UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

COGNITIVE FIT IN VISUALIZING BIG DATA

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

EMRE YETGIN
Norman, Oklahoma
2015

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/14568

COGNITIVE FIT IN VISUALIZING BIG DATA

A DISSERTATION APPROVED FOR THE
MICHAEL F. PRICE COLLEGE OF BUSINESS

BY

_____
Dr.Teresa Shaft, Chair

_____
Dr. Laku Chidambaram

_____
Dr. Matthew Jensen

_____
Dr. Radhika Santhanam

_____
Dr. Chris Weaver

© Copyright by EMRE YETGIN 2015
All Rights Reserved.

Table of Contents

List of Tables vi

List of Figures viii

Abstract ix

Chapter 1: Introduction 1

Chapter 2: Theory Development 6

Visualization 6

Cognitive Fit Theory 11

Cognitive Fit Between Data Analysis Tasks and Visualizations 14

Characteristics of Big Data 25

Chapter 3: Methodology 31

Experiment 31

Participants 32

Stimulus Materials 33

Independent Variables 33

Task Type 33

Visualization Type 35

Volume 37

Variety 39

Control Variables 41

Color Blindness 41

Task Familiarity 41

Motivation 42

Visualization Ability 42
Dependent Variables 43
Eye Tracker Data 43
Task Performance 45
Pilot Study 46
Procedure 47
Chapter 4: Results 50
Eye Tracker Results 52
Task Performance Results 63
Chapter 5: Summary and Conclusions 77
Discussion 77
Limitations and Future Research Directions 85
Conclusions 90
References 92
Appendix A: Stimulus Materials 99
Appendix B: Solution Accuracy Calculation 105
Information Retrieval Tasks 105
Information Comparison Tasks 106
Information Integration Tasks 107
Appendix C: Visualization Ability Measure 109

List of Tables

Table 1. Major Visualization Taxonomies 7

Table 2. Major Data Analysis Taxonomies 17

Table 3. Examples for High-Level Data Analysis Tasks Taxonomy 18

Table 4. Summary of Experimental Treatments 32

Table 5. Data Analysis Problems Used in the Experiment 35

Table 6. Manipulation Checks 40

Table 7. Descriptive Statistics for Eye Tracker Data 45

Table 8. Descriptive Statistics for Information Retrieval Tasks 50

Table 9. Descriptive Statistics for Information Comparison Tasks 51

Table 10. Descriptive Statistics for Information Integration Tasks 51

Table 11. Descriptive Statistics for Control Variables 51

Table 12. Correlation Matrix for Dependent and Control Variables 52

Table 13. Multivariate Tests for Information Retrieval Fixation Count and Information Retrieval View Time 53

Table 14. Univariate Between-Subjects Effects for Information Retrieval Fixation Count and Information Retrieval View Time 55

Table 15. Multivariate Tests for Information Comparison Fixation Count and Comparison Retrieval View Time 57

Table 16. Univariate Between-Subjects Effects for Information Comparison Fixation Count and Information Comparison View Time 58

Table 17. Multivariate Tests for Information Integration Fixation Count and Information Integration View Time 61

Table 18. Univariate Between-Subjects Effects for Information Integration Fixation Count and Information Integration View Time 62
Table 19. Hypothesis Testing 64
Table 20. Multivariate Tests for Information Retrieval Solution Time and Information Retrieval Solution Accuracy 66
Table 21. Univariate Between-Subjects Effects for Information Retrieval Solution Time and Information Retrieval Solution Accuracy 67
Table 22. Multivariate Tests for Information Comparison Solution Time and Information Comparison Solution Accuracy 70
Table 23. Univariate Between-Subjects Effects for Information Comparison Solution Time and Information Comparison Solution Accuracy 71
Table 24. Multivariate Tests for Information Integration Solution Time and Information Integration Solution Accuracy 74
Table 25. Univariate Between-Subjects Effects for Information Integration Solution Time and Information Integration Solution Accuracy 75
Table 26. Results of Hypothesis Tests 76

List of Figures

Figure 1. The Cognitive Fit Model (Adapted from Vessey, 1991) 12

Figure 2. Conceptual Research Model 30

Figure 3. Volume X Variety Interaction on Information Retrieval Fixation Count..... 56

Figure 4. Variety X Volume Interaction on Information Comparison View Time..... 59

Figure 5. Visualization X Variety Interaction on Information Comparison View Time 60

Figure 6. Visualization X Volume Interaction on Information Retrieval Solution Accuracy 68

Figure 7. Visualization X Volume Interaction on Information Comparison Solution Accuracy 72

Figure 8. Treatment 1 (Discrete, Low Volume, Low Variety)..... 99

Figure 9. Treatment 2 (Continuous, Low Volume, Low Variety)..... 99

Figure 10. Treatment 3 (Discrete, High Volume, Low Variety) 100

Figure 11. Treatment 4 (Continuous, High Volume, Low Variety)..... 100

Figure 12. Treatment 5 (Discrete, Low Volume, High Variety) 101

Figure 13. Treatment 6 (Continuous, Low Volume, High Variety)..... 101

Figure 14. Treatment 7 (Discrete, High Volume, High Variety) 102

Figure 15. Treatment 8 (Continuous, High Volume, High Variety) 102

Figure 16. Treatment 9 (Singular, Low Variety)..... 103

Figure 17. Treatment 10 (Multiple, Low Variety) 103

Figure 18. Treatment 11 (Singular, High Variety) 104

Figure 19. Treatment 12 (Multiple, High Variety)..... 104

Figure 20. Image Pairs Used for Measuring Visualization Ability 109

Abstract

This dissertation examines the consequences of cognitive fit in visualizing big data. Specifically, it focuses on the interplay between different types of business data analysis tasks and visualization methods, and how the defining characteristics of big data (i.e., volume and variety) moderate the outcomes concerning data analysis performance (i.e., solution time and solution accuracy). A 12-cell repeated-measures laboratory experiment (n=145) using eye trackers is conducted to test the hypotheses. Data analysis performance is observed to improve when the information emphasized by a visualization method matches the specific information requirements for a data analysis task. Such improvements in data analysis performance are further amplified when the visualized information has high volume and variety.

This dissertation contributes to the literature in at least three ways. First, it improves our understanding of cognitive fit and how it manifests in analysts' problem solving behaviors when using visualization tools. This is done by analyzing participants' eye movement and gaze fixation patterns while they work with different types of data analysis tasks and visualization methods. Based on this analysis, this study proposes an objective method for assessing and measuring cognitive fit. Second, this study maps visualization characteristics to business data analysis task types, and informs the choice of visualization tools among an ever-increasing number of alternatives for supporting the complex problems faced by big data analysts. Third, this dissertation extends the cognitive fit theory to the big data context and highlights the relative importance of cognitive fit in this setting by demonstrating that increases in volume and variety amplify the task performance consequences of cognitive fit. The

limitations of the experiment conducted for this dissertation and the future research opportunities they present are discussed. The findings of this dissertation also can inform the development of new visualization tools and techniques based on task and data characteristics.

Keywords: Big Data Analytics, Business Data Analysis, Cognitive Fit, Eye Tracker, Image Theory, Information Extraction, Volume, Variety, Visualization.

Chapter 1: Introduction

In today's business world, organizations have to gather and analyze big data effectively, to create and maintain a certain level of business advantage (LaValle, Lesser, Shockley, Hopkins, and Kruschwitz, 2011). This is the reason why 65% of today's enterprise senior executives think that their organizations will become irrelevant and/or uncompetitive if they do not embrace big data soon (Columbus, 2015), with large organizations like General Electric spending over a billion dollars for developing their big data collection, storage, and analytics capabilities (Catts, 2012). In fact, organizations that are better at big data-driven decision-making are both more profitable and more productive than their competitors (McAfee and Brynjolfsson, 2012). This happens because big data provides organizations with many opportunities for unprecedented business insights, such as getting to know their customer base and understanding their spending habits better than ever before (Eaton, Deroos, Deutsch, Lapis, and Zikopoulos, 2012). Yet, the increases in the volume, variety, and velocity of a typical big dataset have made it more challenging to manage and make sense of the information it contains, compared to the traditional datasets organizations have been relying on before (Chen, Chiang, and Storey, 2012). These inherent characteristics make big data especially difficult to analyze using traditional methods and tools, such as simple data warehousing (Eaton et al., 2012).

Visualization (i.e., representing data visually on charts or maps) has long been an aid in aggregating otherwise incomprehensible information and presenting it in a way that can provide insights that are difficult if not impossible to obtain through other means (e.g., lists, tables, or summarizing statistics). This occurs because visualizing

information can facilitate data analysis by extending individuals' working memory and by making it easier to interpret the entirety of information relative to textual or numerical representation (Ware, 2004). However, despite its potential importance to big data analysis, the opportunities and challenges visualization presents in a big data setting remain mostly uninvestigated (Chen et al., 2012).

This is not to say that there has been no research on visualizing big data. Researchers have developed numerous visualization tools and techniques (e.g., McNab, Hess, and Valacich, 2011; Rinzivillo, Pedreschi, Nanni, Giannotti, Andrienko, and Andrienko, 2008) that are tailored for specific tasks or contexts (e.g., for emergency response dispatch systems or for analyzing physical trajectories), but whether these tools and techniques can be effectively used for other big data analysis tasks or contexts is not clear. In fact, one type of visualization tool that is very useful for a certain analysis task can be quite detrimental for another (e.g., see Goswami, Chan, and Kim, 2008 for mixed outcomes of visualization tools in spreadsheet error correction).

In this dissertation, I argue that the cognitive fit between the type of information required by a data analysis task and the information that is emphasized by a visualization tool determines the tool's usefulness. There has been minimal research on the interplay between data analysis task and visualization types in the context of big data analytics, and the consequences concerning data analysis performance. Building on this gap, the objective of this dissertation is to understand how visualizations can facilitate or hinder big data analysis, approaching from a cognitive fit perspective. Accordingly, the research question driving this study is:

RQ: How can visualization facilitate or hinder big data analysis, based on cognitive fit?

In answering this question, this dissertation seeks to highlight the relative importance of cognitive fit in a big data context and to inform the choice of visualization tools among an ever-increasing number of alternatives. Improving our understanding of how and why certain types of visualizations provide better support for different types of big data analytics tasks can facilitate the solution of the complex problems faced by big data analysts today, such as relating vast amounts of social media data (e.g., customers' 'like's, comments, locations, and browsing and searching behaviors) to customers' purchasing behaviors.

The study of cognitive fit has been limited to assumptions and experimental manipulations in past research (e.g., Dennis and Carte, 1998; Goswami et al., 2008; Vessey, 1991). In these studies (e.g., Vessey, 1991), cognitive fit was traditionally manipulated via experimental treatments based on theoretical arguments (e.g., graphical representations were expected to provide better cognitive fit for spatial tasks, compared to tabular representations). Then, cognitive fit was inferred to exist when the expected task performance improvements were observed, or when the participants self-reported that one type of visualization provided better support over another, without identifying or observing the exact mechanism through which cognitive fit affected task performance.

This dissertation extends cognitive fit theory to account for the impacts of visualization techniques and big data characteristics for different types of data analytics tasks, increasing the theory's robustness. In doing so, it contributes to the literature by

further improving our understanding of cognitive fit and how it manifests in analysts' problem solving behaviors when using visualization tools, hence identifying the mechanism through which it influences data analysis task performance. This is done by analyzing the eye tracker data collected during a laboratory experiment, and by studying participants' eye movement and gaze fixation patterns as they perform different types of business data analysis tasks while using different types of visualizations. Specifically, cognitive fit is assessed through the efficiency with which participants extract information from a given visualization while they solve the data analysis problems. Based on this analysis, this study proposes an objective method to capture cognitive fit, independent of participants' recall and reporting biases, in an effort to open the black box of cognitive fit in the context of big data visualization. The results of this study can also inform the development of new visualization tools based on task and data characteristics, plus guide researchers and analysts in mapping visualization methods to data analysis task types.

The rest of this dissertation is organized as follows: The next chapter (i.e., Chapter Two) summarizes the literature reviews conducted for identifying different types of visualizations and business data analysis tasks, and describes the two high-level taxonomies used to classify visualization methods and data analysis task types in this dissertation. It also introduces the Cognitive Fit theory, which is used as the rationale for the explanation regarding why certain types of visualizations are expected to provide better decision-making support for different types of data analysis tasks, based on the match between the information emphasized by visualizations and the information required by the tasks. Then, the defining characteristics of big data (Chen et al., 2012;

Eaton et al., 2012; McAfee and Brynjolfsson, 2012) are introduced, and their expected impacts on the task performance consequences (Eppler and Mengis, 2004) of cognitive fit are discussed.

Chapter Three introduces the methods used for this study and the laboratory experiment conducted to test the hypotheses and the research model. Specifically, it discusses in detail the pilot and main studies conducted, experimental procedures and manipulations, experimental treatments and stimulus materials, participants, plus the independent variables, the control variables, and the dependent variables used in the analyses.

Chapter Four describes the two different sets of analyses performed for testing the hypotheses, and presents the results for the hypothesis tests. Chapter Five provides a summary of the findings of this study, and discusses the theoretical and practical implications plus the limitations of this dissertation and suggested future research directions.

Chapter 2: Theory Development

This dissertation focuses on the cognitive fit between visualizations and business data analysis tasks in the context of big data analytics. Therefore, it is important to examine the different types of visualizations and business data analysis tasks identified in the literature, plus the defining characteristics of big data that make it unique and especially challenging to analyze. Accordingly, this chapter first describes the literature reviews conducted for visualization and data analysis task types, and then discusses the inherent characteristics of big data, which are expected to intensify the task performance consequences of the cognitive fit between visualizations and business data analysis task types.

Visualization

Visualization is defined as the computer-supported use of visual processing to gain better understanding of information (Card and Mackinlay, 1997). Due to the advantages it provides for data analysis, visualization has been a major component of decision support systems since the mid 1980s (Li, Feng, and Li, 2001). Traditionally, visualizations have been studied and categorized according to the type or characteristics of the data they are capable of or designed for representing. Table 1 provides a summary list of the major visualization taxonomies in the literature.

Table 1. Major Visualization Taxonomies

Taxonomy Basis	Categories of Visualization	Reference
Levels of data	3; (elementary, overall, intermediate)	Bertin, 1981
Data type × task type	7 × 7; (1D, 2D, 3D, temporal, multi-dimensional, tree, network) × (overview, zoom, filter, details-on-demand, relate, history, extracts)	Schneiderman, 1996
Data type × feedback type × form of interactivity	3 × 3 × 2; (raw, constructed, converted) × (past states, current state, potential states) × (direct manipulation, indirect manipulation)	Tweedie, 1997
Data stages and transformation	7; (data stage, data transformation, analytical abstraction stage, visualization transformation, visualization abstraction stage, visual mapping, view stage)	Chi and Riedl, 1998
Data type	2; (scientific visualization, information visualization)	Gershon, Eick, and Card, 1998
Design space	8; (scientific visualization, GIS-based visualization, multi-dimensional plots, multi-dimensional tables, information landscapes and spaces, node and link diagrams, trees, and text transforms)	Card, Mackinley, and Schneiderman, 1999
Visualization operators and techniques	36; (not listed due to space considerations)	Chi, 2000
Data type × modification × data structure × positioning	2 × 2 × 3 × 3; (raw, derived) × (original, distorted) × (ordered, hierarchical, network) × (overlapping, space-filling, separation)	Ward, 2002
Data type × data relationship structure × task type × interactivity type × user skill × context	3 × 5 × 7 × 2 × 2 × 5; (object, attribute, meta) × (linear, circular, ordered, unordered, lattice) × (overview, zoom, filter, details-on-demand, relate, history, extract) × (textual, graphic) × (novice, expert) × (experience, history, intent, need, device)	Pfitzner, Hobbs, and Powers, 2003
Design model × display attributes	2 × 3 × n ; (discrete, continuous) × (given, constrained, chosen)	Tory and Moller, 2004
Complexity × content area × point of view × thinking aid type × representation type	2 × 6 × 3 × 2 × 2; (low, high) × (data, information, concept, metaphor, strategy, compound knowledge) × (detail, overview, detail and overview) × (convergent, divergent) × (process, structure)	Lenger and Eppler, 2007

The most common categorization (Card, Mackinlay, and Schneiderman, 1999; Gershon, Eick, and Card, 1998), cited about 4,000 times, breaks visualizations down into two main branches: information visualization and scientific visualization. Information visualization refers to the abstract representation of non-physical (e.g., financial) information (i.e., information without an inherent mapping to physical space) (Card et al., 1999), while scientific visualization is usually based on physical information (i.e., information based on physical space coordinates) regarding concrete objects (e.g., geographical or anatomical data) and thus involves an inherent spatial component (Card and Mackinlay, 1997).

In a similar fashion, other major taxonomies have also categorized visualizations according to the dimensionality (i.e., (one-, two-, and three-dimensional data, temporal data, multi-dimensional data, tree data, and network data) (Schneiderman, 1996), levels (i.e., elementary, overall, or intermediate) (Bertin, 1981), or the kind (i.e, raw data, constructed data [data values derived from others], and converted data [data values converted into a new form]) (Tweedie, 1997) of information they can represent. Researchers have also expanded on these taxonomies by taking into account additional data and visualization attributes such as data stages and transformation (i.e., data stage, data transformation, analytical abstraction stage, visualization transformation, visualization abstraction stage, visual mapping, and view stage) (Chi and Riedl, 1998), design space (i.e., scientific visualization, GIS-based visualization, multi-dimensional plots, multi-dimensional tables, information landscapes and spaces, node and link diagrams, trees, and text transforms) (Card et al., 1999), visualization operators and techniques (Chi, 2000), data modification (i.e., original and distorted) and positioning

(i.e., overlapping, space-filling, and separation) (Ward, 2002), and data complexity (i.e., low and high) (Lenger and Eppler, 2007).

Nevertheless, such traditional categorizations still interrelate and overlap substantially, because they mostly focus on contextual data characteristics (Tory and Moller, 2004). This narrow focus in studying visualizations has limited researchers from investigating the cognitive match visualizations provide to data analysts relying on them when solving different types of data analysis problems (Tory and Moller, 2004). Note that task type has also been considered, though rarely and to a limited extent (Pfitzner, Hobbs, and Powers, 2003; Schneiderman, 1996), as a part of some these visualization taxonomies. Furthermore, researchers investigating cognitive fit (e.g., Goswami et al., 2008) have demonstrated that the nature of a problem-solving task determines the extent to which a given type of visualization can support that task. Therefore, investigating the interplay between task characteristics and the nature of data being visualized provides a unique opportunity to understand how visualizations can better support big data analytics, based on the match between visualization characteristics and data analysis task types. This match is especially critical when data analysts are faced with ever-increasing amounts and types of information, as information overload can worsen the consequences of mis-matched visualizations (Eppler and Mengis, 2004).

Regardless of the data type or characteristics, the purpose of computer-supported visualization is to amplify cognition by visually and interactively representing otherwise plain/nonvisual data (Card et al., 1999). Specifically, it supports data analysis by extending analysts' working memory and by providing visual patterns,

which are easier to interpret than nonvisual information such as text or numbers (Ware, 2004). Such support can be especially helpful and even necessary in the context of big data analysis, given that data analysts are faced with ever-increasing amounts and types of information.

The inherent size and complexity of a typical "big" dataset make it more challenging to analyze and interpret using simple traditional methods and tools (Eaton et al., 2012). Unlike "regular" data, which can be analyzed by hand with pen and paper at the expense of time and efficiency, "big" data necessitates the use of computers, as it is usually too large, too complex, and too unstructured to display in its entirety. Thus, analysts increasingly rely on technology to help them visualize and analyze the data in novel ways. Yet, even though the findings of past research have established the benefits of appropriate visualization on analysis and decision performance (e.g., Dennis and Carte, 1998; McNab et al., 2011; Vessey, 1991), the variety of visualization methods plus the challenges and benefits they present in a big data context make the choice of visualization more difficult than in other contexts.

Big data analysts are expected to analyze large amounts and various types of data (e.g., sales figures, inventory stock levels, customer traffic, social media posts, online reviews and complaints, etc.) concurrently to discover unintuitive trends or to solve relatively complex problems (McAfee and Brynjolfsson, 2012). With numerous types of visualization available, it becomes more important yet more difficult to choose the "best" one for analyzing and gaining insights from the represented information. Even seemingly similar analysis tasks might require different visualization approaches, depending on the nature of the information that needs to be emphasized. For instance,

while a simple bar chart might be easier to use for identifying the inventory stock level of a product at a certain location, a geographical heat map might be more appropriate for examining the distribution of its inventory stock levels across the nation. The difference between the two lies in how information is represented and emphasized in the visualization (e.g., discretely in a simple bar chart vs. aggregately in a heat map), and which representation provides a more suitable emphasis (i.e., better cognitive fit) for the given data analysis task. Hence, there is no single "best" type of visualization that can be used for all different types of analyses, as the nature and requirements of a specific analysis task, and thus cognitive fit, determine which type of visualization will be most appropriate. The cognitive fit theory is now discussed in detail.

Cognitive Fit Theory

The extended cognitive fit model (Sinha and Vessey, 1992) suggests that the most appropriate and effective visualization technique for a specific data analysis task is the one that represents and emphasizes the information type that is required by the given task. Past MIS research (e.g., Dennis and Carte, 1998; Goswami et al., 2008; Vessey, 1991) has utilized cognitive fit theory in explaining how certain types of visualizations (e.g., tables vs. graphics) are indeed more suitable for certain types of tasks (e.g., symbolic vs. spatial) (Vessey, 1991), plus that the congruence between information requirements and the mode of information representation has important implications for task performance.

Figure 1 illustrates the general problem-solving model that the cognitive fit theory is based on. At the heart of the cognitive fit theory lies the concept of "mental

representation". A mental representation refers to the way that a data analysis problem is represented in an analysts mind, and it is determined by the specific data analysis task and its information requirements. The cognitive fit theory suggests that to be efficient, visualizations need to represent information in the most compatible way with the mental representation an analyst requires to solve a data analysis problem. In other words, data analysis problems can be solved in the most efficient manner when there is a match between the information emphasized by the visualization (i.e., problem representation) and the type of the information required by the data analysis task (i.e., problem-solving task). On the other hand, when there is a mismatch between the problem representation and problem solving task requirements, analysts have to transform either their mental representation or the problem representation to derive solutions to the data analysis problem, which deteriorates their data analysis performance (Vessey, 1991), resulting in slower and less accurate decisions. This implies, due to cognitive fit, that the choice and format of visualization can be quite consequential for data analysis task performance. Thus, to understand how visualizations can facilitate or hinder data analysis tasks, we need to study the cognitive fit between the mental and visualized representations of information.

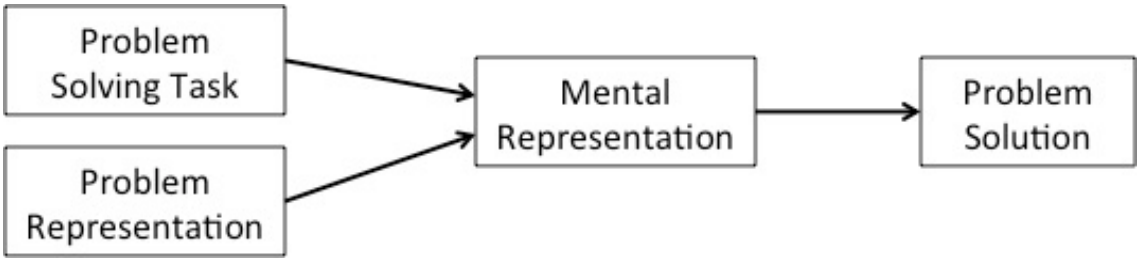


Figure 1. The Cognitive Fit Model (Adapted from Vessey, 1991)

Data analysis efficiency and performance both depend on the cognitive fit between the mental and visualized representations (Vessey, 1991). This occurs because a consistent visual representation will allow “"immediate" information extraction at a single glance with no need to move the eyes or attention.” (Green, 1998; 10) An incompatible visual representation, on the other hand, will provide no cognitive fit and require analysts to spend greater cognitive effort transforming the visualized representation (Vessey, 1991). In this case, the analysts will struggle plus spend more time and effort looking at different parts of the visualization(s) to extract the relevant information while ignoring or discarding the rest (Umanath and Vessey, 1994). Thus, I argue that cognitive fit can be observed through the efficiency with which the analysts scan the visualization and extract information.

The efficiency and ease of information extraction can be captured by using neurophysical tools (e.g., fMRI, EKG, EEG, or eye trackers), which can in turn inform our understanding of cognitive fit better than possible with self-reported measures (Dimoka, Banker, Benbasat, Davis, Dennis, Gefen, Gupta, Ischebeck, Kenning, Pavlou, Müller-Putz, Riedl, vom Brocke, and Weber, 2012). The data obtained through such tools are “generally not susceptible to subjectivity bias, social desirability bias, and demand effects” and “are particularly valuable for measuring IS constructs that people are either unable, uncomfortable, or unwilling to truth- fully self-report … [such as] complex cognitive processes” (Dimoka et al., 2012, p. 680). Eye trackers are particularly relevant for studying cognitive fit because they capture the efficiency with which participants are able to extract information from a given visualization. By analyzing participants’ gaze fixations and eye movements, it is possible to assess

whether if they are able to pinpoint the relevant information and discard the irrelevant rest of information with minimal gaze movements and effort, indicating cognitive fit, or if they struggle to identify the relevant information and separate it from the irrelevant information, hence spending more time viewing the visualization and having to fixate their gaze on different distinct spots due to the lack of cognitive fit. Therefore, I argue that eye trackers can be used to objectively assess cognitive fit:

H1: Cognitive fit will be manifested in eye movement patterns such that when there is cognitive fit between the task and visualization, analysts will have less frequent eye movements and fewer but longer gaze fixations.

