Toward Personalized Visualizations

A dissertation

submitted by

Alvitta Ottley, B.S., M.S.

In partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Computer Science

TUFTS UNIVERSITY

May 2016

Advisor: Remco Chang

For Chizoba and Titan.

Toward Personalized Visualizations

Alvitta Ottley

Advisor: Remco Chang

As we enter the age of big and messy data, visualizations have emerged as an integral part of data analysis and decision making. Visualizations serve as an extension to the analysts cognition, and the computer supports the human by managing large amounts of data, performing computationally difficult tasks, and providing relevant feedback to the human. As a result, analysts are now able to solve historically challenging problems in many areas including health, business and military.

Unfortunately, no two humans are the same, and a single human's behavior may even change from time to time due to fatigue, emotional and situational changes. Despite the overwhelming evidence of the impact of individual differences, most systems are built using a one-size-fits-all approach. To better support the analysis, the computer must be able to adapt and adjust to its user, and intelligently tailor situationally appropriate information to the user in a timely manner. Such systems represent the next generation of user interfaces in which the user is not forced to adapt to the system's interface and design, but is free to interact with the system in ways that are natural and in tune with her individual needs. This dissertation works toward creating such next generation visualization systems by exploring: (1) how individual differences such as personality affect the ways people use visualizations to solve problems, (2) how we can automatically detect these traits from a user's interaction with visualization tools, and (3) how we can leverage users' individual cognitive differences in a mixed-initiative system.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We are all different. Our background, experience and culture all affect the way we think, behave and solve problems. Despite this fundamental truism, majority of our visualization tools are still being built with a one-size-fits-all approach. This traditional approach to visualization design has mainly focused on exploring how data is mapped to visual forms [9, 25, 108] and leveraging perceptual psychology to understand how people perceive these new designs [117, 156, 166, 167]. As a result, much like with many other user interfaces, visualization users are often forced to adapt their reasoning and analysis strategies to the tools' design.

This approach to visualization design in fast becoming inadequate [172]. In recent years, visualizations have emerged as an integral part of data analysis and decision-making. Visualizations now serve as an extension to the analyst's cognition. They are now being used to solve high-impact problems in many areas including health, business and military, and an analyst will typically interact one or more visualization to explore or reason about large and complex data. In order to support this process, it is ever important to know the user.

Realizing this, a new area of research has emerged in the Visualization community. Researchers have recently demonstrated how individual differences such as perceptual speed [3, 32, 30, 153, 154], spatial ability [24, 26, 90, 124, 162, 163, 175], experience [30, 39, 104] and personality [15, 59, 60, 122, 125, 174, 177] can significantly affect how users approach a problem and their ability to use visualization tools. Most importantly,

they highlight that there is no single visual design that suits every user. However, this research is still in its infancy and has primarily focused on correlating individual differences to users speed and accuracy when using visualization tools. While these do highlight differences between groups of users, they give us little insight into *how* or *why* groups of users differ. Many open questions remain. For instance, *Do user groups adapt different problem-solving strategies? How can we choose between existing tools?* and *How can we design new tools to better facilitate users' needs?*

The work in this dissertation aims to answer these questions and explores how visualization tools can better support a user's cognitive process. Knowledge of broad differences between user groups will guide the design for a specific tool and suggest customization options in a single system. This however is non-trivial as there has been mounting evidence that there is no single type of visualization user. Complicating this process further, a single user's behavior may even change from time to time due to fatigue, emotional and situational changes.

To better support the user, the computer must then be able to detect and adjust to its user, and intelligently tailor situationally appropriate information to the user in a timely manner. Such user adaptive systems represent the next generation of user interfaces in which the user is not forced to adapt to the system's interface and design, but is free to interact with the system in ways that are natural and in tune with her individual needs. Existing research suggests that it is possible to detect a visualization user. For instance, Gotz and Wen [58] showed that it is possible to infer a user's intent from their interactions to provide visualizations that are better suited for the detected task. Dou et al. [39] showed that user's strategies can be manually extracted with accuracies up to 80%. Work in this dissertation (Chapter 7) will show that it is possible to infer user attributes in real time.

Still, creating adaptive visualizations is an open challenge, and one important hurdle is deciding what to do once a users profile is detected. In general, there are two ways to think about adaptations: 1) the users profile is detected and the system performs real-time back-end adaptations, and 2) the users profile is detected and the system performs real-time front-end adaptations. In Chapter 8 we discuss a framework for designing personalized adaptive systems.

Altogether, the work presented in this dissertation aims to: (1) identify key individual differences that affect users performances with visualizations, (2) automatically detect individual differences by using machine learning techniques for user modeling, and (3) explore the viability of personalized adaptive visualization systems. Addressing these, I have conducted a series of interdisciplinary research projects that combine knowledge from various fields including: Information Visualization, Human-Computer Interaction, Machine Learning, Psychology and Health Decision Making. My research leverages the power of human computation to conduct large-scale analyses with the aim of solving current, real-world problems.

1.1 Outline

This dissertation is organized as follows:

- Chapter 2 reviews related work and sets the stage for the work presented in this dissertation. It includes an overview of prior work investigating the impact of individual differences on performance with visualization tools, as well as prior work on user modeling in the Visualization community.

- Chapter 3 provides empirical evidence for the importance of considering individual differences in a high impact real-world application domain. It demonstrates how spatial ability influences performance with visualizations for medical decision making.

- Chapter 4 explores other individual differences, and identifies the personality trait locus of control as a key factor for predicting performance on various hierarchical visualizations.

- Chapter 5 further examines the relationship between locus of control and performance with hierarchical visualizations. It demonstrates how psychological priming can be used to temporarily manipulate users' locus of control to impact their performance.

- Chapter 6 examines how individual differences impact user search strategies. It explains *how* and *why* groups of users differ through a manual analysis of interaction

log data.

- Chapter 7 builds on the work in Chapter 6 by demonstrating how we can use machine learning techniques to automatically infer user attributes. It also investigates the viability of real-time adaptation.

- Chapter 8 introduces a model of individual cognitive differences and then discusses techniques for designing adaptive visualizations.

- Chapter 9 summarizes the key contributions of this dissertation.

Chapter 2

Background

In recent years, an overwhelming body of research has demonstrated how individual differences impact people's ability to use information visualization and visualization systems [15, 24, 26, 32, 30, 31, 60, 59, 162, 163], and a growing number of researchers have advocated for better understanding of these effects [172, 176]. They demonstrate that the cognitive differences between users may result in dissimilarities in the way information is internalized and thus the way users approach problems.

2.1 Individual Differences and Visualization

Much of the existing work in the Visualization field on the impact of individual differences has focused on the influence of perceptual ability on visualization use. For instance, Conati and Maclaren [32] found that a users perceptual speed influences performance using two difference Geographical Information Systems. They found that users with high perceptual ability performed better with star graphs while users with low perceptual speed performed better with radar graphs.

Allen [3] too found a connection between perceptual speed and system effectiveness. He found that perceptual speed was a significant predictor of spatial scanning ability in search performance. His findings suggest that users often fail to optimize their visualization use for greater search efficiency, and his work calls for the development of user models to automatically guide users toward optimal strategies. Similarly, Conati et al. [30] found

that a number of individual differences, including perceptual speed, influences performance across different layouts of interactive ValueCharts. Using eye tracking technology, Toker et al. [154] demonstrated how perceptual speed and verbal working memory influences interactions with bar graphs and radar graphs. Other work by Toker et al. [153] found that perceptual speed not only impacts performance on visualization tools but also preference for difference tools.

In addition to performance, a number researchers have also explored the impact of individual differences on preference for visual designs. For instance, Nov and Ye [120] found that the *resistance to change* personality trait can be a significant predictor of the perceived ease of use of a visual interface. Peck et al. [128] used a brain imaging technology called functional near-infrared spectroscopy (fNIRS) to explore how a user's cognitive load can affect their performance with visualizations. They measured users' brain activity as they made comparisons using both bar graphs and pie charts and found that half of their participants had lower cognitive load with pie charts, and the other half had lower cognitive load with bar graphs. Their brain activity predictions also matched their users' perceived ease of use.

These findings have begun to build the case that some individual differences in visual analytics can be attributed to a complex interaction between the mental model suggested by a visual layout and the users own cognitive style. However, most previous studies that show an effect of individual differences either examine a single system or compare two or more systems with numerous differences. In order to apply knowledge of individual differences to visualization design, it is necessary to make a clearer connection between personality groups and exactly which factors lead them to better performance with one visualization over another. Context for these connections can come from psychology research on how personality relates to problem solving and decision making.

2.2 Personality and Problem Solving

Problem solving is the process by which we bridge the gap between the perceived and a desired outcome. Since this process often requires complex thought, researchers have

long been investigating the effect of personality on problem solving and decision making processes. These investigations have identified a series of personality traits that impact problem solving including extraversion, neuroticism, and locus of control. For example, when approaching a problem, persons who score high on the judgment scale of the Myers-Briggs Personality Inventory (i.e., those who are decisive and are quick at making decisions) prefer a problem to be concise and well structured [112]. On the other hand, persons who score high in the perceptive scale are more concerned with seeing all sides of a problem and prefer flexibility.

Disposition on the extraversion scale also affects problem solving [112]. Individuals with high extraversion have greater tendencies to be sociable and require engagement with others. When given a problem, extraverts are more likely to discuss the problem in order to ascertain clarity and understanding, while introverts are more likely to take time and think about the problem before they begin. This implies that introverts would generally take longer before attempting a solution but are more likely to have a well-defined path toward finding that solution. Furthermore, introverts are more likely to take time to understand important concepts while extraverts require feedback on the correctness of their ideas.

While extraversion and judgment can be used as a predictor of a users approach to a problem, similar tendencies can also be deduced from an individuals neuroticism orientation. Neuroticism measures a persons degree of emotional stability. Individuals who score high on the neuroticism scale are more prone to experiencing negative emotions such as stress and anxiety. One study by Uziel [160] has investigated the correlation of neuroticism and extraversion and suggests a negative correlation between them when affective states are considered. An individuals neuroticism disposition can also be a predictor of their problem-solving approach [23]. Persons with high measures of neuroticism are more like to have lower perceived problem-solving skills and are less likely to make decisions when risks are involved. Farley [46] found a curvilinear correlation between neuroticism and time spent solving a problem. People with high levels of neuroticism also take the most time to solve problems, and individuals with average neuroticism are significantly faster on problem-solving tasks than the combined low and high scoring people.

Several studies have also established correlations between locus of control and neu-

roticism [85, 165]. Both traits are multidimensional constructs and are comprised of similar traits such as anxiety and self-esteem. It is therefore not surprising that locus of control may also affect problem solving ability and style. Of the personality dimensions explored in this thesis, locus of control has been found to have the most pronounced correlation to problem solving using visualizations.

2.3 Locus of Control

Locus of control (LOC) [140] measures the degree to which an individual feels in control of, or controlled by external events. Using the Rotter construct [140], individuals are scored on a 23-point scale where the two extreme ends of the scale are categorized as internal locus of control and external locus of control. Persons who are internally oriented on the scale (Internals) believe that external events are contingent upon their own actions while persons who are externally oriented on the scale believe that events are controlled by powerful beings.

The impact of LOC has been explored for many decades beyond the Visualization and HCI communities, and research suggests that the differences between Internals and Externals are quite vast. Internals tend to have a strong sense of self-efficacy allowing them to take control even when faced with difficult problems. Conversely, Externals believe that they have no control over external events, making them far more likely to adapt to situations. However, because of this perceived lack of control, Externals are also more likely to give up when faced with difficulty.

Past research corroborates this. Internality has been shown to correlate with increased effectiveness at work [84], better academic performance [47] and greater ability to cope with stress [4]. LOC also affects learning style. Cassidy and Eachus [20] showed that Internals are more likely to practice deep learning, while Externals are more likely to practice surface learning. This implies that there is a correlation between locus of control and general problem solving techniques which suggests a potential effect of locus of control on problem solving using visualizations.

In the medical community, LOC has been shown to affect patients' recovery out-

comes. Fisher and Johnstion [48] found that users with external LOC were more likely to become discouraged and give in to their disability. In the visualization community, Green and Fisher [59] showed a significant correlation between LOC and users' speed and accuracy when using hierarchical visualizations. They reported that locus of control can be used to predict completion times as well as insight when using the two visualization tools.

2.4 Spatial Ability

Another main factor that have been shown to influence visualization use is *spatial ability*. Spatial ability in general refers to the ability to mentally represent and manipulate two- or three-dimensional representations of objects. Spatial ability is a cognitive ability with a number of measurable dimensions, including spatial orientation, spatial visualization, spatial location memory, targeting, disembedding and spatial perception [90, 162]. People with higher spatial ability can produce more accurate representations and maintain a reliable model of objects as they move and rotate in space.

There is considerable evidence that these abilities affect how well a person can reason with abstract representations of information, including visualizations. For example, Tversky et al. [159, 158] examined how individual differences in ability affects the extraction of structure and function from diagrams. They showed that participants with high spatial ability create mental models that integrate both structure and function, while participants with low spatial ability form mental models where structure is separate from function. Vicente et al. [163] found that low spatial ability corresponded with poor performance on information retrieval tasks in hierarchical file structures. They found that in general high spatial ability users were two times faster than low spatial ability users and that low spatial ability users were more likely to get lost in the hierarchical file structures.

Chen and Czerwinski [24] found that participants with higher spatial ability employed more efficient visual search strategies and were better able to remember visual structures in an interactive node-link visualization. Velez et al. [162] tested users of a three-dimensional visualization and discovered that speed and accuracy were dependent on several factors of spatial ability. Similarly, Cohen and Hegarty [26] found that users'

spatial abilities affects the degree to which interacting with an animated visualization helps when performing a mental rotation task, and that participants with high spatial ability were better able to use a visual representation rather than rely on an internal visualization.

This body of research shows that users with higher spatial ability are frequently more effective at using a variety of visualizations. Taken together, they suggest that high spatial ability often correlates with better performance on tasks that involve either searching through spatially arranged information or making sense of new visual representations. Additionally, there is evidence that high spatial ability makes it easier to switch between different representations of complex information. Ziemkiewicz and Kosara [175] tested users' ability to perform search tasks with hierarachy visualizations when the spatial metaphor implied in the task questions differed from that used by the visualization. Most participants performed poorly when the metaphors conflicted, but those with high spatial ability did not. This confirms findings that spatial ability plays a role in understanding text descriptions of spatial information [37].

2.5 User Modeling

The goal of inferring information about a user based on their interactions is generally referred to as "user modeling". The first attempts at user modeling trace back to the late 1970's and were targeted at improving dialog systems so that they are more conversational [27, 129, 164]. Over the past decades, user modeling has grown into an active area of research in a variety of domains such as Artificial Intelligence, Data Mining, Machine Learning and Human-Computer Interaction [1, 11, 18, 94, 95, 113]. Research in this area not only includes techniques for learning about users, but it also encompasses adaptation strategies. Although the field is broad, the goal of this dissertation is narrower in that it specifically emphasizes learning users' reasoning processes and characteristics with visualization tools, as opposed to modeling users' tasks.

Perhaps most related to the focus of this dissertation is the research in the domain of *analytic provenance* in the visual analytics community [136]. Researchers in analytic provenance believe that the analysis process during an analytic task is just as important as

the analysis product [119]. Through analyzing a user's interactions, researchers in analytic provenance seek to identify *how* a user discovers insight and how the same procedures can be stored and reapplied to automatically solve other similar problems [170, 86].

Many systems have been developed in the visual analytics community for logging, storing, and analyzing a user's interactions and activities. For example, the GlassBox system by Cowley et al. [36] records low-level events generated by the interface (such as copy, paste, window activation, etc.). At a higher level, VisTrails captures the user's steps in a scientific workflow [8]. Finally, at a model level, Endert et al. showed that user interactions can be analyzed systematically, and directly used to perform model steering operations [44]. Similarly, other researchers have demonstrated that a user's interactions can be used to infer parameters of analytical models, which can then be presented visually [14, 55, 170]. The work in this dissertation shares synergistic activities with these prior works in that it also seeks to extract higher-level information from low-level user interactions. However, the goal of this dissertation is to develop techniques that can automatically classify users on their characteristics.

Much of the existing work in the visual analytics community on connecting the ways users solve problems with their cognitive abilities has been based on eye tracker data [5, 100, 107, 151]. For example, Lu et al. demonstrated how eye gaze data can be used to determine important or interesting areas of renderings and automatically select parameters to improve the usability of a visualization system [107]. Steichen et al. explored the use of eye tracking data to predict visualization and task type [151]. With varying degrees of accuracy they were able to predict: (1) a user's cognitive traits: personality, perceptual speed and visual working memory, (2) the difficulty of the task, and (3) the visualization type. Similarly, Toker et al. [154] demonstrated how perceptual speed and verbal working memory influences interactions with bar graphs and radar graphs. These findings are particularly important for visual analytics tasks as previous research has shown that users' cognitive traits can be used as predictors of speed and accuracy [59]. Instead of using eye gaze data, in this work we forgo specialized sensors and analyze mouse interactions[1].

[1]Recent work suggests that mouse movements in some interfaces are strongly correlated with eye movements [33, 73]

In the HCI community, Gajos et al. developed the SUPPLE system that can learn the type and degree of a user's disability by analyzing mouse interaction data and generate dynamic and personalized interfaces for each specific user [51]. Although the intended scenario is in the domain of accessibility, the approach and methods developed by Gajos et al. can be generalized to other interfaces as well. In the web usage mining community, researchers have used click stream data for modeling and predicting users' web surfing patterns [41, 96, 97, 150]. Some of the techniques developed for these web mining applications could be adapted to extend work in the dissertation.

Chapter 3

Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability

This chapter is based on the paper:

> **Alvitta Ottley**, Evan M Peck, Lane T Harrison, Daniel Afergan, Caroline Ziemkiewicz, Holly A Taylor, Paul KJ Han, and Remco Chang. Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability. In *Visualization and Computer Graphics, IEEE Transactions on*, 22(1) pages 529-538. 2016. [124]

One advantage of visualization is that it can make complex or abstract concepts easier to grasp by making them visible. In theory, an effective visual representation can make traditionally difficult problems more concrete and easier to understand. In practice, however, nuanced contextual factors heavily impact the effectiveness of visualizations, making the right representation difficult to establish.

In order to build systems that can better facilitate a users' cognitive processes, we must first understand *which* individual differences can impact reasoning with visualizations and *how*. In this chapter, we begin by exploring a real-world decision making problem to il-

lustrate the practical impact of this dissertation. We investigate how problem representation can influence reasoning with medical statistics and how such reasoning can be mediated by cognitive abilities. The findings presented in this chapter have direct implications for medical risk communication.

3.1 Introduction

As the medical field transitions toward evidence-based and shared decision making, effectively communicating conditional probabilities to patients has emerged as a common challenge. To make informed health decisions, it is essential that patients understand health risk information involving conditional probabilities and Bayesian reasoning [53]. However, understanding such conditional probabilities is challenging for patients [40]. Even more alarming, the burden of communicating complex statistical information to patients is often placed on physicians even though studies have shown that most struggle with accurate estimations themselves [40].

Still, both physicians and patients make life-critical judgments based on conditional probabilities. Deficits in diagnostic test sensitivity and specificity (intrinsic characteristics of the test itself) can lead to false negative and false positive test results which do not reflect the actual state of an individual. For low-prevalence diseases, even a highly specific test leads to false positive results for a majority of test recipients. Unless a patient fully understands the uncertainties of medical tests, news of a negative result can lead to false reassurance that treatment is not necessary, and news of a positive result can bring unjust emotional distress.

Consider the following mammography problem [56]:

"The probability of breast cancer is 1% for women at age forty who participate in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography.

A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?"

Misinterpretation of this and other medical test statistics can have serious adverse consequences such as under- and over-diagnosis [115, 169, 168] or even death. However, there are currently no effective tools for mitigating this problem. Despite decades of research, the optimal methods for improving interpretation of diagnostic test results remain elusive, and the available evidence is sparse and conflicting.

Prior work indicates that visualizations may be key for improving performance with Bayesian reasoning problems. For example, results from Brase [13] and work by Garcia-Retamero and Hoffrage [54] suggest that visual aids such as Euler diagrams and icon arrays hold promise. Researchers have also explored visualizations such as decision trees [50, 109], contingency tables [28], "beam cut" diagrams [56] and probability curves [28], and have shown improvements over text-only representations. However, when researchers in the visualization community extended this work to a more diverse sampling of the general population, they found that adding visualizations to existing text representations did not significantly increase accuracy [114, 123].

Given the contradictory findings of prior research, we aim to identify factors that influence performance on Bayesian reasoning tasks. We hypothesize that these discrepancies are due to differences in problem representations (textual or visual), as well as the end users' innate ability to reason through these problems when using visualizations. In particular, we propose that the phrasing of text-only representations can significantly impact comprehension and that this effect is further confounded when text and visualization are incorporated into a single representation. Furthermore, motivated by prior work [24, 88, 162, 163, 175], we also hypothesize that individual differences (i.e., spatial ability) are mediating factors for performance on Bayesian reasoning tasks.

To test our hypotheses, we conducted two experiments to investigate how problem representation and individual differences influence performance on Bayesian reasoning tasks. The first experiment focused on text-only representations and how phrasing can impact accuracy, while the second explores how individual differences and representations that combine text and visualization affect performance.

With Experiment 1 we show that wording can significantly affect users' accuracy

and demonstrate how probing[1] can help evaluate different representations of Bayesian reasoning problems. Combining techniques that have previously been tested independently, our results show an increase in the accuracy of the mammography problem from the previously reported 6% [114] to 42%. Our findings demonstrate how the phrasing of a Bayesian problem can partially explain the poor or inconsistent results of prior work and provide a baseline text-only representation for future work.

In Experiment 2, we tested six different representations including a new text-only representation that uses indentation to visually depict set relations (*Structured-Text*), a storyboarding visualization that progressively integrates textual information with a frequency grid visualization (*Storyboarding*), and a visualization-only representation (*Vis-Only*). The results of our second experiment show that altering the amount of information shown in text and visualization designs can yield accuracies as high as 77%. However, we found that adding visualizations to text resulted in no measurable improvement in performance, which is consistent with prior work in the visualization community by Micallef et al. [114].

Examining our study population further, we found that spatial ability impacts users' speed and accuracy on Bayesian reasoning tasks, with high spatial ability users responding significantly faster and more accurately than low spatial ability users. Analyzing accuracy with respect to spatial ability, we discovered that users with high spatial ability tend to perform better than users with low spatial ability across all designs, achieving accuracies from 66%-100%. We discuss the implications of these findings for text and visualization design, and how these methods may impact the communication of conditional probability in the medical field and beyond.

We make the following contributions to the understanding of how phrasing, visualizations and individual differences influence Bayesian reasoning:

- We identify key factors that influence performance on Bayesian reasoning tasks and explain the inconsistent and conflicting findings of prior work.
- We show that the phrasing of textual Bayesian reasoning problems can significantly affect comprehension, and provide a benchmark text-only problem representation

[1] Instead of asking a single question (typically the true positive rate), you ask a series of questions designed to guide the user through their calculations.

that allows future researchers to reliably test different designs of Bayesian reasoning problems.

- We demonstrate that a user's spatial ability impacts their ability to solve Bayesian reasoning problems with different visual and textual representations. Our findings provide guidance on how to design representations for users of varying spatial ability.

3.2 Background

There is a substantial body of work aimed at developing novel, more effective methods of communicating Bayesian statistics. Still, there is no authoritative method for effectively communicating Bayesian reasoning, and prior results are inconsistent at best. Below we survey some of these findings.

Gigerenzer and Hoffrage [56] in their seminal work explored how text-only representations can be improved using natural frequency formats. They explored the use of phrases such as *96 out of 1000* instead of *9.6%*, hypothesizing that natural frequency formats have greater perceptual correspondence to natural sampling strategies [56]. Their findings demonstrate that using natural frequency significantly improves users' understanding of Bayesian reasoning problems.

A series of studies has also been conducted to investigate the efficacy of using visualizations to aid reasoning. Various types of visualizations have been tested, including Euler diagrams [13, 89, 114], frequency grids or icon arrays [54, 89, 114, 123, 144], decision trees [50, 144, 149], "beam cut" diagrams [56], probability curves [28], contingency tables [28, 29] and interactive designs [155]. While some researchers have compared several visualization designs [13, 114, 123], many of these visualizations were proposed and tested separately. It is still not clear which best facilitates Bayesian reasoning.

For instance, recent work by Garcia-Retamero and Hoffrage [54] investigated how different representations (text versus visualization) affect the communication of Bayesian reasoning problems to both doctor and patients. They conducted an experiment where half of the participants received natural frequency formats and the other half received percentages. A further division was made within these groups; half of the participants received

the information in numbers while the other half were presented with a visualization (a frequency grid). Their results confirmed the prior results of Gigerenzer and Hoffrage [56] showing that users are more accurate when information is presented using natural frequency formats. With their visualization condition, and they were able to achieve overall accuracies of 62%, one of the highest reported accuracies.

Work by Brase [13] compared various visualizations for communicating Bayesian reasoning. In a comparative study, he analyzed participants' accuracies when three different visualizations (icon arrays, Euler diagrams and discretized Euler diagrams [2]) were added to textual information. Like natural frequency formats, discrete items represented by the icon array were expected to correspond with humans' perception of natural sampling, thus improving Bayesian reasoning. In contrast, Euler diagrams were expected to enhance the perception of the nested-set relations that are inherent in Bayesian reasoning problems. The discretized Euler diagram was designed as a hybrid of the two.

Brase found that icon arrays had the best overall accuracy rate (48%), suggesting that they best facilitate Bayesian reasoning. However there were some inconsistencies with the visualization designs used by Brase [114]. For instance, the Euler diagram was not area proportional but the hybrid diagram was, and the number of glyphs in the hybrid diagram differed from the number of glyphs in the frequency grid [114]. Noticing this, researchers in the visualization community extended this work by designing a new, consistent set of visualizations and surveying a more diverse study population.

Micallef et al. [114] used a combination of natural frequency formats, icon arrays and Euler diagrams to improve the designs of Brase [13]. Instead of surveying university undergraduates, they recruited crowdsourced participants via Amazon's Mechanical Turk in an effort to simulate a more diverse lay population [114]. Their study compared 6 different visualization designs and found no significant performance differences among them. Reported accuracies for the 6 designs ranged from 7% to 21% and the control condition (a text-only representation with natural frequency formats) yielded an overall accuracy of only 6%.

[2] These are Euler diagrams with discrete items. They were designed as hybrid diagrams that combine both the natural sampling affordance of icon arrays and the nested-set relations affordance of traditional Euler diagrams.

They also found no statistically significant performance difference between the control text-only condition and any of the conditions with visualization designs. This finding implies that simply adding visual aids to existing textual representation did not help under the studied conditions. Their follow up work adds yet another dimension. In a second experiment, they reported significant improvements in the accuracies for their visualization conditions when numerical values for the text descriptions were removed. This finding suggests that presenting both text with numerical values and visualization together may overload the user and result in incorrect inferences.

The findings of Micallef et al. [114] suggest a possible interaction between textual information and visualization when representing Bayesian reasoning problems. One possible explanation for this interaction is that both the mental model required to interpret the textual information in a Bayesian reasoning problem and the mental model required to understand a visualization can compete for the same resources [88]. As more information is presented, a user's performance can degrade since more items will be held in the user's spatial working memory [68]. In addition to explaining the inconsistencies among prior work by exploring different wording and visualization representations, this chapter aims to understand how spatial ability mediates performance on Bayesian reasoning problems.

3.2.1 Spatial Ability and Visualizing Bayesian Reasoning

In Bayesian reasoning domain, Kellen [88] found that spatial ability was relevant to the understanding of visualizations of Bayesian reasoning. He used Euler diagrams and investigated how problem complexity (the number of relationships presented in an Euler diagram) impacts users' performance. His findings suggest that spatial ability may moderate the effect of visualizations on understanding. However, his work only investigated spatial ability as it relates to Bayesian reasoning and the number of relationships depicted in an Euler diagram.

Still, like the prior reported accuracy findings, the reported results on the effects of spatial ability on understanding Bayesian reasoning have been contradictory. Micallef et al. [114] too investigated the effects of spatial ability. They compared six text and visualization conditions and one text-only condition but found no significant effect of participants'

spatial abilities.

3.3 Research Goals

The body of existing work presented in this chapter paints a complex portrait of visualization and Bayesian reasoning. First, the results of the prior works are inconsistent. The reported accuracies of the baseline text-only conditions differed significantly: Brase [13] reported 35.4%, Garcia-Retamero and Hoffrage [54] reported 26% while Micallef et al. [114] reported accuracies of only 6%. Second, prior work suggests an interaction between textual information and visualization when they are combined into a single representation.

In order to progress this important area of research, we must first identify factors that affect a user's ability to extract information from text and visualization representations of Bayesian reasoning problems. Thus, the primary research goal for this work is to disambiguate the discrepancies among prior works' results. We hypothesize that the observed discrepancies among prior work are largely due to differences in problem representations. In particular, we hypothesize that the phrasing of text-only representations impacts comprehension. Furthermore, we posit that while visualizations can be effective tools for communicating Bayesian reasoning, simply appending visualizations to complex textual information will adversely impact comprehension.

In the succeeding sections, we present the results of two experiments that were designed to investigate the interacting effect of both problem representation and spatial ability on communicating Bayesian reasoning problems. Together, these experiments address the question of how users make sense of Bayesian reasoning problems under different, and sometimes competing, representations of complex information. Our first experiment establishes a baseline, text-only condition and investigates how various forms of problem phrasing impacts accuracies. With our second experiment, we explore the interaction between textual information and visualizations when they are combined in a single representation, and the effect of users' spatial ability on their performance.

3.4 Experiment 1: Text-Only Representations

A survey of the prior work reveals many inconsistencies among Bayesian problems used for assessing Bayesian reasoning. Many past experiments have used their own Bayesian problems, with differing scenarios, wordings and framings. For instance, in their work, Gigerenzer and Hoffrage [56] used 15 different Bayesian problems, each with differing real-world implications and potential cognitive biases associated with them (e.g. being career oriented leads to choosing a course in economics, or carrying heavy books daily relates to a child having bad posture). Micallef et al. [114] and Garcia-Retamero and Hoffrage [54] each used three different Bayesian problems (with only one in common). Brase [13] used a single Bayesian problem not previously tested by other researchers.

Of the existing work, Brase [13] reported the highest accuracies for his text-only condition, with 35.4% of his participants reaching the correct Bayesian response. In addition to using natural frequencies, Brase [13] used **probing** as a means of evaluating the effectiveness of representations. Probing is a technique by which a series of questions are posed to the user that are designed to guide that user through the calculation process. Cosmides and Tooby [35] proposed that probing can be used to help users uncover information that is necessary for solving Bayesian inference problems and thereby improves performance. Rather than asking the participant to calculate the true positive rate from the given information directly (the task that is traditionally given), they used probing to guide their participants' Bayesian calculations. Ultimately, probing was designed to assess whether the user understands the information as it is presented instead of their mathematical skills. Following Brase [13], we examined probing as one of our study conditions.

In his study, Brase [13] also used a **narrative** - a generalizable, hypothetical scenario. Instead of presenting information about a specific disease such as breast cancer, he presented a fictional narrative, introducing a population in which individuals are exposed to a new disease (*"Disease X"*). By using a hypothetical population and a generic disease name, we hypothesize that this generalizes the problem and may have mitigated biases related to a certain disease, thus impacting accuracies.

In addition to these two techniques (probing and narrative), we adapted framing

principles for reducing the complexity of text representations [35, 157]. Prior studies suggest that **framing** can significantly impact decision making with probability problems [35]. For example, saying *10 out of 100 people will have the disease* versus *90 out of 100 people will not have the disease* can elicit very different responses [157], and presenting both frames can help mitigate biases known as framing effects [157]. Using both frames also has the advantage that it explicitly states relationships in the problem that are implicit in the original text.

3.4.1 Design

In line with our research goals of disambiguating contradictory results in previous research, our first experiment examines how these three techniques (*framing*, adding a *narrative* and using *probing*) can be combined to reduce the complexity of Bayesian reasoning problems. In the context of Bayesian problems, the term complexity can have different meanings, for instance: the number of relationships in the problem [88], the number of steps needed to solve the problem, or the amount of information to be integrated or reconstructed [35]. In the current work, we define complexity as the difficulty of extracting information. This hinges on the notion that the simplicity of a task partly depends on the *how* the information is presented. We believe that is it important first to establish a baseline text representation (i.e. **no visualization**) before we consider the effect of adding visualizations.

We conducted an online study and tested three different text-only representations of Bayesian reasoning problems:

Text$_{orig}$ For our base condition, we chose the mammography problem (see Table 3.1 $Text_{orig}$) since it has been used in many studies [40, 56, 54, 88, 114] and tests a skill of great importance and generalizability [40]. This specific mammography problem was used by Gigerenzer and Hoffrage [56] and Micallef et al. [114], and includes the base rate, true positive rate, and false positive rate. The expected answer was *8 out of 103*.

Text$_{probe}$ The information presented in this condition is exactly the same as $Text_{orig}$ but uses probing instead of asking for the true positive rate directly (see Table 3.1 $Text_{probe}$).

