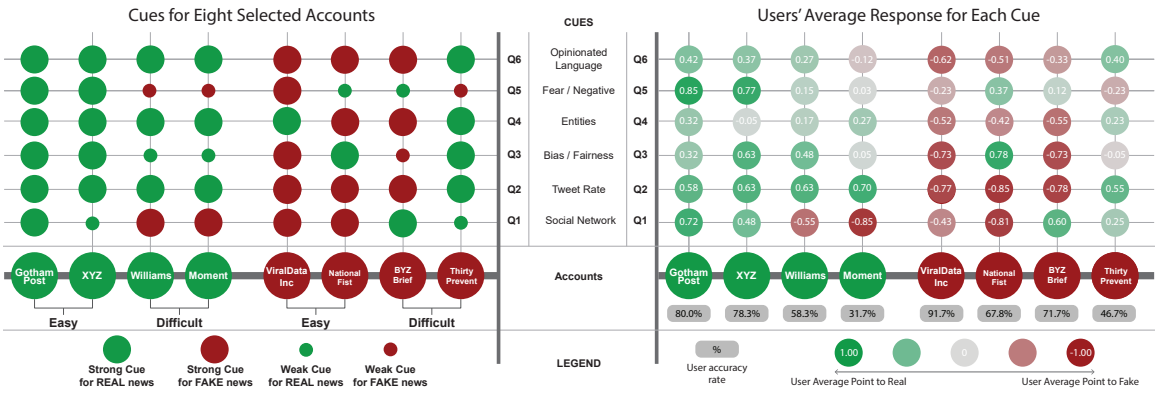


Figure 4.3: Available cues for selected accounts (column) and users' response regarding the importance of these cues (row, Q1-Q6). Left: Shows each of the eight selected accounts as well as the cues available for each of them. Right: Shows average of importance for each cue per account based on participants' responses. Values in gray circles below each account name show average accuracy for predicting that account correctly. The left figure is purely based on the (conflicting) information presented in the cues and is independent from user responses. The right figure based on the user responses on the importance of each cue coincides with the information in the left table.

names were anonymized by assigning them integer identifiers. Given the in-lab nature of the user studies and time limitations, we selected eight accounts that participants would investigate and would label as either real or fake based on their own judgments. The accounts were chosen to cover a range of different cues and degrees of uncertainty. We based our selection of experiment stimuli on classic studies in confirmation bias [28, 186].

Due to institutional concerns, we have masked the names of those accounts while preserving the nature of their naming. The eight selected accounts (4 real and 4 fake) with their masked names and description are shown in Table 4.3. The source of the description is Wikipedia and identifying information was removed to anonymize the accounts. Our goal in selecting the experiment stimuli was to enable participants to make decisions about a wide range of content with the aid of varied, sometimes conflicting, cues.

4.4.3 Experiment Tasks

To test the effect of confirmation bias, we designed an experiment with three experimental conditions: Confirm, Disconfirm, and Control. In the Confirm condition, participants were given a set of six cues about the grayed out accounts (i.e.,the eight selected accounts shown in Table 4.3) and were explicitly asked to *confirm* a given hypothesis that all grayed-out accounts were fake accounts. Similarly, in the Disconfirm condition, participants were explicitly asked to *disconfirm* the given hypothesis that all gray accounts were fake. Our third experiment condition was the Control, where the participants were simply asked to judge the veracity of the accounts; they were given neither the initial hypothesis nor the set of six cues. Following classic psychology studies in confirmation bias [19] where the information presented to the participants has inherent uncertainty, we added the element of uncertainty to the cues. We provided six cues (Q1-Q6) to the participant, of which three cues pointed to the account being real and three cues pointing to the account being fake. Each cue corresponds to a view in the Verifi interface.

The decisions that participants needed to make for the gray accounts involved answering (True/False/Did Not Investigate) for each of the six statements listed below. Each statement is the same as the cue presented to the participant in the confirm and disconfirm condition; the purpose of the statements is to gather information on which cues the participants relied on when making decisions for a certain account.

Q1 This account is predominantly connected to real news accounts in the **social network** graph. This characteristic is typically associated with known real news accounts.

Q2 The average **rate of tweeting** from this account is relatively low (less than 70 tweets per day).

Q3 On the language measures, this account tends to show a higher ranking in **bias**

measure and fairness measure. This characteristic is typically associated with known real news accounts.

Q4 This account tends to focus on a subset of polarizing **entities** (people, organizations, locations) such as Barack Obama or Muslims as compared to focusing on a diverse range of entities.

Q5 On the language measures, this account tends to show a low ranking in **fear and negative** language measures.

Q6 The tweets from this account contain **opinionated language**. This characteristic is typically associated with known fake news accounts.

These statements and the cues given to the participants at the beginning of the experiment (along with the hypothesis) are the same. Based on our data collection and analysis, statements Q1, Q3, and Q5 point to an account while being a real account and while the rest of the statements (Q2, Q4, and Q6) point to the account being fake. For certain statements, we explicitly included information characterizing whether the cue pointed to the account being real or fake (as shown in Q1, Q3, and Q6). The presentation of these cues were deliberately chosen to add to the uncertainty of information presented to the users. In addition to asking participants about their decision-making process on each statement listed above, we also asked the users to rate the importance of each view in the Verifi interface in making those decisions (the Accounts View, Social Network View, Tweet Panel View, and Entity View) on a scale of 1 to 7. Additionally, we asked participants to indicate the confidence of their decision on a scale of 1 to 7 for each account, as well as an optional, free-form response section where participants could provide any additional information as a part of their analysis. All these questions were part of a pop-up form that was displayed when the participants clicked the "Choice" button shown alongside the account number in the

Accounts View (Figure 4.1A). The responses to this form were captured in a database upon submitting the form during the task.

The information regarding each gray account and its cues is summarized in Figure 4.3 (Left). For simplicity of presentation, green circles indicate a cue pointing towards account being real, red circles indicate a cue pointing towards account being fake. The overview of how they score on cues demonstrate how the accounts exhibit different levels of difficulty for decision-making. For example, all evidence pointed to the *@GothamPost* account being real, which means that ideally, upon investigation, a participant would answer True for Q1, Q3 and Q5 and False for Q2, Q4 and Q6 when making their decision for that account. However, other real accounts chosen for investigation had more uncertainty in the cues. Notably, the *@MOMENT* account was chosen as one of the difficult accounts since it had a misleading social network cue (Q1) in that it had only one connection to a fake news account. For the fake news accounts chosen, *@ViralDataInc* had all evidence pointing towards the account being fake (except Q4, which means that the tweets from this account covered a diverse range of entities). This would make *@ViralDataInc* easier to judge as fake than, for instance, *@ThirtyPrevent*, which exhibits many more misleading cues.

4.4.4 Experiment Procedure and Participants

We recruited participants via in class recruitment, email to listservs, and the psychology research pool at our institution. Once signed up, participants came to the lab for a duration of one hour. After the informed consent procedure, participants viewed two training videos designed for this experiment. The first video introduced the interface and explained the different views. The second video provided a task example to determine the veracity of a sample account not used in the study. Both videos were identical across all conditions. After this training, participants completed a pre-questionnaire consisting of questions related to their demographics (age, gender, education), familiarity with visual analytics, social media, and Big-5 personality

questions [146]. The participants were then assigned the task and asked to complete the task in 30 minutes. After completing their task, participants completed a post-test questionnaire which included six vignettes to assess participant's propensity to confirmation bias in general [19].

Sixty participants completed the study, evenly split into three treatment groups. Participant ages were between 18 and 41 (mean=24.7). The gender distribution was 45% male and 55% female. A majority of the participants were undergraduates (65%), followed by Master's (16.7%), Ph.D. (8.3%), and others (10%). The distributions of the participants between computing (48.3%) and non-computing majors (51.7%) was relatively even.

4.5 Data Analysis Methods

In this section, we introduce the analysis methods applied to our experiment data to answer the two research questions.

To address RQ1, namely, "are there significant differences in the way participants interact with the data and their resulting judgments based on the experiment condition?", we use one-way analysis of variance (ANOVA) for testing and post-hoc Tukey's honest significant difference (HSD) test to determine significance (α=0.05). Our experiment design is a between-subjects design with one level: the experimental condition.

To address RQ2 regarding the effects of uncertainty, we used two logistic regressions to explore the effects of uncertainty (in cues, accounts, confidence, and treatment groups) had on users' decision-making. Each regression included a different dependent variable: users' accuracy (1 = correct decision, 0 = incorrect decision) and fake determination (1 = fake prediction, 0 = real prediction). This analysis allows us to determine which factors were most important and aligned with our expectation in terms of direction. For example, as mentioned in the Experiment Stimuli section, cues Q1, Q3, and Q5 were selected to point to real accounts, suggesting a negative

relationship with fake prediction (or less than 1 log odds ratios). Alternatively, cues Q2, Q4, and Q6 were selected to point to fake accounts (i.e., positive relationship or greater than 1 log odds ratios). In addition, we can also identify which cue was most important in decision-making as the one with the largest (in absolute magnitude) coefficient. In addition to the cues, we also include dummy variables for the account-level (using @XYZ as the reference level) as well as include users' confidence level and treatment group (Control group is the reference level) to understand if these factors played an additional role in the users' decisions.

4.6 Analyses Results

In this section, we describe our findings and results. The detailed discussion about the implications of these findings is in the Discussion section.

4.6.1 RQ1: Testing the Effects of Confirmation Bias

Table 4.4 shows the user accuracy rate and fake prediction rate across all three experiment conditions. We found no significant differences between the experimental conditions, on a diverse range of factors. Participants in all three conditions did not differ on the number of accounts labeled as fake and the number of accounts labeled as real ($p>0.05$ for both). We tested the accuracy rate and found no significant difference in the rate of accuracy across experimental conditions ($p>0.05$). In addition, we tested whether the participants interacted differently with the data, depending upon the experiment condition. To test this hypothesis, we computed the total time spent for participants in each condition, including time spent interacting with the data presented in each view in Verifi (e.g., Social Network View, Accounts View). We found no significant differences in the amount of time spent overall or in any specific panel on the interface across the three conditions.

Table 4.4: User accuracy and Fake prediction across conditions.

	Control	Confirm	Disconfirm
Accuracy	60.4%	73.8%	63.1%
Fake Prediction	54.1%	55.0%	51.9%

4.6.2 RQ2: Measuring the Impact of Uncertainty

While we did not find significant differences in users' decisions (e.g., accuracy) between experiment conditions, we expect differences in accuracy and fake prediction given uncertainty in cues for each account. Based on the cues in Figure 4.3 Left, we categorize accounts into two types: Easy and Difficult. These categories are based on how each account scores on the six cues and are independent from users' responses. In this section, we describe regression analysis to analyze the effect of cues and account on users' decision-making. We then present thematic analysis of users' comments regarding their decisions.

Regression Analysis: Our results provide evidence that the prevalent factors in users' decision-making were the cues and the accounts. Table 4.5 provides the log odds ratios for the independent variables by each regression. We observe three findings. First, in general cues have a significant effect on users' fake prediction and accuracy. For the cues, we recoded the responses to indicate whether the cue was used consistent or not (e.g., depending on the direction of the cue relative to fake or real accounts). We find that the opinionated, fear, and social network cues were the most important in explaining correct decisions when used consistently. Alternatively for explaining Fake decisions, we find that log odds ratios align to the cue direction as mentioned in Figure 4.3. For example, cues Q2, Q4, and Q6 point to the account being fake and we find the log odds ratios above one, although only Q4 and Q6 are statistically significant.

Second, we find that certain accounts had a significant effect on both users' accuracy

	Correct?	Type	Group	Comment	category
1	Yes	real	easy	Several language features are consistent with predominantly real accounts	quantitative
2				News appears more factual reporting rather than opinionated discussion of events, which leads me to believe it is a real news account.	quantitative
3			difficult	This account does not seem to deal much with controversial topics, and although it has a lower loyalty score, it has a high fairness score and high bias, which are normally indicative of real accounts.	quantitative + qualitative
4				While this account only has one connection and it's to a fake account, I didn't notice anything suspicious in the tweets. The People and Organizations view only showed topics that are normally discussed in the news and nothing overly controversial.	quantitative + qualitative
5		fake	easy	A lot of the tweets were not even news but simply them stating their opinions about a variety of issues.	qualitative
6				Only follows one account, tends to only tweet about one topic (Trump), and it's all negative and uses opinionated language.	quantitative
7			difficult	High fairness but low loyalty. Little amount of tweets (seemed inconsistent). Very high anger. When looking at the network, it was associated with a wide range of different accounts.	qualitative + qualitative
8				This account is 100% angry, with a low tweet amount. This user also doesn't focus on that many people within their tweets.	quantitative + qualitative
9	No	real	easy	Compared timeline of tweets as other tweets. The timeline and tweet content about taking Mosul for this account do not match with other "real" news.	quantitative
10				For this account, language within the tweets tipped me to believing this is a fake news account, or at least an extremely conservative or right-leaning (with high bias) news account. Wordage like "marxist left mainstream media" for instance.	qualitative
11			difficult	Contains a lot of opinionated language in it's tweets.	qualitative
12				..Despite the high tweet rate, their bias and subjectivity scores were high, which tends to relate to fake accounts. That added to the fact that it's only linked to another fake account and some verified accounts led me to believe this is a fake.	quantitative
13		fake	easy	Though this is very opinionated, it leans towards an overall criticism of America, as opposed to an organization attempting to sway a constituency.	qualitative
14				Admittedly, personal bias played a role in deciding the "real"ness of this account as the information in the tweets, though not seemingly produced by big media, appears real, though not unbiased.	qualitative
15			difficult	Connected to real accounts and has lower subjectivity.	quantitative
16				Although there was a high rating of anger, it seems as though none of the tweets expressed any anger or high bias.	quantitative + qualitative

Figure 4.4: A sample of users' comments about their decisions. Highlighted text shows users' mention of either a qualitative or quantitative reason. Green denotes reasons/cues pointing to the account being real while red pointing to being fake.

and fake prediction. This observation implies that some accounts were more difficult and systematically over or under predicted as fake. For example, @MOMENT has a very low log odds ratio for users' accuracy as users overwhelmingly incorrectly predicted @MOMENT, a real-difficult account, as fake (as indicated by its high log odds ratio for fake prediction).

Last, we find that confidence has no significant relationship in explaining accuracy or fake decisions. While there may be a univariate relationship between confidence and user decisions, this may likely be explained through the account level dummy variables as confidence also varied by accounts. Also, we find the Confirm condition maintains a weakly significant effect on accuracy relative to the Control group (reference level for treatments).

Thematic Analysis of Comments: Our regression analysis revealed that cues played an important role in users' decision making on misinformation. When cues point to conflicting directions of an account being real or fake, users are more likely to arrive at inaccurate decisions. In each decision, users had the option to leave comments in regards to their decisions. These comments are extremely valuable

in helping us decipher users' rationales. We examined all comments (95 total) and thematically categorized users' strategies. Our analysis focuses on how different usage on all or a subset of the cues affect their decision making. Similar to our quantitative analysis, we evaluate these themes through the lens of cue uncertainty and account difficulties.

Our thematic analysis classified comments into three categories: *Quantitative (32 comments), Qualitative (37), and Qualitative + Quantitative (26). We categorized mentions of social network connection, language feature score, and tweet timeline as quantitative. Any mention related to entities and users' understanding of the text of tweets such as "opinionated language," "news-like text," and "style of text" were considered qualitative. The quantitative and qualitative dimensions extracted from the comments aligned well with the six cues provided to the participants.*

Easy Accounts: Easy accounts (column 1, 2, 5, 6 in Figure 4.3 left) are the ones with most cues pointing to the accounts being either real or fake; thus leading many users to correct decisions. Fifteen comments for the easy accounts mentioned quantitative cues such as language features scores (Figure 4.4, row 1) and social network connections (Figure 4.4, row 6) as the basis of their decisions. 12 of these comments led to correct decisions. Seventeen comments focused on the qualitative cues such as opinionated language or entities, e.g., one real account decision based on "factual reporting" and a fake account decision due to seeming "too opinionated" (Figure 4.4, rows 2 and 5).

Difficult Accounts: Difficult accounts (column 3, 4, 7, 8 in Figure 4.3 left) are the ones with the cues pointing to contradicting directions, resulting in more uncertainties in decision making. Seventeen comments focusing on quantitative cues such as fewer social network connections to other real news accounts for some real-difficult accounts yielded eleven inaccurate decisions (Figure 4.4, rows 12 and 15). Furthermore, twenty comments focused on qualitative cues such as users' notion of opinionated language,

in which seven cases it drove them to wrong decisions (Figure 4.4, row 11). Finally, fourteen comments focused on both quantitative and qualitative cues with only three of them yielding wrong decisions. In two of these cases, users decided to disregard the account's anger ranking (Figure 4.4, row 16).

We observe that when users leverage both quantitative and qualitative cues with a thorough analysis of an account, they are more likely to make an accurate decision. Most comments contained a mix of qualitative and quantitative analysis (including language features, social network connections, and opinionated language) helped users to come to the correct decisions (Figure 4.4, rows 3, 4, 7 and 8).

4.7 Discussion and Future Work

Our goal was to assess the effect of confirmation bias and uncertainties on the investigation of misinformation using visual analytics systems. Although our post-questionnaire vignette, based on prior psychology research [19] showed that most of our users demonstrated a high level of confirmation bias, our experiment did not find significant differences between the experiment conditions. One explanation would be the hypothesis (all eight accounts are fake) we gave the participants did not resonate with them. If we had asked the participants to form their own hypothesis of the eight accounts being either real or fake by going through an example account, they may have been more invested in the hypothesis and inclined to confirm or disconfirm it. Another explanation involves the use of Verifi, the visual analytics system that empowers users' decision-making by allowing users to interatively analyze multiple aspects of the news accounts. Often, people are instructed to 'slow down' and inspect information more critically [18] as an antidote to falling for confirmation bias. The Verifi interface could have played a role of somewhat mitigating confirmation bias in our experiments. This will be the subject of our follow-up studies.

We observe that participants' responses to the cues were consistent with the account uncertainties/difficulties. Figure 4.3-Right shows how users' average cue responses

matched our original understanding of these accounts. Moreover, our regression analysis shows that certain cues significantly affected our users' decisions (Q4-Q6) more than others. Opinionated language which had the strongest effect on users fake prediction stands out as an important lesson learned for future attempts to address misinformation. The fact that we allowed the opinionated cue to be purely based on users' understanding of tweet texts, opens a whole new research question: How can we help users' to more objectively identify/quantify opinionated language?

Furthermore, we find that uncertainty affected our users' prediction accuracy. Our research shows that when a combination of quantitative and qualitative cues are presented clearly and with minimal uncertainty, users are successful in correctly differentiating between fake and real news accounts. In order to be resilient to these uncertainties, it is essential to take effective measures to communicate these uncertainties, motivate users to not be anchored on specific cues, and to holistically focus on a combination of qualitative and quantitative evidence. We plan to conduct a followup experiment with adding uncertainty of the cues to the visual analytic system to test this hypothesis. One limitation of our study was the number of accounts chosen. Due to the time duration of our study (one hour), we decided to ask each participant to make decisions about eight accounts with varying difficulties. In order to test whether our results can be generalized, we plan to conduct a follow-up study that focuses on annotating a larger number of randomized accounts. The current study provides guidance on how we would instruct human coders to categorize all accounts based on the cues into different difficulty levels.

4.8 Conclusion

This chapter introduces a visual analytics system, Verifi, along with an experiment to investigate how individuals make decisions on misinformation from Twitter news accounts. We found that the account difficulty as mixed cues indicating real versus fakeness has a significant impact on users' decisions. The Verifi system is the first

visual analytics interface designed to empower people in identifying misinformation. Findings from our experiment inform the design of future studies related to decision-making around misinformation aided by visual analytics systems.

Table 4.5: Log odds ratios for each independent variable in two logistic regressions. The Accuracy column is 1 = Correct, 0 = Incorrect Decision. The Fake column is the user's prediction: 1 = Fake, 0 = Real. The @accounts variables use @XYZ as the reference level and the Group variables use the Control Group as the reference level.

	Dependent Variable	
Independent Variable	**Accuracy**	**Fake**
(Intercept)	0.18**	0.21**
Social Network Cue (Q1)	2.03***	0.99
Tweet Rate Cue (Q2)	1.24	1.06
Fairness Cue (Q3)	1.30*	0.74**
Entity Cue (Q4)	1.43*	1.77***
Fear Cue (Q5)	1.53***	0.90
Opinionated Cue (Q6)	2.78***	2.74***
@ViralDataInc (Fake Easy)	9.86***	117.96***
@NationalFist (Fake Easy)	1.90	9.7***
@GothamPost (Real Easy)	0.95	0.84
@Williams (Real Difficult)	0.90	2.13*
@MOMENT (Real Difficult)	0.36**	5.70***
@BYZBrief (Fake Difficult)	4.91***	24.21***
@ThirtyPrevent (Fake Difficult)	3.70**	18.89***
User Confidence	1.14	0.86
Confirm Group	1.97**	0.91
Disconfirm Group	1.16	0.75

*** = 99%, ** = 95%, * = 90% confidence

CHAPTER 5: INVESTIGATING EFFECTS OF VISUAL ANCHORS ON DECISION-MAKING ABOUT MISINFORMATION.

5.1 Introduction

Visual Analytics (VA) combines statistical and machine learning techniques with interactive visualizations to facilitate high-level decision-making on large and complex data. An important attribute of an effective VA system is the support of *exploratory visual analysis* [187, 188]. Many VA systems designed for exploratory visual analysis often employ coordinated multiple views (CMV) to provide functionality including details-on-demand, linked navigation, and small multiples [189]. These VA systems offer the user flexibility to use the VA system to solve problems through many possible strategy paths and "have a dialogue with the data" [125]. However, user flexibility– like in CMV systems–can introduce trade-offs as well [190]. Zgraggen *et al.* [53] find too much freedom in visualization systems can lead to spurious insights and high rates of false discoveries, also known as the multiple comparisons problem or the **forking paths problem** [52]. Pu and Kay [52] define the forking paths problem in visualizations as "unaddressed flexibility in data analysis that leads to unreliable conclusions." They argue cognitive biases may be one reason for users' susceptibility to the forking paths problem. In this chapter, we consider the problem within the scope of one such cognitive bias, anchoring bias, and the possible effect pre-task training can have on the complex decision-making task of social media misinformation identification using a CMV VA system.

Cognitive biases are the result of the over-reliance of heuristics, or rules-of-thumb, for decision-making tasks to make decisions with relative speed [26]. An emerging topic within the VA community considers the role of cognitive biases in VA decision-

making [23, 191, 192]. Cognitive biases have been shown to affect decision-making processes in predictably faulty ways that can result in sub-optimal solutions when information is discounted, misinterpreted, or ignored [26]. One cognitive bias relevant to exploratory visual analysis with VA systems is **anchoring bias**. It refers to the human tendency to rely too heavily on one and most likely the first piece of information offered (the "anchor") when making decisions [193]. Past studies from psychology and cognitive science have focused on numerical anchoring, in which an initial numerical value anchors judgment and the subsequent adjustment with updated information [193, 27, 194]. Cho *et al.* [32] provided evidence anchoring transfers to VA; specifically **visual anchoring**, which is the over reliance on a single or subset of views during exploratory visual analysis.

To situate our experiment in real-world decision-making tasks with VA systems, we selected the application of misinformation identification. Recently, the topic of combating misinformation has received much attention in many fields including machine learning, psychology, journalism, and computational social science [195, 196, 164]. While a variety of fully automated techniques have been developed, more direct interaction like laboratory experiments with users on misinformation decision-making is needed [197].

Our work makes the following salient contributions:

1. We conducted an empirical study on the effects of visual anchoring in decision-making. Specifically, we investigated misinformation in social media in a between-subjects design laboratory experiment with 94 participants.

2. Introduction and formalization of strategy cues and visual anchors as treatments to intervene within the visualization training process.

3. Careful integration of strategy cues from psychology literature as hypotheses to test the interaction between visual anchoring and providing hypotheses in

visual decision-making tasks.

4. Quantitative analysis on factors that affect anchoring bias in VA to measure visual anchors' impact on user decisions, confidence, time spent, and interactions.

Understanding the effect of cognitive biases like anchoring in visual analysis serve as an important first step to raising awareness and possibly mitigating cognitive biases with visual analysis. At the end of the chapter, we connect findings from our experiment to practices of interacting with participants on a newly designed visual analytic systems. The findings of our experiments shed more light on how and when anchoring effects can occur in visual analytic systems and call for more careful consideration of training users or designing tutorials for a visual analytic system.

5.2 Background

In this section, we review past research on cognitive biases in visualizations. We also review literature that motivated our experiment design and research questions.

5.2.1 Anchoring & Cognitive Biases in VA

A cognitive bias is a systematic and involuntary cognitive deviation from reality [198, 192].[1] Introduced by Tversky and Kahneman [26], cognitive biases have since had a long history of investigation by various social scientists [199]. Ellis and Dix [22] provide an early case on exploring the role of cognitive biases in VA and, more recently, several papers have provided theoretical frameworks or taxonomies of cognitive biases in VA [168, 23, 191, 192]. Empirically, VA research on cognitive biases have developed studies on a variety of biases including attraction [81], selection [82], availability [83], and confirmation bias [30, 87]. In our study, we explore the phenomenon of anchoring [26], which is the tendency to focus too heavily on one piece of information when making decisions. Originally considered in the task of open-ended numerical

[1] Dimara *et al.* [192] provide a detailed discussion that "reality" stems from normative models of decision-making, which in itself leads possible controversies on assessing cognitive biases.

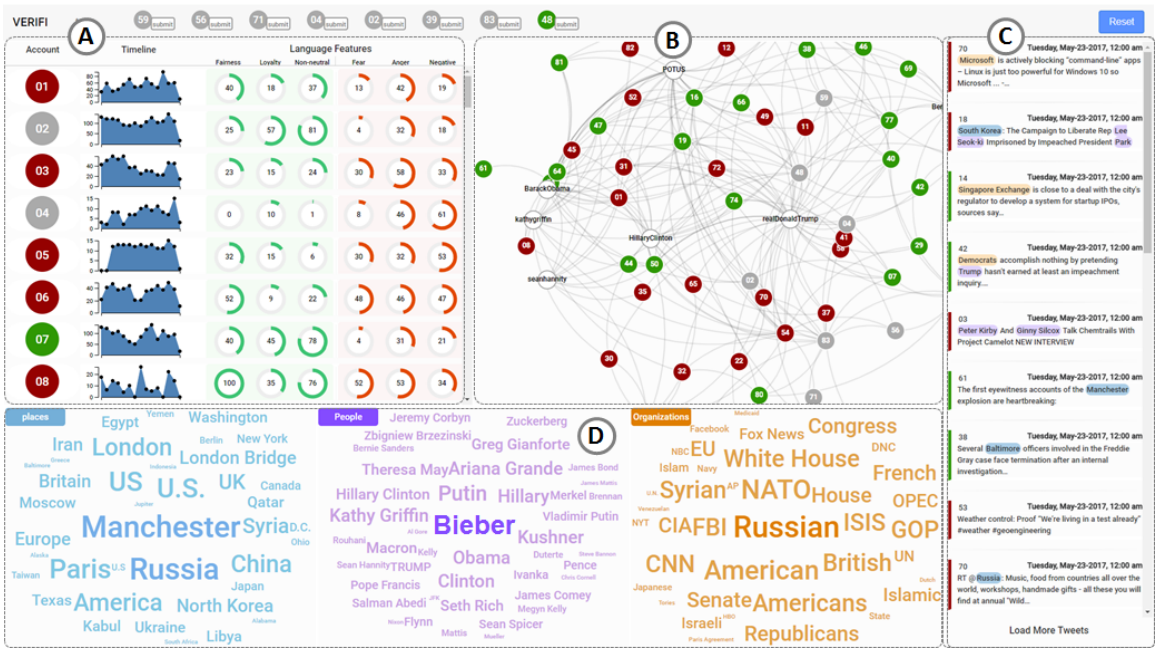


Figure 5.1: Screenshot of Verifi. Verifi is comprised of four views: (A) Language Features View, (B) Social Network View, (C) Tweets Panel View, and (D) Entities View. Progress Bar and Form Submit buttons are at the top.

decisions [26, 27], anchoring has been studied in VA both in MTurk studies using scatterplots [84] and more complex, lab-based experiments using a CMV system [32].

As one of the first studies on the effect of anchoring in VA, Cho *et al.* [32] employs an open-ended task of identifying protest-related events from social media data. They analyzed the impact of visual anchors on the reliance of views and analysis paths; however, users decisions were more affected by classical numerical anchors within the task. The impact of visual anchoring on user performance was not measured and limited information was captured at the time of each decision.

5.2.2 Strategy Cues in Psychology Experiments

Our motivation for providing strategy cues in the experiment design is rooted in prior research from psychology [186, 200, 201, 202, 203]. In their influential conceptual model for exploratory analysis, Pirolli and Card argued that identifying an exploration strategy or hypothesis is one technique that users of visual analytic systems can benefit from [204]. To illustrate, research has demonstrated that users preferred to

devote attention to stimuli that matched a given hypothesis or template, even in the presence of alternate, more optimal strategies [186]. Amer *et al.* [200] designed experiments in which participants were given explicit and implicit spatio-temporal cues in a visual event coding task and found systematic effects of the explicit and implicit cues on users' attention within the visual analytic system and how these cues affected processing of information.

5.2.3 Possible Training Induced Biases

A recent survey of visualization evaluation practices from the Vis Community highlighted that many publications need to observe more evaluation reporting rigor by providing important methodological details [205]. In particular, there is a lack of consistent reporting on how the participants were trained (by experimenters, with or without a script, training videos, example strategies to complete the task, etc.). In our experiment, visual anchors are introduced during training the participants on how to use a visual analytic system to investigate misinformation. We will investigate the impact of the visual anchors on users' performance and behavior during analysis. In summary, how participants were trained may significantly impact their task completion, thus we argue for more consistent reporting of these details.

5.3 Experiment

In this section, we outline our experiment design including reviewing Verifi, the VA system used in the experiment, our research questions, variables, and hypotheses.

5.3.1 The Verifi System

For our study we use Verifi [30] (Figure 5.1), an interactive CMV system for identifying Twitter news accounts suspected of spreading misinformation. Verifi includes four views: Social Network, Language Features, Tweet Panel, and Entities. Each view provides users with different features in detecting misinformation [167]. The Social Network and the Language Features views are the two primary views; the

Entity View and Tweet Panel are secondary views. Following Cho *et al.* [32], we selected Verifi to test visual anchoring as its an example of a complex CMV system. CMV systems inherently require users to make choices on which views to use and strategies to switch between views. Accordingly, visual anchoring may occur in such systems when a user is biased into over relying on one view and possibly leading to a sub-optimal decision.

The system includes two weeks of tweets from 82 Twitter news accounts. Each account name was converted to an integer code (1 to 82) and annotated as a misinformation account (red), real news outlet (green), or requiring user decision (grey). The annotations are based on independent, third-party sources.[2] Each user's task is to make a decision on the veracity (real or suspected of spreading misinformation) for eight grey accounts within a one-hour session. Following [30], the eight accounts were qualitatively selected to provide a range of difficulty level as well as consistent and inconsistent information to challenge users in their decision-making processes. Table 5.1 provides the actual Twitter handles of the eight gray accounts along with a brief description.

5.3.2 Experiment Design

We highlight **two critical design decisions** in our experiment compared to Cho *et al.* [32]:

1. We collect direct feedback from users in a submission form (Figure 5.2) to capture input at the time of each decision (e.g., view importance, strategies, and open-ended comments). Unlike Cho *et al.* [32] who captured users' decision on paper and after the task, we designed the Form Submit view (Figure 5.2) to collect information regarding the factors that influenced each decision.