To discuss the cognitive fit between different types of visualizations and business data analysis tasks, it is important to understand the match between the information emphasized by visualizations and the information required by data analysis tasks. Therefore, the visualization and data analysis task taxonomies utilized in this dissertation are now discussed in detail, with a focus on the information emphasis provided by specific visualizations and the information format required by different business data analysis tasks.

Cognitive Fit Between Data Analysis Tasks and Visualizations

The different visualization taxonomies identified in the literature were introduced in the previous section. In this dissertation, I utilize a high-level taxonomy developed by Tory and Moller (2004) that categorizes visualizations as discrete vs. continuous. This taxonomy was chosen for two main reasons: First of all, this high-level categorization of visualization is based on data model representation, rather than data

characteristics. In other words, its focus is on the format of the information represented by visualizations (i.e., the way in which data is structured and presented in context), rather than the typical focus on the characteristics of the raw, context-free data they are based on. Implicit is the assumption that all raw data can be visualized either discretely or continuously, depending on the conceptual data model that is to be represented (Tory and Moller, 2004). For example, a list of stores and their inventory stock levels can be visualized discretely as a simple bar chart in which the inventory stock level of each store is represented individually as a bar, or it can be represented continuously as an inventory stock level heat map in which the physical regions (e.g., stores, cities, states, or countries) are gradually colored according to their inventory stock levels. Therefore, this high-level approach to categorizing visualizations, as opposed to the traditional scientific vs. information visualization distinction or other taxonomies that categorize visualizations based on raw data types or characteristics, subsumes the other taxonomies because it does not rely on specific data characteristics, thus rendering it applicable across different datasets and contexts.

Second, this high-level taxonomy of visualizations allows the examination of the cognitive fit between the visualized data model and the conceptual data model in the data analysts' minds. This occurs because it enables us to study the congruence between the problem representation (i.e., the way in which information is emphasized by visualization) and the analysts' mental representation dictated by the data analysis task. This congruence would not be consistently observable by relying on a categorization of visualizations based on raw data characteristics, because the same raw data with fixed characteristics can be visualized in multiple ways that emphasize different aspects of

information, as argued before. In other words, categorizing visualizations based on raw data characteristics, without considering the data model represented by the visualizations, prevents us from studying the cognitive fit they provide for the conceptual data models required for data analysis tasks.

As with visualization, there has been considerable taxonomical research about data analysis tasks. Table 2 provides a summary list of the major data analysis task taxonomies in the literature. Most of this research has divided data analysis tasks into context-specific analysis activities. For example, Jarvenpaa and Dickson (1988) investigated and compared the different activities involved in managerial decision-making (i.e., summarizing data, showing trends, comparing points and patterns, showing deviations, point/value reading), while Pirolli and Card (2005) examined the activities involved in expert sensemaking and intelligence analysis (i.e., search and filter, read and extract, schematize, build case, tell story, re-evaluate, search for support, search for evidence, search for relations, and search for information). Other researchers have also identified the tasks involved in presenting intelligent graphics (i.e., value lookup, within comparison, between comparison, distribution, correlation, and indexation) (Roth and Mattis, 1990), screen and report design (i.e., intraset pattern recall and point value recall) (Umanath, Scamell, and Das, 1990), and simple decision-making using graphical and tabular representations (i.e., spatial and symbolic) (Vessey, 1991).

Table 2. Major Data Analysis Taxonomies

Categories of Data Analysis Tasks	Reference
3; (specific amount recall, static comparison, dynamic comparison)	Washburne, 1927
3; (descriptive, normative, prescriptive)	Bell, Raiffa, and Tversky, 1988
5; (summarizing data, showing trends, comparing points and patterns, showing deviations, point/value reading)	Jarvenpaa and Dickson, 1988
6; (value lookup, within comparison, between comparison, distribution, correlation, indexation)	Roth and Mattis, 1990
2; (intraset pattern recall, point value recall)	Umanath, Scamell, and Das, 1990
11; (identify, locate, distinguish, categorize, cluster, distribution, rank, compare, within and between relations, associate, correlate)	Wehrend and Lewis, 1990
2; (spatial, symbolic)	Vessey, 1991
7; (overview, zoom, filter, details-on-demand, relate, history, extracts)	Schneiderman, 1996; Pfitzner, Hobbs, and Powers, 2003
3; (Information Retrieval, Information Comparison, Information Integration)	Zhang, 1996
15; (associate, background, categorize, cluster, compare, correlate, distinguish, emphasize, generalize, identify, locate, rank, reveal, switch, encode)	Zhou and Feiner, 1998
10; (retrieve value, filter, compute derived value, find extremum, sort, determine range, characterize distribution, find anomalies, cluster, correlate)	Amar, Eagan,and Stasko, 2005
10; (search and filter, read and extract, schematize, build case, tell story, re-evaluate, search for support, search for evidence, search for relations, search for information)	Pirolli and Card, 2005

The repetitive list of tasks identified across these taxonomies can broadly be summarized in three categories; extracting individual or aggregate data values, identifying the patterns and relationships in data, and comparing and/or integrating different dimensions of data. These data analysis activities can benefit from different types of visualizations providing different emphases on information, as they

fundamentally differ from one another in terms of the cognitive approaches and the mental representations they require. For instance, extracting individual data values requires an analyst to identify and isolate specific data points among others, while identifying patterns and relationships requires the analyst to view the data points as an aggregate whole.

To study the impacts of cognitive fit between business data analysis tasks and computer-generated visualizations, this dissertation utilizes a high-level data analysis task taxonomy (Zhang, 1996) based on an analysis of relational information displays (i.e., representations of information that display multiple dimensions of data in relation to one another). According to this taxonomy, there are three major types of data analysis tasks, overlapping with the three broad categories of data analysis activities summarized in the previous paragraph: Information Retrieval, Information Comparison, and Information Integration tasks (Zhang, 1996). Table 3 provides two examples for each one of tasks based on this taxonomy. These tasks are described in detail in the following paragraphs.

Table 3. Examples for High-Level Data Analysis Tasks Taxonomy

Context / Data Provided	**Demographics by city**	**Daily sales and inventory stock levels by store**
Task Type	**Analysis Example 1**	**Analysis Example 2**
Information Retrieval	What is the population for City A?	What is the total sales amount for Store #33 during Black Friday?
Information Comparison	Which city has a larger population, A or B?	How does the annual sales performance of Store #34 compare with the Black Friday sales performance of Store #33?
Information Integration	Which city or state has the smallest employment-to-population ratio?	Which store or region has the best net sales to inventory ratio?

As with the visualization taxonomy, this high-level task taxonomy was chosen for two main reasons: First, this taxonomy can be applied across different contexts unlike others (e.g., computing correlations is not necessarily applicable for investigating the geographical distribution of inventory stock levels). More specifically, this high-level taxonomy subsumes others because it encapsulates what has been consistently identified as the basic data analysis activities in several different contexts and taxonomies (see Table 3). For instance, Information Retrieval maps onto specific amount recall, value lookup, and reading and extracting; Information Comparison maps onto static and dynamic comparison, comparing points and patterns, and ranking; and Information Integration maps onto encoding, calculating correlations, and computing derived values.

Second, as with the high-level visualization taxonomy utilized in this dissertation, this approach to categorizing business data analysis task types enables us to study the cognitive fit between the data analysis task and visualization types by classifying data analysis tasks according to the different cognitive processes and behaviors, and hence the mental representations, they require. Each one of the data analysis task types identified in this taxonomy (i.e., Information Retrieval, Information Comparison, and Information Integration tasks), plus their information requirements and how they can be better supported by certain types of visualizations, are now discussed.

Firstly, Information Retrieval tasks typically require analysts to search for and extract particular information along a specified dimension. This means that analysts have to identify and isolate a specific data point, usually in the presence of many others

(Zhang, 1996). Doing so requires the analysts to be able to distinguish between the data points, and extract relevant information while ignoring the rest. Thus, this type of a task will best be supported by a visual representation that either highlights specific data points or presents them in an unambiguous and distinctive manner, so that the analysts can identify the data points relatively easily and tell them from one another. For these types of tasks, discrete visualization of information, in which data points are explicitly and singularly represented, are expected to be more appropriate as opposed to a continuous visualization of information, in which data points are more difficult to isolate and identify because they are aggregately visualized as lines, areas, patterns, or shades of colors.

From a cognitive fit perspective (Sinha and Vessey, 1992), the visual representation provided by discrete visualizations is expected to be consistent with the requirements of Information Retrieval tasks. Using discrete visualizations, relevant information that needs to be retrieved can be extracted from the visualization in the most effective manner possible, while the rest of the information can be easily ignored. On the other hand, continuous visualizations will provide an incompatible aggregate representation, requiring the analyst to spend greater cognitive effort locating and isolating the target data point from the rest of the aggregated information. The difference between these two conditions is that the first one enables the analyst to almost immediately or automatically extract the relevant information with minimal movement of the eyes and attention, consistent with an efficient visualization (Bertin, 1983; Green, 1998). In the second condition, the analyst has to scan the aggregate visualization to locate the exact point to which the required information corresponds.

This movement of the eyes and attention while scanning is not automatic and requires additional mental effort, which can disrupt information extraction from the visualization (Woods, 1991). Therefore, I propose that:

H2: For Information Retrieval tasks, discrete visualizations will provide a better cognitive fit than continuous visualizations, resulting in (a) more accurate and (b) faster decisions.

Information comparison tasks, on the other hand, involve contrasting two or more pieces of information along the same dimension (i.e., within) or different dimensions (i.e., between) with the same scale. This type of a task requires the analyst to compare two or more data points, and determine the magnitude of their difference along the specified dimension. Here, the focus is on assessing the difference between the data points, rather than identifying their individual values. Therefore, an Information Comparison task can be accomplished only by determining the difference between multiple data points, without having to determine their exact values. Visualization can support this task and make it more efficient to the extent that it enables analysts to determine easily how close, or far, the data points are on the dimension of interest. Thus, analysts performing Information Comparison tasks are expected to benefit more from a continuous visualization, in which the data points could be represented aggregately on the same scale, making their differences easier to immediately notice, as opposed to a discrete visualization that represents data points in isolation.

It is also possible for analysts first to identify individually and extract the specified data points from discrete visualization(s) as in an Information Retrieval task and then compare them, but this approach will take more time and could be less

accurate, since there are more steps in which errors can be made (i.e., locate and extract information from each data point, convert the information onto a common scale if necessary, and finally compare the information). For an Information Comparison task, discrete visualizations provide an incompatible visual representation because they require the analyst to adapt the mental representation required to assess a single piece of information (i.e., the difference between data points) to that for extracting multiple pieces of information (i.e., individual values of the data points) and comparing them. This action requires greater cognitive effort (Umanath and Vessey, 1994). Stated from a cognitive fit perspective, the representation provided by continuous visualizations is expected to be more consistent with the mental representation required by Information Comparison tasks, compared to the representation provided by discrete visualizations. Hence, I propose that:

H3: For Information Comparison tasks, continuous visualizations will provide a better cognitive fit than discrete visualizations, resulting in (a) more accurate and (b) faster decisions.

The third and final type of task, namely Information Integration tasks, require analysts to gather and integrate information from two or more dimensions, and thus might necessitate the use of distinct visualizations to represent each one of the dimensions. However, multiple dimensions of information (e.g., sales figures, inventory stock levels, and geographical coordinates) also can be displayed on a singular visualization by overlaying one layer of dimension on another (such as by displaying sales and/or inventory stock levels on a map) or by utilizing multiple axes. Many companies still take the former approach by continuing to rely on legacy data

warehouses and investing millions of dollars on developing "classic" analytics dashboards (Davenport and Dyche, 2013). These dashboards typically provide multiple visualizations of predefined "key performance measures" (KPM) or "key performance indicators" (KPI) that are used to summarize and assess businesses' performance.

Contrary to this widespread approach taken by practitioners, previous research based on Image Theory (Bertin, 1983) suggests that singular images inherently are more efficient in conveying information than figurations (i.e., constructions of multiple graphics) (Crossland, Wynne, and Perkins, 1995). Image Theory argues that individuals extract information from visualizations based on their perception of the correspondences between different data dimensions represented by the variables (Green, 1998). This happens in three stages; (1) in the "external identification" stage, the analyst determines what data is being visualized, (2) in the "internal identification" stage, the analyst determines which data dimension is mapped onto each visual variable (e.g., the horizontal and vertical axes), and (3) in the last stage, the analyst perceives the correspondences (e.g., correlation) between the data dimensions being visualized. Singular images are more efficient, because they permit almost immediate extraction of information, with minimal time spent in each of these stages.

From this point of view, visualizations are deemed to be efficient to the extent that they allow immediate extraction of specific information without having to scan through the entire information presented visually. Accordingly, a single image combining all the specified dimensions is expected to provide a more efficient and compatible representation for an Information Integration task, than do several separate visualizations displayed at the same time, such as in dashboards. For instance, since

many practical business analysis tasks include a spatial information component (e.g., customer demographics and addresses, retail or warehouse sites, inventory and shipment locations) (Crossland et al., 1995; Card and Mackinlay, 1997), they are expected to benefit from a single visualization on the geographical coordinate system, which inherently integrates spatial information with any other information it represents.

Compared to singular visualizations, dashboards providing several visualizations at the same time are expected to be inherently less efficient for Information Integration tasks. This happens simply because in this case analysts have to scan and gather individual data from each one of the visualizations and then mentally integrate the information as required by the task. Doing so requires greater time to be spent in all three information extraction stages for each one of the visualizations, as they will not necessarily be consistent in terms of what data is represented and how. In this case, the analysts will have to spend greater cognitive effort for the overall analysis task, as they will have to transform their mental representations, possibly several times, to extract information from each visual representation. Thus, even though singular overlaid visualizations might appear to be more complex, the cognitive fit theory suggests that a single visualization combining all relevant dimensions will be more efficient than multiple simple visualizations, in the context of Information Integration tasks:

H4: For Information Integration tasks, singular visualizations will provide a better cognitive fit than multiple visualizations, resulting in (a) more accurate and (b) faster decisions.

Because the focus of this dissertation is on how cognitive fit can facilitate or hinder big data analytics, the reasons why big data is especially challenging to analyze

need to be investigated. Therefore, the following section introduces the defining characteristics of big data, and discusses how these characteristics are expected to amplify the hypothesized task performance consequences of the cognitive fit between visualizations and data analysis task types.

Characteristics of Big Data

As argued before, "big" data is more challenging to analyze than "regular" data due to its inherent characteristics. Thus, the cognitive fit between the data models represented by visualizations and those required by the data analysis tasks becomes more important in the context of big data and the challenges it presents. Despite the lack of a commonly accepted definition, the term "big data" is mostly used to refer to data that cannot be easily analyzed by traditional tools or processes (Eaton et al., 2012). There are three definitional characteristics of big data that separate it from "regular" data and make it inherently more challenging to analyze; namely, its (1) Volume, (2) Variety, and (3) Velocity (Chen et al., 2012; Eaton et al., 2012; McAfee and Brynjolfsson, 2012). Each one of these characteristics are now discussed.

Volume refers to the amount of information in a dataset, and as the name suggests, "big" data usually refers to considerably large amounts of data in the order of magnitude of petabytes (i.e., quadrillion [10^{15}] bytes) or even exabytes (i.e., quintillion [10^{18}] bytes). This sheer amount of information can easily cause "information overload" (Eppler and Mengis, 2004), which is one of the major problems that analysts face while dealing with a typical big data set (Chen et al., 2012; Manyika et al., 2011). This phenomenon is said to occur when the information load (i.e., the information that must

be processed to accomplish an analysis task) exceeds an analyst's processing capacity (Hiltz and Turoff, 1985).

Variety refers to the rich diversity of data types (e.g., unstructured text, pictures, videos, GPS location data, various sensor readings, etc.) entailed in big datasets. Due to the continuous increase in use of mobile devices and social media networks, an ever-growing amount and variety of user generated content (e.g., product reviews, 'like's, comments, check-ins, photo and video uploads, etc.) is being captured by organizations and added to their "big" datasets (VijayaBaskaran, 2013). Not only is it more difficult for analysts to realize the patterns and relationships among such variety of data (Eaton et al., 2012), but the diverse and fragmented nature of the information contained in these datasets can also contribute to information overload (Tzabbar, 2009).

Velocity refers to the speed with which data is created, and it is becoming more common for big datasets to be updated in near real-time. Organizations collect more and more real-time data, such as transaction details, locations of customers, or the number of cars in parking lots, in hopes for gaining rapid insights and competitive advantage (McAfee and Brynjolfsson, 2012). However, most of big data analytics is still performed on static datasets (VijayaBaskaran, 2013) due to technological and practical limitations regarding collecting, storing, aggregating, and displaying such amounts of information in real-time. Thus, the consequences of velocity for visualizing big data are excluded from the scope of this study, and suggested as a future research topic.

Researchers have observed that information overload is consistently detrimental to analysis and decision performance, which is usually evident in increased processing times and/or decreased decision quality (e.g., Iselin, 1988; Speier, Valacich, and

Vessey, 1999; Chan, 2001; Gao, Zhang, Wang, and Ba, 2012). A review of past literature (Eppler and Mengis, 2004) suggests that the amount (i.e., volume) and diversity (i.e., variety) of the information that an analyst has to process are both major factors that contribute to information overload. In other words, the very characteristics of a typical big data can easily exacerbate information overload, making big data inherently more difficult to analyze.

Aggregation of information, such as by visualization, long has been recommended as a way to prevent or mitigate information overload (Ackoff, 1967; Meharia, 2012). Even though visualization might help analysts interpret and provide insights into regular datasets, the utter amount and variety of visually represented information can still be overwhelming for analysts working on big datasets. Visualizing such high volumes and large varieties of information can result in over-plotting (i.e., the over-accumulation of data points to the extent that they obscure the underlying data values and relationships) (Grolemund and Wickham, 2015) and render the visualizations uninterpretable (Palaniappan, 2014). Therefore, to investigate how the defining characteristics of big data (i.e., volume and variety) influence the task performance consequences of cognitive fit, the scope of this dissertation is limited to interpretable visualizations representing manageable volumes and varieties of information.

In this domain, visualizations providing a compatible representation with the task requirements will provide a stronger advantage, whereas the large amount and variety of information will worsen the consequences of incompatible visual representations by making it further difficult to retrieve and compare information. For

instance, larger volumes of data typically require an increased visualization range and/or a decreased level of details (Pajarola, 1998), simply because there is more information to represent. Providing a larger range of visualization or decreasing the level of detail can make it harder to notice the differences between two data points, especially if they are relatively close to each other, because their difference will be smaller in comparison to the irrelevant rest of the information, which will occupy a larger space.

Similarly, larger variety of information can make it more difficult for analysts to distinguish between the overlaid dimensions of information (e.g., multiple lines in a graph with different scales) as well as between individual data points. This occurs again because there will be more irrelevant information contained in the visualization that analysts will have to scan, separate from the relevant information, and ignore, resulting in greater time and cognitive effort spent for the data analysis task. In short, I predict that cognitive fit will play an even more important role as the volume and variety of data make visualizations complex enough to diminish the elemental gains they provide:

H5: For Information Retrieval tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has larger volume.

H6: For Information Retrieval tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has larger variety.

H7: For Information Comparison tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has larger volume.

H8: For Information Comparison tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has larger variety.

Big data analytics typically are summarized as collections of visualized key metrics and relationships through dashboards (Chen et al., 2012; Eaton et al., 2012; Davenport and Dyche, 2013). However, as argued before, such collections of multiple visualizations are not as effective as singular visualizations that can overlay the relevant dimensions for a particular Information Integration task. To make matters worse, big datasets with larger variety of data have a greater number of dimensions, which have to be represented using a greater number (and possibly variety) of distinct visualizations. I argue that the efficiency gain provided by singular visualizations will be more pronounced when there are more dimensions to represent, as analysts will have to struggle with extracting and integrating information from an even larger number of distinct visualizations without a singular comprehensive visualization available. Thus, I propose that:

H9: For Information Integration tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has larger variety.

Figure 2 below summarizes these hypotheses and depicts the conceptual research model. The following chapter describes the laboratory experiment conducted to test this model and the hypotheses.

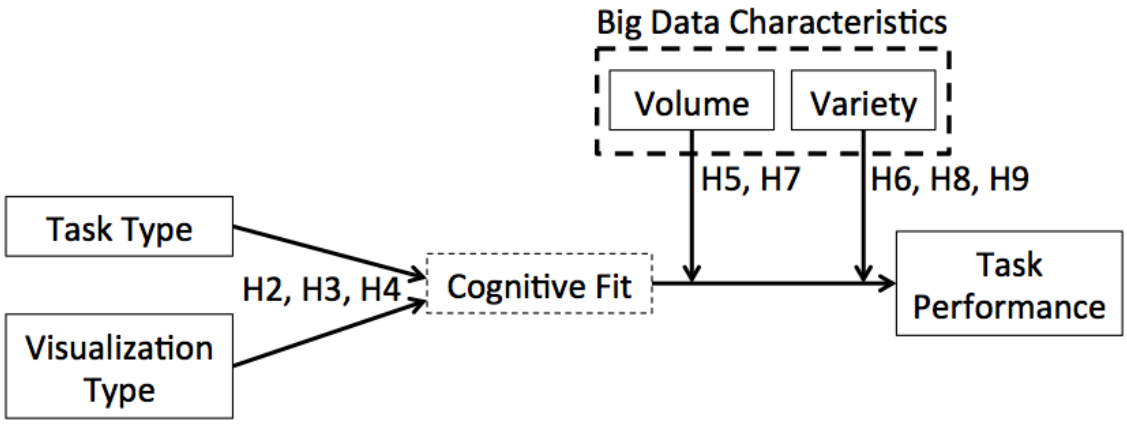


Figure 2. Conceptual Research Model

Chapter 3: Methodology

Experiment

As this study investigates the consequences of cognitive fit in visualizing big data, it was important to observe how task performance would be affected for different types of actual big data analytics tasks when different types of visualizations were provided to the analysts. Hence, a laboratory experiment was conducted using a large financial dataset obtained from an online credit marketplace, the Lending Club (www.lendingclub.com). This dataset contained financial information (i.e., number of loan applications, loan amount, loan interest rate, and loan applicants' annual income) regarding over a million loan applications across the United States of America.

To test the research model and hypotheses, participants were asked to solve four or eight business data analysis problems (see Task Type) based on various types of visualized financial data about the loans issued by the Lending Club. A simple executable program was coded in C# to provide the participants with the instructions, questions, and visualizations used in the experiment. These visualizations (see Visualization Type and Appendix A), based on the financial information about loan applications, were developed with data visualization software Tableau. The developed experimental materials (i.e., task instructions, questions, and visualizations) were revised and finalized after a pilot study was conducted (see Pilot Study).

Table 4 summarizes the experimental design and treatments for this study. To test the cognitive fit between different data analysis tasks and different types of visualizations, the experiment followed a combined 2 (Visualization Type: Discrete vs. Continuous) x 2 (Volume: High vs. Low) x 2 (Variety: High vs. Low) full-factorial

between-subjects design for Information Retrieval and Information Comparison tasks (Task Type was manipulated within-subjects), and a 2 (Visualization Type: Singular vs. Multiple) x 2 (Variety: High vs. Low) between-subjects design for Information Integration tasks. Put differently, two distinct experiments were conducted where the participants were randomly assigned to one of the first eight or last four treatments, and worked either on both Information Retrieval and Information Comparison tasks, or only on Information Integration tasks.

Table 4. Summary of Experimental Treatments

Task Type	Treatment	Visualization Type	Volume	Variety
Information Retrieval & Information Comparison	1	Discrete	Low	Low
	2	Continuous	Low	Low
	3	Discrete	High	Low
	4	Continuous	High	Low
	5	Discrete	Low	High
	6	Continuous	Low	High
	7	Discrete	High	High
	8	Continuous	High	High
Information Integration	9	Singular	N/A	Low
	10	Multiple	N/A	Low
	11	Singular	N/A	High
	12	Multiple	N/A	High

Participants

The participants were recruited from four different mid-level undergraduate courses (MIS2113 – Computer Based Information Systems; MIS3223 – Financial Data Modeling; MIS3353 – Databases/Accounting Information Systems; MIS3373 – Systems Analysis/Design Theory) in the Price College of Business. Extra course credit worth approximately one percent of their final course grade was offered to the students in exchange for their participation in the experiment. A total of 145 students from nine

different sections participated in the experiment, of whom 48.97% were female, with an average age of 21.19 (s.d. = 2.69) and an average of 2.89 (s.d. = 1.19) years of education after high school.