Table 3.1: The three questions used in Experiment 1

Text_{orig}	10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. *Here is a new representative sample of women at age forty who got a positive mammography in routine screening.* *How many of these women do you expect to actually have breast cancer? ____ out of ____*
Text_{probe}	10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10 women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. *Imagine 1000 people are tested for the disease.* *(a) How many people will test positive? ____* *(b) Of those who test positive, how many will actually have the disease? ____*
$\text{Text}_{diseaseX}$	There is a newly discovered disease, Disease X, which is transmitted by a bacterial infection found in the population. There is a test to detect whether or not a person has the disease, but it is not perfect. Here is some information about the current research on Disease X and efforts to test for the infection that causes it. There is a total of 1000 people in the population. Out of the 1000 people in the population, 10 people actually have the disease. Out of these 10 people, 8 will receive a positive test result and 2 will receive a negative test result. On the other hand, 990 people do not have the disease (that is, they are perfectly healthy). Out of these 990 people, 95 will receive a positive test result and 895 will receive a negative test result. *Imagine 1000 people are tested for the disease.* *(a) How many people will test positive? ____* *(b) Of those who test positive, how many will actually have the disease? ____*

The participant is first probed for the expected number of people who will be tested positive (*103*) then she is probed for the true positive count (*8*). The two probed questions used in the current design are:

Positive Count How many people will test positive?

True Positive Count Of those who test positive, how many will actually have the disease?

Text$_{diseaseX}$ For this condition, we adopted a narrative similar to Brase [13] for the mammography problem, and similarly to $Text_{probe}$ we used probing. In addition to using these two techniques, this condition also provides the user with both positive and negative frames of the problem. Instead of only providing the base rate, the true positive rate, and the false positive rate, the text included the true negative and the false negative rates (see Table 3.1 $Text_{diseaseX}$). It is important to note that no *new* data was added. The true negative and false negative rates were implicit in the $Text_{orig}$ and $Text_{probe}$ conditions.

3.4.1.1 Participants

We recruited 100 online participants (37 for $Text_{orig}$, 30 for $Text_{probe}$ and 33 for $Text_{diseaseX}$) via Amazon's Mechanical Turk. Mechanical Turk is a virtual market place that allows individuals or companies to post small jobs for completion and where people can easily work for a small remuneration. There are over 100,000 "workers" from around the world registered with this service, who are able to choose jobs from a pool and are upon completion [139].

For this reason, Mechanical Turk is attractive to Human-Computer Interaction and Visualization researchers as it facilitates the recruitment of a more diverse study population [67]. One of the biggest concerns when using this tool is the possibility of participants

Table 3.2: Demographics for Experiment 1

N	100
Gender	Female: 35%, Male: 65%
Education	High School: 13%, College: 42%, Graduate School: 25%, Professional School: 17%, Ph.D.: 1%, Postdoctoral: 2%
Age	$\mu : 33.63, \sigma : 11.8$, Range: 19 - 65

randomly clicking through the study only to be paid. Researchers have compensated for this by providing bonuses for correct responses and participants are paid only upon completing the study [99]. This has been proven effective and has been adapted for our study.

Participants received a base pay of $.50 with a bonus of $.50 and $.05 for each correct answer on the main task and surveys respectively. The total possible renumeration was $2.80 which is comparable to the U.S. minimum wage. Participants completed the survey via an external link and performed tasks using an online experiment manager developed for this study. Using this tool, participants were regulated by their Mechanical Turk worker identification number and were only allowed to complete the experiment once. Table 3.2 summarizes the demographic information for Experiment 1.

3.4.1.2 Procedure

After selecting the task from the Mechanical Turk website, participants were instructed to navigate to a specified external link. Once there, they entered their Mechanical Turk worker identification number which was used both for controlling access to the experiment manager and for remuneration. After giving informed consent, participants were randomly presented with one of the three Bayesian reasoning problems. They were instructed to take as much time as needed to read and understand the information provided as they would have to answer questions based on the information and that bonuses will be paid for each correct answer. To separate the time spent reading the question from the time spent actually solving the problem, the question was not visible until they clicked the appropriately labeled button to indicate that they were ready. The participants were once again instructed to take as much time as needed and enter the answers in the space provided. The timer ended when the participant entered an answer. Any edits to their responses extended the recorded time. Once they submitted the main task, they completed a short demographic questionnaire.

3.4.2 Results

For our analysis, responses were only deemed correct if participants entered the expected response for **both** probed questions.

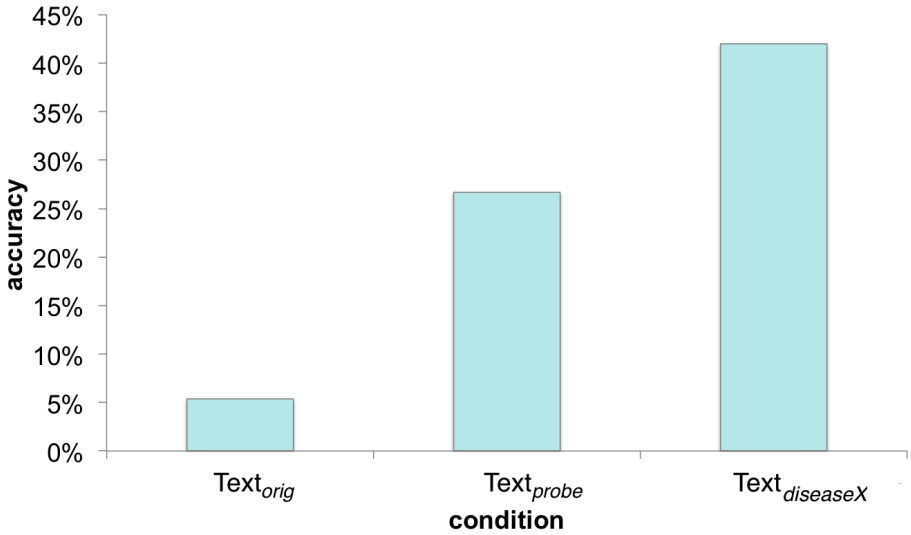


Figure 3.1: Accuracies across all conditions in Experiment 1. Combining probing and narrative techniques proved to be effective for reducing the overall complexity of the text and increasing accuracy.

With the $Text_{orig}$ condition, we successfully replicated prior results of Micallef et al. [114] with an accuracy rate of 5.4% as compared to their reported 6% for text-only representations. Modifying the original question by using probing ($Text_{probe}$), we presented participants with questions that were easier to understand [35]. Consistent with prior work by Cosmides and Tooby [35], we found that this small change yielded a significantly higher accuracy rate of 26.7%.

Finally, by changing the problem text with our $Text_{diseaseX}$ condition, we successfully replicate Brase's [13] results with an accuracy rate of 42.4% as compared to his reported 53.4%. A chi-square test was conducted across all participants and revealed significant differences between the accuracy rates of the three conditions($\chi^2(2, N = 100)$=13.27 $p = 0.001$). Performing a pairwise chi-square test with a Bonferroni adjusted alpha ($\alpha = 0.017$), we found significant differences between $Text_{orig}$ and $Text_{probe}$ ($\chi^2(1, N = 67) = 5.9$, $p = 0.015$), and $Text_{orig}$ and $Text_{diseaseX}$ ($\chi^2(1, N = 70) = 13.56$, $p < 0.001$).

3.4.3 Discussion

In our first experiment, we found that by simply changing how the problem was presented, we observed an improvement in participants' overall accuracy from 5.4% to 42.4%. We

adapted techniques such as probing, which nudges the user to think more thoroughly about the problem, adding a narrative which generalized the problem, and presenting both frames for mitigating framing effects.

Taken together this gives us insight into how lexical choices of text-only representations of Bayesian reasoning problems govern their effectiveness and may at least partially explain the poor or inconsistent accuracies observed in previous work. By using probing alone, our results showed a significant improvement over our base condition which used direct questioning. This suggests that assessment techniques for Bayesian reasoning problems should be thoroughly scrutinized.

Participants were even more accurate when the stimulus combined all three techniques (probing, narrative and framing). This finding provides initial evidence that even with text-only representations (i.e. without visualization aids), the phrasing of the problem can impact comprehension.

Indeed, there were several factors that potentially contributed to the increase in communicative competence we observed for $Text_{diseaseX}$. For example, using the generic term *Disease X* instead of a specific disease may have mitigated biases introduced by the mammography problem. Alternatively, the observed increase in accuracy could be attributed to the overall readability of the text or the amount of data presented in the conditions (the $Text_{diseaseX}$ condition presented the user with slightly more explicit data than the $Text_{orig}$ and $Text_{probe}$ conditions). Deciphering these was beyond the scope of this project, but will be an important direction for future work.

In the following study, we further address our research goals by investigating the effect of adding visualizations for representing Bayesian reasoning tasks. We use our results from this initial experiment by adopting $Text_{diseaseX}$ as a baseline text-only representation for evaluating different text and visualization designs.

Table 3.3: Table showing the 6 conditions used in Experiment 2

Control-Text

There is a total of 100 people in the population. Out of the 100 people in the population, 6 people actually have the disease. Out of these 6 people, 4 will receive a positive test result and 2 will receive a negative test result. On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy). Out of these 94 people, 16 will receive a positive test result and 78 will receive a negative test result.

Complete-Text

There is a total of 100 people in the population. Out of the 100 people in the population, 6 people actually have the disease. Out of these 6 people, 4 will receive a positive test result and 2 will receive a negative test result. On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy). Out of these 94 people, 16 will receive a positive test result and 78 will receive a negative test result.

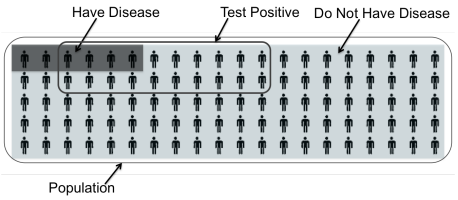
Another way to think about this is... Out of the 100 people in the population, 20 people will test positive. Out of these 20 people, 4 will actually have the disease and 16 will not have the disease (i.e., they are perfectly healthy). On the other hand, 80 people will test negative. Out of these 80 people, 2 will actually have the disease and 78 will not have the disease (i.e., they are perfectly healthy).

Structured-Text

There is a total of 100 people in the population.
Out of the 100 people in the population,
6 people actually have the disease. Out of these 6 people,
4 will receive a positive test result and
2 will receive a negative test result.
On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy). Out of these 94 people,
16 will receive a positive test result and
78 will receive a negative test result.

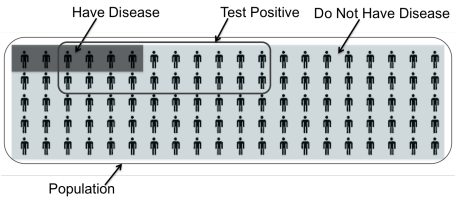
Another way to think about this is...
Out of the 100 people in the population,
20 people will test positive. Out of these 20 people,
4 will actually have the disease and
16 will not have the disease (i.e., they are perfectly healthy).
On the other hand, 80 people will test negative. Out of these 80 people,
2 will actually have the disease and
78 will not have the disease (i.e., they are perfectly healthy).

Vis-Only



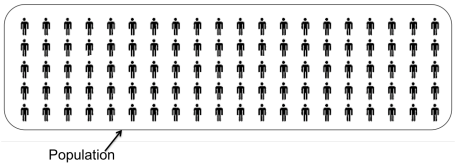
Control+Vis

There is a total of 100 people in the population. Out of the 100 people in the population, 6 people actually have the disease. Out of these 6 people, 4 will receive a positive test result and 2 will receive a negative test result. On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy). Out of these 94 people, 16 will receive a positive test result and 78 will receive a negative test result.

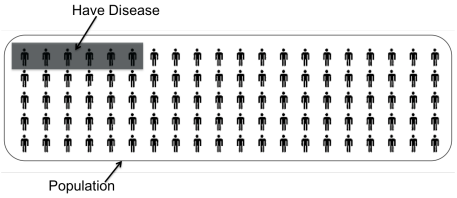


Storyboarding

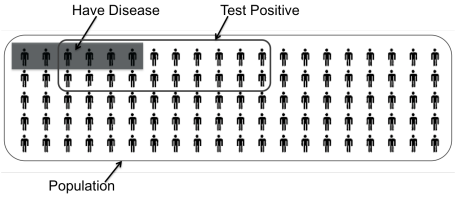
There is a total of 100 people in the population.



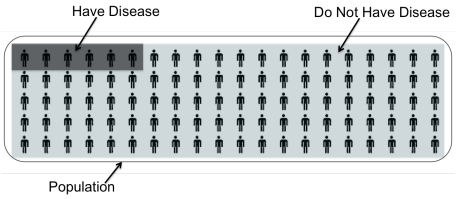
Out of the 100 people in the population, 6 people actually have the disease.



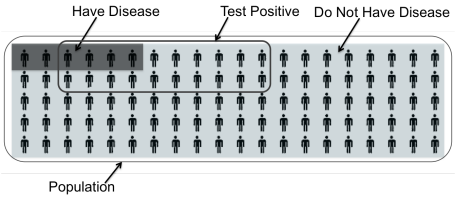
Out of these 6 people, 4 will receive a positive test result and 2 will receive a negative test result.



On the other hand, 94 people do not have the disease (i.e., they are perfectly healthy).



Out of these 94 people, 16 will receive a positive test result and 78 will receive a negative test result.



3.5 Experiment 2: Text and Visualization

Although visualization has been suggested as a solution to the Bayesian reasoning problem, recent findings suggest that, across several designs, simply adding visualizations to textual Bayesian inference problems yields no significant performance benefit [114, 123]. Micallef et al. [114] also found that removing numbers from the textual representation can improve performance. The findings of this prior work suggest an interference between text and visualization components when they are combined into a single representation.

Differing from prior work which focused mainly on comparing different visualization designs [114], our second experiment aimed to progress Bayesian reasoning research by further investigating the effect of presenting text and visualization together. We examined the amount of information presented to the user and the degree to which the textual and visual information are integrated. Grounded by the baseline condition established in Experiment 1 (Table 3.3 **Control-Text**), we tested representations that gradually integrate affordances of visualizations or the visualization itself.

One affordance of visualizations is that relationships that are implicitly expressed in text are often explicated in visual form. Visualizations make it easier to "see" relationships among groups. To bridge this information gap, we gradually expanded the text-only representation to explicate implied information and relationships.

Secondary to our main research goals and motivated by the prior work demonstrating a connection between spatial ability and visual design [89, 175], our second experiment also aimed to understand *how* nuances in spatial ability affect users' capacity to use different representations of Bayesian reasoning problems. Since prior research suggests that low spatial-ability users may experience difficulty when both the text and visual representations are presented [175], we hypothesize that low spatial-ability users would be more adept at using representations which integrated affordances of the visualization but not the visualization itself. On the other hand, we hypothesize that high spatial-ability users will benefit greatly from representations which merge textual and visual forms, as they are more likely to possess the ability to effectively utilize both representations.

3.5.1 Design

To test our hypotheses, we present the Text$_{diseaseX}$ condition from Experiment 1, using a variety of representations. Our intent was to manipulate the total amount of information presented, as well as the coupling between the problem text and visual representation. Consistent with Text$_{diseaseX}$, each condition in Experiment 2 began with an introductory narrative:

> There is a newly discovered disease, Disease X, which is transmitted by a bacterial infection found in the population. There is a test to detect whether or not a person has the disease, but it is not perfect. Here is some information about the current research on Disease X and efforts to test for the infection that causes it.
>
> The format of the questions asked were also consistent with the Text$_{diseaseX}$:
>
> (a) How many people will test positive? ____
>
> (b) Of those who test positive, how many will actually have the disease? ____

3.5.1.1 Conditions

There were a total of 6 conditions which were randomly assigned to our participants (see Table 3.3 for the exact stimuli).

Control-Text As the name suggests, this is our control condition and uses the same text format as was presented in the **Text**$_{diseaseX}$ condition of Experiment 1.

Complete-Text In this condition, the text is expanded to present all possible relationships and framings of the problem, which is a common affordance of visualizations. It is important to note that the text still presents the same amount of information as *Control-Text* (i.e. the base rate, the true positive rate, the false positive rate, the false negative rate and the true negative rate), however, it presents the data both with respect to having the disease and being tested positive (see Table 3.3 *Complete-Text*).

Structured-Text Here, we further improve the text by integrating another affordance of visualizations. Like *Complete-Text*, the text in this condition enumerates all possible

relationships and framings of the problem, however, we enhanced the text by adding visual cues to the representation. Instead of using long-form paragraphs, we used indentation to clarify relationships. Similar to spreadsheets of raw data, the spatialization of the information makes the relationships more apparent.

Vis-Only With this condition, we establish a baseline for using visualizations. With the exception of the introductory narrative mentioned above, there is no additional text in this condition.

While researchers have investigated numerous visualization designs for representing Bayesian reasoning (see Section 3.2), there is still no consensus on which is best. In fact, recent research in the visualization community comparing the effectiveness of 6 different visualization designs found no significant difference between them [114].

That said, for this visualization-only condition, we chose to represent the information using an icon array visualization (see Table 3.3). A number of researchers have explored their utility for risk communication and Bayesian reasoning [13, 52, 53, 54, 63, 66, 89, 114, 123, 144] and icon-arrays are often used in the medical community for representing risk information.

The specific icon-array used in this work consists of a 5 by 20 grid of anthropomorphic figures. We adapted a sequential layout for the different sets (as opposed to a random layout which has been previously used for representing uncertainty [63]) and we used in-place labeling for ease of reference. This is similar to the design used by Brase [13].

Control+Vis Mirroring prior work [13, 114, 123] that investigated the utility of adding visualization designs to Bayesian problems, here we simply added a visualization to our control text-only representation. The information for this condition is represented using both the *Control-Text* description and the icon array visualization from *Vis-Only*.

Storyboarding This condition was designed to simplify Bayesian reasoning by gradually integrating the textual and visual components of *Control+Vis*. Such storytelling tech-

Table 3.4: Demographics for Experiment 2

N	377
Gender	Female: 34.2%, Male: 65%, Unspecified: .8%
Education	High School: 22.3%, College: 51.5%, Graduate School: 19.6%, Professional School: 4.2%, Ph.D.: 1.6%, Postdoctoral: .5%, Unspecified: .3%
Trained$^+$	Yes: 12.2%, No: 87.5%, Unspecified: .3%
Age	$\mu : 31, \sigma : 9.87$, Range: 18 to 65
Spatial Ability	$\mu : 8.60, \sigma : 5.25$, Range: -3.75 to 20
Numeracy	$\mu : 4.23, \sigma : 1.22$, Range: 0 to 6

$^+$Participants received statistics training

niques are becoming increasingly popular in recent years [75, 76, 145] and have even been recently referred to as "the next step for visualization" [98].

For our *Storyboarding* design, no information was added, but the information is presented sequentially, allowing for temporal processing. The text shown in this condition is exactly the same as *Control-Text* and the final visualization is the same as *Vis-Only*.

3.5.1.2 Cognitive Ability Measures

We measured participants' spatial ability using the paper folding test (VZ-2) from Ekstrom, French, & Hardon [42]. This survey consists of 2 3-minute sessions with 10 questions each. A similar version of the test has been used as a standard technique to compare spatial ability to Bayesian reasoning skills in other studies [88, 114]. Consistent with prior work [88, 114], a participant's spatial ability score was calculated by summing the number of correct answers minus the total number of incorrect answers divided by four.

Consistent with prior studies investigating the effectiveness of Bayesian reasoning problem representations [54, 89, 114], we measured participants' numerical skills. This was measured using Brown et al.'s 6-question test [16]. Prior research has demonstrated a correlation between numerical skills and understanding natural frequencies [22] and numerical skills has been shown to correlate with one's ability to understand medical risk information [16, 138].

3.5.1.3 Participants

We recruited 377 participants (61-65 per condition) via Amazon's Mechanical Turk who had not completed Experiment 1. The recruitment and remuneration techniques used for this experiment follows that of Experiment 1 (see Section 3.4.1.1). Table 3.4 summarizes our participants' demographics.

3.5.1.4 Procedure

The procedure for this experiment also follows that of Experiment 1 (see Section 3.4.1.2) except for the following changes. After the main tasks, in addition to the demographics survey, participants completed VZ-2 to measure spatial ability and the numerical skill survey.

3.5.2 Hypotheses

Following our high-level claims that text complexity, visual representation, and spatial ability affect accuracy on Bayesian reasoning, we form the following hypotheses:

H1 Prior work shows no performance increase when visualizations are simply appended to existing text representations [114]. Therefore, removing the textual description completely from the visualization condition will mitigate this effect and *Vis-Only* will be more effective than *Control-Text* and *Control+Vis*.

H2 Since participants are given an increased amount of information in the text, users will perform better on *Structured-Text* and *Complete-Text* than on *Control-Text*.

H3 High spatial-ability users will perform better overall than low spatial-ability users. In the absence of a visualization, we expect that high spatial-ability users will be better able at mentally constructing a visual representation than low spatial-ability users. Additionally, we expect that high spatial-ability users will more easily create connections between the text and visual representations when they are presented together than low spatial-ability users.

H4 Designs that include both text and vis (*Storyboarding, Control+Vis*) will require higher spatial ability than text-only designs (*Complete-Text, Structured-Text*) and *Vis-Only*.

3.5.3 Results

While recent work [114] has advocated for a more continuous or fine-grained approach to assessing users' accuracy on Bayesian reasoning tasks (for instance, reporting the differences between users' responses and correct answers in terms of a log ratio), we report our accuracies in terms of the percentages of correct exact answers. Choosing this binary approach of assessing accuracy has two advantages: (1) it allows us to directly compare our results across the prior body of work as they have all (including [114]) reported their accuracies similarly, and (2) this course-grained approach is especially user-friendly for comparing representations with substantial accuracies as seen in the subsequent sections.

Consistent with Experiment 1, the proceeding analyses focus only on participants who answered both questions correctly (see Section 3.5.1 for the exact questions asked). In an effort to further simulate a lay population, our analysis excluded participants who reported to have had statistical training.

3.5.3.1 Accuracy Across Designs

Across all conditions, the average accuracy was remarkably high; 63% of the participants correctly answered both questions. Figure 3.2 summarizes the accuracies across all conditions. *Complete-Text*, *Structured-Text* and *Vis-Only* yielded the highest overall accuracies ranging from 71% to 77%. Along with *Control-Text*, *Storyboarding* yielded the lowest overall accuracies with only 51% and 49% respectively of the participants responding correctly to the questions.

We performed a chi-square analysis to test for differences in accuracy across the six conditions. The test revealed that the percentage of participants who correctly answered both questions differed by design ($\chi^2(5, N= 330) = 17.2$, $p = 0.004$). We then performed all pairwise chi-square tests with a Bonferroni adjusted alpha ($\alpha = 0.003$) to identify the specific designs that deferred. The analysis revealed a significant differences between only *Storyboarding* and *Structured-Text* ($\chi^2(1, 114) = 8.8$, $p < 0.003$).

We found no significant difference in accuracy between *Control-Text* and *Vis-Only*, and we found no difference in accuracy among *Control-Text*, *Complete-Text* and *Structured-*

Text. Consistent with prior findings [114, 123], we also found no significant difference between the *Control-Text* and *Control+Vis* conditions. This suggests that under the studied conditions, using visualizations (with or without textual information) and increasing the amount of explicit textual information in text-only designs did not improve performance, thereby rejecting both **H1** and **H2**.

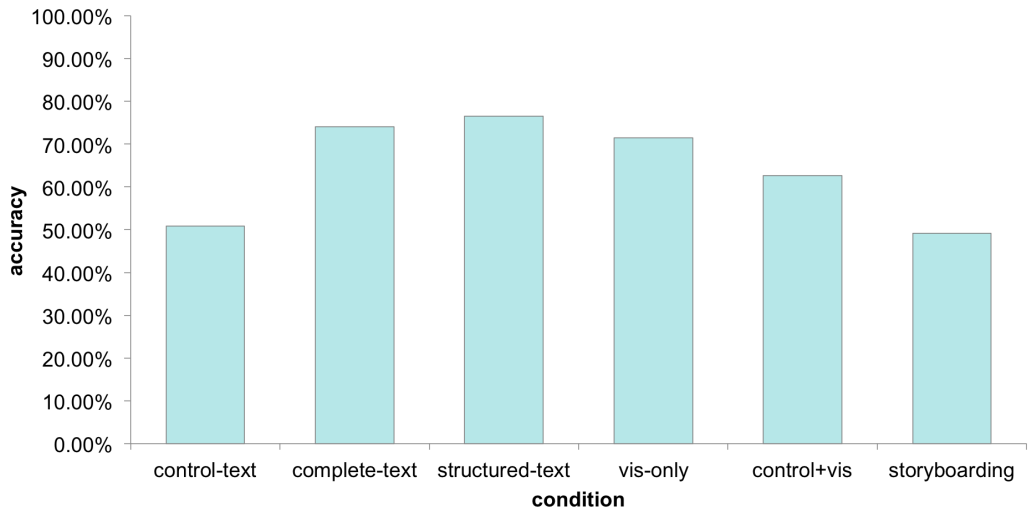


Figure 3.2: Accuracies across all conditions. We found that participants were most accurate with *Structured-Text*, *Complete-Text* and *Vis-Only*.

3.5.3.2 Spatial Ability and Accuracy

To test our hypothesis that spatial ability affects participants' capacity to extract information from the different representations, we performed a binary logistic regression to predict participants who correctly answered both questions using their spatial abilities score as a predictor. A test of the resulting model against the constant model was statistically significant at $p < 0.001$, indicating that spatial ability highly correlates with participants' ability to answer the questions accurately. Prediction success overall was 71.5% (87.1% for predicting those who responded correctly and 44.6% for predicting those who responded incorrectly).

For a more specific analysis, we split users into two groups (spatial_{low} and spatial_{high}) based on a median split of their spatial abilities scores ($\text{spatial}_{low} < 9$, N = 170 and $\text{spatial}_{high} >= 9$, N = 160). Confirming **H3**, the overall accuracy for the spatial_{high} group was 78.8% while

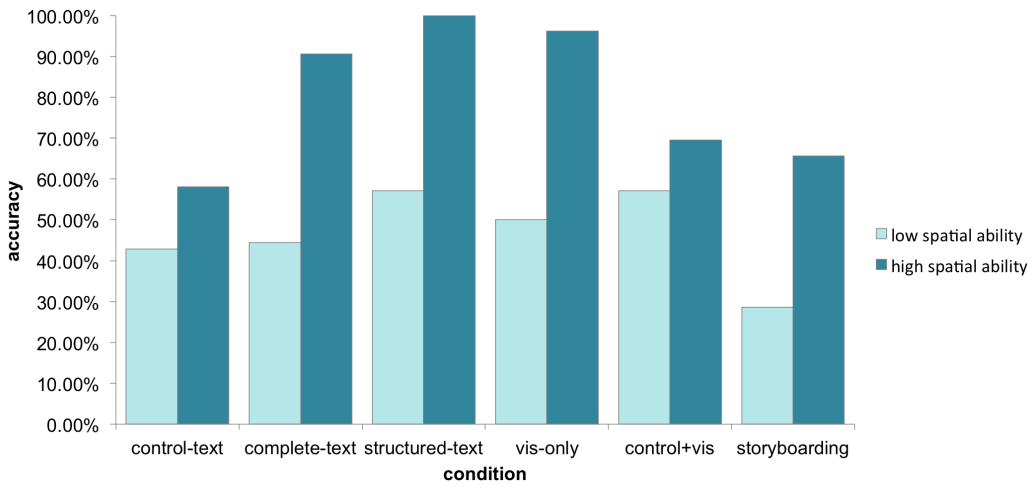


Figure 3.3: Average accuracy for the low and high spatial ability groups for each design. Overall, we found that high spatial users were much more likely to correctly answer the question prompts.

spatial$_{low}$ was 46.9%. Figure 3.3 summarizes the groups' accuracies for each of the six conditions.

We then performed separate chi-square analyses testing for significant differences between the accuracies of the six conditions for the spatial$_{low}$ group and the spatial$_{high}$ group. The chi-square test for the spatial$_{high}$ group was significant ($\chi^2(5, \text{N} = 170) = 26.3$, $p < 0.001$) and multiple comparisons with the Bonferroni adjusted alpha ($\alpha = 0.003$) revealed significant differences between:

- *Control-Text* & *Complete-Text* ($\chi^2(1, N{=}63) = 8.8$, $p < 0.003$)
- *Control-Text* & *Structured-Text* ($\chi^2(1, N{=}54) = 12.7$, $p < 0.001$)
- *Control-Text* & *Vis-Only* ($\chi^2(1, N{=}57) = 11.07$, $p < 0.001$)
- *Structured-Text* & *Storyboarding* ($\chi^2(1, N{=}65) = 13.5$, $p < 0.001$)

These results indicate that for the spatial$_{high}$ group, *Structured-Text*, *Complete-Text*, and *Vis-Only* resulted in improved performance over our control condition (*Control-Text*), confirming **H2** and partially supporting **H1**. More generally, the results imply that spatial ability must be considered when evaluating the effectiveness of Bayesian reasoning designs.

Performing a similar analysis with the spatial$_{low}$ group, we found no significant difference between the accuracies for the six conditions ($\chi^2(5, N{=}160) = 6.5$, $p = 0.262$). This

indicates that the accuracies for the spatial$_{low}$ group were similar across the all conditions, suggesting that the proposed designs were ineffective for low spatial-ability users.

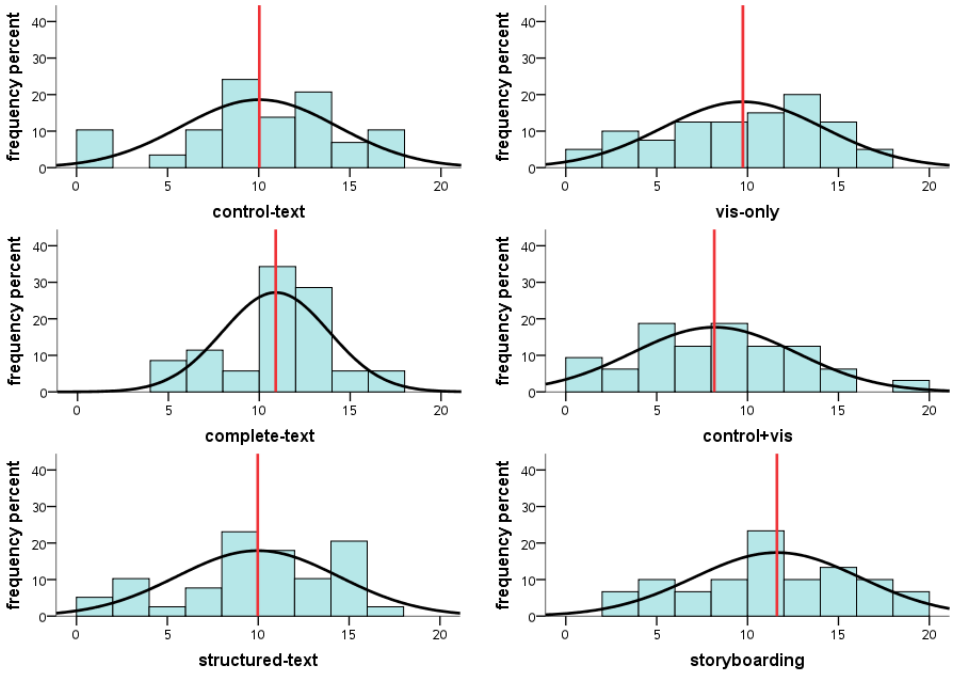


Figure 3.4: Histograms showing the distribution of spatial ability scores for participants who correctly answered both questions across the six conditions. The graphs provide preliminary evidence that *Complete-Text* and *Storyboarding* may require higher spatial ability to use them.

3.5.3.3 Spatial Ability Across Designs

Given our findings that participants' spatial ability affects their likelihood of correctly answering the question prompts using a given representation, we hypothesize that we can now use spatial ability as a tool for ranking representations based on their complexity. In particular, we use spatial ability as a proxy for measuring and comparing the extraneous cognitive load necessary to effectively use each representation. Figure 3.4 shows the distribution for spatial ability scores for the correct users on the six conditions.

Prior to our analysis, we removed two outliers whose spatial ability score was more that two standard deviations from the mean score for their respective conditions. We then conducted a one-way Analysis of Variance (ANOVA) to test for differences between the spatial ability scores of participants who correctly answered both questions for each of the six conditions. Our model was statistically significant ($F(5, 206) = 2.57$, $p = 0.028$)

suggesting that the spatial ability scores differed across conditions.

Post hoc comparisons using Fisher's least significant difference (LSD) indicated that the mean scores of the following conditions differed significantly:

- *Control-Text* & *Control+Vis* ($p = 0.042$)
- *Complete-Text* & *Control+Vis* ($p = 0.005$)
- *Storyboarding* & *Control+Vis* ($p = 0.002$)

This finding supports our hypothesis that some representations may require higher spatial ability to use them. However, we partially reject **H4**. *Control+Vis* had the lowest average indicating that this representing may be most suitable for users with lower spatial ability. Conversely, we found that the average spatial ability score for correct users on *Storyboarding* was higher than all other conditions, suggesting that *Storyboarding* was the most difficult representation to use.

3.5.4 Discussion

The results of our Experiment 1 demonstrated how phrasing of Bayesian problems can influence performance. In Experiment 2, we examined whether we could improve problem representations by enhancing text or combining it with visualization. In an effort to bridge the information gap between text and visual representations, we studied text-only representations that clarified information that usually is more easily seen in a visualization. Our *Complete-Text* design sought to decrease this information gap by enumerating all probability relationships in the text and our *Structured-Text* design used indentations to visualize these relationships. Still, when spatial ability was not considered, we found that adding more information did not benefit users.

