LET THE ROBOT DO IT FOR ME: ASSESSING VOICE AS A MODALITY FOR VISUAL ANALYTICS FOR NOVICE USERS

By

JONATHAN LUIS PULLIZA

A Dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Communication, Information and Media

Written under the direction of

Dr. Chirag Shah

And approved by

New Brunswick, New Jersey

October, 2020

ABSTRACT OF THE DISSERTATION

LET THE ROBOT DO IT FOR ME: ASSESSING VOICE AS A MODALITY FOR VISUAL ANALYTICS FOR NOVICE USERS

by JONATHAN LUIS PULLIZA

Dissertation Director: Dr. Chirag Shah

The growth of Visual Analytics (VA) systems has been driven by the need to explore and understand large datasets across many domains. Applications such as Tableau were developed with the goal of better supporting novice users to generate data visualizations and complete their tasks. However, novice users still face many challenges in using VA systems, especially in complex tasks outside of simple trend identification, such as exploratory tasks. Many of the issues stem from the novice users' inability to reconcile their questions or representations of the data with the visualizations presented using the interactions provided by the system.

With the improvement in natural language processing technology and the increased prevalence of voice interfaces, there is a renewed interest in developing voice interactions for VA systems. The goal is to enable users to ask questions directly to the system or to indicate specific actions using natural language, which may better facilitate access to functions available in the VA system. Previous approaches have tended to build systems in a screen-only environment in order to encourage interaction through voice. Though they did produce significant results and guidance for the technical challenges of voice in VA, it is important to understand how the use of a voice system would affect novice users within their most common context instead of moving them into new environments. It is also important to understand when a novice user would choose to use a voice modality when the traditional keyboard and mouse modality is also available.

This study is an attempt to understand the circumstances under which novice users of a VA system would choose to interact with using their voice in a traditional desktop environment, and whether the voice system better facilitates access to available functionalities. Given the users choose the voice system, do they choose different functions than those with only a keyboard and a mouse? Using a Wizard of Oz set up in the place of an automated voice system, we find that the participants chose to use the voice system because of its convenience, ability to get a quick start on their work, and in some situations where they could not find a specific function in the interface. Overall function choices were not found to be significantly different between those who had access to the voice system versus those who did not, though there were a few cases where participants were able to access less common functions compared to a control group. Participants refrained from choosing voice because their previous experiences with voice systems had led them to believe all voice systems were not capable of addressing their task needs. They also felt using the voice system was incongruent with gaining mastery of the underlying VA system, as the convenience of using the voice system could lead to its use as a crutch. Participants then often chose to struggle with the visual interface instead of using the voice system for assistance.

ACKNOWLEDGMENTS

My dissertation would not be possible without the many smart folks who have been gracious enough to offer their time and support. Below is a small subset, but there are many others who have contributed to my work. I am notoriously overly brief in my writing, and any people omitted will hopefully be forgiving.

First, my dissertation advisor Dr. Chirag Shah. He reached out to me early on to join the InfoSeeking Lab, which made a huge difference in the path of my research. I am naturally a bit rough around the edges and more focused on building and breaking things, but he has helped me refocus my ideas and think of the bigger picture. This work began in January 2019 when he somehow incepted the idea of multimodality into my mind. His guidance and forceful encouragement have been appreciated. The other members of our lab have been a source of friendly and intelligent support, from overcoming technical hurdles to general commiserating. Thank you to all of our members, especially: Dr. Ruoyuan Gao, Dr. Souvick Ghosh, Dr. Jiqun Liu, Soumik Mandal, Dr. Matthew Mitsui, Dr. Manasa Rath, Shawon Sarkar, Shannon Tomlinson, and Dr. Yiwei Wang.

My dissertation committee members all deserve recognition for their support of this work, especially during such a challenging time. Dr. Nina Wacholder has been listening to my ideas from the start of my studies and has been very generous with her time and expertise. She told me to make sure whenever I ask a research question to make sure that the findings will be interesting even if the answer is negative. That has come into play a lot since then, but especially in this particular work. Dr. Mark Aakhus was quick to offer a challenging question that would catch me off guard but always made me think deeper about the work of research and the role of the researcher. His insightfulness is only matched by his love of passionate discourse, so having him on my committee has always been fun. Dr. Melanie Tory served as my outside committee member, though her work in Visual Analytics would have been critical to this study regardless.

The advice offered was much needed and appreciated, especially in thinking about all of the design concerns and tradeoffs. Dr. Nicholas Belkin served on my qualifying exam committee, and discussions and arguments I have had with him still resonate with me on a weekly basis.