Stimulus Materials

There were twelve different experimental treatments, as summarized in Table 4. The visualizations (see Visualization Type) provided for each one of these twelve treatments are included in Appendix A. Each one of the manipulations in these treatments (i.e., Task Type, Visualization Type, Volume, and Variety) is now discussed in detail.

Independent Variables

Task Type

As the focus of this study is on the cognitive fit that different types of visualizations provide for different types of business data analysis tasks, it was important to observe the participants solve each one of the task types identified in this study, while being provided with different visualizations. Accordingly, the task type in this experiment was manipulated by asking the participants to solve different types of data analysis problems with different information requirements, consistent with the definitions of Information Retrieval, Information Comparison, and Information Integration tasks.

For Information Retrieval tasks, participants were asked to retrieve a single value corresponding to a specific data point (e.g., the number of loan applications in a

certain state). Consistent with the definition for Information Retrieval tasks, these tasks required the participants to extract only one dimension (i.e., type) of information from the visualizations. For Information Comparison tasks, participants were asked to compare or rank the values of multiple data points (e.g., comparing the number of loan applications in two different states or ranking the states by the number of loan applications). Consistent with the definition for Information Comparison tasks, these tasks required the participants to compare two or more data values across only one dimension (i.e., type) of information from the visualizations. For Information Integration tasks, participants were asked to estimate a data value that was not directly represented in the visualization (e.g., loan amount to applicant income ratio), based on the provided pieces of information (e.g., loan amount and applicant income, displayed separately). Consistent with the definition for Information Integration tasks, these tasks required the participants to extract multiple (i.e., two) dimensions (i.e., types) of information from the visualizations, and calculate a new dimension of information (i.e., a ratio of the two dimensions). Table 5 provides a complete list of the data analysis problems (12 total; 3 for each data analysis task type) used in the experiment.

Participants in the first eight treatment conditions (see Table 4) were asked to solve a total of eight data analysis problems (i.e., four each for Information Retrieval [IR1, IR2, IR3, and IR4] and Information Comparison [IC1, IC2, IC3, and IC4] tasks), and participants in the remaining four treatment conditions were asked to solve a total of four Information Integration tasks (II1, II2, II3, and II4). Participants were randomly assigned to treatments, and the order of the data analysis problems was randomized for each participant.

Table 5. Data Analysis Problems Used in the Experiment

Information Retrieval Tasks	
Task	*Data Analysis Problem*
IR1	How many loans were issued in Florida?
IR2	How many loans were issued in Colorado?
IR3	How many states have more than 45 / 2,000 loans issued?
IR4	How many states have less than 45 / 2,000 loans issued?
Information Comparison Tasks	
Task	*Data Analysis Problem*
IC1	In which state was the most number of loans issued?
IC2	In which state was the least number of loans issued?
IC3	Which are the top three states with the most number of loans issued?
IC4	How many more loans were issued in Florida than Colorado?
Information Integration Tasks	
Task	*Data Analysis Problem*
II1	Which state has the highest loan amount to applicant annual income ratio on average?
II2	Which state has the lowest loan amount to applicant annual income ratio on average?
II3	Among the three states with the lowest average applicant annual income, which state has the lowest loan amount issued?
II4	Among the three states with the highest average applicant annual income, which state has the highest loan amount issued?

Visualization Type

As the focus of this study is on the cognitive fit that different types of visualizations provide for different types of business data analysis tasks, it was important to observe the participants solve the data analysis problems while being provided with each one of the different types of visualizations identified in this study. However, different aspects of the visualization type (i.e., Discrete vs. Continuous or Singular vs. Multiple) were hypothesized to affect the cognitive fit provided for Information Retrieval (H2) and Information Comparison (H3) tasks, and for Information Integration (H4) tasks. Accordingly, the visualization type in the

experiment was manipulated separately for Information Retrieval and Comparison tasks (as discrete vs. continuous) and for Information Integration tasks (as singular vs. multiple). Participants in the first eight treatments, who worked on Information Retrieval and Information Comparison tasks, were provided with either discrete or continuous visualizations. Participants in the remaining four treatments, who worked on Information Integration tasks, were provided with either singular or multiple visualizations.

Discrete visualizations represented data in isolation (i.e., as individual data points for each loan application), while continuous visualizations represented them aggregately as a whole (e.g., total number of loan applications for each state represented through the shades of colors on a heat map, as shown in Figure 9 in Appendix A). Singular visualizations represented two to four different kinds of information (i.e., number of loan applications, annual income, interest rate, and loan amount) overlaid on a single graphic, while these information dimensions were represented individually by distinct graphics in the multiple visualization condition. The full set of stimulus materials for different types of visualizations is provided in Figures 8 through 19 in Appendix A.

The effectiveness of the visualization manipulation was assessed via seven-point Likert-type items during the pilot study (see Pilot Study). Due to the difference in the manipulation of visualizations, different manipulation check items were used for the discrete vs. continuous visualization manipulation and the singular vs. multiple visualization manipulation. Participants in the first eight treatments, where visualizations were manipulated as discrete vs. continuous, responded to five seven-

point Likert-type items ("Each loan application was displayed explicitly.", "Each loan application was displayed individually.", "Each loan application was represented by a discrete symbol.", "The number of loan applications was summarized by state.", "The number of loan applications was combined by state.") after being provided with each one of the eight visualizations in a random order. A statistically significant difference ($p<0.002$) was observed for all five items when the responses for different visualization types (i.e., discrete vs. continuous) were contrasted within-subjects (see Table 6). Hence, the discrete vs. continuous visualization manipulation was deemed effective.

Participants in the last four treatments, where visualizations were manipulated as singular vs. multiple, responded to four seven-point Likert-type items ("All data were represented on a single visualization (i.e., on a single map).", "There was only one visualization (i.e., a single map) that displayed all of the data.", "Each type of data was represented on a distinct visualization.", "There were two or more maps, each of which displayed a different type of data.") after being provided with each one of the four visualizations in a random order. A statistically significant difference ($p<0.044$) was observed for all four items when the responses for different visualization types (i.e., singular vs. multiple) were contrasted within-subjects (see Table 6). Hence, the singular vs. multiple visualization manipulation was deemed effective.

Volume

Volume was manipulated by providing the participants with different amounts of visually represented data for the Information Retrieval and Information Comparison tasks. Participants in the low volume condition were provided with 1,000 distinct data

points (i.e., loan applications), whereas participants in the high volume condition were provided with 300,000 distinct data points. Volume was not manipulated for Information Integration tasks (see Table 4), as the amount of information being visualized (i.e., volume) was not hypothesized to affect the cognitive fit that different types of visualizations (i.e., singular vs. multiple) provide for Information Integration tasks. The reason is that the difference in the format of information representation between singular and multiple visualizations is based on variety (i.e., the number of information dimensions being represented on a singular visualization or as distinct visualizations) alone, and not on the volume (i.e., amount) of the information being visualized.

The effectiveness of the volume manipulation was assessed via three seven-point Likert-type items ("There was a large number of loan applications displayed.", "There was a high volume of loan applications.", "It was difficult to estimate the total number of loan applications being shown.") during the pilot study (see Pilot Study). Participants in the first eight treatments, where volume was manipulated as low vs. high, responded to these three items after being provided with each one of the eight visualizations in a random order. A statistically significant difference ($p<0.037$) was observed for all three items when the responses for different levels of volume (i.e., low vs. high) were contrasted within-subjects (see Table 6). Hence, the low vs. high volume manipulation was deemed effective.

Variety

Variety was manipulated by providing the participants with different kinds/dimensions of visually represented information. Participants in the low variety condition were provided with one dimension of information (i.e., number of loan applications) for the first eight treatments (i.e., discrete vs. continuous visualization manipulation) and two dimensions of information (i.e., number of loan applications and annual income) for the remaining four treatments (i.e., singular vs. multiple visualization manipulation). The low variety condition for the singular vs. multiple visualization manipulation contained two dimensions of information instead of one, because at least two dimensions of information are required for them to be overlaid on a singular visualization. Participants in the high variety condition were provided with three (i.e., number of loan applications, interest rate, and loan amount) and four (i.e., number of loan applications, annual income, interest rate, and loan amount) dimensions of information respectively for the discrete vs. continuous and singular vs. multiple visualization manipulations.

The effectiveness of the variety manipulation was assessed via two seven-point Likert-type items ("There was only one kind [two kinds] of data being displayed.", "Only a single type [two types] of data was [were] displayed.") during the pilot study (see Pilot Study). Participants in all treatments, where variety was manipulated as low vs. high, responded to these two items after being provided with each one of the four or eight visualizations in their condition, in a random order. A statistically significant difference ($p<0.007$) was observed for both items when the responses for different

levels of variety (i.e., low vs. high) were contrasted within-subjects (see Table 6). Hence, the low vs. high variety manipulation was deemed effective.

Table 6. Manipulation Checks

Item	Mean (Std. Error)	Mean (Std. Error)	Within-Subject Effect	
VISUALIZATION (Treatments 1-8)	**Discrete**	**Continuous**	**F**	**Sig.**
"Each loan application was displayed explicitly."	4.800 (.317)	3.067 (.350)	14.200	.002
"Each loan application was displayed individually."	5.083 (.321)	2.817 (.320)	31.448	.000
"Each loan application was represented by a discrete symbol."	5.367 (.233)	3.267 (.298)	43.891	.000
"The number of loan applications was summarized by state."	2.883 (.338)	5.983 (.188)	43.777	.000
"The number of loan applications was combined by state."	2.733 (.339)	5.567 (.255)	31.462	.000
VISUALIZATION (Treatments 9-12)	**Singular**	**Multiple**	**F**	**Sig.**
"All data were represented on a single visualization (i.e., on a single map)."	6.500 (.164)	1.438 (.220)	185.939	.000
"There was only one visualization (i.e., a single map) that displayed all of the data."	6.438 (.175)	1.500 (.189)	189.121	.000
"Each type of data was represented on a distinct visualization."	4.750 (.401)	6.000 (.299)	6.034	.044
"There were two or more maps, each of which displayed a different type of data."	1.562 (.320)	6.188 (.353)	89.561	.000
VOLUME (Treatments 1-8)	**Low**	**High**	**F**	**Sig.**
"There was a large number of loan applications displayed."	4.900 (.268)	5.417 (.289)	5.309	.037
"There was a high volume of loan applications."	4.650 (.293)	5.217 (.263)	7.549	.016
"It was difficult to estimate the total number of loan applications being shown."	4.667 (.294)	5.233 (.332)	6.704	.021
VARIETY (Treatments 1-8)	**Low**	**High**	**F**	**Sig.**
"There was only one kind of data displayed."	5.550 (.411)	1.650 (.226)	62.969	.000
"Only a single type of data was displayed."	5.450 (.406)	1.700 (.261)	56.519	.000
VARIETY (Treatments 9-12)	**Low**	**High**	**F**	**Sig.**
"There were only two kinds of data displayed."	4.750 (.807)	1.375 (.157)	15.417	.006
"Only two types of data were displayed."	4.312 (.744)	1.438 (.175)	13.869	.007

Control Variables

Color Blindness

Since this laboratory experiment relied on colorful visualizations (see Appendix A), participants were asked if they were color-blind ("Are you colorblind?") before starting the experiment, in an attempt to rule out a potential confound on their data analysis performance. Participants were also instructed to indicate the type of colorblindness they had (e.g., anomalous trichromacy, dichromacy, or monochromacy). However, none of the participants reported having colorblindness.

Task Familiarity

Participants' familiarity with visual data analysis tasks was included in this experiment as a control variable to rule out alternative explanations regarding their analysis performance. Participants were asked to report their familiarity by responding to three survey items at the beginning of the experiment ("How familiar are you with extracting information from visual representations of data such as charts, graphs, infographic maps, etc.?", "How much experience do you have with analyzing visual representations of data such as charts, graphs, infographic maps, etc.?", "How frequently do you analyze visual representations of data such as charts, graphs, infographic maps, etc.?"). Their responses were combined (Cronbach's Alpha = 0.81) to form a mean score of task familiarity.

Motivation

Participants' motivation for visual data analysis was included in this experiment as a control variable to rule out alternative explanations regarding their analysis performance. Participants were asked to report their motivation for visual data analysis by responding to three survey items at the beginning of the experiment ("How important to you is the subject of visual data analysis?", "How relevant to you is the subject of visual data analysis?", "How pertinent to you is the subject of visual data analysis?") Their responses were combined (Cronbach's Alpha = 0.85) to form a mean score of motivation.

Visualization Ability

Individuals' visualization ability (i.e., their ability to interpret and analyze information from visualizations) can affect their visual analysis performance (Shen et al., 2012). Thus, participants' visualization ability was controlled for by using a previously validated measure adapted from Shen et al. (2012). Participants were provided with six image pairs, and asked to determine whether if the image on the right represented an accurate 3-D rotation of the image on the left for each image pair. The image pairs and the instructions for the visualization ability measure are provided in Appendix C. Each correct answer was coded as "1" and each incorrect answer as "0", and participants' answers were combined into a visualization ability score (out of six) to be used as a control variable. Participants' average visualization ability score was 4.94 (st. dev. = 1.33).

Dependent Variables

Recall that Hypothesis 1 argues that cognitive fit will manifest in participants' eye movement patterns, while Hypotheses 2 through 9 argue about the cognitive fit between different types of data analysis tasks and visualization methods and how the task performance consequences will be affected by big data characteristics. Therefore, two types of dependent variables (i.e., eye tracker data and task performance) were required to test the hypotheses. Each one of these two dependent variables is explained in detail below.

Eye Tracker Data

"Eye tracking tools can capture whether a user finds it difficult to identify information by observing how her or his eyes wander aimlessly on a computer screen" (Dimoka et al., 2012, p. 685). To assess cognitive fit in this experiment, a Tobii TX-300 eye tracker with a 300 Hz sampling rate was used to capture participants' information extraction efforts. Two types of data were collected via the eye tracker: *View Time* and *Fixation Count.*

View Time is the total time a participant spent looking at a given area on their screen for each data analysis task. To capture View Time, an area of interest (AOI) must first be defined on the screen used for the experiment. The AOI needs to cover the entire range of information provided to the participants to capture all of their information extraction efforts. It is important to capture the information extraction efforts for the entire visualization, as opposed to only the relevant parts of it for a given problem (e.g., only the top three states with the highest number of loan applications),

because participants' efforts in viewing and discarding the irrelevant information (e.g., the states with fewest or no loan application) is also a factor affecting (i.e., decreasing) the efficiency of their information extraction. Therefore, the specific AOI defined for this experiment was the whole visualization (i.e., the entire map or maps; See Appendix A for the visualizations used as stimulus material) provided to the participants to solve the data analysis problems. View Times provide the measure of how much time each participant spent gazing at the visualization while extracting the information required for each data analysis task. Greater time spent viewing the visualization indicates greater cognitive effort and, thus, less efficiency in extracting information (Parasuraman & Manzey, 2010).

Fixation Count refers to the number of times a participant fixated their gaze on a given area on the screen. Based on the AOI described for View Time, the number of times a participant focused on the entire area of a given visualization was measured for each data analysis task. Higher Fixation Counts are indicative of greater levels of cognitive effort and lower efficiency of information extraction, consistent with greater View Times.

Due to excessive movement during the experiment, the eye tracker data for 6 participants were rendered unusable and discarded. Therefore, the final sample size for the eye tracker analyses was 139 (n=91 for Information Retrieval and Information Comparison tasks, and n=48 for Information Integration tasks). For each one of the three task types (i.e., Information Retrieval, Information Comparison, and Information Integration), participants' View Times and Fixation Counts across the four data analysis problems (see Table 5 for a full list of the problems) were summed to calculate the total

View Time and Fixation Count. Due to their departures from normality based on the results of skewness and kurtosis analyses, a square-root transformation was applied to these six variables (i.e., View Time and Fixation Count for each one of the three task types). After the transformation, the skewness and kurtosis for all variables were found to be within acceptable limits, not exceeding the values of 3 and 10, respectively (Kline, 2010). Table 7 provides the descriptive and normality statistics for the eye tracker data (i.e., View Times and Fixation Counts) for each task type, both before and after the square-root transformation.

Table 7. Descriptive Statistics for Eye Tracker Data

	Minimum	Maximum	Mean	St. Dev.	Skewness		Kurtosis	
Variable	**Statistic**	**Statistic**	**Statistic**	**Statistic**	**Statistic**	**Std. Error**	**Statistic**	**Std. Error**
IRFC	52.00	1149.00	274.86	149.41	2.33	.25	11.53	.50
ICFC	78.00	757.00	253.05	105.16	1.70	.25	5.26	.50
IIFC	120.00	997.00	429.63	202.88	.87	.34	.62	.67
IRVT	15.34	383.23	95.67	50.99	2.13	.25	9.79	.50
ICVT	25.11	283.17	83.07	39.10	1.96	.25	6.96	.50
IIVT	29.39	306.80	122.56	58.73	.96	.34	1.37	.67
Sqrt(IRFC)	7.21	33.90	16.06	4.14	.75	.25	2.56	.50
Sqrt(ICFC)	8.83	27.51	15.61	3.09	.81	.25	1.87	.50
Sqrt(IIFC)	10.95	31.58	20.17	4.83	.26	.34	-.10	.67
Sqrt(IRVT)	3.92	19.58	9.48	2.42	.66	.25	2.20	.50
Sqrt(ICVT)	5.01	16.83	8.90	1.97	.90	.25	2.18	.50
Sqrt(IIVT)	5.42	17.52	10.76	2.63	.20	.34	.27	.67

IRFC = Information Retrieval Fixation Count
IRVT = Information Retrieval View Time
ICFC = Information Comparison Fixation Count
ICVT = Information Comparison View Time
IIFC = Information Integration Fixation Count
IIVT = Information Integration View Time

Task Performance

Two dimensions of task performance (i.e., solution time and accuracy) were measured to capture the tradeoff between participants' speed and accuracy in solving the data analysis tasks. Such an approach is consistent with prior research assessing and

comparing task performance for different forms of visualizations (e.g., Dennis and Carte, 1998). Solution time was measured as the number of seconds a participant took to answer a data analysis problem. Participants' solution time across the four data analysis problems for each task type was averaged to calculate a mean solution time for each task type.

To assess participants' solutions' objective accuracy, they were assigned a score out of 100 and they were given partial credit depending on how far off their solution was from the correct answer, consistent with previous cognitive fit studies (e.g., Dennis and Carte, 1998; Shaft and Vessey, 2006). Solution accuracy was assessed differently for each one of the data analysis problems (see Appendix B for the grading procedure), because different task types required different types of answers (e.g., a numerical answer vs. a list of three states). Participants' solution accuracy across four data analysis problems was then averaged for each task type.

Pilot Study

The stimulus materials used in the experiment were finalized after a pilot study was conducted with 10 graduate students from the Price College of Business. Based on the feedback obtained from these participants, the language used in some of the questions and instructions was revised to improve clarity. For instance the words "(i.e., on a single map)" were added to the end of one of the manipulation check items for visualization type: "All data were represented on a single visualization (i.e., on a single map)." Furthermore, borderlines were added to the images used for the visualization ability measure (See Visualization Ability and Figure 20 in Appendix C) because one

participant suggested that the image pairs were difficult to identify without clear borders separating them.

Another pilot study was conducted with 23 undergraduate students from the Price College of Business to perform the manipulation checks for the visualization type (discrete vs. continuous or singular vs. multiple), volume (low vs. high), and variety (low vs. high) manipulations. Participants were randomly assigned to the first eight (n=15) and last four (n=8) treatments, maintaining an assignment ratio of 2:1, because the participants in the first eight treatments responded to the manipulation check items for eight different visualizations while the participants in the last four treatments responded to these items for only four different visualizations. The manipulation checks are described in detail in the Independent Variables section, and their results are reported in Table 6. No changes to the experimental materials were deemed necessary based on the manipulation check pilot study.

Procedure

The participants were recruited via in-class announcements. Interested participants were instructed to make an appointment for the experiment. Upon showing up to their appointments, being greeted by the experimenter, and providing electronic consent to participate in the experiment, recruited participants were randomly assigned to one of the twelve experimental treatments (see Table 4). After they provided electronic consent to participate in the study, participants first answered a survey about the control variables (i.e., task familiarity, color blindness, motivation, and visualization ability; See Control Variables). Then, participants completed a training session to

ensure that they were familiar with the experimental procedures before starting the actual experiment.

The training session involved solving two data analysis problems that were very similar to the actual problems used in the experiment. To prevent potential learning effects (i.e., participants' performance improving as a result of repeated use), it is especially important to provide such training to participants before they work on the actual experimental tasks (e.g., McNab et al., 2011; Shaft & Vessey, 1995; Yetgin et al., 2015).

Depending on the experimental treatment they were assigned to (see Table 4), participants were asked to solve different types of data analysis problems (see Task Type), while being provided with different visualizations (see Visualization Type). After solving each data analysis problem in their experimental treatment, participants indicated their confidence in their answer. Once the experiment was completed, participants responded to a second survey about their demographic information before being released.

While the participants performed the experimental tasks, an eye tracker was used to capture where exactly they were looking on their screens. The eye tracker was calibrated for each participant prior to the experiment, by asking the participant to follow with their eyes a red circle that moved around on their screens. The experiment and data collection commenced after successful calibration of the eye tracker. Two types of data were collected via the eye tracker (see Eye Tracker Data): *View Times* (i.e., the time participants' spend looking at a specific area of interest on the screen) and *Fixation Counts* (i.e., the number of times participants fixate their gaze on an area of

interest). These data were used to assess cognitive fit, based on the patterns of how the participants moved and fixated their eyes and attention while preforming the tasks with different visualizations. Less frequent movement and fixation of the eyes were expected to indicate greater cognitive fit and efficiency in solving the data analysis tasks, as argued in the first hypothesis. This is consistent with the suggestion that eye tracker data can indicate the level of difficulty with which participants extract information from the visualizations on their computer screens (Dimoka et al., 2012).

Chapter 4: Results

Recall that Hypothesis 1 concerns the manifestation of cognitive fit through participants' eye movement patterns, while Hypotheses 2 through 9 concern the task performance implications of cognitive fit and big data characteristics. Therefore, two sets of analyses were performed to test the entire set of hypotheses, with View Time and Fixation Count as the dependent variables for testing Hypothesis 1, and Solution Accuracy and Solution Time as the dependent variables for testing Hypotheses 2 through 9. These tests are explained in detail in the following paragraphs. Table 8, Table 9, and Table 10 provide the descriptive statistics for the dependent variables for Information Retrieval, Information Comparison, and Information Integration tasks, respectively. Table 11 provides the descriptive statistics for the control variables (i.e., task familiarity, motivation, and visualization ability) across all experimental treatments.

Table 8. Descriptive Statistics for Information Retrieval Tasks

	View Time	Fixation Count	Solution Accuracy	Solution Time
Treatment	Mean (s.d.)	Mean (s.d.)	Mean (s.d.)	Mean (s.d.)
1	9.53 (2.07)	15.06 (3.03)	50.58 (6.60)	31.86 (11.29)
2	9.62 (2.25)	16.93 (3.40)	38.22 (24.10)	34.99 (13.04)
3	8.73 (2.52)	13.96 (4.11)	9.61 (13.47)	30.23 (14.31)
4	8.71 (2.70)	15.23 (4.80)	55.01 (14.38)	33.88 (16.64)
5	8.54 (1.97)	14.32 (3.24)	43.47 (13.78)	27.83 (10.66)
6	9.48 (2.19)	16.39 (3.51)	40.23 (12.71)	32.10 (12.21)
7	11.61 (2.97)	20.00 (5.33)	22.72 (9.91)	50.31 (36.41)
8	10.24 (2.08)	17.87 (3.35)	50.27 (23.42)	37.04 (9.73)

Table 9. Descriptive Statistics for Information Comparison Tasks

	View Time	Fixation Count	Solution Accuracy	Solution Time
Treatment	Mean (s.d.)	Mean (s.d.)	Mean (s.d.)	Mean (s.d.)
1	10.12 (2.09)	16.61 (3.69)	79.70 (3.47)	39.71 (13.34)
2	8.23 (1.39)	15.42 (2.61)	85.98 (9.59)	29.88 (8.12)
3	8.75 (1.90)	15.08 (2.71)	56.49 (11.55)	34.21 (11.09)
4	7.52 (1.72)	13.82 (2.88)	80.05 (12.67)	27.43 (8.72)
5	8.89 (1.47)	14.86 (2.97)	68.54 (8.45)	34.01 (7.52)
6	8.48 (1.36)	15.16 (1.89)	77.08 (12.34)	29.76 (7.05)
7	9.48 (1.77)	16.92 (2.61)	51.38 (12.18)	37.84 (12.78)
8	10.00 (2.93)	17.57 (4.35)	76.08 (12.64)	42.47 (20.49)

Table 10. Descriptive Statistics for Information Integration Tasks

	View Time	Fixation Count	Solution Accuracy	Solution Time
Treatment	Mean (s.d.)	Mean (s.d.)	Mean (s.d.)	Mean (s.d.)
9	8.74 (1.99)	16.55 (3.71)	67.50 (15.36)	36.24 (12.67)
10	11.92 (3.22)	22.44 (5.69)	69.03 (26.86)	63.31 (27.06)
11	10.57 (1.51)	19.61 (2.53)	62.92 (21.36)	50.57 (10.86)
12	11.80 (2.41)	22.08 (4.78)	69.44 (24.11)	60.11 (22.67)

Table 11. Descriptive Statistics for Control Variables

	Task Familiarity	Motivation	Visualization Ability
Treatment	Mean (s.d.)	Mean (s.d.)	Mean (s.d.)
1	4.21 (1.43)	4.45 (1.04)	5.55 (0.69)
2	4.94 (1.18)	5.19 (1.09)	4.25 (1.82)
3	4.45 (0.90)	4.06 (1.10)	4.27 (1.56)
4	3.92 (1.00)	4.28 (1.03)	4.67 (1.44)
5	4.67 (0.88)	4.39 (1.18)	5.45 (0.69)
6	4.50 (1.40)	4.28 (0.92)	5.17 (0.83)
7	4.82 (1.00)	4.36 (1.01)	4.82 (1.17)
8	4.52 (1.09)	4.76 (1.12)	4.64 (1.80)
9	3.94 (1.73)	3.81 (1.21)	5.25 (1.06)
10	3.78 (1.32)	3.83 (1.05)	4.50 (1.83)
11	4.75 (0.75)	4.50 (1.25)	5.33 (1.15)
12	4.17 (1.61)	4.42 (1.00)	5.25 (1.06)

Eye Tracker Results

To test the first hypothesis (i.e., H1), multivariate analyses of covariance (MANCOVA) were performed separately for Information Retrieval, Information Comparison, and Information Integration tasks. For each one of these three task types, View Time and Fixation Count were included as the dependent variables in a MANCOVA. MANCOVA was the appropriate method for these analyses, because the dependent variables (i.e., View Time and Fixation Count for each task type) are conceptually related (i.e., reflecting efficiency of information extraction) and are highly (i.e., above 90%) correlated (see Table 12). Task familiarity, motivation, and visualization ability were included as control variables in the MANCOVAs. If significant multivariate effects were observed, univariate tests were then performed to determine the nature of these effects. The multivariate and univariate tests for Information Retrieval, Information Comparison, and Information Integration tasks are now explained in detail.