We observed similar results with our visualization conditions. Although we hypothesized that the *Vis-Only* design would be more effective than *Control-Text* and *Control+Vis*, our results did not support this hypothesis. Again, when spatial ability was not considered, adding visualizations (with or without textual information) did not improve performance. However, a closer examination of our results adds nuance to this finding when individual differences are considered.

3.5.4.1 Spatial Ability Matters

While the lack of overall difference across conditions was unexpected, factoring in the effect of spatial ability helped shed light on these findings. Across all visualizations, spatial ability was a significant indicator of accuracy and completion times. We found that users with low spatial ability generally performed poorly; the accuracy of high spatial-ability users was far higher than the accuracy of low spatial-ability users (78.8% v. 46.9%). Relative to the *Control-Text* condition, for high spatial users, the *Structured-Text, Complete-Text* and *Vis-Only* designs were extremely effective, yielding accuracies of 100%, 90% and 96% respectively. These unprecedented accuracies suggest that, for users with high spatial ability, these designs can solve the problem of Bayesian reasoning. However, it is interesting to note that effective designs were "pure" designs (i.e., they did not combine text and visualizations). This finding contradicts prior work in Psychology which demonstrates a multimedia advantage (i.e. providing both text and visual representation) for comprehension and memory [17].

3.5.4.2 Text+Vis Interference

For high spatial-ability users, we found that representations that combined text and visualization (*Control+Vis, Storyboarding*) actually impeded users' understanding of conditional probability when compared to text-only (*Complete-Text, Structured-Text*) or *Vis-Only* conditions. Despite the fact that high spatial-ability users performed comparatively poorly with the *Control+Vis* design (accuracy decreased by nearly 30% when compared to *Complete-Text, Structured-Text*, and *Vis-Only*), such disparity in accuracy was not observed with low spatial ability users using *Control+Vis*. One possible explanation relies on considering the problem as a mental modeling task. Users with low spatial ability may have simply chosen the representation in *Control+Vis* (text or visualization) that best fit their understanding of the problem. On the contrary, high spatial-ability users may have attempted (and failed) to integrate the text and visualization representations in order to find the correct answer. This hypothesis would be in line with Kellen's [89] hypothesis that text and visual representations in a complex problem may compete for the same mental resources, increasing

the likelihood of errors.

The *Storyboarding* design proved to be an enormous obstacle for the user. Performing analysis to investigate the spatial ability scores required to successfully extract information from the six designs revealed that *Storyboarding* demand higher spatial ability scores than the other designs. While it is intended to gradually guide users through the Bayesian reasoning problem, the different steps may have inadvertently introduced distractors to the information that the user is truly looking for and/or forced users into a linear style of reasoning that was incongruent with their mental model of the problem. This added complexity increased cognitive load to a point that accuracy for all users suffered.

Still, such storytelling techniques have been shown to be effective for communicating real world data [75, 76, 98, 145]. The tasks in this study, however, go beyond typical information dissemination, as users had to understand information known to be inherently challenging for most people. Future work could investigate the utility of storytelling techniques for similar reasoning tasks.

3.6 Summary

Effectively communicating Bayesian reasoning has been an open challenge for many decades, and existing work is sparse and sometimes contradictory. In this chapter we presented results from two experiments that help explain the factors affecting how text and visual representations contribute to performance on Bayesian problems. With our first experiment, we showed that the wording of text-only representations can significantly impact users' accuracies and may partly be responsible for the poor or inconsistent findings observed by prior work.

Our second experiment examined the effects of spatial ability on Bayesian reasoning tasks and analyzed performance with a variety of text and visualization conditions. We found that spatial ability significantly affected users ability to use different Bayesian reasoning representations. Compared to high spatial-ability users, low spatial-ability users tended to struggle with Bayesian reasoning representations. In fact, high spatial-ability users were almost two times more likely to answer correctly than low spatial-ability users. Addition-

ally, we found that text-only or visualization-only designs were more effective than those which blend text and visualization.

Ultimately, our results not only shed light on how problem representation (both in text phrasing and combining text and visualization) can affect Bayesian reasoning, but also question whether one-size-fits-all visualizations are ideal. Further study is needed to clarify how best to either adapt visualizations or provide customization options to serve users with different needs. The results from these studies can be used for real-world information displays targeted to help people better understand probabilistic information. They also provide a set of benchmark problem framings that can be used for more comparable future evaluations of visualizations for Bayesian reasoning. Further work in this domain can have significant impact on pressing issues in the medical communication field and other domains where probabilistic reasoning is critical.

3.7 Dataset

To facilitate future work, participants' data are made available at: http://github.com/TuftsVALT/Bayes.

Chapter 4

Personality and Performance on Visualizations

This chapter is based on the paper:

> Caroline Ziemkiewicz, **Alvitta Ottley**, R. Jordan Crouser, Ashley Rye Yauilla, Sara L. Su, William Ribarsky, and Remco Chang. How Visualization Layout Relates to Locus of Control and Other Personality Factors. In *Visualization and Computer Graphics, IEEE Transactions on*, 19(7) pages 1109-1121. 2013 [177]. This is joint work with Caroline Ziemkiewicz. My specific contributions included: performed data analysis, and co-wrote paper.

The work presented in Chapter 3 provides empirical evidence of the impact of individual differences on a real-world reasoning task with visualizations. It showed that a user's spatial ability is a mediating factor, and that spatial ability can be a used as a predictor of performance across different representations of a Bayesian reasoning problem. Furthermore, we saw that none of the tested conditions were ideal for all users, highlighting the importance of tailoring designs to individuals. In this chapter, we extend this work by investigating other individual differences that may impact performance on visualizations.

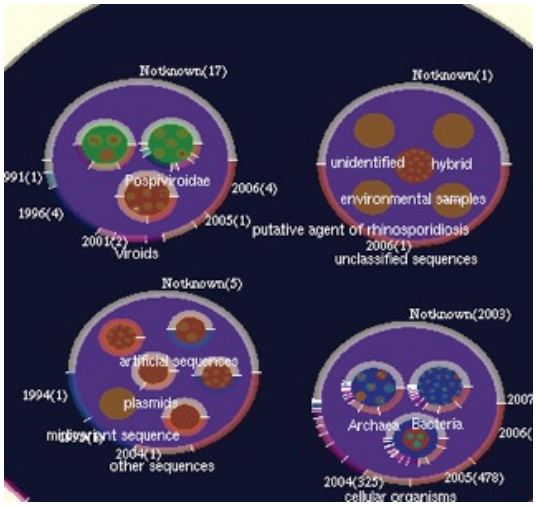
4.1 Introduction

Recent research indicates that personality plays an important role in performance when using a visualization system [59] and that measures of some personality traits can serve as predictors of a user's willingness to adapt their mental model to various visual metaphors [175]. What these findings underscore is the importance of considering individual differences such as users' personality when designing new visualizations. Given the increasing number of known individual differences that affect performance [59, 60, 65, 105, 174], it is likely that one or more can cause a user to perform less optimally on a visual design.

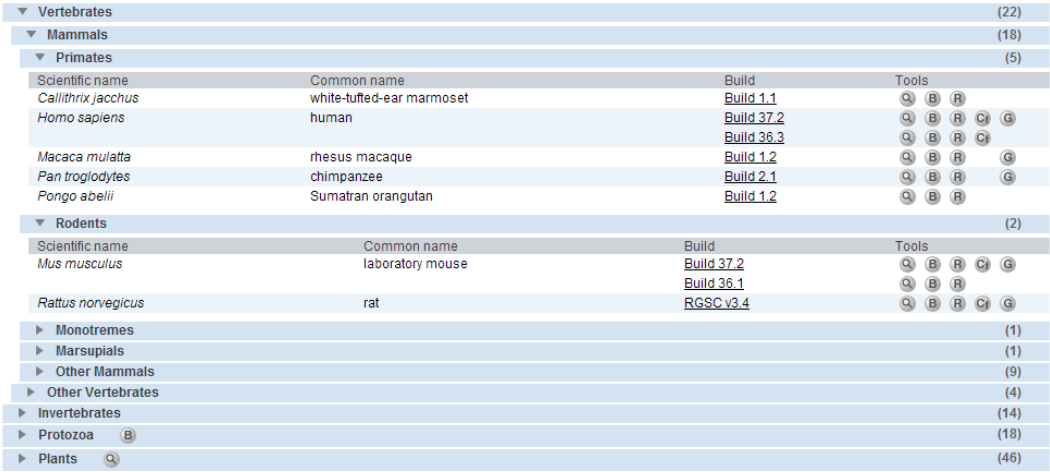
In this chapter, we investigate the cognitive trait locus of control (LOC) the its effect on user performance with varying data visualization layouts. Recall that locus of control [140] measures the degree to which a person attributes outcomes to themselves (internal locus of control) or to outside forces (external locus of control). Green and Fisher [59] originally found a correlation between locus of control and performance on two real-world data exploration systems. However, the two systems differed on many dimensions, including the use of color, labeling, interaction, and layout style, therefore making it difficult to generalize the findings.

We hypothesize that layout style is the key variable that determines the interaction between locus of control and compatibility with different system designs. The definition of layout here encompasses any differences in the spatial arrangement and presentation of marks in a visualization. This is to be distinguished from differences in the visual encoding, that is, how individual data variables are mapped to individual graphical variables, such as color or size.

To test our hypothesis, we conducted on online study with 240 subjects who varied in their locus of control. Participants are presented with four variations of a hierarchy visualization showing phylogenetic data (Figure 4.2). These designs include a view that employs a list-like organizational structure (V1), a view that presents the hierarchy in a strong containment metaphor (V4), and two designs that lie between these extremes (V2 and V3). We hypothesize that users with a more external locus of control are more willing to adapt their thinking to unfamiliar visual metaphors than those with an internal locus of



(a) GVis



(b) NCBI Map Viewer

Figure 4.1: The two interfaces used in Green et al.'s study [59] of personality differences in visual analytics use.

control. We show how locus of control can predict performance on inferential task questions using these interface designs. Specifically, we test the hypothesis that an individual with a more internal locus of control will show a performance decrease when using layouts with a strong containment metaphor, while those with a more external locus of control will not show this decrease. Our findings can inform the design of visualization interfaces adapted to an individuals needs.

4.2 Background

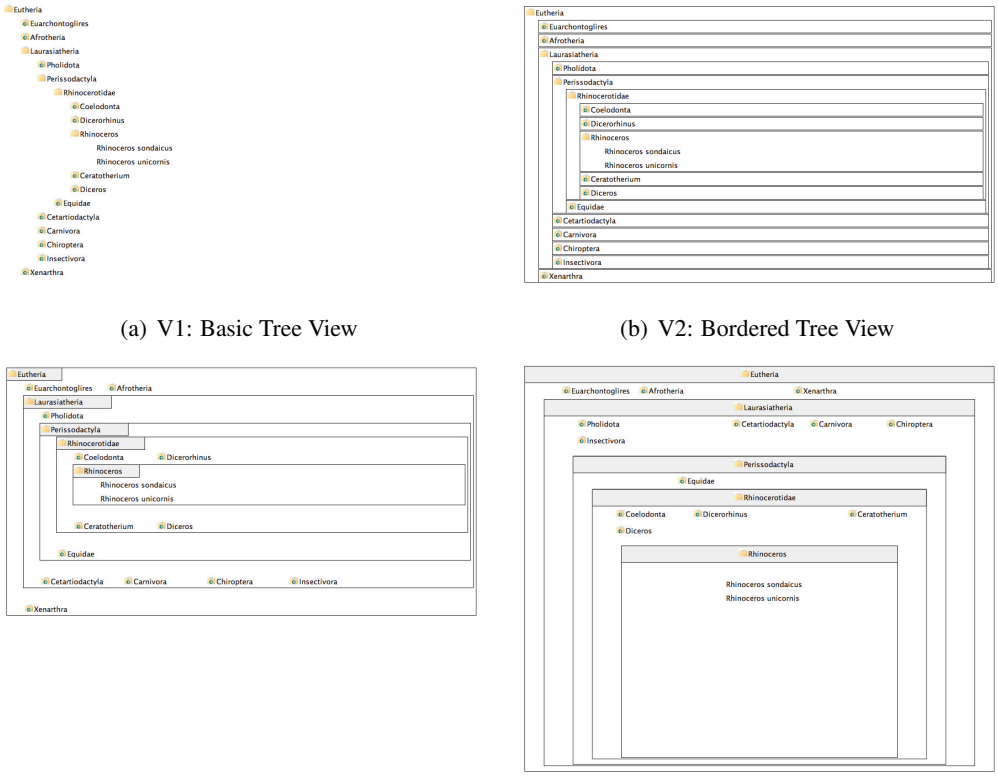
Green et al.'s work studied the effect of locus of control and other personality dimensions on both procedural and inferential learning in a GVis, a visual analytics interface (Figure 4.1(a)) versus NCBI Map Viewer, a more traditional web interface (Figure 4.1(b)). The procedural tasks they studied involved searching for a specific piece of information in a genomic database, while inferential tasks were those in which a user had to make a more open-ended comparison between two items. In both cases, the tasks were prompted by questions of the kind found in a typical usability study. The findings from both experiments suggested that the web table interface was more conducive to answering procedural questions and the findings from their first experiment suggest that the visual interface is more conducive to answering inferential questions.

In the first experiment, they found that participants with an external locus of control completed inferential tasks more quickly than those with an internal locus of control. This effect was more pronounced in GVis. In the second experiment, they studied only procedural tasks and found that, in contrast to inferential tasks, those with an internal locus of control completed procedural tasks more quickly.

In this chapter, we build upon these findings by seeking to isolate the factors of the interface structure that cause the different effects between those with an internal or external locus of control. The amount of data we have about how different user types react to different interfaces is rapidly increasing. Our goal as visualization researchers is to make sense of this data within the context of models of the user. In order to apply this knowledge to improve design, we need to know not just what differences exist between users, but why.

4.3 Experiment

Research by Green et al. [59, 60] suggests that locus of control influences an individual's use of a complex visualization system. However, it seems counterintuitive that a personality trait with no known connection to visual or spatial ability should have any consistent influence over such a complex relationship. We propose that rather than an interaction between



(a) V1: Basic Tree View

(b) V2: Bordered Tree View

(c) V3: Indented Boxes View

(d) V4: Nested Boxes View

Figure 4.2: The four visualizations used in the study. Each view is showing the same portion of one of the phylogenetic tree datasets.

between locus of control and specific complex visualizations, the observed pattern may in fact be a correlation between locus of control and visual layout.

For example, consider the nested circles used in the GVis system from work by Green et al. [59, 60]; these structural elements are visually dominant due to their unusual shape and large size with respect to the surrounding textual elements. In contrast, Map Viewer uses a more subtle indentation-based structural expression which is dominated by text. Apart from using different visual encodings and interaction styles, these two designs represent significantly different visual layouts of the same underlying data. Exploring how those layouts differ may help explain why locus of control interacts significantly with them.

To more closely investigate this correlation, we conduct a comparable study in which the test visualizations are more tightly controlled. We restrict the variation between our four test interfaces to visual layout style, holding interaction metaphor and visual encoding consistent across all interfaces. We hypothesize that, even under this simplified

setting, participants with a more internal LOC will have difficulty with layouts that depend on a strong containment metaphor, while participants with a more external LOC will show a greater willingness to adapt to a variety of visual layouts.

To test this hypothesis, we performed a user study in which participants were asked to answer search and inferential questions about data in four simple hierarchy visualizations (Figure 4.2). Like Green et al., we measured personality traits of the participants beforehand in order to test whether Locus of Control affects a participant's ability to use these visualizations. The four views were designed to express an increasingly visually explicit containment metaphor for the hierarchy, ranging from a list-like view that only used indentation to show hierarchical structure to a view that used large nested rectangles.

Green et al. [59] used two real-world systems in their work, which has the benefit of providing a realistic testing environment. However, this also makes it difficult to isolate exactly which aspects of the two designs prompted the differing user behavior they found. As the long-term goal of this work is to assist designers in choosing how to display information for varying user types, knowing exactly which elements of the design should be altered is vitally important. Therefore, our intention in designing the visualizations used in our study was, as much as possible, to isolate the factor of layout style which we hypothesized to be the key to these differences.

Before viewing the data, each participant was given a personality test, including the Locus of Control Inventory [140]. Each participant was then presented with a series of tasks to perform on each visualization. The order in which the visualizations were presented was randomized, and the user's ability to successfully accomplish tasks using each visualization was then recorded.

4.3.1 Participants

We recruited 240 participants over Amazon's Mechanical Turk service. Altogether, it took approximately two days to collect data from all participants. Of the 240 participants, four did not report their age or gender. Of the rest, there were 124 males and 112 females. Self-reported age ranged from 18 to 62, with a mean of 26.7 ($\sigma = 9.5$). Our participants reported an average Locus of Control score of 3.61 ($\sigma = .59$). This is slightly lower than

scores reported in other publications that use this particular scale [62, 102]. For example, Lapierre and Allen [102] find a mean Locus of Control of 4.03 ($\sigma = .61$) for a participant pool of 205 employees in various professions. This difference may reflect the broader demographics of Mechanical Turk workers versus participants in traditional psychological studies. Locus of control in particular is often studied in the context of work or education, meaning the participants in these studies may have different educational or economic backgrounds than the general population. However, as we did not collect such demographics from our participants, we can only speculate on this point.

4.3.2 Materials

Participants were initially given two questionnaires to measure the aspects of their personality which are relevant to our hypotheses: a scale to measure the Big Five personality dimensions of Extraversion and Neuroticism and a Locus of Control scale to measure the degree to which they see themselves as in control of or controlled by external events. Both scales were taken from the International Personality Inventory Pool [57] and were combined to form a 40-question survey. Neuroticism and Extraversion were included for comparison with Green et al.'s results, but are not the focus of the current analysis.

Green et al.'s study used two fully functional data exploration systems with many differences between them. For our study, we wish to isolate as much as possible the variable of layout style, and and so we created a set of four very specific visualizations (Figure 4.2). The first of these, V1 (Figure 4.2(a)) displays a tree in a simplified Windows Explorer style, using only indentation to indicate hierarchical relationships. This is representationally similar to the webpage organization used by Map Viewer in Green et al.'s work. The fourth view, V4 (Figure 4.2(d)) uses a nested boxes display that relies heavily on the visual metaphor of *hierarchy as containment* [178]. Although it uses rectangles rather than circles and a very different interaction style, this view is representationally similar to the nested bubbles of the GVis visual analytics system.

Between these two extremes, we designed two intermediate views. V2 (Figure 4.2(b)) is very similar to the indentation style of V1, but adds borders around the tree nodes to suggest a containment metaphor. V3 (Figure 4.2(c)) breaks the strictly vertical layout style

used in V1 in favor of a horizontal layout closer to that used in V4, but still employs indentation to organize the levels of the hierarchy. These views are intended to provide cases that interpolate between the two layout styles used in Green et al.

Our use of such simplified views may raise the concern that our results do not directly apply to more realistic visual analytics scenarios. However, since we are partially attempting to replicate results from a study that employed real-world analysis systems, we feel this approach is complementary to previous work. If our results are similar to those found by Green et al, it would demonstrate that this simplification still maintains the important differences between the systems used in their study.

Furthermore, if a trend can be found to increase from V1 to V4, it would show that our intermediate views do indeed capture the major differences between the two views. Since the intermediate views primarily differ in the degree to which they express either a list-like or containment metaphor, this could support our argument that the finding is largely based on the different user groups' willingness to adapt to one metaphor over the other. That said, layout is a complex factor that by nature is made up of many dimensions. While this study design is intended to keep these layout differences as controlled as possible, there are like other aspects of the layout, such as data-ink ratio or size of visual elements, which cannot be entirely ruled out as factors. Nonetheless, this study design can at least test whether layout factors in general can lead to a locus of control effect without differences in visual encoding or interaction.

Apart from these specific design differences, we attempted as much as possible to maintain consistency between the four views. They all use the same font size and folder icons. Each visualization also has the same interaction style, based on a collapsing folders metaphor such as that seen in a standard desktop file system. This may somewhat bias the results in favor of V1, which most closely resembles the interfaces used in file systems. Nonetheless, we argue that maintaining interaction consistency is important enough for isolating design factors that this is worth the tradeoff.

In addition, we implemented the restriction that only one subtree could be open at one time. If a user expanded one branch of the hierarchy and then attempted to expand a node in an unconnected branch, the first branch would automatically collapse. This was

Table 4.1: The size of the four datasets used in the study.

Dataset Name	Number of Species	Non-leaf Nodes
Amphibia	92	94
Aves	112	145
Eutheria	92	94
Lepidosauria	99	126

intended to keep the amount of potentially visible information consistent across the four views. Having several subtrees open is fairly easy in V1, which is purely vertical, but it would be difficult or impossible to open an infinite number of subtrees in V4 without making the lower-level nodes too small to display an entire label. Finally, to see genome data about individual species, the participant hovered the mouse over the species name to bring up a tooltip. Our goal with this simple if not necessarily intuitive interaction style was to keep the four views as consistent as possible except in how they visually organize the space.

The datasets presented in these four visualizations were four subsets of the full taxonomic tree from the National Center for Biotechnology Information's Genome database [118]. Each dataset consists of a phylogenetic tree where the leaf nodes are individual species. At the leaf level, there is some data on the genome mapping data available for that species, such as the date the entry was updated and the number of proteins and genes in the database. This is similar to the data used in Green et al., but does not include all the information found in the Map Viewer subset of the same database. In our case, we chose to show less data at the leaf level in order to present more data overall and more complex trees. The four datasets had, on average, 98.75 leaf nodes (i.e. individual species) and 114.75 non-leaf nodes in the phylogenetic tree. Details for each individual dataset are shown in Table 4.1. There was some variety in the branching factors and overall structure of the trees, although this was not carefully controlled for. However, since the datasets were ultimately balanced with respect to the view types, these differences should not affect our results substantially.

We considered the unfamiliarity of the datasets to be beneficial to our study, since we could trust that participants would need to consult the views in order to answer the task

Table 4.2: The eight task questions seen in the study.

Dataset	Question Type	Question
Amphibia	search	Within the classification "Batrachuperus," which species was most recently updated?
	inferential	Under "Anura," find the classification "Bufo" and note the subclasses it contains. There is another classification under "Mesobatrachia" that has something notable in common with "Bufo." Find that classification.
Aves	search	Under the classification "Falco," find the species with a "Length" value over 18000.
	inferential	Looking in "Sphenisciformes," find the classification "Eudyptula" and note the species under it. Now look in "Threskiornithidae" for a classification that has something notable in common with "Eudyptula."
Eutheria	search	Within the classification "Tarsius," find the species which was most recently updated.
	inferential	Under "Caniformia," find the classification "Canis" and note the subclasses and species it cotains. Now find another classification under "Ursidae" that has something notable in common with "Canis."
Lepidosauria	search	Under the classification "Bipes," find the species with the lowest "Length" value.
	inferential	Within "Scincomorpha," find the classification "Lacerta" and note the species under it. Now look in "Crotalinae" for a classification which has something in common with "Lacerta."

questions we presented them. Following Green et al. [59], these questions were divided into search tasks and inferential tasks. In both cases, tasks took the form of questions that participants were expected to consult the visualization to answer. This is similar to the methodology used in most visualization evaluation studies. These two question types are meant to represent simple data lookup and more complex analytical tasks, although they are simplified versions of the real-world versions of these tasks. We expect to see more differences in the inferential questions, since these are more likely to require understanding of the structure of a dataset rather than simple navigation ability.

The search questions asked the participants to find a single species within a classification that had a certain property. For example, "Within the classification 'Batrachuperus,' which species was most recently updated?" The verbal metaphor for tree structure used in these questions was varied between a levels and containment metaphor, to avoid any potential confound of metaphor compatibility [178]. Participants were asked to write the name of the species they found within a text field. The inferential questions were more open-

ended, asking the participant to find a certain classification, then find another classification in another part of the taxonomy that had something in common with the first. For example:

> Under "Anura," find the classification "Bufo" and note the subclasses it contains. There is another classification under "Mesobatrachi" that has something notable in common with "Bufo." Find that classification.

This was usually a similar pattern of names or numbers of species that fell under the classification. For example, the classification "Bufo" contained a classification also called "Bufo," and the correct answer was another classification that contained a child node with the same name as itself. Since correctness may be harder to judge in such a question with a free text response, participants in this case chose their response from a list of four multiple choice answers (as well as a "None of the Above" response). Each dataset was associated with two task questions: one search question and one inferential question. Participants therefore saw eight task questions altogether (Table 4.2).

4.3.3 Procedure

After selecting the study task from the Mechanical Turk website, participants were first asked to fill out the 40-question personality scale by rating each item in the scale from 1 (strongly disagree) to 5 (strongly agree). Once they were done, they read instructions on the main portion of the experiment.

The main portion of the experiment consisted of four sessions, one with each of the four views. The sequence in which participants saw these views was counterbalanced to prevent ordering effects. Each view was randomly associated with one of the four datasets. Each participant saw all four datasets, but they appeared with equal probability in all four views.

When they first saw the visualization, participants were asked to familiarize themselves with it for as long as they wanted. When they were done, they clicked on a button labeled "Start" and were presented with the search question. They again had as long as they needed to interact with the visualization and find the answer. When they were satisfied with their answer, they clicked on a "Ready to Answer" button and were presented with a text

field to fill in their answer. At this point, interaction with the visualization was locked. The purpose of this was to separate response time into the actual interaction time, without the time needed to type in the answer.

After answering the search question, the participant was presented with the inferential question for the same dataset. We presented the questions in this order since the inferential questions are naturally more difficult than the search tasks, and answering the search question would presumably give the user time to learn the basics of navigating the interface. The procedure for the inferential question was the same as for the search question, except that participants were presented with a set of multiple choice responses instead of a text field. These responses were presented in random order and were not visible during the interaction period.

Once both questions were answered, participants were shown a brief four-question preference survey on how much they liked the visualization. This survey included statements such as, "The system was easy to use," and "I enjoyed using this system." As in the personality scales, participants rated how much they agreed with each of these statements on a five-point scale. After completing all four visualization sessions, participants were asked to provide their age, gender, and any comments about the study in a form on the Mechanical Turk site.

We measured the time a participant took in their initial training period, the time taken in the interaction period, and the time taken to record their response, as well as whether their response was correct or incorrect. In addition, we calculated each participant's Locus of Control based on their personality scale responses, and averaged their preference survey responses to generate a Preference Score for each of the visualizations they used.

4.4 Results

Our task questions proved to be quite difficult, with an overall accuracy of 68.6% correct responses on search tasks and 47.1% on inferential tasks. This difficulty, and the large amount of time spent interacting with the views, should be kept in mind when interpreting the fol-

lowing results. Across all participants and question types, no view condition was more or less difficult in terms of accuracy ($\chi^2(3, N = 1919) = 3.7$, $p = .29$) or correct response time ($F(3, 1109) = .57$, $p = .63$). As our primary interest is in how these results varied with a participant's personality scores, further analysis focuses on participants grouped by personality type. Generally, we found support for our hypothesis that participants with a more internal LOC would have more difficulty with views more similar to V4. While we found that participants with a more external LOC did perform very well with V4, we did not find a corresponding trend in which they performed more poorly on views similar to V1.

4.4.1 Effects of Locus of Control

We initially divided participants into three groups based on their score on the Locus of Control scale. Participants with a score lower than one standard deviation from the mean (i.e. less than 3.01) were classified as *external LOC* users. Those with a score greater than one standard deviation from the mean (i.e. greater than 4.21) were classified as *internal LOC* users. The rest were classified as *average LOC* users.

The independent variables we analyzed to test our hypothesis were Question Type (search task versus inferential task), View Condition (V1, V2, V3, or V4), and LOC Group (external, average, or internal). Our overall model was therefore a 2x4x3 Univariate Analysis of Variance (ANOVA). This overall model was significant for the variable of Correct Response Time ($F(23, 1109) = 10.67$, $p < .01$). Correct Response Time only included correct responses, and represents only the time that a participant spent interacting with the view, not the time spent writing or choosing her response.

In addition, there were significant main effects of question type ($F(1, 1109) = 158.86$, $p < .01$) and Locus of Control ($F(2, 1109) = 5.38$, $p < .01$). Unsurprisingly, search questions were answered faster than the more difficult inferential questions. Overall, internal LOC participants were slower at answering questions correctly ($M = 183.4$ seconds, $\sigma = 11.6$) than external LOC participants ($M = 147.9$ seconds, $\sigma = 10.3$). There was also a significant interaction between view condition and question type ($F(3, 1109) = 3.01$, $p < .05$) and between question type and LOC group ($F(2, 1086) = 5.93$, $p < .01$). The reason for the latter is that internal LOC participants are much slower than the others on

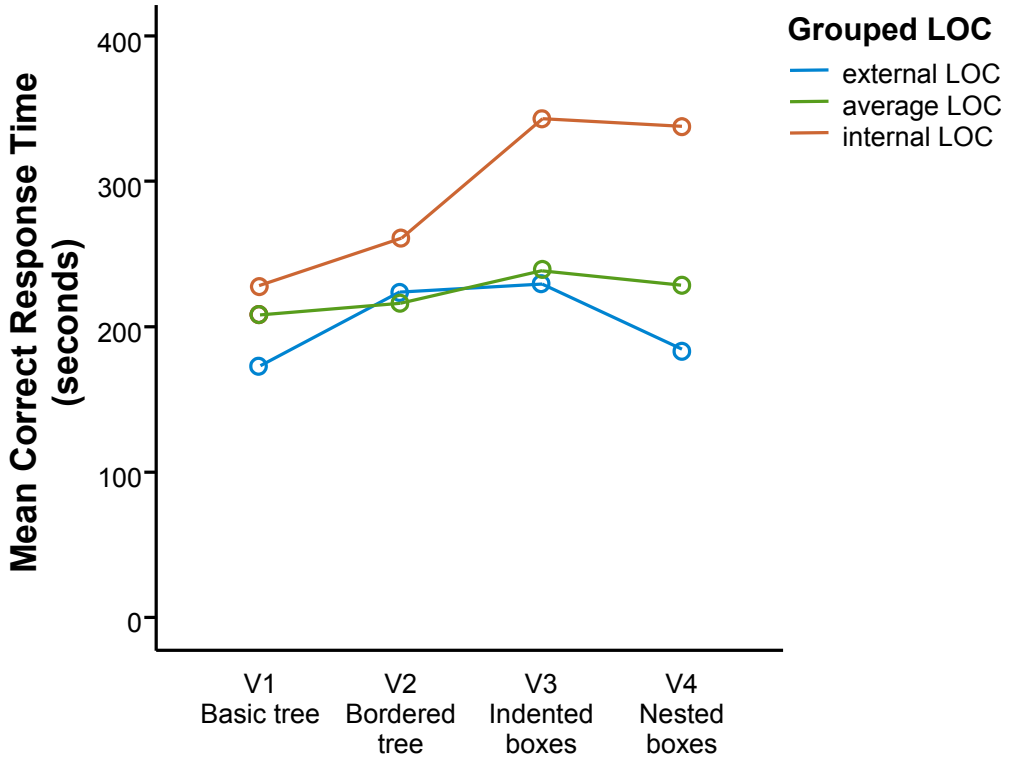


Figure 4.3: Response time for correct answers only across the four view conditions and three Locus of Control (LOC) groups. Participants with a highly internal LOC, who see themselves as in control of external events, were much slower than other participants at answering question in V4, a visualization that uses a strong nested-boxes visual metaphor. Participants with a highly external LOC, who see themselves as controlled by outside events, are relatively more likely to perform quickly on V4.

inferential tasks, but are the same speed when answering search tasks. The source of the interaction between view and question type is that, while inferential questions were answered more slowly than search questions in all views, this difference was significantly larger in V3. For inferential questions, V3 produced the slowest response times of all views, while it produced the fastest for search tasks.

The results relevant to our main hypothesis are summarized in Figure 4.3. We found that participants with an external LOC answer questions in V4 (the nested boxes) faster than other participants, although they answer questions in V1 faster as well. Internal LOC participants show a clear trend of slower performance from V1 to V4, although there is no distinction for these participants between V3 and V4. Average LOC participants show

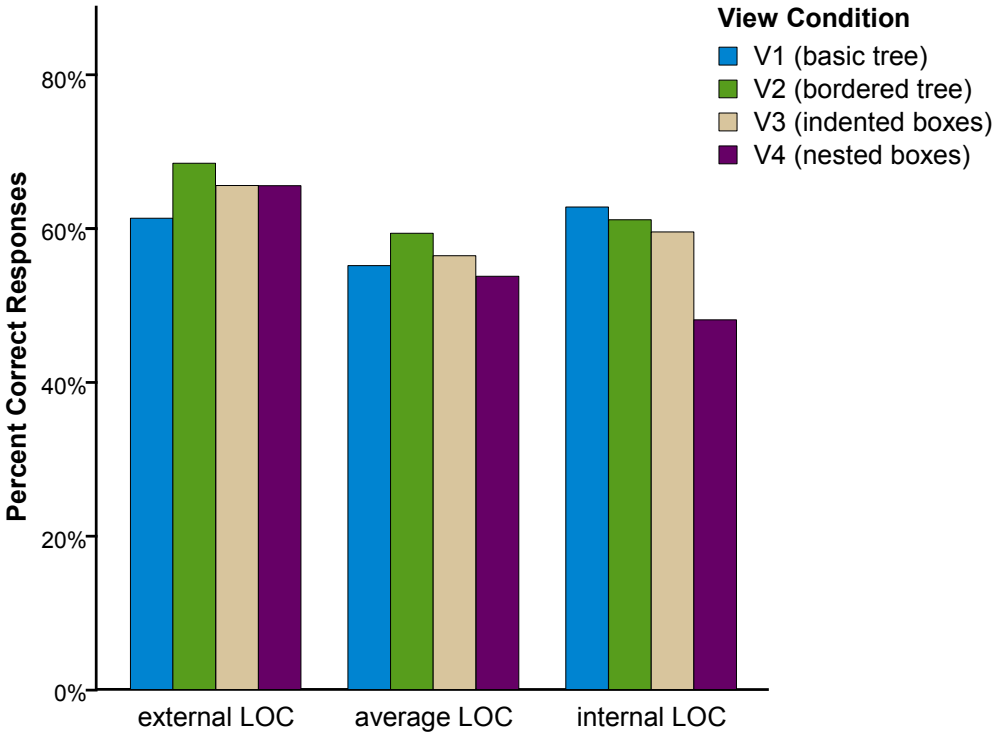


Figure 4.4: Percentage of correct answers across the four view conditions for participants grouped by their Locus of Control score. Participants with a more external locus of control were more accurate overall, while the other groups performed poorly with V4.

no response time difference between the four views.