[2]Suspicious accounts are based on four websites as provided in [167]. 31 real news accounts are provided through the following links: https://tinyurl.com/yctvve9h and https://tinyurl.com/k3z9w2b

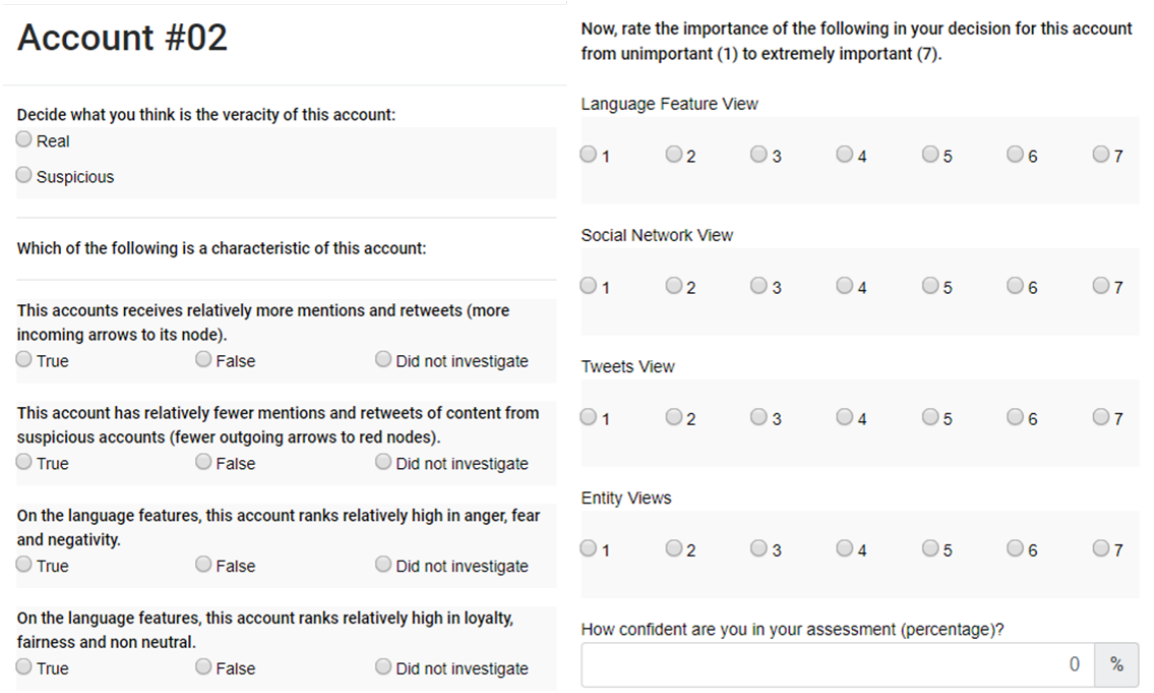


Figure 5.2: Form Submit view of Verifi for Account #02 (@ABC). This pop-up provides an interface for the user decisions and feedback per account (e.g., strategy cues use, view importance, and open-ended comments (not shown).)

2. We provide **strategy cues**, or hypotheses, as a secondary condition in the form of written statements that reinforce functionality for each primary view in Verifi [30]. Strategy cues are initial hypotheses, provided on paper to users, of possible relationships between the data elements in a specific view. Strategy cues align to confirmatory data analysis as they provide a mechanism to control for possible hypotheses of the task and its functionality with an anchored view (e.g., real news accounts have lower anger, fear, or negativity). Strategy cues serve as interaction variables to the visual anchor as they may enhance the anchor's effect on a view if the user follows the view's strategy cues.

To analyze the effects of visual anchors and strategy cues in decision-making, we conducted a between-subjects, repeated measures laboratory experiment. Figure 5.3 provides the experiment flow. Each user's task is to make a decision on the veracity (real or suspicious) of eight grey Twitter accounts (see Table 5.1). Users submit

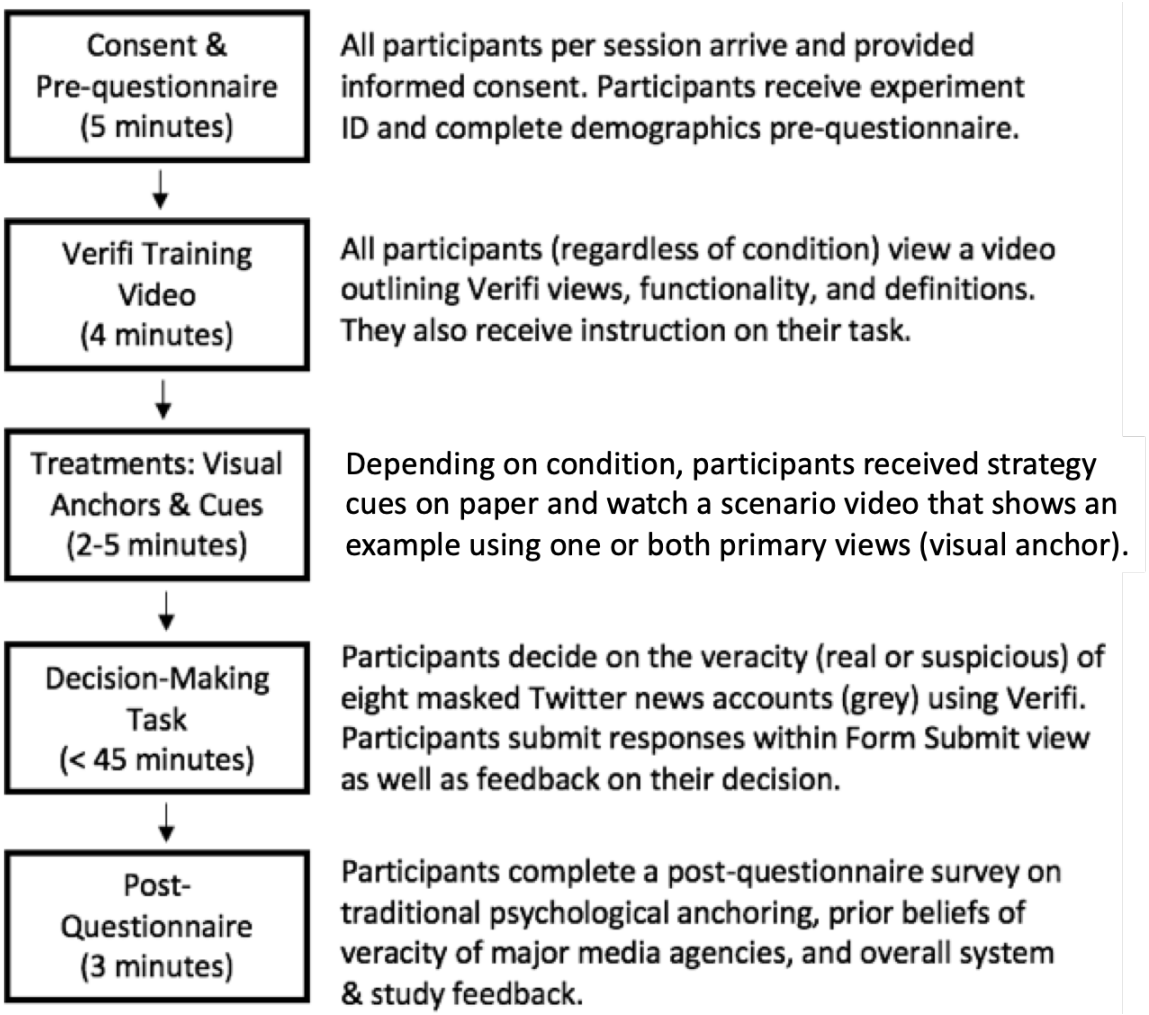


Figure 5.3: The experiment flow for each participant session.

their decisions in the Form Submit view (Figure 5.2) along with their ratings of the visualization views and strategy cues. To control for learning effects, we randomized the order of the account icons in the Progress Bar per participant.

94 users participated in our study. The gender distribution was 68% male and 32% female. Users' ages were between 21 and 56 (M = 28.7). A majority of users were pursuing a master's degree (n = 83), followed by undergraduate (n = 5), graduate certificate (n = 5), and Ph.D. (n = 1). Students were recruited through extra credit incentives offered in six courses: Visual Analytics (n = 40), Natural Language Processing (n = 25), Advanced Business Analytics (n = 14), Human Behavior Modeling

Table 5.1: Eight Twitter news accounts for users' decisions (i.e., grey accounts in the interface). Accounts were anonymized in the study.

News Outlet	Description
@zerohedge	A financial blog with aggregated news and editorial opinions
@AddInfoOrg	An anti right-wing news blog and aggregator
@NatCounterPunch	An alternative media magazine and online news aggregator
@SGTreport	Anti corporate propaganda outlet with exclusive content and interviews
@ABC	A news division of a major broadcasting company
@nytimes	An American newspaper with worldwide influence and readership
@TIME	An American weekly news magazine
@Reuters	An international news agency

(n = 6), Applied Machine Learning (n = 6), and Social Media Communications (n = 3).

Each participant session was capped at 45 minutes and averaged 27.1 minutes (SD = 7.524). Each session is identified through a participant ID and interactions (e.g., clicks, hovers, and scrolls) were saved to a MongoDB database. Computer specifications (browser, output/zoom) were controlled for to avoid them as confounding factors. Our study was approved under our institution's Institutional Review Board (IRB) policies (IRB #17-0251).

5.3.3 Research Questions

We investigate how users may be visually anchored on different *views* in a CMV system and how they might be anchored on specific interaction *strategies* based upon the training given to them. How does visual anchoring affect user performance, confidence and data coverage? Accordingly, our main research questions (RQs) are:

RQ1: What is the effect of visual anchors and strategy cues on participant perfor-

mance (i.e., accuracy, speed, and confidence) and ratings (e.g., view importance and strategy usage)?

To analyze RQ1 from a participant-level, we use aggregated[3] non-parametric bootstrapped confidence intervals [207]. In our results, we focus on effects sizes rather than p-values [208] and follow conventions provided by Dragicevic [206]. Then we employ a hierarchical model to consider both participant and task-level effects on user accuracy and confidence. Following Kay *et al.*'s [208] recommendation for Bayesian methods in HCI, we use Bayesian mixed-effects regressions with weakly-informed priors [5].

RQ2: Can users' analysis process (e.g., interaction logs) be linked to participant performance outcomes to infer user strategies?

For RQ2, we estimate condition effect sizes of user time spent per view and coverage metrics [23] using mean bootstrapped confidence intervals. To identify user behaviors with the coverage and time spent metrics, we used Ward's D2 Agglomerative Hierarchical Clustering [2] to cluster users and features using the R package `heatmaply` [209]. To determine the optimal number of clusters for the rows (features) and columns (users), we used the maximal average silhouette width method on the cophenetic distance of the dendrogram [210]. The algorithm detected five clusters on the user-level, as identified by the five colors in the horizontal dendrogram. We then annotated the five clusters based on common attributes shared by users within a cluster.

5.3.4 Independent Variables (experimental conditions)

For our experiment design, we developed six treatments in three condition groups: Control, Balanced, and Partial (Table 5.2). The **Control** group did not receive any visual anchor (i.e., scenario video). The **Balanced** group received a visual anchor that reviewed a strategy using both primary views. The **Partial** group received a

[3]Given HCI's focus on people not tasks, Dragicevic [206] advocates for calcuating confidence levels on a participant-level, not a task-level.

Table 5.2: Experiment treatments by condition groups.

	Visual Anchor	Strategy Cues	Users	Decisions
Control	None	None	14	112
	None	1S, 2S, 1L, 2L	15	120
Balanced	SN -> LF	1S, 2S, 1L, 2L	17	134
	LF -> SN	1S, 2S, 1L, 2L	16	128
Partial	SN Only	1S, 2S	15	119
	LF Only	1L, 2L	17	135

visual anchor that covered only one primary view but not the other.

In addition, the difference between each group condition was the *strategy cues* (or hypotheses) given to participants that reinforce each primary view. Each *strategy cue* is a hypothesis on how to identify real news accounts that aligns to one of the two primary views in the Verifi: Language Features view **(L)** and Social Network view **(S)**. The Language Features view presents predictive linguistic features for each account, such as fairness, loyalty, anger, and fear. The Social Network view provides retweet and mention relationships [30]. The cues are:

Cue 1L: "On the *language measures*, real news accounts tend to show a higher ranking in loyalty, fairness, and non-neutral."

Cue 2L: "On the *language measures*, real news accounts tend to show a lower ranking in anger, fear and negativity."

Cue 1S: "In the *social network graph*, real news accounts are less likely to mention and retweet content from suspicious accounts (fewer outgoing arrows to red nodes)."

Cue 2S: "In the *social network graph*, real news accounts tend to receive more mentions and retweets (more incoming arrows to their nodes)."

5.3.5 Dependent Variables

We have two types of dependent variables: decision and behavioral metrics (see Figure 5.4). Decision metrics are provided by users and can be divided into two groups: primary and secondary outcomes.

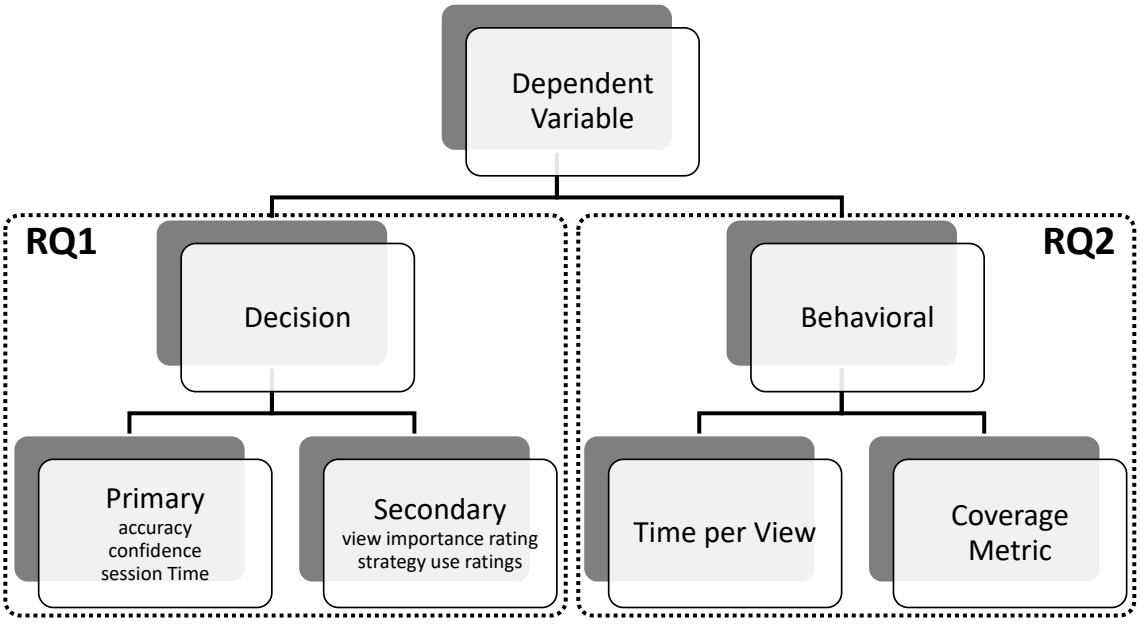


Figure 5.4: Dependent variable groups in our experiment.

Primary outcomes: We evaluated user performance based on three primary outcomes: (1) **accuracy** in correctly identifying misinformation accounts, (2) **confidence** of each misinformation decision as range from 100 (perfectly confident) to 1 (perfectly not confident), and (3) **session time** that is the time of the decision as minutes from the start of the experiment.

Secondary outcomes refer to the four **view importance ratings** and four **strategy use ratings** provided directly by each user at the time of each decision. The importance ratings use 1 (unimportant) to 7 (extremely important) Likert scale and the strategy cue ratings use a True, False, or Did Not Investigate value (see Figure 5.2).[4]

In addition to decision metrics, we also consider participants' actions as dependent variables in two types of behavioral metrics.

Time spent metrics. We measured users' time spent per view through a mouse

[4]Consistent with [30], we recoded strategy cue ratings to ensure whether the cues were consistent or not, depending on whether the account was Real or Misinformation. In this way, the cues can be interpreted as 1 = cue used consistently, -1 = cue used inconsistently, 0 = cue not investigated.

enter-exit log tracking. By using the enter-exit periods and allocating that to each view, we were able to measure participants time spent in the five views (two primary, two secondary, and Form Submit view).

Coverage metrics. Following Wall *et al.* [23], we created coverage metrics to measure participant use of key interface functionality. Specifically, we consider six primary actions: progress bar click, LF sort (combined for red/green features), and SN hovers (for grey, green, and red accounts).[5]

5.3.6 Hypotheses

Based on the RQ's, we developed the following hypotheses:

H1: Balanced visual anchor users will have the highest accuracy as users will have more information on how to use both primary views. These users will use the primary views more than the secondary views as compared to the Control groups.

H2: Partially visual anchored users will have the worst accuracy as their anchors are one-sided. These users will disproportionately interact with the view associated with their anchor. By failing to consider the opposite view, their performance will diminish.

H3: When given scenario videos that include both primary views (i.e., Balanced group), order matters. The first view provided will have a larger effect than the second, leading to an increased use (time, coverage) of the first view introduced. To evaluate, we compare performance within the two Balanced conditions.

H4: Strategy cues will improve performance, confidence, and shorten session as more information is helpful. In this hypothesis, we'll compare treatments to the Control treatment with no cues.

[5]We removed hovers less than one second after a previous action to remove unintentional actions.

5.4 Results

5.4.1 RQ1: Effects of Visual Anchoring and Strategy Cues on User Level

Contrary to **H1**, we do not find evidence that the Balanced visual anchored groups (70.2%-71.88%) have the highest accuracy. In fact, we find that the Control groups (i.e., no visual anchor/scenario video) performed just as well in terms of accuracy. Figure 5.5 provides the means and bootstrapped confidence intervals for the primary outcomes relative to the experiments conditions. Alternatively, we find some evidence in support of **H2** as the Partial-Social Network treatment had a lower accuracy (M = 61.3%) than the other groups, but not outside of 95% bootstrapped CIs. We do find evidence that the visual anchors provide a positive effect on user confidence relative to the Control conditions.

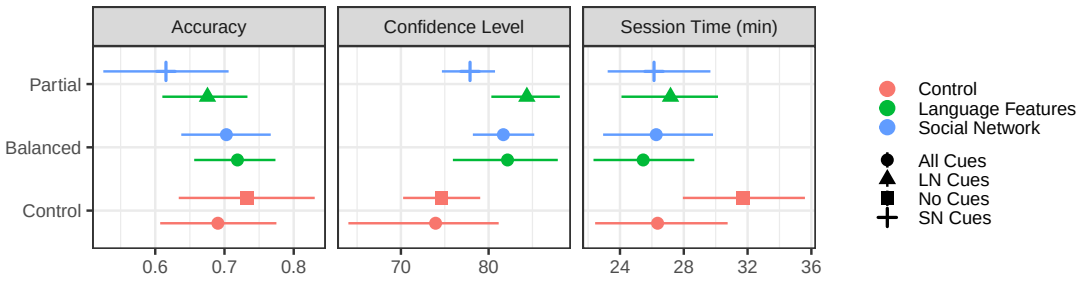


Figure 5.5: Primary outcomes means and bootstrapped 95% confidence intervals on a user-level (n = 94).

For **H3**, we find little difference in primary outcomes between the two Balanced groups, indicating that order doesn't appear to affect final decision outcomes. For **H4**, we find no evidence that the strategy cues provided an advantage in accuracy, in fact the opposite as the Control/no cues condition has the highest average accuracy. We find the cue groups do tend to have higher accuracy, but their effects may interact with the visual anchors as we find little difference between the Control groups. Last, it does seem that cues may shorten the session as the Control/no cues group, the only without any cues, had the highest average session time (M = 31.7), well above

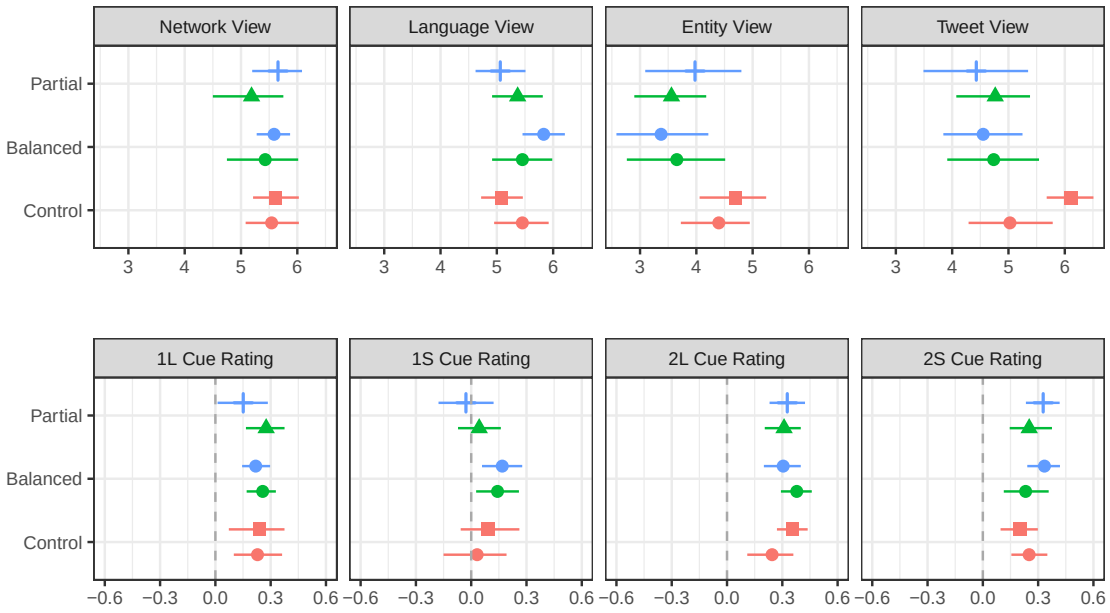


Figure 5.6: Secondary outcomes means and bootstrapped 95% confidence intervals on a user-level (n = 94). The figure uses the same color and shape encodings as Figure 5.

all other groups (ranging from 25.4 to 27.2). We'll explore this more in **RQ2** results when we decompose the session time by views.

As for the secondary outcomes, we find evidence for **H1** that visual anchors seem to diminish users' value of the secondary views. For example, Balanced (and Partial) anchored users tend to rate both the Entities and Tweet view less than Control groups (Figure 5.6 top row). In fact, we find the Control/no cues condition valued the Tweet view the highest (M = 6.1), suggesting that without any anchors or cues, users valued the qualitative secondary view the most (i.e. reading individual tweets).

Last, we find little variance across cue ratings by the six conditions (Figure 5.6 bottom row). Most average ratings range from 0.2 - 0.3, indicating a slightly above average (0) use of the cue in their decision. The one exception is **1S**, in which all but the Balanced groups average rating was within 0 for its confidence interval.

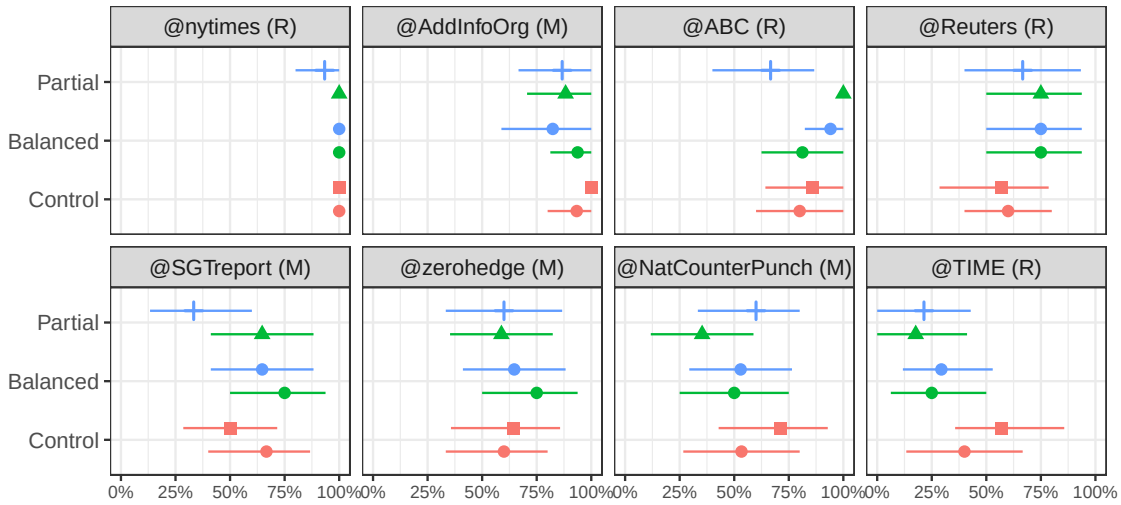


Figure 5.7: Accuracy by Twitter account and bootstrapped 95% confidence intervals on decision-level (n = 748). (R) indicates a "real" news account and (M) indicates a "misinformation" account. The figure uses the same color and shape encodings as Figure 5.

5.4.2 RQ1: Effects of Visual Anchoring and Strategy Cues on User & Task Level

One weakness of the user-level analysis is that it ignores the task-level. In Figure 5.7, we find that user accuracy varied drastically by each account (task). For example, nearly all participants correctly predicted @nytimes while most users incorrectly predicted @TIME, especially those receiving visual anchors. To consider both the user- and task-level, we use mixed-effects regressions for both accuracy and confidence.[6]

We use a Bayesian generalized linear mixed-effects regression for each of the two outcome values using the R packages `brms` [211] and `tidybayes` [212]. Our fixed effects are each *treatment* (Table 5.2), *time of decision* (in minutes), and their interactions.[7] For the random effects, we use *account* (Table 5.1) and *participant*. We use account as a random effect given the variability in difficulty from the qualitative

[6]We did not investigate total session time due to the problem of allocating time to each actions for each decision. Therefore, we only investigate accuracy and confidence as dependent variables in regression.

[7]We do not report the fixed effects of time of decision as we did not have a prior hypothesis to evaluate. However, these effects can be observed 04-regressions.Rmd | .html in the supplemental materials.

ground truth [30].[8]

For each regression we use a slight variant depending on the outcome variable format. For accuracy, a binary 1 (correct) or 0 (incorrect) variable, we use a logistic mixed-effects regression. Alternatively, confidence is a continuous variable between 0 (no confidence) to 1 (perfect confidence) and, hence, we use a linear mixed-effects model.

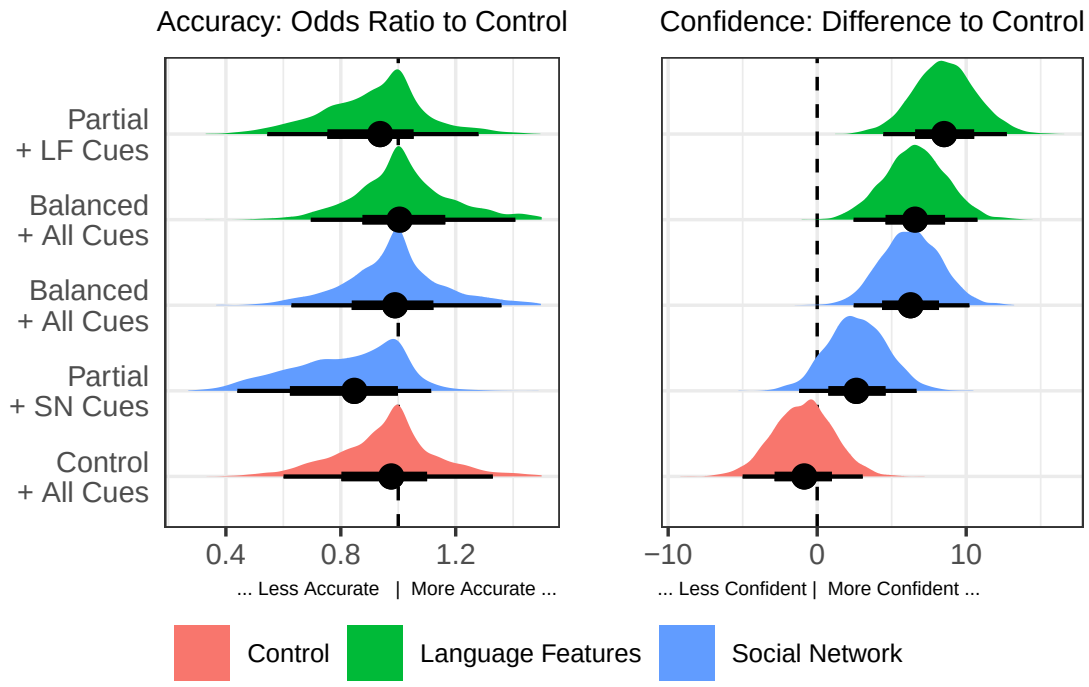


Figure 5.8: Posterior distributions of differences in means of user accuracy and confidence level. For both plots, the conditions are relative to the Control (no cues) treatment. CIs of differences are at 95% and 66%.

From Figure 5.8, we find that the treatments had a strong effect on user confidence but a smaller effect on accuracy. For instance, the Partial (LF cues) and the Balanced (SN) had effects larger than the 66% CI compared to the reference level, Control (no cues) treatment. This provides evidence that the visual anchors tend to produce higher user confidence levels. However, for accuracy, we find small effects of the

[8]We only included participant as a random effects for confidence, not accuracy, following a significant effect via ANOVA testing with frequentist mixed-effects modeling. See 04-regressions.Rmd | .html in the supplemental materials.

treatments as nearly all odds ratio CIs are within 1 (i.e., as likely as the reference level). The one exception is the Partial (SN Cues) treatment in which its 95% CI is nearly out of 1.

5.4.3 RQ2: Time Spent & Coverage Metrics

To evaluate the behavioral effect of visual anchors, we explore effect sizes using bootstrapped confidence intervals to identify differences in participants' time spent and coverage metrics, Figure 5.9 and 5.10). To consider **H1** and **H2**, we compare the visual anchored groups, Balanced and Partial, to the Control groups. First, we find that the visual anchored groups tended to spend more time on the Language View than the Control groups; however, time on the Network View was mixed as the Control Groups spent around 8-9 minutes on average, nearly the same as the anchored groups. This provides some evidence for our hypotheses, but only for Language View anchoring. The one anchored group that spent little time in the Language View was the Partial-Social Network treatment, where users averaged only 4 minutes (M = 4.01 minutes) as compared 6 to 7.5 minutes for the other anchored groups. This makes sense given these users' anchors only included the Social Network cues and videos, not the Language treatments.

Considering users' coverage metrics, we consistently find that visual anchored groups (except the Partial-Social Network treatment) had many more Language interactions (Green and Red sort) than the Control groups. However, Social Network interactions (i.e., hovers) are similar between the visual anchored groups and the controls. Both of these points lead to partial evidence for **H1** and **H2**. That is, we find that users can be visually anchored to the Language View, but not the Social Network View.

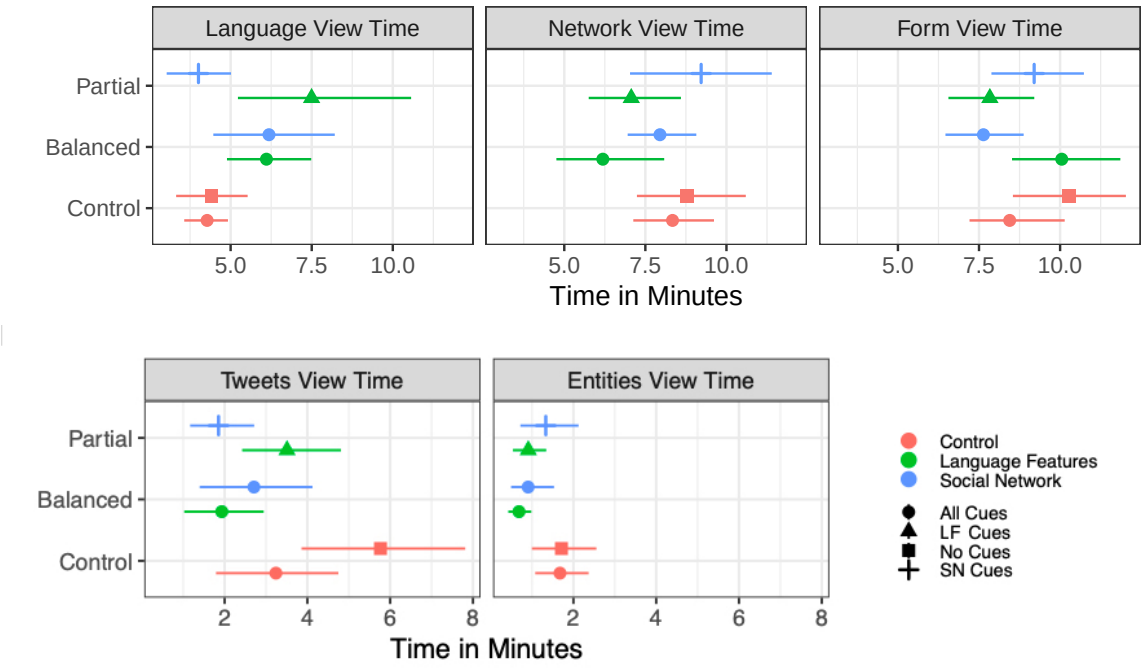


Figure 5.9: Time spent per view means and bootstrapped 95% confidence intervals on user-level (n = 94).