Table 12. Correlation Matrix for Dependent and Control Variables

Pearson Correlation	IRFC	ICFC	IIFC	IRVT	ICVT	IIVT	Task Fam.	Motiv.	Vis. Ability
IRFC	1.00								
ICFC	.50**	1.00							
IIFC	N/A	N/A	1.00						
IRVT	.96**	.54**	N/A	1.00					
ICVT	.41**	.94**	N/A	.51**	1.00				
IIVT	N/A	N/A	.98*	N/A	N/A	1.00			
Task Familiarity	.15	.12	.02	.16	.08	.08	1.00		
Motivation	.15	.25*	.16	.16	.20	.19	.50**	1.00	
Visualization Ability	-.17	.00	-.05	-.09	.07	-.05	.20*	.02	1.00

** Correlation significant at the 0.01 level (2-tailed).
* Correlation significant at the 0.05 level (2-tailed).

For Information Retrieval tasks, visualization (discrete vs. continuous), volume (low vs. high), and variety (low vs. high) were entered as the independent variables. Table 13 shows the results for the multivariate tests performed with Information Retrieval Fixation Count (IRFC) and Information Retrieval View Time (IRVT) as the dependent variables. Visualization (Pillai's Trace = 0.174, $F = 8.317$, $p < 0.001$), Variety (Pillai's Trace = 0.122, $F = 5.491$, $p < 0.006$), and the Visualization X Variety (Pillai's Trace = 0.088, $F = 3.794$, $p < 0.027$) and Volume X Variety (Pillai's Trace = 0.088, $F = 3.717$, $p < 0.029$) interactions had significant multivariate effects on the dependent variables. The corrected model for Information Retrieval Fixation Count ($F(10,80)=2.724$, $p<0.006$) was significant, with an adjusted R-squared of 0.161 and a partial Eta-squared of 0.254. The p-value of the model for Information Retrieval View Time was slightly above conventional levels of significance ($F(10,80)=1.875$, $p<0.061$). Hence, the effects on Information Retrieval View Time are not interpreted.

Table 13. Multivariate Tests for Information Retrieval Fixation Count and Information Retrieval View Time

Effect	Pillai's Trace	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Intercept	.381	24.364	2	79	.000	.381	48.728	1.000
Motivation	.029	1.197	2	79	.308	.029	2.394	.255
Task Familiarity	.019	.773	2	79	.465	.019	1.545	.177
Visualization Ability	.119	5.360	2	79	.007	.119	10.720	.828
Visualization	.174	8.317	2	79	.001	.174	16.634	.957
Volume	.015	.603	2	79	.550	.015	1.207	.147
Variety	.122	5.491	2	79	.006	.122	10.981	.837
Visualization * Volume	.010	.397	2	79	.673	.010	.795	.112
Visualization * Variety	.088	3.794	2	79	.027	.088	7.588	.675
Volume * Variety	.086	3.717	2	79	.029	.086	7.435	.666
Visualization * Volume * Variety	.033	1.363	2	79	.262	.033	2.727	.286

Table 14 shows the results for the univariate tests performed with Information Retrieval Fixation Count (IRFC) and Information Retrieval View Time (IRVT) as the dependent variables. A significant Volume X Variety interaction effect on Information Retrieval Fixation Count was observed between participants ($F(1,80)=6.872$, $p<0.010$). As shown in Figure 3, participants who were provided with high volume and variety of information had the highest fixation counts (i.e., they moved their gaze most frequently, fixating on the most number of distinct points), suggesting that they struggled the most while trying to extract information from the visualizations they were provided. However, low volume and variety of information, or high volume information with low variety, did not result in as high fixation counts. Taken together with the finding that visualization type did not have a significant main or interaction effect on the View Time and Fixation Count for Information Retrieval tasks, these results suggest that Variety had the strongest impact on the efficiency with which the participants extracted information from the visualizations while solving Information Retrieval tasks. Therefore, Hypothesis 1 was not supported for Information Retrieval tasks, because visualization type (i.e., singular vs. continuous) was not observed to affect the efficiency of information extraction, hence the cognitive fit, for these tasks.

Table 14. Univariate Between-Subjects Effects for Information Retrieval Fixation Count and Information Retrieval View Time

Source	DV	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2	Noncent. Parameter	Observed Power
Corrected Model	IRFC	404.747	10	40.475	2.724	.006	.254	27.244	.950
	IRVT	103.454	10	10.345	1.875	.061	.190	18.752	.816
Intercept	IRFC	656.253	1	656.253	44.173	.000	.356	44.173	1.000
	IRVT	188.125	1	188.125	34.100	.000	.299	34.100	1.000
Motivation	IRFC	9.102	1	9.102	.613	.436	.008	.613	.121
	IRVT	6.700	1	6.700	1.215	.274	.015	1.215	.193
Task Familiarity	IRFC	13.730	1	13.730	.924	.339	.011	.924	.158
	IRVT	2.889	1	2.889	.524	.471	.007	.524	.110
Visualization Ability	IRFC	58.002	1	58.002	3.904	.052	.047	3.904	.497
	IRVT	8.383	1	8.383	1.520	.221	.019	1.520	.230
Visualization	IRFC	4.556	1	4.556	.307	.581	.004	.307	.085
	IRVT	1.429	1	1.429	.259	.612	.003	.259	.079
Volume	IRFC	17.552	1	17.552	1.181	.280	.015	1.181	.189
	IRVT	5.505	1	5.505	.998	.321	.012	.998	.167
Variety	IRFC	86.410	1	86.410	5.816	.018	.068	5.816	.664
	IRVT	16.694	1	16.694	3.026	.086	.036	3.026	.405
Visualization * Volume	IRFC	11.639	1	11.639	.783	.379	.010	.783	.141
	IRVT	3.692	1	3.692	.669	.416	.008	.669	.128
Visualization * Variety	IRFC	7.880	1	7.880	.530	.469	.007	.530	.111
	IRVT	.002	1	.002	.000	.986	.000	.000	.050
Volume * Variety	IRFC	102.100	1	102.100	6.872	.010	.079	6.872	.736
	IRVT	29.837	1	29.837	5.408	.023	.063	5.408	.632
Visualization * Volume * Variety	IRFC	40.584	1	40.584	2.732	.102	.033	2.732	.372
	IRVT	13.308	1	13.308	2.412	.124	.029	2.412	.335
Error	IRFC	1188.523	80	14.857					
	IRVT	441.352	80	5.517					
Total	IRFC	25530.000	91						
	IRVT	8840.320	91						
Corrected Total	IRFC	1593.270	90						
	IRVT	544.806	90						

Information Retrieval Fixation Count (IRFC) R Squared = .254 (Adjusted R Squared = .161)
Information Retrieval View Time (IRVT) R Squared = .190 (Adjusted R Squared = .089)

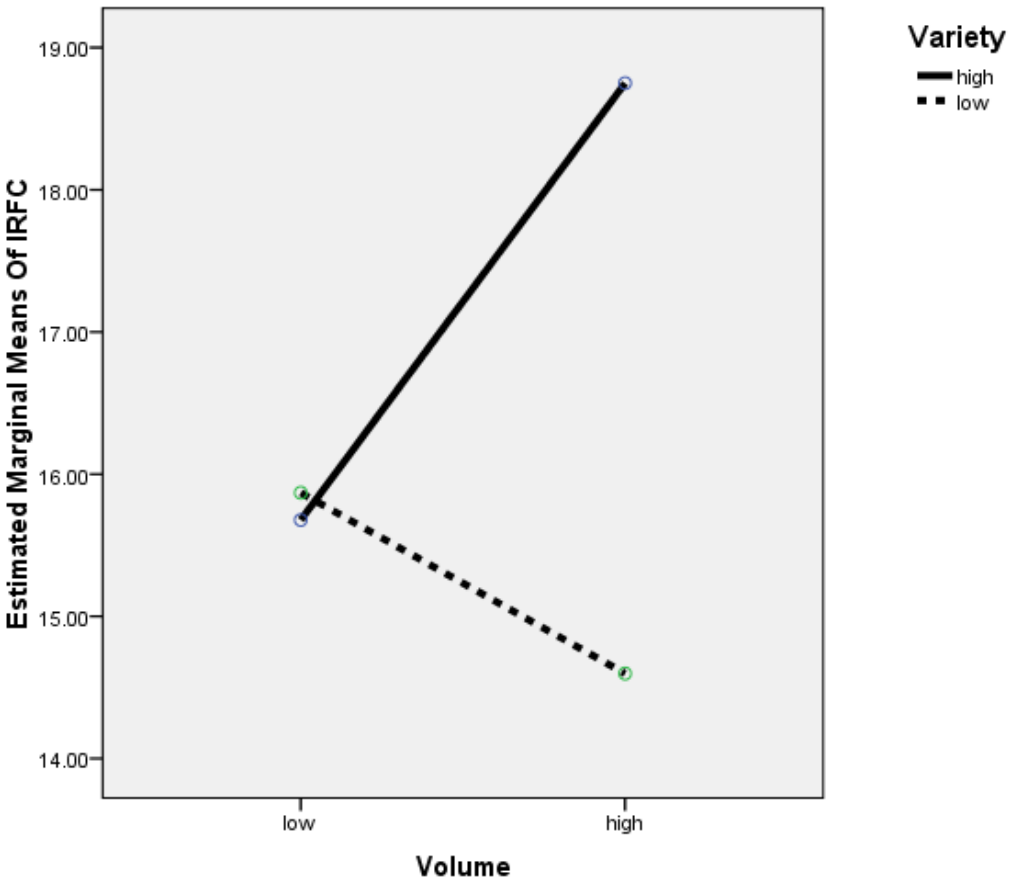


Figure 3. Volume X Variety Interaction on Information Retrieval Fixation Count

For Information Comparison tasks, visualization (discrete vs. continuous), volume (low vs. high), and variety (low vs. high) were entered as the independent variables. Table 15 shows the results for the multivariate tests performed with Information Comparison Fixation Count (ICFC) and Information Comparison View Time (ICVT) as the dependent variables. Visualization (Pillai's Trace = 0.173, F = 8.238, $p < 0.001$) and the Visualization X Variety (Pillai's Trace = 0.085, F = 3.662, $p < 0.030$) and Volume X Variety (Pillai's Trace = 0.073, F = 3.115, $p < 0.050$) interactions had significant multivariate effects on the dependent variables. The corrected model for Information Comparison View Time ($F(10,80)=2.320$, $p<0.019$) was significant, with

an adjusted R-squared of 0.128 and a partial Eta-squared of 0.225. The p-value of the model for Information Comparison Fixation Count was slightly above conventional levels of significance ($F(10,80)=1.908$, $p<0.056$). Hence, the effects on Information Comparison Fixation Count are not interpreted.

Table 15. Multivariate Tests for Information Comparison Fixation Count and Comparison Retrieval View Time

Effect	Pillai's Trace	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Intercept	.389	25.100	2	79	.000	.389	50.200	1.000
Motivation	.054	2.274	2	79	.110	.054	4.548	.450
Task Familiarity	.014	.568	2	79	.569	.014	1.137	.141
Visualization Ability	.014	.579	2	79	.563	.014	1.158	.143
Visualization	.173	8.238	2	79	.001	.173	16.476	.955
Volume	.026	1.071	2	79	.348	.026	2.142	.232
Variety	.029	1.184	2	79	.311	.029	2.369	.252
Visualization * Volume	.058	2.410	2	79	.096	.058	4.820	.473
Visualization * Variety	.085	3.662	2	79	.030	.085	7.323	.659
Volume * Variety	.073	3.115	2	79	.050	.073	6.229	.584
Visualization * Volume * Variety	.001	.048	2	79	.953	.001	.096	.057

Table 16 shows the results for the univariate tests performed with Information Comparison Fixation Count (ICFC) and Information Comparison View Time (ICVT) as the dependent variables. According to the univariate, between-subjects tests, participants in the continuous visualization condition (mean=8.526, s.d.=2.069) had significantly shorter View Times ($F(1,80)=5.073$, $p<0.027$), indicating that they took a shorter amount of time to extract information compared to the participants in the discrete visualization condition (mean=9.310, s.d.=1.843). Furthermore, significant

Volume X Variety ($F(1,80)=4.772$, $p<0.032$) and Visualization X Variety ($F(1,80)=4.810$, $p<0.031$) interaction effects on Information Comparison View Time was observed between participants.

Table 16. Univariate Between-Subjects Effects for Information Comparison Fixation Count and Information Comparison View Time

Source	DV	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2	Noncent. Parameter	Observed Power
Corrected Model	ICFC	169.839	10	16.984	1.908	.056	.193	19.076	.824
	ICVT	80.259	10	8.026	2.320	.019	.225	23.200	.904
Intercept	ICFC	445.321	1	445.321	50.017	.000	.385	50.017	1.000
	ICVT	140.480	1	140.480	40.609	.000	.337	40.609	1.000
Motivation	ICFC	39.953	1	39.953	4.487	.037	.053	4.487	.553
	ICVT	15.441	1	15.441	4.464	.038	.053	4.464	.551
Task Familiarity	ICFC	.998	1	.998	.112	.739	.001	.112	.063
	ICVT	1.479	1	1.479	.427	.515	.005	.427	.099
Visualization Ability	ICFC	.316	1	.316	.035	.851	.000	.035	.054
	ICVT	.107	1	.107	.031	.861	.000	.031	.053
Visualization	ICFC	8.820	1	8.820	.991	.323	.012	.991	.166
	ICVT	17.551	1	17.551	5.073	.027	.060	5.073	.605
Volume	ICFC	4.565	1	4.565	.513	.476	.006	.513	.109
	ICVT	.214	1	.214	.062	.804	.001	.062	.057
Variety	ICFC	20.910	1	20.910	2.349	.129	.029	2.349	.328
	ICVT	7.992	1	7.992	2.310	.132	.028	2.310	.324
Visualization * Volume	ICFC	.063	1	.063	.007	.933	.000	.007	.051
	ICVT	2.277	1	2.277	.658	.420	.008	.658	.126
Visualization * Variety	ICFC	20.947	1	20.947	2.353	.129	.029	2.353	.329
	ICVT	16.639	1	16.639	4.810	.031	.057	4.810	.582
Volume * Variety	ICFC	54.399	1	54.399	6.110	.016	.071	6.110	.685
	ICVT	16.508	1	16.508	4.772	.032	.056	4.772	.579
Visualization * Volume * Variety	ICFC	.137	1	.137	.015	.901	.000	.015	.052
	ICVT	.002	1	.002	.000	.982	.000	.000	.050
Error	ICFC	712.267	80	8.903					
	ICVT	276.749	80	3.459					
Total	ICFC	23173.000	91						
	ICVT	7573.330	91						
Corrected Total	ICFC	882.106	90						
	ICVT	357.008	90						

Information Comparison Fixation Count (ICFC) R Squared = .193 (Adjusted R Squared = .092)
Information Comparison View Time (ICVT) R Squared = .225 (Adjusted R Squared = .128)

As shown in Figure 4, participants who were provided with high volume and variety of information had the highest view times, suggesting that they struggled the

most while trying to extract information from the visualizations. However, low volume and variety of information, or high volume information with low variety, did not result in such shorter view times. Finally, high volume and low variety information recorded the lowest view times, indicating that Variety had a stronger impact than Volume on the efficiency with which the participants extracted information from the visualizations while solving Information Comparison tasks, as with Information Retrieval tasks.

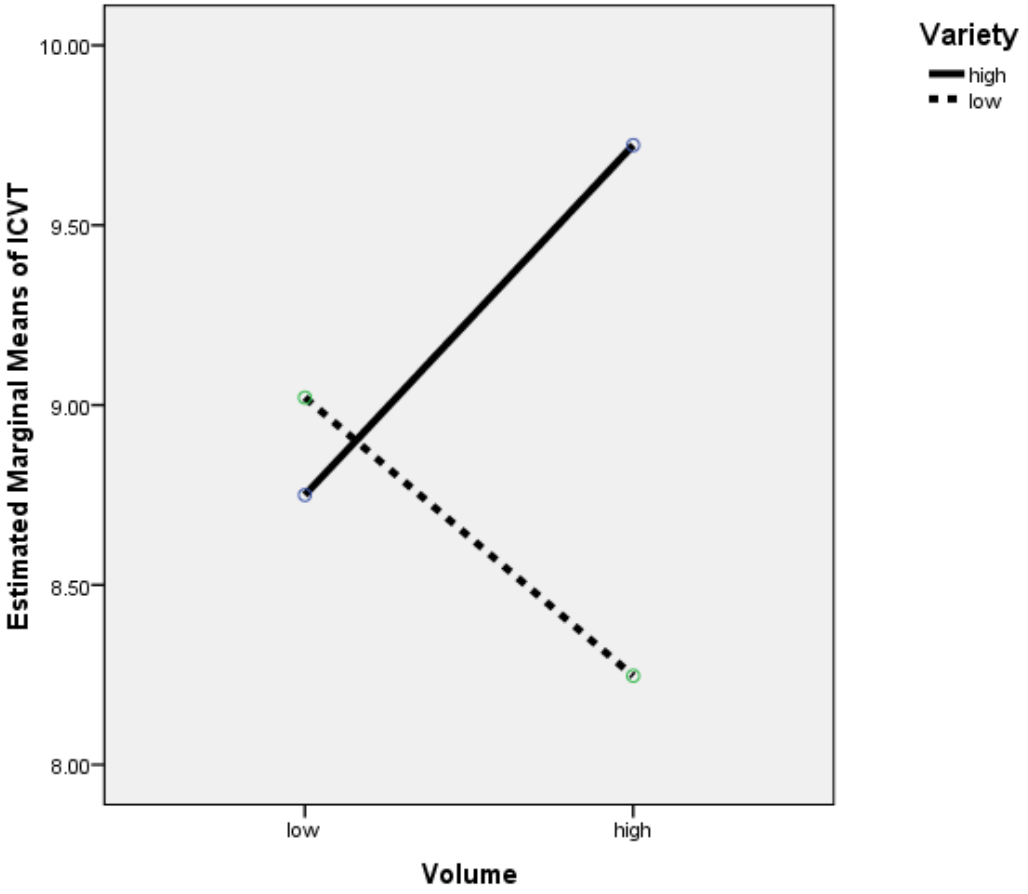


Figure 4. Variety X Volume Interaction on Information Comparison View Time

Figure 5 shows that, when provided with a low variety of information, participants in the discrete visualization spent more time extracting relevant information

from the visualizations for Information Comparison tasks, compared to the participants in the continuous visualization condition. This gap was completely closed when the participants were provided with a high variety of information. These results support the argument that continuous visualizations provide a better cognitive fit for Information Comparison tasks than do discrete visualizations, but this advantage only exists when the visualized information has low variety. Therefore, Hypothesis 1 was supported for Information Comparison tasks.

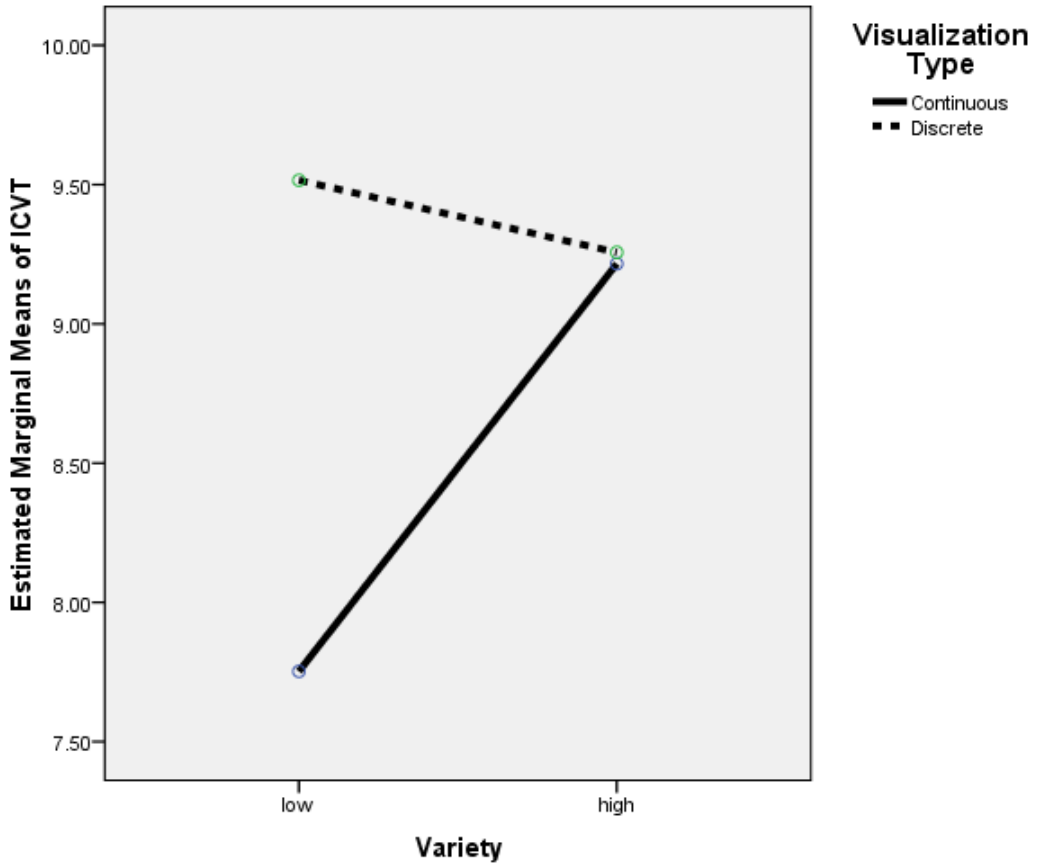


Figure 5. Visualization X Variety Interaction on Information Comparison View Time

For Information Integration tasks, visualization (singular vs. multiple) and variety (low vs. high) were entered as the independent variables. Table 17 shows the results for the multivariate tests performed with Information Integration Fixation Count (IIFC) and Information Integration View Time (IIVT) as the dependent variables. Visualization (Pillai's Trace = 0.198, F = 4.951, p < 0.012) had the only significant multivariate effect on the dependent variables. The corrected models for Information Integration Fixation Count (F(6,41)=2.412, p<0.043) and Information Integration View Time (F(6,41)=2.458, p<0.040) were both significant, with adjusted R-squares of 0.153 and 0.157 and partial Eta-squares of 0.261 and 0.265, respectively.

Table 17. Multivariate Tests for Information Integration Fixation Count and Information Integration View Time

Effect	Pillai's Trace	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Intercept	.383	12.436	2	40	.000	.383	24.873	.994
Motivation	.022	.442	2	40	.646	.022	.884	.117
Task Familiarity	.068	1.461	2	40	.244	.068	2.922	.294
Visualization Ability	.001	.023	2	40	.977	.001	.046	.053
Visualization	.198	4.951	2	40	.012	.198	9.901	.779
Variety	.019	.387	2	40	.682	.019	.774	.108
Visualization * Variety	.042	.876	2	40	.424	.042	1.752	.190

Table 18 shows the results for the univariate tests performed with Information Integration Fixation Count (IIFC) and Information Integration View Time (IIVT) as the dependent variables. According to the univariate, between-subjects tests, participants in the singular visualization condition (mean=9.658, s.d.=1.963) had significantly shorter View Times (F(1,41)=9.862, p<0.003), indicating that they took a shorter amount of time to extract information compared to the participants in the multiple visualizations

condition (mean=11.863, s.d.=2.783). Furthermore, participants in the singular visualization condition (mean=18.080, s.d.=3.480) had significantly lower Fixation Counts than the participants in the continuous visualization condition (mean=22.258, s.d.=5.142) did ($F(1,41)=10.134$, $p<0.003$). There were no other significant main or interaction effects on the dependent variables. These results support the argument that singular visualizations provide a better cognitive fit for Information Integration tasks than do multiple visualizations. Therefore, Hypothesis 1 was supported for Information Integration tasks.