In addition, we found that raw Locus of Control score correlates significantly with correct response time on nested boxes but no other condition ($r(104) = .23$, $p < .05$). For search tasks, there was no significant response time difference for any participants between the four views.

To test overall accuracy, we used a Pearson's chi-square test on correctness and grouped Locus of Control. Participants with an external LOC answered more questions correctly overall ($\chi^2(2, N = 1919) = 7.7$, $p < .05$), possibly due to their better performance on conditions V3 and V4. These results are summarized in Figure 4.4.

4.4.2 Results by Big Five Personality Factors

As with Locus of Control, we divided participants into high, low, and average groups for both Extraversion and Neuroticism using one standard deviation from the mean as dividing

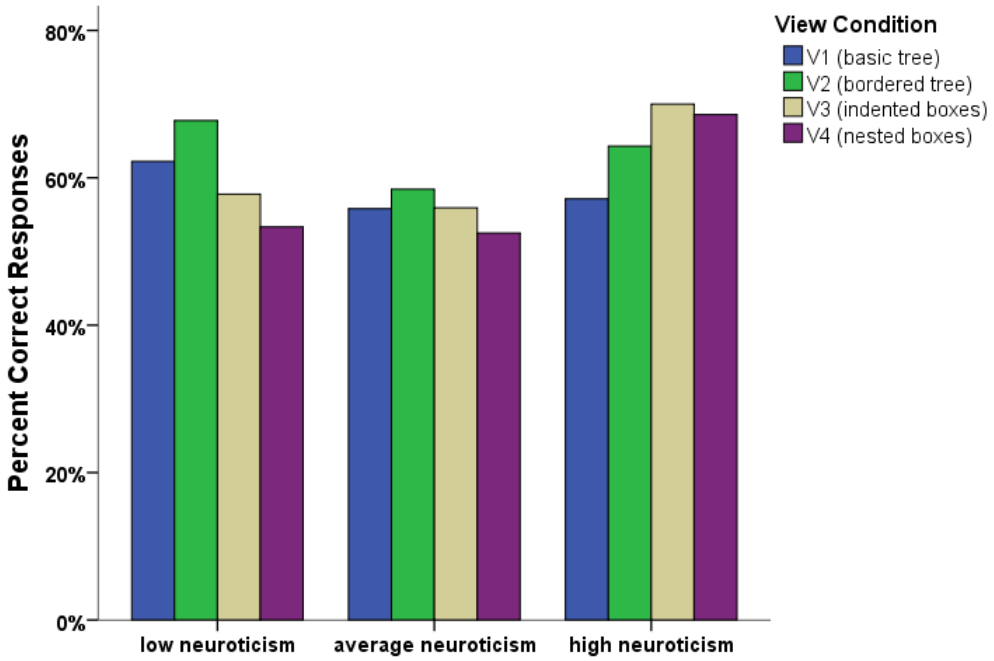


Figure 4.5: Percentage of correct answers in both question types across the four view conditions for participants grouped by their neuroticism score. More neurotic participants performed more accurately with visualizations that used a more container-like layout, while participants in other groups displayed the opposite trend.

points. This split participants into three groups based on Extraversion: introverted (less than 2.29), average extraversion, and extraverted (greater than 3.86). For Neuroticism, these groups were low neuroticism (less than 2.02), average neuroticism, and high neuroticism (greater than 3.53).

We found no overall effects of either of the two Big Five measures on correct response time. As with LOC, we used a Pearsons chi-square test on correctness and grouped personality type to test differences in accuracy. Introverts were more accurate across all four views and both question types than extraverts ($\chi^2(2, N = 1919) = 35.15$, $p < .001$) A similar effect was found for participants with high Neuroticism scores ($\chi^2(2, N = 1919) = 9.24$, $p < .01$).

In addition, highly neurotic participants were significantly more accurate than the other groups in the high-structure V4 condition ($\chi^2(2, N = 1919) = 6.12$, $p < .05$). In general, the more neurotic participants seemed to answer a higher percentage of questions

accurately as views became more container like, while the other groups showed the opposite trend (Fig. 4.5). There was no equivalent significant effect for extraversion. Additionally, unlike Green and Fisher [59], we found no effect of either extraversion or neuroticism on response time for search tasks.

4.4.3 Other findings

In addition to the personality variables that we measured, we analyzed our results based on the demographics and preference information we collected from participants. According to a repeated measures ANOVA on the preference scores for each visualization, V1 and V2 were rated significantly more positively than V3 and V4 ($F(3,1109) = 15.09$, $p < .001$). These preference ratings displayed no significant correlation with accuracy or response time. There was also no relationship between locus of control and preference ratings for any of the four views.

Female participants answered more questions correctly than male participants ($t(231) = -2.31, p < 05$). We found that the age of participants correlated positively with overall response time ($r(1,893) = .13, p < 001$), so that older participants took more time interacting with the interfaces. Age had no effect, however, on accuracy, and did not correlate with any of the personality measures. In the following sections, we will discuss these implications of these findings for our hypothesis and for visualization design in general.

4.5 General Discussion

Through this study, we demonstrated previously reported effects of locus of control in a simplified design that isolates layout style. This supports our main hypothesis that the way visual elements are spatially arranged is a significant factor of design that interacts with locus of control. This also clarifies the earlier findings and suggests an explanation for this effect rooted in the relationship between locus of control and use of external representations.

4.5.1 Replication and Expansion of Previous Findings

Our results replicated those of Green et al. [59] in some cases but not completely. Error rates across the two experimental designs are not directly comparable, as in their case, participants were allowed to try as many times as needed to answer a question correctly, with each mistake recorded as an error. However, in both our study and theirs, we found that participants made more errors overall in the containment-metaphor visualization than in the more list-like view (V1 in our case, Map Viewer in theirs). More significantly, in both our work and theirs, participants with an external LOC responded faster to inferential questions than did participants with an internal LOC, particularly in a visualization with a containment-based visual metaphor (V4 in our case, GVis in theirs).

Together, these two sets of findings provide evidence that the effect of locus of control on visualization use is a robust one. Users with either an internal or external locus of control show performance differences in general on data exploration tasks, and additionally, each group performs better with different visualization styles. This suggests that locus of control is a variable that merits further study, and that personality differences are a valuable topic of research in visualization.

In general, we did not replicate their findings on search question response times. They found that internal LOC participants responded faster to search questions in GVis. However, their completion times for these tasks included any incorrect responses and subsequent guesses made by participants, so for these questions our response times are not directly comparable. This was not the case for the inferential questions, where they recorded only the time to make a single response.

Since our search questions were relatively difficult, it is unlikely that the lack of an effect in these questions is only a matter of performance differences not appearing in easier tasks. We speculate that the inferential questions forced the users to think in terms of the structure of the data to a greater degree. The search questions may have simply measured a participant's ability to navigate the interface quickly, while the inferential questions asked them to characterize parts of the data in an open-ended fashion. Participants may have interpreted these questions in a variety of ways, allowing the structural elements of the

visualization design to play a greater role in their thought process.

Although external LOC participants were faster than other participant groups at answering questions in V4, they were equally fast in V1. This finding does not fit our original hypothesis that internal and external participants prefer different types of visual layouts. It may be that the very high familiarity of tree menus like V1 created a training effect that caused it to break the overall pattern. However, given the evidence, we cannot conclude that external participants perform better with containment views than with list-like views. Rather than a clear trend of group preference, a better interpretation of our results may be that external LOC participants are generally better able to answer inferential types of questions using unusual visualization layouts. An experiment that removes the potential confound of a highly familiar view would be needed to test which interpretation is better supported.

Familiarity may also explain the higher preference scores across all participants for V1 and V2, although it is interesting that there was no correspondence between preference and performance. It is also possible that this lack of a relationship may reflect the fact that our participants were paid a bonus for correct responses, and therefore had an incentive to perform well despite disliking the interface. In any case, people may have felt that V3 and V4 were especially confusing due to their unusual appearance, but they were just as capable of answering questions with these interfaces.

4.5.2 Findings on Other Personality Factors

Compared to Green and Fisher [59], we found fewer notable differences between participants categorized by neuroticism or extraversion. In fact, these personality dimensions had no significant effect on response time, which is where the most dramatic effects of locus of control were found. On the other hand, we did find that these dimensions influenced participants accuracy on search tasks.

In the case of neuroticism, our results provide support for Green and Fishers finding that more neurotic participants generally perform better on search tasks with visual interfaces. The highly neurotic participants were significantly more accurate overall, aggregating all four views and both question types. As suggested by Green and Fisher, this may

be explained by the theory that people with more neurotic or trait-anxious personalities tend to be more attentive to problem-solving tasks up to a certain level of complexity [77].

However, we also found that this effect was especially pronounced in the case of V3 and V4, the most container-like of the four views. Notably, the low-neuroticism participants performed more poorly on these two views than on V1 and V2, which likely contributes to the overall relationship between neuroticism and accuracy (Fig. 4.5). This mirrors the findings on locus of control (see Figs. 4.4 and 4.5) and suggests that less neurotic participants, like the more internal ones, are less able to make sense of these types of visual layouts. It is possible that the greater attentiveness of these users makes it easier for them to learn an unfamiliar interface. Lower neuroticism also correlates with a more internal locus of control, and it is possible that the two scales measure similar qualities that both indicate different aspects of a users unwillingness or inability to adapt to unusual external representations. An alternate explanation is that the neurotic participants may put more pressure on themselves to complete a question correctly rather than abandoning a task due to an unfamiliar visualization.

In the case of extraversion, our results seemingly diverge from Green and Fishers. They found that more extraverted participants responded more quickly to search tasks, and we found that more introverted participants were more accurate on all task types. One possibility is that, under some circumstances, extraverted users respond more quickly but less accurately. That said, we did not find any significant results regarding response time in the current work, so this remains speculative. However, this hypothesis is supported by previous work on introversion and problem solving which finds that more introverted people tend to take more time to think through problems [112]. This extra time may have been particularly helpful in a situation where they had to reason with unfamiliar visual interfaces and data sets.

Since extraversion showed no significant relationship with view type, the overall profile of a participant in our study who performed well with the more containment-like visual layouts is someone who had an external or average locus of control and was highly neurotic. As neuroticism measures emotional stability, and locus of control the degree to which a participant feels in command of situations, these findings may indicate that feeling

somewhat out of control can be an advantage when it comes to making sense of novel visualization designs. It is worth noting that both of these dimensions tend to be negatively correlated with job performance and other practical outcomes [84]. Taken together, this suggests the possibility that persons who struggle with more standard tasks are better suited to thinking with complex visual representations. This possibility warrants further investigation. At the same time, a novel visualization design, while helpful for some users, may be a hindrance to users who already perform a task well using their own methods.

Further work is needed to understand these patterns of users who perform better or more poorly as visual layouts tend toward a strong containment metaphor. It is necessary to examine whether these findings can be generalized beyond this specific task and data type as well as whether they can be generalized to different sets of visual metaphors. Nonetheless, we argue that our findings, particularly those on locus of control, serve as a step toward a better understanding of the externalization theory of visualization and how it works for different types of users.

4.5.3 Locus of Control and External Representations

By isolating the variable of layout style in this study, we have shown that simplified layout changes alone can produce locus of control effects similar to those found in a more realistic user study. This provides evidence that layout is a key factor in mediating this effect, independent of the effects of interaction style, visual encoding, and general differences between traditional interfaces and visual analytic systems. Although the current work does not directly support a causal explanation, we can speculate on why the different layouts we designed interacted with participant locus of control the way they did. We argue that the explanation for the behavior we observed is that participants with an external locus of control are willing to adapt more readily to visualization styles that employ a strong structural metaphor.

The four views in our study range from one (V1) which is dominated by white space and uses subtle visual organization to one (V4) which is dominated by screen elements expressing a visual metaphor of nested boxes. We argue that these views essentially represent a progression from visually implicit hierarchical structure to structure made visually ex-

plicit through layout and the use of lines and fills. V1 (Figure 4.2(a)) uses only a single dimension of spatial organization (horizontal indentation) and makes little distinction between leaf nodes and parent nodes. These factors make the hierarchical structure of the data less explicit. For example, two sibling nodes (such as Rhinocerotidae and Equidae in Figure 4.2(a)) can be visually separated to the point that their relationship is not immediately obvious. Because of this, we argue that V1 is the least structurally dominant of the four views.

After V1, each visualization in the sequence adds at least one visual element used to draw attention to the hierarchical structure of the data. The borders in V2 (Figure 4.2(b)) slightly emphasize the nesting of child nodes within parent nodes; for example, Rhinoceros and its two children are clearly grouped. V3 (Figure 4.2(c)) takes this further by highlighting the names of parent nodes and arranging child nodes horizontally to emphasize a spatial metaphor of containment. Finally, V4 (Figure 4.2(d)) gives the parent node a strong visual emphasis with centering and a title bar and collects the child nodes together at the top of the parent to make sibling relationships obvious. This is the least "list-like" of the four views and uses the most "ink" and screen space to express hierarchical relationships.

The connection between this use of explicit visual structure and locus of control may be explained by the body of thought that views visualization as an external mental representation. The field of distributed cognition [70] sees mental processes such as problem solving and memory as relying not only on knowledge stored in the mind, but also on knowledge stored in a person's environment, in the form of physical objects, information artifacts, and other people. Applying this perspective to visualization, Liu et al. [106] have argued that a primary benefit of visualization is the externalization of information. Externalization makes problem-solving more efficient and accurate by substituting quicker perceptual processes for cognitive processing of information. Furthermore, studies have shown that the form of these representations can influence problem-solving strategies.

Users with an external locus of control are those who expect the outside world to dominate their fate. Taking a distributed cognition approach, it is possible that these users also rely more heavily on outside representations, rather than internal mental representations, when solving problems or making sense of information. Conversely, those with an

internal locus of control may prefer to perform cognitive tasks more internally, relying less on external representations. In general, this makes people with an internal locus of control more adept at problem-solving and learning, as the extensive literature on locus of control has shown. However, it is possible that this same tendency to rely on internal representations may make it more difficult to use the complex external representations found in a visual analytics system.

If this is indeed the case, a visualization with a highly explicit and unfamiliar visual structure may be more jarring for an internal LOC user. Someone with an external locus of control may be more willing by nature to adapt her thinking to the external representation, while the user with an internal locus may be going through a more difficult process of fitting the external representation to her own ideas of what the data is like. Cassidy and Eachus's work [20], discussed in Section 4.2, implies that the "surface learning" approach taken by external LOC students is academically harmful, and based on well-established findings in locus of control research, this does seem to be the case in general. However, this very tendency to focus on surface structure may be beneficial in the context of learning a new visualization system. Our external LOC participants were just as fast with a novel visualization as they were with the kind of indented list they see on their computer desktop every day. There may be other ways of interpreting these effects, and for now, this is a hypothesis for future research, not a firm conclusion. Nevertheless, this potential ability to make an advantage out of a personality style that is usually considered problematic suggests intriguing future directions in the application of visualization to learning.

Taken together with previous work, these findings contribute to the case for an externalization-based view of how people perform complex tasks with a visualization. Furthermore, they imply that this externalization process varies greatly between people and situations, which may be a significant factor in the difficulty of controlling and intepreting evaluations of visualization systems. We discuss the implications of these findings for design and evaluation in the following section.

4.6 Design Implications

This study and others like it provide mounting evidence that personality and design style can have a significant effect on whether a user accepts a visualization design. It is possible that a user's personality can serve as shorthand for subtle cognitive style differences that are not easily measureable otherwise, but which gain importance in the exploratory context of visualization use. When we give users a novel visualization, we are essentially asking them to give up some control over their thinking processes. Some users will find this helpful, while others may find it a hindrance. We argue that a visualization designer should have a sense of how willing a given user will be to take on an external representation, and know how to design a visualization that makes it more or less difficult to ignore the structural aspects of that representation.

Based on our findings, a useful guideline for adaptation would be to increase the amount of explicit structure for users that might have a more external locus of control. Users with a very internal LOC will most likely perform best with a visualization style that uses simple spatial organization and minimal borders, outlines, and other grouping elements. In practice, this type of design may correspond to the maximized "data-ink ratio" argued for by Tufte [156]. External LOC users, on the other hand, may perform more efficiently with a visualization style that violates this classic guideline by including more non-functional elements such as borders, fills, and outlines to call attention to a specific information structure. In addition, this type of user may have an easier time working with visualizations that use a two-dimensional spatial layout to organize information.

Although it would usually be impractical to directly measure a user's locus of control and adapt the visualization accordingly, it is still possible to use this principle to guide design. Our perspective sees locus of control as predicting the degree to which a user will, by innate disposition, prefer her own internal mental models (internal LOC) versus being willing to adapt to an external representation (external LOC). This general principle, then, may apply in other situations where users are likely to prefer a pre-existing mental model or problem-solving process for reasons other than personality. Expert users, for example, may be more resistent to visualizations with highly explicit structure. Likewise, a user group

with a highly standardized analysis process should be given visualizations with a low structural emphasis. A user group which is likely to approach a problem in a more exploratory mode may find it easier to work with a visualization that makes the structural organization of data more explicit.

Visualization, especially in the context of more complex, open-ended problems, involves explorations that can require search, organizing and filtering, inference-building, and iteration in combinations that cannot be predicted beforehand. The results presented in this paper and the design guidelines suggested will be important in the construction of effective visualization tools and methods for this open-ended cognitive process. In addition to telling us something about where adaptation to personality type can be significant, the results tell us where adaptation is less so, permitting the freer use of visual representations and interaction styles determined by other factors. All this will help lead to the development of a model of human reasoning in the presence of automated analysis, which is of central importance to the emerging field of visual analytics.

4.7 Summary

In this chapter, we have contributed findings on how users with different personality types react to varying layout styles used in a hierarchy visualization. We found evidence that systematic differences in layout style can indeed influence a user's response time and accuracy with different types of visualizations that are informationally equivalent but differ in layout. These findings seem to fit a pattern in which users with a more external locus of control are more efficient at using a visualization which uses a highly explicit visual metaphor than users with a more internal locus of control. We hope that these findings can serve as a step towards better understanding of why subtle differences between users' personality styles can have a surprising influence on visualization use.

Chapter 5

Manipulating and Controlling for Personality Effects on Visualization Tasks

This chapter is based on the paper:

> **Alvitta Ottley**, R. Jordan Crouser, Caroline Ziemkiewicz, and Remco Chang. Manipulating and controlling for personality effects on visualization tasks. In *Information Visualization*, 14(3) pages 223-233. 2015. [122]

Work in the previous chapters demonstrate how individual differences can impact speed and accuracy on visualizations. Given the increasing number of known individual differences that impact interaction, it is likely that one or more cognitive factors will cause a user to perform at less than optimal levels on a visual interface. While we have laid a critical foundation for understanding the role of cognitive processes and individual differences in visualization, concretizing the intuition that each user experiences a visual interface through an individual cognitive lens is only half the battle.

The proposed solution is to calibrate the user interface to each individual by intelligently adapting its design based on passive or explicit user input. Previous work has found

that adaptive systems result in performance or satisfaction gains, when it responds to motor abilities [51], vision [51], or brain sensing [2, 148], among others. However, indiscriminately modifying the design of an interface may prevent the user from establishing a clear mental model of the system, decreasing the user's effectiveness and increasing feelings of loss of control.

To avoid this pitfall, it is important to have a thorough understanding of the connection between individual differences and performance on visualizations. In this chapter, we borrow techniques from Psychology and we investigate the impact of manipulating users' personality on observed behavior when using a visualization. We demonstrate how priming techniques can be used to nudge a user's cognitive state and predictably influence performance. We also discuss design implications and the potential benefits of such a technique.

5.1 Introduction

Researchers in HCI and Psychology have investigated the efficacy of priming an individual's personality with the intent of temporarily influencing behavior. Studies include using emotionally-charged visual stimuli to inspire creativity [105] and eliciting varying levels of conformity by having users read words with positive or negative connotations [45]. Results indicate that noninvasive priming tasks can result in significant behavioral changes. In light of these findings, we posit that through priming, we may also be able to elicit performance changes when using visualization systems, specifically changes in speed and accuracy.

We focus our study on the effects of priming a well-established personality trait known as *locus of control* (LOC). Research in personality psychology suggests that an individual's LOC may vary over time, and may even be intentionally manipulated [48, 83]. We hypothesize that we can significantly influence a user's speed and accuracy on visualization tasks by using priming techniques. Specifically, we expect that prompting an average user to be more internal will make them exhibit the behavior of internal participants from Chapter 4, and we expect a reverse effect when prompting an average user to be more external. Similarly, we posit that prompting internal participants to be more external and external participants to be more internal will lead to a reduction of differences between the

groups, if not a full reversal of the original effect.

To test our hypothesis, we replicate the experimental design in Chapter 4 using existing priming techniques to manipulate LOC [48]. After priming, we measure users' performance as they complete search and inferential tasks on two hierarchical visualizations. We recruited three hundred online subjects with varying LOC scores via Amazon's Mechanical Turk and demonstrate that a small priming task is sufficient to significantly impact performance when performing complex tasks.

5.2 Background

There exist many well-vetted techniques for priming personality factors. For example, Epley et al. used nonconscious priming to elicit conformity to social pressures [45]. Participants were given a scrambled sentence task containing words related to either conformity or rebellion, and were told that they were performing a pilot test. They found that this priming task was sufficient to elicit conformity in a subsequent social scenario. Chalfoun et al. introduced subliminal cues to enhance users' learning capabilities when using tutoring systems [21]. They found that by including positive cues such as hints to encourage inductive thinking, they were able to increase reasoning ability and improve decision making when solving logic problems.

Lewis et al. studied the effect of affective priming on creativity by having participants choose images for their background when using drawing applications [105]. Participants were asked to select an image from one of the following randomly-assigned sets of priming stimuli: positive (e.g. smiling babies), neutral (e.g. hammers), negative (e.g. images of dead bodies) and control (no image). They found that individuals who were positively primed produced the most creative drawings as well as the highest quantity of drawings, providing evidence that affective priming can have a positive influence on productivity and creativity.

Harrison et al. extended their study of affective priming by measuring its impact on performance using information visualization [65]. Participants were asked to read a short story designed to elicit either positive or negative emotional responses. They were

then asked to perform standard graphical perception tests. The results indicated that positive affective priming can be used to improve accuracy when interpreting certain visualizations [65].

5.2.1 Priming LOC

Similar priming techniques have been used to manipulate the personality trait *locus of control* (LOC). As previously stated, LOC measurers the extent to which a person believes that external events are influenced by their own behavior. Using the Rotter construct [140], individuals are scored on a 23-point scale where the extreme high and low ends of the scale are categorized as *internal* and *external*, respectively. Persons who score higher on the scale (*internals*) believe that events are contingent upon their own actions, while persons who score lower on the scale (*externals*) believe that events are controlled by external forces or supernatural beings.

Research in psychology has shown that higher LOC scores correlate with increased effectiveness at work [84], better academic performance [47] and greater ability to cope with stress [4]. A person's orientation on the locus of control scale also affects their learning style. Cassidy and Eachus [20] showed that internals are more likely to practice deep learning, while externals are more likely to practice surface learning. This implies that there is a relationship between locus of control and general problem solving techniques, which suggests a potential effect of locus of control on problem solving using visualizations.

Fisher et al. [48] used priming to investigate the effects of psychological intervention targeting LOC in persons with disabilities. In this study, patients with chronic lower back pain were randomly primed to score higher (measure more internal) or lower (measure more external) on the LOC scale. Researchers primed participants' LOC by asking experience recall questions: participants were asked to describe either times when they felt in control (thereby increasing their LOC score) or times they did not feel in control (thereby decreasing their LOC score). They found a significant difference in LOC scores before and after the application of this priming technique. They then assessed patients' perceived and actual physical ability using a lifting task. They reported that patients who were primed to be more internal spent more time on average performing the lifting task and selected heav-

ier weights than the externally-primed group. Patients primed to be more external were significantly more likely to decline to participate in the lifting task. These results indicate that LOC can be primed using experience recall, and also demonstrates that priming LOC can influence behavior.

5.3 Experiment

To test our hypotheses we replicated the study in Chapter 4, holding constant the views, datasets, and questions to enable us to make accurate comparisons between the two results. We conducted a targeted study extending the prior work by applying the experience recall techniques introduced by Fisher et al. [48] to manipulate participants' LOC. We measured participants' baseline LOC prior to the main task using a 23-point Rotter LOC Scale [140] and used this score to assign priming groups:

- Participants who scored **higher than 15** (designated *internal*) were given a task designed to **decrease** LOC score and assigned to group $I^{\rightarrow E}$. Participants of this group were expected to exhibit performance measures that are similar to the average participants reported in Chapter 4.

- Participants who scored **lower than 10** (designated *external*) were given a task designed to **increase** LOC score and assigned to the group $E^{\rightarrow I}$. These participants were also expected to exhibit performance measures that are similar to the average participants of the previous study.

- Participants who scored **between 10 and 15** (designated *average*) were randomly given a priming task and assigned to the appropriate group, either $A^{\rightarrow I}$ or $A^{\rightarrow E}$. We anticipated that average users who were primed internal or external, would exhibit performance measures similar to the internal users of the previous study, while those who were primed external would exhibit performance measures similar to the external users.

For simplicity, we used only the most extreme views (see Figure 5.1) presented in Chapter 4, as their results were most compelling for these views. The order in which the

(a) V1: Basic Tree View

(b) V4: Nested Boxes View

Figure 5.1: The two visualization used in the current study. They were two of the four visualizations designed in Chapter 4. V1 was designed to have a list-like metaphor while V4 was designed to be container-like.

views were presented was randomized, and participants were asked to complete a search and an inferential task for each view.

5.3.1 Participants

We recruited 300 participants via Amazon's Mechanical Turk service. Mechanical Turk is an online market place that allows individuals to be recruited to complete small tasks for remuneration. Though the service is increasingly being used as a research tool because a large number of diverse participants can be recruited in a short period of time [78, 67], it is not without reservations. There are indeed some factors that may affect the validity of the data gathered from online resources such as vote flooding and lack of incentive for completion [99], but the absence of these factors have heightened the appeal for the use of Mechanical Turk in Human-Computer Interaction and Visualization research. It is especially useful for studies such as this, where there is a ground truth for evaluating the results [91] and incentives can be given for correct responses [99].

Of the 300 recruited, we discarded the results of 71 participants for failure to complete the task as required. Participants' data were also discarded if their interaction times were impractical (less than 10 seconds) and they also had no correct responses. The average LOC score was 12 on a 23-point scale ($\sigma = 4.49$) and there were 59 externals, 106 users with average LOC and 36 internals.

5.3.2 Materials

Prior to beginning the experiment, each participant was given the a 29-question LOC personality survey. This survey is adapted from Rotter [140] and is comprised of a group of 23 forced-choice questions such that participants must choose either the external or internal response which best describes their belief. The remaining 6 questions are filler questions designed to disguise the purpose of the test. Once complete, a user was scored by counting the number of internal statements selected, with score of 0 meaning extreme external and a score of 23 meaning extreme internal [1]. The same locus of control test was administered at

[1] This is reverse scored from the original Rotter survey which counts the number of external responses instead of internal responses.

the end of the experiment to maintain consistency between the scoring.

5.3.2.1 Priming

Participants were assigned to one of two priming tasks which were slight modifications of those used by Fisher et al. [48]. In the first condition, participants were asked to describe times when they felt in control, which was designed to increase their LOC score (shifting their LOC toward *internal*). In the second condition, participants were asked to recall times when they felt they were not in control, which was designed to lower their LOC score (shifting their LOC toward *external*). Because our study was conducted via Amazon's Mechanical Turk, users completed priming tasks by entering at least three free-text examples of 100 words each to ensure effective priming. Below are the two priming stimuli used for this study.

Priming Question 1 (Increase Locus of Control)

"We know that one of the things that influence how well you can do everyday tasks is your sense of control over problems you face. The more control you believe you have, the better you will succeed at the things you try and do. If you feel optimistic and able to make the best of your situations, you will do very well. In the spaces provided below, give 3 examples of times when you have felt in control and achieved things well. Each example must be at least 100 words long."

Priming Question 2 (Reduce Locus of Control)

"We know that one of the things that influence how well you can do everyday tasks is the number of obstacles you face on a daily basis. If you are having a particularly bad day today, you may not do as well as you might on a day when everything goes as planned. Variability is a normal part of life and you might think you can't do much about that aspect. In the spaces provided below, give 3 examples of times when you have felt out of control and unable to achieve something you set out to do. Each example must be at least 100 words long."

5.3.2.2 Views

Instead of using the four visualizations as described in Chapter 4, we simplified this study and used only the two extreme visualizations, V1 and V4. V1 displays data in a list-like fashion similar to that of Internet Explorer where hierarchy relationship is represented by indentation, and V4 employs a *hierarchy as containment* [178] metaphor and uses nested boxes to represent relationship. Like in Chapter 4 we limited the exploration of the visualization so that only one sub-tree can be explored and if the user attempted to explore another, the previously explored sub-tree would collapse to its original position. This maintains consistency with respect to the maximum amount of information that is displayed by each view. We also maintained consistency between the visualizations by keeping icons, text sizes and interaction styles constant for the two views.

5.3.2.3 Datasets

The datasets presented in this study are also the same as the ones used in Chapter 4. The previous work used four subsets of the full taxonomic tree from the National Center for Biotechnology Information's Genome database [118]. For every leaf, we present information for that species such as the date when the entry was last updated and the number of proteins and genes stored in the database. Each dataset consisted of a phylogenetic tree where each leaf node is a species. The four datasets had an average of 98.75 leaf nodes (individual species) and 114.75 non-leaf nodes (Table 4.1).

5.3.2.4 Tasks

Consistent with the previous study, participants were then asked to complete a search task and an inferential task using each view. For each task, users were presented with a question and were expected to explore the visualization to retrieve the answer (Table 4.2). The search task was a simple look-up task whereby the user was asked to find information about a single species within a specific category. For example,

> Under the classification "Falco," find the species with a "Length" value over 18000.

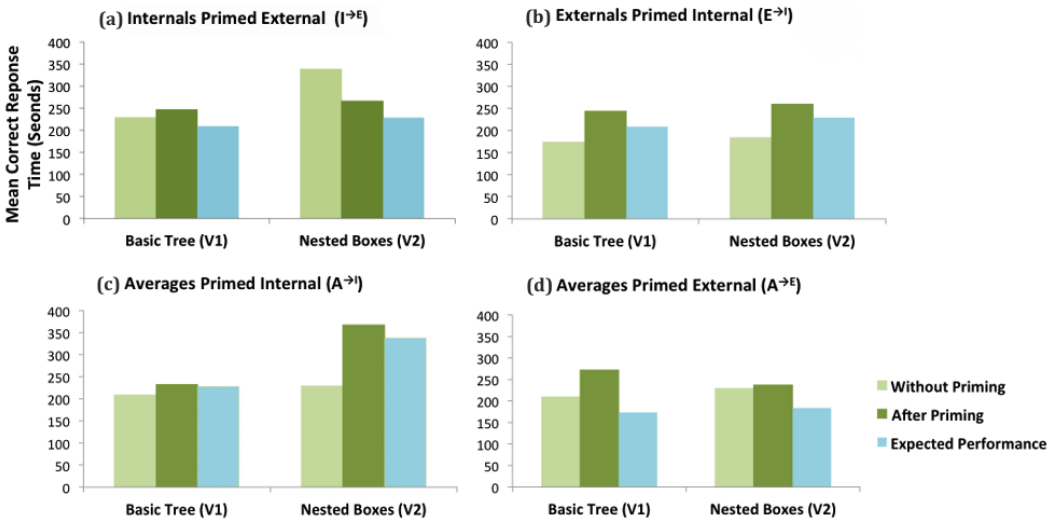


Figure 5.2: Mean correct response times on inferential task questions across the two views for each of the four priming groups. The average participants were successfully primed to behave as internal participants, while the internal and external participants were successfully primed to be more average.

The inferential task asked the user to first find a specific classification and then find another classification with similar properties, forming a more complex analytical task. For example:

> Looking in "Sphenisciformes," find the classification "Eudyptula" and note the species under it. Now look in "Threskiornithidae" for a classification that has something notable in common with "Eudyptula".