The School of Communication and Information at Rutgers University has been my other home for the last few years. It always felt like a small community of folks who shared much more common ground then our eventual studies would indicate. The school funded my studies and this particular experiment, so this is a product of its design and generosity. Before this program, I was an MSIS student at the University of North Carolina at Chapel Hill. I was a member of the first Visual Analytics course in the program that was designed and taught by Dr. David Gotz, so I had a great start. Special thanks to Dr. Robert Capra, Dr. Jaime Arguello, Dr. Arcot Rajasekar, and Dr. Stephanie W. Haas, who were all helpful in exposing me to the Information Science field.

Finally, I would like to thank my family. My father who valued hard work and critical thinking, my mother who was unconditionally caring, and my sister who was always good for a laugh. My wife is really the only reason I pushed myself to take on a Ph.D. and commit this much of my life to this work. She has somehow managed to become a better person every day I have known her, and I will do my best to earn the love and support she offers me.

DEDICATION

I dedicate this dissertation to my wife, who has always encouraged me to be better, and my son, whose birth pushed me to finish.

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Background

Visual Analytics (VA) focuses on how to support users in cognitively challenging data tasks by providing interactions with a visual interface and improved data mining techniques. The growth of Visual Analytics (VA) systems has been driven by the need to explore and understand large datasets across many domains and as "the ability to collect data far outstrips the ability to analyze the collected data" (Thomas & Cook, 2006). There is an expectation that domain experts not formally trained in constructing visualizations can use these systems to complete their tasks. Applications such as Tableau, Qlik Sense, and Many Eyes (Viegas et al., 2007) were developed with the goal of better supporting these novice users to generate data visualizations and complete their tasks. However, novice users still face many challenges in using VA systems, especially in complex tasks outside of simple trend identification such as investigative or exploratory tasks (Kang, Gorg, & Stasko, 2009). These challenges have been investigated and categorized in studies such as Grammel et al., (2010) and Kwon et al. (2011), with many of the issues stemming from the novice users' inability to reconcile their questions or representations of the data with the visualizations presented using the interactions provided by the system. This gap is equivalent to the "gulf of execution," as described by Norman (2013), where users have to discover how to interact with a system and what the system is capable of doing.

A possible contributing factor to these challenges is the use of WIMP (Windows, Icons, Menus, Pointer) interfaces for these systems, where direct manipulation of objects is necessary through keyboard and mouse interactions with menus and buttons. The HCI community has identified many weaknesses within WIMP, including a tendency towards deep hierarchical menus that hide functionality and overwhelming amounts of buttons and widgets. Lee et al., (2012) highlights these weaknesses as they pertain to VA system designers, as the addition of visual objects such as buttons or widgets may overwhelm users within a limited visual environment. As designers increase the number of functions available to users to address the complexity of data problems, they expect users to develop further their mastery of the system to interact optimally. As summarized by Sun et al. (2014), these modern tools offer a broad array of capabilities made to be wielded by experts, while for novices it is akin to "expecting someone to know how to build a house by simply sending them to Home Depot" (Sun et al., 2014).

There is increasing pressure within the visualization community to look outside of the traditional WIMP interface paradigm and to create new ways for users to interact with VA systems. With the improvement in natural language processing technology and the increased prevalence of voice interfaces, there is a renewed interest in developing voice interactions for VA systems. A report out of the 2008 NSF workshop "Enabling Science Discoveries through Visual Exploration" stated that "there is a strong desire for conversational interfaces that facilitate a more natural means of interacting with science" (Ebert et al., 2008). The goal is to enable users to ask questions directly to the system or to indicate specific actions using natural language. The system is then more responsible for mapping the intentions of the user to the appropriate system actions. Users would then be able to focus more on their tasks instead of how to navigate the graphical interface. The development of voice as a modality for VA is technically challenging, with systems like Articulate (Aurisano et al., 2016) and Orko (Srinivasan & Stasko, 2018) highlighting the technical impediments and possibilities of building this type of system in a screen-only environment. Though they did produce significant results and guidance for the technical challenges of voice in VA, it is important to understand how the use of a voice modality would affect novice users within their most common context instead of moving them into new environments. It is also important to understand when a novice user would choose to use a voice modality when the traditional keyboard and mouse modality is also available.