5.4.4 RQ2: Clustering Users based on Interactions

We find users' actions can provide indications of different interaction behaviors (Figure 5.11). For example, consider the 'Slow and Steady' cluster. In Figure 5.11, these users are mostly yellow, indicating a high rank across all metrics. These users were very active, exploring the entire interface's functionality for an extended period of time. On the other hand, the 'Fast and Quick' group is mostly dark blue as they ranked low in coverage metrics and time spent. The bottom two rows of the dendrogram provide the treatment conditions for each user. Comparing these rows to the clusters, we find some evidence for **H1** and **H2**. Take 'Anchored to Social Network' group as an example. Only one user who was treated with a LF visual anchor (dark blue) is within this cluster. As we would expect, many are SN groups (light red) that received the SN visual anchors. However, what is peculiar is the number of Control users (dark blue), particularly those without any strategy cues. Perhaps one interpretation is that these users are naturally drawn to the social network view

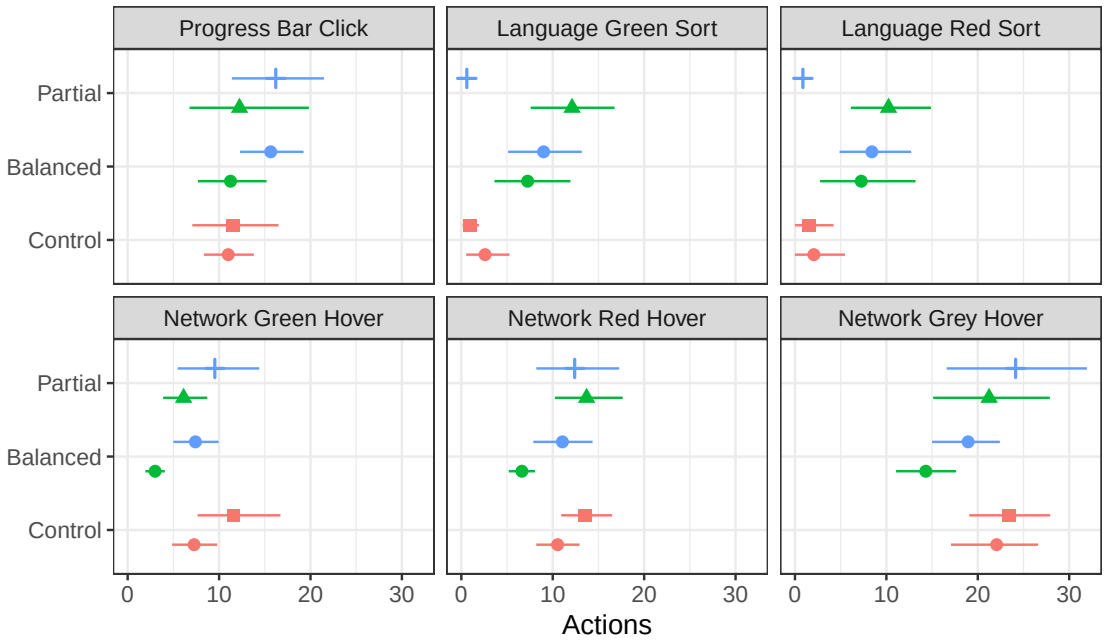


Figure 5.10: Coverage metrics means and bootstrapped 95% confidence intervals on user-level (n = 94). The figure uses the same shape and color encodings as Figure 9.

more than other views.

Descriptive statistics can also provide more context on each cluster. We find that the 'Slow and Steady' cluster users averaged much longer session times (M = 36.0 minutes). These users tended to have longer initial exploration periods, as they averaged nearly 10 minutes before their first decision submission. As context, other users typically made their first decision between 3 and 7 minutes. We also find that these users actively used the Progress Bar (M = 21.5 times), indicating a more organized strategy and using both primary views frequently. Interestingly, this cluster has, on average, the highest accuracy of 82.8%. Alternatively, we identified two clusters as users who focus more on either the SN (#1) or LF (#2). For example, cluster #1 spent 2.3x more time on the Social Network view than the Language Features view while the opposite holds for cluster #2.

Cluster validation: To validate the clusters, we compared them to post-questionnaire and decision data that was not included in the clustering process. For instance, we

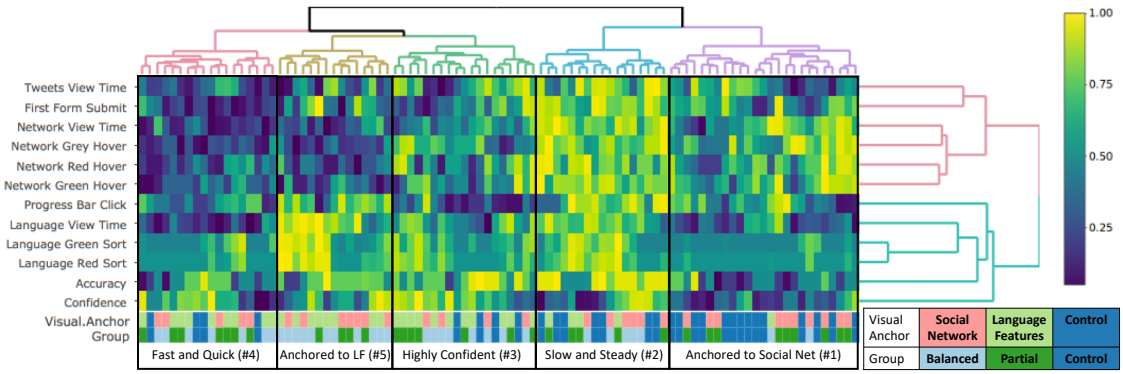


Figure 5.11: Heatmap clustering of interaction logs (Ward.D2 [2]) by columns (users) and rows (metrics). Each column is normalized for its percentile ranks. Users with a high feature rank are yellow while users with a low rank usage are dark blue. The bottom two rows indicate user's group and anchor condition. Both metrics were not used in clustering and provided for comparison.

find that the clusters provide a range of different ratings for the language features and social network functionality in the post-questionnaire. Users in the 'Anchored to Social Network' (#1), 'Highly Confident' (#3), and 'Fast and Quick' (#4) generally preferred the social network over the language features. However, the 'Anchored to Language Features' cluster (#5) was the only cluster to prefer, on average, the LF over SN. Alternatively, we find distinct differences in user motivation, interest, and challenge between clusters like 'Slow and Steady' (#2) and 'Fast and Quick' (#4). The 'Slow and Steady' cluster tended to be the most motivated, interested, and challenged out of all of the clusters. This makes sense given their longer session times and heavy usage. On the other hand, the 'Fast and Quick' cluster was the least motivated and interested. Perhaps lack of interest led to shorter session times and may factor in their lower accuracy.

Last, we explored the user-level interaction logs through scatter plots to validate our clusters. Figure 5.12 provides a scatter plots of fifteen user sessions. In each plot, a dot represents an action for each of the six views across session time (x-axis) and view (y-axis), with slight y-axis jittering to avoid overlapping actions. Each column includes three user sessions per cluster and chart row order represents, in descending

order, highly accurate to inaccurate users.[9]

We were able to identify general patterns and outliers from these plots. For example, the left-most column provides three users who are clustered to the 'Anchored to Social Network' group. These users tend to have many more actions in the Social Network view as compared to the Language Features, Tweet Panel, or Entities view. They seldomly use the Progress Bar (e.g., S104 and C1 use it somewhat while S108 never used the Progress Bar). Alternatively, we find examples in the 'Slow and Steady' group to have much longer user sessions, lasting well over thirty minutes (some even near forty minutes or more). These users tend to use a combination of all views like the Language Features, Social Network, and even the Tweet Panel views. Alternatively, we were able to identify outlier behaviors, like L103, who almost exclusively used the Language Features view. Even more interesting, the user waited until the end of the session to make all decisions.

Post-Questionnaire Feedback. We also evaluated open-ended feedback from users to assess user strategies. For instance, some participants identified a lack of trust in the language features because of a lack of clarity of their composition: "I did not like making a decision based on you saying whether the language measures were good or bad, I wanted to understand the language measures better." Others commented on the need for additional interface features, like a help menu, to aid in this intensive cognitive process: "it would be beneficial to have a 'help' section ON the platform to look at when needing the reminder of things the video mentioned." Some users commented on the usability of views in general, like the Entities and Tweet View. For example, one user commented "I didn't really understand the need of entities to determine fake articles." While another user admitted that "I did not use the tweets or entity features of the interface." Both comments explain users' limited use of that view but was expected given the limited training to functionality for these

[9]See 03-logs.html in the supplemental materials for all 94 users' plots.

views.

5.5 Discussion and Limitations

In this section, we discuss implications of our findings on VA evaluation practices as well as consider limitations of our study along with avenues of future work.

5.5.1 Implications for VA Evaluation Practices

Our findings are informative for guidance on training and tutorial during visualization evaluation with human subjects. Our findings show that visual anchors and strategy cues can significantly impact users' confidence and time spent investigating in each view when performing tasks. Anchoring to a subset of views may lead to the over-reliance on (often incomplete) information presented in those views, thus preventing users from getting a comprehensive picture.

Such anchoring effects could occur due to how participants are trained to use the visual interface before carry out the tasks. First, providing a general training video is a good idea, however, careful considerations are needed when devising a script or training video. The experimenter may want to make sure that all important features/views get equal coverage in the script and video.

Since our experiments show that visual anchors can indeed impact multiple performance metrics (confidence, accuracy, time to decision), we would like to raise awareness of participants possibly being unintentionally anchored and suggest careful consideration on how to train users to use a visual interface.

5.5.2 Limitations and Future Work

While we attempted to avoid negative impacts to validity, there are several limitations to generalizing our results.

First, there are limits to studying users' behavior through interactions. A different approach to tracking visual anchoring could be through eye tracking to detect users' attention directly rather than through interactions. However eye tracking too presents

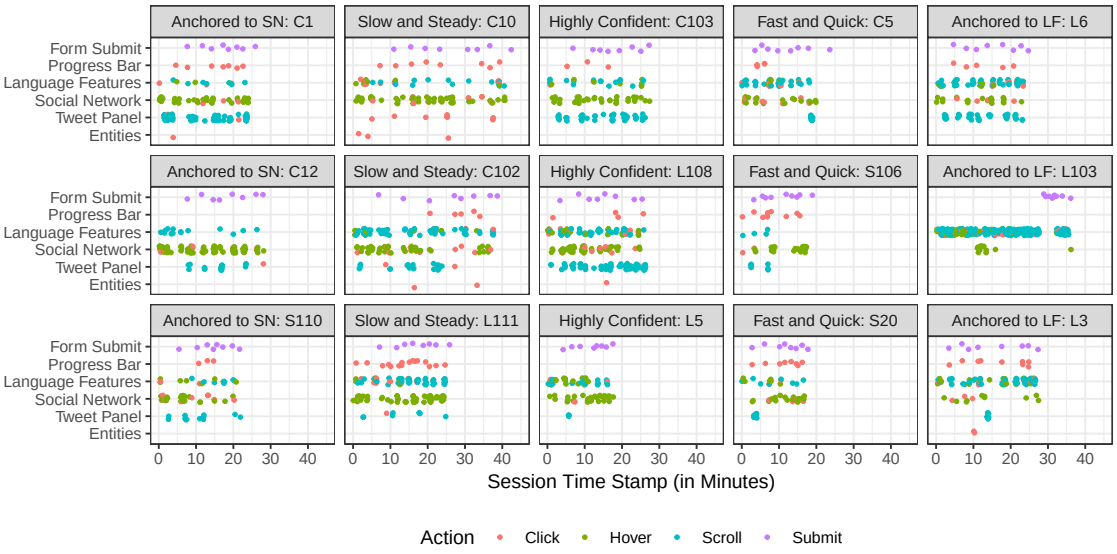


Figure 5.12: Experiment interaction logs of Verifi. Each plot is a user's interaction log. Each dot is a user action: click (red), hover (green), scroll (blue), and submit (purple). The x-axis is the time of the action. The y-axis is the respective view associated with that action. The order corresponds to critical functionality (e.g., Form Submit) to primary view (e.g., Language Features vs. Social Network) to secondary views (e.g., Tweet Panel or Entities). Chart columns indicate user-level strategies based on user-level dendrogram clustering. Chart row order represents, in descending order, highly accurate users (7+ out of 8, top row), average users (5-6 out of 8, middle row), and inaccurate users (4 or less of 8, bottom row).

challenges of its accuracy (especially for a multi-view interface). Future work in visual anchors should consider additional ways to analyze and measure user interactions.

Second, given the complex nature of the interface, we recruited highly trained students in computer science and data science who had some experience in visual analytics, machine learning, or social media communications. Consequently, our results may not generalize for a broader population (i.e., no experience in visual analytics). Future work could develop simpler interfaces that could be more appropriate for testing within broader participants pools like crowdsourcing (e.g., MTurk). Related, the choice of accounts can affect the difficulty of the decision-making task. If we selected different Twitter accounts, we may find our treatments have a different effects for our decision-making task.

Third, our study did not consider manipulating the interface design. While the training process differed between groups, all users received the same interface. As argued by Pu and Kay [52], design may have a significant effect on the forking paths problem as well. A future study could provide control interface layouts to identify the marginal value of each view in the decision-making process (e.g., testing whether the strategy cues with only the Tweet view – which mimics everyday social media usage – can measure a baseline accuracy). With such a baseline, a more precise estimate of the effect of the visualizations can be inferred. Another possible design enhancement could include adding uncertainty, like [5], encoding to the visualization (e.g., confidence intervals of each account based on past users' accuracy).

Last, cognitive science has developed Bayesian, rational computational models for understanding cognitive biases like numerical anchoring [194]. Such theoretical models–like Wu *et al.* [94]–can provide testable hypotheses that may aid future studies of cognitive biases in visual analytics. Two promising avenues to facilitate such cognitive modeling is through the incorporating prior knowledge [213, 107] and the addition of incentives and decision-theory within visualization tasks [5].

5.6 Conclusion

In this chapter, we presented an experiment on the role of visual anchoring in misinformation decision-making in a CMV VA system. We find that providing visual anchors and strategy cues can greatly affect users' confidence but have mixed results on users' speed and decision accuracy. Visual anchors can also play a role in secondary outcomes like users' view importance ratings and use of provided strategy cues. Last, exploration of user interaction logs can provide hints to users' strategies and the effects such treatments can have for certain users. While we find that some users are susceptible to such anchoring, others can ignore such treatments–perhaps due to uncertainty or a lack of trust–leading user attributes like motivation or interest can explain more of the users' knowledge seeking behaviors.

CHAPTER 6: A BAYESIAN COGNITION APPROACH FOR BELIEF UPDATING OF CORRELATION JUDGMENT THROUGH UNCERTAINTY VISUALIZATIONS

6.1 Introduction

Correlation judgement is an important topic and has been recently studied by the data visualization community [214, 215, 216, 217]. Understanding how people perceive correlations from data is necessary for the design of effective visualizations like scatterplots. Visualization researchers have investigated perceptual constraints on correlation judgment, including the use of Weber's Law [215, 217], a log-linear model augmented with censored regression and Bayesian methods [216], and other visual features [214]. While these empirical studies and models provide valuable insights and recommendations for correlation visualization design, they can be expanded to consider other factors that affect people's understanding of variable relationships.

One such factor is a user's prior beliefs when interpreting a correlation visualization. Previous studies often examine the perception of correlations between unnamed variables to avoid the effects of prior knowledge [218, 217] so that participants' beliefs about the variables do not influence their judgements. However, in practice, people rely on prior knowledge when interpreting and learning from correlation visualizations. As a result, it is important to investigate how prior beliefs affect the perception and interpretation of correlations. In addition to prior beliefs, another factor related to correlation judgement that warrants more research is uncertainty communication. Recently, visualization researchers have argued for the importance of uncertainty communication in information visualization [219]. Uncertainty communication techniques like hypothetical outcome plots (HOPs) [98, 99] provide methods

to visualize uncertain data for general audiences.

The experiments in this chapter build on previous research on correlation judgement by examining the impact of prior beliefs and uncertainty communication. We explore the following research questions: (1) how do prior beliefs impact one's correlation judgement? (2) how do people adjust their beliefs when the correlation visualization aligns or conflicts with their prior belief? (3) when uncertainty communication is incorporated in a correlation visualization, are users more or less likely to adjust their beliefs based on the conveyed relationship?

We also use Bayesian cognitive modeling [97] to quantitatively model how people interpret newly observed data in light of existing prior knowledge. Bayesian cognitive modeling offers a principled framework to understand how people interpret visualizations in light of prior beliefs [213] and how such beliefs should be updated with new information from a data visualization through Bayesian reasoning [109, 220, 95]. This provides a normative framework for evaluating the effects of visualization on beliefs, including the impact of uncertainty communication on users' interpretations of data [103, 5, 213] and the presence of biases that impair data-driven decision making [95].

Building such Bayesian cognitive models requires an accurate understanding of people's prior beliefs. Existing techniques for eliciting priors about correlations have a number of limitations, including a reliance on expert statistical knowledge related to correlation coefficients and their relationship to data [221, 222, 223]. In this chapter, we first evaluate a novel graphical elicitation method, "Line + Cone", for eliciting beliefs about the correlation between two variables through interactive data visualizations. With the proposed elicitation method, we conduct two experiments to study how people update beliefs about bivariate relationships when seeing correlation visualization with and without uncertainty representation.

This chapter bridges several areas of past work on correlation judgment, belief elicitation, and uncertainty visualization, while also drawing on recent methods for

modeling belief change using the framework of Bayesian inference. Specifically, this chapter's contributions are:

- Study 1: Introduce and validate the graphical "Line + Cone" method for eliciting prior beliefs about bivariate correlations, which is then used in the subsequent studies to measure belief change.
- Study 2: Compare differences in belief updating across correlation visualization with and without uncertainty communication.
- Study 3: Explore differences in users' belief update when the correlation visualization (with and without uncertainty communication) is congruent or incongruent with their prior beliefs.

Analysis of Study 1 showed that the "Line + Cone" belief elicitation method can be used to estimate peoples' mental representations of the correlation compared to a recent, more labor-intensive approach from cognitive science for measuring subjective belief distributions [224]. Study 2 revealed that participants updated their beliefs more effectively, and felt more confident, after observing visualizations with representations of uncertainty. In Study 3 we found evidence to support the hypothesis that people exhibit less belief change when seeing correlation visualizations that are incongruent with their prior beliefs. These results lay the groundwork for quantitative theories of how visualizations guide, and in some cases distort, how people learn about correlations through data visualization.

6.2 Background

6.2.1 Correlation perception and the effects of prior beliefs

A common task in visual analytics is assessing the relationship between two or more variables, often as a scatterplot [225]. In statistics, such relationships are typically quantified as correlations. However, statistics like Pearson correlation can be

misleading. For example, Anscombe's quartet [226] demonstrates that hidden patterns in the data are obscured by identical statistics. Even for expert data analysts, visual data inspection is an important part of the analysis process. Past psychology studies have considered how perceptual processing of scatterplots can affect an individual's understanding of correlations [227, 217, 218]. Building off that research, InfoVis researchers have identified scatterplots as an effective technique in discriminating correlations [228], testing correlation perception with Weber's law through additional techniques [215, 216], and identifying visual features in correlation perception [214]. However, these studies have not considered how prior beliefs affect individual's perception of variable relationships.

Research in psychology shows that prior beliefs have a strong influence on people's interpretation of uncertain data [73, 74, 75, 76], especially for correlations [77, 78]. A central theory that explains why prior beliefs are important is the dual-process account of reasoning [79, 18]. This theory posits that fast heuristic processes (System 1) competes with slower analytic processes (System 2) that can affect logical decisions. Evans *et al.* [80] suggested that belief bias [79, 73] could occur as "within-participant conflict" between the two systems when participants tend to agree with an argument based on whether or not they agree with the conclusion rather than its logical conclusion. Alternatively, other research focused on theory-motivated reasoning bias based on "congruent" and "incongruent" evidence relative to an individuals' belief systems [74]. These theories motivate design aspects in Study 2 and 3.

6.2.2 Uncertainty visualizations

Uncertainty visualizations are important as they enable better decision-making by conveying the possibility that a point estimate may vary [103]. More recently, research in InfoVis has provided innovative techniques like Hypothetical Outcome Plots (HOPs) [98, 99], frequency based representations [229, 5], visual semiotics [230], and design guidelines [108] for visualizing uncertainty. Alternatively, other visualiza-

tion researchers have studied important application aspects of uncertainty visualizations including hurricane prediction through ensemble modeling [231, 232], comparing users' prior beliefs congruence to social data [213], how uncertainty evaluation is prone to error [233], and its potential to improve one's ability to make predictions about replications of future experiments [234].

6.2.2.1 Eliciting correlation beliefs

Psychologists have used a variety of approaches to elicit beliefs about correlations. Initial research used two-step procedure to elicit participant's correlation belief [235, 77]: (1) determine relationship direction (positive or negative) and (2) rate the strength of the relationship. Later methods expanded on this approach by including Likert Scales, Spearmans's correlation, probability of concordance, and conditional quantile estimates [236, 221, 222, 237, 238]. However, there are several shortcomings with the previous approaches. Some methods only elicit beliefs about central tendency without capturing degree of uncertainty, while methods which do elicit uncertainty are labor-intensive [223]. Most methods rely on some background knowledge of statistics [221, 222], including how to interpret correlation coefficients, thus limiting their applicability to non-expert populations.

Cognitive scientists have developed a related technique for eliciting subjective belief distributions named Markov Chain Monte Carlo with People (MCMC-P; [224, 239]). Inspired by algorithms for MCMC estimation [237], MCMC-P as an approach to estimate a person's subjective belief distribution through sampling. In Study 1, we use MCMC-P as an elicitation benchmark to our proposed Line + Cone belief elicitation technique and outline this technique in Section 4.

6.2.3 Bayesian cognitive modeling in data visualizations

Cognitive modeling in visualization initially was studied as a subset of visuospatial reasoning in how individuals derive meaning from external visual representations

[92]. Visualization researchers have integrated similar ideas to understand visualization cognitive processes through insight-based approaches [45] and top-down modeling [93, 46]. More recently, InfoVis researchers have used Bayesian models to understand cognitive processing of visualizations [94, 95]. Cognitive scientists have demonstrated the importance of Bayesian modeling to understanding individual decision-making [96, 97]. In this approach, an individual has some prior belief that is updated when the individual consumes additional data, resulting in their posterior beliefs. Bayesian cognition models have been used to understand deviations from optimal belief updating due to conservatism, sample-based inference (approximation) and "resource-rational" interpretations of cognitive bias [60].

To our knowledge only two previous InfoVis studies [94, 95] have combined belief elicitation with a Bayesian cognitive modeling framework. Wu *et al.* [94] examined whether people integrated prior probabilities with data in an optimal manner. They found that priors influenced predictions in a manner consistent with Bayesian inference, although to a lesser extent than predicted by the model. However, a limitation to this study was that participants were given a prior; therefore, prior beliefs cannot be examined. In contrast, Kim *et al.* [95] empirically measured participants' prior beliefs about the a target proportional quantity and used those priors to calculate the normative posterior given the data that was presented. In aggregate, participants' judgments were consistent with predictions derived from Bayesian inference, although less so for large data sets. However, participants expressed greater uncertainty in their judgments than expected from the Bayesian model. Further, the authors connect such Bayesian modeling and belief elicitation with recent research on visualizing uncertainty through techniques like HOPs [98, 99]. Our work extends their framework but considering correlation beliefs rather than proportional values.

6.3 Research Questions and Analysis Methods

Our primary research question is the effect of providing uncertainty communications on users' belief updating in correlation visualization. In order to address this research question, we conducted a sequence of three experiments with latter ones building on the earlier studies.

A key to understanding users' belief update is the ability to accurately and intuitively capture such beliefs. **Study 1** evaluates the Line + Cone elicitation method relative to Markov Chain Monte Carlo with People (MCMC-P) [224], a belief elicitation method from cognitive science. After validating the Line + Cone method, we apply it in the next two experiments to address the main research question. In **Study 2**, we explore the effect of correlation visualizations with and without uncertainty representation on belief updating. Our primary hypothesis is that visualizations with uncertainty representation will overall lead to less belief updating about the correlation between two variables. Findings from Study 2 provides partial evidence to support the primary hypothesis. To expand on the findings, we are interested in further understanding users' belief update when the data visualization was deliberately manipulated based on users' prior beliefs. Therefore, **Study 3** extends Study 2's design but introduces a treatment that alters the data provided to participants to be either congruent or incongruent with their prior beliefs. We then evaluate the degree to which individuals update their beliefs when data provided either conflicts or aligns with their prior and whether the presence of uncertainty visualizations interact with that effect.

To analyze the results of Study 2 and 3, we employ mixed effects models to identify differences between treatments. The mixed effects models control for individual heterogeneity assumed between participants and the datasets (variable pairs) provided to participants. To explain the findings from the mixed effects models, we evaluate whether Bayesian cognitive models can be used to predict users' posterior beliefs

under different experiment treatment.

6.4 Study 1: Evaluating Line + Cone Elicitation

Our goal in Study 1 (see preregistration[1]) was to develop and validate the Line + Cone visual interface for eliciting prior beliefs about the correlation between two variables. In selecting our approach, we aimed to measure beliefs about both the *most likely* correlation between variables and the *degree of uncertainty*, without a need for statistics domain knowledge or numerical reasoning (see Section 2.2.1). We assessed the convergent validity of the Line + Cone method by comparing it to a higher resolution, but more labor-intensive, approach to eliciting subjective beliefs: Markov Chain Monte Carlo with People (MCMC-P; [224, 239]). MCMC-P resembles common sampling-based estimation algorithms such as Metropolis-Hastings in which a chain of states are sampled from an underlying probability distribution. In MCMC-P, state transitions are determined by asking participants to make forced-choice comparisons of the likelihood of possible values of the target parameter in many trials (usually in the range of 100 or more).

In our experiment we elicited prior beliefs about five sets of variables using both MCMC-P and the Line + Cone method. We created variable sets to cover a range of plausible correlationspossible divergent prior beliefs. For example, we expected that for the relationship *Weight x Price of diamonds* most participants would believe there is a strong positive correlation, while there may be less consensus about the relationship *Vaccination rate x Rate of illness*. Based on participants' responses we estimated the mean and confidence interval of their subjective prior belief (i.e., the relative likelihood of possible correlations between two variables). We then examined the degree to which the resulting prior means and CIs were correlated across the two methods.

[1] http://aspredicted.org/blind.php?x=zp7hr3

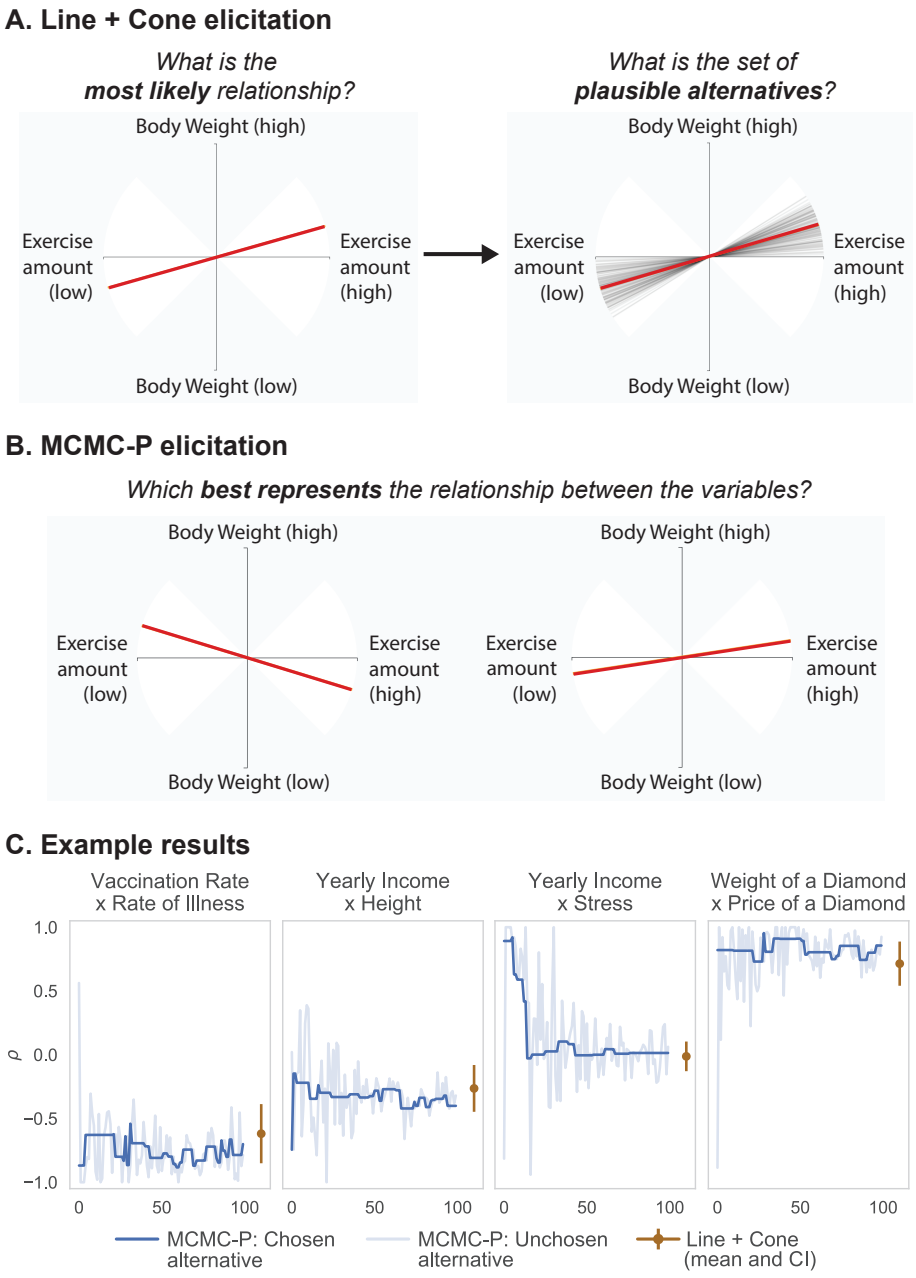


Figure 6.1: Elicitation methods in Study 1. **A:** For the Line + Cone elicitation, participants first recorded the belief about the most likely relationship between two variables (red line), then adjusted the set of plausible alternatives based on their uncertainty (gray lines). **B:** For the MCMC-P elicitation, participants responded to a series of two-alternative forced choices in which they judged which of two lines was more likely to represent the true relationship between the variables. **C:** Example comparison of elicitation results for a participant in Study 1. Dark blue lines indicate the chain of chosen alternatives from MCMC-P across 100 trials. Light blue lines indicate unchosen alternatives. The corresponding mean and CI from the Line + Cone elicitation is shown at the right of each plot.

6.4.1 Study Design

The experiment involved a within-subjects manipulation of elicitation method (Line + Cone vs. MCMC-P). Participants' beliefs were elicited for the same set of five variable sets (Table 6.1) using each method in a blocked presentation. The order of elicitation methods and variable sets within each block were randomized for each participant.

6.4.1.1 Line + cone elicitation

We designed a visual interface in which the mean and CI are directly elicited through the user's interaction. Each elicitation involves a two-step procedure (Figure 6.1A). First, the user selects the orientation of a red line according to their belief about the most likely relationship between the variables. Second, the user adjusts the width of the uncertainty cone. The uncertainty cone was depicted by gray lines which were draws from a Normal distribution centered on the most likely correlation (red line) and truncated at -1 and 1. Participants were instructed to adjust the cone such that the lines captured the range of "plausible alternatives" for the relationship between the variables.