5.3.3 Procedure

Participants were first asked to complete the LOC survey. Once this was completed, their locus of control score was calculated and each participant was then issued one of the two experience recall questions. The main portion of the study consisted of two sessions, one with each view, and for each session, participants were presented with a search task then an inferential task. After completing both sessions, they once again completed the LOC survey for which the order of the questions were altered.

We measured each participant's training times, interaction times and times taken to record their responses. Additionally, we recorded their initial LOC score as well as their post survey LOC scores.

5.4 Results

For all but one group ($A^{\rightarrow I}$), the priming prompts were successful at influencing the participants' LOC scores in the desired direction. We ran t-tests on the pre-test and post-test LOC scores for each of the priming groups, and in each case the change was significant at a $p < .01$ level. The group $A^{\rightarrow I}$ was not significant with $p = .3$. Although we observed the statistically significant changes with the other groups, on average the mean difference was small in each case, with the mean magnitude of difference being $M = 1.69$ in the average group, $M = 1.45$ in the internal group, and $M = 1.37$ in the external group.

5.4.1 Impact on Response Time

Although the change in LOC scores was not dramatic, our analysis revealed strong evidence that priming successfully caused performance changes on the visualization tasks. Consistent with findings in Chapter 4, we found significant differences in response time for inferential questions but not for search questions. Therefore, in the following analyses, we refer only responses to inferential questions unless otherwise indicated.

We found a high degree of variability among our subjects in terms of overall response time to inferential questions ($M = 263.5s, \sigma = 240.8$). For this reason, we chose to analyze our primary hypothesis using a repeated measures design. To test our main hypothesis, we performed a repeated measures ANOVA on Correct Response Time using a 2×4 mixed design of Visual Layout (within-subjects) by Priming Group (between-subjects). The ANOVA uncovered no significant main effect of Visual Layout, perhaps because the differences in performance across priming groups counteracted the overall differences in the effectiveness of each layout type. However, this test revealed a significant interaction between Visual Layout and Priming Group, $F(3,57) = 2.85, p < .05$. This finding indicates that different priming groups showed significantly different patterns of performance between the two views.

Analysis indicates that introducing a priming stimulus was generally successful in affecting participants' performance. For group $I^{\rightarrow E}$, the response time difference between V1 and V4 was much smaller than those reported in Chapter 4, which indicates that par-

ticipants with an internal baseline LOC were successfully primed to elicit the behavior of average users reported in the previous study (see Figure 5.2(a)). Likewise, priming the group $E^{\rightarrow I}$, successfully elicited performance measures similar to the average users of the previous study (see Figure 5.2(b)).

Priming was equally effective at influencing the performance of participants with an average baseline LOC. In Chapter 4 we reported that average-scoring participants showed no significant difference in response time or accuracy across the four visualization views. As described in section 5.3.2.1, we primed half of these participants to be more external and half of them to be more internal. The response time results for both groups indicate that priming was indeed successful in causing performance differences:

- Group $A^{\rightarrow I}$ performed much more slowly on V4, with $M = 456.2s$ and $\sigma = 541.4$, as compared to V1, with $M = 215.5s$ and $\sigma = 177.4$ (see Figure 5.2(c))
- Group $A^{\rightarrow E}$ was conversely faster with V4, with $M = 231.7s$ and $\sigma = 137.1$ than with V1, with $M = 348.8s$ and $\sigma = 438.1$, though to a lesser degree (see Figure 5.2(d))

The behavior of group $A^{\rightarrow I}$ mimics the results for highly internal participants reported in Chapter 4. While the behavior of group $A^{\rightarrow E}$ is clearly distinct from the average-scoring participants reported in Chapter 4, it does not replicate the results for their highly externals. Instead of showing no difference in response time between V1 and V4, ($t(46) = 2.01, p = .68$).

5.4.2 Impact on Accuracy

In addition to speed, we also examined participants' accuracy levels across views and priming groups. While we found no significant effects directly related to our initial hypothesis, we did find an unexpected result: when counts of accurate responses were aggregated across question type and visual layout, we found a significant effect of priming group on accuracy, $\chi^2(3, N = 916) = 19.94, p < .001$. Upon closer analysis, it is clear that this effect stems from the fact that participants in group $E^{\rightarrow I}$ were far more accurate overall with 63.5% correct responses. Differences in accuracy across the other three priming groups were not

statistically significant: $I^{\rightarrow E}$ 49.2%; $A^{\rightarrow E}$, 45.1%; and $A^{\rightarrow I}$, 46.5%. These are comparable to the overall accuracy reported in Chapter 4. While the previous study found that external participants were slightly more accurate overall than other participants, the magnitude of the difference was not comparable to that found in the current study.

Perhaps since the priming effect size was small on average, there was no significant correlation between amount of priming and performance difference. One notable finding is that those participants who did show a priming effect, i.e. those whose post-test score showed a difference in the direction predicted by the priming stimulus, were significantly more accurate overall. We counted accurate responses across all four questions (including search questions) and compared this accuracy count between the group that demonstrated an expected score change and the group that did not. The result was significant, $t(227) = 2.26, p < .05$. It is possible that this indicates that these participants were paying more attention to all parts of the study, affecting both the degree of their priming and their performance on the visualization tasks.

5.4.3 Other Findings

Along with our primary hypotheses, we also examined whether other factors had an effect on performance between the two views. While these are not directly related to our main hypothesis, they do provide context for our findings. One question we studied was whether the base LOC reported by the participant at the beginning of the study affected their performance, as it did in Chapter 4. We found weak evidence that this may be the case. To test this question, we studied the correlation between pre-test LOC score and the extent of inferential response time difference between the tree view and the nested boxes view. We named this calculated variable "tree advantage", and note that it is a similar measure to the differences tested by our repeated measures ANOVA. Positive values of tree advantage correspond to faster performance on the basic tree view (V1) thus signifying a behavior like internals of previous studies. Negative values correspond to faster performance on the nested boxes view (V4) and a value of 0 means there was not performance difference between the two views (Figure 5.3).

We found no direct correlation between base LOC and tree advantage. However, the

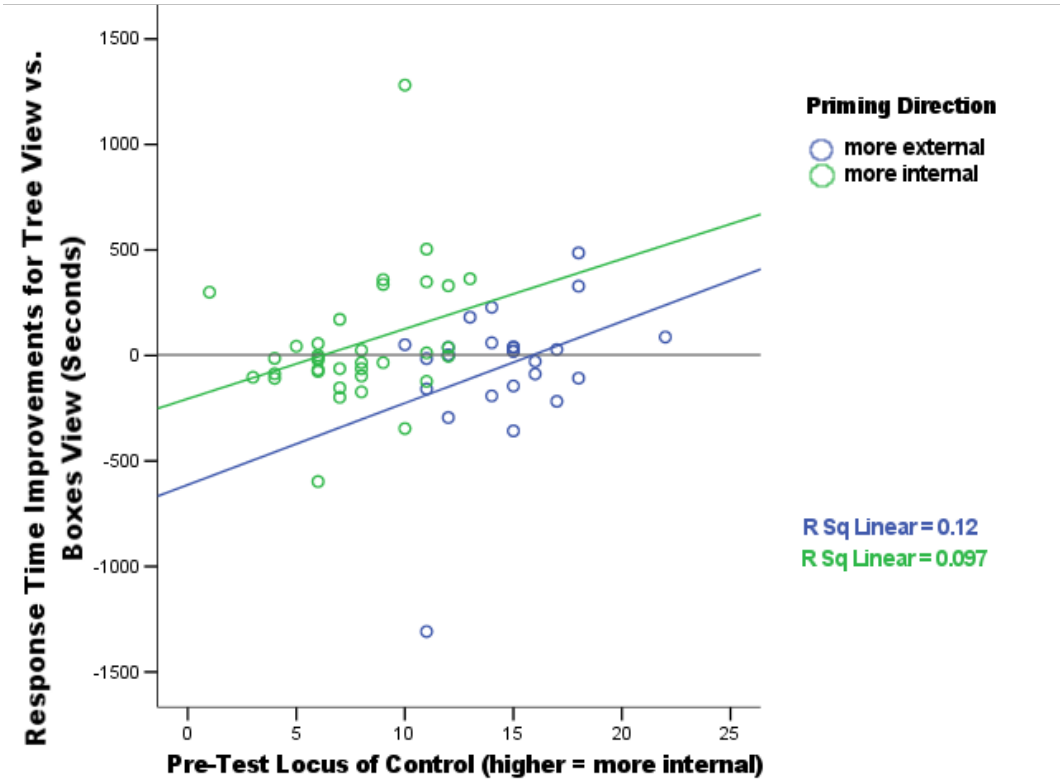


Figure 5.3: When users are divided by the direction in which they were primed, a correlation approaching (but not reaching) significance was found between pre-test LOC and the amount by which a participant performed faster with the basic tree view (V1). A higher score on the LOC scale indicates a more internal LOC. A positive value on the Y-axis indicates that the user was faster at answering inferential questions with the basic tree view, while a negative value means that the user was faster at answering inferential questions with the nested boxes view. Within each priming condition, there was a near-significant effect such that a more internal LOC was associated with a greater performance advantage for the basic tree view.

effect was less ambiguous when we split the participants by priming direction. Among participants who were primed to be more internal, the correlation was positive and approached significance ($r(35) = .31, p = .07$), and the same was true among participants who were primed to be more external ($r(26) = .35, p = .08$). A higher score on the LOC scale corresponds to a more internal locus. Therefore, a potential positive correlation suggests that within each group, there may be an effect such that the more internal a participant is, the more of a performance advantage the basic tree view offers, and vice versa. This suggests that users who measure high on the LOC scale (extreme internals) show little or no performance change after being primed.

One reason this effect doesn't appear when the two groups are aggregated is that, while the two trends are parallel, they are offset in such a way that the trend disappears when they are combined. While this effect is not statistically significant in the existing data, it does raise the possibility that, even in the presence of perceived control priming, innate LOC does still affect visualization performance difference. Taken together with the fact that the difference between pre- and post- LOC scores were small, there is a possibility that the performance difference seen may have been due to some other factor. A closer examination of the priming responses revealed that majority of the participants responded to the stimuli with emotionally charged life stories, and it is also possible that performance differences could somewhat be due to changes in participants' emotional state. A more focused experiment would be needed to verify this claim.

5.5 Discussion

Our results confirm that we can successfully use priming to alter users' performance during complex tasks. We were able to both reduce differences between extreme user groups as well as create differences between average users using priming. It is important to note that the change in performance is both *statistically* and *practically* significant. By asking internal users to recall times when they did not feel in control we are able to effect a remarkable improvement in performance: the response gap between the two views for internal users were reduced from 110 seconds (roughly two minutes) to just 20 seconds. Notably, we

also found that priming affected a user's response time more than their accuracy. This is also true of previous work on metaphor priming and individual differences [175]. This is evidence that interactions between visual style and a user's frame of mind may be more relevant in situations where efficiency is important.

For all but one group ($A^{\rightarrow E}$) we observed the expected completion times. While the average completion time for the $A^{\rightarrow E}$ group was slower than the external users from Chapter 4, the evaluation revealed no significant difference between their performance on the two views . This is still consistent with external users from Chapter 4. Overall, the observed completion times for the current study were slightly slower than the those reported in Chapter 4. One possible explanation could be differences between the time of day and the time of the week when the HITs were posted.

Although we reported significant changes in completion times, the mean change in F was small for groups that showed significant differences between pre-test and post-test LOC scores. Feedback from participants suggests this may be partly due to the fact that some users remembered their responses from earlier and tried to answer consistently. This could be an artifact of the Rotter LOC survey where for each question, the user chooses one of two statements which best describes them. For instance, one question asked users to choose between a statement that suggests that *there will always be someone who doesn't like you* and one which suggests that *if you can't get people to like you then you don't understand how to get along with others*. While we believe priming does affect perceived control of one's environment, the Rotter LOC may not be the appropriate tool for measuring small changes. Future work could investigate the use of other LOC surveys which uses a Likert scale instead of a forced-choice scale.

To some extent, these results complicate previous findings on personality effects in visualization. We found no evidence that we are able to prime users with an extreme baseline LOC to adopt the behaviors of the other extreme using this technique. However, we were able to successfully prompt extreme users to exhibit behavior similar to averages. It is possible that users with a strong tendency in a certain personality trait can only be coaxed out of that tendency to a limited degree. This indicates that while volatile individual differences are still important factors to consider during design and evaluation, priming is not a

panacea. In addition to furthering our understanding of the role of individual differences in future applications, these findings also shed light on the results of previous studies. In the next section, we discuss how this study relates to the design and evaluation of visualizations.

5.6 Implications

This and previous studies underscore that evaluating tools to help people think is a complicated endeavor. Our results suggest that traditional efficiency measures of speed and accuracy may not capture all of what we value in a visualization. While accuracy alone may not reflect the actual difficulty of a task, interaction time proves to be far too sensitive to minor changes in user inclination to provide generalizable information about a system. Evaluation must therefore go beyond simply analyzing the efficiency of a visual design but should also include methods that analyzes the user's cognitive factors.

5.6.1 Evaluation

The way people think and solve problem is often situation-dependent. It is entirely possible that subtle aspects of user study procedures, task question design, and even a researcher's behavior can initiate unintentional cognitive priming and contribute a participant feeling more or less in control. If task performance can be affected by a user's cognitive state, this kind of unintentional priming could harm the validity of evaluation results. This recalls Ziemkiewicz and Kosara's finding [178] that metaphors used in the wording of task questions can interact with visualization layout in an evaluation setting.

Priming can also be intentional. While we only focused on LOC for this study, previous work highlights how other cognitive states can also affect performance on visualization systems [65, 177]. In some cases, the interaction of cognitive states can negatively affect both a user's speed and accuracy and therefore negatively affect evaluation. One practical application of priming is that we may be able to negate these disadvantages. Before evaluating visualizations, researchers can explicitly or subconsciously nudge the user into a specific frame of mind. By subtly presenting a positive news article as participants wait to begin the experiment or displaying a positive picture, researchers could affectively

prime participants, making them better suited to perform certain tasks. Researchers can also administer an "unrelated" pre-task to disguise priming stimuli.

5.6.2 Design

One possibility for visualization design is to prompt a user into a certain frame of mind better suited for the tool at hand. For example, verbal or textual elements such as instructions could also be tuned to temporarily prime the user, improving their capacity for working with a specific interface type. A system could use language in its instruction texts that primes the user to adopt a different frame. It may even be possible to design elements in a more subtle way. Lewis et al. demonstrated how the use of images can influence a user's emotional state [105]. Subtly including images in an interface design may also improve the performance of someone who is having a bad day. Indeed, such prompting may be implicitly at work in existing designs and may affect other cognitive states. Understanding this process better may make it possible to automate some of this design work.

Future systems can be equipped with a better understanding of users' cognitive states and automatically nudge users into a specific frame of mind. Priming may be used to counteract biases and encourage a user to experience an interface with a new perspective. They also proposed the use of priming in the design of collaborative systems. While is it often advantageous to have different perspectives, it is sometimes necessary for collaborators to see an interface as everyone else does. Priming has the potential to unify their conflicting perspectives when they exist.

That said, individual differences research remains necessary. One of our findings is that some users are simply more susceptible to prompting than others. Identifying these less flexible user groups and how they respond to varying visualization designs is still important if we are to completely understand how priming techniques can be used to control personality effects. While we have focused on LOC in this work, similar priming techniques exist for other cognitive states. These may also provide opportunities for controlling for individual differences within evaluation studies. However, it is first necessary to determine whether these cognitive states affect performance on visualization tasks.

5.7 Summary

In this chapter, we demonstrated that by manipulating a user's LOC, we can prompt them to exhibit significantly different behavioral patterns. Our results also highlighted the sensitivity of evaluation measures such as response time to small situational changes. These findings help build toward understanding personality factors can affect the ways humans solve problems with visualizations and contribute to the development of systems that are robust to the effects of individual differences. This research also helps build toward a symbiosis between the system and the user, where not only do users adapt their systems to better suit their analytical needs, but systems can also encourage adaptation by the user to enhance performance.

Chapter 6

Locus of Control as a Predictor of User Strategy

This chapter is based on the paper:

> **Alvitta Ottley**, Huahai Yang, and Remco Chang. Personality as a Predictor of User Strategy: How Locus of Control Affects Search Strategies on Tree Visualizations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, pages 3251-3254. 2015. [125]

The work in Chapter 5 provides compelling evidence of the connection between locus of control and performance with visualizations. We saw that locus of control impacts both speed and accuracy, and that we can predictably influence performance by priming locus of control. This indicates a strong relationship between locus of control and performance. Is it then possible to customize or design an interface based on locus of control? Such customization would hinge on the assumption that locus of control not only impacts speed and accuracy, but it also influences *how* a user interacts with an interface. In this chapter, we explore the feasibility of personalized visualizations by analyzing users' interactions to investigate the relationship between locus of control, visual design, and strategies.

6.1 Introduction

User interactions are critical to the success of visual analytic tools, and interactive interfaces as a whole [152]. It is through user interaction that humans leverage their curiosity, intuition, and creativity to discover patterns, relationships, and other phenomena within data. Interaction fosters visual data exploration. Additionally, users can test and assert hypotheses through their interactions [132, 152]. Thus, gaining a thorough scientific understanding of how users can benefit from interactions in visualization is an important research area [132].

The ability to interact in order to explore information has important implications with regards to enabling the cognitive processes involved in gaining insight into data. Such processes, which can be broadly categorized as sensemaking tasks [133, 92], involve a series of cognitive manipulations and transformations of the data. These tasks fundamentally involve transferring the domain expertise and hypotheses from a user to a system through user interaction. Thus, designers of visual analytic tools often aim to create user interactions that enable and support this cognitive flow of the analytic process [43]. The challenge then is understanding to what extent the interactions a user performs reflects their strategies, mental models, and analytic processes. Further, how much of these cognitive artifacts can be recovered through the analysis of the user interaction?

The review of prior work presented in Chapter 2 suggests that this is possible. For example, Dou et al. [39] showed that user attributes such as strategies, methods, and findings can be retrieved by manually reviewing interaction logs *alone*, with 60-80% accuracy. Similarly, Gotz and Wen observed that users performed similar short, iterative sequences of analytic steps to accomplish a specific analytic goal [58]. Although each user's specific steps might differ, the patterns can be manually classified into higher-level action sequences and detected in a real-time adaptive system.

This chapter expands on previous work and we hypothesize that we can successfully infer user's strategies and cognitive traits from their interaction logs. We hypothesize that the observed differences in speed and accuracy of locus of control groups are due to differences in their strategies. By better understanding how individuals use visualizations, the community can begin to design tools that target users' specific cognitive needs. We

focus on how LOC impacts users' search strategies - their data exploration path as they perform searches using a given visualization. We present results from a user study where we captured participants' mouse interactions as they completed simple search tasks using two popular hierarchical visualizations (Figure 6.1). Overall, we found a strong correlation between LOC and strategies. An initial analysis found that Externals were almost twice as fast as Internals when using the indented tree visualization and conversely, Internals were almost twice as fast as Externals when using the dendrogram.

Further analysis demonstrated that the search patterns for the two groups of users (Internals and Externals) differed significantly, and even within groups, we found that search patterns differed across visualizations. Altogether, our results provide evidence that Externals are less efficient when a visualization affords open exploration than when a visualization provides a guided/restricted exploration.

We make the following contributions:

- We demonstrate how LOC affects users' speed and search strategies.
- We introduce recommendations for designing personalized visualizations based on the observed user strategies.

6.2 Related Work

The impact of LOC has been explored for many decades beyond the Visualization and HCI communities, and research suggests that the differences between Internals and Externals are quite vast and that LOC can impact the way we physically interact with the world. Internals tend to have a strong sense of self-efficacy allowing them to take control even when faced with difficult problems. Conversely, Externals believe that they have no control over external events, making them far more likely to adapt to situations. However, because of this perceived lack of control, Externals are also more likely to give up when faced with difficulty.

Past research corroborates this. Internality has been shown to correlate with increased effectiveness at work [84], better academic performance [47] and greater ability to cope with stress [4]. LOC also affects learning style. Cassidy and Eachus [20] showed that

Internals are more likely to practice deep learning, while Externals are more likely to practice surface learning. In the medical community, LOC has been shown to affect patients' recovery outcomes. Fisher and Johnstion [48] found that users with external LOC were more likely to become discouraged and give in to their disability.

6.2.1 Inferring User Attributes

Existing work in the community on inferring user attributes has been mainly twofold. One school of researchers has focused on analyzing users eye-tracking data while other haves analyze mouse interaction data. We summarize these below (A more in-depth review of prior work can be found in Chapter 2).

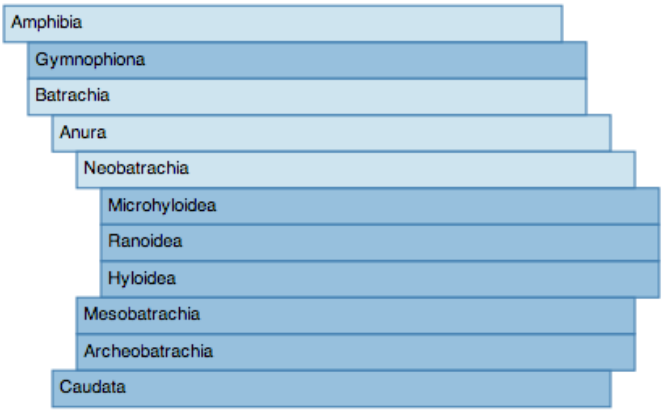
Lu et al. demonstrated how eye gaze data can be used to determine important or interesting areas of renderings and automatically select parameters to improve the usability of a visualization system [107]. Steichen et al. explored the use of eye tracking data to predict visualization and task type [154, 151]. With varying degrees of accuracy they were able to predict: (1) a user's cognitive traits: personality, perceptual speed and visual working memory, (2) the difficulty of the task, and (3) the visualization type.

In the HCI community, Gajos et al. developed the SUPPLE system that can learn the type and degree of a user's disability and generate dynamic and personalized interfaces for each specific user [51].

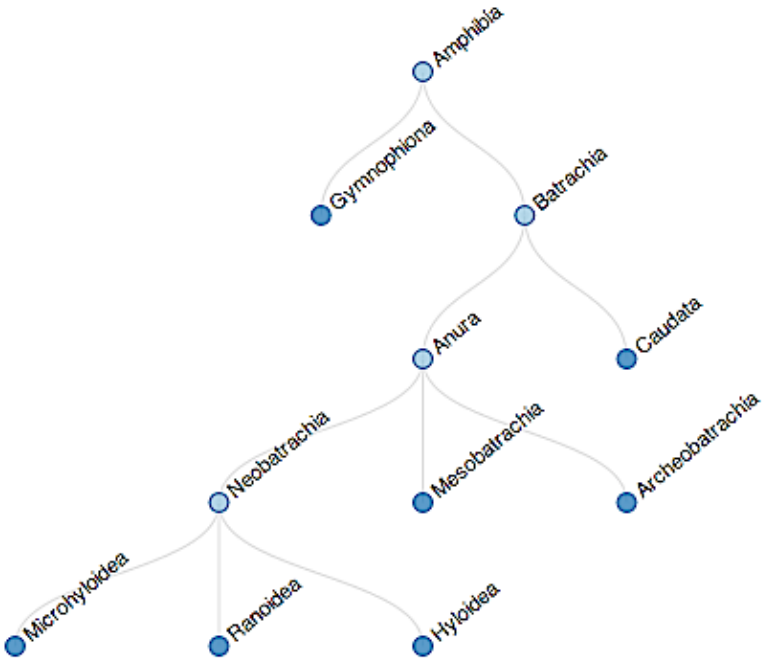
6.3 Experiment Design

Grounded by existing work, our experiment examines exactly how LOC impacts strategies. To test participants' LOC, we used the LOC inventory from the International Personality Inventory Pool [57]. We recruited 54 participants over Amazon's Mechanical Turk service (28 males). Participants' age ranged from 20 to 60 with an average age of 33 years. Only 16% of our participants self reported to be familiar with interactive visualizations and 66.7% self reported to have at least a college education.

Each participant was only able to complete the task once and was paid a base pay of $1 plus a bonus for every section successfully completed. A successful participant received



(a) Indented Tree



(b) Dendrogram

Figure 6.1: The two visualizations used for our experiment.

a total of $5. This method of renumeration was chosen to dissuade participants from simply clicking through the task to be paid, thus increasing the reliability of our data.

6.3.1 Visualizations

Participants completed search tasks using two hierarchical visualizations: an indented tree (Figure 6.1(a)) and a dendrogram (Figure 6.1(b)). These particular visualizations were chosen because they are typical representations of hierarchical data [147], are commonly used in real world scenarios and have been extensively studied in the visualization community [59, 122, 177]. The indented tree uses indentation to depict hierarchy while the dendrogram uses a classic node-link structure.

With the exception of the layout, all other the design features were consistent across the two visualizations. Participants were able to explore the datasets by clicking parent nodes to expand their children. Clicking an already expanded node "hides" that subtree. If a user clicks a parent node to expand a "hidden" subtree, the subtree would be restored to its former state.

6.3.1.1 Datasets

We used two different datasets; both were subsets of a full taxonomic tree retrieved from National Center of Biotechnology Information's Genome database [118]. Each dataset was a phylogeny tree where the leaf nodes were actual species. By hovering over a leaf node, participants were able to access attributes of the species.

6.3.1.2 Tasks

For the main task, participants performed two simple search tasks (one with each visualization) where they were instructed to find a species under a specified classification with a specified attribute. For example:

> Under the classification "Pelophylax", find the species with the lowest length value.

6.3.2 Procedure

Prior to beginning the experiment, participants completed the LOC survey. The main task was divided into two sessions, one for each visualization. For each session, the participants

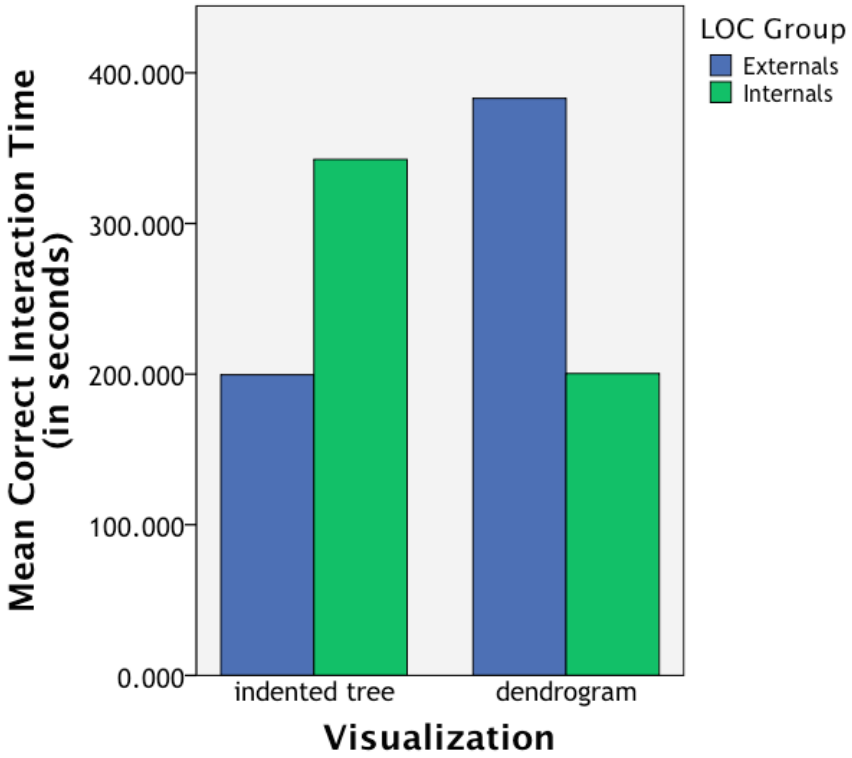


Figure 6.2: Mean correct response time for the two LOC groups across the two visualization designs.

were first given an opportunity to interact and familiarize themselves with the visualization (a third dataset was used for the trials). They were then instructed to take as long as they needed to interact with the visualization to find the target species.

Once they found the species, they clicked "Ready to Answer" and were then able to enter the species name in the text field provided. At this point, they were no longer able to interact with the visualization and we recorded their search times, mouse clicks, hovering and scroll events. After the main tasks, each participant completed usability and demographic surveys.

6.4 Results

Each participant performed 2 search tasks (one with each visualization), resulting in a total of 108 trials. To filter the inherent "noise" in our Mechanical Turk data, our analysis only includes trials where the participant successfully found the target (91 trials).

The average time spent completing the task was 259 seconds ($\sigma = 207.7$) with an average of 75 clicks ($\sigma = 11$), and the overall accuracy rate was 68% (75.6% with the indented tree and 62% with the dendrogram). While participants were slightly more accurate with the indented tree, the observed difference was not statistically significant ($\chi^2(1, N{=}\ 91){=}\ 1.92, p = .166$).

6.4.1 Interaction Time

We divided participants into two groups based on the median split of their LOC scores. We then ran a 2x2 Univariate Analysis of Variance on the log interaction times to test for differences between the two LOC groups (Externals vs. Internals) across the two visualization designs (indented tree vs. dendrogram). Our overall model was significant (F(3, N=91) = 3.189 p = .03) and the interaction between LOC group and visualization was also statistically significant (F(1, N=91) = 9.297 p = .003).

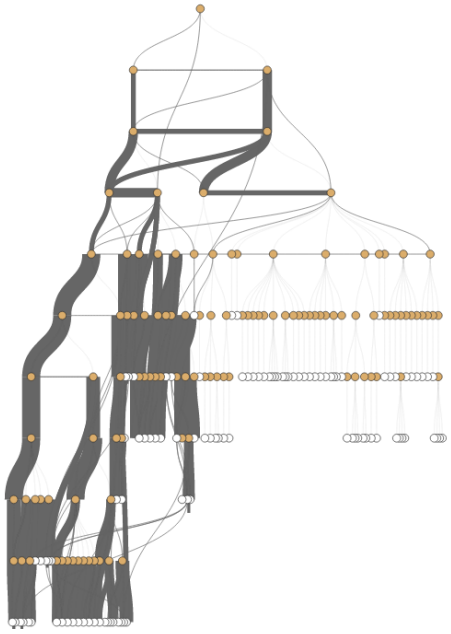
Figure 6.2 summarizes our interaction time findings. Taking a closer look at interaction times, we found that Externals were almost two times faster with the indented tree visualization (Figure 6.1(a)) than the Internals. The Externals were also almost two times faster with the indented tree than they were using the dendrogram (Figure 6.1(b)). Interestingly, we observed the exact opposite effect with the Internals.

6.4.2 Strategies

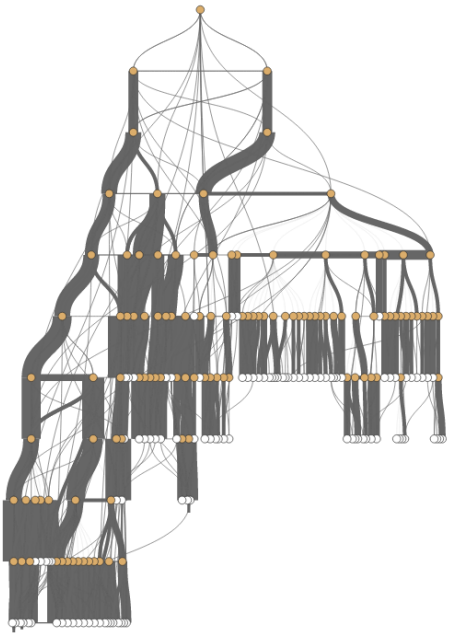
During the experiment, a user explored a non-leaf node by clicking to expand or collapse its children and explored a leaf node by hovering to view its attributes. For our analysis we recorded each participant's mouse click and mouse hover events.

With this data, we were then able to reconstruct each user's exploration path as they searched for the target species. This analysis was performed manually using a visualization tool that we developed specifically for this study, and utilizes animation and juxtaposition to compare different exploration paths. Figure 6.3 summarizes our findings.

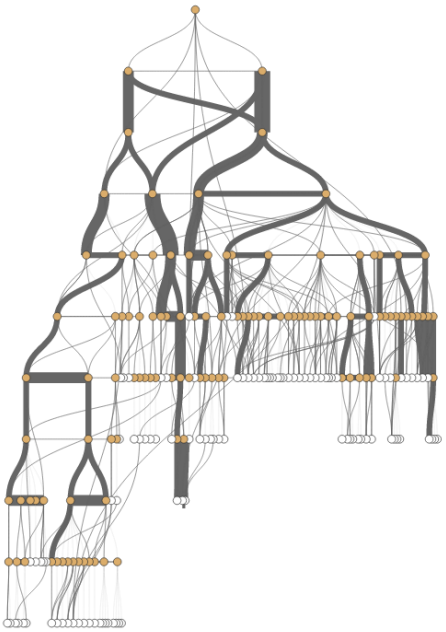
In general, we found that visualization design influences a user's strategies. When using the indented tree visualization, most participants explored the tree in a top-down



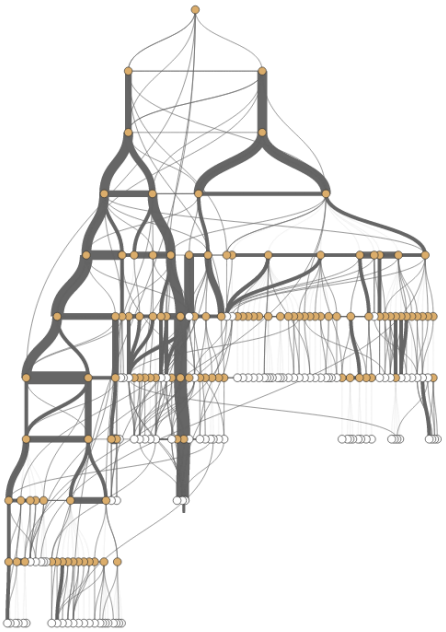
(a) Externals with indented tree



(b) Internals with indented tree



(c) Externals with dendrogram



(d) Internals with dendrogram

Figure 6.3: The figures above show the aggregated exploration paths for users who were successful in locating the target. We grouped users based on a median split of their LOC scores and the visualization design used. For ease of comparison, we used a dendrogram to visualize the exploration paths of both visual designs used in our experiment. The weight of the links represent the percentage of users from that group who traversed that link during their search.

fashion, resulting in a search strategy that resembles a depth-first search. Search patterns with the dendrogram were usually less structured and were often a combination of depth-first and breadth-first search.