This study is an attempt to answer the question under what circumstances would a novice user of a VA system choose to interact with their voice in a traditional desktop environment. Voice has been developed for VA systems as a possible solution to interaction issues for novice users and to address the “gulf of execution” problem that limits their ability to utilize proper functions. Given the users choose the voice system, do they choose different functions than those with only a keyboard and a mouse? Interactions are compared in regards to how and why users chose a particular mode of interaction as they go through exploratory tasks. The remainder of the dissertation is organized as follows: The next section of Chapter 1 provides a statement of the problem of interest. Chapter 2 reviews the literature concerning novice users of VA systems, WIMP and alternative interfaces, and voice as an input modality. The model is drawn from the literature on how a novice user would utilize a voice system for VA tasks is discussed in Chapter 3, along with the research questions of the study. The methodology is covered in Chapter 4, with the results presented in Chapter 5. The conclusions of this study are discussed in Chapter 6, and the limitations of this study as they affect those conclusions are discussed in Chapter 7. The final section Chapter 8 discusses the implications of this study further and possible future work.

1.2 Problem Statement

The purpose of this dissertation is to understand the outcome of providing an additional voice input modality to novice VA users, given that users also have the choice of using their keyboard or mouse. Providing a voice system may be an opportunity to alleviate some of the interaction issues faced by novice users and to better facilitate access to functions available in the VA system. This dissertation is about the choices made in a multimodal environment, where the strategy by which users may choose to utilize a particular modality in a particular situation is of interest. The strategies used and the difficulties encountered can serve as a roadmap for future developers of VA systems on the circumstances in which users are more likely to utilize a voice modality and how the system can offer better support.

Although there has been research into the use of VA systems by novice users and using a voice interface for visualizations, there is still an opportunity to examine the effect of adding voice to a traditional desktop VA system on novice users. Some previous studies limit users to only prefabricated visualizations, limiting the scope of their interactions and reducing the amount of functionality they would choose to access. Though effective in inducing voice commands from users, previous research conducted within a voice and screen only environment does not reflect how users would interact when in their most common workspaces with a keyboard and mouse. This experimental set up also ignores the possibility that novice users may employ a strategy in choosing the voice input modality over another input when both are present. This study will incorporate a voice system that focuses on providing the system functionalities available to the user that would normally be accessed through keyboard or mouse actions. Voice is the focus of this study mainly due to its current ubiquity in HCI design, as more systems are adding voice modalities with varying levels of intelligence that process and respond to user requests. The goals of this study are then to find 1) when do novice users choose to use a voice modality to interact with a VA system? and 2) does adding a voice system change the functions novice users access when interacting with a VA system?

1.3 Summary

Developers of VA systems have increased the number of functionalities available to address more complex tasks, but this may come at the cost of further isolating users who have never used the system before. The use of WIMP interface designs further exacerbates this problem, as functionality is either added to the limited visual space or hidden in hierarchical menus. The use of voice as an input modality is seen as a possible remedy for these issues, although the behavior of novice users when given the choice of using their voice or using a traditional keyboard and mouse modality has not been studied. This dissertation aims to understand when novice users choose to use voice over traditional keyboard and mouse inputs, as well as what effect this has on the functionalities chosen by the users during interactions with the VA system.

Chapter 2

Literature Review

This chapter attempts to review 1) past methods and findings in studying the interactions between novice users and VA systems, 2) relevant problems with traditional WIMP interfaces, 3) Natural User Interfaces and Multimodality, and 4) Voice interfaces for VA systems.

2.1 VA Systems

The following section defines the field of Visual Analytics (VA) in terms of its value to users by providing a means of interacting with external representations of mental models, which is modeled as a sensemaking process. This section then focuses on novice users of VA systems and their particular issues found in previous studies pertaining to their mental models of systems and interactions.

2.1.1 Defining VA Systems

Visual Analytics (VA) focuses on how to support users in cognitively challenging tasks through interactions with a visual interface and improved data mining techniques. It utilizes visual metaphors and interaction to leverage data mining technology while helping users to build the appropriate cognitive structures for their goals. VA depends on the value of external cognition, where the process of creating and interacting with external representations of concepts reduces cognitive load and improves overall task performance (Scaife & Rogers, 1996). The core problem of the field is how to organize and draw conclusions from large and heterogeneous data sets that are difficult for either a human or machine alone to handle. As summarized by Endert et al.

(2014), “[t]he rationale is that data is too large for purely visual methods, requiring the use of data processing and mining; yet, the desired tasks are too exploratory for purely analytical methods” (Endert et al., 2014). Thus, human cognitive abilities are supposed to work together with data mining techniques, with visualization and interaction as the medium in-between the parties.