Table 18. Univariate Between-Subjects Effects for Information Integration Fixation Count and Information Integration View Time

Source	DV	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2	Noncent. Parameter	Observed Power
Corrected Model	IIFC	285.960	6	47.660	2.412	.043	.261	14.475	.757
	IIVT	85.983	6	14.331	2.458	.040	.265	14.747	.766
Intercept	IIFC	489.747	1	489.747	24.790	.000	.377	24.790	.998
	IIVT	128.722	1	128.722	22.077	.000	.350	22.077	.996
Motivation	IIFC	17.885	1	17.885	.905	.347	.022	.905	.153
	IIVT	5.126	1	5.126	.879	.354	.021	.879	.150
Task Familiarity	IIFC	.663	1	.663	.034	.856	.001	.034	.054
	IIVT	.180	1	.180	.031	.861	.001	.031	.053
Visualization Ability	IIFC	.099	1	.099	.005	.944	.000	.005	.051
	IIVT	.004	1	.004	.001	.979	.000	.001	.050
Visualization	IIFC	200.199	1	200.199	10.134	.003	.198	10.134	.875
	IIVT	57.498	1	57.498	9.862	.003	.194	9.862	.866
Variety	IIFC	10.738	1	10.738	.544	.465	.013	.544	.111
	IIVT	3.963	1	3.963	.680	.414	.016	.680	.127
Visualization * Variety	IIFC	34.211	1	34.211	1.732	.196	.041	1.732	.251
	IIVT	10.464	1	10.464	1.795	.188	.042	1.795	.258
Error	IIFC	809.990	41	19.756					
	IIVT	239.051	41	5.831					
Total	IIFC	20622.000	48						
	IIVT	5882.820	48						
Corrected Total	IIFC	1095.951	47						
	IIVT	325.035	47						

Information Integration Fixation Count (IIFC) R Squared = .261 (Adjusted R Squared = .153)
Information Integration View Time (IIVT) R Squared = .265 (Adjusted R Squared = .157)

With the exception of Information Retrieval tasks, the results of the analyses so far indicate significant differences in the efficiency of information extraction, as measured through the eye tracker data (i.e., View Time and Fixation Count), between participants who were provided with different types of visualizations. Specifically, when solving Information Comparison tasks, participants were more efficient in extracting information from continuous visualizations compared to discrete visualizations. As for Information Integration tasks, efficiency of information extraction was greater with singular visualizations compared to multiple visualizations. Taken together, these findings suggest that the cognitive fit between visualizations and data analysis task types does manifest in the efficiency of information extraction (i.e., as fewer gaze fixations and less time viewing the visualization), providing overall support for Hypothesis 1.

Task Performance Results

To test the remaining hypotheses (i.e., H2-H9), several MANCOVAs were performed with the two aspects of task performance (i.e., solution time and accuracy) as the dependent variables. This approach is consistent with previous cognitive fit studies (e.g., Goswami et al., 2008; Vessey and Galletta, 1991). Visualization type, volume, and variety were included as the independent variables. Task familiarity, motivation, and visualization ability were included as the control variables. Each hypothesis was tested by limiting the sample to the participants who worked on a specific task type, and then testing for a main or interaction effect, as summarized in Table 19. Similar to the

procedure for testing Hypothesis 1, univariate tests were then performed to interpret the observed multivariate effects.

Table 19. Hypothesis Testing

Hypotheses	Filter by Task Type	Test for Effect
H2: For **Information Retrieval** tasks, discrete visualizations will provide a better cognitive fit than continuous visualizations, resulting in (a) more accurate and (b) faster decisions.	Information Retrieval	Visualization Type
H3: For **Information Comparison** tasks, continuous visualizations will provide a better cognitive fit than discrete visualizations, resulting in (a) more accurate and (b) faster decisions.	Information Comparison	Visualization Type
H4: For **Information Integration** tasks, singular visualizations will provide a better cognitive fit than multiple visualizations, resulting in (a) more accurate and (b) faster decisions.	Information Integration	Visualization Type
H5: For **Information Retrieval** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger volume**.	Information Retrieval	Visualization Type × Volume
H6: For **Information Retrieval** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger variety**.	Information Retrieval	Visualization Type × Variety
H7: For **Information Comparison** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger volume**.	Information Comparison	Visualization Type × Volume
H8: For **Information Comparison** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger variety**.	Information Comparison	Visualization Type × Variety
H9: For **Information Integration** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger variety**.	Information Integration	Visualization Type × Variety

To test hypotheses H2, H5, and H6, the analysis included the participants who solved Information Retrieval tasks. A MANCOVA was performed with solution time (Information Retrieval Solution Time) and accuracy (Information Retrieval Solution Accuracy) as the dependent variables. Visualization (discrete vs. continuous), volume (low vs. high), and variety (low vs. high) were included as the independent variables. Task familiarity, motivation, and visualization ability were modeled as the control variables.

Table 20 shows the results for the multivariate tests performed with Information Retrieval Solution Time (IRST) and Information Retrieval Solution Accuracy (IRSA) as the dependent variables. Visualization (Pillai's Trace = 0.166, $F = 8.489$, $p < 0.001$), Volume (Pillai's Trace = 0.143, $F = 7.064$, $p < 0.001$), and their interaction (Pillai's Trace = 0.355, $F = 23.413$, $p < 0.001$) had significant multivariate effects on the dependent variables. The corrected model for Information Retrieval Solution Accuracy ($F(10,86)=9.128$, $p<0.001$) was significant, with an adjusted R-squared of 0.458 and a partial Eta-squared of 0.515. The model for Information Retrieval Solution Time was not significant ($F(10,86)=1.527$, $p<0.144$). Therefore, the effects on Information Retrieval Solution Time are not interpreted.

Table 20. Multivariate Tests for Information Retrieval Solution Time and Information Retrieval Solution Accuracy

Effect	Pillai's Trace	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Intercept	.245	13.811	2	85	.000	.245	27.621	.998
Motivation	.014	.610	2	85	.546	.014	1.219	.149
Task Familiarity	.035	1.527	2	85	.223	.035	3.054	.317
Visualization Ability	.019	.828	2	85	.440	.019	1.656	.188
Visualization	.166	8.489	2	85	.000	.166	16.977	.961
Volume	.143	7.064	2	85	.001	.143	14.128	.921
Variety	.022	.965	2	85	.385	.022	1.930	.213
Visualization * Volume	.355	23.413	2	85	.000	.355	46.826	1.000
Visualization * Variety	.016	.700	2	85	.500	.016	1.399	.165
Volume * Variety	.032	1.397	2	85	.253	.032	2.793	.292
Visualization * Volume * Variety	.053	2.383	2	85	.098	.053	4.766	.469

Table 21 shows the results for the univariate tests performed with Information Retrieval Solution Time (IRST) and Information Retrieval Solution Accuracy (IRSA) as the dependent variables. According to the univariate, between-subjects tests, participants in the discrete visualization condition (mean=33.282, s.d.=20.873) had a lower solution accuracy ($F(1,86)=15.627$, $p<0.001$) than the participants in the continuous visualization condition (mean=45.845, s.d.=19.905). In other words, compared with the participants in the discrete visualization condition, participants in the continuous visualization condition performed better with Information Retrieval tasks, suggesting that continuous visualizations provided a better cognitive fit for Information Retrieval tasks, contrary to what was expected. Therefore, Hypothesis 2 was not supported.

Table 21. Univariate Between-Subjects Effects for Information Retrieval Solution Time and Information Retrieval Solution Accuracy

Source	DV	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2	Noncent. Parameter	Observed Power
Corrected Model	IRST	4358.271	10	435.827	1.527	.144	.151	15.273	.715
	IRSA	22344.087	10	2234.409	9.128	.000	.515	91.276	1.000
Intercept	IRST	2162.703	1	2162.703	7.579	.007	.081	7.579	.777
	IRSA	5726.414	1	5726.414	23.393	.000	.214	23.393	.998
Motivation	IRST	319.284	1	319.284	1.119	.293	.013	1.119	.182
	IRSA	55.312	1	55.312	.226	.636	.003	.226	.076
Task Familiarity	IRST	128.761	1	128.761	.451	.504	.005	.451	.102
	IRSA	566.895	1	566.895	2.316	.132	.026	2.316	.325
Visualization Ability	IRST	428.501	1	428.501	1.502	.224	.017	1.502	.228
	IRSA	15.570	1	15.570	.064	.801	.001	.064	.057
Visualization	IRST	145.004	1	145.004	.508	.478	.006	.508	.109
	IRSA	3825.353	1	3825.353	15.627	.000	.154	15.627	.974
Volume	IRST	1034.280	1	1034.280	3.625	.060	.040	3.625	.469
	IRSA	2183.812	1	2183.812	8.921	.004	.094	8.921	.840
Variety	IRST	542.628	1	542.628	1.902	.171	.022	1.902	.276
	IRSA	40.353	1	40.353	.165	.686	.002	.165	.069
Visualization * Volume	IRST	313.375	1	313.375	1.098	.298	.013	1.098	.179
	IRSA	10680.240	1	10680.240	43.629	.000	.337	43.629	1.000
Visualization * Variety	IRST	293.916	1	293.916	1.030	.313	.012	1.030	.171
	IRSA	137.422	1	137.422	.561	.456	.006	.561	.115
Volume * Variety	IRST	490.011	1	490.011	1.717	.194	.020	1.717	.254
	IRSA	362.352	1	362.352	1.480	.227	.017	1.480	.225
Visualization * Volume * Variety	IRST	470.662	1	470.662	1.649	.202	.019	1.649	.246
	IRSA	916.445	1	916.445	3.744	.056	.042	3.744	.481
Error	IRST	24540.105	86	285.350					
	IRSA	21052.443	86	244.796					
Total	IRST	146027.959	97						
	IRSA	193737.962	97						
Corrected Total	IRST	28898.376	96						
	IRSA	43396.531	96						

Information Retrieval Solution Time (IRST) R Squared = .151 (Adjusted R Squared = .052)
Information Retrieval Solution Accuracy (IRSA) R Squared = .515 (Adjusted R Squared = .458)

A significant Visualization X Volume interaction effect on solution accuracy was observed between participants ($F(1,86)=43.629$, $p<0.001$). As shown in Figure 6, participants who were provided with continuous visualizations performed even better when the visualized information had higher volume, while participants in the discrete visualization condition suffered a great decrease in their solution accuracy as the

amount of visualized information increased. Therefore, Hypothesis 5 was not supported. However, in line with the previous finding that continuous visualizations might provide a better cognitive fit for Information Retrieval tasks, this pattern indicates that the effects of cognitive fit on decision accuracy are amplified when a larger volume of information is being visualized, consistent with the rationale behind Hypothesis 5a. Variety did not have any significant main or interaction effects on the dependent variables, hence failing to support Hypothesis 6.

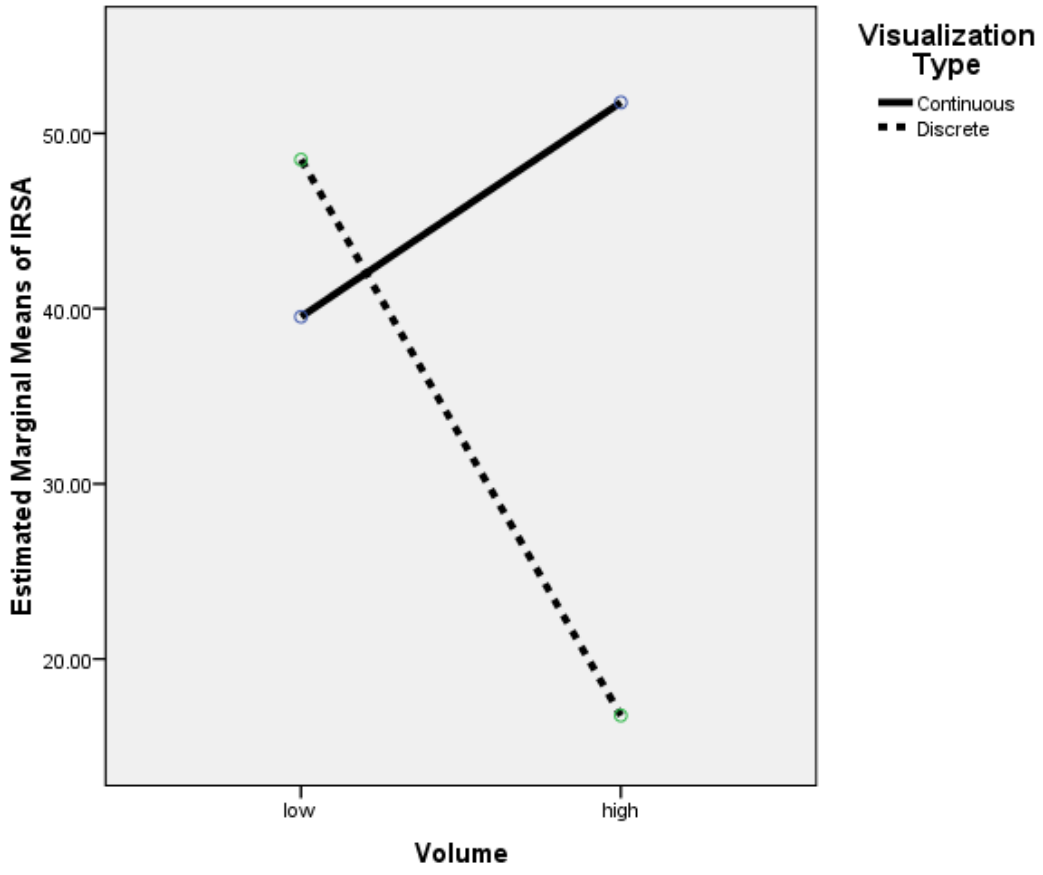


Figure 6. Visualization X Volume Interaction on Information Retrieval Solution Accuracy

To test hypotheses H3, H7, and H8, the analysis included the participants who solved Information Comparison tasks. A MANCOVA was performed with solution time (Information Comparison Solution Time) and accuracy (Information Comparison Solution Accuracy) as the dependent variables. Visualization (discrete vs. continuous), volume (low vs. high), and variety (low vs. high) were included as the independent variables. Task familiarity, motivation, and visualization ability were modeled as the control variables.

Table 22 shows the results for the multivariate tests performed with Information Comparison Solution Time (ICST) and Information Comparison Solution Accuracy (ICSA) as the dependent variables. Visualization (Pillai's Trace = 0.371, $F = 25.053$, $p < 0.001$), Volume (Pillai's Trace = 0.278, $F = 16.358$, $p < 0.001$), Variety (Pillai's Trace = 0.117, $F = 5.622$, $p < 0.005$), and the Visualization X Volume interaction (Pillai's Trace = 0.193, $F = 10.135$, $p < 0.001$) had significant multivariate effects on the dependent variables. The corrected models for Information Comparison Solution Time ($F(10,86)=2.220$, $p<0.024$) and Information Comparison Solution Accuracy ($F(10,86)=11.753$, $p<0.001$) were both significant, with adjusted R-squares of 0.113 and 0.528 and partial Eta-squares of 0.205 and 0.577, respectively.

Table 22. Multivariate Tests for Information Comparison Solution Time and Information Comparison Solution Accuracy

Effect	Pillai's Trace	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Intercept	.591	61.450	2	85	.000	.591	122.900	1.000
Motivation	.070	3.184	2	85	.046	.070	6.369	.595
Task Familiarity	.004	.166	2	85	.847	.004	.332	.075
Visualization Ability	.007	.284	2	85	.753	.007	.569	.094
Visualization	.371	25.053	2	85	.000	.371	50.106	1.000
Volume	.278	16.358	2	85	.000	.278	32.717	1.000
Variety	.117	5.622	2	85	.005	.117	11.244	.848
Visualization * Volume	.193	10.135	2	85	.000	.193	20.270	.983
Visualization * Variety	.032	1.420	2	85	.247	.032	2.841	.297
Volume * Variety	.039	1.731	2	85	.183	.039	3.462	.354
Visualization * Volume * Variety	.012	.526	2	85	.593	.012	1.052	.134

Table 23 shows the results for the univariate tests performed with Information Comparison Solution Time (ICST) and Information Comparison Solution Accuracy (ICSA) as the dependent variables. According to the univariate, between-subjects tests, participants in the discrete visualization condition (mean=37.825, s.d.=12.348) took significantly longer ($F(1,86)=6.512$, $p<0.012$) to solve the Information Comparison tasks compared to the participants in the continuous visualization condition (mean=32.170, s.d.=13.085). Furthermore, participants in the discrete visualization condition (mean=65.175, s.d.=14.963) had a lower solution accuracy than the participants in the continuous visualization condition (mean=79.880, s.d.=12.115) did ($F(1,86)=46.295$, $p<0.001$). Therefore, Hypothesis 3 was supported.

Table 23. Univariate Between-Subjects Effects for Information Comparison Solution Time and Information Comparison Solution Accuracy

Source	DV	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2	Noncent. Parameter	Observed Power
Corrected Model	ICST	3307.666	10	330.767	2.220	.024	.205	22.200	.891
	ICSA	13258.576	10	1325.858	11.753	.000	.577	117.532	1.000
Intercept	ICST	1579.936	1	1579.936	10.604	.002	.110	10.604	.896
	ICSA	12250.734	1	12250.734	108.597	.000	.558	108.597	1.000
Motivation	ICST	765.571	1	765.571	5.138	.026	.056	5.138	.611
	ICSA	109.923	1	109.923	.974	.326	.011	.974	.164
Task Familiarity	ICST	31.338	1	31.338	.210	.648	.002	.210	.074
	ICSA	16.644	1	16.644	.148	.702	.002	.148	.067
Visualization Ability	ICST	3.901	1	3.901	.026	.872	.000	.026	.053
	ICSA	59.893	1	59.893	.531	.468	.006	.531	.111
Visualization	ICST	970.320	1	970.320	6.512	.012	.070	6.512	.713
	ICSA	5222.505	1	5222.505	46.295	.000	.350	46.295	1.000
Volume	ICST	140.995	1	140.995	.946	.333	.011	.946	.161
	ICSA	3695.156	1	3695.156	32.756	.000	.276	32.756	1.000
Variety	ICST	417.273	1	417.273	2.801	.098	.032	2.801	.380
	ICSA	1038.839	1	1038.839	9.209	.003	.097	9.209	.851
Visualization * Volume	ICST	202.976	1	202.976	1.362	.246	.016	1.362	.211
	ICSA	2073.200	1	2073.200	18.378	.000	.176	18.378	.989
Visualization * Variety	ICST	399.455	1	399.455	2.681	.105	.030	2.681	.367
	ICSA	12.144	1	12.144	.108	.744	.001	.108	.062
Volume * Variety	ICST	281.143	1	281.143	1.887	.173	.021	1.887	.274
	ICSA	155.915	1	155.915	1.382	.243	.016	1.382	.213
Visualization * Volume * Variety	ICST	97.310	1	97.310	.653	.421	.008	.653	.126
	ICSA	54.439	1	54.439	.483	.489	.006	.483	.106
Error	ICST	12813.746	86	148.997					
	ICSA	9701.542	86	112.809					
Total	ICST	135523.158	97						
	ICSA	530005.802	97						
Corrected Total	ICST	16121.412	96						
	ICSA	22960.118	96						

Information Comparison Solution Time (ICST) R Squared = .205 (Adjusted R Squared = .113)
Information Comparison Solution Accuracy (ICSA) R Squared = .577 (Adjusted R Squared = .528)

A significant Visualization X Volume interaction effect on solution accuracy was observed between participants ($F(1,86)=18.378$, $p<0.001$). As shown in Figure 7, participants in the discrete visualization condition suffered a great decrease in their solution accuracy as the amount of visualized information increased, while the performance of the participants in the continuous visualization condition remained relatively stable. This finding suggests that, as the volume of visualized information

increases, the negative impacts of visualizations that do not provide cognitive fit are amplified. Therefore, Hypothesis 7a was supported.

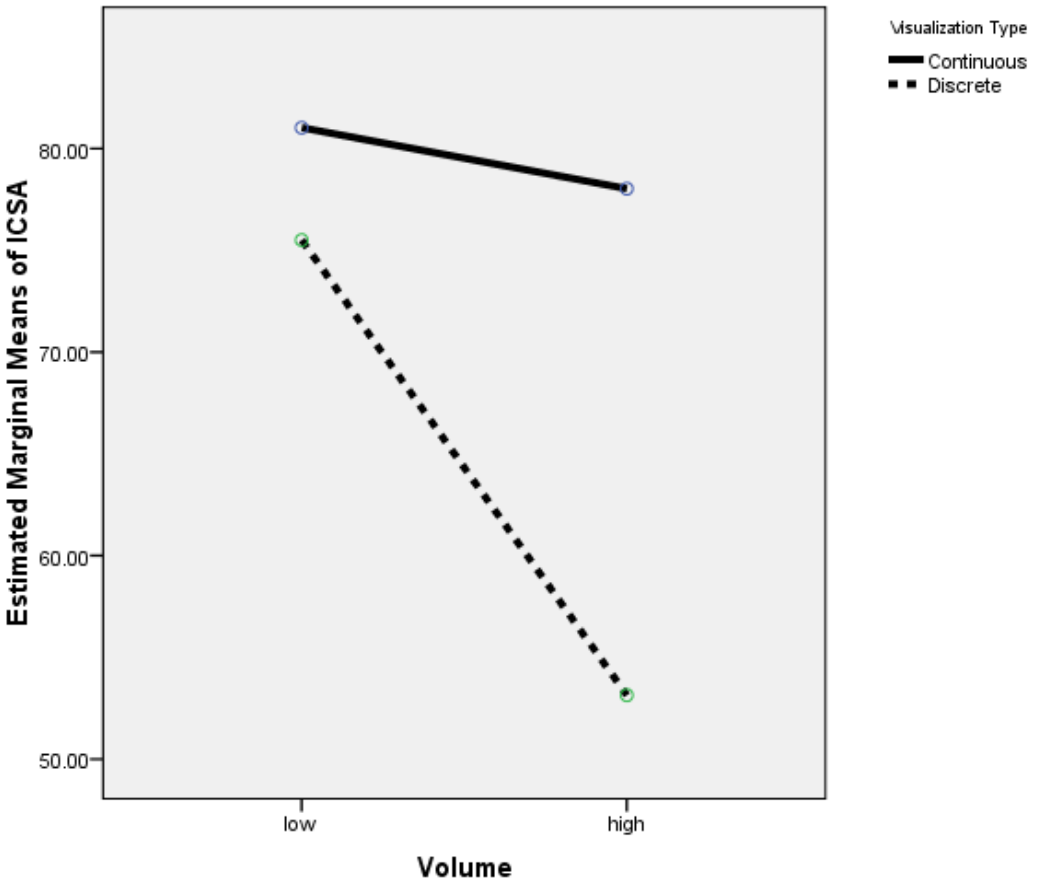


Figure 7. Visualization X Volume Interaction on Information Comparison Solution Accuracy

Although the participants in the low variety condition (mean=75.148, s.d.=15.639) had a higher solution accuracy than the participants in the high variety condition (mean=69.142, s.d.=14.802) did ($F(1,86)=9.209$, $p<0.003$), Variety was not observed to have any significant interaction effects on the dependent variables. Therefore, Hypothesis 8 was not supported.

To test hypotheses H4 and H9, the analysis included the participants who solved Information Integration tasks. A MANCOVA was performed with solution time (Information Integration Solution Time) and accuracy (Information Integration Solution Accuracy) as the dependent variables. Visualization (singular vs. multiple) and variety (low vs. high) were included as the independent variables. Task familiarity, motivation, and visualization ability were modeled as the control variables. Volume was not included as an independent variable in these models because it was not manipulated for Information Integration tasks (see Table 4 for a summary of the experimental treatments).

Table 24 shows the results for the multivariate tests performed with Information Integration Solution Time (IIST) and Information Integration Solution Accuracy (IISA) as the dependent variables. Visualization (Pillai's Trace = 0.199, $F = 4.966$, $p < 0.012$) had significant multivariate effects on the dependent variables. The corrected model for Information Integration Solution Time ($F(6,41)=2.517$, $p<0.036$) was significant, with an adjusted R-squared of 0.162 and a partial Eta-squared of 0.269. The corrected model for Information Integration Solution Accuracy was not significant ($F(6,41)=1.867$, $p<0.110$).