Again, we separated participants into LOC groups. When using the indented tree visualization, Externals tended to be very strategic and explored the tree in top-down fashion, following the search strategy afforded by the visualization design (Figure 6.3(a)). While this strategy logically follows the indented tree's top-down design, we found that Internals were far less likely to adopt this strategy (Figure 6.3(b)). This resulted in a more exploratory but less effective search strategy.

Conversely, when using the dendrogram visualization, we found that Externals adopted more sporadic search techniques, resulting in much slower interaction times (Figure 6.3(c)). On the other hand, Internals performed a combined depth- and breadth- first search that proved quite effective with the dendrogram (Figure 6.3(d)).

6.5 Discussion

While we seldom think about a user's personality traits when designing interfaces, the current work demonstrates why they must be considered. Between the tested designs, no design was suitable for all users. We found that Externals performed better with the indented tree but the Internals were more effective with the dendrogram. When asked which visualization they preferred using, 78% of Externals reported to prefer the indented tree over the dendrogram and 63% of Internals said they preferred the dendrogram.

These results are consistent with the LOC construct. Internals prefer control and tend to struggle when a visualization does not fit the their mental model of a problem [177]. While both visualizations were fully exploratory, the design of the indented tree best facilitates a top-down exploration and users who adapted this strategy were fastest (see figure 6.3(a)). This design property is subtle, still it proved to be a hindrance for the Internals. Conversely, since Externals in general do not believe that they can influence external events, they are more adaptable and were able to adopt the top-down search strategy afforded by the indented tree.

The scenario was reversed for the dendrogram visualization. The dendrogram proved to be more versatile and as a result, the Internals excelled with this visualization. They were able to freely search in a manner that best fit their mental model of the data. Unfortunately, this versatility was incompatible with the Externals. Their perceived lack of control coupled with the lack of guidance may have been too overwhelming.

Our design implication follows directly from this finding. When designing for Internals, designers should ensure that designs are flexible to users' mental models. Internals perform better when a visualization allows them to explore the data freely and doesn't impose a strategy. However, the needs of Externals are quite different. While they are in general more flexible and are likely to adapt to novel designs, our findings suggest that adapting to a flexible system may be overwhelming. Unlike Internals who take control of difficult situations, Externals are more likely to feel hopeless and may give up. Thus, it is essential to provide guidance (implicit or explicit) to Externals.

That said, LOC is a multidimensional construct and there are also many other cognitive traits and states that may impact performance on visualization systems [127]. Thus, it is important for us to better understand how these factors, individually and collectively, impact the usability of our designs. There is also evidence that certain traits can be detected from users' interactions [15, 151]. Future work can expand this existing research by exploring how systems can automatically detect and adapt to an individual's specific needs.

6.6 Summary

Altogether, we have demonstrated that individual differences, specifically the personality trait locus of control, do impact both interaction times and search strategies. We believe that our work is a significant step toward fully understanding *how* individual differences affect visualization use and how we can begin to design visualizations that better facilitate users' cognitive needs.

Chapter 7

Finding Waldo: Learning about Users from their Interactions

This chapter is based on the paper:

> Eli T. Brown, **Alvitta Ottley**, Helen Zhao, Quan Lin, Richard Souvenir, Alex Endert, Remco Chang. Finding Waldo: Learning about Users from their Interactions. *Transactions on Visualization and Computer Graphics (TVCG)*, 2014 [15]. This is joint work with Eli Brown and Helen Zhao. My specific contributions included: helped conceive and design experiment, performed data analysis, and co-wrote paper.

As we saw in Chapter 6, interactions with a visualization can reflect important high level information about users, such as their strategies and reasoning processes. However, such manual analysis is limited in terms of scale and timeliness. In order to truly leverage the information contained in interaction logs, automated methods must be developed and evaluated based on their ability to infer high level information about users. The work in this chapter demonstrates on a small visual analytics task that it is possible to automatically extract high-level semantic information about users and their analysis processes. Specifically, by using well-known machine learning techniques, this chapter shows that we can: (1) predict a user's task performance, and (2) infer some user personality traits. Furthermore, we establish that these results can be achieved quickly enough that they could be applied to

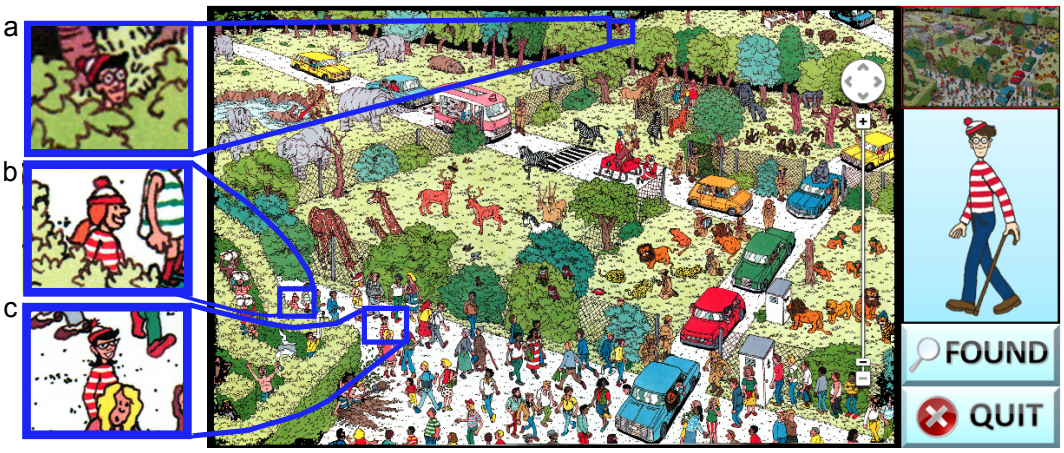


Figure 7.1: Screenshot of the interface used in the study. Users perform the Where's Waldo [64] search task by finding Waldo, the man drawn on the right, in the image. The navigation commands are simple: zoom in or out, pan up, down, left or right.

real-time systems.

7.1 Introduction

Visual analytics systems integrate the ability of humans to intuit and reason with the analytical power of computers [87]. At its core, visual analytics is a collaboration between the human and the computer. Together, the two complement each other to produce a powerful tool for solving a wide range of challenging and ill-defined problems.

Since visual analytics fundamentally requires the close collaboration of human and computer, enabling communication between the two is critical for building useful systems [152]. While the computer can communicate large amounts of information on screen via visualization, the human's input to an analytic computer system is still largely limited to mouse and keyboard [103]. This human-to-computer connection provides limited bandwidth [80] and no means for the human to express analytical needs and intentions, other than to explicitly request the computer to perform specific operations.

Researchers have demonstrated that although the mouse and keyboard appear to be limiting, a great deal of a user's analytical intent and strategy, reasoning processes, and even personal identity can be recovered from this interaction data. Machine learning researchers have recovered identity for re-authenticating specific users in real time using statistics over

raw mouse interactions [111, 135, 137, 173] and keyboard inputs [101], but classified only identity, no user traits or strategies. In visual analytics, Dou et al. [39] have shown that strategies can be extracted from interaction logs alone, but at the cost of many hours of tedious labor. Unfortunately these manual methods are not feasible for real-time systems to adapt to users. The techniques needed to learn about users and their strategies and traits in real time do not exist to our knowledge.

In this paper, we demonstrate on a small visual analytics subtask that it is indeed possible to automatically extract high-level semantic information about users and their analysis goals. Specifically, by using well-known machine learning techniques, we show that we can: (1) predict a user's task performance, and (2) infer some user personality traits. Further (3), we establish that these results can be achieved quickly enough that they could be applied to real-time systems.

Our conclusions draw from an online experiment we conducted to simulate a challenging visual search task that one might encounter as a component of a visual analytics application with the game *Where's Waldo* (see Figure 7.1). The participants were given a visualization enabling a set of interactions (panning and zooming) to explore the image and find the character Waldo. During the participants' search process, we collect a wide range of information about their interactions, including the state of the visualization, and the time and location of all mouse events.

Inspired partly by our visualization of the user paths through the search image, we used this low-level interaction data to create three encodings that capture three major aspects of visual analytics: data, user and interface. The encodings are: (1) *state-based*, which captures the total state of the software based on what data (portion of the image) is showing, (2) *event-based*, which captures the user's actions through statistics of the raw mouse activity, and (3) *sequence-based*, which encodes sequences of clicks on the interface's buttons. The encoded information is then analyzed using well-known machine learning techniques such as support vector machines (SVM) and decision trees to classify groups of users with performance outcomes and individual differences.

The results of our analyses demonstrate that we can indeed automatically extract users' task performance, and infer aspects of their personality traits from interaction data

alone. Further, task performance can be estimated quickly enough to be used in a real-time system. Depending on which data encoding with its corresponding machine learning algorithm, we attain between 62% and 83% accuracy at differentiating participants who completed the task quickly versus slowly, with state-based yielding up to 83%, event-based up to 79% accuracy, and sequence-based 79%.

Beyond predicting performance, we demonstrate that using the same techniques, we can infer aspects of the user's personality factors, including locus of control, extraversion, and neuroticism. With the goal of uncovering more intrinsic user factors, we applied the same techniques to classify participants on personality traits, and found promising signal. In particular, we can classify users based on three of their personality traits: locus of control, extraversion, and neuroticism with 61% to 67% accuracy. The correlation between these three personality traits and the participants' performance are consistent with previous findings in the visual analytics community on individual differences [59, 122, 177].

Finally, on applying the techniques in real-time, we show that accurate prediction of the user's task performance and personality traits can be achieved after observing users for a limited time period. Using the same encoding and analysis techniques described above, we build classifiers based on a limited amount of the user's interaction logs. We demonstrate encouraging results for employing this technology in real-time systems, e.g. with only two minutes of observation on a task that requires an average of nearly eight minutes to complete, we can correctly classify the users with an average of 84% of the final accuracy.

We make the following contributions:

- We show that participants can be classified as fast or slow at the visual search task by applying machine learning to three encodings of participants' interaction data: (1) state-based, (2) event-based, and (3) sequence-based.
- We apply these same techniques to classify participants based on personality traits and demonstrate success for the traits locus of control, extraversion and neuroticism.
- We evaluate the plausibility of applying this work to real-time systems by providing results using shorter timespans of data collection.

7.2 Experiment Design

To investigate what interaction data encodes about users of a system, we sought a task that would simulate realistic tasks, and be difficult enough that people would have to think about how to solve it (engage strategies). Adopting a large visual search task satisfies our criteria: it is easy to explain to participants, but not easy to do, and it is a basic component of typical visual analytics tasks. Specifically, we chose *Where's Waldo* [64], a famous children's game consisting of illustration spreads in which children are asked to locate the character Waldo. Finding Waldo is not easy thanks to the size of the image, which is large enough to require panning and zooming, and the fact that it is craftily drawn to provide a challenge. However, the target is known and there is a guarantee that the task is possible.

We performed an online experiment, collecting interaction data as our participants searched for Waldo in a large image (for our interface, see Figure 7.1). While Waldo is attired in a distinct red and white striped pattern (see Figure 7.1: his full image appears in the panel on the right and his placement in the full spread is shown in inset (a)), he is sometimes hidden behind objects, and the illustrations are filled with distractors specifically designed to mislead the user (e.g., scene elements covered in red and white stripes or the characters shown in Figure 7.1 insets (b) and (c)). To locate Waldo, users have to visually filter unimportant data, making him sometimes difficult to find. This difficulty is also analogous to real-life applications of visual search, where the target item may be partly occluded or obscured by objects of similar color, shape or size.

7.2.1 Task

In the main task, participants were presented with a Where's Waldo poster and were asked to navigate the image by clicking the interface's control bar (Figure 7.1). The control bar was designed to resemble Google Maps' interface and afforded six interactions: *zoom in*, *zoom out*, *pan left*, *pan right*, *pan up* and *pan down*. However, unlike Google Maps, our interface does not allow dragging, rather all actions occur through mouse clicks only.

The zoom levels for the interface range from 1 to 7 (level 1 being no zoom and level 7 being the highest magnification possible). The full image has resolution 5646 by 3607

pixels. At zoom level 1, the full image is shown. At zoom level k, the user sees proportion $1/k$ of the image. Panning moves the display by increments of $1/2k$ pixels.

The interface also includes two buttons not used for navigation: *Found* and *Quit*. When the target is found, the participant is instructed to first click *Found* then click on the target. The user must then confirm the submission on a pop-up alert. We require multiple clicks to indicate Waldo has been found to encourage participants to actively search for the target instead of repeatedly testing many random guesses. If the participant clicks *Found* but does not click on the correct location of Waldo, the click is logged, but nothing happens visually. Unless the participant quits the application, the experiment does not terminate until Waldo is found correctly.

7.2.2 Data Collection

For our analysis, we recorded as much mouse activity as possible, including both mouse click and mouse move events. Mouse click events on interface buttons were logged with a record of the specific button pressed and a time stamp. Similarly, we recorded the interface coordinates of the mouse cursor and the timestamp for every mouse move event.

To establish labels for our machine learning analysis of performance outcomes and personality traits, we recorded both completion time and personality survey scores for each participant. Because researchers have shown [59, 122, 177] that the personality factors locus of control (LOC), a measure of perceived control over external events, and neuroticism and extraversion are correlated with performance on complex visualization tasks, the survey was chosen to collect those traits. Specifically, we use a twenty-seven-question survey which includes the Locus of Control (LOC) Inventory (five questions) [57] and the Big Five Personality Inventory (twenty questions) [38] intermingled. The survey also includes two attention checks which require participants to give an obvious and precise response. These were used to filter participants who did not pay attention while completing the survey.

7.2.3 Participants

We recruited online unpaid volunteers, totaling 118 who successfully completed the task by finding Waldo, of whom 90 successfully completed a personality survey and passed an attention check. Women comprised 39 percent, and men 61 percent. Each participant used his or her own computer and completed the task via the Internet. They were required to have basic computer skills and to have never seen the poster in the experiment before. The participants had a median education level of a master's degree. Ages range from 18 to 45 ($\mu = 24$ and $\sigma = 2.8$). Average task completion time was 469.5 seconds ($\sigma = 351.9$).

7.2.4 Procedure

Participants were first asked to complete the personality surveys by rating a series of Likert scale questions on a scale of 1 (strongly disagree) to 5 (strongly agree). Next, participants read the instructions for the main portion of the experiment and were shown the main interface (Figure 7.1). They were instructed to manipulate the image by using the six buttons on the control bar to find Waldo using as much time as needed and told their completion time would be recorded. Once they had successfully found the target, they completed a basic demographic survey.

7.3 Hypotheses

We collected data at the lowest possible level to ensure that we captured as much information about the participants' analysis process as possible. Over the next four sections we discuss how we first visualize this data, then create encodings to capture different aspects of the participants' interactions based on three core aspects of visual analytics: data, user, and interface. Specifically we encode (1) the portion of the data being displayed, as high-level changes in program state, (2) low-level user interactions, in the form of complete mouse-event data, and (3) interface-level interactions, as sequences of button clicks on the interface's controls. We analyze our data with these encodings with machine learning to evaluate the following hypotheses. First, we hypothesize that participants who are quick at completing the task employ different search strategies from those who are slow, and that

these differences are encoded in a recoverable way in the interactions; second, that we can analytically differentiate users' personality traits based on interactions; and third, that these differentiations can be detected without collecting data for the entire timespan of the task, but instead can be found using a fraction of the observation time.

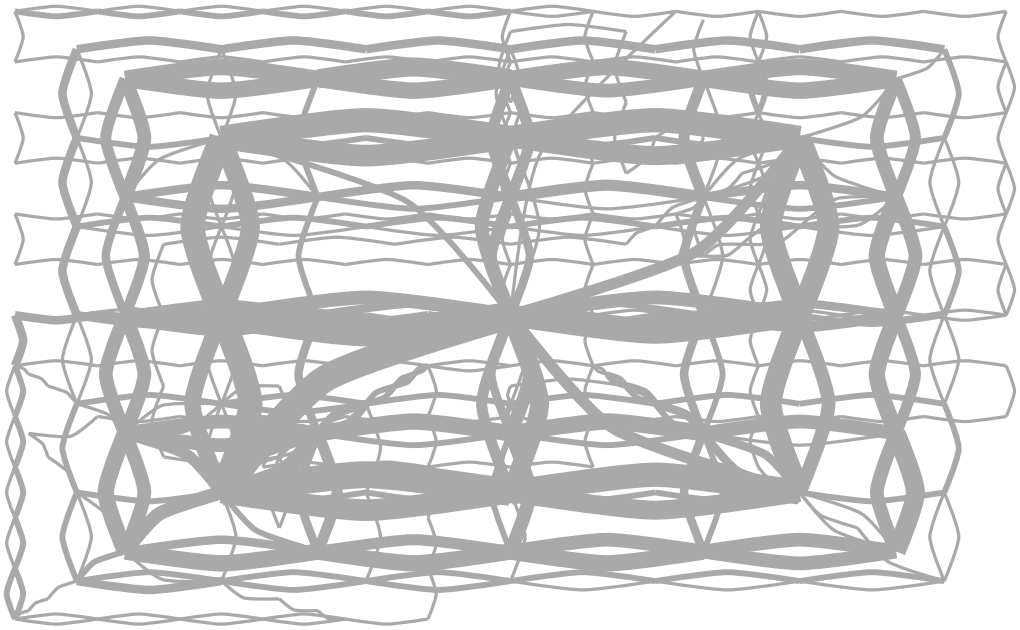
7.4 Visualizing User Interactions

To explore our hypothesis that we can detect strategies employed by different groups of participants, we first visualize their interactions. Figures 7.2 and 7.3 show example visualizations of user movement around the Waldo image. The area of the visualization maps to the Waldo image. Each elbow-shaped line segment represents a transition from one user view of the image to another, i.e. from a view centered on one end-point of the line to the other. Where these lines intersect with common end-points are viewpoints of the image experienced by the participant while panning and zooming. The lines are bowed (elbow shaped) to show the direction of movement from one viewpoint to the next. Lines curving below their endpoints indicate movement toward the left, and those curving above indicate movement to the right. Bowing to the right of the viewpoints indicates movement toward the bottom, and bowing left indicates movement toward the top.

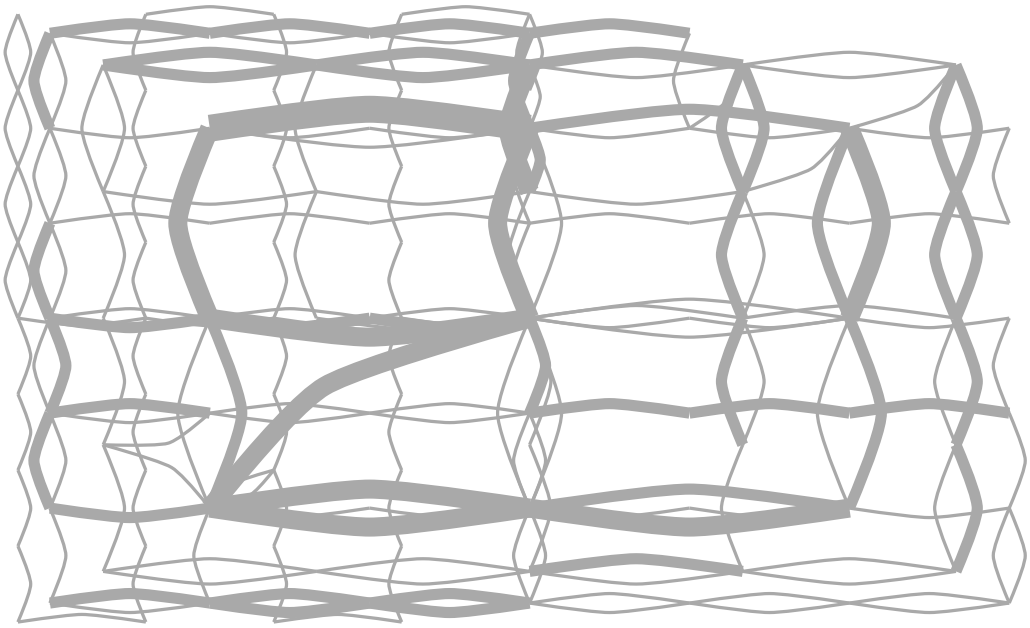
Zoom levels of viewpoints are not explicitly encoded, but the set of possible center points is determined by the zoom level. High zoom levels mean center points are closer together, so shorter-length lines in the visualization indicate the user was exploring while zoomed in. Note that diagonal movement through the Waldo image is not possible directly with the participants' controls. Instead, diagonal lines in the visualization are created because of zooming, i.e. when zooming out requires a shift in the center point.

This visualization can be used to show the movements made by a whole group of users by counting, for each flow line, the number of users who made the transition between the two corresponding viewpoints in the correct direction. In our examples, we are showing such aggregations for four different groups of users. In each case, the thickness of the lines encodes how many users in the group made that transition.

The two sub-figures of Figure 7.2 compare users who were fast versus slow at com-

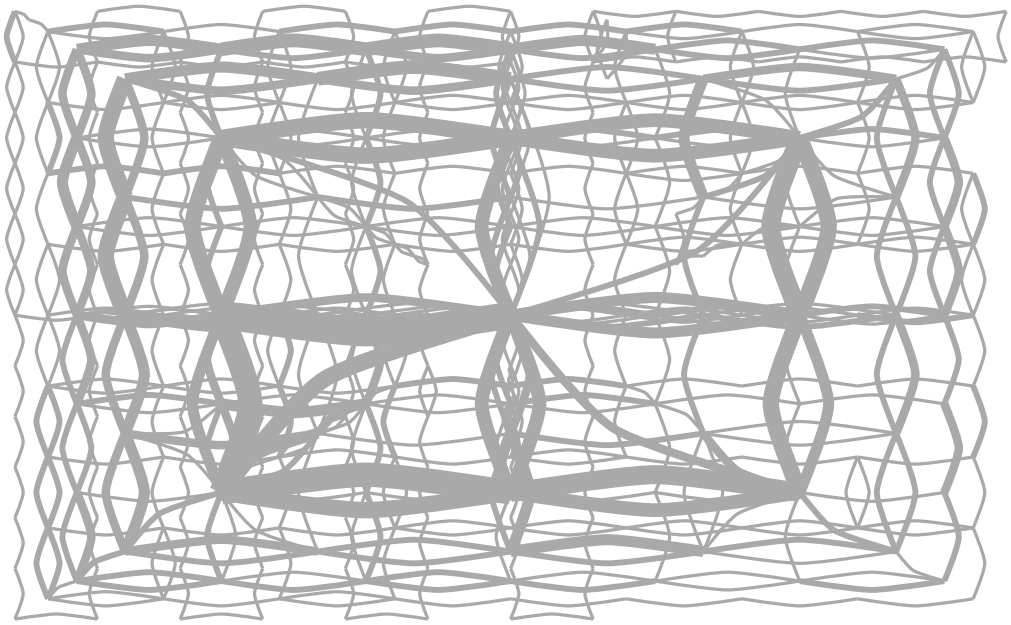


(a) Slow

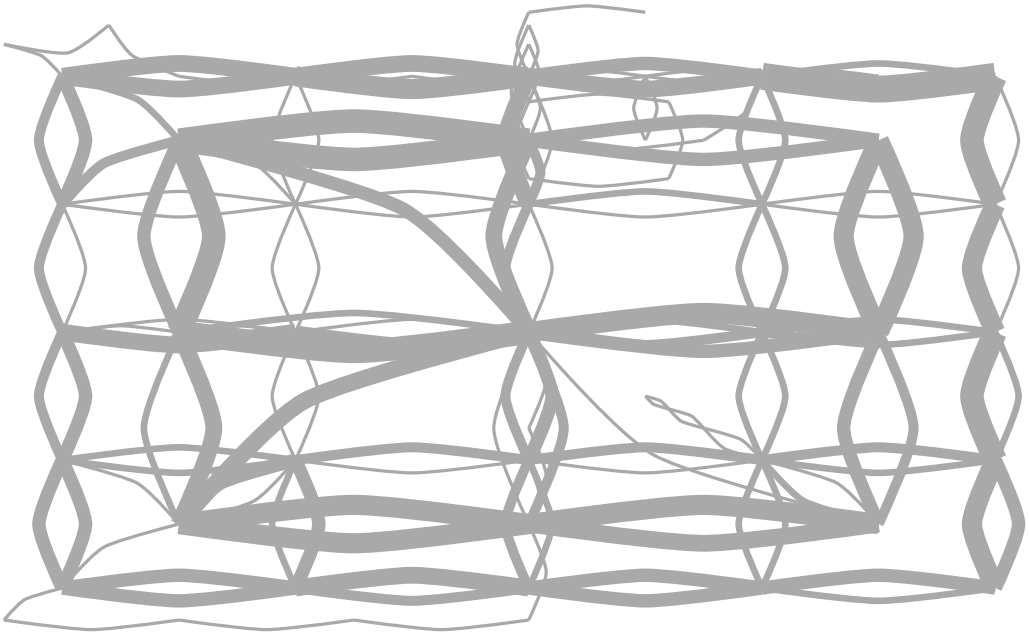


(b) Fast

Figure 7.2: Visualizations of transitions between viewpoints seen by participants during the study (see Section 7.4). Subfigures (a) and (b) show slow and fast users respectively, as determined by the mean_nomed splitting method (see Section 7.5).



(a) External LOC



(b) Internal LOC

Figure 7.3: Visualizations of transitions between viewpoints seen by participants during the study (see Section 7.4). Subfigures (a) and (b) are split with the mean_nomed method (see Section 7.5) based on locus of control, a personality measure of a person's perceived control over external events on a scale from externally controlled to internally controlled.

pleting the task. Users were considered fast if their completion time was more than one standard deviation lower than the mean completion time, and correspondingly considered slow with completion times more than one standard deviation above the mean (for further explanation see Section 7.5). Users who were slow produce a finer-grain set of lines, indicating they made more small movements through the image using a higher zoom level and saw more of the Waldo image in higher detail. Further, the extra lines in the lower left of Figure 7.2 (a) as compared to Figure 7.2 (b) suggest that these slower participants were led astray by the distractors in the image, e.g. the people wearing similar clothing to Waldo seen in Figure 7.1, insets (b) and (c).

Evidence of different strategies is also salient when visualizing results based on some personality factors. The personality trait locus of control (LOC) has been shown to affect interaction with visualization systems [59, 122, 177]. Figures 7.3 (a) and (b) visualize differences between participants with external (low) versus internal (high) LOC. In these subfigures, we see that the external group zoomed in much further on average, while the internal group performed more like the fast group and was able to find Waldo with a smaller set of viewpoints.

These observations are readily seen through these visualizations, but cannot be seen from inspection of the data, nor from machine learning results. Encouragingly, these visualizations hint that there are patterns to uncover in the data. The rest of this work explains our analytical results in extracting them automatically with machine learning.

7.5 Completion Time Findings

In Section 7.4, we presented visual evidence that our collected interaction data encodes differences between groups of participants. However, being able to tell fast users from slow is more useful if it can be done automatically. In this section, we delve into the data with analytical methods, using machine learning to build predictors of task performance outcomes. In particular, we adopt two common machine learning algorithms, decision trees [116], which learn hierarchical sets of rules for differentiating data, and support vector machines (SVMs) [72], which learn hyperplanes that separate data points of different classes in the

Table 7.1: Completion Time Classifiers - results for state space, edge space and mouse events were achieved using support vector machines. The n-gram space results use decision trees. These results were calculated using leave-one-out cross validation.

Data Representation	Class Split	Accuracy (%)
state space	mean_nomed	83
	mean	79
edge space	mean_nomed	83
	mean	63
mouse events	mean_nomed	79
	mean	62
n-gram space	mean_nomed	79
	mean	77

data space. We apply these, to three representations of the interaction data, created to capture different aspects of how users interacted with the system.

Specifically, we tested three categories of representations of the participants' interactions, corresponding to some core aspects of visual analytics (data, user, and interface): the views of the image data participants encountered during their task (state-based), their low-level mouse events (event-based), and their clicks on interface controls (sequence-based). In this section we briefly explain how we derive the target participant groups used for our machine learning results, then show, for each representation of the data, our results at predicting if a given user would be fast or slow in completing the task.

We establish two different methods for labelling our participants based on the collected data. Our analyses aim to classify participants into discrete classes, fast and slow, but our recorded data includes only each participant's actual completion time. The first discretization method is to apply the mean completion time (469.5 seconds) as a splitting point: participants with a completion time lower than the mean are assigned to the 'fast' group, and higher to 'slow'. Participants with scores exactly equal to the mean are excluded from the data. In our results, this splitting method is indicated as *mean*. In the second method, we assume that participants whose scores are within one standard deviation of the mean have 'average' performance and we exclude them from the study, labelling the rest as above. We refer to this approach as the 'no-medium' splitting method, indicated in results tables as *mean_nomed*. The no-medium method allows us to see that stronger patterns emerge for participants with more extreme performance.

7.5.1 State-Based Analysis

In the visualization of participants' movement through the Waldo image (see Section 7.4), differences across groups of participants in how they examine the data become salient. This discovery would be more broadly applicable if the differences could be determined automatically. We create two data representations emulating these visual forms to search for patterns that differentiate users based on what parts of the image they chose to look at. In the "state space" encoding, we capture the portion of the data viewed as each participant navigated the Waldo image. In the "edge space" encoding, we capture transitions participants made between viewpoints of the image. Applying support vector machines (SVM) yields high-accuracy classifiers of completion time with both representations.

The state space encoding can be represented by a vector space. We consider the set $s \in S$ of all visual states (given by view position and zoom) that were observed by any user during completing the task. We create a set of vectors u_i, one representing each user, such that $u_i = (count_i(s_1), count_i(s_2), \ldots, count_i(s_{|S|}))$, where $count_i(s_j)$ indicates the number of times user i landed on state j. For the data from the Waldo task, this process yields a vector space in 364 dimensions.

A similar vector space expresses the transitions between viewpoints of the visualization, encoding how participants moved the viewpoint around the image in their search for Waldo. Their strategies may be encapsulated by how they directed the view during their search. In this vector space, the set $t \in T$ consists of all transitions made between any viewpoints by any participant while completing the task. If each viewpoint is represented by the location of its center, x, then $T = \{(k,m)\}$ where any participant made the transition $x_k \rightarrow x_m$ from position x_k to position x_m while searching for Waldo. Each individual user's vector is constructed as $v_i = (count_i(t_1), count_i(t_2), \ldots, count_i(t_{|T|})$, where $count_i(t_j)$ indicates the number of times user i made transition t_j. The dimensionality of our derived transition-based vector space (edge space) is 1134. The zoom levels are not explicitly encoded, but the set of possible center points is determined by the zoom level. This feature space is most closely related to the visualization described in Section 7.4 and seen in Figure 7.2.

The calculated vectors are used as a set of data features for input to an SVM [161], a widely-applied machine learning method that works on vector space data. SVMs are both powerful and generic, and work by discovering an optimal hyperplane to separate the data by class. For this work we focus on results from the default implementation in the machine learning software package Weka [61], which means a linear hyperplane, and slack parameter $c = 1$. This choice of an out-of-the-box classifier is intended to demonstrate that these results can be achieved in a straightforward manner.

Table 7.1 shows the accuracy of our completion time predictions, calculated via leave-one-out cross validation. Both state and edge space provide strong completion-time prediction results, with maximum accuracies of 83%. However, these classifiers can only take into account high-level changes in the software, as opposed to the lower-level physical actions that may characterize different participants, which leads us to investigate different encodings for further analyses.

7.5.2 Event-Based Analysis

Users move their mouse throughout the process of working with a visual analytic system. Sometimes they move the mouse purposefully, e.g. to click on a control, other times they hover over regions of interest, and sometimes they make idle movements. Where the state and edge space encodings fail to make use of this information, the event-based data encoding described in this section derives from the most raw interaction information available to capture innate behavioral differences.

Previous machine learning work has shown that mouse event data contains enough information to re-authenticate users for security purposes [111, 135, 137, 173]. We adapted the data representation of Pusara et al. [135] for our interaction data by calculating their set of statistics over event information. Because we are predicting completion time, we removed any statistics that we found to be correlated with completion time. Table 7.2 shows the set of functions we used to encapsulate the participants' mouse movements and raw clicks. This set includes statistics on click information (number of clicks and time between clicks), raw button click information (percentage of clicks on a particular button, e.g., "% Left" refers to the percentage of button clicks on the "Pan Left" button), and derived mouse

Table 7.2: Features calculated for SVM analysis of mouse movement and raw mouse click data. μ, σ, and μ_3' refer to the mean, standard deviation, and third statistical moment. Pairwise indicates functions of pairs of consecutive events.