6.4.1.2 MCMC-P elicitation

Markov Chain Monte Carlo with People (MCMC-P) is used to estimate subjective belief distributions based on a series of choices between two alternatives. In our task, each alternative represents a potential correlation between a pair of variables. For each variable set there were 100 choice trials. On each trial, the participant was shown two lines representing potential correlations (Figure 6.1B). Participants were instructed to select the alternative which was more likely to represent the true relationship. On the first choice trial the alternatives were two randomly selected correlations, one positive and one negative. In subsequent trials, the choice set included the alternative chosen on the previous trial and a *proposal* generated from a Normal distribution centered

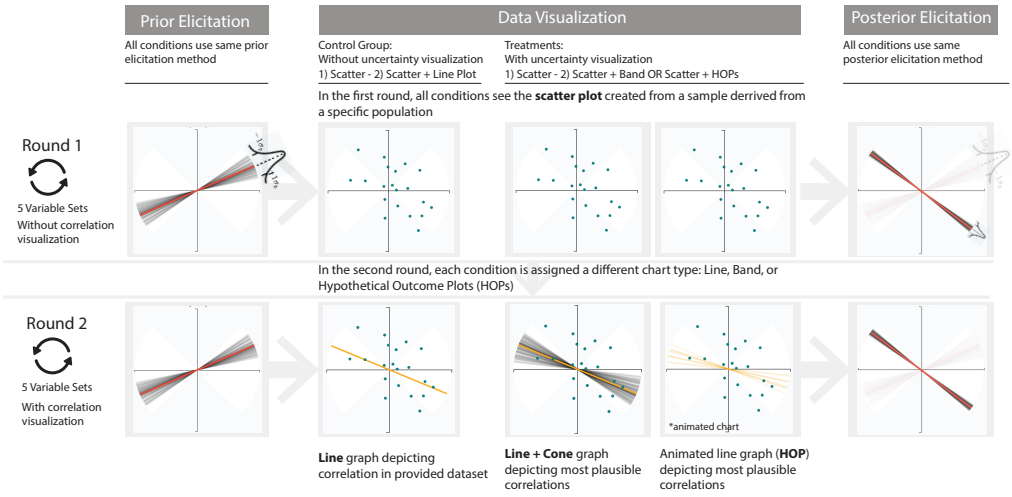


Figure 6.2: Study 2 Design. Each user goes through ten variable sets (five variables for two rounds) and elicit their belief before and after seeing data visualizations about each variable set. In Round 1, the user views five variable sets through only scatterplots. In Round 2, the user is randomly assigned to either Line, Cone, or HOP visualization treatments and views the remaining five variable sets.

on the previous choice. The width of the proposal distribution was adaptively tuned based on how often a participant accepted new proposals (see [240]). Each block resulted in a chain of alternatives that were chosen by the user (Figure 6.1C). The prior mean was calculated as the mean of the sampling chain, while the CI was the range between the 2.5% and 97.5% quantiles.

6.4.2 Participants

$N = 152$ participants were recruited from Amazon Mechanical Turk. Participants earned \$2.00 upon completion of the task, which took an average of 25.4 minutes ($SD = 12.2$). Per our pre-registration, we used several measures of task engagement to decide whether to exclude a participant. We excluded 55 participants who failed an attention check question and 36 participants who made nonsensical or incomplete responses to a set of open-ended questions regarding how they would respond to real-world situations. We also excluded 35 participants who met pre-specified exclusion criteria based on responses in the MCMC-P elicitation, including response streaks, response alternation, and response time. After accounting for all exclusions, $N = 92$

participants were included in the analysis.

6.4.3 Results and Discussion of Study 1

Table 6.1: Correlations between prior means and CIs elicited through Line + Cone and MCMC-P methods in Study 1.

	Prior mean		Prior CI	
Variable set	Pearson r	p-value	Pearson r	p-value
Weight x Price of diamonds	.26	**.012**	.34	**.001**
Exercise amount x Body weight	.37	< **.001**	.29	**.005**
Yearly income x Height	.12	.26	.27	**.010**
Yearly income x Stress	.45	< **.001**	.30	**.003**
Vaccination rate x Rate of illness	.40	< **.001**	.39	< **.001**

Our primary question was whether the belief distributions elicited with the Line + Cone method correlated with those generated using our MCMC-P procedure. We calculated Pearson correlations between the prior means and CIs for each variable set (Table 6.1). Elicited prior means were significantly correlated for 4 of the 5 variable sets, with the *Yearly income X Height* variable set the only exception. Prior CIs elicited from the two methods were significantly correlated in all 5 variable sets. These results suggest that our visual Line + Cone elicitation method is able to capture variation in beliefs about correlations across different variable sets, including beliefs about the most likely relationship as well as the degree of uncertainty, while being less labor-intensive than MCMC-P and requiring less statistics domain knowledge than existing elicitation methods.

6.5 Study 2: Belief updating with and without uncertainty representations

In the second study, we applied the Line + Cone elicitation method to examine belief change in the context of correlation visualization. We evaluated whether the type of visualization impacted the degree to which people updated their beliefs. Specifically, our **Study 2 Main Hypothesis** [2] was that correlation visualizations which include representations of the uncertainty in the true population correlation would

[2] http://aspredicted.org/blind.php?x=39yn5g

lead to less belief updating when people's prior beliefs were inconsistent with the presented data. This hypothesis is motivated by research on confirmation bias [72, 19] showing that people overweight evidence that is consistent with their prior beliefs. Uncertainty visualizations, by giving credence to a range of possible relationships (including less likely relationships that are more similar to a person's prior belief) may lead to less belief updating compared to visualizations that only represent the most likely a posteriori relationship. As a secondary hypothesis, we hypothesize that datasets with small and moderate correlations lead to less belief updating compared to datasets with stronger correlations.

6.5.1 Study Design

We employed a mixed design with a between-subjects manipulation of the visualization type (with and without uncertainty representation) and a within-subjects manipulation of the sample correlation of data presented to participants. In each trial participants reported their belief about the relationship between a set of variables, both before and after they experienced a data visualization All participants completed two rounds of five trials. In the first round the datasets were visualized as scatterplots to all participants (**Scatter** condition). In the second round the scatterplots were augmented with a visualization of the predicted population correlation based on the given dataset. Participants were randomly assigned to one of the following conditions (Fig 6.2):

- **Line**: A line representing the most likely population correlation was superimposed on the scatterplot [3]
- **Cone**: The line appeared with an uncertainty cone which represents the 95% confidence interval for the population correlation

[3]Note that the Line condition does not contain an uncertainty representation while the Cone and HOP conditions do.

- **HOP**: Hypothetical outcome plots (HOPs, [98, 99]) were used to present animated draws from the 95% confidence interval for the correlation

6.5.1.1 Datasets

We created two groups of five variable pairs that covered a range of population correlations between -0.9 to 0.9. We then generated 100 random samples for each variable pair based on the population correlation. The participants were told that the dataset is a *a sample of data collected from the real world.* [4] All points were re-centered with a mean of zero on each variable. All participants saw the same data points for each variable pair. The order of the variable pairs was randomized for each participant.

Note that population correlations were specified for each variable pair based on agreement among the authors (see examples in Figure 6.3). Our assumptions about the correlations of these variables may not reflect the ground truth relationship, and may differ from participants' beliefs. However, because we measure each individual's prior beliefs, we can assess whether belief updating was affected by any mismatch between their prior and the sample correlation.

6.5.1.2 Elicitation, attention check procedures, and collected data

Each trial consisted of a prior elicitation, correlation visualization, and posterior elicitation. For both elicitation steps we used the Line + Cone method validated in Study 1 (Figure 6.1A). Each elicitation resulted in three measurements: the most likely correlation (μ) and the lower and upper bounds of the uncertainty cone (b_{lower}, b_{upper}). All three values were bounded between $\rho = -1$ and $\rho = +1$.

We designed practice questions to familiarize participants with the Line + Cone elicitation. Participants answered test questions to ensure that they understood how to interpret the elicitation interface, including the direction of a correlation and the

[4] Due to random sampling, the sample correlations differed slightly from the specified population correlation.

degree of uncertainty captured with the cone. We also included attention check questions (same as in Study 1) to screen inattentive respondents or other invalid data [241].

In Study 2 and 3, we also collected basic demographic data, duration of each trial, and the error count of users in the instructions section.

6.5.2 Participants

Participants were recruited from Amazon Mechanical Turk. For all studies we required that participants were located in the U.S. and had a 95% or above approval rating. Participants earned $1.80 upon completion of the task, which took an average of 25.7 minutes ($SD = 14.6$) to complete. Per our pre-registration, we excluded any participants due to: failed attention check questions ($n = 35$); technical errors ($n = 15$); or task completion in less than 5 minutes ($n = 38$). This left $n = 212$ participants for the analysis (Line: 74; Cone: 64; HOP: 74).

6.5.3 Results

For the analysis we built three mixed effects models using R's `lme4` package for two linear regressions and R's `glmmTMB` for a beta regression. We used the normal approximation to calculate p-values of fixed effects using t-values produced by `lme4`.[5]

Dependent & Independent Variables: We considered three dependent variables (DV): (1) the absolute belief difference, (2) the difference in uncertainty, and (3) belief distance from the model's predicted posterior mean. For our independent variables (IV), we included the Visualization treatment (Line, Cone, HOP, and Scatter) and the absolute correlation of the generated data for the variable sets (see Figure 3)

Model Specification: For each model, we included the visualization treatment and the absolute correlation of the data as fixed effects. For the visualization treatment, the Scatter condition is the omitted reference condition. We treated the sample

[5]The code used is included in our supplemental materials.

correlation as a categorical variable and used zero absolute correlation as the omitted reference condition. We included the unique variable set and the participant id as random effects.

6.5.3.1 Beliefs about variable pairs

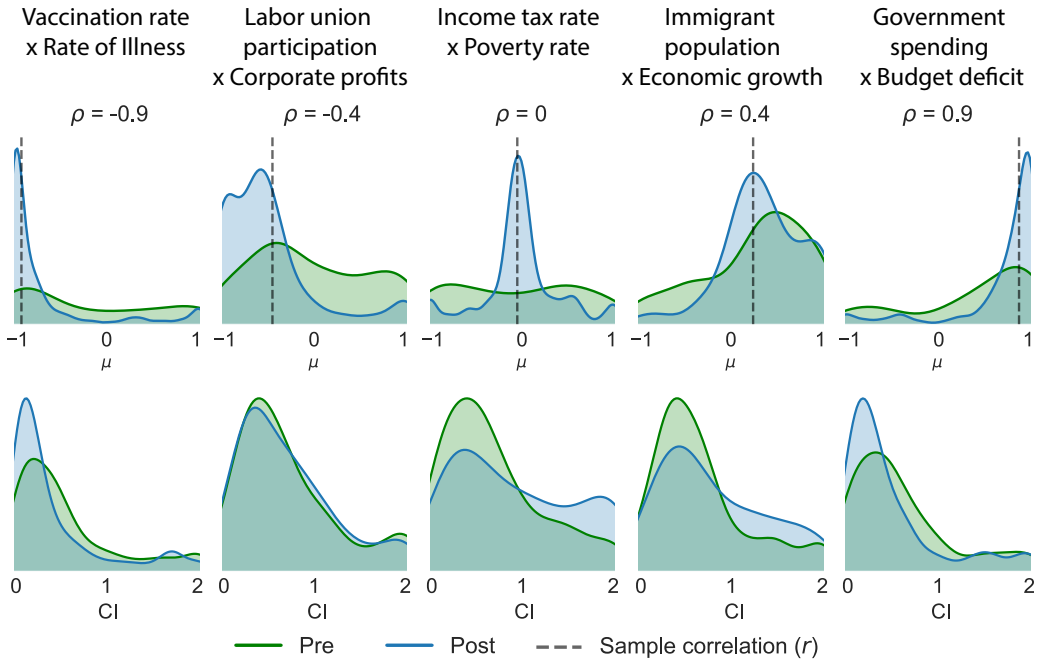


Figure 6.3: Density plots of means (top row) and CIs (bottom row) of elicited belief distributions for selected variable sets in Study 2. Dashed lines indicate the sample correlation of the dataset presented to participants.

We first examined participants' beliefs before and after experiencing the data visualization. Figure 6.3 displays pre- and post-treatment judgments about the most likely correlation (μ, top row) and uncertainty (CI, bottom row) for five of the ten variable pairs, aggregated across visualization treatments. With respect to the mean correlation μ, prior judgments (green density plots) were largely consistent with the relationship that was designated for each variable pair, such that the modal prior belief was close to the sample correlation. This suggests that the datasets presented were congruent with most participants' prior belief about the relationship between the variables. One notable exception was *Income tax rate X Poverty rate*, where

the designated correlation was $\rho = 0$ but prior beliefs were relatively uniformly distributed from -1 to +1. Post-treatment beliefs about the same variable sets (blue density plots) strongly shifted toward the sample correlation of the observed dataset (dashed lines) for all variable sets. The plots for the CIs reveal that the strength of the sample correlation also affected changes in uncertainty. CIs decreased after seeing strongly correlated datasets ($\rho = \pm 0.9$) but in some cases increased following data visualizations with weaker relationships. We report more detailed analysis of how the uncertainty changed in different treatments in section 5.3.3.

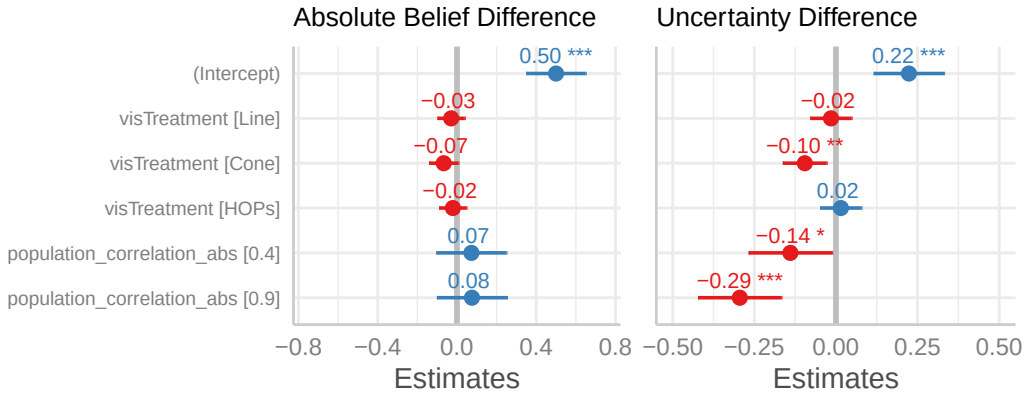


Figure 6.4: Study 2 fixed effects coefficients for absolute belief difference (left) and uncertainty difference (right). Error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: *** 99.9%, ** 99%, * 95%. For visTreatment, the reference category is the Scatter condition.

6.5.3.2 Change in beliefs about most likely relationship

We used linear mixed effects regression to model the effect of visualization conditions and population correlation on the absolute change in beliefs about the most likely correlation ($|\mu_{post} - \mu_{pre}|$. There were no significant effects (Figure 6.4, left), though the Cone condition showed marginally smaller changes in beliefs compared to the Scatter condition ($\beta = -0.07 \quad [-0.14, 0.01], \quad z = -1.779, \quad p = 0.075$). Thus, while participants clearly shifted their beliefs about the most likely correlation in response to observed datasets (Figure Figure 6.3), contrary to our expectations we

did not find that the degree of belief change differed by visualization treatment or population correlation.

6.5.3.3 Change in uncertainty

Mixed effects linear regression was used to model the effects of visualization condition and population correlation on the change in uncertainty ($|CI_{post} - CI_{pre}|$). As shown in Figure 6.4 right, relative to the Scatter condition, the Cone condition exhibited greater reduction in uncertainty ($\beta = -0.10\ [-0.16, -0.03],\ z = -2.782,\ p < .01$). In other words, participants assigned to the cone Condition felt less uncertain (more confident) with their input. There was no difference in the Line ($\beta = -0.02\ [-0.08, 0.05],\ z = -0.468,\ p = 0.640$) or HOP condition ($\beta = 0.02\ [-0.05, 0.08],\ z = 0.473,\ p = 0.636$). In addition, more extreme sample correlations had a greater impact on belief change: Compared to $\rho = 0.0$, there was a greater reduction in uncertainty for $\rho = .4$ (($\beta = -0.14\ [-0.27, -0.01],\ z = -2.138,\ p < .05$) and $\rho = .9$ ($\beta = -0.29\ [-0.42, -0.17],\ z = -4.513,\ p < .001$) variable sets.

6.5.3.4 Accuracy of posterior beliefs

We examined the accuracy of participants' posterior mean (μ_{post}) compared to the sample correlation of the observed datasets. In the Scatter condition, posterior means were biased to be more extreme for moderately positive and negative sample correlations. Relative to the $\rho = 0$ variable sets, absolute error was higher for $\rho = \pm 0.4$ variable sets ($\beta = .29\ [.19, .41], z = 5.46, p < .001$) but did not differ from $\rho = \pm 0.9$ variable sets ($\beta = .002\ [-.11, .11], z = .03, p = .96$) The remaining visualization conditions led to more accurate beliefs across the full range of sample correlations. Compared to the Scatter condition, the absolute error was lower in all three visualization conditions (Line: $\beta = -.22\ [-.34, -.10], z = -3.50, p < .001$; Cone: $\beta = -.34\ [-.47, -.21], z = -4.96, p < .001$; HOP: $\beta = -.25\ [-.38, -.13], z = -3.97, p < .001$).

6.5.4 Bayesian belief updating model

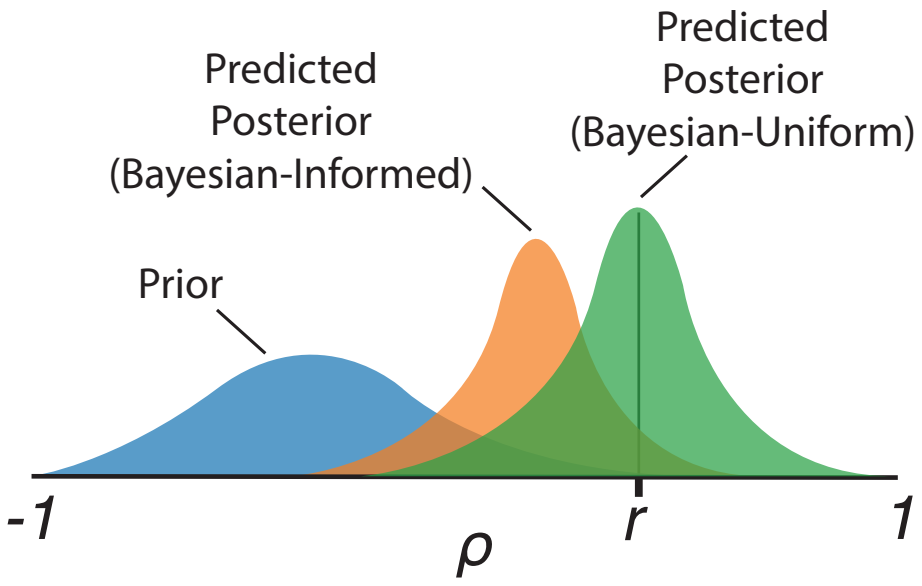


Figure 6.5: The Bayesian cognitive models predict how beliefs should change based on the sample correlation and the participants' prior beliefs. For a dataset with sample correlation r, the Bayesian-Informed model predicts the posterior distribution integrates the new evidence with the person's prior belief according to Bayes rule. The Bayesian-Uniform model assumes a uniform prior over possible correlations, predicting that the posterior mean will be at r.

In this section we use Bayesian cognitive modeling to investigate the influence of prior beliefs on the belief updating process. Under the principles of Bayesian inference, people should integrate new evidence about a correlation with their prior beliefs about that relationship. Bayesian models provide a normative benchmark for how beliefs *should* change depending on the strength of the evidence and participants' uncertainty. For instance, a person who is confident that variables are negatively correlated may only shift their beliefs a small amount after seeing a dataset with a positive sample correlation. A second person who is highly uncertain about the relationship, however, may be more strongly influenced by the same data and report posterior beliefs that are closely matched to the sample correlation. This framework also allows us to identify when people systematically fail to adjust their beliefs as predicted by the Bayesian model. Returning to the main hypothesis of Study 2, if

uncertainty representations cause smaller adjustments to beliefs, this will correspond to larger divergence between participants' elicited posterior beliefs and the predictions of the Bayesian model compared to other conditions.

Having elicited prior beliefs about each set of variables, we examined whether participants' posterior beliefs (following the data visualization) could be predicted by a normative Bayesian model. The model uses Bayesian inference to predict a posterior belief distributions over possible population correlations, ρ, based on an observed dataset and a particular prior (see [242] for similar model formulation).

We evaluated two variants of the model that differed only in their prior. The **Bayesian-Informed** model relied on the participant's elicited prior to calculate the normative posterior distribution after observing a dataset. The prior belief was modeled as a bounded Normal distribution, $\rho \sim BoundedNormal(\mu_{pre}, \sigma_{pre}, [-1, 1])$, where μ_{pre} and σ_{pre} are the mean and standard deviation of the participant's elicited prior. The observed bivariate data X was modeled as having been generated from a standardized multivariate Normal distribution with mean of zero and standard deviation of 1 on each dimension (see [242]),

$$X \sim MultivariateNormal(\begin{bmatrix} 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^{-1}). \tag{6.1}$$

Under the **Bayesian-Uniform** model, the prior was a uniform distribution over the correlation coefficient, $\rho \sim Uniform(-1, 1)$. The mean and 95% CI of the posterior distribution for this model is equivalent to the values used for the visualizations in the Line, Cone, and HOP treatments. Predicted posterior distributions for ρ were estimated for both models using MCMC with the PyMC3 library [243] with two chains of 20,000 samples and 1000 burn-in iterations. Lastly, we compared the elicited posteriors to the elicited priors, absent any belief updating. We refer to this baseline as the **Prior-only** model in the results below. Figure 6.5 outlines the difference

between each model as hypothetical density plots.

The relative fit of the models reflects the weight of prior beliefs in the updating process, with the Bayesian-Informed model representing the normative integration of priors with new evidence. If people relied only on the visualization without accounting for prior beliefs, their elicited posteriors should be best fit by the Bayesian-Uniform model. In contrast, if they did not adjust beliefs upon observing a dataset, the Prior-only model should provide the closest match to posterior beliefs.

6.5.4.1 Model comparison

Following [95] we evaluated each model's performance with two metrics: mean absolute error (MAE) between the predicted and elicited posterior means; and Kullback-Liebler distance (KLD) (Figure 6.6A). These measures are complementary in that MAE captures the magnitude of differences in beliefs independently of the amount of uncertainty, while KLD measures correspondence across the entire belief distributions. We used mixed effects linear regression to compare MAE and KLD with model and visualization type as fixed effects and random effects for participants and variable sets.

In terms of MAE there were significant effects of visualization treatment ($\chi^2(1,3) = 33.17$, $p < .001$) and model ($\chi^2(1,2) = 333.97$, $p < .001$), but no interaction ($\chi^2(1,6) = 7.55$, $p = .27$). Pairwise comparisons indicated that MAE was lower under both the Bayesian-Informed and Bayesian-Uniform models than the Prior-only model in all four visualization treatments (all $p < .001$). The Bayesian-Uniform model achieved lower MAE than the Bayesian-Informed model in the Scatter ($z = -4.52$, $p < .001$), Cone ($z = -2.8$, $p = .01$), and HOP ($z = -2.48$, $p = .04$) conditions, but the two did not differ in the Line condition ($z = -2.22$, $p = .07$). Comparing the best-fitting Bayesian-Uniform model across visualization treatments showed that MAE was higher in the Scatter group than the Cone ($z = 3.14$, $p = .01$) and Line groups ($z = 2.94$, $p = .02$), but not significantly different from the HOP group

($z = 2.54$, $p = .05$).

For KLD there were significant effects of visualization treatment ($\chi^2(1, 3) = 67.57$, $p < .001$), model ($\chi^2(1, 2) = 31.64$, $p < .001$), and model $\times$ treatment interaction ($\chi^2(1, 6) = 32.49$, $p < .001$). In the Scatter condition, KLD of the Prior-only model was lower than the Bayesian-Informed model ($z = -3.32$, $p = .003$), but did not differ from the Bayesian-Uniform conditions ($z = -.87$, $p = .66$). This indicates that the Bayesian model was relatively unsuccessful at predicting the posterior distribution in the Scatter condition, failing to outperform the baseline Prior-only model. In the remaining conditions (Line, Cone, HOP), the Bayesian-Uniform model had lower KLD than both the Prior-only and Bayesian-Informed models (all $p < .022$). As was the case for MAE, the KLD of the Bayesian-Uniform model was higher in the Scatter condition than the other conditions (all $p < .001$), but did not differ among the Line, Cone, and HOP groups. This supports the earlier finding that the accuracy of posterior beliefs was poorer in the Scatter condition compared to the other treatments.

The predictions of the three models diverge most when there is a discrepancy between participants' priors and the sample correlation of the observed dataset. We therefore explored how the fit of each model depended on the absolute distance between the prior mean and the sample correlation (Figure 6.6B). At small distances the three models have comparable MAE and KLD, while the advantage for the Bayesian-Uniform model grows with increasing distance between the prior and sample correlation. The poorer fit of the Bayesian-Informed model indicates that participants discounted their priors when they observed a dataset with a drastically different correlation. Notably, at small distances KLD was lowest for the Prior-only model. This result suggests that when people observed a dataset that was consistent with their prior, they were less likely to update their beliefs as predicted by either Bayesian model.

6.5.5 Discussion of Study 2

Results of the regression analysis and cognitive modeling showed that visualizations with representations of the population correlation (Line, Cone, and HOPs) led to greater accuracy in posterior beliefs compared to the Scatter condition. In addition, higher correlations led to larger reductions in uncertainty, potentially because stronger relationships are easier to detect in scatterplots [217, 218] and are associated with less uncertainty in the population correlation. We found initial evidence for this updating process using the Bayesian cognitive model, showing that when the sample correlation presented to participants was far from their prior mean, they strongly adjusted their beliefs to reflect the pattern in the data (Figure 6.6B).

We did not find support for our main hypothesis that uncertainty visualizations would be associated with smaller changes in beliefs. On the contrary, the Cone visualization (with a cone of "plausible alternatives" representing uncertainty in the correlation based on the data) led to greater reductions in uncertainty. This result suggests that the explicit representation of uncertainty provided by the Cone visualization leads to greater confidence about the true relationship compared to the other visualization types. Interestingly, we did not find a similar effect on uncertainty change in the HOP condition, possibly due to the transient nature of the animated uncertainty cone.

There were two shortcomings of the present study that may have limited our ability to detect differences in belief updating between conditions. First, participant's prior beliefs largely aligned with the sample correlation, leading to many cases with little room for participants to adjust their beliefs. Second, the relatively large sample size of the datasets ($n = 100$) meant there was relatively little uncertainty about the population correlation. This may explain why the Bayesian-Uniform model provided the best fit to elicited posteriors, such that the sample correlation had a stronger influence than individuals' priors. **Study 3** was designed to further explore how these

factors affect belief change. We manipulated the data provided to be either congruent or incongruent with the user's elicited prior belief. In addition, we manipulated the amount of data uncertainty by varying the sample size.

6.6 Study 3: How correlation congruence and uncertainty affect belief updating

The hypothesis of Study 2 was that people would exhibit less belief change when they experienced visualizations with representations of uncertainty. The main hypothesis for Study 3[6] extends this further to predict that viewers of uncertainty representations would exhibit smaller changes in beliefs when correlation visualizations are *incongruent* with users' prior belief and when the dataset has a smaller sample size.

6.6.1 Study Design

For Study 3, we extended the design of Study 2 by explicitly manipulating the congruence of the sample correlation (**factor 1**) with a user's prior belief and the amount of uncertainty (**factor 2**). Both above factors are within-subjects while the visualization treatment remains a between-subject factor. Figure 6.7 summarizes the design of Study 3. For each variable pair, participants saw datasets that were either congruent or incongruent to their prior beliefs:

- **Congruent** datasets: Random samples were drawn from a multivariate normal distribution with correlation *0.25* away from the prior mean. For example, if a participant's prior mean was 0.85, the data was sampled from a distribution with population correlation of 0.6 $(0.85 - 0.25)$. In this condition a user always saw sample correlations with the same sign as their prior belief.

- **Incongruent** datasets: Random samples from a multivariate normal distribution with correlation value that is *1.0* away from the prior mean. For example, if the prior mean was 0.6, the data was sampled from a distribution with a

[6] http://aspredicted.org/blind.php?x=x7ph2u

population correlation of *-0.4*. In this condition, participants saw datasets with the opposite correlation sign from their prior belief.

We also manipulated the number of samples in the datasets for specific variable pairs (10 points vs. 100 points). Datasets with 10 points result in greater uncertainty as measured by the 95% confidence interval. As in Study 2, participants were randomly assigned to visualization conditions of Line, Line + Cone and HOPs. Given Study 2's results that users achieved better accuracy with all three visualization types, we omitted the Scatter condition.

6.6.1.1 Datasets, elicitation, and attention check procedures

For Study 3 we selected variable pairs from the results of a pilot study. With 50 pilot participants, we elicited prior belief and uncertainty about 30 variable pair candidates, then categorized variables into a 2 X 2 grid of high/low social consensus on correlation and uncertainty. [7] With lessons learned on users' beliefs about the variable pairs from Study 2 (section 5.3.1), we aimed to select pairs that cover a range of distributions of beliefs about the mean correlation and uncertainty. We selected four variables with either high / low correlation consensus and high / low uncertainty. Study 3 used the same elicitation process, instructions, and attention checks as Study 2.

6.6.2 Participants

Participants were recruited from Amazon Mechanical Turk. Participants earned $1.80 upon completion of the task, which took an average of 22.9 minutes (SD = 12.28) to complete. Per our pre-registration, we excluded any participants who: failed attention check questions ($n = 12$); technical errors ($n = 95$); or completed the entire task in less than 5 minutes ($n = 11$). This left $n = 267$ participants for the analysis (Line: 89; Cone: 92; HOP: 86).

[7] Social consensus was measured as the standard deviation of prior means, while average uncertainty was measured as the mean CI.

6.6.3 Results

Dependent & Independent Variables: Similar to Study 2, we considered three dependent variables: (1) the absolute belief difference, (2) the difference in uncertainty, and (3) the user's belief distance from the model's predicted posterior. For our independent variables (IV), we created two features based on our variable conditions from Figure 6.7. First, we defined **pre-belief distance** as the distance between users' prior elicitation and the correlation of the provided sample, which is larger when a participant is provided incongruent datasets. Next, we defined **sample uncertainty** as the size of uncertainty shown to users resulting from the sample size. In doing so, we used continuous IVs ranging from zero to two rather than binary variables. For reference, we provide kernel density plots (Figure 6.8 and 6.9) for the two IV's partitioned by its respective condition categories.

Model Specification: We employed three mixed effects models as in Study 2 (see Section 5.3). For each model, we included the interaction terms between the visualization treatment, the pre-belief distance, and the sample uncertainty as fixed effects. For the visualization treatment, the Line condition is the omitted reference condition.

6.6.3.1 Change in belief about most likely relationship

For absolute belief difference (Figure 6.10, left), we found the largest effect to be pre-belief distance ($\beta = 0.73[0.64, 0.81], z = 2.40, p < .001$), indicating that users updated their beliefs more when they viewed incongruent datasets.

There were significant interactions between pre-belief distance and visualization type, such that there were smaller belief changes when the data was incongruent in both the Cone ($\beta = -0.12 \quad [-0.19, -0.05], \quad z = -3.3, \quad p < .001$) and HOPs ($\beta = -0.11 \quad [-0.18, -0.04], \quad z = -3.16, \quad p < .01$) conditions relative to the Line condition. This finding is in line with our hypothesis that in the incongruent condi-

tion, users would show smaller update in their belief when uncertainty representations are present.

Finally, while the HOP condition led to slightly larger changes compared to the Line condition ($\beta = 0.10 \quad [0.02, 0.17], \quad z = 2.400, \ p < 0.05$), this condition had a negative interaction with sample uncertainty such that beliefs shifted less after seeing smaller datasets ($\beta = -0.07 \quad [-0.13, -0.01], \ z = 2.181, \quad p < .05$). We did not find corresponding effects for the Cone condition. This difference between the Cone and HOP visualizations might suggest that uncertainty is more evident in larger uncertainty amounts when using the HOP technique. This is potentially due to the lack of a fixed representation of most likely correlation in the HOP technique as opposed to the Cone technique.

6.6.3.2 Uncertainty change

In our regression of the uncertainty difference (Figure 6.10, right), we found that users in the Cone condition exhibited more reduction in uncertainty than the Line condition ($\beta = -0.19 \quad [-0.27, -0.10], \quad z = -4.166, \quad p < .001$), replicating the effect seen in Study 2. There was not a significant effect in the HOPs condition ($\beta = -0.07 \ [-0.16, 0.02], \ z = -1.514, \ p = 0.130$).

Pre-belief distance had no effect on the uncertainty difference in any condition. However, sample uncertainty had a positive effect on changes in uncertainty ($\beta = 0.13 \ [0.02, 0.23], \quad z = 2.391, \quad p < .05$). We also found that the Cone visualization condition had larger effects on the uncertainty difference when interacting with sample uncertainty ($\beta = 0.20 \ [0.12, 0.28], \ z = 4.784, \ p < .001$). The HOPs condition also showed a positive interaction with uncertainty difference when interacting with datasets with larger sample uncertainty ($\beta = 0.13 \ [0.05, 0.21], \ z = 3.019, \ p < .01$). These findings suggest that participants in the Cone condition showed more overall reduction in posterior uncertainty compared to the Line treatment but the HOP condition did not show similar effects. Interestingly, when dealing with larger uncertainty

(10 data points), the presence of an uncertainty representation resulted in an increase in users' uncertainty. This finding suggests that both visualization techniques convey uncertainty when uncertainty amounts are larger, but users' experience of the HOP condition is similar to the Line condition when dealing with datasets with smaller uncertainty. Perhaps this is due to users' inability to perceive small angular movements of the line.