Table 24. Multivariate Tests for Information Integration Solution Time and Information Integration Solution Accuracy

Effect	Pillai's Trace	F	Hypothesis df	Error df	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power
Intercept	.169	4.078	2	40	.024	.169	8.156	.691
Motivation	.080	1.739	2	40	.189	.080	3.477	.343
Task Familiarity	.008	.152	2	40	.859	.008	.305	.072
Visualization Ability	.154	3.648	2	40	.035	.154	7.297	.640
Visualization Variety	.199	4.966	2	40	.012	.199	9.932	.781
Visualization * Variety	.044	.925	2	40	.405	.044	1.850	.199
	.053	1.119	2	40	.337	.053	2.237	.233

Table 25 shows the results for the univariate tests performed with Information Integration Solution Time (IIST) and Information Integration Solution Accuracy (IISA) as the dependent variables. According to the univariate, between-subjects tests, participants in the multiple visualizations condition (mean=61.709, s.d.=24.466) took significantly longer ($F(1,86)=9.288$, $p<0.004$) to solve the Information Integration tasks compared to the participants in the singular visualization condition (mean=43.407, s.d.=13.664). Therefore, Hypothesis 4b was supported. Variety did not have any significant main or interaction effects on the dependent variables, hence failing to support Hypothesis 9.

Table 25. Univariate Between-Subjects Effects for Information Integration Solution Time and Information Integration Solution Accuracy

Source	DV	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta^2	Noncent. Parameter	Observed Power
Corrected Model	IIST	5944.645	6	990.774	2.517	.036	.269	15.104	.778
	IISA	4777.930	6	796.322	1.867	.110	.215	11.201	.626
Intercept	IIST	2950.383	1	2950.383	7.496	.009	.155	7.496	.762
	IISA	484.500	1	484.500	1.136	.293	.027	1.136	.180
Motivation	IIST	629.062	1	629.062	1.598	.213	.038	1.598	.235
	IISA	914.143	1	914.143	2.143	.151	.050	2.143	.298
Task Familiarity	IIST	111.485	1	111.485	.283	.597	.007	.283	.081
	IISA	16.494	1	16.494	.039	.845	.001	.039	.054
Visualization Ability	IIST	4.150	1	4.150	.011	.919	.000	.011	.051
	IISA	3165.555	1	3165.555	7.421	.009	.153	7.421	.758
Visualization	IIST	3655.386	1	3655.386	9.288	.004	.185	9.288	.845
	IISA	513.271	1	513.271	1.203	.279	.029	1.203	.188
Variety	IIST	177.199	1	177.199	.450	.506	.011	.450	.100
	IISA	581.202	1	581.202	1.363	.250	.032	1.363	.207
Visualization * Variety	IIST	893.070	1	893.070	2.269	.140	.052	2.269	.313
	IISA	2.649	1	2.649	.006	.938	.000	.006	.051
Error	IIST	16136.553	41	393.574					
	IISA	17488.477	41	426.548					
Total	IIST	154673.706	48						
	IISA	239168.879	48						
Corrected Total	IIST	22081.199	47						
	IISA	22266.408	47						

Information Integration Solution Time (IIST) R Squared = .269 (Adjusted R Squared = .162)
Information Integration Solution Accuracy (IISA) R Squared = .215 (Adjusted R Squared = .100)

Table 26 presents an overall summary of the results of hypothesis tests. The following chapter provides a discussion of these results and their theoretical and practical implications.

Table 26. Results of Hypothesis Tests

Hypotheses	Results
H1: **Cognitive fit** will be manifested in eye movement patterns such that when there is cognitive fit between the task and visualization, analysts will have less frequent eye movements and fewer but longer gaze fixations.	Supported (for Information Retrieval and Information Comparison tasks)
H2: For **Information Retrieval** tasks, discrete visualizations will provide a better cognitive fit than continuous visualizations, resulting in (a) more accurate and (b) faster decisions.	Not supported (contradicted)
H3: For **Information Comparison** tasks, continuous visualizations will provide a better cognitive fit than discrete visualizations, resulting in (a) more accurate and (b) faster decisions.	Supported
H4: For **Information Integration** tasks, singular visualizations will provide a better cognitive fit than multiple visualizations, resulting in (a) more accurate and (b) faster decisions.	Supported (b)
H5: For **Information Retrieval** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger volume**.	Supported (a)
H6: For **Information Retrieval** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger variety**.	Not supported
H7: For **Information Comparison** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger volume**.	Supported (a)
H8: For **Information Comparison** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger variety**.	Not supported
H9: For **Information Integration** tasks, the effect of cognitive fit on the (a) accuracy and (b) speed of decisions will be greater when the represented data has **larger variety**.	Not supported

Chapter 5: Summary and Conclusions

Discussion

This dissertation studies how the cognitive fit between different types of business data analysis tasks and different visualization techniques can affect task performance, in the context of big data analytics. To do so, a laboratory experiment (n=145) was conducted, and data analysis task and visualization types plus big data characteristics (i.e., volume and variety) were manipulated. While the participants were working on the data analysis problems using different visualizations, their information extraction behaviors (i.e., their gaze movements and fixation counts) were captured via an eye tracker. Cognitive fit was then assessed through the efficiency with which participants extracted information from the provided visualizations. Based on the results of this experiment, this dissertation contributes to the literature in at least three broad avenues. A summary of the findings of this dissertation and the associated contributions are now discussed in detail.

First of all, the results of this study confirm that cognitive fit manifests through the efficiency with which analysts extract information from visualizations. Even though the cognitive fit theory and how cognitive fit affects task performance have been extensively studied in the past three decades, this is the first study in which cognitive fit was captured objectively rather than being manipulated or assumed as a part of the experimental design. This approach (i.e., capturing the physiological correlates of cognitive fit and misfit by using neurophysiological tools such as eye trackers) has recently been suggested as a novel method for improving our understanding of cognitive fit and designing better systems and decision aids (e.g., data analysis tools)

(Dimoka et al., 2012). Capturing cognitive fit via eye trackers also enables us to minimize common method bias by not relying exclusively on self-reported measures (Dimoka et al., 2011), and decreases our susceptibility to other biases such as demand effects, plus social desirability and subjectivity biases (Dimoka et al., 2012). Therefore, this dissertation contributes to the cognitive fit literature both theoretically and methodologically. Theoretically, it extends the cognitive fit model to account for the consequences of big data characteristics (i.e., volume and variety) and different visualization techniques for different types of data analysis tasks. In doing so, this dissertation improves our understanding of how and why cognitive fit manifests in analysts' problem solving behaviors and consequently affects their task performance. Prior to this study, the cognitive fit model (Vessey, 1991) only considered the problem representation and the problem-solving task, while the extended cognitive fit model additionally included the mental representation of the problem (Shaft and Vessey, 2006). Methodologically, this dissertation contributes to the literature by proposing and validating an objective method for assessing cognitive fit through data analysts' gaze patterns, consistent with the suggestions of Dimoka et al. (2012).

It is important to note that cognitive fit was observed to manifest through the efficiency of information extraction only for Information Comparison and Information Integration tasks, and not for Information Retrieval tasks. As shown in Table 8, the mean View Times and Fixation Counts for discrete and continuous visualization conditions were very close to one another, without an excessive amount of variance. This indicates that there was no significant advantage provided by one type of visualization over the other in terms of the efficiency of information extraction for

Information Retrieval tasks. Furthermore, the fact that significant effects of visualization type were observed for Information Comparison and Information Integration tasks suggests that the sample was large enough to reveal the significance of visualization type effects.

Recall that Information Retrieval tasks were the simplest of all three data analysis task types examined in this dissertation. Thus, one potential explanation for this finding is that Information Retrieval tasks, which required the participants to extract only one dimension of information, were simple enough that they could be solved equally efficiently with discrete and continuous visualizations. Put differently, it is possible that participants were able to easily transform the represented information (i.e., overcome cognitive misfit) for Information Retrieval tasks, even when the representation was inconsistent with the problem-solving task requirements. Nevertheless, this finding deserves further investigation, specifically regarding how visualizations can better facilitate information extraction for Information Retrieval tasks.

The second avenue in which this dissertation contributes to the literature is the mapping of business data analysis task types to visualization characteristics, in terms of cognitive fit. First of all, the findings of this study indicate that, contrary to what was hypothesized, continuous visualizations provide better cognitive fit for Information Retrieval tasks, compared to discrete visualizations. This was evident in the difference of task performance when participants' solution accuracy was compared between the discrete (33%) and continuous (46%) visualization conditions. This finding suggests that continuous visualizations that present data in aggregation (e.g., as shades of colors

on a map), while providing the corresponding data values on scales below the visualization (see Appendix A for the visualizations used in the experiment), provide better support for Information Retrieval tasks, compared to discrete visualizations that present data in isolation (i.e., as individual data points).

A possible explanation for this finding is that the volume of information visualized in this experiment was sufficiently high, even in the low volume condition (i.e., 1,000 distinct data points), that participants were better able to estimate data values from continuous visualizations and the scales provided below, as opposed to identifying and counting the individual data points on the discrete visualizations. This explanation is further supported by the finding that participants who were provided with continuous visualizations had an even higher solution accuracy (51%) when the visualized information had higher volume, while higher volume resulted in a considerable decrease in solution accuracy (17%) for the participants who were provided with discrete visualizations. Regardless, these results suggest that continuous visualizations provide better decision-making support for Information Retrieval tasks.

The second finding regarding the mapping of business data analysis task types to visualization characteristics is that continuous visualizations provide better cognitive fit for Information Comparison tasks, compared to discrete visualizations, as expected. This was evident in higher solution accuracy and faster decision times when the participants' task performance was compared between the discrete (65%; 38s) and continuous (80%; 32s) visualization conditions. This finding suggests that continuous visualizations that present data in aggregate, better facilitate the comparison of data

points, as required for Information Comparison tasks, compared to discrete visualizations that present data in isolation.

The third finding regarding the mapping of business data analysis task types to visualization characteristics is that singular visualizations provide better cognitive fit for Information Integration tasks, compared to multiple visualizations, also as expected. This was evident in the difference of task performance when participants' solution time was compared between the singular (43s) and multiple (62s) visualization conditions. This finding suggests that faster decisions can be made with singular visualizations that overlay the relevant dimensions of information, compared to multiple distinct visualizations representing each one of the information dimensions, such as in dashboards.

Overall, the empirical findings regarding cognitive fit suggest that at least one dimension of task performance (i.e., solution time and/or accuracy) can be improved by choosing a matching type of visualization for a given data analysis task. The results of this dissertation indicate that continuous visualizations better support decision-making for both Information Retrieval and Information Comparison tasks, compared to discrete visualizations. Furthermore, the results also indicate that singular visualizations provide better decision-making support for Information Integration tasks, compared to multiple visualizations like dashboards. These findings provide important implications for data analysts that rely on visualizations to solve business data analysis problems; data analysts first need to determine the type of the data analysis task they are working on, before selecting a specific kind of visualization to use. Then, as the results of this study indicate, using continuous visualizations results in better decisions for Information

Retrieval and Information Comparison tasks, while singular visualizations have been observed to result in faster decisions for Information Integration tasks. The practical implication of this finding is that the use of dashboards, as is the frequent practice today (Davenport and Dyche, 2013) is not efficient, especially for Information Integration tasks.

The third avenue in which this dissertation contributes to the literature is the identification of the role that characteristics of big data play in influencing the task performance consequences of cognitive fit. Specifically, this dissertation observed how the two defining characteristics of big data (i.e., volume and variety of information) affect the impacts of cognitive fit on two different aspects of task performance (i.e., solution time and solution accuracy). The overall findings of this research indicate that high volume and high variety of information both amplify the difference in task performance between the visualizations that provide cognitive fit and those that do not, for a given type of business data analysis task.

For Information Retrieval tasks, continuous visualizations were observed to provide better decision-making support, compared to discrete visualizations. As discussed before, the difference in Information Retrieval solution accuracy for discrete vs. continuous visualizations was amplified when a larger volume of information was being visualized. However, manipulating the variety of the visualized information was not observed to affect directly or indirectly the task performance for Information Retrieval tasks. Nevertheless, these results support the argument that the task performance consequences of cognitive fit for Information Retrieval tasks are amplified

in the context of big data analytics, due to the high volume of information being visualized.

For Information Comparison tasks, continuous visualizations again were observed to provide better decision-making support, compared to discrete visualizations. As the volume of visualized information increased, participants in the discrete visualization condition suffered a considerable decrease in their solution accuracy (76% to 63%), while the task performance of the participants in the continuous visualization condition decreased very slightly and remained relatively stable (81% to 78%). This finding suggests that, as the volume of visualized information increases, the negative impacts of visualizations that do not provide cognitive fit were amplified. As with Information Retrieval tasks, the variety of the visualized information was not observed to influence the task performance consequences of cognitive fit for Information Comparison tasks. However, participants in the low variety condition (75%) had higher solution accuracy than the participants in the high variety condition (69%) for Information Comparison tasks, regardless of the visualization type. Therefore, these results support the arguments that big data is especially challenging to analyze (due to high variety of information), and that the task performance consequences of cognitive fit for Information Comparison tasks are amplified in the context of big data (due to high volume of information).

For Information Integration tasks, singular visualizations were observed to provide better decision-making support, compared to multiple visualizations. However, the variety of visualized information was not observed to affect directly or indirectly the task performance for Information Integration tasks. Nevertheless, since big data

analytics are still frequently performed through dashboards (i.e., multiple visualizations) (Chen et al., 2012; Eaton et al., 2012; Davenport and Dyche, 2013), this finding provides an important insight to data analysts, that this common approach of relying on dashboards might result in inferior analytics performance compared to if singular visualizations are used.

Overall, these findings suggest that when visualizing high volumes and large varieties of information, it is even more consequential and thus more important to choose a visualization type that properly supports the data analysis task in hand. Recall that the research question driving this dissertation regards the facilitation of big data analytics by visualizations that provide cognitive fit. The results of this dissertation indicate that continuous visualizations can better facilitate big data analytics, compared to discrete visualizations, when the analysts are faced with Information Retrieval and Information Comparison tasks. In addition, singular visualizations were observed to better facilitate big data analytics, compared to multiple visualizations, when the analysts are working on Information Integration tasks. Considering that the use and importance of big data analytics is growing rapidly in today's business environment (Columbus, 2015; Eaton et al., 2012), the results of this dissertation provide important insights for decision-makers regarding how to make the best use of this asset. Nevertheless, these results were obtained through a tightly controlled laboratory experiment, which is subject to certain limitations. These limitations, plus how future research can address them and build on the findings of this dissertation, are now discussed in detail.

Limitations and Future Research Directions

The method of objectively assessing cognitive fit via eye trackers developed in this study, as suggested by Dimoka et al. (2012), provides researchers with an unprecedented opportunity to better understand how cognitive fit affects technology users' task performance. In this dissertation, the cognitive fit that different visualizations provide for certain data analysis tasks was examined by assessing the efficiency with which users extracted information from a decision-aid tool that provided different visualizations of information. It would be beneficial for future researchers to study cognitive fit via eye trackers in other contexts, and with professional users of decision-aid tools. Doing so could improve our understanding of the role extensive experience and habits of the users, plus the technological characteristics of other decision aid-tools (e.g., recommendation tools, expert systems, aggregators, and collaboration tools) play in affecting cognitive fit and its task performance consequences. Such research could also lead to the design of technological decision-aid tools that better facilitate data analysis and decision-making in various contexts.

One of the limitations of this dissertation is that even though the defining characteristics of big data (i.e., volume and variety) were manipulated as low vs. high (i.e., as 1,000 vs. 300,000 distinct data points) in a tightly-controlled laboratory experiment, the participants were not performing the data analysis tasks using an actual big dataset that might have contained billions or trillions of records. The experiment was designed in this matter to ensure that the participants were performing the analysis tasks within the realm of interpretable visualizations and that they were not overwhelmed by the visualizations of such quantities of data points. This experimental

design allowed the study of how the defining characteristics of big data (i.e., volume and variety) influence the task performance consequences of cognitive fit without exposing the participants to uninterpretable and unmanageable visualizations. However, this experimental design also limits us from observing how an actual big dataset being visualized influences the impacts of cognitive fit on task performance. The findings and implications of this dissertation are thus limited to interpretable visualizations that contain up to hundreds of thousands of data points. It remains unexplored how or if visualizations can facilitate big data analytics when much larger volumes and varieties of information are visualized. Therefore, it would be beneficial for future researchers to replicate, confirm, and expand the findings of this dissertation in actual big data settings.

This dissertation also has several other limitations that suggest future research opportunities. First of all, this study is subject to the common limitations of experimental research. Although the sample size in this study was particularly large compared to similar eye tracker studies (e.g., the sample size in Cyr, Head, Larios, and Pan (2009) was 22, and the sample size in Djamasbi, Siegel, Skorinko, and Tullis (2011) was 30), the sample consisted of undergraduate students. However, participants' task familiarity, motivation, and visualization ability were controlled for to rule out the possible explanations that their lack of familiarity, incentive, or ability affected their task (i.e., data analysis) performance. Furthermore, all participants were thoroughly familiarized with the experimental data analysis procedures through an extensive training session before they started working on the actual data analysis tasks, as explained in the Procedures section. Therefore, participants' familiarity with the

experimental procedures was carefully established and deemed sufficient for this study, and the effects of sampling students were expected to be minimal (DeSanctis, 1988). Nevertheless, studying professional (big data) analysts and decision-makers, as previously suggested, provides an opportunity to understand the role long-term experience and habits play in affecting cognitive fit and task performance.

Second, the data analysis task types were strictly manipulated as Information Retrieval, Information Comparison, or Information Integration tasks in this study. This allowed experimental control and random assignment to the experimental treatments, plus enabled the investigation of the cognitive fit provided by different visualization types for specific data analysis tasks. However, this dissertation did not take into account the potentially different approaches participants could have taken to solve these data analysis tasks. For instance, it is possible to transform or decompose Information Integration tasks into multiple other tasks (e.g., by first calculating the ratio between two information dimensions and then treating the task at hand as an Information Retrieval task), which could potentially influence the efficiency with which the participants solved the data analysis tasks. Therefore, one future research direction would be to investigate participants' different problem-solving approaches and how visualizations can better support certain activities during the transformation or decomposition of the business data analysis tasks.

Furthermore, business data analysts are sometimes faced with relatively ambiguous tasks, such as data exploration or discovery (Lurie and Mason, 2007), that require only a basic understanding of the data be established and do not necessarily involve retrieving information or computing data values based on the visualizations,

such as for the tasks examined in this dissertation. I acknowledge that these tasks are also important components of big data analytics, and suggest that future research investigate the cognitive fit and decision-making support that different visualization types can provide for unclassified data analytics tasks, such as data exploration. Multiple visualizations, such as dashboards, might be appropriate for such exploratory tasks (Chandler, 2007).

Third, the visualization types in this study were strictly manipulated as Discrete vs. Continuous or Singular vs. Multiple. To maintain consistency across the stimulus material and experimental tasks, all visualizations were presented on a geographical map of the United States of America, and were as large as the eye-tracker monitor permitted (i.e., roughly 1400x900 pixels). This also allowed the experimental visualizations and data analysis tasks to be consistent with the majority of practical business data analytics tasks, which include a spatial or geographical information component such as the locations of customers or inventory (Crossland et al., 1995; Card and Mackinlay, 1997). As a consequence, there were a vast number of loan applications plotted over large metropolitan areas (e.g., New York City or San Francisco), especially when a high volume of information was visualized (e.g., see Figure 10 in Appendix A). This could potentially have confounded the results because the over-crowding of such areas could have made it even more difficult for the participants to extract information from the visualizations, beyond the effects of high volume alone. Therefore, this limitation of the specific visualizations used in this dissertation (see Appendix A) should be taken into account when the results are interpreted. Future research is

warranted to investigate different sizes and types of Discrete vs. Continuous and Singular vs. Multiple visualizations that do not rely on geographical maps.

Another major avenue for future research is the development of data analysis tools and methods based on the insights that this dissertation provides. The results of this study indicate that continuous visualizations are superior to discrete visualizations, and that singular visualizations outperform multiple presentations, in terms of providing decision-making support for certain types of data analysis tasks within the domain of interpretable visualizations. Considering that analysts are increasingly expected to solve a multitude of different types of complex data analysis problems (McAfee and Brynjolfsson, 2012), it could be fruitful to design visualization tools that enable the analysts to rapidly add or remove data dimensions and switch from one type of visualization to another as they work on different types of business data analysis tasks. It would also be beneficial for future researchers to study new visualization tools and techniques, and the cognitive fit they provide for different types of data analysis tasks, as improvements in technology allow us to visualize information in novel and more complicated ways.

One last avenue for future research is the investigation of the role that the third defining characteristic of big data (i.e., velocity) plays in the context of business data analytics. This dissertation focused on the volume and variety of information, the two main defining characteristics of big data (Eaton et al., 2012), because the majority of today's big data analytics is still performed on static datasets (i.e., snapshots of data) due to technological and practical limitations (VijayaBaskaran, 2013). Nevertheless, as analyzing high velocity big data in real-time becomes feasible, the ability to do so is

expected to be a key competitive asset and differentiator to organizations (Eaton et al., 2012). Future studies could examine how velocity impacts the task performance consequences of cognitive fit, and how to mitigate the challenges velocity presents for big data analytics through proper approach to visualization.

Conclusions

This dissertation examines how task performance is affected by the cognitive fit between different types of visualizations and data analysis tasks, and how these effects are amplified in the context of big data analytics. The results of this study provide important implications for researchers and practitioners, and contribute to the literature in at least three ways. First, this dissertation proposes an objective method to assess cognitive fit, which can be used in future research to further improve our understanding of cognitive fit and how it can be better facilitated by technology in various contexts. Second, the results of this study map visualization characteristics to business data analysis tasks, providing a better understanding of how visualizations can facilitate data analysis and guiding the choice of visualization types among an ever-increasing number of alternatives. Finally, this study extends cognitive fit theory to the big data context and highlights the relative importance of cognitive fit in this setting by demonstrating that the choice of visualization methods is especially consequential for high volume and large variety information settings.

In conclusion, this dissertation provides empirical evidence supporting the argument that the match between the information emphasized by a visualization tool and the type of information required by a data analysis task determines the tool's

usefulness for that task. This match, or cognitive fit, has greater consequences when a larger amount and/or more different kinds of information are visualized. The results of this study can inform visualization tool design and choice for a variety of data analysis tasks, benefiting researchers and practitioners alike who are interested in (big) data analytics.

References

Ackoff, R. L. (1967). Management misinformation systems. *Management Science* 14, pp. 147–156.

Allen, G., and Parsons, J. (2010). Is query reuse potentially harmful? Anchoring and adjustment in adapting existing database queries. *Information Systems Research*, *21*(1), pp. 56-77.

Amar, R., Eagan, J., and Stasko, J. (2005). Low-level components of analytic activity in information visualization. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (pp. 111-117). IEEE.

Bell, D. E., Raiffa, H., and Tversky, A. (1988). Descriptive, normative, and prescriptive interactions in decision making. *Decision making: Descriptive, normative, and prescriptive interactions*, *1*, 9-32.

Bertin, J. (1981). *Graphics and graphic information processing*. Walter de Gruyter.

Bertin, J. (1983) *The semiology of graphics*. University of Wisconsin Press, Madison, WI.

Card, S. K., and Mackinlay, J. (1997,). The structure of the information visualization design space. In *Information Visualization, 1997. Proceedings., IEEE Symposium on* (pp. 92-99). IEEE.

Card, S.K., Mackinlay, J.D., and Shneiderman, B. (1999). Readings in Information Visualization: Using Vision to Think., Morgan Kaufmann Publishers, San Francisco, CA.

Catts, T. (2012). GE's Billion-Dollar Bet on Big Data. Retrieved April, 2015, from http://www.bloomberg.com/bw/articles/2012-04-26/ges-billion-dollar-bet-on-big-data

Chan, S. Y. (2001). The use of graphs as decision aids in relation to information overload and managerial decision quality. Journal of Information Science 6, pp. 417–426.

Chandler, N. (2007). Q&A: Important Integration Considerations for Scorecards, Dashboards and Portals. Retrieved April, 2015, from https://www.gartner.com/doc/509017/qa-important-integration-considerations-scorecards

Chen, H., Chiang, R. H., and Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), pp. 1165-1188.

Chi, E. H. H., and Riedl, J. T. (1998). An operator interaction framework for visualization systems. In *Information Visualization, 1998. Proceedings. IEEE Symposium on* (pp. 63-70). IEEE.

Chi, E. H. H. (2000). A taxonomy of visualization techniques using the data state reference model. In *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on* (pp. 69-75). IEEE.

Columbus, L. (2015). 56% Of Enterprises Will Increase Their Investment In Big Data Over The Next Three Years. Retrieved April, 2015, from http://www.forbes.com/sites/louiscolumbus/2015/03/22/56-of-enterprises-will-increase-their-investment-in-big-data-over-the-next-three-years/

Crossland, M. D., Wynne, B. E., and Perkins, W. C. (1995). Spatial decision support systems: An overview of technology and a test of efficacy. *Decision Support Systems*, *14*(3), pp. 219-235.

Cyr, D., Head, M., Larios, H., and Pan, B. (2009). Exploring human images in website design: A multi-method approach. *Management Information Systems Quarterly, 33*(3), pp. 539-566.

Davenport, T. H., and Dyche, J. (2013). Big Data in Big Companies. *International Institue of Analytics,May 2013*.

Davern, M., Shaft, T., and Te'eni, D. (2012). Cognition matters: Enduring questions in cognitive IS research. *Journal of the Association for Information Systems*, *13*(4).