2.1.2 External Cognition

VA depends on the value of external cognition, where the process of creating and interacting with external representations of concepts reduces cognitive load and improves overall task performance. Much of this modeling is an extension of the work of Larkin and Simon (1987). They examined how people solve problems by building internal representations of information and then create and use external artifacts. Scaife and Rogers (1996) summarize the main arguments for external cognition into three categories: computational offloading, re-representation, and graphical constraining. Computation offloading is the difference in the amount of effort required to solve a problem using an external representation versus solely in the user’s mind. A typical example is geometry, where Larkin and Simon (1987) find that users are much more adept at solving geometric problems using a diagram versus solely in their minds. Re-representation is the ability to change the external representation of an abstract structure in order to make the problem easier to solve, such as multiplying numbers using decimal representations versus roman numerals. Graphical constraining refers to how a graphical representation inherently limits the inferences that can be made about the underlying abstract subject. Attributes such as relations or comparative size that are reflective of the real-world item then inform users on what is possible as they work through a problem. Constraining the possible inferences helps facilitate more efficient hypothesis testing. Each of these benefits does not replace the cognitive work necessary on the part of users but supplement and support them as they move through the process.

2.1.3 Sensemaking for VA

The goal of VA systems is to support the cognitive work of a user in conjunction with interactive visualizations, so modeling the user's cognitive process is key to understanding user behavior. The predominant model for this process is sensemaking, which underlies much of the research in VA and serves as a theory for modeling the cognitive process that occurs during user interactions. Endert et al. (2012) go as far as to state that the purpose of the field of Visual Analytics is to support "sensemaking of large, complex datasets through interactively exploring visualizations generated by statistical models" (Endert et al., 2012). Sensemaking has been defined and framed by many researchers, with the most common theme in VA research being the building and use of some mental representation of a concept using information. Models of sensemaking in the VA community stem from the work of Russell et al. (1993), who define the goal of sensemaking as the development of a schema that the user can use to better structure and understand information to perform a specific information task at a lower cognitive cost. The process is then "searching for a representation and encoding data in that representation to answer task-specific questions" (Russell et al., 1993). Their model was later extended by Pirolli and Card (2005), who define VA as an iterative process that involves "information gathering, data preprocessing, knowledge representation, interaction and decision making" (Pirolli & Card, 2005) and model the process as an interlocking loop between sensemaking and information foraging. The operations required by the task are then the driver of what models the user must build in order to be successful. Andrews and North (2012) understand sensemaking to be a cognitively demanding but necessary task that involves "searching out information, breaking it down, and then piecing it back together into an understandable, compelling whole" (Andrews & North, 2012). The task at hand does not need to be well structured, since sensemaking is a process of "incremental formalism," where "structure evolve with the growing knowledge and understanding of the information that is being structured" (Andrews & North, 2012).

Assuming that users are going through a sensemaking process with VA systems, the challenge is then to measure how well a VA system supports this process as users move through their work and towards their goals. De Liddo et al. (2012) posit that sensemaking takes tends to take the form of hypothesis testing, where analysts draw connections within the data to develop a visual narrative that can become an artifact to be communicated to others. Therefore, when observing users, it is expected to see them interact with a system in an iterative manner, as they test hypotheses and build their understanding of the data. Each interaction is then an opportunity to assess the quality of support provided by the VA system in helping users achieve their task goals.

2.1.4 Interaction in VA

A key feature of VA is the emphasis on interaction, as it is the medium by which the synthesis of human cognition and algorithmic computing is possible. Though visualization as a means for external cognition is a powerful tool in developing insight, it is often insufficient in supporting users with large amounts of data. As Pike et al. (2009) state, "[a]s visual analytics is concerned with the relationship between visual displays and human cognition, merely developing novel visual metaphors is rarely sufficient" (Pike et al., 2009). They define multiple levels of interaction, including at the most basic level the use of controls that a user has available to manipulate an interface, as well as emerging fields of study in interaction such as ubiquitous computing and capturing user intentionality in real-time. For VA system design, a common heuristic to evaluate the basic functionalities of a VA system is the Visual Information Seeking Mantra by Shneiderman (2003), which is "overview first, zoom and filter, then details-on-demand." While this emphasizes the visual interface and interaction in exploring data, Keim et al., (2010) expand this for VA to capture more of the analytical process to "Analyze First - Show the Important - Zoom, Filter, and Analyze Further - Details on Demand."