6.6.3.3 Accuracy of posterior beliefs

We used beta regression to model the effects on the distance of users' posterior beliefs from the true sample correlation. We found that pre-belief distance had the largest positive effect on users' post-belief distance ($\beta = .34 \quad [0.16, 0.52] \quad z = 17.254, \quad p < .01$). In other words, posterior beliefs were less similar to the sample correlation when the dataset was incongruent with users' prior beliefs. We also found that compared to the Line condition, the HOP condition had a positive effect on posterior distance when viewing a dataset with more uncertainty ($\beta = 0.29 \quad [0.14, 0.45], \quad z = 2.181, \quad p < .01$). This might be due to the lack of a fixed most-likely correlation representation in the HOPs condition, therefore when sample uncertainty is larger, users are more prone to larger distances (errors) in their judgements.

6.6.4 Bayesian belief updating model

We used the models from Study 2 to examine how prior beliefs influenced belief updating in Study 3. In general, the best fit to elicited posteriors in terms of both MAE and KLD was achieved by the Bayesian-Uniform model in all conditions (Figure 6.11). Incongruent trials provide a strong comparison of the Bayesian-Informed and Bayesian-Uniform models because they involve datasets that conflict with participants' prior beliefs. If people integrate new evidence with their elicited prior, they should show smaller shifts in beliefs in Incongruent trials than expected under

the Bayesian-Uniform model. However, as was seen in Study 2, posterior distributions were best-fit by the Bayesian-Uniform model, suggesting a stronger influence of the data visualization on posterior beliefs. Notably, the only condition in which the two models performed comparably on Incongruent trials was the Cone treatment, where there were no differences in MAE ($t(722.63) = 1.33$, $p = 0.18$) or KLD ($t(731.98) = 1.78$, $p = 0.08$), indicating that Cone visualizations produced belief updates that more closely aligned with the normative prediction of the Bayesian-Informed model.

6.6.5 Discussion of Study 3

We predicted that people exposed to uncertainty visualizations (Cone and HOP conditions) would exhibit less belief change compared to those without uncertainty (Scatter and Line conditions). We found strong support for this hypothesis in Study 3 when participants saw data that was incongruent with their prior beliefs. Both the Cone and HOP treatments were associated with smaller belief updates compared to the Line condition which did not represent uncertainty about the correlation. Uncertainty visualizations also affected whether there were shifts in participants' degree of uncertainty. Relative to the Line condition, Cone visualizations led to greater reductions in uncertainty for large datasets, whereas uncertainty did not change when datasets were small. Similar (albeit weaker) effects were present for HOP visualizations.

Finally, we replicated the modeling results from Study 2, showing that posterior beliefs were best-fit by the predictions of the Bayesian-Uniform model. Although this does not imply that participants completely disregarded their prior beliefs, it indicates that the data visualizations tended to have a stronger influence on posterior beliefs than expected from a normative Bayesian perspective. The Cone visualization was the only condition in which the Bayesian-Informed model performed comparably to the Bayesian-Uniform model. This result suggests an alternative interpretation of the

smaller degree of belief updating in that condition when faced with incongruent data. Rather than representing an irrational failure to modify beliefs akin to confirmation bias, the Cone condition may be most effective for striking the appropriate balance between new data and prior beliefs.

6.7 Discussion, Future Work, and Conclusion

In this chapter, we study the effect of prior belief and uncertainty representations on correlation judgement. In Study 1 we developed the Line + Cone method for eliciting people's beliefs about the correlation between two variables, including their degree of uncertainty. The Line + Cone method serves as a good choice for eliciting users' beliefs about bivariate relationships for future studies of correlation judgement. In addition to capturing users' beliefs about the correlation means (commonly done in previous correlation judgement studies), results from all three studies demonstrate that it is also important to capture users' uncertainties about their judgements. In Studies 2 and 3, we used the Line + Cone method to investigate belief updating in the context of data visualization. We found that visualization conditions with uncertainty communication led to less belief updating compared to visualizations without uncertainty, especially when the presented correlation visualization is incongruent with users' prior beliefs. An important conclusion is that judgements are affected by the existence of uncertainty depictions. How we encode uncertainty (e.g., Cone vs. HOPs), also affects users' belief and uncertainty change. As the visualization community pays more attention to the importance of uncertainty representations and elicitation, it is important to be cognizant to the affects of such techniques on users' judgements.

In our studies we applied a Bayesian cognition framework to understand how people update their beliefs about bivariate correlations with different types of visualizations. Recent studies have applied insights from Bayesian cognitive modeling to understand how people integrate new data with their existing knowledge [97, 60]. The Bayesian

framework provides normative benchmarks that can be used to evaluate whether people optimally revise their beliefs given their existing uncertainty and the strength of new evidence conveyed through a visualization [95]. We used Bayesian models to compare participants' posteriors to three benchmarks: no change in beliefs (Prior-only model); the normative posterior when taking into account the elicited prior (Bayesian-Informed model); and the normative posterior when disregarding the prior (Bayesian-Uniform model). In both Studies 2 and 3, elicited posterior distributions were best-described by the Bayesian-Uniform model, suggesting that the characteristics of the visualized dataset had a stronger influence on posterior beliefs than expected under the Bayesian-Informed model.

There are several possible explanations for why posterior beliefs appeared to underweight participants' priors. One possibility is that people have a different interpretation of the cone representation which is used to elicit their uncertainty. In order to minimize demands on numerical or probabilistic reasoning, participants were simply instructed to adjust the cone to capture the range of "plausible alternatives" for the correlation between the variables. In Study 1 we found support for the claim that this method captures participants' uncertainty, but there may nevertheless be a mismatch between the elicited distribution and participants' subjective beliefs such that people are more uncertain than indicated by their elicited priors.

We found other evidence that people updated beliefs in a way consistent with Bayesian inference. In Study 2, users reduced their uncertainty to a greater extent for more extreme sample correlations. In Study 3, uncertainty increased when people saw small datasets ($n = 10$) compared to large datasets ($n = 100$), even in the Line condition which lacked an explicit representation of the correlation uncertainty. Participants also expressed greater uncertainty in the posterior beliefs than predicted by the Bayesian models, echoing the findings of Kim *et al.* [95].

These studies provide the groundwork for investigating how people interpret data

that is relevant to strongly-held or favored beliefs. Prior beliefs can distort the perception of new evidence, as is seen in widespread evidence of confirmation bias [72, 19, 244]. Using intuitive, visual belief elicitation methods in conjunction with Bayesian cognitive models offer a promising path toward understanding the causes of such biases in data visualization.

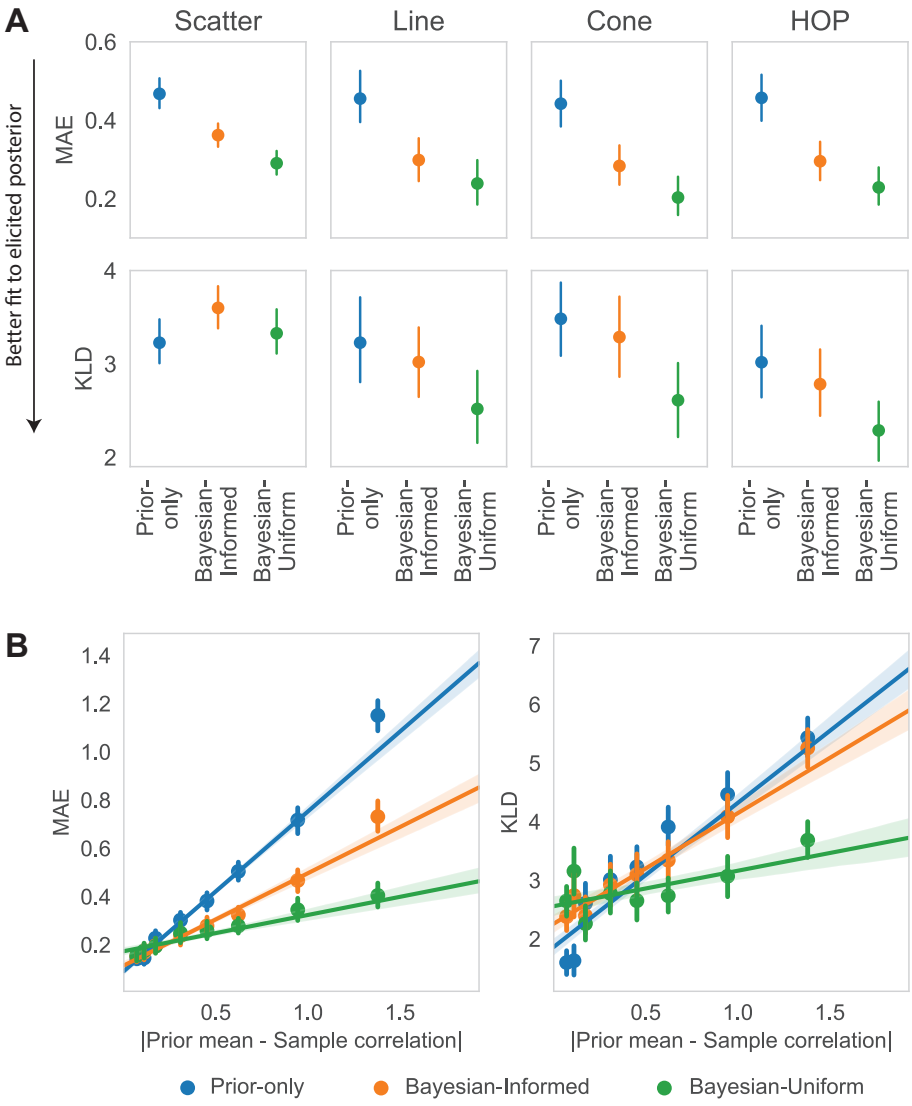


Figure 6.6: **A:** MAE and KLD between elicited posterior and predictions of Prior-only, Bayesian-Informed, and Bayesian-Uniform models. **B:** Model performance as a function of the absolute distance between the elicited prior mean and the sample correlation.

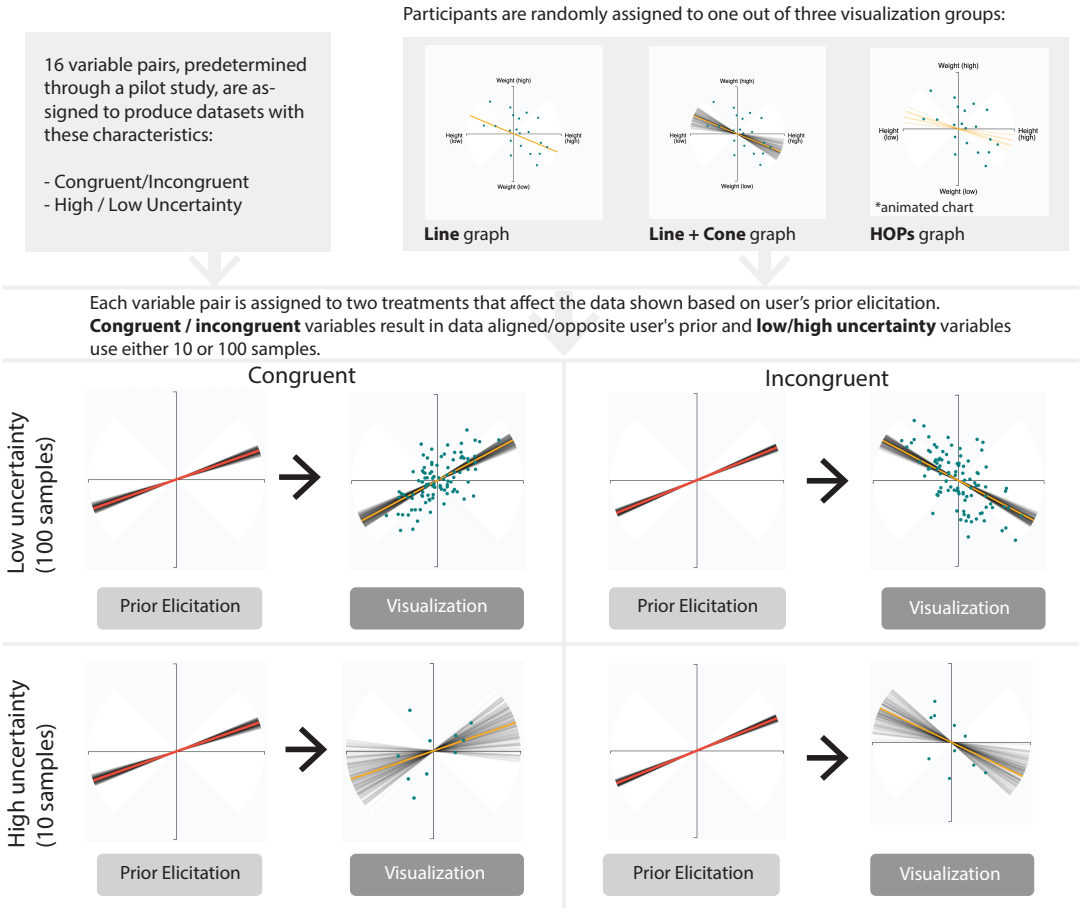


Figure 6.7: Study 3 design. Like Study 2, users elicit their beliefs about correlations of variable pairs before and after seeing data visualizations. Users are randomly assigned to Line, Cone, and HOP visualization treatments. The datasets are generated based on users' prior elicitation as either congruent/incongruent and 10 or 100 data points.

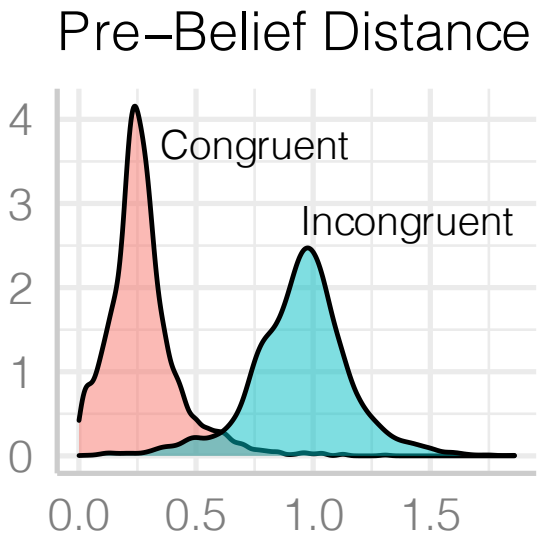


Figure 6.8: Kernel density plots for pre-belief distance and sample uncertainty values by congruent or incongruent conditions (pre-belief distance).

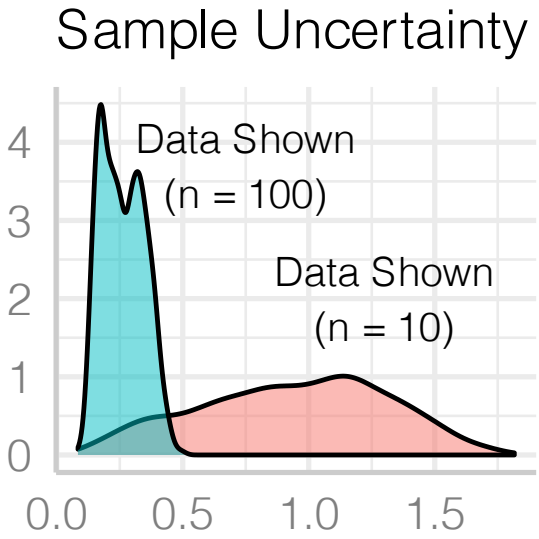


Figure 6.9: Kernel density plots for pre-belief distance and sample uncertainty values by data shown ($n = 100$ or $n = 10$).

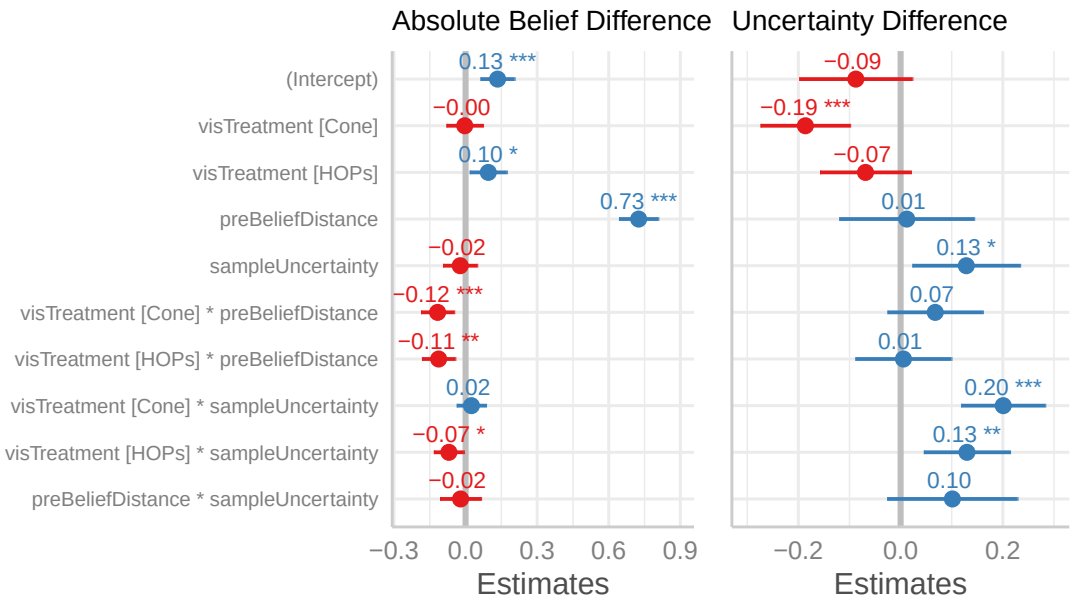


Figure 6.10: Study 3 fixed effects coefficients from analyzing absolute belief difference (left) and uncertainty difference (right). The error bars indicate 95% confidence intervals. Asterisks indicate statistical significance than zero using p-values: *** 99.9%, ** 99%, * 95%. For visTreatment, the reference category is the Line condition.

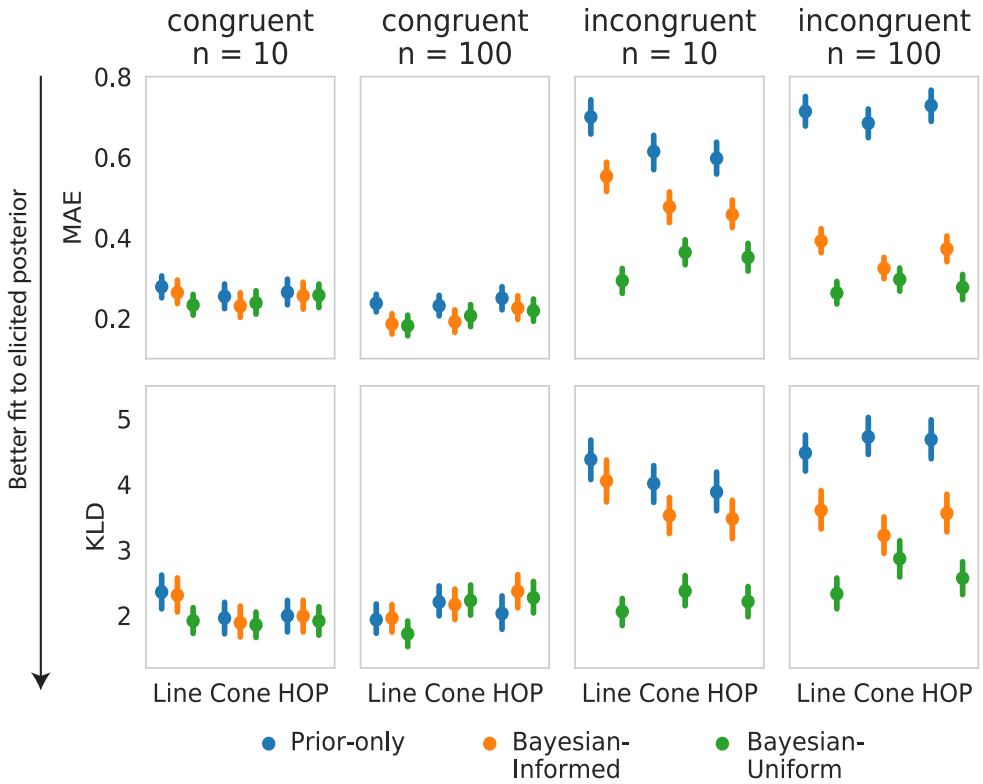


Figure 6.11: MAE and KLD by model for Study 3.

CHAPTER 7: EFFECT OF UNCERTAINTY VISUALIZATIONS ON MYOPIC LOSS AVERSION AND THE EQUITY PREMIUM PUZZLE IN RETIREMENT INVESTMENT DECISIONS

7.1 Introduction

As companies move towards choice-based retirement investment accounts like IRAs and 401(k) plans, Americans now have $20 trillion dollars in such accounts indicating the colossal importance of smart retirement investing.[1] A recent report by the National Institute on Retirement Security highlighted that the shift from pensions to 401(k) plans has pushed more retirement risk onto individual workers [245]. At the same time, individual investors are increasing their adoption of investing on digital platforms (e.g., online or mobile retirement account tracking). The combination of long-term retirement planning and access to short-term trends on digital investment platforms increases the importance of human-computer interaction in the design choices for web-based and mobile applications by financial companies.

Financial decision-making is central to retirement investing as typical investors make decisions about how to allocate funds across a wide range of assets that vary in risk (e.g., stocks vs. bonds). A seminal study by Mehra and Prescott [4] found a surprising reluctance to take on risk, in that standard economic models could not account for the large historical premium for riskier investments (the "equity premium puzzle"). Benartzi and Thaler [21] theorized that individuals deviate from the predictions of neoclassical economic theory due to two factors, an oversensitivity to the possibility of losses, and evaluation of returns over short time periods, a combination they referred to as *myopic loss aversion*. Benartzi and Thaler [3] showed that myopic

[1] `https://www.statista.com/statistics/940498/assets-retirement-plans-by-type-usa/`

loss aversion emerges when making investment decisions with simple visualizations of the distribution of returns (i.e., bar charts). They found that investors allocated less in stocks when shown returns over a 1-year evaluation period due to aversion to short-term losses. Their results suggest that the method for visualizing investment performance can have a dramatic effect on individuals' willingness to take on risk.

Recently, the field of information visualization has proposed multiple visualization techniques for including uncertainty representations of data. Different visual encodings can change how users perceive and interpret uncertainty, and in turn, the decisions they make. Our study aims to measure the effect that different uncertainty visualizations/representations have on myopic loss aversion in long term (retirement) financial decision-making. The experiment is motivated by the results of Benartzi and Thaler [3]; we first try to replicate their findings and further expand it by testing the effect of a range of new uncertainty visualizations (e.g., quantile dotplot [246, 5, 31] and hypothetical outcome plot [98, 99, 100]) on myopic loss aversion.

More specifically, our study seeks to address the following research questions: **RQ1:** Do crowdsourced investors exhibit myopic loss aversion when presented investment returns that are aggregated over a range of evaluation periods (1, 5, 10, 15, 20, 25, 30 year)? RQ1 aims to first determine if we can replicate the findings of the Bernatzi and Thaler [3] study (1 year vs. 30 year). The visual encoding employed in evaluating RQ1 is a sorted bar chart as in [3]. RQ1 extends [3] by evaluating the impact of intermediate evaluation periods on myopic loss aversion, providing a more fine-grained understanding of how decisions change alongside visual representations of risk. **RQ2**: Do different uncertainty representations affect myopic loss aversion in retirement asset allocation? We included six different uncertainty representations, one bar chart visualization, and one tabular representation for RQ2. In addition to evaluating the effect of different visual encodings, RQ2 also explores how evaluation periods and visual encodings may interact.

To address these research questions, we present findings from a two-round crowdsourced online experiment depicted in Figure 1. Round 1 is to see if we can replicate the findings of Benartzi and Thaler [3] on myopic loss aversion in an online setting. In round 2 participants were assigned to one of the eight different treatments with different uncertainty visualizations. Consistent with Bernatzi and Thaler [3], with bar charts we find evidence of myopic loss aversion as participants opted for much less stock allocation for 1 year than 30 year evaluation period. Similarly, we observe a positive monotonic relationship between evaluation period and stock allocation (and expected return). Interestingly, we found that the type of uncertainty representation significantly affects participants' stock allocation and their expected returns. Simpler and more intuitive uncertainty visualizations (with extrinsic annotation [229]) led to higher stock allocation that were closer to optimal. In contrast, uncertainty representations that draw attention to risk/volatility (e.g. HOPs) yielded lower stock allocations and tended to increase the equity premium which could amount to hundred of thousands of dollars less at retirement for an average investor. We did not observe significant interactions between the visual encoding and evaluation period, except for the density plot. Our qualitative analysis of user comments also sheds light on their visual reasoning strategies.

Situated in the task of maximizing returns for long-term retirement planning—a common, high-stakes example of repeated decisions under uncertainty—our results demonstrate that common uncertainty visualization with simple uncertainty representations lead to more optimal investment allocation and expected returns. Grounded in economic theories including prospect theory, myopic loss aversion, and the equity premium puzzle, these findings shed light on the important role of uncertainty visualization in financial decision making including retirement planning. Our work connects behavioral economic theories with information visualization and highlights the need for more research in the visualization community in order to provide tangible

recommendations on the use of uncertainty visualization to make better investment decisions for lay audiences.

7.2 Background Work

7.2.1 Economic Theory in Long Term Investing

Three tenets from modern economic theory provide the motivation for myopic loss aversion [21, 3]: Samuelson's gamble [247], lifetime portfolio selection [248, 249], and the equity premium puzzle [4]. During lunch one day, the noted American economist Paul Samuelson offered MIT colleagues a gamble: If the colleague guessed a coin flip correctly they would win \$200, but would lose \$100 if incorrect. This gamble has a positive expected value of $.5 \times \$200 + .5 \times -\$100 = \$50$. One colleague responded that he wouldn't accept the bet once but would accept the bet 100 times. Soon after Samuelson wrote a mathematical proof [247] showing that such a preference was irrational and demonstrated loss aversion. For example, the faculty member turned down the bet because "I would feel the \$100 loss more than the \$200 gain" [21]. It seems that for some individuals the fear of a loss outweighed the expected likelihood of a gain.

Another important tenet from economic theory of investments include theorems by Merton (1969) [248] and Samuelson (1975) [249] that under certain assumptions like random walk of asset prices [250] and constant relative risk aversion utility functions, "asset allocation should be independent of the time horizon of the investor" [3]. This counter-intuitive notion predicts that a 35 year old and a 64 year old should choose the same allocation for retirement investing. This idea is directly connected to the differentiation between the evaluation and investment (horizon) periods, which we discuss in Figure 7.1.

However, one major issue to economic theory impeded modern financial theory: the equity premium puzzle. Mehra and Prescott [4] studied the implications of economic theory for the difference between relatively risk-free government-backed bonds (e.g.,

US treasury bill) and relatively more risky stocks. They found a surprising "equity premium" in the form of excess returns for taking on the relatively higher risk of equity investing. Comparing such implications to historical returns of government bonds and stocks, Mehra and Prescott [4] found that to account for the average 6% equity premium, such a framework would imply an extremely risk-averse and implausible representative investor [251]. Such challenges open up new behavioral explanations for the equity premium puzzle: myopic loss aversion [21].

7.2.2 Investing Decisions in Behavioral Economics

Neoclassical economic theory is grounded on the assumption of a perfectly rational decision maker [252, 253]. However, in recent decades the field of behavioral economics challenged this assumption on three fronts: unbounded rationality, unbounded willpower, and unbounded selfishness [254]. Coined by Herbert Simon [62], bounded rationality refers to the acknowledgment that individual decisions are fundamentally constrained by factors like scale, time, and cognitive ability. This work led to the discovery of heuristics, or mental shortcuts, which people use to find solutions for decisions otherwise thought to be intractable. Tversky and Kahneman [26] argued that such heuristics can lead to systematic errors in decision-making, or cognitive biases. They further extended such work with prospect theory, a descriptive theory of decision-making under uncertainty [255]. A main assumption of prospect theory is that utility is derived not from objective metrics of value, but instead as gains and losses relative to some reference point. Prospect theory captures people's inherent risk aversion or "loss aversion", the notion that individuals are affected more by a loss than a gain of the same magnitude. A key implication is that changes in the *framing* of a choice in terms of gains or losses—despite no differences in the underlying economic problem—can exert a strong influence on decisions [256].

Benartzi and Thaler [21] proposed that a narrow framing of investment decisions focused on short-term outcomes rather than long-term (aggregated) returns, in com-

bination with loss aversion, provides a solution to the equity premium puzzle. Evaluating returns over a short time window highlights the possibility of losses and leads to more risk aversion, even in the context of investment decisions with long time horizons. Using prospect theory, Benartzi and Thaler [21] found that the equity premium observed by Mehra and Prescott was consistent with an evaluation period of 1 year.

As Benartzi and Thaler noted, while loss aversion might be considered a matter of individual preference (or a "fact of life"), the evaluation period is often dictated by the environment or "choice architecture" [257]. Accordingly, subsequent work has shown that myopic loss aversion is reduced when investors are provided aggregated returns over longer evaluation periods [3, 258] or when they have less frequent opportunities to change their allocations [259]. These effects exemplify a common theme in behavioral economics: Subtle features of the choice architecture can "nudge" people toward behaviors that are consistent with a policy objective [257], such as maximizing retirement savings through the use of default contributions [260]. For instance, in a recent study using a simulated retirement investment task, Camilleri *et al.* [261] found that presenting dynamic risk information, such that the evaluation period was aligned with the time left until retirement, encouraged reliance on a "smart default" plan which invested more in riskier (higher growth) investments early on. These results underscore the importance of interaction design in investment platforms. Providing frequent updates on short-term changes in fund performance and making it easy to reallocate may increase engagement, but these same factors likely amplify the harmful effects of myopic loss aversion on long-term returns [262].

7.2.3 Visualization in financial decisions and uncertainty visualization

Past research in data visualization and human-computer interaction has examined the role of data visualizations in improving financial decision-making [263, 264, 265, 266, 267]. Relevant for our study, Gunaratne and Nov developed a user interface for retirement decision-making based on endowment effect and loss aversion. They

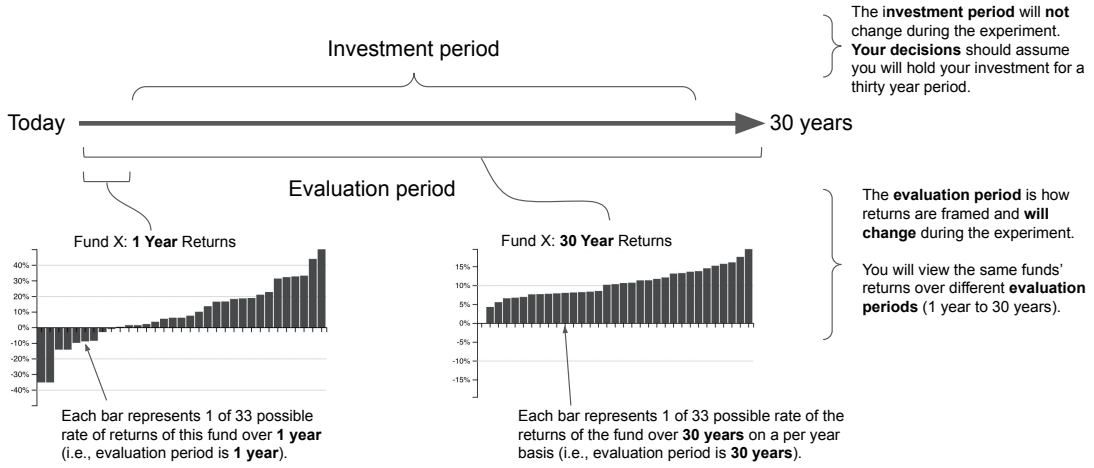


Figure 7.1: Evaluation period versus investment period. We provided this figure to participants to distinguish between these critical concepts. Participants were instructed and incentivized to invest over a 30 year investment period for all decisions. However, following [3] we manipulate the evaluation period of the returns shown to participants (e.g., 1 year to 30 years). Benartzi and Thaler [3] argue that economic theory predicts rational investors would not differ between these two decisions but that myopic loss aversion could explain why the decisions are not consistent.

showed that highlighting the long-term implications of users' decisions could lead to adjustment of their investment allocations and reduction of loss aversion. [268]. Other research has considered the role of portfolio allocation or financial literacy. Rudolph *et al.* designed a financial planning study using a simple visual analytics system (FinVis) to aid in portfolio allocation [264]. Using a simple table as a control group, they found university students made more optimal allocation decisions using FinVis as compared to the control (table of returns and standard deviation). Lusardi *et al.* provides visual analytics tools for financial literacy [269]. Additional research has also considered the intersection of visualization and financial decisions in relation to cognitive biases [270] and risk premium [271]. But one gap of this research on financial decision-making in visualization research has been the connection with research on visualizing uncertainty in which different techniques for encoding uncertainty such as Hypothetical Outcome Plots [98] or Quantile Dot Plots [5] are shown to have different effects on users' belief-updating and decision outcomes [246, 5, 99, 100, 34]. Kay *et al.* noted that uncertainty can be intrinsic or extrinsic to the representation [229];

we found this categorization informative when interpreting our experiment results. Overall, evaluating uncertainty visualizations are fraught with many challenges [233]. Alternatively, HCI research on the gamification of uncertainty decisions has shown that too much uncertainty information can lead to unnecessary risk-taking [272]. For a comprehensive survey on uncertainty visualizations, we recommend Padilla, Hullman, and Kay [31]. Our experiment design is informed by research on uncertainty visualization and modern behavioral economic theories.