Dennis, A. R., and Carte, T. A. (1998) "Using Geographical Information Systems for Decision Making: Extending Cognitive Fit Theory to Map-based Presentations," *Information Systems Research* (9:2), pp. 194-203.

DeSanctis, G. (1988). Small group research in information systems: Theory and method. Paper presented at the Harvard Colloquium on Experimental Research in Information Systems, University of British Columbia.

Dimoka, A., Pavlou, P. A., and Davis, F. D. (2011). Research Commentary – NeuroIS: The Potential of Cognitive Neuroscience for Information Systems Research. *Information Systems Research*, *22*(4), 687-702.

Dimoka, A., Banker, R.D., Benbasat, I., Davis, F.D., Dennis, A.R., Gefen, D., Gupta, A., Ischebeck, A., Kenning, P., Pavlou, P.A., Müller-Putz, G., Riedl, R., vom Brocke, J., and Weber, B. (2012). On the Use of Neurophysiological Tools in IS Research: Developing a Research Agenda for NeuroIS. *Management Information Systems Quarterly, 36*(3), pp. 679-702.

Djamasbi, S., Siegel, M., Skorinko, J., & Tullis, T. (2011). Online viewing and aesthetic preferences of generation Y and the baby boom generation: Testing user web

site experience through eye tracking. *International Journal of Electronic Commerce, 15*(4), pp. 121-158.

Eaton, C., Deroos, D., Deutsch, T., Lapis, G., and Zikopoulos, P. (2012), Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data. New York: McGraw-Hill.

Eppler, M. J. and Mengis, J. (2004). The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines, *The Information Society, 20(5),* pp. 325–344.

Fischer, S., Lowe, R. K., and Schwan, S. (2008). Effects of presentation speed of a dynamic visualization on the understanding of a mechanical system. *Applied Cognitive Psychology*, *22*(8), pp. 1126-1141.

Frankel, F., and Reid, R. (2008). Big data: distilling meaning from data. *Nature, 455*(7209), p. 30.

Gao, J., Zhang, C., Wang, K., and Ba, S. (2012). Understanding online purchase decision making: The effects of unconscious thought, information quality, and information quantity. *Decision Support Systems*, *53*(4), pp. 772-781.

George, J. F., Duffy, K., and Ahuja, M. (2000). Countering the anchoring and adjustment bias with decision support systems. *Decision Support Systems, 29,* pp. 195-206.

Gershon, N., Eick, S. G., and Card, S. (1998). Information visualization. *Interactions*, *5*(2), pp. 9-15.

Goswami, S., Chan, H.C., and Kim, H.W. (2008). The role of visualization tools in spreadsheet error correction from a cognitive fit perspective. *Journal of the Association for Information Systems, 9*(6), pp. 321-343.

Green, M. (1998). Toward a perceptual science of multidimensional data visualization: Bertin and beyond. *ERGO/GERO Human Factors Science*, *8*.

Grolemund, G., and Wickham, H. (2015). Visualizing Complex Data with Embedded Plots. *Journal of Computational and Graphical Statistics, 24*(1), pp. 26-43.

Heer, J., Bostock, M., and Ogievetsky, V. (2010). A tour through the visualization zoo. *Commun. ACM*, *53*(6), pp. 59-67.

Hiltz, S.R. and Turoff, M (1985). Structuring computer-mediated communication systems to avoid information overload, Communications of the ACM, vol. 28, no. 7, pp. 680-689.

Iselin, E. R. (1988). The effects of information load and information diversity on decision quality in a structured decision task. Accounting, Organizations and Society 13, pp. 147–164.

Jarvenpaa, S. L., and Dickson, G. W. (1988). Graphics and managerial decision making: Research-based guidelines. *Communications of the ACM*, *31*(6), 764-774.

Kline, R. B. (2010). Principles and Practices of Structural Equation Modeling, New York: The Guilford Press.

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., and Kruschwitz, N. (2011). Big data, analytics and the path from insights to value. *MIT Sloan Management Review, 52*(2), pp. 21-32.

Lengler, R., and Eppler, M. J. (2007,). Towards a periodic table of visualization methods for management. In *IASTED Proceedings of the Conference on Graphics and Visualization in Engineering (GVE 2007), Clearwater, Florida, USA*.

Li, T., Feng, S., and Xia Li, L. (2001). Information visualization for intelligent decision support systems. *Knowledge-Based Systems*, *14*(5), pp. 259-262.

Lurie, N. H., and Mason, C. H. (2007). Visual Representation: Implications for Decision Making. *Journal of Marketing*, *71*(1), pp. 160-177.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute.

McAfee, A., and Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), pp. 60-66.

McNab, A. L., T. J. Hess, and J. S. Valacich. (2011). Designing Emergency Response Dispatch Systems for Better Dispatcher Performance, *AIS Transactions on Human-Computer Interaction* (3) 1, pp. 26-55.

Meharia, P. (2012). Use Of Visualization In Digital Financial Reporting: The Effect Of Sparkline. Theses and Dissertations--Business Administration. Paper 1. http://uknowledge.uky.edu/busadmin_etds/1

Pajarola, R. (1998). Large scale terrain visualization using the restricted quadtree triangulation. In *IEEE Visualization '98 Proceedings* (pp. 19-26).

Palaniappan, R. (2014). Data Visualization: Creating Mind's Eye. *Handbook of Research on Cloud Infrastructures for Big Data Analytics*, pp. 322-351.

Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 52*(3), pp. 381-410.

Pfitzner, D., Hobbs, V., and Powers, D. (2003). A unified taxonomic framework for information visualization. In *Proceedings of the Asia-Pacific symposium on Information visualisation-Volume 24* (pp. 57-66). Australian Computer Society, Inc..

Pirolli, P., and Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis* (Vol. 5, pp. 2-4).

Rinzivillo, S., Pedreschi, D., Nanni, M., Giannotti, F., Andrienko, N., and Andrienko, G. (2008). Visually driven analysis of movement data by progressive clustering. *Information Visualization*, *7*(3-4), pp. 225-239.

Roth, S. F., and Mattis, J. (1990). Data characterization for intelligent graphics presentation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 193-200). ACM.

Shaft, T. M., and Vessey, I. (1995). Research report—The relevance of application domain knowledge: The case of computer program comprehension. *Information Systems Research, 6*(3), pp. 286-299.

Shaft, T. M., and Vessey, I. (2006). The role of cognitive fit in the relationship between software comprehension and modification. *Management Information Systems Quarterly, 30*(1), pp. 29-55.

Shen, Wei-Cheng, Carswell, M.C., Santhanam, R., and Bailey, K. (2012). Emergency Management Information Systems: Could Decision Makers be supported in Choosing Display Formats?. *Decision Support Systems, 52*(2), pp. 318-330.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on* (pp. 336-343). IEEE.

Speier, C., Valacich, J. S., and Vessey, I. (1999). The influence of task interruption on individual decision making: An information overload perspective. Decision Sciences 30, pp. 337–359.

Tory, M., and Moller, T. (2004). Rethinking visualization: A high-level taxonomy. In Information Visualization, 2004. INFOVIS 2004. IEEE Symposium on (pp. 151-158). IEEE.

Tweedie, L. (1997). Characterizing interactive externalizations. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems* (pp. 375-382). ACM.

Tzabbar, D. (2009). When does Scientist Recruitment Affect Technological Repositioning? *The Academy of Management Journal, 52*(5), pp. 873-896.

Umanath, N. S., Scamell, R. W., and Das, S. R. (1990). An Examination of Two Screen/Report Design Variables in an Information Recall Context*. *Decision Sciences*, *21*(1), pp. 216-240.

Umanath, N. and Vessey, I. (1994). Multi-Attribute Data Presentation and Human Judgement: A Cognitive Fit Perspective, *Decision Sciences, 25*(5/6), pp. 795-824.

Vessey, I. (1991). Cognitive Fit: A Theory‐Based Analysis of the Graphs Versus Tables Literature. *Decision Sciences*, *22*(2), pp. 219-240.

Vessey, I., and Galletta, D. (1991). Cognitive fit: An empirical study of information acquisition. *Information Systems Research*, *2*(1), pp. 63-84.

VijayaBaskaran, R. (2013). An Analysis Of Emerging Trends In Big Data And Discretionary Opportunities For Indian Bpo Industry. *International Journal of Information Technology & Computer Sciences Perspectives*, *2*(2), pp. 441-451.

Ward, M. O. (2002). A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, *1*(3-4), pp. 194-210.

Ware, C., (2004) Information Visualization: Perception for Design. Morgan Kaufmann Publishers, San Francisco, CA.

Washburne, J. N. (1927). An experimental study of various graphic, tabular, and textual methods of presenting quantitative material. *Journal of Educational Psychology*, *18*(6), pp. 465-476.

Wehrend, S., and Lewis, C. (1990). A problem-oriented classification of visualization techniques. In *Proceedings of the 1st Conference on Visualization'90* (pp. 139-143). IEEE Computer Society Press.

Woods, D. D. (1991). The cognitive engineering of problem representations. In J. Alty and G. Weir (Eds.), *Human-computer interaction in complex systems* (pp.169-188). London: Academic.

Yau, N. (2013). Data Points: Visualization That Means Something. John Wiley & Sons.

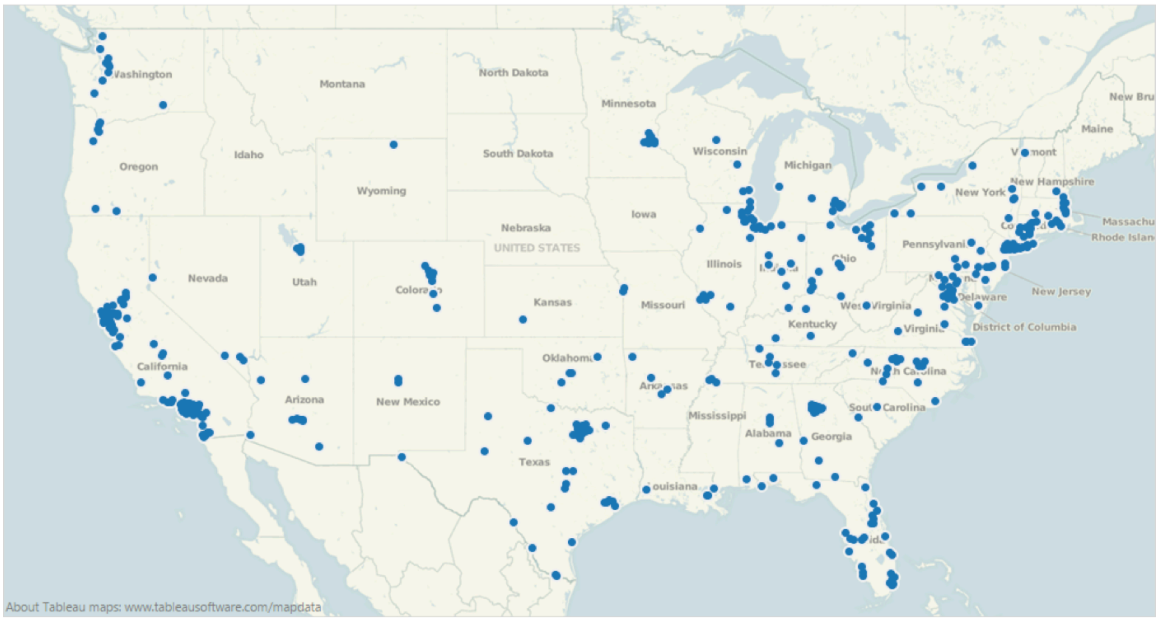
Yetgin, E., Jensen, M. L., and Shaft, T. M. (2015). Complacency and Intentionality in IT Use and Continuance. *AIS Transactions on Human-Computer Interaction, 7*(1), pp. 17-42.

Zhang, J. (1996). A representational analysis of relational information displays. International Journal of Human-Computer Studies, 45(1), pp. 59-74.

Zhou, M. X., and Feiner, S. K. (1998). Visual task characterization for automated visual discourse synthesis. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 392-399). ACM Press/Addison-Wesley Publishing Co.

Zhu, B., and Watts, S. A. (2010). Visualization of network concepts: The impact of working memory capacity differences. *Information Systems Research*, *21*(2), pp. 327-344.

Appendix A: Stimulus Materials



* Each circle denotes a loan issued

Figure 8. Treatment 1 (Discrete, Low Volume, Low Variety)

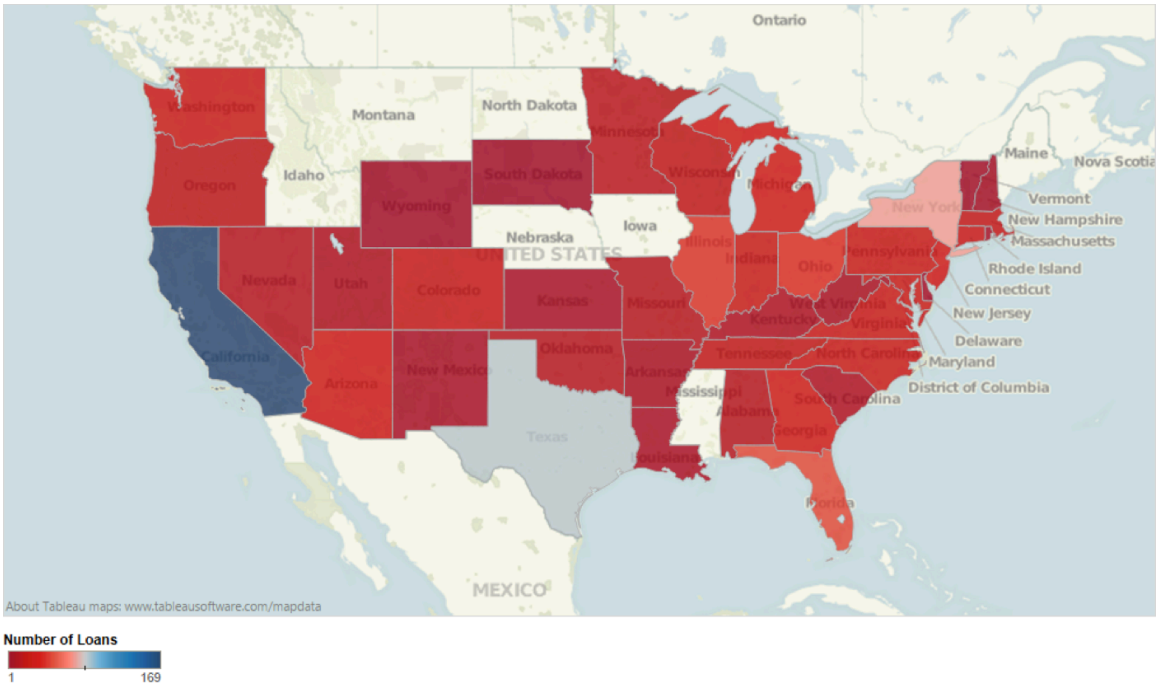
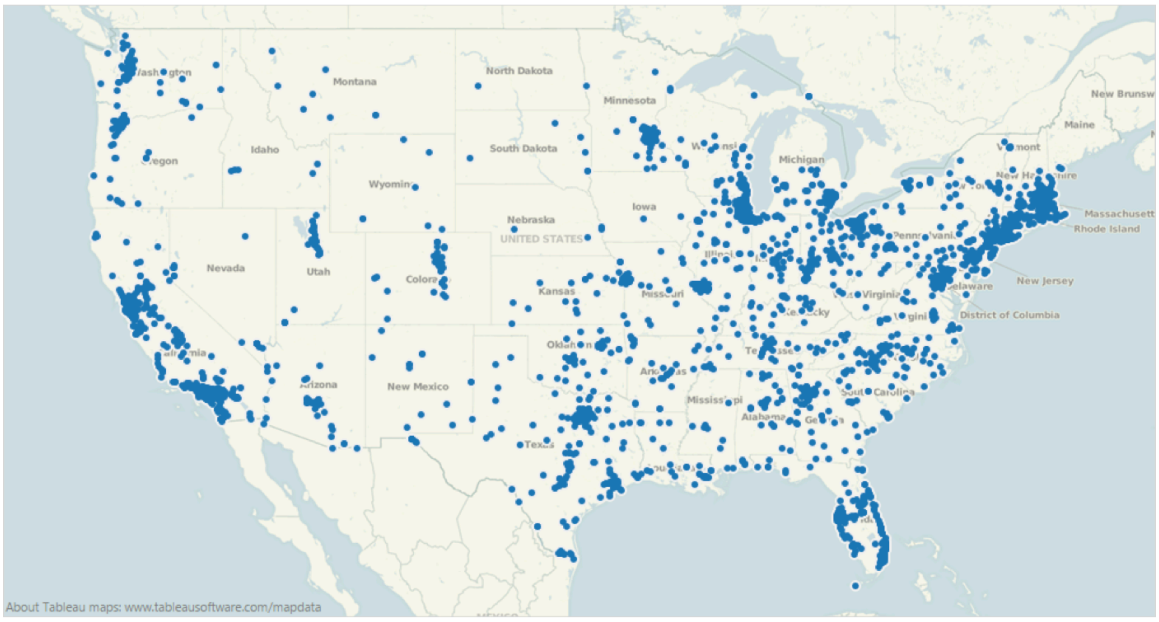
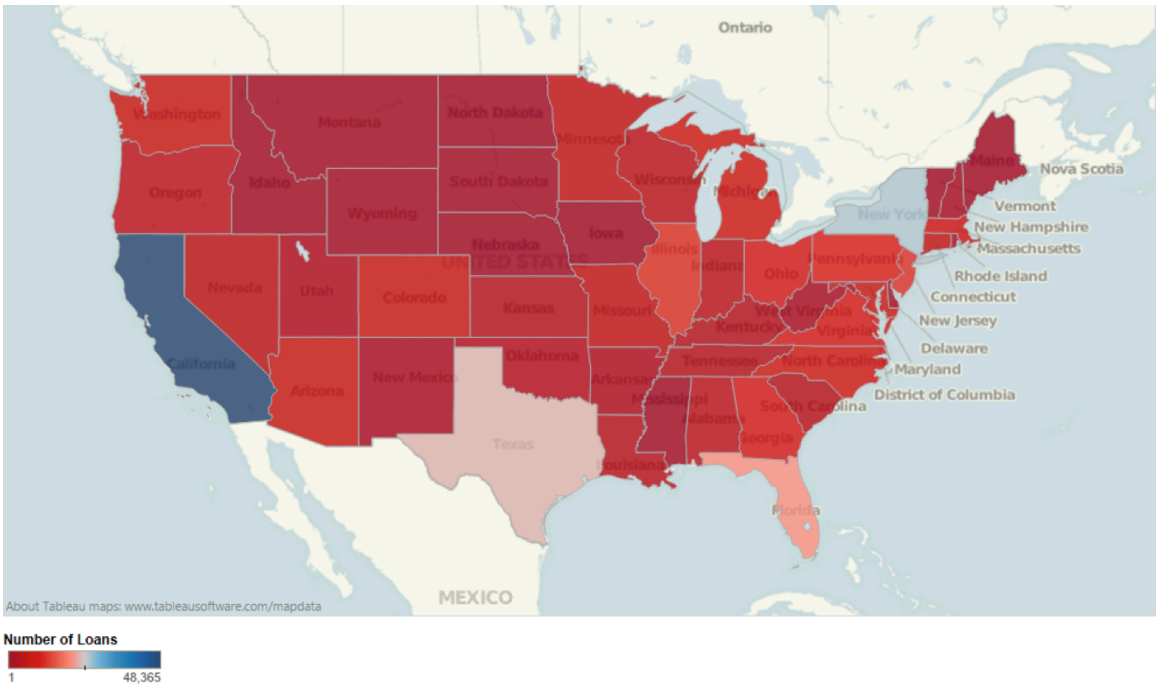


Figure 9. Treatment 2 (Continuous, Low Volume, Low Variety)



* Each circle denotes a loan issued

Figure 10. Treatment 3 (Discrete, High Volume, Low Variety)



Number of Loans

1 48,365

Figure 11. Treatment 4 (Continuous, High Volume, Low Variety)

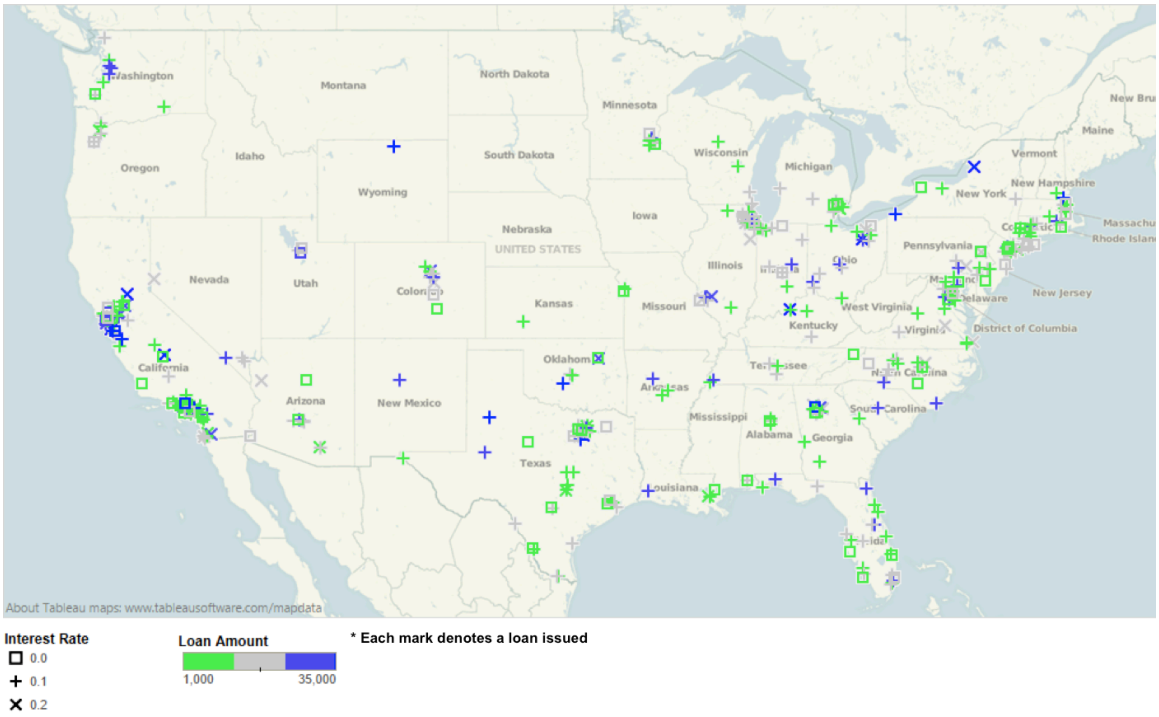


Figure 12. Treatment 5 (Discrete, Low Volume, High Variety)

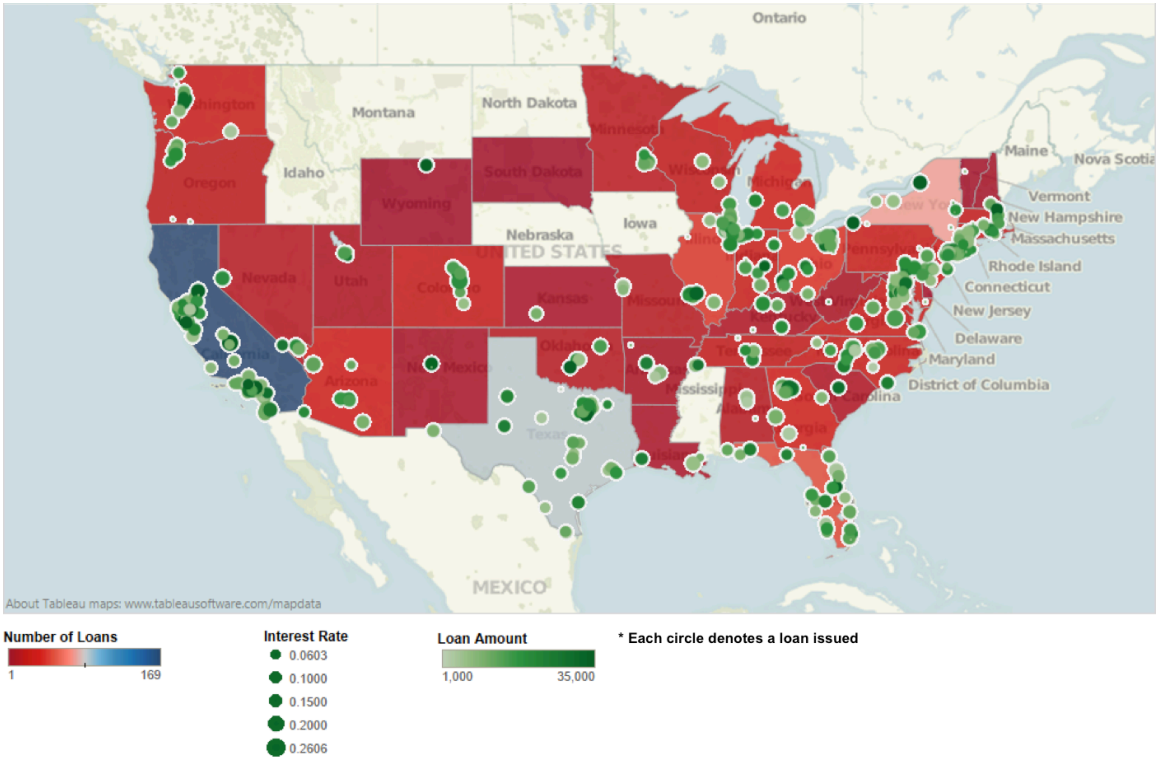


Figure 13. Treatment 6 (Continuous, Low Volume, High Variety)