Click Event Features	Move Event Features
Clicks per second	Movements per second
Avg. time between clicks	Pairwise Euclidean distance (μ, σ, μ_3')
% Left, %Right	Pairwise x distance (μ, σ, μ_3')
% Up, % Down	Pairwise y distance (μ, σ, μ_3')
% Zoom in, % Zoom out	Pairwise speed (μ, σ, μ_3')
% Found, % Quit	Pairwise angle (μ, σ, μ_3')
% Clicks on Image	

movement information (such as the number of moves, and the mean, standard deviation and third statistical moment of the distance and angle between them). The set does not include total counts of clicks on different buttons or the total number of mouse movement events, because those were strongly correlated with the total completion time. In total, we use twenty-seven features, listed across the two columns of Table 7.2.

As with the state-space representations, we apply SVMs to the mouse-event data. Table 7.1 shows the accuracy achieved with the mouse-event data using SVM classifiers, calculated using leave-one-out cross-validation. This approach manages a maximum score of 79%, which shows that there is strong signal in this low-level mouse data. The input features may reflect subconscious mouse movement habits more than actual intended actions, so the results indicate that the differences between populations may be driven by innate differences in approach or cognitive traits. Even though none of the features is individually correlated with the completion time, these low-level interaction statistics taken together are enough to analytically separate fast from slow users.

7.5.3 Sequence-Based Analysis

The most direct representation of a user's process may be the sequence of direct interactions with software. Clicks are conscious actions that represent the user's intentions, and thus building classifiers based only on these semantically relevant interactions may provide more insight into why and how participants' analytical strategies differ. For our sequence-based analysis, we examine the sequences of button clicks used by participants to achieve the task

of finding Waldo. We connect n-grams, a method from information retrieval for extracting short subsequences of words from collections of documents, to decision trees, a class of machine learning algorithms that produces human-readable classifiers.

7.5.3.1 N-Grams and Decision Trees

The n-gram method from information retrieval is intended for text, so an n-gram feature space must start with a string representation of data. We assign a unique symbol to each of the seven buttons in the interface: 'L' for pan left, 'R' for right, 'U' for up, 'D' for down, 'I' for zoom in, 'O' for out, and 'F' for declaring Waldo found. Each participant's total interactions are thus given by an ordered string of symbols. We derive an n-gram vector space by considering each symbol a word, and each participant's sequence of words a document. Each dimension in the vector space then corresponds to one n-gram (i.e. one short sequence of user actions). Participants are represented by a vector of counts of the appearances of each n-gram in their interaction strings.

In our analyses we apply the J48 decision tree algorithm and NGramTokenizer from Weka [61] to classify participants based on task performance, and report accuracy scores from leave-one-out cross validation. The effectiveness of n-grams is sensitive to the choice of n. We empirically chose a combination of 2- and 3-grams as we found that to best balance accuracy and expressiveness of our eventual analytic output. Our results at classifying participants on completion time are shown in Table 7.1, revealing a top accuracy of 77% calculated with leave-one-out cross validation.

7.5.3.2 Decision Tree Interpretation

One advantage to using a decision tree with n-grams is that the resulting classifier is human-readable. Figure 7.4 shows the decision tree produced for the completion time data in n-gram space, using a mean split for classes. Each internal node shows a sequence of button clicks and the branches are labeled with the number of occurrences needed of that n-gram to take that branch. We can make several inferences about strategy from this tree. The root node indicates the strongest splitting criteria for the data. In this case, that node contains "L D", the n-gram corresponding to a participant clicking "Pan Left" then "Pan Down". If

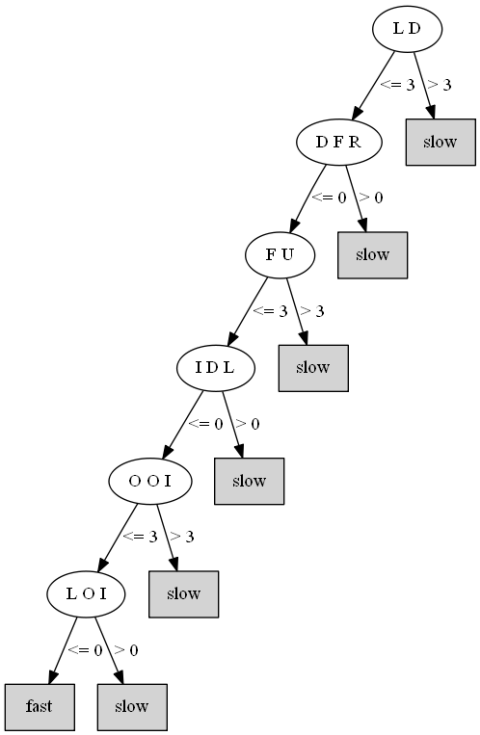


Figure 7.4: This is the decision tree generated as a classifier for fast versus slow completion time with mean class splitting. Each internal node represents an individual decision to be made about a data point. The text within the node is the n-gram used to make the choice, and the labels on the out-edges indicate how to make the choice based on the count for a given data point of the n-gram specified. Leaf nodes indicate that a decision is made and are marked with the decided class.

that sequence was clicked more than three times by anyone, that indicated the person would finish slowly. This makes sense because Waldo is in the upper right of the image. Moving in the wrong direction too many times can be expected to slow down progress at the task.

The "F U" and "D F R" nodes are also revealing. The "F" corresponds to telling the program that Waldo is found. These "F" button presses are not the last action, meaning they do not correspond to correctly finding Waldo. Instead, these sequences show participants' false guesses. Thus the tree suggests that participants who made several false guesses finished the task more slowly.

Finally, the "O O I" and "L O I" nodes correspond to behavior where the participant zoomed out and then back in again. The "O I" component could indicate participants zooming out to gain context before zooming in again. Alternatively, the same subsequence could indicate participants zooming out and immediately back in, wasting time.

Table 7.3: Personality Classifiers - all of these results are with SVM except when using n-grams, which we pair only with decision trees

Data Representation	Class Split	Accuracy (%)
LOC		
n-gram	mean	67
Neuroticism		
mouse events	mean_nomed	62
edge space	mean_nomed	64
Extraversion		
edge space	mean	61

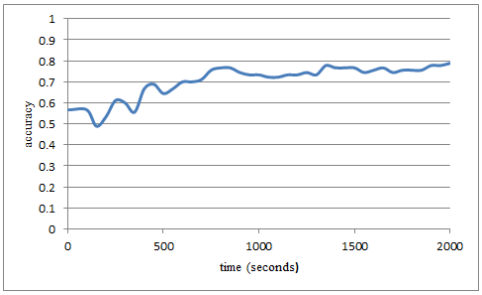
The readability of this technique shows promise for identifying trends in strategies and behaviors. We cannot guarantee that these interpretations reflect the participants' actual intentions, but rather submit these as possible reasons for what is shown in the tree. The real power of using n-grams and decision trees on interaction sequences is that it makes this type of hypothesizing possible, leading to deeper investigation when it is beneficial to understand how people are solving a given task.

7.6 Personality Findings

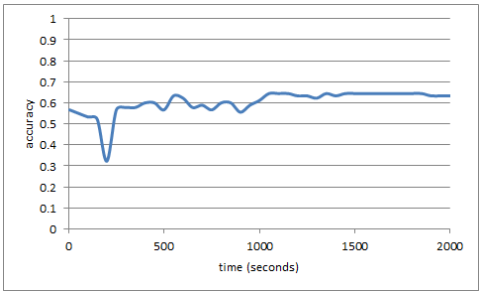
The work in the previous chapters demonstrate that users will use a visualization system differently based on their personality traits. Motivated by these findings, we explore the efficacy of extracting personality traits from interactions. Specifically, we apply the same data encodings and machine learning algorithms used for the completion time analyses to predict users based on their personality traits.

Instead of classes derived from completion time, we separate users into *low* and *high* groups based on their scores on each personality inventory: locus of control, extraversion, agreeableness, conscientiousness, neuroticism and openness to experience. Consistent with our completion time analysis, we test both mean and mean_nomed splits (see Section 7.5). Table 7.3 summarizes our analysis results.

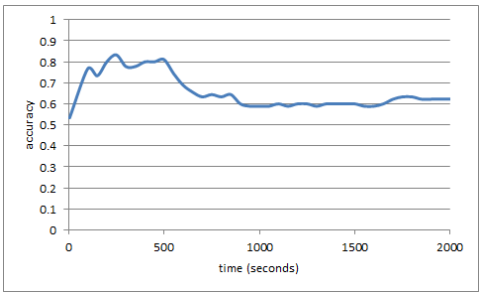
Across several techniques, we successfully classified users based on their LOC, neuroticism, and extraversion scores. Of the personality traits, our techniques were best with LOC, yielding classification accuracies as high as 67%. This supports the findings of



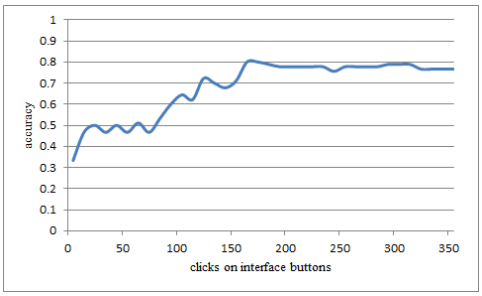
(a) State Based



(b) Edge Based



(c) Mouse Events



(d) Sequence Based

Figure 7.5: Graphs showing the ability to classify participants' completion time as a function of the extent of data collected. The x-axis represents the number of seconds of observation, or number of clicks for the sequence based data. The y-axis is the accuracy achieved after that amount of observation. Accuracy values are calculated with leave-one-out cross validation, and use the mean splitting method (see Section 7.5).

Chapter 4 that of the personality traits, LOC was the strongest predictor of users' performance on visualization search tasks. Consistent with our findings, prior work also found significant effects with neuroticism and extraversion [59, 177].

7.7 Limited Observation Time

The participants in our study were given as much time as they needed to complete the Waldo task. So far, the presented results have taken advantage of the full timespan of the collected data from their interactions to classify them. Investigating the minimal timespan required for this type of analysis is crucial for potential real-time applications, so we evaluated our classifiers' performance as a function of the data collection time.

Figure 7.5 shows, for each of the different data representations, graphs of how task performance classification improves (on trend) with more observation time, i.e. more infor-

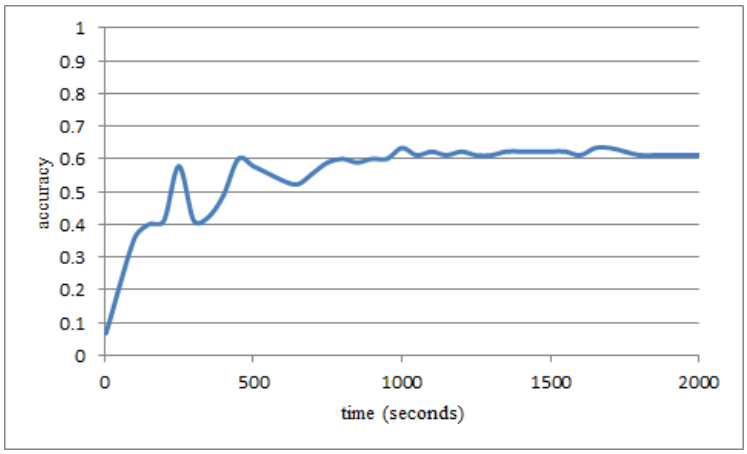


Figure 7.6: This graph shows the dependence of the ability to classify the personality trait extraversion on the amount of time the participants are observed. The x-axis represents the number of seconds of observation. The y-axis is the accuracy achieved after that amount of time. This example uses the edge space encoding and the mean splitting method (see Section 7.5). Accuracy values are calculated with leave-one-out cross validation.

mation available. Figure 7.6 shows one example of this behavior from personality trait classifiers. The x-axis is the amount of data collected, and the y-axis is the accuracy achieved by training and classifying with that amount of data. For all but the sequence-based analysis, the x-axis represents time. For the button click sequences, the x-axis is based on the number of clicks instead. Leave-one-out cross validation (LOOCV) and the mean-based class definition are used for all these results.

These graphs demonstrate two things. First, accuracy scores comparable to the final score can be achieved with much less than the maximum time. Note that close to the mean completion time, the encodings are achieving much of their eventual accuracy scores: state-based, 64% instead of its eventual 83%; edge-based, 60% compared to 63%; and sequence-based, 61% as opposed to 77%. These correspond to 77%, 95% and 79% of their final accuracy percentage scores, respectively.

Second, as expected, in most cases using more data allows for better results. In the case of the mouse event data, the accuracy peaks before reaching the average participant's finishing time, about 470 seconds.

7.8 Extended Results

In this work, we focused on machine learning results produced with off-the-shelf algorithms to emphasize that they could be re-applied in a straightforward way. However, in the course of our investigation, we applied a number of additional customizations to find the most accurate classifiers possible with our data representations. These extended results can be found in Section 7.11. In the appendix, we explain the additional methods we used and show the results we achieved by customizing the classifiers. We show cases in which our tuning produced higher-accuracy classifiers, and revealed signal with feature spaces or class splitting criteria that otherwise could not encode certain traits.

7.9 Discussion and Future Work

In this work, we have shown, via three encodings, that interaction data can be used to predict performance for real-time systems, and to infer personality traits. Our performance predictions ranged in accuracy from 62% to 83%. On personality traits, we were able to predict locus of control, extraversion, and neuroticism with 61% up to 67% accuracy. Further, we found that with only two minutes of observation, i.e. a quarter of the average task completion time, we can correctly classify participants on performance at up to 95% of the final accuracy.

Given the above results, there are some fascinating implications and opportunities for future work. In this section, we discuss the choice of task and data representations, and how they may be generalized, differences between the personality results versus those for completion time, and future work.

7.9.1 The Waldo Task and Our Encodings

The Where's Waldo task was chosen because it is a generic visual search task. It is a simple example of an elementary sub-task that comes up often in visual analytics: looking for a needle in a haystack. The user can manipulate the view, in this case with simple controls, and employ strategy to meet a specific task objective. In this section we address how this

experiment and our analyses may scale to other systems. Because our set of encodings is based on three core aspects of visual analytics, data, user, and interface, we frame the extensibility of our approach in terms of data and interface.

The data in our experiment is the image in which participants search for Waldo. At a data scale of twenty-megapixels, our state-based interaction encodings, which are closely tied to the data because they capture what parts of the image a participant sees, reach hundreds of features to over 1000 features. As the size of the data (image) increases, the state space and edge space may not scale. However, the event-based and sequence-based encodings depend only on the interface, and thus could scale with larger image data.

Conversely, the interface in our experiment is a simple set of seven buttons. Increasing the complexity of the interface affects the event-based and sequence-based encodings. The mouse-event features include statistics about how often each button is pressed, and the sequence-based encoding requires a different symbol for each button. While these two encodings may not be able to scale to meet increased interface complexity, the state-based encoding is unaffected by the interface and thus could scale with the number of controls.

The three encodings we used in this chapter can all be extracted from the same interaction logs. Each one of them provides enough information to recover task performance efficiently. Because of their complementary relationships with the core concepts of interface and data, their strength, as an ensemble, at learning from interaction data is not strictly constrained by either interface or data complexity.

The scalability of the ensemble of encodings raises the possibility that our approach could be generalized to other visual search tasks and to complex visual analytics tasks. In particular, since users' interactions in visual analytics tasks have been shown to encode higher-level reasoning [39], we envision that our technique could be applied to other subtasks in visual analytics as well. Specifically, we consider the Waldo task as a form of the data search-and-filter task in the Pirolli and Card Sensemaking Loop [133]. We plan on extending our technique to analyzing user's interactions during other phases of the analytic process such as information foraging, evidence gathering and hypothesis generation.

7.9.2 Personality

Being able to demonstrate that there is signal in this interaction data that encodes personality factors is exciting. However, none of the results for personality factors are as strong as those for completion time. Not only are the overall accuracy scores lower, but we found that in examining the time-based scores (as in Section 7.7), for many personality factors, there was not a persistent trend that more data helped the machine learning (Figure 7.6 shows one of the stronger examples where there is a trend).

While the prediction accuracies are low, our results are consistent with prior findings [151] in the human-computer interaction community on individual differences research. Taken together, this suggests that although personality and cognitive traits can be recovered from users' interactions, the signals can be noisy and inconsistent. In order to better detect these signals, we plan to: (1) explore additional machine learning techniques, like boosting [142] for leveraging multiple learners together, and (2) apply our techniques to examine interactions from more complex visual analytics tasks. We expect the latter to amplify results as Chapter 4 have shown that personality trait effects are dampened when the task is simple. In Chapter 4, for complex inferential tasks, the effects were more pronounced and potentially easier to detect.

7.9.3 Future Work

This work is a first step in learning about users live from their interactions, and leaves many exciting questions to be answered with further research. The ability to classify users is interesting on its own, but an adaptive system could test the feasibility of applying this type of results in real time. For example, since locus of control can affect how people interact with visual representations of data (Chapter 4) , a system that could detect this personality trait could adapt by offering the visualization expected to be most effective for the individual user. Different cognitive traits may prove more fruitful for adaptation, but even completion time could be used to adapt, by giving new users advice if they start to follow strategies that would lead to their classification as slow.

Further, of the data representations we evaluated, only the mouse events, the lowest-

level interactions, encode any information about time passing during the task. The other representations do not encode the time between states or button presses, but that information could be useful for a future study. For our sequence-based analysis, our approach was to pair n-grams with decision trees for readability, but there are plenty of existing treatments of sequence data that remain to be tried for this type of data classification on visual analytic tasks, including sequence alignment algorithms, and random process models, e.g. Markov models. Finally, in this work we focused on encoding one aspect of data or interface at a time, but combining feature spaces could be powerful. In fact, in experimenting with a feature space that leverages multiple types of encodings, we achieved 96% accuracy on completion time with mean_nomed splitting[1].

The experimental task was a simple version of a basic visual analytics sub-task. Our results could be strengthened by expanding the experiment to test Waldo in different locations, or different stimuli like maps with buildings and cars. The breadth of applicability could be evaluated by testing other elementary visual analytics tasks such as using tables to find data or comparing values through visual forms.

Our plans to extend this work expand on three fronts: (1) evaluating additional personal traits, like cognitive factors such as working memory, to our analyses, (2) trying further machine learning algorithms and encodings to learn from more of the information being collected, like the times of the interactions and (3) extending experiments with different tasks including deeper exploration of visual search. We believe there are many opportunities to extend this work, both experimentally and analytically.

7.10 Summary

In this chapter, we presented results of an online experiment we conducted where we recorded participants' mouse interactions as they played the game Where's Waldo. We broke the users into groups by how long it took them to find Waldo (completion time) and their personality traits. Visualizing the participants views of the data, we showed that there are differences in strategies across groups of users. We then applied machine learning tech-

[1] Specifically, we tested a modified state space encoding where the zoom level information is replaced by an identifier of the button click that caused the state.

Table 7.4: Additional SVM Results - all results are calculated using leave-one-out cross validation.

Data Representation	Class Split	Classifier	Accuracy (%)
Completion Time			
edge space	mean_nomed	SVM_{poly}	87
	mean	SVM_{poly}	72
mouse events	mean_nomed	SVM_{poly}	88
	mean	SVM_{poly}	82
LOC			
edge space	mean	SVM_{poly}	62
state space	mean_nomed	SVM_{poly}	63
state $space_{PCA}$	mean_nomed	SVM	63
Neuroticism			
edge space	mean_nomed	SVM_{poly}	68
state $space_{PCA}$	mean_nomed	SVM	68

niques, and demonstrated that we can accurately classify the participants based on their completion time using multiple representations of their interactions: visualization states, low-level mouse events, and sequences of interface button clicks. By examining artifacts of our machine learning work with these sequences, we were able to identify short subsequences of interactions that identify groups of users. These human-readable classifier results hint at user strategies across groups. We were also able to detect and classify the participants based on some personality factors: locus of control, extraversion, and neuroticism. Finally, we showed the dependence of the machine learning results on the observation time of the participants.

7.11 Appendix: Extended Results

Though we demonstrated that completion time and personality can be modeled from raw interactions can be done directly with off-the-shelf tools using default settings, further attention to detail can yield stronger classifiers. In this appendix, we discuss some additional results that we achieved by tuning the algorithms, including applying principal component analysis (PCA), and optimizing the parameters of support vector machines (SVMs).

The SVM algorithm is sensitive to a slack parameter [34] and to the choice of kernel. Common practice is to address this by using a parameter search to find the best parameter

values [72]. In the context of deploying the best possible classifier for a given dataset, that entails simply trying different choices of the parameter (or sets of parameters) and evaluating the classifiers until the best can be reported. Since our goal is instead to evaluate the classifiers and encodings themselves for this type of data, we take the approach of validating the algorithm of classifier+param-search. As usual for cross validation, the data is split into k folds. Each fold takes a turn as the test data, while the other folds are used for training, providing k samples of accuracy to be averaged for a total score. In testing a classifier+param-search algorithm, the algorithm being evaluated on one fold is one that chooses a parameter by testing which value produces the best classification result. To evaluate "best classification result", another (nested) cross validation is needed. The original fold's training data is split into folds again and cross validation is used over those inner folds to pick the optimal parameter. Weka implements a more sophisticated version of this practice that allows optimizing two parameters at once (generally referred to as grid search) and uses optimizing heuristics to limit the number of evaluations [130]. We have used this implementation to run a grid search that optimizes over (1) slack parameter and (2) degree of polynomial for kernel (1 or 2, i.e. linear or quadratic). In Table 7.4, this classifier is called SVM_{poly}. This table shows highlights of the results that we produced with this technique.

Another helpful factor in working with SVMs on high-dimensional data is principal component analysis. PCA projects the high-dimensional data into a lower-dimensional space defined by the eigenvectors of the original data. The number of eigenvectors is chosen to make sure that 95% of the variance in the data is accounted for in the low-dimensional approximation[2]. Applying PCA to the data space was particularly helpful in data representations like state space, which has a high degree of dimensionality. In Table 7.4, data representations with PCA applied are indicated by the subscript PCA.

Overall, the results in Table 7.4 show cases in which our tuning produced higher-accuracy classifiers, and revealed signal with feature spaces or class splitting criteria that otherwise could not encode certain traits. The completion time results for the edge space and mouse event feature spaces are improvements of up to 32%. Specifically with edge space encoding and mean split, SVM_{poly} offers 82% accuracy instead of 62% with off-

[2]We used the Weka filter implementation with this option enabled.

the-shelf SVM. In our earlier analyses, we did not find sufficient signal to report on LOC with any state-based encodings, but using PCA or parameter search makes that possible. Through applying standard methods for tuning SVM, we gained higher accuracy over our existing results, and demonstrated connections between encodings and traits that were otherwise obscured.

Chapter 8

Discussion

The effects of individual differences on user interaction is a topic that has been explored for the last 25 years in HCI. Though the progress made by this dissertation is undeniable, one limitation is that every cognitive factor that affects visualization performance is not considered. For instance, in Chapter 3 we focused on spatial ability alone and did not consider how differences in personality, working memory, or previous experience can also affect visualization performance. As stated by Yi in his position statement in 2010, the visualization community has yet to employ a comprehensive and standardized model for measuring individual differences such that researchers can better understand how factors in individual differences interact with each other and with existing evaluation techniques [172].

In this chapter, we make a first-step toward a solution by introducing the ICD^3 Model (Individual Cognitive Differences) - a 3-dimensional model that encompasses the cognitive facets of individual indifferences. A necessary step in attempting to define a model of individual cognitive differences was to seek an underlying structure of previous research by identifying which factors are dependent and which are independent of one another. By surveying the existing literature, we propose that individual differences can be categorized into three orthogonal dimensions: cognitive traits, cognitive states and experience/bias.

Cognitive traits are user characteristics that remain constant during interaction with a visual analytic system. Factors such as personality, spatial visualization ability, and perceptual speed are all examples of cognitive traits. These have been shown to correlate with

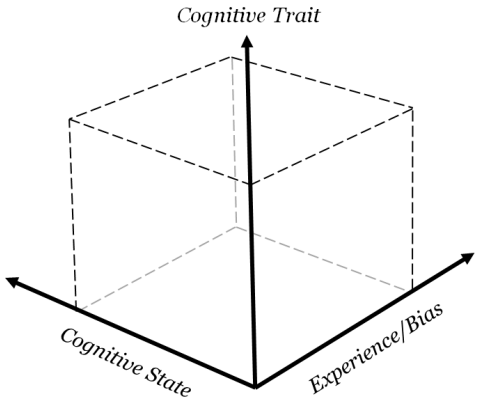


Figure 8.1: The ICD^3 categorizes individual cognitive differences in three orthogonal dimensions: Cognitive Traits, Cognitive States, and Experience/Bias

a user's ability to interact with a visualization [24, 32, 59, 162, 177] and can be generalized to predict the behavioral patterns of users with different cognitive profiles.

Cognitive states, on the other hand, are the aspects of the user that may change during interaction and include situational and emotional states, among others. Research has shown that a user's performance can be significantly altered by changes in their emotional state [6, 49, 93, 131, 141, 146], and the importance of combining workload with performance metrics has been noted for decades [74, 126, 171]. Although cognitive states are difficult to measure because of their volatility, they provide important contextual information about the factors affecting user performance that can not be described through cognitive traits alone.

Cognitive states and traits can describe a significant portion of a user's cognitive process but they are not comprehensive; experience and biases can also affect cognition. Intuitively, we think of experience and bias separately, but they both describe learned experiences that can affect behaviour when familiar problems arise, and are therefore not orthogonal. Although there has been little work about the impact of experience/bias on interaction with visual analytics systems, previous studies have shown that learned behavior such as confirmation bias can significantly affect performance and decision-making [69].

Taken together, these three dimensions can create a model that encapsulates the cognitive aspects of individual differences (Figure 8.1). Similar to how analyzing state

and trait alone would disregard potential performance gains from expertise, ignoring any one dimension of the model would also result in an incomplete description of performance. For example, analyzing only expertise and traits ignores changes that may be triggered by workload or emotions (cognitive state). Thus, the model is only complete if all three dimensions are considered. By using ICD^3, evaluators can identify what factors must be controlled in an experiment and which should be included as independent variables. The community can also begin to evaluate visualizations using this common platform and be able to better reproduce and extend each other's research.

8.1 The ICD^3 Model

In light of the three dimensions that we have discussed, we believe that a structured model would be useful in order to describe individual cognitive differences when users interact with visualizations.

We therefore propose a three dimensional model that is composed of cognitive traits, cognitive states, and experience/bias - the ICD^3 model. In this model, each orthogonal dimension would represent an individual difference of a user thereby allowing researchers to describe or perhaps even predict a user's ability to interact with a visualization, by knowing where that individual lies along the three different axes. This would allow for not just isolated cognitive factors, but for the *interaction* of the user's different cognitive abilities.

Figure 8.2 gives a hypothetical example of a user looking at percentage judgments in treemaps. The cognitive state in this example is the user's workload, their cognitive trait is their working memory capacity, and their experience/bias is how experienced they are with treemaps. An ICD^3 model would show that if the user is an expert, has a low workload, and has a high working memory capacity, then they have higher performance and abilities with percentage judgment in treemaps. Conversely, if the user is overloaded with work, has a naturally low working memory capacity, and has no experience of treemaps, then they will be less effective in performing that task.

After defining the visualization, task, and cognitive factors, a set of experiments can then be run in which participant workload, working memory capacity, and experience

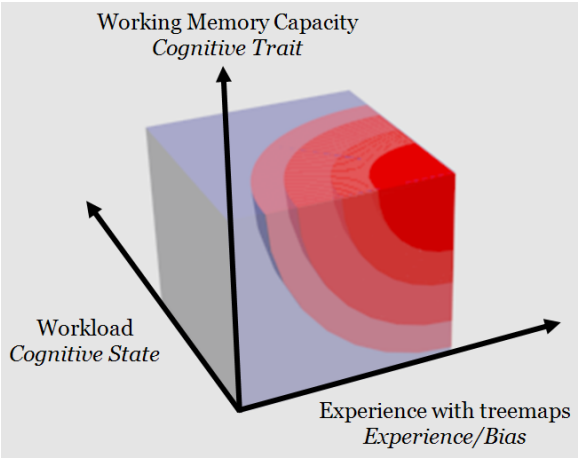


Figure 8.2: An example of how an ICD^3 model might be constructed. We map the interaction of workload, working memory capacity, and experience on performance of percentage judgments in treemaps. Darker red represents better performance.

is varied. For each interaction of factors, performance is recorded in the instance at the appropriate coordinates. Given enough data, we construct a descriptive topology of performance for a task and visualization.

Unfortunately, the interaction of cognitive facets is ordinarily much more nuanced than depicted in Figure 8.2. For the sake of simplicity, we chose working memory capacity, workload, and experience because their impact on performance is relatively straightforward. But in practice, we have little knowledge of how other combinations of state, trait, and experience/bias influence interaction with a visualization.

For example, some studies have shown that extraverts and introverts perform differently when they receive positive or negative feedback about a task, thus modifying their cognitive state [12]. Introverts tend to perform well when given positive feedback and worse when given negative feedback. Reciprocally, extraverts perform worse than intraverts given positive feedback, but their performance improves under negative feedback. This exemplifies why it is important to consider the interaction of state and trait.

However, other studies have suggested that people with an external locus of control (LOC), which is correlated with extraversion [121], perform better in visualizations where they have had no previous experience than people with an internal LOC [177]. This demonstrates how trait and experience can interact to influence performance.

Each of these examples provide a two dimensional snapshot of how cognitive dimensions can impact performance. But how do we combine the knowledge of these two studies? How would performance be impacted when an experienced intravert is given negative feedback, or an inexperienced extravert is given positive reinforcement during a task? Thus, a key attribute of the ICD^3 model is that limiting the scope of evaluation to any two of the three described dimensions leaves an incomplete and potentially misleading model of performance:

- Analyzing state and trait without experience ignores performance gains by expertise
- Analyzing state and experience without trait ignores interaction differences that are driven by personality or inherent cognitive strengths (e.g. spatial ability)
- Evaluating experience and trait without state disregards the moment-to-moment cognitive changes in the user that could be driven by emotion or workload

While instances of the ICD^3 model should be constructed for a explicit task and visualization, we imagine that the interaction of certain cognitive factors will be generalizable across visual forms (and tasks). In the next section, we explore the implications of the ICD^3 for design.

8.2 Toward Adaptive Interfaces

One important advantage of understanding individual users' cognitive states, traits, and biases as a cohesive structure is that this opens up the possibility of developing adaptive, mixed-initiative visualization systems [152]. Similar mixed-initiative systems were proposed in the HCI community by Horvitz in 1999 [71]. As noted by Thomas and Cook in *Illuminating the Path* [152], an important direction in advancing visual analytics research is the development of an automated, computational system that can assist a user in performing analytical tasks. However, most visualization systems today are designed in a one-size-fits-all fashion without the ability to adapt to different users' analytical needs into the design.

Creating such mixed-initiative visualization systems is particularly difficult as visualization are often designed to support complex thought and decision-making. Still, there

is some evidence that successful adaptive systems can significantly improve a user's ability in performing complex tasks. In the recent work by Solovey et al. [148], the authors show that with the use of a brain imaging technology (fNIRS) to detect a user's cognitive states the system can adapt the amount of automation and notably improve the user's ability in performing a complicated robot navigation task. Afergan et al. also used fNIRS to detect mental states of unmanned aerial vehicles (UAVs) operators as the completed complex UAV navigation tasks [2]. They designed a system that successfully detects when the operator is in the state of boredom or high workload, and automatically increase or reduce workload depending on the operator's cognitive load [2].

It is clear that adaptive systems can offer new possibilities for visualization research and development, but more work is necessary to model *how* and *when* a system should adapt to a user's needs. In general, there are two ways to perform real-time system adaptations: 1) **dynamic back-end adaptations**, and 2) **dynamic front-end adaptations**. Both concepts have existed for decades in the AI and HCI communities [79].

Dynamic back-end adaptation: This is an unobtrusive approach to adaptation. In these systems, the current display remain unchanged but the system performs additional tasks in the background to support the user. For instance, a system could predict a user's search strategy and perform pre-fetching or pre-computation when the data are large. Existing work in the Database community shows that this is possible. Battle et al. demonstrated how pre-fetching mechanisms can be informed through analyzing users' interactions [7]. The system can also provide help or addition information in a separate window. From the HCI community, the Ambient Help [110] system displays information that may be relevant to a user's current task, and work by Billsus et al. [10] suggests improving such proactive information systems by allowing users to adjust their obtrusiveness.

Dynamic front-end adaptation: In these systems, the display will dynamically adjust itself to the support a specific user or task. For instance, to support exploratory tasks a system could assist a user by highlighting exploration paths that a user may be unlikely to explore, or recommend visualizations that are better suited for a detected analysis pattern.

Relevant to this type of adaption, Sears and Schneiderman proposed split menus where the most frequently used menu items would percolate to the top of the list in order to facilitate faster access [143]. A popular example of similar menu adaptation is the Microsoft Smart Menus which was introduced in the Windows 2000 operating system. Also the in HCI community, Jefferson and Harvey successfully demonstrated how a system's graphical presentation can be adapted based on a user's color blindness [81, 82]. Recent work by Carenini et al. explored different ways of adapting information visualizations and demonstrated how interventions can significantly improve performance [19].