7.3 Research questions and hypotheses

The core research question of this work follows Benartzi and Thaler [3]: "How do investors think about investment decisions over long horizons, and how do their choices depend on the way in which risk and return data are presented?" From this question, we derive two research questions.

RQ1: Replicating Benartzi and Thaler (1999), do crowdsourced investors exhibit myopic loss aversion when presented with a 1 year versus a 30 year evaluation period? More broadly, what is the effect of different evaluation periods on myopic loss aversion?

Design and Hypothesis: MTurk participants who own at least one financial asset participate in a within-subject asset allocation decision between stock and bond returns (both names masked) over seven evaluation periods (i.e. 1, 5, 10, 15, 20, 25, 30 year) for a fixed 30 year investment period. Myopic loss aversion [21, 3] predicts that individuals are more risk averse for shorter evaluation periods (e.g., 1 year) than longer periods, resulting in a higher asset allocation to less risky assets (e.g., bonds) than is optimal for long planning horizons like retirement.

RQ2: Does visualization with uncertainty representation affect myopic loss aversion (i.e., retirement asset allocation) and do uncertainty representations interact with evaluation periods?

Design and Hypothesis: Consistent with past designs for uncertainty visual-

izations [229, 5, 31], we design a mixed experiment with between-subjects (different uncertainty visualization) and repeated measures within-subjects (seven evaluation periods). We expect that visualizations with intrinsic uncertainty representations like frequency framing (e.g., dotplot) and animated plots (e.g., hypothetical outcome plots) will result in better returns and less myopic loss aversion.

7.4 Methods

We designed a pre-registered[2] experiment to test how uncertainty visualizations impact allocation decisions for simulated long term (30 year) retirement investments. Following Benartzi and Thaler [3], we frame individuals' decisions as dividing an investment between two assets. The names of the assets are masked but they correspond to standard benchmarks for bonds (10 year United States Treasury) and stocks (S&P 500).[3] We compared the bar chart visualization from Benartzi and Thaler [3] with a range of alternative uncertainty representations.

7.4.1 Investment task and experiment design

The experiment included two rounds. The first round was a within-subjects manipulation of evaluation period using the bar chart visualization from [273] (see example in Figure 7.2). Because we aimed to replicate [3], the first two decisions were based on 1 year or 30 year evaluation periods presented in random order. The remaining decisions involved five different evaluation periods in a random order (5, 10, 15, 20, and 25 years). In the second round, each user was randomly assigned to one of eight uncertainty visualization conditions. The presentation order followed the same scheme as in Round 1 (1 and 30 years in the first two trials, followed by the remaining evaluation periods). Participants were not provided immediate feedback about their decisions (i.e., simulated returns based on their allocations) in order to avoid any learning effects.

[2] `https://aspredicted.org/blind.php?x=sz8j4b`

[3] Data is from Aswath Damodaran and available at `http://pages.stern.nyu.edu/~adamodar/New_Home_Page/datafile/histretSP.html`.

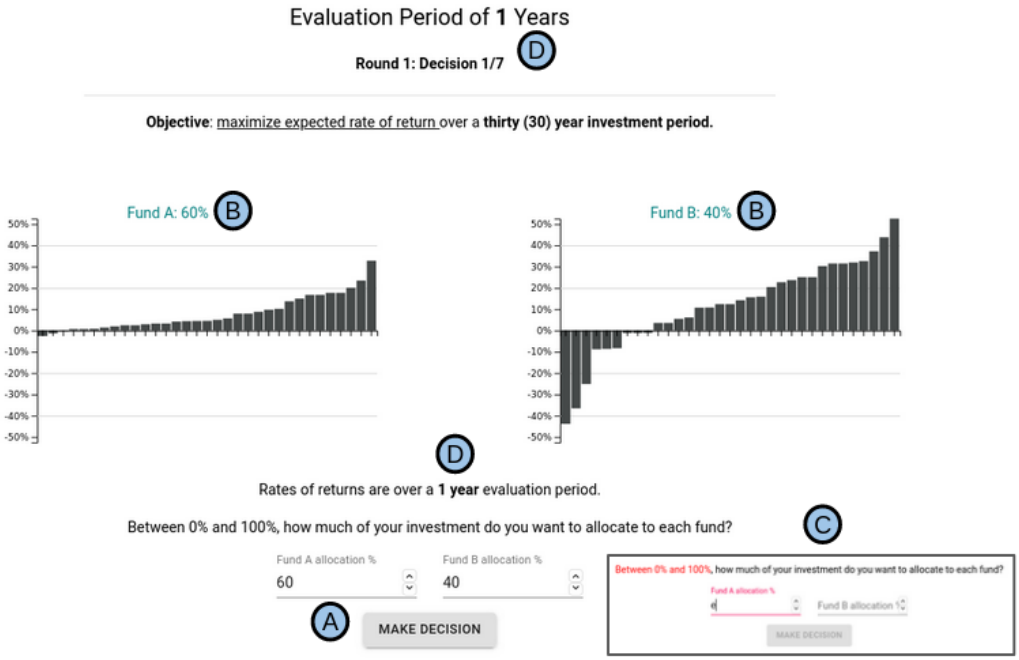


Figure 7.2: A depiction of the experiment interface. This example shows the round 1 (bar chart) and 1 year evaluation period decision. The user inputs their allocation (A) for each evaluation period (D) that updates chart titles (B) and input is controlled for invalid responses (C).

We designed a react.js custom web interface that used D3 for visualizations, node.js for the server, mongoDB for the database and the app was deployed on heroku.[4] Figure 7.2 provides a screenshot of the investment task interface for an evaluation period of 1 year. The charts show the distribution of simulated returns for each fund over the evaluation period. The left-right position of the two assets (bonds vs. stocks) was randomized on each trial. The participant inputs an allocation in either of two boxes (A) that are reactive (i.e., sum of the two boxes always equal 100%). As the user enters valid value (0 to 100 in 1 increments), the "Make Decision" button becomes active and the user can proceed to the next allocation decision. To ensure understanding of their decision, each chart's title updates real time based on participant's decision (B). If the user enters invalid input (e.g., "e"), "Between 0% and 100%" is highlighted in red and "Make Decision" becomes inactive, preventing the

[4]The experiment is available at `https://retirement-study-1.herokuapp.com/` and the interface code will be released publicly on github.

user from moving to the next decision until a valid response is provided. For each decision, the interface updates the evaluation period and emphasize it in two places (D).

There are three intentional deviations from the original experiment by Benartzi and Thaler [3]. First, we provide incentives to align to performance. This is critical to ensure participants have a vested interest in performing this task. This change is especially important given the use of crowd sourced workers. Second, for round 1, we use the bar chart design from [273] which lays out each fund in its own bar chart horizontally rather than the original design which was one grouped bar chart with each fund being a different bar. We did this for design preferences. Third, the original study used 34 bars for the 1 year return and 50 bars for the 30 year return. For simplicity, we used 33 draws (e.g., bars, dots) for easy mental computation that each draw is approximately 3% of the data. We also controlled 33 draws across visualizations and evaluation periods for consistency.

7.4.1.1 Simulating expected returns

Following [3], we showed participants data generated from a historical simulation using bootstrapped sampling with replacement. To calculate average and annualized returns by evaluation period N, we used the geometric mean for each sample i:

$$\text{Geometric Mean} = \left(\prod_{i=1}^{N} (1 + Returns_i) \right)^{\frac{1}{n}} - 1 \tag{7.1}$$

Participants decided the allocation between stocks and bonds for a simulated 30 year investment for retirement. Incentives were established by running a 10,000 bootstrap with replacement sample returns with a 30 year investment period. We then calculated what the expected return would be for 101 different possible stock allocations (0 to 100 in integer increments), with the remainder allocated to bonds. When participants made a decision, we determined their incentive by randomly selecting one

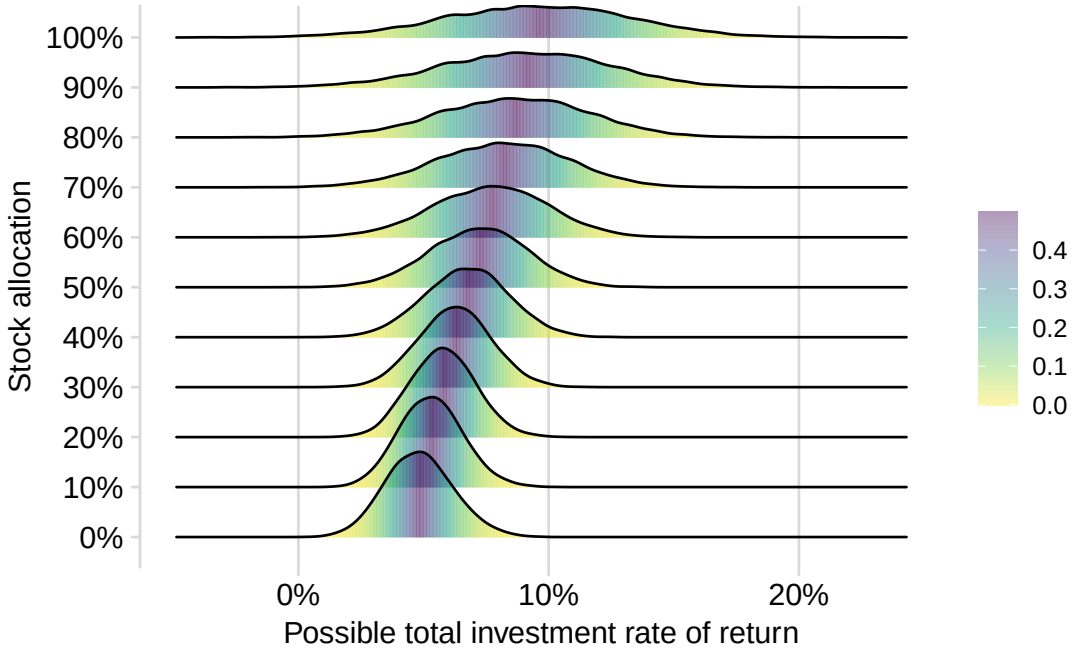


Figure 7.3: Historical simulation of returns for different stock allocation (S&P500) decisions over 30 year investment period in 10% increments. Bond allocation (10 year US Treasury) is 1 minus stock allocation. We use an empirical cumulative density function with viridis palette to indicate density. 100% stocks is the optimal allocation for maximizing expected returns and is consistent with the equity premium puzzle [4].

of the conditional returns given their chosen allocation and the 30-year investment period.

We used the same two asset simulation across 30 years to derive the distribution of portfolio outcomes and calculated the average return for each of the possible allocation combinations. Figure 7.3 shows the distribution of simulated returns for stock allocations in increments of 10%. The color represents density with mean/median corresponding to the darker areas. Aligned with the equity premium puzzle [4], 100% stock allocation has the highest expected return and is the optimal decision if maximizing the expected return. The expected return is also a monotonic function of the stock allocation decision.

One drawback of using stock allocation as a dependent variable is that it does

not reflect the participants' expected returns and how that decision compares to the optimal strategy of choosing 100% stocks. To address this drawback, we follow Fernandes *et al.* [5] and convert stock allocations into the ratio of the expected return relative to the optimal strategy (i.e., expected return for 100% stocks). We then use this ratio of expected return to optimal expected return as our decision metric and dependent variable in our regression analyses.

7.4.1.2 Uncertainty representations

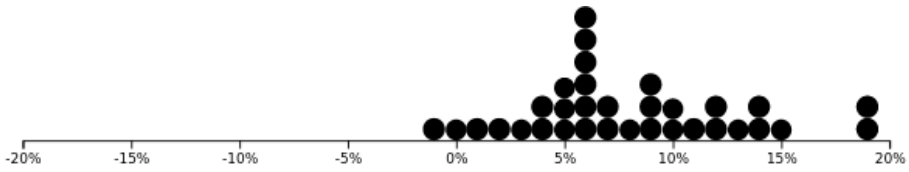
In the second round of the task we evaluated the effect of different uncertainty representations on retirement financial decision making. Our choices of the uncertainty representations were grounded in prior work [229, 274, 275] such that the representations were designed to be "glanceable" and to allow "quick in-the-moment decisions to be made" for mobile devices without significant training. The resulting uncertainty representations cover a design space characterized by frequency framing [246, 5], point-interval [100], animation [98, 100], and controls (round 1 bar chart and table). To ensure consistency across the visualizations, all of the discrete plots use the same 33 returns from bootstrap sampling with replacement. We discuss the rationale of our decision to include each of the visualizations below as our eight treatments.

Table: As a control condition, we provided a table of the returns to compare uncertainty visualizations to a treatment with no data visualization. This enables us to make measurements of what is the marginal effects of data visualizations over the raw data itself. The tables provides the data in ascending order from lowest (top left) to highest (bottom) returns.

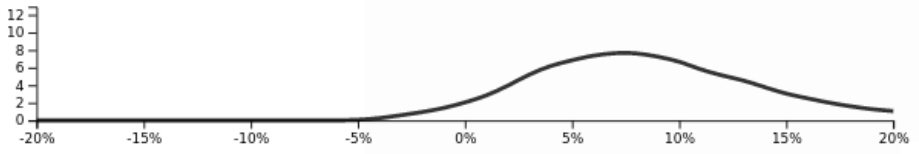
-0.27%	0.90%	1.50%	2.98%	3.89%	4.24%	4.29%	5.24%	5.54%	5.98%
6.06%	6.19%	6.47%	6.65%	6.96%	6.97%	7.38%	7.84%	8.89%	9.30%
9.96%	9.99%	10.55%	10.86%	11.84%	12.08%	12.50%	13.04%	14.15%	14.32%
15.97%	19.16%	19.98%							

Bar chart: We repeated the same visualization from round 1 (see Figure 7.2) to enable a between-subjects comparison of the bar chart versus other uncertainty visualizations.

Dot plot: Past research [229, 5, 100, 31] has shown that frequency framing [276] can improve the understanding of probabilities better than other representations in a variety of tasks. We provide a Wilkinson dot plot [277] to display each dot as one of the 33 possible sampled returns.[5]

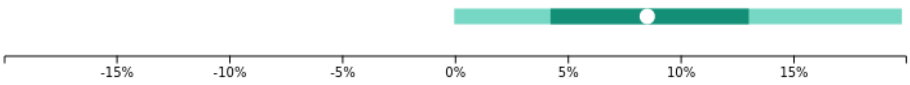


Probability density: A popular data visualization to represent uncertainty on distribution is the (probability) density plot. The purpose of a density is to visualize an underlying data distribution through an approximate continuous curve. This enables a smooth representation that is common in probability distributions and can be used in spatial plots [278].

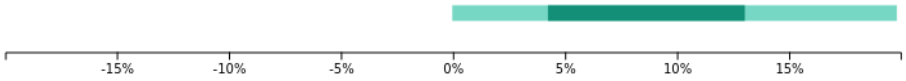


[5]While similar to a quantile dot plot[229, 5], we label it a Wilkinson plot instead of a quantile dot plot as it was generated from a discrete (historical) distribution, not from a continuous distribution.

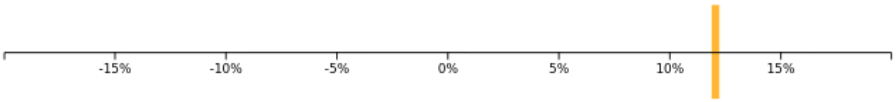
Point-Interval: Recent research in uncertainty visualizations on effect size judgments has shown that providing means to uncertainty visualizations has possible biasing effects [100]. To test for such an effect in our task, we developed the point-interval condition that showed participants intervals that represent a range containing 66% (dark green) and 95% (light green) of the possible (bootstrapped) outcomes as well as a point estimate of the mean.



Interval: Similar to the point-interval, we also provided one condition in which we provided the same intervals (66% and 95%) but without a point (mean) estimate. In this condition, without a separate point for the mean, users could only mentally estimate the mean as the midpoint of the intervals.

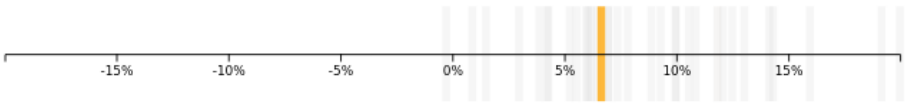


HOP: While static visualizations like error bars are the most common uncertainty visualization, hypothetical outcome plots (HOPs) are designed to focus on the user experiencing uncertainty information through animated draws. HOPs have been studied in a variety of applications including identifying trends [99], effect size judgments [100], and correlation judgment [34].



HOP + Strip: We also provide a hybrid HOP that combines a static distribution (strip plot) overlaid with the HOP. The motivation of this plot is that we expect

individuals will perform better in this condition than the HOP only as it reduces the cognitive load to storing into short term memory the sampled distribution and enables the user to focus attention on the draws relative to the sampled distribution (strip plot).



7.4.2 Participants

We recruited participants through Amazon Mechanical Turk. Given our focus on financial decision-making, we targeted MTurk workers who report owning at least one of four financial products: stocks, bonds, mutual funds, or electronic traded funds (ETFs). To ensure a high quality of performance, participants were either a MTurk Master ($n = 42$) or a non-Master with a HIT acceptance rate of 97% or better ($n = 179$). Following [3], we targeted a sample size of 25 participants per condition (target 200 total across eight conditions). 221 participants completed the study as we expected exclusions. After applying our pre-registration exclusions, we ended with 198 total participants. The average total compensation (with bonus) was $2.46 and participants took on average 14.6 minutes to complete the study. The average age was 37.4 years and 27% identified as female. Out of the entire sample, 67.6% reported owning a retirement investment account.

7.4.3 Procedure

In addition to the main investment task, the experiment included the following components:

Pre-questionnaire: We asked five classical questions on risk aversion. The first two questions ask participants whether they would accept Paul Samuelson's gamble [247] (flip a fair coin and guess correctly get $200 or lose $100 if incorrect) either once

(question 1) or 100 times (question 2). The next two questions measured whether the participant exhibited behavior aligned with prospect theory from Kahneman and Tversky [255]. Participants were asked two questions that provide the same possible payouts but framed as either potential gains (question 3) or losses (question 4). Last, we asked participants to choose between a gamble between a 50/50 chances of winning either $100,000 or $50,000 and sure investment of different payouts that aligned to different coefficients of relative risk aversion (CRRA). Mankiw and Zeldes [251] found that to account for the levels of equity premium in the past, neoclassical economic theory predicts that investors would prefer a certain payoff of $51,209 (CRRA = 30) to a 50/50 bet paying either $50,000 or $100,000.

Post-questionnaire: We required participants to answer six closed ended demographics questions to measure sex, age, education, ownership of financial asset (e.g., stock, bond, ETF, mutual fund), ownership of retirement investment account (e.g., 401k, IRA, Roth IRA), and satisfaction with the study. We also solicited (optional) open ended user feedback on the study.

Payment: All participants who completed the study received at least $1.00 (base) + a bonus of up to $3.50 based on simulated performance (up to $0.25 per 14 trial). A participant's bonus from a trial was based on the quintile of simulated performance of their allocation, with the bonus increasing by $0.05 at each quintile (i.e., lowest quintile: $0.05; highest quintile: $0.25).

Attention/Learning Checks: We included a CAPTCHA after the pre-questionnaire to screen for bots. To check for understanding, we asked three questions following round 1 instructions: how many years in future is hypothetical investments; what is the basic task; what leads to higher incentives. Participants could not move forward if incorrect but could modify their answers until they provided the correct answers. To check for knowledge of the bar chart interpretation, in round 1 participants are asked to hover their mouse over the largest bar and provide the value of that bar.

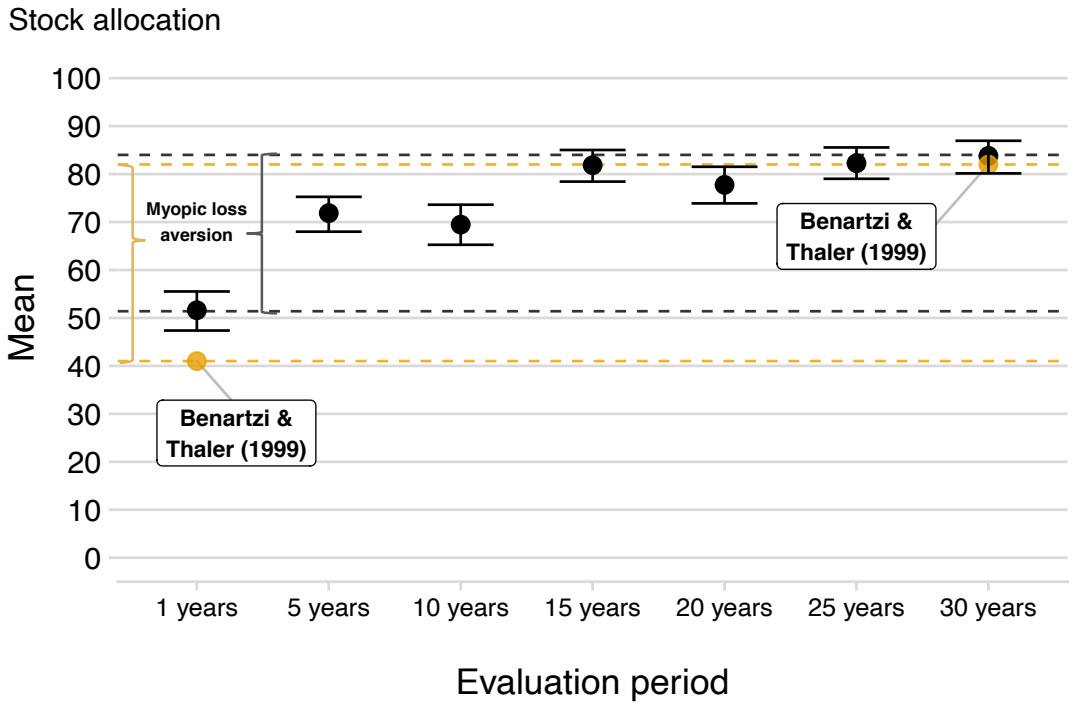


Figure 7.4: Round 1 mean stock allocation and bootstrapped 95% confidence intervals ($n = 198$) by evaluation period by participant. The orange points are the original values from Benartzi and Thaler [3]. Dotted lines are means for 1 and 30 year evaluation periods and the arrows indicate the allocation difference which we measure as myopic loss aversion.

Participants could not proceed without the correct answer (rounded to nearest whole number). Lastly, in the post-questionnaire we asked participants to write a qualitative response in 1-2 sentences to describe the study's task.

User strategy feedback: Recent work has shown users having challenges with uncertainty visualizations through suboptimal strategies or switching strategies [100]. Following [100], we asked participants the following qualitative question to elicit feedback on user strategies after each round: "How did you use the charts to complete the task? Please do your best to describe what sorts of visual properties you looked for and how you used them?"

7.4.4 Analysis approach

Following Fernandes *et al.* [5], we use a mixed effects Bayesian beta regression. We used a beta regression given that the ratio of the expected return to the optimal

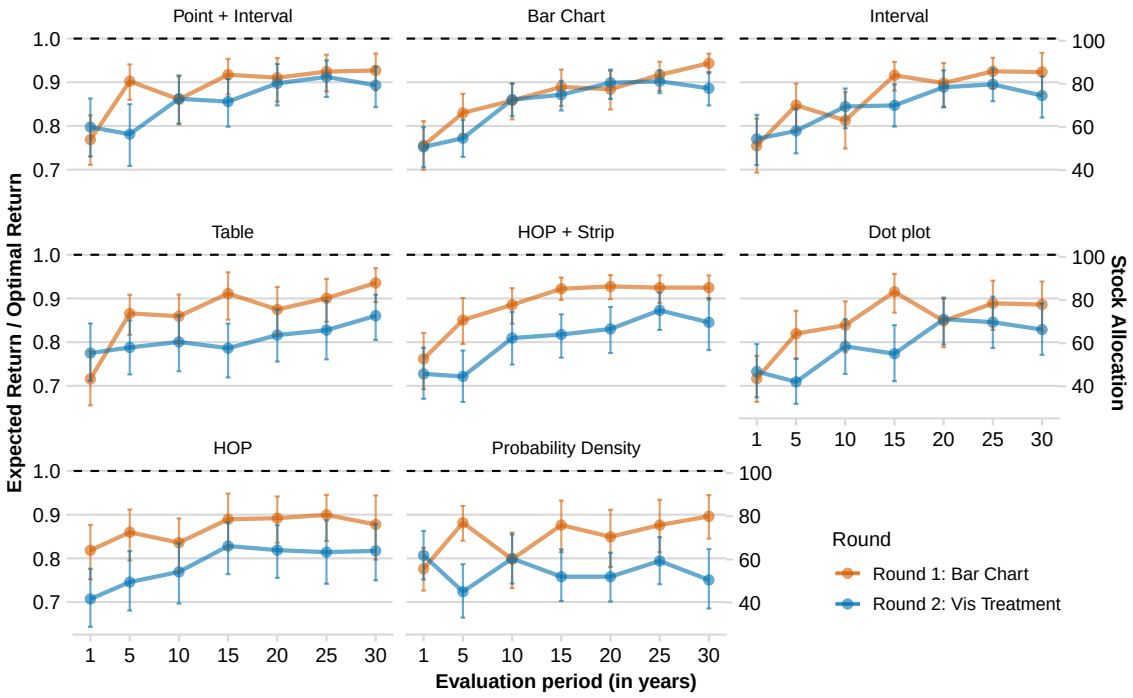


Figure 7.5: Mean and bootstrapped 95% confidence intervals for participants' expected / optimal return (left axis) and stock allocation (right axis). We provide results for both round 1 (bar chart only) and round 2 (visualization treatment). The dotted line indicates the optimal strategy (100% stocks).

expected return are values between 0 and 1. We included the participant ID as a random effect given that decisions are repeated by participant. For fixed effects, we considered both evaluation period and visualization treatment and interaction between evaluation period and treatment.[6] For our priors, we considered both non-informative priors and priors from Fernandes *et al.* [5] but there were no substantive differences. For model fitting and visualization we used R packages **tidyverse** [279], **brms** [211] and **tidybayes** [280].

To aid in the interpretation of our results, we convert the dependent variable into the expected investment value for a hypothetical investor at retirement. Consistent with our average age (37), we assume someone who will retire in 30 years which typically occurs in the United States between 65 and 67. Also we assume an initial

[6]Per our pre-registration, we considered two variants with and without the interaction and decided to include the interaction due to a lower model AIC.

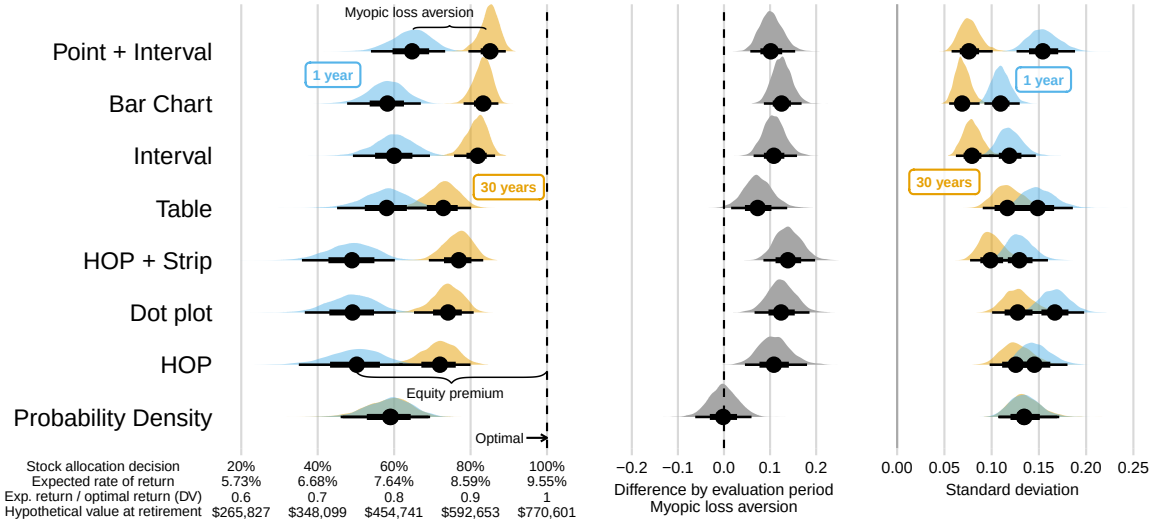


Figure 7.6: Round 2 posterior mean (mu), mean differences, and standard deviations by 1 year (blue) and 30 year (orange) evaluation periods by treatment. We provide multiple conversions of the DV (expected return / optimal return) including the expected return, the stock allocation, and the retirement balance for a hypothetical 37 year old to with $50,000 initial investment (subject to other assumptions).

investment balance ($50,000) similar to the average retirement savings for that age group ($48,710). For simplicity, we assume no role of taxes, no need for liquidity (e.g., cash out), zero discount rate / no inflation, and no additional contributions or withdrawals. Future work could relax these assumptions.

7.5 Results

RQ1: Consistent with Bernatzi and Thaler [3], we find evidence of myopic loss aversion with bar chart visualizations (Round 1), as participants had a significantly lower stock allocation for 1 year versus 30 year evaluation period (Wilcoxon paired test on the median stock allocation, $V = 2020.5$, $z = -9.188$, $p < .001$). Figure 7.4 provides the bootstrapped 95% confidence intervals ($n = 198$) for mean stock allocation. Notably, stock allocations for 1-year and 30-year evaluation periods were similar to the original study by Bernatzi and Thaler [21], adding support for the robustness of the effect in our task and with a crowdsourced participant pool.

We also find a positive monotonic relationship between evaluation period and stock allocation (and expected return) (Figure 7.4). We noticed that the most significant

allocation is the 1 year evaluation period in which average stock allocation was around 40-60% (or 0.7-0.8 expected ratio). This makes sense as the 1 year had the largest amount of negative returns for stocks. We also find that participants' stock allocation tends to level off around 15 year evaluation period, especially for round 1. Figure 7.5 shows the monotonic relationship between evaluation period and participant's expected return normalized by optimal return from both round 1 (bar chart only) and round 2 (visualization treatment). We grouped the plots based on the similarity of second round performance and alignment between round 1 and 2. For example, consider the top row (best performing) plots had similar results in round 1 and 2. The Point + Interval and Interval plots led to the best expected returns comparing to the other uncertainty visualizations. But in the next two rows we can see a decrease of means into round 2, indicating that these visualizations may have hurt performance and increased the equity premium. One interesting exception is the probability density treatment, in which (especially for round 2) stock allocation was nearly flat across all evaluation periods. We'll examine this more carefully in our RQ2 modeling and in our discussion.

RQ2: Figure 7.6 provides posterior estimates of the effects of each visualization by the dependent variable and its conditional mean and standard deviation. To measure the effect of uncertainty visualizations on investment decisions, we provide each treatment with posterior samples of the mean for the 1 year (blue) and 30 years evaluation periods. These plots show the credible intervals of the conditional mean (mu) and standard deviations for each condition. As noted by Fernandes *et al.* [5], since these plots are conditional values, they show the mean and standard deviation for a typical participant given their treatment and evaluation period.

First, we find evidence of differences in performance (equity premium) by visualization. Participants with the point + interval, bar chart, and interval visualizations had on average higher stock allocations and associated higher expected rates of re-

turn. Using that expected rate of return for the hypothetical retirement investor (37 year olds), we can estimate those allocations would lead to over $600,000 value at retirement. Compare this example with participants who used the probability density. Consistent with results in **RQ1**, the model predicts decisions for the probability density with little variation. Those participants choose 60% stocks, which on average would have led to around $450,000 investment at retirement, a nearly two-thirds less value at retirement for 30 year evaluation period.