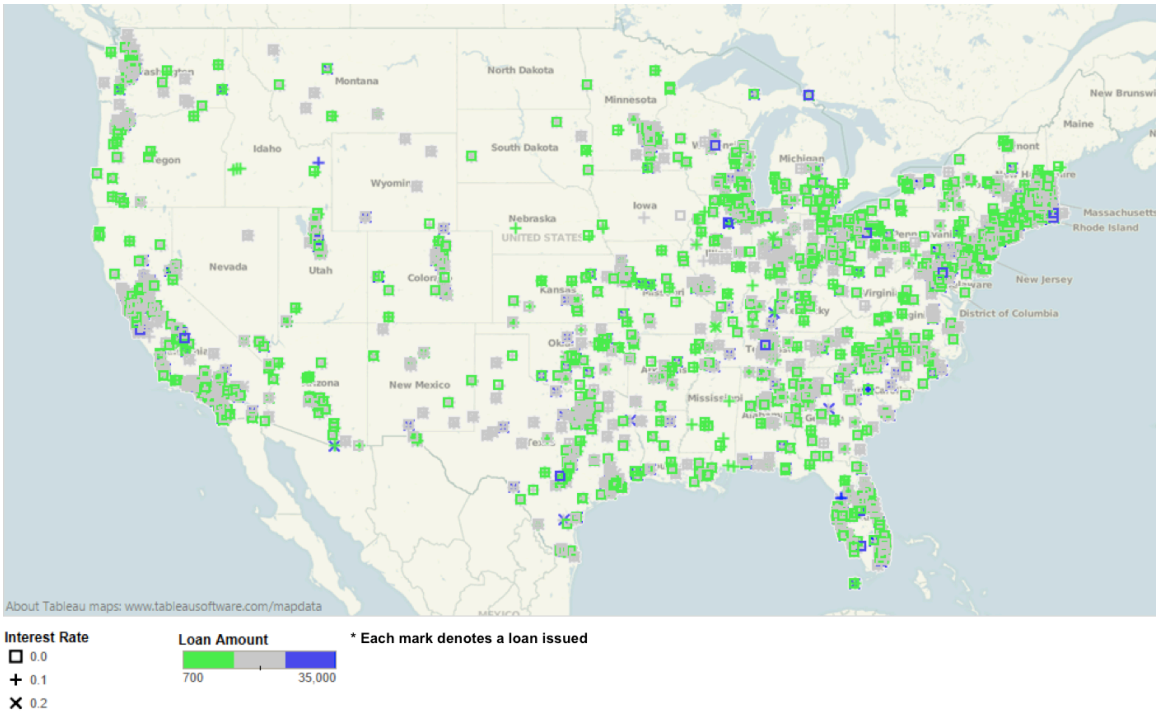


Figure 14. Treatment 7 (Discrete, High Volume, High Variety)

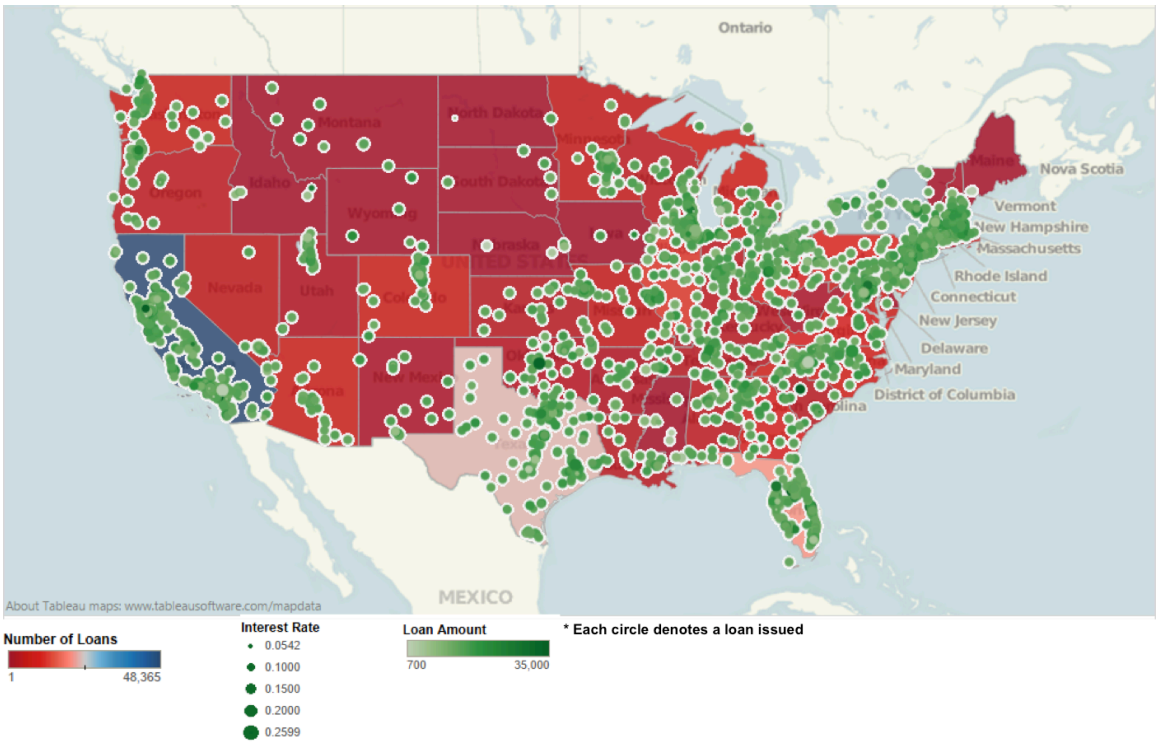


Figure 15. Treatment 8 (Continuous, High Volume, High Variety)

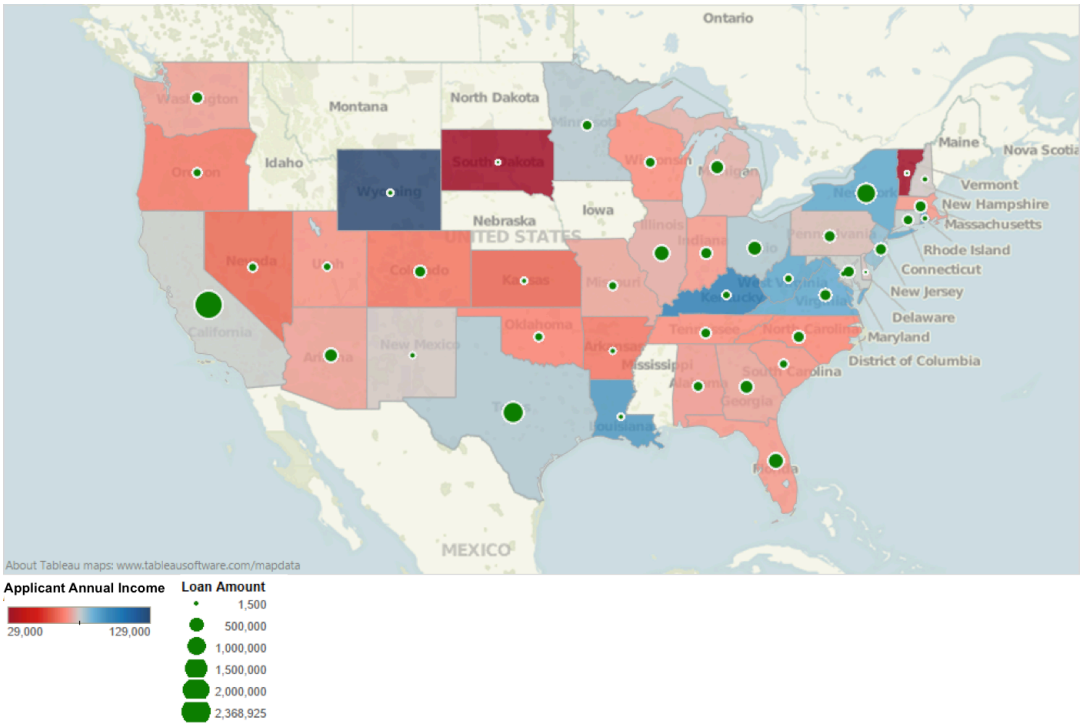


Figure 16. Treatment 9 (Singular, Low Variety)

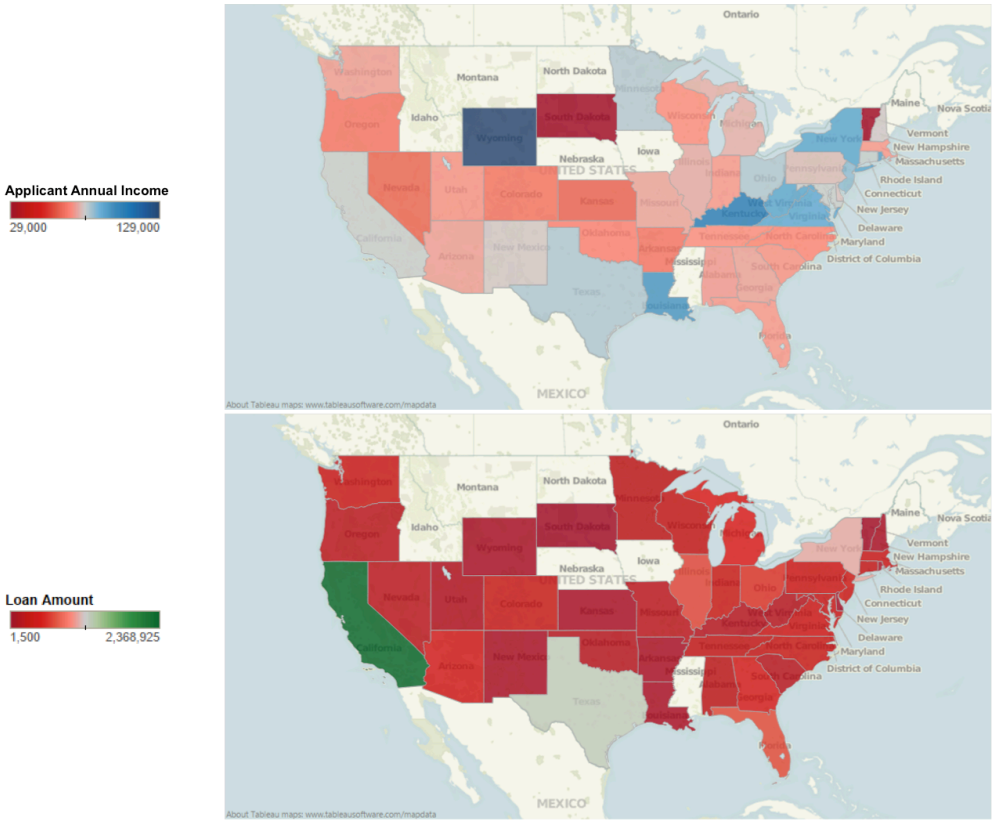


Figure 17. Treatment 10 (Multiple, Low Variety)

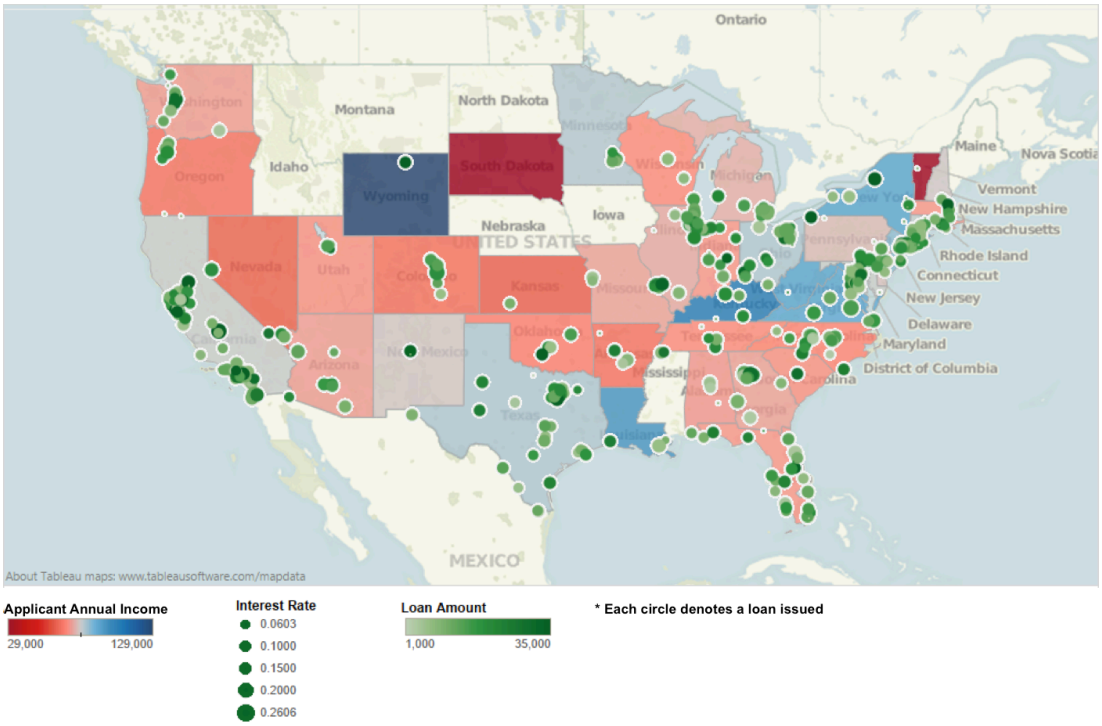


Figure 18. Treatment 11 (Singular, High Variety)

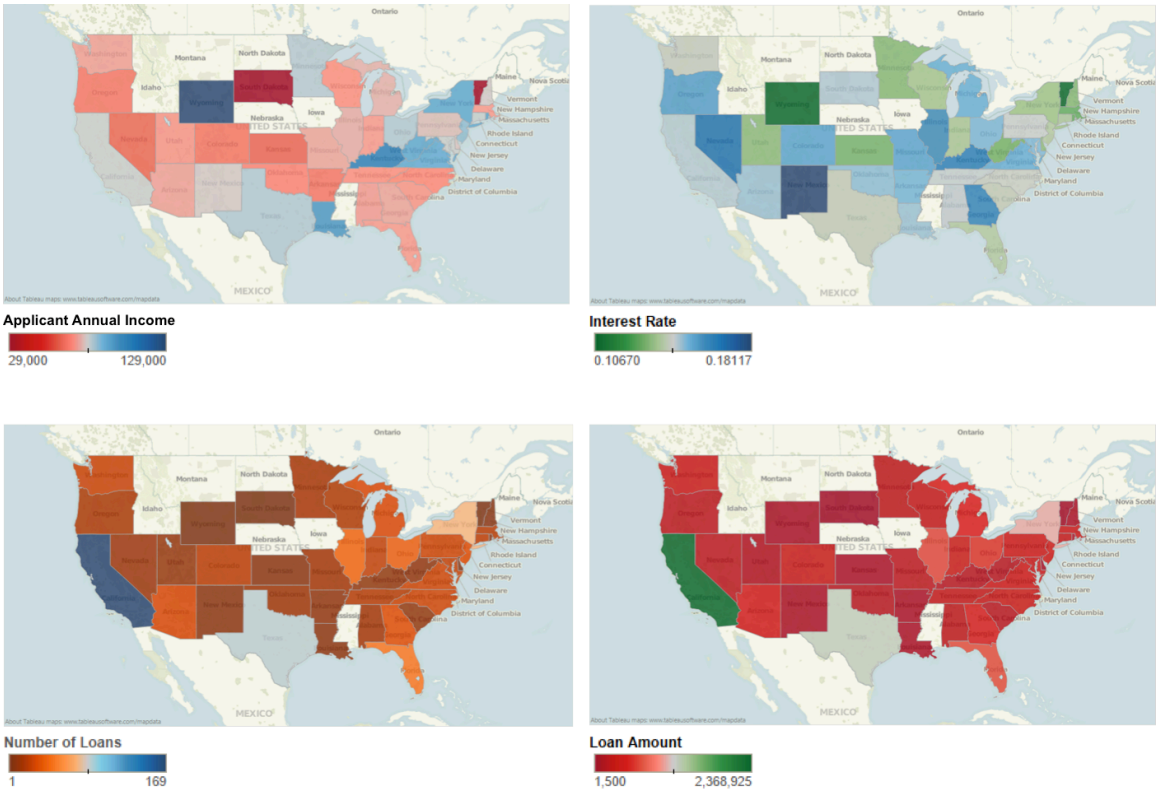


Figure 19. Treatment 12 (Multiple, High Variety)

Appendix B: Solution Accuracy Calculation

Participants were asked to solve four data analysis problems for each task type (see Task Type and Table 5). These tasks required different types of answers, such as numerical answers (for IR1, IR2, IR3, IR4, and IC4), the name of a state (IC1, IC2, II1, II2, II3, and II4), or a list of three states (IC3). To be able to assess the accuracy of participants' solutions for each task consistently and relative to one another, each solution was assigned a score out of 100, consistent with past Cognitive Fit research (e.g., Dennis and Carte, 1998; Shaft and Vessey, 2006). Completely correct solutions received a score of 100% and completely incorrect solutions received a score of 0%. Participants were given partial credit for partially correct answers, similar to the way task performance was assessed in previous Cognitive Fit studies (e.g., Shaft and Vessey, 2006). The amount of partial credit depended on how close the participants' solution was to the correct answer (e.g., how many of the three states they were able to guess correctly). The grading procedure for each data analysis task is explained in detail below.

Information Retrieval Tasks

Recall that all Information Retrieval tasks required a numerical answer (i.e., the number of loans or states). To calculate the accuracy of participants' solutions for these tasks, first the absolute differences between participants' answers (PA) and the correct answer (CA) for each Information Retrieval task were calculated. Then, error percentages were calculated by dividing these absolute differences by the correct answers. Finally, participants were assigned a score for each task by subtracting the

respective error percentages from 100%. If a participant's answer was more than 100% off from the correct answer, which would result in a negative score, they were assigned a score of 0%. In other words, if a participants' solution was off from the correct answer by a magnitude of the correct answer in either direction, their solution was considered completely inaccurate and they received no partial credit.

Below is the formula that was used to calculate Information Retrieval scores for each one of the four tasks (n):

IR(n)score = 100% - (| PA - CA | / CA)

Average Information Retrieval task accuracy for each participant was calculated by using the following formula:

Information Retrieval Solution Accuracy = (IR1score + IR2score + IR3score + IR4score) / 4

Information Comparison Tasks

Task IC1 required the participants to name the state in which the most number of loans were issued. For this task, states were first ranked in descending order by their number of loans issued. Participants' answers were then assigned a rank based on this list, with the correct answer having the first rank. This rank was then converted into a percentage score so that the top rank would be assigned a score of 100% and the lowest rank would be assigned 0%. For example, if a participants' solution ranked third on the list of 41 states that were displayed, they were assigned a score of 95%, as each rank after the first state on the list suffered a 2.5% penalty (100/40) with the 41st rank receiving a score of 0. If the state a participant named was not ranked on the list (i.e.,

was not included in the visualization), they also received a score of 0%. Task IC2 was graded by following the same procedure, except the states were ranked in ascending order this time, because the task required the participants to name the state with the least number of loans issued.

Task IC3 required the participants to list the top three states with the most number of loans issued. For this task, participants were assigned a score out of three, based on how many of the top three states they were able to correctly guess. These scores were then converted into percentage scores so that 3/3 correct states would be assigned a score of 100%, 2/3 would be assigned a score of 66.67%, 1/3 would be assigned a score of 33.33%, and 0/3 would be assigned a score of 0%.

Because it required a numerical answer (i.e., the number of loans), the scores for task IC4 were calculated by following the same procedure for grading Information Retrieval tasks (i.e., by subtracting absolute error percentages from 100%).

Average Information Comparison task accuracy for each participant was calculated by using the following formula:

Information Comparison Solution Accuracy = (IC1score + IC2score + IC3score + IC4score) / 4

Information Integration Tasks

For task II1, participants were asked to name the state with the highest loan amount to applicant annual income ratio on average. Similar to the procedure for scoring task IC1, states were first ranked in descending order by their loan amount to applicant annual income ratio. Participants' answers were then assigned a rank based on

this list, with the correct answer having the first rank. This rank was then converted to a percentage score so that the top rank would be assigned a score of 100% and the lowest rank would be assigned 0%, as with task IC1. Task II2 was graded by following the same procedure, except the states were ranked in descending order this time, because the task required the participants to name the state with the lowest loan amount to applicant annual income ratio.

Tasks II3 (and II4) required the participants to name the state with the lowest (or highest) loan amount issued among the three states with the lowest (or highest) average applicant annual income. For these tasks, participants were assigned a score out of three, based on the rank of their answer among the three states with the lowest (or highest) average applicant annual income. If a participant's answer was not among these three states, they were assigned a score of 0%. These scores were then converted into percentage scores, similar to the procedure for IC3, so that the first rank would be assigned a score of 100%. For example, if a participants' answer ranked third (i.e., 3/3) among the three states, they received a score of 33.33%, whereas they would have received a score of 0% if their answer was not among the three states with the lowest (or highest) average applicant annual income.

Average Information Integration task accuracy for each participant was calculated by using the following formula:

Information Integration Solution Accuracy = (II1score + II2score + II3score + II4score) / 4

Appendix C: Visualization Ability Measure

Participants were asked to answer the following question (adapted from Shen et al., 2012) to obtain a measure of their visualization abilities, which was controlled for to rule out alternative explanations regarding task performance.

Please mentally rotate the objects below and answer the question:

Does the figure on the right show an accurate rotation of the figure on the left?

Choose Yes or No.

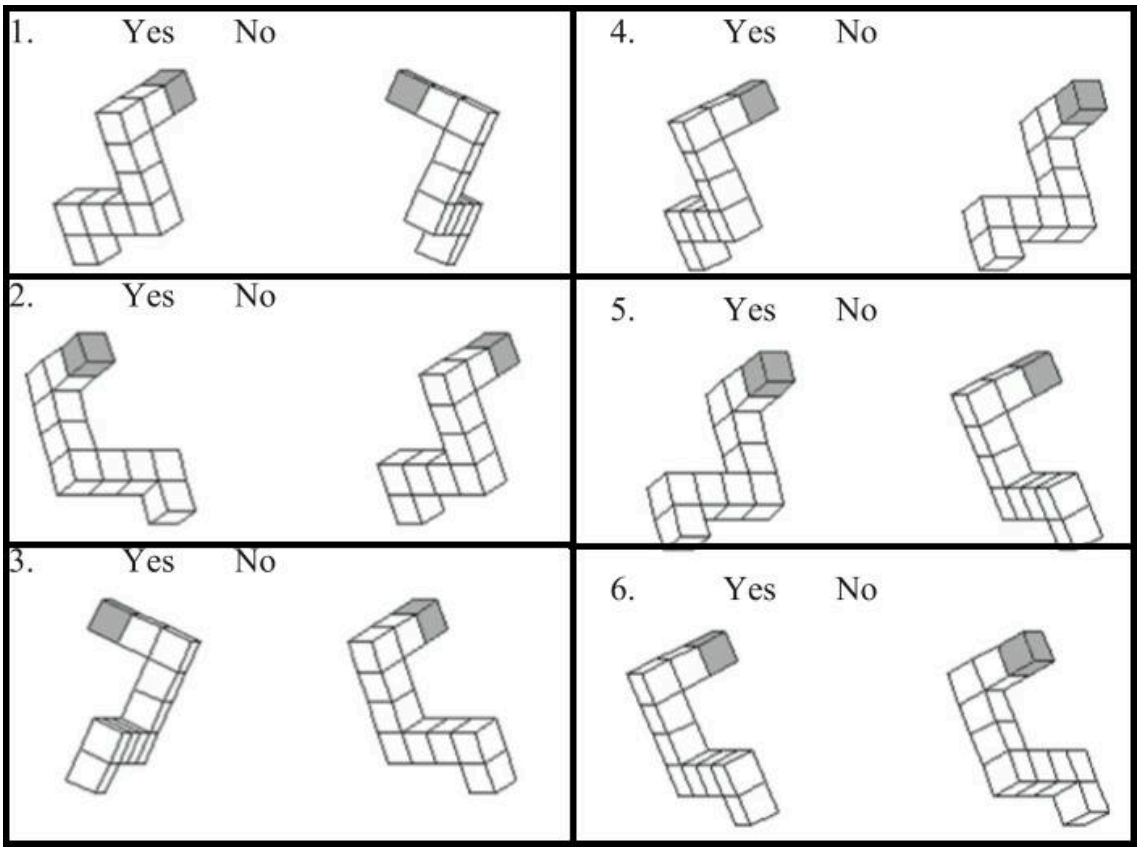


Figure 20. Image Pairs Used for Measuring Visualization Ability