By monitoring the user and intelligently tailoring situationally appropriate information to the user, we can create next generation user interfaces that better support the user's analytics process. Furthermore, we can begin to design visualization systems that actually leverage the acuity of the human visual system as well as our capacity to understand and reason about complex data. Such systems may be able to overcome (or even leverage) some of the limitations imposed by the human brain such as limited working memory, bias, and fatigue.

8.3 Limitations and Future Work

Creating a precise model of individual differences is a daunting task. From the literature, we see that even the slightest deviations between people can influence performance on a visualization. Cognitive states may interact with and manipulate each other - for example, emotional state has been shown to impact working memory - and people simultaneously bring many traits and experiences to the table each time they see a visualization. Furthermore, there are almost certainly cognitive traits, states, and experiences that impact interaction significantly more than others.

While we do not believe that these problems impact the orthogonality of our proposed model, it illuminates the potential dependency of factors within each dimension, increasing the difficulty of predicting human interaction. We highlight at least two future areas of research that will be critical to addressing these challenges.

First, discovering new and unobtrusive methods to capture cognitive state, trait,

and experience/bias will ultimately drive research in individual cognitive differences. For example, in Chapter 7, we saw that we can detect user attributes by analyzing their click steam data. Recent advances in non-intrusive physiological sensors that detect emotional states, such as the Affectiva Q-Sensor [134], will enable future studies into the impact of emotional state and visualization performance. In real-world scenarios, it is unrealistic to expect users to be subjected to a deluge of forms and intrusive monitoring equipment. The simple act of filling out personality surveys or applying brain sensing equipment is enough to potentially modify cognitive state (or introduce biases) before interaction. It should be a central goal to develop new and effective ways in which we can automatically detect these states.

Second, finding dominant individual cognitive factors both within dimensions and between dimensions should limit the sheer volume of cognitive tests necessary to describe interaction. For example, if participants have a low working memory capacity, their locus of control might not matter given a certain task and a visualization. If this is true, then detecting and adapting to locus of control may be unnecessary. Similarly, we suspect that a person's experiences and biases may impact performance more than many other cognitive traits and states. Thus, if we know a person is an expert at a simple task, emotional state might be irrelevant. Identifying these dominant factors should reduce the number of interactions between cognitive factors.

The generalizability of cognitive states, cognitive traits, experiences/biases on performance in visualization has yet to be seen. As a result, the ICD^3 model takes a conservative approach by specifying an exact set of cognitive factors and requiring tests to be performed on a fixed task and fixed visualization. By identifying important factors or important interactions between factors, we hope to construct new metrics in the future that are more predictive of interaction with a visualization.

8.4 Summary

We have made initial steps towards a model that captures the various cognitive aspects that affect visualization performance by dividing them into three dimensions: cognitive states,

cognitive traits, and experience/bias. Furthermore, we have discussed how each of these dimensions are orthogonal to each other, meaning that during visualization interaction, a user may exhibit different values for states, traits, or experience/biases. Each of the dimensions are found to influence cognitive processes related to visualization, such as reasoning and perception. The ICD^3 model provides a sample space for experiments involving visualizations, so that we may form a better understanding of the cognitive underpinnings of visualization.

Chapter 9

Conclusion

In this dissertation, we laid the foundation for developing next generation visualization systems which will detect and adapt to a specific user's cognitive needs. We demonstrated how individual differences can impact reasoning with real-world visualizations for medical risk communication, and saw that spatial ability influences reasoning across different representations of medical test statistics. We also saw that the personality trait locus of control is a significant predictor of speed and accuracy with a variety of layout styles for hierarchical visualizations. Chapter 5 demonstrated the solidity of the relationship between locus of control and performance by showing how changes in locus of control can predictably influence performance. Altogether, the work in Chapters 3 , 4 and 5 reveal how nuances in individual characteristics can determine the effectiveness of a visualization tool.

In Chapters 6 and 7, we explored the feasibility of detecting user attributes from their interactions with a visualization tool. We saw that locus of control was a significant predictor of search strategies with hierarchical visualizations, and demonstrated how we can leverage advances in Machine Learning to automatically detect user attributes. We found that it is indeed possible to detect user attributes in real-time, thus supporting the promise of adaptive visualization systems.

Finally, we introduced the ICD^3 model which categorizes individual cognitive differences into three orthogonal dimensions: Cognitive Traits, Cognitive States, and Experience/Bias. We demonstrated the utility of this model for comparing and evaluating visualizations with respect to the impact individual cognitive differences, and discussed different

adaptive techniques for visualization systems. In closing we examined the model's limitations and avenues for future work. Through this research, we have made significant advances toward creating personalized visualization systems.

Bibliography

[1] Ahmad Abdel-Hafez and Yue Xu. A survey of user modelling in social media websites. *Computer and Information Science*, 6(4):59, 2013.

[2] Daniel Afergan, Evan M Peck, Erin T Solovey, Andrew Jenkins, Samuel W Hincks, Eli T Brown, Remco Chang, and Robert JK Jacob. Dynamic difficulty using brain metrics of workload. In *In Proc. ACM CHI 2014, ACM Press*. Citeseer, 2014.

[3] Bryce Allen. Individual differences and the conundrums of user-centered design: Two experiments. *Journal of the American Society for Information Science*, 51(6):508–520, 2000.

[4] Carl R. Anderson. Locus of control, coping behaviors, and performance in a stress setting: A longitudinal study. *Journal of Applied Psychology*, 62(4):446–451, 1977.

[5] Gennady Andrienko, Natalia Andrienko, Michael Burch, and Daniel Weiskopf. Visual analytics methodology for eye movement studies. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 18(12):2889–2898, 2012.

[6] F Gregory Ashby, Vivian V Valentin, and AU Turken. The effects of positive affect and arousal on working memory and executive attention. *Advances in Consciousness Research*, 44:245–288, 2002.

[7] Leilani Battle, Remco Chang, and Michael Stonebraker. Dynamic prefetching of data tiles for interactive visualization. *ACM Special Interest Group on Management of Data (SIGMOD)*.

[8] Louis Bavoil, Steven P Callahan, Patricia J Crossno, Juliana Freire, Carlos E Scheidegger, Cláudio T Silva, and Huy T Vo. VisTrails: enabling interactive multiple-view visualizations. In *Visualization, 2005. VIS 05. IEEE*, pages 135–142. IEEE, 2005.

[9] Jacques Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[10] Daniel Billsus, David M Hilbert, and Dan Maynes-Aminzade. Improving proactive information systems. In *Proceedings of the 10th international conference on Intelligent user interfaces*, pages 159–166. ACM, 2005.

[11] Pradipta Biswas and Peter Robinson. A brief survey on user modelling in hci. In *Proc. of the International Conference on Intelligent Human Computer Interaction (IHCI) 2010*, 2010.

[12] John Boddy, Annabel Carver, and Kevin Rowley. Effects of positive and negative verbal reinforcement on performance as a function of extraversion-introversion: Some tests of gray's theory. *Personality and Individual Differences*, 7(1):81–88, 1986.

[13] Gary L Brase. Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, 23(3):369–381, 2009.

[14] Eli T Brown, Jingjing Liu, Carla E Brodley, and Remco Chang. Dis-function: Learning distance functions interactively. In *Proceedings of the IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 83–92. IEEE, 2012.

[15] Eli T Brown, Alvitta Ottley, Helen Zhao, Quan Lin, Richard Souvenir, Alex Endert, and Remco Chang. Finding waldo: Learning about users from their interactions. *IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS*, 20(12), 2014.

[16] Sandra M Brown, Julie O Culver, Kathryn E Osann, Deborah J MacDonald, Sharon Sand, Andrea A Thornton, Marcia Grant, Deborah J Bowen, Kelly A Metcalfe, Harry B Burke, et al. Health literacy, numeracy, and interpretation of graphical breast cancer risk estimates. *Patient education and counseling*, 83(1):92–98, 2011.

[17] Tad T Brunyé, Holly A Taylor, David N Rapp, and Alexander B Spiro. Learning procedures: The role of working memory in multimedia learning experiences. *Applied Cognitive Psychology*, 20(7):917–940, 2006.

[18] Peter Brusilovsky and Eva Millán. User models for adaptive hypermedia and adaptive educational systems. In *The adaptive web*, pages 3–53. Springer-Verlag, 2007.

[19] Giuseppe Carenini, Cristina Conati, Enamul Hoque, Ben Steichen, Dereck Toker, and James Enns. Highlighting interventions and user differences: informing adaptive information visualization support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1835–1844. ACM, 2014.

[20] Simon Cassidy and Peter Eachus. Learning style, academic belief systems, self-report student proficiency and academic achievement in higher education. *Educational Psychology*, 20(3):307–320, 2000.

[21] Pierre Chalfoun and Claude Frasson. Subliminal priming enhances learning in a distant virtual 3d intelligent tutoring system. *IEEE Technology and Engineering Education (ITEE)*, 3(4):125–130, 2008.

[22] Gretchen B Chapman, Jingjing Liu, et al. Numeracy, frequency, and bayesian reasoning. *Judgment and Decision Making*, 4(1):34–40, 2009.

[23] Judy M Chartrand, Melissa L Rose, Timothy R Elliott, Cheri Marmarosh, and Susan Caldwell. Peeling back the onion: Personality, problem solving, and career decision-making style correlates of career indecision. *Journal of Career Assessment*, 1(1):66–82, 1993.

[24] Chaomei Chen and Mary Czerwinski. Spatial ability and visual navigation: An empirical study. *New Review of Hypermedia and Multimedia*, 3(1):67–89, 1997.

[25] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[26] Cheryl A Cohen and Mary Hegarty. Individual differences in use of external visualisations to perform an internal visualisation task. *Applied Cognitive Psychology*, 21(6):701–711, 2007.

[27] Philip R Cohen and C Raymond Perrault. Elements of a plan-based theory of speech acts. *Cognitive science*, 3(3):177–212, 1979.

[28] William G. Cole. Understanding bayesian reasoning via graphical displays. In *ACM SIGCHI Bulletin*, volume 20, pages 381–386. ACM, 1989.

[29] William G Cole and Janet E Davidson. Graphic representation can lead to fast and accurate bayesian reasoning. In *Proceedings/the... Annual Symposium on Computer Application [sic] in Medical Care. Symposium on Computer Applications in Medical Care*, pages 227–231. American Medical Informatics Association, 1989.

[30] Cristina Conati, Giuseppe Carenini, Enamul Hoque, Ben Steichen, and Dereck Toker. Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. In *Computer Graphics Forum*, volume 33, pages 371–380. Wiley Online Library, 2014.

[31] Cristina Conati, Giuseppe Carenini, Dereck Toker, and Sébastien Lallé. Towards user-adaptive information visualization. In *AAAI*, pages 4100–4106, 2015.

[32] Cristina Conati and Heather Maclaren. Exploring the role of individual differences in information visualization. In *Proceedings of the working conference on Advanced visual interfaces*, pages 199–206. ACM, 2008.

[33] Lynne Cooke. Is the mouse a "poor man's eye tracker"? In *Annual Conference-Society for Technical Communication*, volume 53, page 252, 2006.

[34] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[35] Leda Cosmides and John Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73, 1996.

[36] Paula Cowley, Lucy Nowell, and Jean Scholtz. Glass box: An instrumented infrastructure for supporting human interaction with information. In *Proceedings of the Thirty-Eighth Annual Hawaii International Conference on System Sciences*, pages 296c–296c. IEEE, 2005.

[37] Rossana De Beni, Francesca Pazzaglia, Valerie Gyselinck, and Chiara Meneghetti. Visuospatial working memory and mental representation of spatial descriptions. *European Journal of Cognitive Psychology*, 17(1):77–95, 2005.

[38] M Brent Donnellan, Frederick L Oswald, Brendan M Baird, and Richard E Lucas. The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, 18(2):192, 2006.

[39] Wenwen Dou, Dong Hyun Jeong, Felesia Stukes, William Ribarsky, Heather Richter Lipford, and Remco Chang. Recovering reasoning process from user interactions. *IEEE Computer Graphics & Applications*, 2009.

[40] David M. Eddy. Probabilistic reasoning in clinical medicine: Problems and opportunities. *In D. Kahneman, P. Slovic, & A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases*, pages 249–267, 1982.

[41] Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. *Transactions on Internet Technology (TOIT)*, 3(1):1–27, 2003.

[42] Ruth B Ekstrom, John W French, Harry H Harman, and Diran Dermen. Manual for kit of factor-referenced cognitive tests. *Princeton, NJ: Educational Testing Service*, 1976.

[43] Niklas Elmqvist, Andrew Vande Moere, Hans-Christian Jetter, Daniel Cernea, Harald Reiterer, and TJ Jankun-Kelly. Fluid interaction for information visualization. *Information Visualization*, 10(4):327–340, 2011.

[44] Alex Endert, Patrick Fiaux, and Chris North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 473–482. ACM, 2012.

[45] Nicholas Epley and Thomas Gilovich. Just going along: Nonconscious priming and conformity to social pressure. *Journal of Experimental Social Psychology*, 35(6):578–589, 1999.

[46] FH Farley. Individual differences in solution time in error-free problem solving. *British Journal of Social and Clinical Psychology*, 5(4):306–309, 1966.

[47] Maureen J. Findley and Harris M. Cooper. Locus of control and academic achievement: A literature review. *Journal of Personality and Social Psychology*, 44(2):419–427, 1983.

[48] Keren Fisher and Marie Johnstion. Experimental manipulation of perceived control and its effect on disability. *Psychology and Health*, 11(5):657–669, 1996.

[49] Barbara L Fredrickson. What good are positive emotions? *Review of general psychology*, 2(3):300, 1998.

[50] Hendrik Friederichs, Sandra Ligges, and Anne Weissenstein. Using tree diagrams without numerical values in addition to relative numbers improves students numeracy skills a randomized study in medical education. *Medical Decision Making*, 2013.

[51] Krzysztof Gajos and Daniel S Weld. Supple: automatically generating user interfaces. In *Proceedings of the Ninth International Conference on Intelligent User Interfaces*, pages 93–100. ACM, 2004.

[52] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology*, 28(2):210, 2009.

[53] Rocio Garcia-Retamero and Mirta Galesic. Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social science & medicine*, 70(7):1019–1025, 2010.

[54] Rocio Garcia-Retamero and Ulrich Hoffrage. Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine*, 83:27–33, 2013.

[55] Supriya Garg, Julia Eunju Nam, IV Ramakrishnan, and Klaus Mueller. Model-driven visual analytics. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 19–26. IEEE, 2008.

[56] Gerd Gigerenzer and Ulrich Hoffrage. How to improve bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684, 1995.

[57] Lewis R Goldberg, John A Johnson, Herbert W Eber, Robert Hogan, Michael C Ashton, C Robert Cloninger, and Harrison G Gough. The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, 40(1):84–96, 2006.

[58] David Gotz and Zhen Wen. Behavior-driven visualization recommendation. In *Proceedings of the 14th international conference on Intelligent user interfaces*, pages 315–324. ACM, 2009.

[59] Tera Marie Green and Brian Fisher. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In *IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 203–210. IEEE, 2010.

[60] Tera Marie Green, Dong Hyun Jeong, and Brian Fisher. Using personality factors to predict interface learning performance. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

[61] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[62] Sarah Hamlyn-Wright, Riccardo Draghi-Lorenz, and Jason Ellis. Locus of control fails to mediate between stress and anxiety and depression in parents of children with a developmental disorder. *Autism*, 11(6):489–501, 2007.

[63] Paul KJ Han, William MP Klein, Bill Killam, Tom Lehman, Holly Massett, and Andrew N Freedman. Representing randomness in the communication of individu-

alized cancer risk estimates: effects on cancer risk perceptions, worry, and subjective uncertainty about risk. *Patient education and counseling*, 86(1):106–113, 2012.

[64] Martin Handford. *Where's Waldo?* Little, Brown Boston, 1987.

[65] Lane Harrison, Drew Skau, Steven Franconeri, Aidong Lu, and Remco Chang. Influencing visual judgment through affective priming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2949–2958. ACM, 2013.

[66] Sarah T Hawley, Brian Zikmund-Fisher, Peter Ubel, Aleksandra Jancovic, Todd Lucas, and Angela Fagerlin. The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient education and counseling*, 73(3):448–455, 2008.

[67] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proc. CHI 2010*, pages 203–212. ACM Press, 2010.

[68] Mary Hegarty. Capacity limits in diagrammatic reasoning. In *Theory and application of diagrams*, pages 194–206. Springer, 2000.

[69] Richards J. Heuer. *Psychology of intelligence analysis*. United States Govt Printing Office, 1999.

[70] James Hollan, Edwin Hutchins, and David Kirsh. Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, 7(2):174–196, 2000.

[71] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166. ACM, 1999.

[72] Chih-Wei Hsu, Chih-Chung Chang, and Chih-Jen Lin. A practical guide to support vector classification. Technical report, National Taiwan University, Taipei 106, Taiwan, April 2010.

[73] Jeff Huang, Ryen White, and Georg Buscher. User see, user point: gaze and cursor alignment in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1341–1350, New York, NY, USA, 2012. ACM.

[74] Weidong Huang, Peter Eades, and Seok-Hee Hong. Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization*, 8(3):139–152, January 2009.

[75] Jessica Hullman and Nicholas Diakopoulos. Visualization rhetoric: Framing effects in narrative visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2231–2240, 2011.

[76] Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. A deeper understanding of sequence in narrative visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2406–2415, 2013.

[77] Michel C Ioannou, Karin Mogg, and Brendan P Bradley. Vigilance for threat: effects of anxiety and defensiveness. *Personality and Individual Differences*, 36(8):1879–1891, 2004.

[78] Panagiotis G Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16–21, 2010.

[79] Julie A Jacko. *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. CRC press, 2012.

[80] Robert JK Jacob, John J Leggett, Brad A Myers, and Randy Pausch. Interaction styles and input/output devices. *Behaviour & Information Technology*, 12(2):69–79, 1993.

[81] Luke Jefferson and Richard Harvey. Accommodating color blind computer users. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 40–47. ACM, 2006.

[82] Luke Jefferson and Richard Harvey. An interface to support color blind computer users. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1535–1538. ACM, 2007.

[83] Marie Johnston, Penny Gilbert, Cecily Partridge, and Jenny Collins. Changing perceived control in patients with physical disabilities: An intervention study with patients receiving rehabilitation. *British Journal of Clinical Psychology*, 31(1):89–94, 1992.

[84] Timothy A. Judge and Joyce E. Bono. Relationship of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology*, 86(1):80–92, 2001.

[85] Timothy A Judge, Amir Erez, Joyce E Bono, and Carl J Thoresen. Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of personality and social psychology*, 83(3):693, 2002.

[86] Nazanin Kadivar, Victor Chen, Dustin Dunsmuir, Eric Lee, Cheryl Qian, John Dill, Christopher Shaw, and R Woodbury. Capturing and supporting the analysis process. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 131–138. IEEE, 2009.

[87] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. Visual analytics: Definition, process, and challenges. In *Proceedings of the IEEE Conference on Information Visualization*, pages 154–175. Springer Berlin Heidelberg, 2008.

[88] Vince J Kellen. The effects of diagrams and relational complexity on user performance in conditional probability problems in a non-learning context. In *Doctoral Thesis*. DePaul University, 2012.

[89] Vince J Kellen, Susy Chan, and Xiaowen Fang. Facilitating conditional probability problems with visuals. In *Human-Computer Interaction. Interaction Platforms and Techniques*, pages 63–71. Springer, 2007.

[90] Doreen Kimura. *Sex and cognition*. MIT press, 2000.

[91] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceedings CHI*, pages 453–456, 2008.

[92] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. A data-frame theory of sensemaking. *Expertise out of context*, pages 113–155, 2007.

[93] Kitty Klein and Adriel Boals. Expressive writing can increase working memory capacity. *Journal of Experimental Psychology: General*, 130(3):520, 2001.

[94] Alfred Kobsa. User modeling: Recent work, prospects and hazards. *Human Factors in Information Technology*, 10:111–111, 1993.

[95] Alfred Kobsa. Generic user modeling systems. *User modeling and user-adapted interaction*, 11(1-2):49–63, 2001.

[96] Pranam Kolari and Anupam Joshi. Web mining: Research and practice. *Computing in science & engineering*, 6(4):49–53, 2004.

[97] Raymond Kosala and Hendrik Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations Newsletter*, 2(1):1–15, 2000.

[98] Robert Kosara and Jock Mackinlay. Storytelling: The next step for visualization. *Computer*, 46(5):44–50, 2013.

[99] Robert Kosara and Caroline Ziemkiewicz. Do mechanical turks dream of square pie charts? In *Proceedings of the 3rd BELIV'10 Workshop: BEyond time and errors: novel evaLuation methods for Information Visualization*, pages 63–70. ACM Press, 2010.

[100] Sébastien Lallé, Dereck Toker, Cristina Conati, and Giuseppe Carenini. Prediction of users' learning curves for adaptation while using an information visualization.

In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 357–368. ACM, 2015.

[101] Terran Lane and Carla E. Brodley. Temporal sequence learning and data reduction for anomaly detection. *ACM Transactions on Information and System Security*, 2(3):295–331, August 1999.

[102] Laurent M. Lapierre and Tammy D. Allen. Control at work, control at home, and planning behavior: Implications for work family conflict. Published online before print in Journal of Management, 2010.

[103] Bongshin Lee, Petra Isenberg, Nathalie Henry Riche, and Sheelagh Carpendale. Beyond mouse and keyboard: Expanding design considerations for information visualization interactions. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2689–2698, 2012.

[104] Sukwon Lee, Sung-Hee Kim, Ya-Hsin Hung, Heidi Lam, Youn-ah Kang, and Ji Soo Yi. How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):499–508, 2016.

[105] Sheena Lewis, Mira Dontcheva, and Elizabeth Gerber. Affective computational priming and creativity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 735–744. ACM, 2011.

[106] Zhicheng Liu, Nancy Nersessian, and John Stasko. Distributed cognition as a theoretical framework for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1173–1180, 2008.

[107] Aidong Lu, Ross Maciejewski, and David S Ebert. Volume composition using eye tracking data. In *Proceedings of the Eighth Joint Eurographics/IEEE VGTC conference on Visualization*, pages 115–122. Eurographics Association, 2006.

[108] Jock Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986.

[109] Laura Martignon and Christoph Wassner. Teaching decision making and statistical thinking with natural frequencies. In *Proceedings of the Sixth International Conference on Teaching of Statistics. Ciudad del Cabo: IASE. CD ROM*, 2002.

[110] Justin Matejka, Tovi Grossman, and George Fitzmaurice. Ambient help. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2751–2760. ACM, 2011.

[111] Roy A. Maxion. Masquerade detection using enriched command lines. In *Proceedings of the IEEE International Conference on Dependable Systems and Networks*, pages 5–14. IEEE, 2003.

[112] Mary H McCaulley. The myers-briggs type indicator: A jungian model for problem solving. *New Directions for Teaching and Learning*, 1987(30):37–53, 1987.

[113] Michael F McTear. User modelling for adaptive computer systems: a survey of recent developments. *Artificial intelligence review*, 7(3-4):157–184, 1993.

[114] Luana Micallef, Pierre Dragicevic, and J Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2536–2545, 2012.

[115] Anthony B Miller, Cornelia J Baines, Ping Sun, Teresa To, and Steven A Narod. Twenty five year follow-up for breast cancer incidence and mortality of the canadian national breast screening study: randomised screening trial. *BMJ: British Medical Journal*, 2014.

[116] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[117] T. Munzner. A nested process model for visualization design and validation. *IEEE Visualization and Computer Graphics*, 15(6):921–928, 2009.

[118] National Center for Biotechnology Information. NCBI genome database. http://www.ncbi.nlm.nih.gov/genome/. Accessed 11/18/2010.

[119] Chris North, Remco Chang, Alex Endert, Wenwen Dou, Richard May, Bill Pike, and Glenn Fink. Analytic provenance: process+ interaction+ insight. In *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems*, pages 33–36. ACM, 2011.

[120] Oded Nov and Chen Ye. Users' personality and perceived ease of use of digital libraries: The case for resistance to change. *Journal of the American Society for Information Science and Technology*, 59(5):845–851, 2008.

[121] Dennis W Organ. Extraversion, locus of control, and individual differences in conditionability in organizations. *Journal of Applied Psychology*, 60(3):401, 1975.

[122] Alvitta Ottley, R Jordan Crouser, Caroline Ziemkiewicz, and Remco Chang. Manipulating and controlling for personality effects on visualization tasks. *Information Visualization*, 2013.

[123] Alvitta Ottley, Blossom Metevier, Paul KJ Han, and Remco Chang. Visually communicating bayesian statistics to laypersons. In *Technical Report*. Tufts University, 2012.

[124] Alvitta Ottley, Evan M Peck, Lane T Harrison, Daniel Afergan, Caroline Ziemkiewicz, Holly A Taylor, Paul KJ Han, and Remco Chang. Improving bayesian reasoning: the effects of phrasing, visualization, and spatial ability. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):529–538, 2016.

[125] Alvitta Ottley, Huahai Yang, and Remco Chang. Personality as a predictor of user strategy: How locus of control affects search strategies on tree visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2015.

[126] Fred GWC Paas and Jeroen JG Van Merriënboer. The efficiency of instructional conditions: An approach to combine mental effort and performance measures. *Human Factors*, 35(4):737–743, 1993.

[127] Evan M Peck, Alvitta Ottley, Beste F Yuksel, Remco Chang, and Lane Harrison. ICD 3: Towards a 3-dimensional model of individual cognitive differences. *ACM BELIV'12: Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, 2012.

[128] Evan M M Peck, Beste F Yuksel, Alvitta Ottley, Robert JK Jacob, and Remco Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 473–482. ACM, 2013.

[129] C Raymond Perrault, James F Allen, and Philip R Cohen. Speech acts as a basis for understanding dialogue coherence. In *Proceedings of the 1978 workshop on Theoretical issues in natural language processing*, pages 125–132. Association for Computational Linguistics, 1978.

[130] Bernhard Pfahringer, Geoff Holmes, and fracpete. Class gridsearch, revision 9733. http://weka.sourceforge.net/doc.stable/weka/classifiers/meta/GridSearch.html, March 2014.

[131] Elizabeth A Phelps, Sam Ling, and Marisa Carrasco. Emotion facilitates perception and potentiates the perceptual benefits of attention. *Psychological science*, 17(4):292–299, 2006.

[132] William A Pike, John Stasko, Remco Chang, and Theresa A O'Connell. The science of interaction. *Information Visualization*, 8(4):263–274, 2009.

[133] Peter Pirolli and Stuart Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of the International Conference on Intelligence Analysis*, volume 5, pages 2–4, 2005.

[134] Ming-Zher Poh, Tobias Loddenkemper, Nicholas C Swenson, Shubhi Goyal, Joseph R Madsen, and Rosalind W Picard. Continuous monitoring of electrodermal activity during epileptic seizures using a wearable sensor. In *Engineering in*

Medicine and Biology Society (EMBC), 2010 Annual International Conference of the IEEE, pages 4415–4418. IEEE, 2010.

[135] Maja Pusara and Carla E Brodley. User re-authentication via mouse movements. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pages 1–8. ACM, 2004.

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

[137] Robert W Reeder and Roy A Maxion. User interface defect detection by hesitation analysis. In *Dependable Systems and Networks, 2006. DSN 2006. International Conference on*, pages 61–72. IEEE, 2006.

[138] Valerie F Reyna, Wendy L Nelson, Paul K Han, and Nathan F Dieckmann. How numeracy influences risk comprehension and medical decision making. *Psychological bulletin*, 135(6):943, 2009.

[139] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*, pages 2863–2872. ACM, 2010.

[140] Julian B. Rotter. Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs: General and Applied*, 80(1):1, 1966.

[141] Gillian Rowe, Jacob B Hirsh, and Adam K Anderson. Positive affect increases the breadth of attentional selection. *Proceedings of the National Academy of Sciences*, 104(1):383–388, 2007.

[142] Robert E Schapire. The strength of weak learnability. *Machine learning*, 5(2):197–227, 1990.

[143] Andrew Sears and Ben Shneiderman. Split menus: effectively using selection frequency to organize menus. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 1(1):27–51, 1994.

[144] Peter Sedlmeier and Gerd Gigerenzer. Teaching bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130(3):380, 2001.

[145] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1139–1148, 2010.

[146] Alexander J Shackman, Issidoros Sarinopoulos, Jeffrey S Maxwell, Diego A Pizzagalli, Aureliu Lavric, and Richard J Davidson. Anxiety selectively disrupts visuospatial working memory. *Emotion*, 6(1):40, 2006.

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

[148] Erin Solovey, Paul Schermerhorn, Matthias Scheutz, Angelo Sassaroli, Sergio Fantini, and Robert Jacob. Brainput: enhancing interactive systems with streaming fnirs brain input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2193–2202. ACM, 2012.

[149] David Spiegelhalter, Mike Pearson, and Ian Short. Visualizing uncertainty about the future. *Science*, 333(6048):1393–1400, 2011.

[150] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *ACM SIGKDD Explorations Newsletter*, 1(2):12–23, 2000.

[151] Ben Steichen, Giuseppe Carenini, and Cristina Conati. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, pages 317–328. ACM, 2013.

[152] James J Thomas and Kristin A Cook. *Illuminating the path: The research and development agenda for visual analytics*. IEEE Computer Society Press, 2005.

[153] Dereck Toker, Cristina Conati, Giuseppe Carenini, and Mona Haraty. Towards adaptive information visualization: on the influence of user characteristics. In *User Modeling, Adaptation, and Personalization*, pages 274–285. Springer, 2012.

[154] Dereck Toker, Cristina Conati, Ben Steichen, and Giuseppe Carenini. Individual user characteristics and information visualization: connecting the dots through eye tracking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 295–304. ACM, 2013.

[155] Jennifer Tsai, Sarah Miller, and Alex Kirlik. Interactive visualizations to improve bayesian reasoning. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 55, pages 385–389. SAGE Publications, 2011.

[156] Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[157] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[158] Barbara Tversky. Functional significance of visuospatial representations. *Handbook of higher-level visuospatial thinking*, pages 1–34, 2005.

[159] Barbara Tversky, Maneesh Agrawala, Julie Heiser, P. U. Lee, Pat Hanrahan, Doantam Phan, Chris Stolte, and MP Daniele. Cognitive design principles: From cognitive models to computer models. In L. Magnani, editor, *Model-based reasoning in science and engineering*, pages 1–20. King's College, 2006.

[160] Liad Uziel. The extraverted and the neurotic glasses are of different colors. *Personality and Individual Differences*, 41(4):745–754, 2006.

[161] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[162] Maria C Velez, Deborah Silver, and Marilyn Tremaine. Understanding visualization through spatial ability differences. In *IEEE Visualization*, pages 511–518. IEEE, 2005.

[163] Kim J Vicente, Brian C Hayes, and Robert C Williges. Assaying and isolating individual differences in searching a hierarchical file system. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 29(3):349–359, 1987.

[164] Wolfgang Wahlster and Alfred Kobsa. *User models in dialog systems*. Springer, 1989.

[165] James J Walsh, John M Wilding, Michael W Eysenck, and John D Valentine. Neuroticism, locus of control, type a behaviour pattern and occupational stress. *Work & stress*, 11(2):148–159, 1997.

[166] C. Ware. *Information visualization: Perception for design*, volume 22. Morgan Kaufmann, 2004.

[167] Colin Ware. *Visual thinking: For design*. Morgan Kaufmann, 2010.

[168] H Gilbert Welch. Overdiagnosis and mammography screening. *Bmj*, 339, 2009.

[169] H Gilbert Welch and William C Black. Overdiagnosis in cancer. *Journal of the National Cancer Institute*, 102(9):605–613, 2010.

[170] Ling Xiao, John Gerth, and Pat Hanrahan. Enhancing visual analysis of network traffic using a knowledge representation. In *Proceedings of the IEEE Symposium On Visual Analytics Science And Technology (VAST)*, pages 107–114. IEEE, 2006.

[171] Yei-yu Yeh and Christopher D Wickens. Dissociation of performance and subjective measures of workload. *Human Factors*, 30(1):111–120, 1988.

[172] Ji Soo Yi. Implications of individual differences on evaluating information visualization techniques. In *Proceedings of the BELIV Workshop*, 2010.

[173] Yingbing Yu. Anomaly detection of masqueraders based upon typing biometrics and probabilistic neural network. *Journal of Computing Sciences in Colleges*, 25(5):147–153, 2010.

[174] Caroline Ziemkiewicz, R Jordan Crouser, Ashley Rye Yauilla, Sara L Su, William Ribarsky, and Remco Chang. How locus of control influences compatibility with visualization style. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 81–90. IEEE, 2011.

[175] Caroline Ziemkiewicz and Robert Kosara. Preconceptions and individual differences in understanding visual metaphors. In *Computer Graphics Forum*, volume 28, pages 911–918. Wiley Online Library, 2009.

[176] Caroline Ziemkiewicz, Alvitta Ottley, R Jordan Crouser, Krysta Chauncey, Sara L Su, and Remco Chang. Understanding visualization by understanding individual users. *Computer Graphics and Applications, IEEE*, 32(6):88–94, 2012.

[177] Caroline Ziemkiewicz, Alvitta Ottley, R Jordan Crouser, Ashley Rye Yauilla, Sara L Su, William Ribarsky, and Remco Chang. How visualization layout relates to locus of control and other personality factors. *Visualization and Computer Graphics, IEEE Transactions on*, 19(7):1109–1121, 2013.

[178] Coraline Ziemkiewicz and Robert Kosara. The shaping of information by visual metaphors. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1269–1276, 2008.