Nevertheless, on average the decisions were still distant from the optimal (dotted line) as most participants choose the majority stocks but with some bonds. One interpretation of this distance is the equity premium, or how much investors would be willing to forgo to not fully invest in stocks (equity). Similarly, we can also measure myopic loss aversion, which would be the difference between the 1 year (blue) and 30 year evaluation periods (orange). We find that myopic loss aversion can vary between 20% drop in stock for most visualizations. However, for the table and especially the probability density, we find less evidence of myopic loss aversion.

Figure 7.7 provides the model's *prediction* of participant's decisions for their expected returns relative to the optimal expected return (i.e., 100% stock allocation) as well as converted values. Similar to Fernandes *et al.* [5], we use a Bayesian framework to enable measurements of the marginal posterior predictions to predict how a random participant would perform in our experiment's task. The top of Figure 7.7 provides the mean expected return / optimal return (red line). Similar to RQ1, we find that on average the mean performance increases with the evaluation period (x axis) as on average participants allocate more stock when viewing returns in a longer evaluation period. One notable exception is the density plot which exhibits a flat mean performance, which is consistent with what we observed in Figure 7.5.

In addition to the mean values, the figure also includes the posterior predictive intervals, or PPIs, for each treatment with three different value ranges: 50% (dark

blue band), 80% (light blue band), and 95% (lightest blue band). Let's consider the Point + Interval treatment to interpret these values. This treatment exhibited on average the highest value relative to the optimal strategy in which about 50% of decisions (dark blue) ranged from 70-90% optimal (1 year evaluation period) to 87-97% optimal (30 year evaluation period). What's interesting across all of these plots is that while optimal (1.00) was possible within 95% PPI values (lightest blue), for most treatments at best the optimal value was only within the 80% PPI range, especially for longer evaluation periods. While myopic loss aversion and the visualization treatment accounts for some of the underallocation to stocks, this result indicates an additional missing factor that limits participants from making the optimal decision of 100% stocks.

7.6 Visual Reasoning Strategies

To understand participants' reasoning in the task, we analyzed participants' qualitative feedback on the strategies they used to arrive at their decisions. These self-reported descriptions were recorded after each round. To expedite the the analysis, we used Non-Negative Matrix Factorization (NMF) topic modeling [281] to categorize comments from each round into several topics. Two researchers then qualitatively evaluated the semantic differences between the topics by reading top representative documents from each topic. This process resulted in several themes that summarize participants' strategies for each round.

In round 1, participants experienced the bar chart visualization of expected percentage returns based on different evaluation periods. Based on 10 extracted topics, we observed three prevailing themes for round 1. The majority of the comments are related to users' strategies to allocate funds based on their perceptions of risk, returns, and balancing trade-offs between the two.

Minimize risk or maximize reward: A group of comments primarily focus on minimizing risk / losses. For example, one participant wrote: *"I looked at the risk*

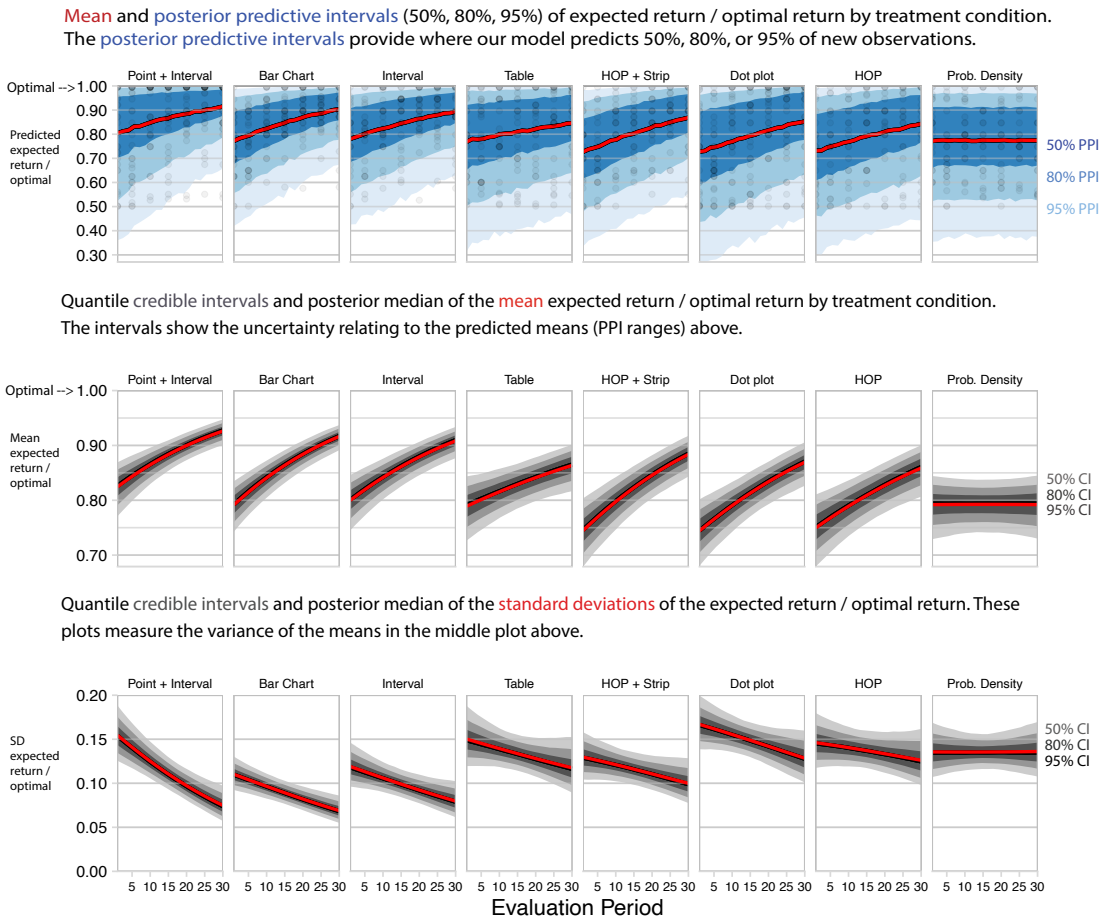


Figure 7.7: Round 2 predicted mean, standard deviations, posterior predictive intervals, and credible intervals from model for expected return / optimal return by treatment and evaluation period. These figures are based on Fernandes *et al.* [5].

and amount of possible losses to make my decision.". Another participant wrote: *'I based it on risk. If a fund had a high likelihood of a negative return, I allocated less money to that fund.'*. One participant explicitly mentioned that although they look at highest returns, they primarily focus on minimizing losses: *"I looked to see which had the best return. I also looked at negative returns. I tended to stay with the option that had little or no negative returns."*. Another group of comments are related to the strategy of maximizing returns. For example, a participant mentioned *"I used the charts to determine what my best chances of the highest return would be."* Similarly, another participant said: *"I looked out for which investment on the chart will give the highest return and allocated a larger percentage of my money to it"*.

Balancing gains and losses: A large portion of the comments described attempts to balance gains versus losses. For example: *"I looked for which charts showed the biggest returns with the least risks of negative returns."* Another participant mentioned how they used the safer fund to hedge their losses: *"[...] I tried to balance reward vs risk and allocate accordingly. I tended to invest more with the aggressive fund in order to maximize profits, but still invested some with the other fund to hedge my gamble."*

Negative bars as a key decision aspect: Other participants looked specifically for negative returns and preferred funds without them (e.g., evaluation periods over 15 years) to maximize returns. For example, one participant said: *"If there were no negative years in either chart, I chose the chart that had the higher returns. If there were negative years in one chart that also had higher returns in other years, I chose 80% or 90% of the riskier chart, with 10% or 20% of the safer chart to balance out the bad years."* Another user had a similar strategy: *"I saw whether there are negative bars. If yes, how many and how long. I tried to imagine the average of an all positive chart and compare that average with other chart."*

In round 2, many responses were consistent with those made in round 1. This indicates some users replicated their strategies in both rounds. For example, one user within the density treatment mentioned that *"I looked to see how much of the distribution was below zero. I viewed the larger area of distribution below zero as being more risky, so I invested more conservatively in those funds. I attempted to invest more heavily in the less risky funds so that I would have a greater chance of not losing money."*. Another user in the table condition had a similar strategy: *"I tended to go with slow growth again. Better if it didn't start out in the negatives cause it could quickly go back in the red zone."*

However, since users were randomly assigned to different visualization treatments, some topics naturally emerged as being related to each visualization treatment. Here

we will describe some interesting emerging themes about different visualization techniques.

Point plot and average rate of return: Some users in the point treatment reported allocating more funds to higher (point) mean returns. For example, one user mentioned: *"I looked at the average return dot. I put all money into the fund with the higher return."* Another mentioned that *"I always chose the fund with the higher average return."*

Interval plot and confidence intervals: Within the Interval treatment, users explained how the dark green area with 66% interval had a much larger influence on their decision, potentially limiting the effects of extreme values in datasets shown to users. For example, one user simply said: *"Wanted the dark green to be highest."*. Similarly, another said: *"I tried to see which chart had the larger opportunity for growth in the dark green zone. I tried to focus more on the dark green zone than the light green zone."*

HOPs and volatility: Within these treatments, some users mentioned they allocated funds with samples with higher returns. For example, one user mentioned: *"I tried to notice if one was higher than the other for most years and allocated more for that one."*. However, some users focused on volatility and made decisions based on funds that changed less. For example, a user mentioned that: *"It was difficult to see the distribution. So I allocated everything to the one that changed the least."* Another user discussed similar strategies: *"I looked at the fluctuations over the period of time and used that to determine the volatility of each investment. I chose the one that I felt was safest to protect my investment but I did always put some in both funds."*

Table, max-min heuristic: Some users reported using a heuristic to find the maximum and minimum returns. These comments were mostly from the Table treatment. For example, one participant wrote: *"I tried to look for the highest and lowest rates in the chart. I struggled to get a good mental picture of each fund, so I tended*

to balance a little more.. Similarly, another user said: *"I checked the highest and lowest return value of each chart and compared them to make decision. Middle ones I ignored."*

7.7 Discussion and Limitations

Participants' performance in the investment task clustered by visualizations with similar encodings. The bar chart and interval plots (including with point) had the highest while HOPs and dot plot had on average lower values. Probability density and table had low myopic loss aversion. But if we consider these plots' performance in round 1 to round 2 from Figure 7, participants decreased their stock allocation indicating that these plots may have increased aversion towards stocks towards more stable bonds.

7.7.1 Simpler, intuitive plots had higher stock allocation

The point + interval, bar chart, and interval plots had the highest stock allocations especially for longer evaluation periods (80%+). From Figure 7.6, for the 30 year evaluation period those conditions also had the lowest standard deviations, which indicate participant consistency near the mean. Consistent with [3], shorter evaluation periods like 1 year decreased stock allocation near to 60-65%. However, for these plots we also find slightly less myopic loss aversion as seen in the difference between the 1 year and 30 year allocations (about 20% less allocation or about 1% less returns) compared to other conditions. To put it into perspective, Figure 7.6 shows that such a difference in allocations could amount to approximately $150,000 at retirement for an average 37 year old investor. We suspect that participants were able to get better mental models of the averages through either direct encoding (point + interval) or easy-to-calculate heuristics like mid-range of intervals or identifying the median bar in the bar charts.

7.7.2 HOPs and dot plots may amplify risk aversion

We find that HOPs and dot plots had lower mean stock allocation (and thus lower expected returns) when comparing both within (round 1 bar chart, Fig. 7.5) and between subjects (round 2 bar chart, Fig. 7.6). We suspect that participants found it more difficult to estimate means or medians visually. For animated plots like HOPs, we found in the feedback participants who overly focused on volatility and may have struggled to identify the mean. However, we think performance with these plots is likely to be context dependent. For example, one simple remedy for these plots would be to overlay mean or medians. We believe with a mean, anchor HOPs and dot plot participants would perform better when optimizing expected returns. Alternatively, different investment objectives may yield more promise for these techniques (e.g., minimum return targets like 5% which may place greater importance on tail risks).

7.7.3 Density and table have mixed results

Similar to HOPs and dot plots, we found that the table and the density plots had lower mean stock allocation than simpler plots. However, table and the density plot differ by displaying little myopic loss aversion as participants didn't change their stock allocation much with longer evaluation periods (see Figure 7.6). But within these two plots, we suspect that participants had trouble differentiating either due to cognitive constraints (table) or continuity smoothing (density).

We find stock allocation (and expected returns) for the table was near the middle but showed some but not significant myopic loss aversion. Unlike HOPs and dot plots, table participants could approximate the means with heuristics like $(min + max)/2$, which could explain the table's higher stock allocation. However, the table group did worse than the simpler plots perhaps because their heuristics weren't as accurate as estimates from simpler plots as the raw average calculation was challenging. We suspect some may have encoded the negative numbers more easily than other plots

which could increase loss aversion. Alternatively, loss aversion may be reduced as the table made it hard to visually measure magnitude (i.e., how large are the losses).

Contrary to the other plots, there was no evidence that probability density plots led to myopic loss aversion. This is despite finding that the same participants exhibited consistent myopic loss aversion in round 1 when presented with bar charts (50% in 1 year and 80% in 30 year, see Figure 7.5). This is compared to a zero-to-negative difference for the same periods in round 2 with the density plot. We suspect that participants may have had difficulty in measuring central tendencies given density's smooth distribution, especially for longer evaluation periods. Like HOPs and dot plot, we suspect overlaying means may increase stock allocation.

7.7.4 Limitations and Future Work

This study focuses on long term retirement investment allocation and makes several assumptions that may limit the results. First, participants' incentives are based on expected results over a thirty year planning horizon. In practice, many retirement investors may want to plan for a near-term withdrawal (e.g., sell stocks for a major purchase or emergency). Second, like [3], we consider only two funds that represent general asset classes (stocks and bonds). This ignores other asset classes like cash, real estate, or riskier assets (e.g., individual stocks, (crypto)currencies, high yield bonds). Third, we assume no effects from taxes or inflation (e.g., zero discount rate), which may affect actual investment behaviors for retirement. Fourth, we use a historical, non-parametric simulation (bootstrap with replacement) of past returns for stimuli and incentives. Instead, parametric approaches like monte carlo simulation based on continuous distributions (e.g., Normal, t-distribution, or fat tail distributions) would produce more uncertain returns and may benefit from visualizations with continuous representations of uncertainty [229, 5].

There are multiple avenues of potential future work. First, the goal of maximizing the expected returns is sensible but in practice retirement investors may have a slightly

different decision problem. For example, many retirement investments are made to meet a balance goal, not necessarily to maximize expected return [268]. Future work could modify the objective and incentivize reaching a simulated goal (e.g., reach 5% annualized returns). Although simple representations of the mean or median appeared to aid performance in our task, different visual encodings which bring attention to the variability in returns or other statistics may be better suited to alternative objectives [100].

Second, we did not provide users feedback for their decision. Future work could explore effects of learning through simulated investment feedback. Such a mechanism could enable measurements of the "explore-exploit" trade-off in allocation decisions. Recent research in cognitive science [282] has used a similar approach to examine short versus long-run strategies which either exploit known information (i.e., experienced returns) or explore different allocation strategies. Another possible direction is the interaction of uncertainty representations with descriptive text that on textual uncertainty [103] or strategy cues [33] to aid users how to interpret or interact with uncertainty representations [100, 31].

Last, the study of allocation with uncertainty visualizations could be expanded to include experts like portfolio managers, shorter investment horizons, or more complex assets. For example, we could incorporate more advanced financial risk measurements like VaR (value-at-risk) or Conditional VaR [283] and measure the interaction providing such metrics can modify the effects of uncertainty representations. Alternatively, we could expand to allocation across many assets and correlation of basket of n-assets, which could also be combined with Bayesian approaches in investment management [284] like Black-Litterman model that incorporate a financial manager's beliefs into a Markowitz modern portfolio theory (MPT) framework [285].

7.8 Conclusion

In this chapter, our contributions include findings from a crowdsourced (MTurk) incentivized mixed design experiment on the effect of uncertainty visualizations have on myopic loss aversion and the equity premium in long-term (retirement) investment decisions. Our results suggest that visualizations could have a large effect in $100,000's balances at retirement for a typical long term investor. While myopic loss aversion has some variation across visualizations, performance remains sub-optimal (below 100% stocks) which suggest future work in providing feedback and learning effects.

CHAPTER 8: CONCLUSIONS

This dissertation provides five experiments to investigate the role of cognitive biases in interactive data visualizations for several tasks for decision-making under uncertainty. This dissertation recasts the importance of a human-centered AI approach [16, 15] and outlines important considerations in the design of future HCAI systems and where cognitive biases may become a potential point of failure in the data analysis process for such interactive data visualization systems.

8.1 Future Work

There is much future work that can be done to further extend the research provided in this dissertation. We'll discuss three possible areas of future work.

8.1.0.1 Explore vs Exploit: Role of Multi-play and Learning Effects

As noted previously, one limitation to our study in Chapter 7 for retirement investment decision-making was that we did not provide users feedback to their decision. This was a design decision to simplify our first study and avoid feedback effects. However, future work could consider multi-play decision under a sampling paradigm [282] in which the system has interactive information foraging functionality like a simulate button. Participants could be observed to determine at what point do they reach a sufficient amount of information gathering (i.e., number of times they click the simulate button for a hypothetical asset simulation) before they're ready to make an allocation decision. There may be additional potential to integrate ideas for portfolio allocation with techniques like reinforcement learning and multi-armed bandits [286] or Thompson sampling [287, 288].

8.1.0.2 NLP + Explanability + Uncertainty Visualizations and Elicitation Techniques

One promising application of future work can be the role of belief bias [20] in evaluating natural language processing explanations through uncertainty representations. Gonzales, Rogers, and Sogaard [289] provide a case study for gradient based explainability approaches in NLP and how argue that belief bias may lead to different conclusions from NLP explanations. Future work could integrate belief elicitation techniques like Line + Range [290] to elicit interval values and the associated uncertainty on those values. In addition, Bayesian cognitive modeling [95] could be used with graphical inference models like latent dirichlet allocation [143] (LDA) that have used Bayesian inference may work for cognitive modeling. Related, other work in NLP [291, 292, 293] have considered labeling for NLP models from a distributional, or uncertain, approach rather than a traditional binary (1 or 0) label. This work too may provide opportunities to use belief elicitation techniques and uncertainty representations for NLP labels that can enhance both NLP model evaluation as well as model training.

8.1.0.3 Developing Graphical Inference Theories

In data analysis, a classic trade-off exists between exploratory versus confirmatory analysis. Exploratory data analysis is more common in interactive data visualization fields like information visualization and visual analytics. However, recent work has shown that too much flexibility in interactive data visualization systems can cause lead to issues of p-hacking or forking paths problems [53, 52] without grounded hypotheses that are common in exploratory data analysis. One recent approach by Hullman and Gelman proposes a more general framework for developing graphical inference theories by highlighting that the concept of a model check in a Bayesian statistical framework may provide a mechanism to unite exploratory and confirma-

tory analysis [25]. They highlight that such an approach provides new opportunities for data visualization designers and empirical researchers. Further, another promising approach to developing more theory on visualization user behaviors for cognitive biases include resource rational approach [60, 294]. Wesslen *et al.* [295] argue for the integration of resource-rational analysis through constrained Bayesian cognitive modeling to understand cognitive biases in data visualizations to provide a more realistic bounded rationality representation of data visualization users. They argue resource-rationality can provide a quantitative framework that can make theoretical predictions for visualization users' decisions that can then be tested empirically. By empirically testing, the original theory can be modified based on experiments through further constraints that reflect more realistic phenomenon and bridge the gap from Marr's computational to algorithmic level of representations [296].

8.2 Closing Remarks

Recent gains in artificial intelligence and machine learning provide many opportunities but also pose great risks due to a lack of humans control if they amplify emerging AI risks like fairness or explainability. Human-centered artificial intelligence [16] provides a promising alternative that aims to increase both automation and human control by focusing AI algorithms around humans instead of humans around AI. Interactive data visualizations have shown many opportunities to develop insight in data analysis through amplifying individual cognition abilities and could have many benefits in HCAI frameworks. However, past research in psychology, cognitive science, and behavioral economics has outlined many problems in human judgment known as cognitive biases that could add noise and problems within HCAI. This dissertation provides five experiments to explore the role of cognitive biases in interactive data visualizations and outlines both evidence of and endeavors to find ways to better harmonize humans' role to design to better HCAI systems and thus enable better decision-making.

REFERENCES

[1] J. R. Firth, "A synopsis of linguistic theory, 1930-1955," *Studies in linguistic analysis*, 1957.

[2] F. Murtagh and P. Legendre, "Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?," *Journal of classification*, vol. 31, no. 3, pp. 274–295, 2014.

[3] S. Benartzi and R. H. Thaler, "Risk aversion or myopia? choices in repeated gambles and retirement investments," *Management science*, vol. 45, no. 3, pp. 364–381, 1999.

[4] R. Mehra and E. C. Prescott, "The equity premium: A puzzle," *Journal of monetary Economics*, vol. 15, no. 2, pp. 145–161, 1985.

[5] M. Fernandes, L. Walls, S. Munson, J. Hullman, and M. Kay, "Uncertainty displays using quantile dotplots or cdfs improve transit decision-making," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 144, ACM, 2018.

[6] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*, pp. 77–91, PMLR, 2018.

[7] S. Hajian, F. Bonchi, and C. Castillo, "Algorithmic bias: From discrimination discovery to fairness-aware data mining," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 2125–2126, ACM, 2016.

[8] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[9] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic books, 2018.

[10] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, *et al.*, "Extracting training data from large language models," *arXiv preprint arXiv:2012.07805*, 2020.

[11] S. Russell, *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?ï¿œï¿œ," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

[13] Z. Tufekci, "Facebook said its algorithms do help form echo chambers, and the tech press missed it," *New Perspectives Quarterly*, vol. 32, no. 3, pp. 9–12, 2015.

[14] Z. Tufekci, "Algorithmic harms beyond facebook and google: Emergent challenges of computational agency," *Colo. Tech. LJ*, vol. 13, p. 203, 2015.

[15] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[16] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 10, no. 4, pp. 1–31, 2020.

[17] D. Kahneman and A. Tversky, "Intuitive prediction: Biases and corrective procedures," tech. rep., DECISIONS AND DESIGNS INC MCLEAN VA, 1977.

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

[19] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises.," *Review of general psychology*, vol. 2, no. 2, p. 175, 1998.

[20] J. S. B. Evans, *Bias in human reasoning: Causes and consequences.* Lawrence Erlbaum Associates, Inc, 1989.

[21] S. Benartzi and R. H. Thaler, "Myopic loss aversion and the equity premium puzzle," *The quarterly journal of Economics*, vol. 110, no. 1, pp. 73–92, 1995.

[22] G. Ellis and A. Dix, "Decision making under uncertainty in visualisation?," in *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2015.

[23] 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 *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

[24] 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.

[25] J. Hullman and A. Gelman, "Designing for interactive exploratory data analysis requires theories of graphical inference," *Harvard Data Science Review*, 2021.

[26] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," *science*, vol. 185, no. 4157, pp. 1124–1131, 1974.

[27] N. Epley and T. Gilovich, "The anchoring-and-adjustment heuristic: Why the adjustments are insufficient," *Psychological science*, vol. 17, no. 4, pp. 311–318, 2006.

[28] P. C. Wason, "On the failure to eliminate hypotheses in a conceptual task," *Quarterly journal of experimental psychology*, vol. 12, no. 3, pp. 129–140, 1960.

[29] I. Cho, R. Wesslen, S. Volkova, W. Ribarsky, and W. Dou, "Crystalball: A visual analytic system for future event discovery and analysis from social media data.," in *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

[30] A. Karduni, R. Wesslen, S. Santhanam, I. Cho, S. Volkova, D. Arendt, S. Shaikh, and W. Dou, "Can you verifi this? studying uncertainty and decision-making about misinformation in visual analytics.," *The 12th International AAAI Conference on Web and Social Media (ICWSM)*, 2018.

[31] L. M. Padilla, M. Powell, M. Kay, and J. Hullman, "Uncertain about uncertainty: How qualitative expressions of forecaster confidence impact decision-making with uncertainty visualizations," *Frontiers in Psychology*, vol. 11, 2020.

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

[33] R. Wesslen, S. Santhanam, A. Karduni, I. Cho, S. Shaikh, and W. Dou, "Investigating effects of visual anchors on decision-making about misinformation," in *Computer Graphics Forum*, vol. 38, pp. 161–171, Wiley Online Library, 2019.

[34] A. Karduni, D. Markant, R. Wesslen, and W. Dou, "A bayesian cognition approach for belief updating of correlation judgement through uncertainty visualizations," *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[35] R. Wesslen, A. Karduni, D. Markant, and W. Dou, "Effect of uncertainty visualizations on myopic loss aversion and equity premium puzzle in retirement investment decisions," *arXiv preprint arXiv:2107.02334*, 2021.

[36] T. L. Griffiths, "Manifesto for a new (computational) cognitive revolution," *Cognition*, vol. 135, pp. 21–23, 2015.

[37] E. Dimara and J. Stasko, "A critical reflection on visualization research: Where do decision making tasks hide?," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

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

[39] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," in *Visual data mining*, pp. 76–90, Springer, 2008.

[40] S. Card, J. Mackinlay, and B. Shneiderman, "Information visualization," *Human-computer interaction: Design issues, solutions, and applications*, vol. 181, 2009.

[41] R. Mazza, *Introduction to information visualization*. Springer Science & Business Media, 2009.

[42] C. Ware, *Information visualization: perception for design*. Elsevier, 2012.

[43] P. Parsons and K. Sedig, "Common visualizations: Their cognitive utility," in *Handbook of human centric visualization*, pp. 671–691, Springer, 2014.

[44] L. M. Padilla, "A case for cognitive models in visualization research," 2018.

[45] T. M. Green, W. Ribarsky, and B. Fisher, "Building and applying a human cognition model for visual analytics," *Information visualization*, vol. 8, no. 1, pp. 1–13, 2009.

[46] 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.

[47] 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.

[48] T. T. Hills, P. M. Todd, D. Lazer, A. D. Redish, I. D. Couzin, C. S. R. Group, *et al.*, "Exploration versus exploitation in space, mind, and society," *Trends in cognitive sciences*, vol. 19, no. 1, pp. 46–54, 2015.

[49] P. Pirolli and S. Card, "Information foraging.," *Psychological review*, vol. 106, no. 4, p. 643, 1999.

[50] B. Christian and T. Griffiths, *Algorithms to live by: The computer science of human decisions*. Macmillan, 2016.

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

[52] X. Pu and M. Kay, "The garden of forking paths in visualization: A design space for reliable exploratory visual analytics," 2018.

[53] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska, "Investigating the effect of the multiple comparisons problem in visual analysis," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 479, ACM, 2018.

[54] A. Kale, M. Kay, and J. Hullman, "Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2019.

[55] A. C. Valdez, M. Ziefle, and M. Sedlmair, "Studying biases in visualization research: Framework and methods," in *Cognitive Biases in Visualizations*, pp. 13–27, Springer, 2018.

[56] M. Pohl, L.-C. Winter, C. Pallaris, S. Attfield, and B. W. Wong, "Sensemaking and cognitive bias mitigation in visual analytics," in *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*, pp. 323–323, IEEE, 2014.

[57] G. Ellis, "So, what are cognitive biases?," in *Cognitive Biases in Visualizations*, pp. 1–10, Springer, 2018.

[58] E. Dimara, G. Bailly, A. Bezerianos, and S. Franconeri, "Mitigating the attraction effect with visualizations," *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[59] P. Parsons, "Promoting representational fluency for cognitive bias mitigation in information visualization," in *Cognitive Biases in Visualizations*, pp. 137–147, Springer, 2018.

[60] F. Lieder and T. L. Griffiths, "Resource-rational analysis: understanding human cognition as the optimal use of limited," *Psychological Science*, vol. 2, no. 6, pp. 396–408, 2018.

[61] J. E. Korteling, A.-M. Brouwer, and A. Toet, "A neural network framework for cognitive bias," *Frontiers in psychology*, vol. 9, 2018.

[62] H. A. Simon, "A behavioral model of rational choice," *The quarterly journal of economics*, vol. 69, no. 1, pp. 99–118, 1955.

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

[64] G. Gigerenzer and W. Gaissmaier, "Heuristic decision making," *Annual review of psychology*, vol. 62, pp. 451–482, 2011.

[65] G. Klein, "Sources of power: How people make decisions. 1998," *MIT Press, ISBN*, vol. 13, pp. 978–0, 1998.

[66] M. G. Haselton and D. Nettle, "The paranoid optimist: An integrative evolutionary model of cognitive biases," *Personality and social psychology Review*, vol. 10, no. 1, pp. 47–66, 2006.

[67] J. Tooby and L. Cosmides, "Evolutionary psychology: Conceptual foundations," *The handbook of evolutionary psychology*, 2005.

[68] J. E. Russo, P. J. Schoemaker, and E. J. Russo, *Decision traps: Ten barriers to brilliant decision-making and how to overcome them.* Doubleday/Currency New York, NY, 1989.

[69] T. D. Wilson, C. E. Houston, K. M. Etling, and N. Brekke, "A new look at anchoring effects: basic anchoring and its antecedents.," *Journal of Experimental Psychology: General*, vol. 125, no. 4, p. 387, 1996.

[70] K. E. Jacowitz and D. Kahneman, "Measures of anchoring in estimation tasks," *Personality and Social Psychology Bulletin*, vol. 21, no. 11, pp. 1161–1166, 1995.

[71] E. Jonas, S. Schulz-Hardt, D. Frey, and N. Thelen, "Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information.," *Journal of personality and social psychology*, vol. 80, no. 4, p. 557, 2001.

[72] C. R. Mynatt, M. E. Doherty, and R. D. Tweney, "Confirmation bias in a simulated research environment: An experimental study of scientific inference," *The quarterly journal of experimental psychology*, vol. 29, no. 1, pp. 85–95, 1977.

[73] J. S. B. Evans and J. Curtis-Holmes, "Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning," *Thinking & Reasoning*, vol. 11, no. 4, pp. 382–389, 2005.

[74] P. A. Klaczynski, "Motivated scientific reasoning biases, epistemological beliefs, and theory polarization: A two-process approach to adolescent cognition," *Child Development*, vol. 71, no. 5, pp. 1347–1366, 2000.

[75] W. C. Sá, R. F. West, and K. E. Stanovich, "The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill.," *Journal of educational psychology*, vol. 91, no. 3, p. 497, 1999.

[76] C. Chinn and W. Brewer, "The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction," *Review of educational research*, vol. 63, no. 1, p. 1, 1993.

[77] D. Billman, B. Bornstein, and J. Richards, "Effects of expectancy on assessing covariation in data:"prior belief" versus "meaning"," *Organizational Behavior and Human Decision Processes*, vol. 53, no. 1, pp. 74–88, 1992.

[78] H. Baumgartner, "On the utility of consumers' theories in judgments of covariation," *Journal of Consumer Research*, vol. 21, no. 4, pp. 634–643, 1995.

[79] J. S. B. Evans, "In two minds: dual-process accounts of reasoning," *Trends in cognitive sciences*, vol. 7, no. 10, pp. 454–459, 2003.

[80] J. S. B. Evans, J. L. Barston, and P. Pollard, "On the conflict between logic and belief in syllogistic reasoning," *Memory & cognition*, vol. 11, no. 3, pp. 295–306, 1983.

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

[82] 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*, pp. 85–95, ACM, 2016.

[83] E. Dimara, P. Dragicevic, and A. Bezerianos, "Accounting for availability biases in information visualization," *arXiv preprint arXiv:1610.02857*, 2016.

[84] 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.

[85] 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*, pp. 555–575, Springer, 2019.

[86] M. Procopio, A. Mosca, C. E. Scheidegger, E. Wu, and R. Chang, "Impact of cognitive biases on progressive visualization," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[87] R. Delpish, S. Jiang, L. Davis, and K. Odubela, "A visual analytics approach to combat confirmation bias for a local food bank," in *International Conference on Applied Human Factors and Ergonomics*, pp. 13–23, Springer, 2018.

[88] J. Baron, *Thinking and deciding*. Cambridge University Press, 2000.

[89] R. F. Pohl, *Cognitive illusions: Intriguing phenomena in judgement, thinking and memory*. Psychology Press, 2016.

[90] D. Kirsh, "Thinking with external representations," *AI & society*, vol. 25, no. 4, pp. 441–454, 2010.

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

[92] B. Tversky, "Visuospatial reasoning," *The Cambridge handbook of thinking and reasoning*, pp. 209–240, 2005.

[93] Z. Liu and J. Stasko, "Mental models, visual reasoning and interaction in information visualization: A top-down perspective," *IEEE Transactions on Visualization & Computer Graphics*, no. 6, pp. 999–1008, 2010.

[94] Y. Wu, L. Xu, R. Chang, and E. Wu, "Towards a bayesian model of data visualization cognition," 2017.

[95] 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*, pp. 1–14, 2019.

[96] T. L. Griffiths and J. B. Tenenbaum, "Optimal predictions in everyday cognition," *Psychological science*, vol. 17, no. 9, pp. 767–773, 2006.

[97] T. L Griffiths, C. Kemp, and J. B Tenenbaum, "Bayesian models of cognition," 2008.

[98] 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, p. e0142444, 2015.

[99] 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.

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

[101] J. R. Anderson, *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1990.

[102] B. Rehder, C. Lewis, B. Terwilliger, P. Polson, and J. Rieman, "A model of optimal exploration and decision making in novel interfaces," in *Conference Companion on Human Factors in Computing Systems*, pp. 230–231, ACM, 1995.

[103] J. Hullman, X. Qiao, M. Correll, A. Kale, and M. Kay, "In pursuit of error: A survey of uncertainty visualization evaluation," *IEEE transactions on visualization and computer graphics*, 2018.

[104] M. Botvinick and T. Braver, "Motivation and cognitive control: from behavior to neural mechanism," *Annual review of psychology*, vol. 66, 2015.

[105] J. von Neumann, O. Morgenstern, H. W. Kuhn, and A. Rubinstein, *Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition)*. Princeton University Press, 1944.

[106] J. Tsai, S. Miller, and A. Kirlik, "Interactive visualizations to improve bayesian reasoning," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 55, pp. 385–389, SAGE Publications Sage CA: Los Angeles, CA, 2011.

[107] 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 & Computer Graphics*, no. 1, pp. 1–1, 2018.

[108] M. Greis, J. Hullman, M. Correll, M. Kay, and O. Shaer, "Designing for uncertainty in hci: When does uncertainty help?," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pp. 593–600, ACM, 2017.

[109] L. Micallef, P. Dragicevic, and J.-D. Fekete, "Assessing the effect of visualizations on bayesian reasoning through crowdsourcing," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 12, pp. 2536–2545, 2012.

[110] G. L. Brase, "Pictorial representations in statistical reasoning," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 23, no. 3, pp. 369–381, 2009.

[111] S. A. Sloman, D. Over, L. Slovak, and J. M. Stibel, "Frequency illusions and other fallacies," *Organizational Behavior and Human Decision Processes*, vol. 91, no. 2, pp. 296–309, 2003.

[112] D. M. Eddy, "Probabilistic reasoning in clinical medicine: Problems and opportunities," 1982.

[113] J. S. Blumenthal-Barby and H. Krieger, "Cognitive biases and heuristics in medical decision making: a critical review using a systematic search strategy," *Medical Decision Making*, vol. 35, no. 4, pp. 539–557, 2015.

[114] L. Cen, G. Hilary, and K. J. Wei, "The role of anchoring bias in the equity market: Evidence from analysts' earnings forecasts and stock returns," *Journal of Financial and Quantitative Analysis*, vol. 48, no. 01, pp. 47–76, 2013.

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

[116] D. Kahneman, "A perspective on judgment and choice: mapping bounded rationality.," *American psychologist*, vol. 58, no. 9, p. 697, 2003.

[117] P.-L. Yu, *Multiple-criteria decision making: concepts, techniques, and extensions*, vol. 30. Springer Science & Business Media, 2013.

[118] J. M. Tien, "Big data: Unleashing information," *Journal of Systems Science and Systems Engineering*, vol. 22, no. 2, pp. 127–151, 2013.

[119] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon, *Visual Analytics: Definition, Process, and Challenges*, pp. 154–175. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.

[120] D. M. Eler, F. V. Paulovich, M. C. F. de Oliveira, and R. Minghim, "Coordinated and multiple views for visualizing text collections," in *2008 12th International Conference Information Visualisation*, pp. 246–251, IEEE, 2008.

[121] H. A. Meier, M. Schlemmer, C. Wagner, A. Kerren, H. Hagen, E. Kuhl, and P. Steinmann, "Visualization of particle interactions in granular media," *IEEE transactions on visualization and computer graphics*, vol. 14, no. 5, pp. 1110–1125, 2008.

[122] N. A. Giacobe and S. Xu, "Geovisual analytics for cyber security: Adopting the geoviz toolkit," in *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 315–316, IEEE, 2011.

[123] D. Keefe, M. Ewert, W. Ribarsky, and R. Chang, "Interactive coordinated multiple-view visualization of biomechanical motion data," *IEEE transactions on visualization and computer graphics*, vol. 15, no. 6, pp. 1383–1390, 2009.

[124] I. Cho, W. Dou, D. X. Wang, E. Sauda, and W. Ribarsky, "Vairoma: A visual analytics system for making sense of places, times, and events in roman history," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 210–219, 2016.

[125] J. C. Roberts, "State of the art: Coordinated & multiple views in exploratory visualization," in *Coordinated and Multiple Views in Exploratory Visualization, 2007. CMV'07. Fifth International Conference on*, pp. 61–71, IEEE, 2007.

[126] 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.

[127] M. D. Plumlee and C. Ware, "Zooming versus multiple window interfaces: Cognitive costs of visual comparisons," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 13, no. 2, pp. 179–209, 2006.

[128] G. Convertino, J. Chen, B. Yost, Y.-S. Ryu, and C. North, "Exploring context switching and cognition in dual-view coordinated visualizations," in *Proceedings International Conference on Coordinated and Multiple Views in Exploratory Visualization-CMV 2003-*, pp. 55–62, IEEE, 2003.

[129] A. M. MacEachren, "Visual analytics and uncertainty: Its not about the data," 2015.

[130] I. Cho, R. Wesslen, S. Volkova, W. Ribarsky, and W. Dou, "Crystalball: A visual analytic system for future event discovery and analysis from social media data," *Visual Analytics Science and Technology (VAST), 2017 IEEE Conference on*.

[131] G. Gigerenzer and H. Brighton, "Homo heuristicus: Why biased minds make better inferences," *Topics in Cognitive Science*, vol. 1, no. 1, pp. 107–143, 2009.

[132] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," in *Utility, probability, and human decision making*, pp. 141–162, Springer, 1975.

[133] M. W. Bennett, "Confronting cognitive anchoring effect and blind spot biases in federal sentencing: A modest solution for reforming a fundamental flaw," *J. Crim. L. & Criminology*, vol. 104, p. 489, 2014.

[134] D. D. Loschelder, R. Trötschel, R. I. Swaab, M. Friese, and A. D. Galinsky, "The information-anchoring model of first offers: When moving first helps versus hurts negotiators.," *Journal of Applied Psychology*, vol. 101, no. 7, p. 995, 2016.

[135] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim, "The role of uncertainty, awareness, and trust in visual analytics," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 240–249, 2016.

[136] A. Nussbaumer, K. Verbert, E.-C. Hillemann, M. A. Bedek, and D. Albert, "A framework for cognitive bias detection and feedback in a visual analytics environment," in *Intelligence and Security Informatics Conference (EISIC), 2016 European*, pp. 148–151, IEEE, 2016.

[137] L. Harrison, D. Skau, S. Franconeri, A. Lu, and R. Chang, "Influencing visual judgment through affective priming," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2949–2958, 2013.

[138] J. F. George, K. Duffy, and M. Ahuja, "Countering the anchoring and adjustment bias with decision support systems," *Decision Support Systems*, vol. 29, no. 2, pp. 195–206, 2000.

[139] M. B. Cook and H. S. Smallman, "Human factors of the confirmation bias in intelligence analysis: Decision support from graphical evidence landscapes," *Human Factors*, vol. 50, no. 5, pp. 745–754, 2008.

[140] X. Jin, Y. Zhou, and B. Mobasher, "A unified approach to personalization based on probabilistic latent semantic models of web usage and content," in *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP'04)*, 2004.

[141] A. S. Das, M. Datar, A. Garg, and S. Rajaram, "Google news personalization: scalable online collaborative filtering," in *Proceedings of the 16th international conference on World Wide Web*, pp. 271–280, ACM, 2007.

[142] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, ACM, 1999.

[143] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[144] T. Iwata, S. Watanabe, T. Yamada, and N. Ueda, "Topic tracking model for analyzing consumer purchase behavior.," in *IJCAI*, vol. 9, pp. 1427–1432, 2009.

[145] "Twitter usage statistics." `http://www.internetlivestats.com/twitter-statistics/`. Accessed: 2017-03-31.

[146] O. P. John, E. M. Donahue, and R. L. Kentle, "The big five inventory-versions 4a and 54," 1991.

[147] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Third international AAAI conference on weblogs and social media*, 2009.

[148] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008.

[149] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[150] M. E. Roberts, B. M. Stewart, and E. M. Airoldi, "A model of text for experimentation in the social sciences," *Journal of the American Statistical Association*, vol. 111, no. 515, pp. 988–1003, 2016.

[151] D. M. Blei and J. D. Lafferty, "A correlated topic model of science," *The Annals of Applied Statistics*, pp. 17–35, 2007.

[152] J. Eisenstein, A. Ahmed, and E. P. Xing, "Sparse additive generative models of text," 2011.

[153] D. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with dirichlet-multinomial regression," *arXiv preprint arXiv:1206.3278*, 2012.

[154] M. Roberts, B. Stewart, and D. Tingley, "stm: R package for structural topic models," *R package version 1.2.1*, 2017.

[155] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson, and D. G. Rand, "Structural topic models for open-ended survey responses," *American Journal of Political Science*, vol. 58, no. 4, pp. 1064–1082, 2014.

[156] C. Fong and J. Grimmer, "Discovery of treatments from text corpora," in *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2016.

[157] L. Howell *et al.*, "Digital wildfires in a hyperconnected world," *WEF Report*, vol. 3, pp. 15–94, 2013.

[158] A. Kott, D. S. Alberts, and C. Wang, "War of 2050: a battle for information, communications, and computer security," *arXiv preprint arXiv:1512.00360*, 2015.

[159] A. Tsang and K. Larson, "The echo chamber: Strategic voting and homophily in social networks," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 368–375, International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[160] N. Mele, D. Lazer, M. Baum, N. Grinberg, L. Friedland, K. Joseph, W. Hobbs, and C. Mattsson, "Combating fake news: An agenda for research and action," 2017.

[161] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, "The spread of fake news by social bots," *arXiv preprint arXiv:1707.07592*, 2017.

[162] M. Allan, "Information literacy and confirmation bias: You can lead a person to information, but can you make him think?," 2017.

[163] J. Soll, "The long and brutal history of fake news," *POLITICO Magazine*, 2017.

[164] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[165] M. Wagner, F. Fischer, R. Luh, A. Haberson, A. Rind, D. A. Keim, W. Aigner, R. Borgo, F. Ganovelli, and I. Viola, "A survey of visualization systems for malware analysis," in *EG Conference on Visualization (EuroVis)-STARs*, pp. 105–125, 2015.

[166] F. Wanner, A. Stoffel, D. Jäckle, B. C. Kwon, A. Weiler, D. A. Keim, K. E. Isaacs, A. Giménez, I. Jusufi, T. Gamblin, *et al.*, "State-of-the-art report of visual analysis for event detection in text data streams," in *Computer Graphics Forum*, vol. 33, 2014.

[167] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas, "Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, pp. 647–653, 2017.

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

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

[170] R. M. Entman, "Framing bias: Media in the distribution of power," *Journal of communication*, vol. 57, no. 1, pp. 163–173, 2007.

[171] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017.

[172] N. Hassan, A. Sultana, Y. Wu, G. Zhang, C. Li, J. Yang, and C. Yu, "Data in, fact out: automated monitoring of facts by factwatcher," *Proceedings of the VLDB Endowment*, vol. 7, no. 13, pp. 1557–1560, 2014.

[173] P. T. Metaxas, S. Finn, and E. Mustafaraj, "Using twittertrails. com to investigate rumor propagation," in *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pp. 69–72, ACM, 2015.

[174] P. Resnick, S. Carton, S. Park, Y. Shen, and N. Zeffer, "Rumorlens: A system for analyzing the impact of rumors and corrections in social media," in *Proc. Computational Journalism Conference*, 2014.

[175] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Hoaxy: A platform for tracking online misinformation," in *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 745–750, International World Wide Web Conferences Steering Committee, 2016.

[176] W. C. Adams, "Whose lives count? tv coverage of natural disasters," *Journal of Communication*, vol. 36, no. 2, pp. 113–122, 1986.

[177] R. M. Entman, *Projections of power: Framing news, public opinion, and US foreign policy*. University of Chicago Press, 2004.

[178] K. Starbird, "Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter.," in *ICWSM*, pp. 230–239, 2017.

[179] H. Wallach, "Computational social science ≠ computer science + social data," *Commun. ACM*, vol. 61, pp. 42–44, Feb. 2018.

[180] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "Linguistic models for analyzing and detecting biased language.," in *ACL (1)*, pp. 1650–1659, 2013.

[181] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pp. 347–354, Association for Computational Linguistics, 2005.

[182] S. Volkova, Y. Bachrach, M. Armstrong, and V. Sharma, "Inferring latent user properties from texts published in social media.," in *AAAI*, pp. 4296–4297, 2015.

[183] B. Liu, M. Hu, and J. Cheng, "Opinion observer: analyzing and comparing opinions on the web," in *Proceedings of the 14th international conference on World Wide Web*, pp. 342–351, ACM, 2005.

[184] J. Graham, J. Haidt, and B. A. Nosek, "Liberals and conservatives rely on different sets of moral foundations.," *Journal of personality and social psychology*, vol. 96, no. 5, p. 1029, 2009.

[185] J. Haidt and J. Graham, "When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize," *Social Justice Research*, vol. 20, no. 1, pp. 98–116, 2007.

[186] J. Rajsic, D. E. Wilson, and J. Pratt, "Confirmation bias in visual search.," *Journal of experimental psychology: human perception and performance*, vol. 41, no. 5, p. 1353, 2015.

[187] J. W. Tukey, *Exploratory data analysis*. Addison-Wesley Pub. Co., Reading, Mass., 1977.

[188] D. A. Keim, "Information visualization and visual data mining," *IEEE transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 1–8, 2002.

[189] T. Munzner, *Visualization analysis and design*. CRC press, 2014.

[190] H. Lam and T. Munzner, "A guide to visual multi-level interface design from synthesis of empirical study evidence," *Synthesis Lectures on Visualization*, vol. 1, no. 1, pp. 1–117, 2010.

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

[192] 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*, 2019.

[193] D. Kahneman, "36 heuristics and biases," *Scientists Making a Difference: One Hundred Eminent Behavioral and Brain Scientists Talk about Their Most Important Contributions*, p. 171, 2016.

[194] F. Lieder, T. L. Griffiths, Q. J. Huys, and N. D. Goodman, "The anchoring bias reflects rational use of cognitive resources," *Psychonomic bulletin & review*, pp. 1–28, 2017.

[195] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[196] G. Pennycook, T. Cannon, and D. G. Rand, "Prior exposure increases perceived accuracy of fake news," 2018.

[197] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[198] R. Pohl, *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press, 2004.

[199] S. E. Blackwell, M. Browning, A. Mathews, A. Pictet, J. Welch, J. Davies, P. Watson, J. R. Geddes, and E. A. Holmes, "Positive imagery-based cognitive bias modification as a web-based treatment tool for depressed adults: a randomized controlled trial," *Clinical Psychological Science*, vol. 3, no. 1, pp. 91–111, 2015.

[200] T. Amer, D. G. Gozli, and J. Pratt, "Biasing spatial attention with semantic information: an event coding approach," *Psychological research*, pp. 1–19, 2017.

[201] W. Wright, D. Sheffield, and S. Santosa, "Argument mapper: Countering cognitive biases in analysis with critical (visual) thinking," in *Information Visualisation (IV), 2017 21st International Conference*, pp. 250–255, IEEE, 2017.

[202] D. M. Shaffer, E. McManama, and F. H. Durgin, "Manual anchoring biases in slant estimation affect matches even for near surfaces," *Psychonomic bulletin & review*, vol. 22, no. 6, pp. 1665–1670, 2015.

[203] D. Bonaretti, M. Ł. Bartosiak, and G. Piccoli, "Cognitive anchoring of color cues on online review ratings," 2017.

[204] 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*, vol. 5, pp. 2–4, McLean, VA, USA, 2005.

[205] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller, "A systematic review on the practice of evaluating visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, pp. 2818–2827, Dec. 2013.

[206] P. Dragicevic, "Fair statistical communication in hci," in *Modern Statistical Methods for HCI*, pp. 291–330, Springer, 2016.

[207] J. O. Wobbrock and M. Kay, "Nonparametric statistics in human–computer interaction," in *Modern statistical methods for HCI*, pp. 135–170, Springer, 2016.

[208] M. Kay, G. L. Nelson, and E. B. Hekler, "Researcher-centered design of statistics: Why bayesian statistics better fit the culture and incentives of hci," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 4521–4532, ACM, 2016.

[209] T. Galili, A. O'Callaghan, J. Sidi, and C. Sievert, "heatmaply: an r package for creating interactive cluster heatmaps for online publishing," *Bioinformatics*, 2017.

[210] T. Galili, "dendextend: an r package for visualizing, adjusting, and comparing trees of hierarchical clustering," *Bioinformatics*, 2015.

[211] P.-C. Bürkner *et al.*, "brms: An r package for bayesian multilevel models using stan," *Journal of Statistical Software*, vol. 80, no. 1, pp. 1–28, 2017.

[212] M. Kay, *tidybayes: Tidy Data and Geoms for Bayesian Models*, 2018. R package version 1.0.0.

[213] Y.-S. Kim, K. Reinecke, and J. Hullman, "Explaining the gap: Visualizing one's predictions improves recall and comprehension of data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1375–1386, ACM, 2017.

[214] F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang, "Correlation judgment and visualization features: A comparative study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 3, pp. 1474–1488, 2018.

[215] L. Harrison, F. Yang, S. Franconeri, and R. Chang, "Ranking visualizations of correlation using weber's law," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1943–1952, 2014.

[216] M. Kay and J. Heer, "Beyond weber's law: A second look at ranking visualizations of correlation," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 469–478, 2015.

[217] R. A. Rensink and G. Baldridge, "The perception of correlation in scatterplots," in *Computer Graphics Forum*, vol. 29, pp. 1203–1210, Wiley Online Library, 2010.

[218] R. A. Rensink, "The nature of correlation perception in scatterplots," *Psychonomic bulletin & review*, vol. 24, no. 3, pp. 776–797, 2017.

[219] J. Hullman, "Why authors don't visualize uncertainty," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 130–139, 2019.

[220] A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. Han, and R. Chang, "Improving bayesian reasoning: The effects of phrasing, visualization, and spatial ability," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 529–538, 2015.

[221] B. C. P. Kraan, "Probabilistic inversion in uncertainty analysis: and related topics," 2002.

[222] A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow, *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

[223] M. Zondervan-Zwijnenburg, W. van de Schoot-Hubeek, K. Lek, H. Hoijtink, and R. van de Schoot, "Application and evaluation of an expert judgment elicitation procedure for correlations," *Frontiers in psychology*, vol. 8, p. 90, 2017.

[224] A. Sanborn and T. L. Griffiths, "Markov chain monte carlo with people," in *Advances in neural information processing systems*, pp. 1265–1272, 2008.

[225] M. Sedlmair, T. Munzner, and M. Tory, "Empirical guidance on scatterplot and dimension reduction technique choices," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2634–2643, 2013.

[226] F. J. Anscombe, "Graphs in statistical analysis," *The american statistician*, vol. 27, no. 1, pp. 17–21, 1973.

[227] J. Meyer, M. Taieb, and I. Flascher, "Correlation estimates as perceptual judgments.," *Journal of Experimental Psychology: Applied*, vol. 3, no. 1, p. 3, 1997.

[228] J. Li, J.-B. Martens, and J. J. Van Wijk, "Judging correlation from scatterplots and parallel coordinate plots," *Information Visualization*, vol. 9, no. 1, pp. 13–30, 2010.

[229] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson, "When (ish) is my bus? user-centered visualizations of uncertainty in everyday, mobile predictive systems," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5092–5103, 2016.

[230] A. M. MacEachren, R. E. Roth, J. O'Brien, B. Li, D. Swingley, and M. Gahegan, "Visual semiotics & uncertainty visualization: An empirical study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2496–2505, 2012.

[231] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr, "Non-expert interpretations of hurricane forecast uncertainty visualizations," *Spatial Cognition & Computation*, vol. 16, no. 2, pp. 154–172, 2016.

[232] L. Liu, A. P. Boone, I. T. Ruginski, L. Padilla, M. Hegarty, S. H. Creem-Regehr, W. B. Thompson, C. Yuksel, and D. H. House, "Uncertainty visualization by representative sampling from prediction ensembles," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 9, pp. 2165–2178, 2016.

[233] J. Hullman, "Why evaluating uncertainty visualization is error prone," in *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pp. 143–151, 2016.

[234] J. Hullman, M. Kay, Y.-S. Kim, and S. Shrestha, "Imagining replications: Graphical prediction & discrete visualizations improve recall & estimation of effect uncertainty," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 446–456, 2017.

[235] D. Jennings, T. M. Amabile, and L. Ross, "Judgment under uncertainty: Heuristics and biases," ch. Informal covariation assessment: Data-based vs. theory-based judgments, Cambridge University Press, 1982.

[236] R. T. Clemen, G. W. Fischer, and R. L. Winkler, "Assessing dependence: Some experimental results," *Management Science*, vol. 46, no. 8, pp. 1100–1115, 2000.

[237] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan, "Statistical methods for eliciting probability distributions," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 680–701, 2005.

[238] S. R. Johnson, G. A. Tomlinson, G. A. Hawker, J. T. Granton, and B. M. Feldman, "Methods to elicit beliefs for bayesian priors: a systematic review," *Journal of clinical epidemiology*, vol. 63, no. 4, pp. 355–369, 2010.

[239] A. N. Sanborn, T. L. Griffiths, and R. M. Shiffrin, "Uncovering mental representations with markov chain monte carlo," *Cognitive psychology*, vol. 60, no. 2, pp. 63–106, 2010.

[240] G. O. Roberts and J. S. Rosenthal, "Optimal scaling for various Metropolis-Hastings algorithms," *Statistical Science*, vol. 16, pp. 351–367, Nov. 2001.

[241] M. Chmielewski and S. C. Kucker, "An MTurk crisis? Shifts in data quality and the impact on study results," *Social Psychological and Personality Science*, p. 1948550619875149, 2019.

[242] M. D. Lee and E.-J. Wagenmakers, *Bayesian cognitive modeling: A practical course*. Cambridge university press, 2014.

[243] J. Salvatier, T. V. Wiecki, and C. Fonnesbeck, "Probabilistic programming in python using pymc3," *PeerJ Computer Science*, vol. 2, p. e55, 2016.

[244] R. J. Heuer, *Psychology of intelligence analysis*. Center for the Study of Intelligence, 1999.

[245] B. T. and D. Doonan, "The growing burden of retirement rising costs and more risk increase uncertainty," *National Institute on Retirement Security*, 2020.

[246] M. Kay and J. Heer, "Beyond weber's law: A second look at ranking visualizations of correlation," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 469–478, 2016.

[247] P. Samuelson, "Risk and uncertainty: A fallacy of large numbers," 1963.

[248] R. C. Merton, "Lifetime portfolio selection under uncertainty: The continuous-time case," *The review of Economics and Statistics*, pp. 247–257, 1969.

[249] P. A. Samuelson, "Lifetime portfolio selection by dynamic stochastic programming," *Stochastic Optimization Models in Finance*, pp. 517–524, 1975.

[250] R. E. Hall, "Stochastic implications of the life cycle-permanent income hypothesis: theory and evidence," *Journal of political economy*, vol. 86, no. 6, pp. 971–987, 1978.

[251] N. G. Mankiw and S. P. Zeldes, "The consumption of stockholders and non-stockholders," *Journal of financial Economics*, vol. 29, no. 1, pp. 97–112, 1991.

[252] H. Landreth and D. C. Colander, *History of economic thought*. Houghton Mifflin College Division, 2002.

[253] J. Fox and A. Sklar, *The myth of the rational market: A history of risk, reward, and delusion on Wall Street*. Harper Business New York, 2009.

[254] S. Mullainathan and R. H. Thaler, "Behavioral economics," tech. rep., National Bureau of Economic Research, 2000.

[255] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk, econometrica, vol. 47, pp 263-291," 1979.

[256] A. Tversky and D. Kahneman, "Rational choice and the framing of decisions," *Journal of business*, vol. 59, no. S4, p. 251, 1986.

[257] R. Thaler and C. Sunstein, "Nudge: The gentle power of choice architecture," *New Haven, Conn.: Yale*, 2008.

[258] R. H. Thaler, A. Tversky, D. Kahneman, and A. Schwartz, "The effect of myopia and loss aversion on risk taking: An experimental test," *The Quarterly Journal of Economics*, vol. 112, no. 2, pp. 647–661, 1997.

[259] A. M. Hardin and C. A. Looney, "Myopic loss aversion: Demystifying the key factors influencing decision problem framing," *Organizational Behavior and Human Decision Processes*, vol. 117, no. 2, pp. 311–331, 2012. Publisher: Elsevier.

[260] R. H. Thaler and S. Benartzi, "Save more tomorrow: Using behavioral economics to increase employee saving," *Journal of political Economy*, vol. 112, no. S1, pp. S164–S187, 2004.

[261] A. R. Camilleri, M. Cam, and R. Hoffmann, "Nudges and signposts: The effect of smart defaults and pictographic risk information on retirement saving investment choices," *Journal of Behavioral Decision Making*, vol. 32, pp. 431–449, 2019.

[262] C. A. Looney and A. M. Hardin, "Decision Support for Retirement Portfolio Management: Overcoming Myopic Loss Aversion via Technology Design," *Management Science*, vol. 55, pp. 1688–1703, Oct. 2009.

[263] A. Savikhin, R. Maciejewski, and D. S. Ebert, "Applied visual analytics for economic decision-making," in *Visual Analytics Science and Technology, 2008. VAST'08. IEEE Symposium on*, pp. 107–114, IEEE, 2008.

[264] S. Rudolph, A. Savikhin, and D. S. Ebert, "Finvis: Applied visual analytics for personal financial planning," in *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 195–202, Citeseer, 2009.

[265] A. Savikhin, H. C. Lam, B. Fisher, and D. S. Ebert, "An experimental study of financial portfolio selection with visual analytics for decision support," in *System Sciences (HICSS), 2011 44th Hawaii International Conference on*, pp. 1–10, IEEE, 2011.

[266] S. Ko, I. Cho, S. Afzal, C. Yau, J. Chae, A. Malik, K. Beck, Y. Jang, W. Ribarsky, and D. S. Ebert, "A survey on visual analysis approaches for financial data," in *Computer Graphics Forum*, vol. 35, pp. 599–617, Wiley Online Library, 2016.

[267] A. C. Savikhin, "The application of visual analytics to financial decision-making and risk management: Notes from behavioural economics," in *Financial Analysis and Risk Management*, pp. 99–114, Springer, 2013.

[268] J. Gunaratne and O. Nov, "Informing and improving retirement saving performance using behavioral economics theory-driven user interfaces," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 917–920, 2015.

[269] A. Lusardi, A. Samek, A. Kapteyn, L. Glinert, A. Hung, and A. Heinberg, "Visual tools and narratives: New ways to improve financial literacy," *Journal of Pension Economics & Finance*, vol. 16, no. 3, pp. 297–323, 2017.

[270] Y. Zhang, R. K. Bellamy, and W. A. Kellogg, "Designing information for remediating cognitive biases in decision-making," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pp. 2211–2220, 2015.

[271] X. Yue, J. Bai, Q. Liu, Y. Tang, A. Puri, K. Li, and H. Qu, "sportfolio: Stratified visual analysis of stock portfolios," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 601–610, 2019.

[272] M. Greis, P. E. Agroudy, H. Schuff, T. Machulla, and A. Schmidt, "Decision-making under uncertainty: How the amount of presented uncertainty influences user behavior," in *Proceedings of the 9th Nordic Conference on Human-Computer Interaction*, pp. 1–4, 2016.

[273] R. H. Thaler and L. Ganser, *Misbehaving: The making of behavioral economics*. WW Norton New York, 2015.

[274] M. F. Jung, D. Sirkin, T. M. Gür, and M. Steinert, "Displayed uncertainty improves driving experience and behavior: The case of range anxiety in an electric car," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, (New York, NY, USA), pp. 2201–2210, Association for Computing Machinery, 2015.

[275] M. Wunderlich, K. Ballweg, G. Fuchs, and T. von Landesberger, "Visualization of delay uncertainty and its impact on train trip planning: A design study," *Computer Graphics Forum*, vol. 36, no. 3, pp. 317–328, 2017.

[276] G. Gigerenzer, "The psychology of good judgment: frequency formats and simple algorithms," *Medical decision making*, vol. 16, no. 3, pp. 273–280, 1996.

[277] L. Wilkinson, "Dot plots," *The American Statistician*, vol. 53, no. 3, pp. 276–281, 1999.

[278] C. Li, G. Baciu, and Y. Han, "Streammap: Smooth dynamic visualization of high-density streaming points," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 3, pp. 1381–1393, 2017.

[279] H. Wickham, M. Averick, J. Bryan, W. Chang, L. D. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, *et al.*, "Welcome to the tidyverse," *Journal of Open Source Software*, vol. 4, no. 43, p. 1686, 2019.

[280] M. Kay, "tidybayes: Tidy data and geoms for bayesian models," *R package version*, vol. 1, no. 0, 2019.

[281] S. Arora, R. Ge, and A. Moitra, "Learning topic models–going beyond svd," in *2012 IEEE 53rd annual symposium on foundations of computer science*, pp. 1–10, IEEE, 2012.

[282] D. U. Wulff, T. T. Hills, and R. Hertwig, "How short-and long-run aspirations impact search and choice in decisions from experience," *Cognition*, vol. 144, pp. 29–37, 2015.

[283] P. Jorion, *Value at risk: the new benchmark for managing financial risk.* The McGraw-Hill Companies, Inc., 2007.

[284] P. N. Kolm, G. Ritter, and J. Simonian, "Black–litterman and beyond: The bayesian paradigm in investment management," *The Journal of Portfolio Management*, 2021.

[285] F. J. Fabozzi, F. Gupta, and H. M. Markowitz, "The legacy of modern portfolio theory," *The Journal of Investing*, vol. 11, no. 3, pp. 7–22, 2002.

[286] Y.-J. Hu and S.-J. Lin, "Deep reinforcement learning for optimizing finance portfolio management," in *2019 Amity International Conference on Artificial Intelligence (AICAI)*, pp. 14–20, IEEE, 2019.

[287] M. Zhu, X. Zheng, Y. Wang, Y. Li, and Q. Liang, "Adaptive portfolio by solving multi-armed bandit via thompson sampling," *arXiv preprint arXiv:1911.05309*, 2019.

[288] Q. Zhu and V. Tan, "Thompson sampling algorithms for mean-variance bandits," in *International Conference on Machine Learning*, pp. 11599–11608, PMLR, 2020.

[289] A. V. Gonzalez, A. Rogers, and A. Søgaard, "On the interaction of belief bias and explanations," *arXiv preprint arXiv:2106.15355*, 2021.

[290] A. Karduni, R. Wesslen, D. Markant, and W. Dou, "Images, emotions, and credibility: Effect of emotional facial images on perceptions of news content bias and source credibility in social media," *arXiv preprint arXiv:2102.13167*, 2021.

[291] T. Chen, Z. Jiang, A. Poliak, K. Sakaguchi, and B. Van Durme, "Uncertain natural language inference," *arXiv preprint arXiv:1909.03042*, 2019.

[292] S. Zhang, C. Gong, and E. Choi, "Capturing label distribution: A case study in nli," *arXiv preprint arXiv:2102.06859*, 2021.

[293] W. Gantt, B. Kane, and A. S. White, "Natural language inference with mixed effects," *arXiv preprint arXiv:2010.10501*, 2020.

[294] T. L. Griffiths, F. Lieder, and N. D. Goodman, "Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic," *Topics in cognitive science*, vol. 7, no. 2, pp. 217–229, 2015.

[295] R. Wesslen, D. Markant, A. Karduni, and W. Dou, "Using resource-rational analysis to understand cognitive biases in interactive data visualizations," *arXiv preprint arXiv:2009.13368*, 2020.

[296] D. Marr, *Vision: A Computational Investigation into the Human Representation ond Processing of Visual Information.* WH Freeman, 1982.