There are many different models of interaction that operate at different levels, as discussed by Pike et al. (2009), with some going beyond just being able to control the behavior of the system. A model for a higher level of interaction between users and VA systems is the reconciliation between the users' views of the data versus the current visual encoding. Ziemkiewicz and Kosara (2008) view interactions with visualization tools as the process by which a user tries to match their internal representation of information to the presented visual representation. They primed users to be more associated with either a tree metaphor or a container metaphor before interpreting different visualizations and found the mismatch to be a major source of frustration and performance errors. Liu and Stasko (2010), in their work on mental models for visual reasoning, concur that the matching of users' internal versus the presented external representation is the primary driver of interaction with systems. They categorize the intentions of visualization interactions as external anchoring, information foraging, and cognitive offloading. The main set of challenges are caused by the users' inability to reconcile two representations given set of interactions that they can access. This coincides with the findings of Dimara and Perin (2020), who find in their large review of the literature that interaction in the field of visualization is defined as the dialogue between the user and the data.

Individual interactions in VA have been evaluated similarly to static visualizations within the cognitive efficiency view, where the goal is to minimize the cognitive cost while still facilitating user goals, such as the interaction cost framework developed by Lam (2008). The framework includes seven categories of costs that affect information visualization use, including decision costs and visual-cluttering costs, which must be addressed when designing an effective VA system. This model assumes that users select and evaluate interactions for the sole purpose of task completion, so learning how to interact with the system is seen as a cost and is only valuable towards this goal. Combining the interaction cost model with the view on mental model matching, interactions that better facilitate the reconciliation between the user's representation and the visual representation presented of the data and are less cognitively costly are therefore of higher quality.

2.1.5 Novice Users of VA Systems

As the users of VA systems have expanded beyond analysts and visualization experts, developers have to address the needs of users who are not already trained in how to use a particular system. A user can be a novice due to inexperience in the domain of the data, the particular VA system, or in VA in general. Morton, et al., (2014) talk about "Data Enthusiasts," a group of users that do not have formal training in data science but use visualization tools in their work, such as journalists who use data to illustrate their stories. For the purposes of this study, a novice user is someone who has little or no experience using a particular VA system. Many of the initial studies and common models in VA are based on the work of expert analysts in domains such as national security and finance, so the expansion in types of users requires reevaluation and adjustments of previous VA models.

Novice users have been found to have particular difficulties in interacting with VA systems. These users have to build a mental model of the VA system at the same time that they are trying to understand the target data set, which makes their initial sessions particularly difficult. This process in which a user questions how to interact with an object is what Norman (2013) calls the "gulf of execution," where users must establish how to interact with an object as well as what can be done with it. Users interact with objects with an intended goal in mind and look for clues as to how to interact from the object itself. Designers can ameliorate this process with "the use of signifiers, constraints, mappings, and a conceptual model" (Norman, 2013) that give proper direction to the users as to how to interact towards their desired goals.

Understanding the interactions available within a system and accessing the appropriate functionality when needed is particularly difficult for novice users in VA systems, as designers increase the number of functions available to users to address the complexity of data problems. Users must then further develop their mastery of the system to optimally interact, which may increase the level of difficulty in completing their targeted task. As summarized by Sun et al. (2014), these modern tools offer an extensive array of capabilities made to be wielded by experts, while for novices it is akin to "expecting someone to know how to build a house by simply sending them to Home Depot" (Sun et al., 2014).

Grammel et al. (2010) examined the challenges novice users encountered while building visualizations, categorizing the issues as translating questions into data attributes, designing visual mappings, and interpretation of the visualization output. These coincide with similar cognitive models such as Liu and Stasko (2010) that emphasize the visualization process of translating the user's understanding of data into a visual representation, as well as interpreting visual representations as they pertain to the patterns and relationships in the underlying phenomena. Grammel et al. (2010) confirmed this is particularly difficult for novice users when using VA systems, though participants in their study were not able to interact with the system directly. Users could only use the system through an intermediary as they talked through their process, which limits the applicability of the findings to understanding the actual issues encountered by novice users during interactions. Those utterances took the form of selecting data attributes or visual representations, and sometimes asking direct questions of the system such as "What are our best sellers?" or "What do we make the most money on?" The requests were not meant to replace keyboard and mouse actions but instead to capture some manipulation of the data set or the visualization presented. The main focus of the study was on the language of those requests in order to better understand the mental models of novice users because, in a direct reference to the work of Norman (2013), "[t]o bridge the gulf of execution, we need to understand the mental model visualization novices have of visualization specification" (Grammel et al., 2010). Even with an expert intermediary for the system, these users still struggled to build and interpret visualizations, so they may be expected to have further difficulty without this direct intervention.

Kwon et al. (2011) extend the work of Grammel et al., (2010) by having the users interact directly with a VA system in order to identify what they call the roadblocks for novice users. They initially implemented a pair analytic method similar to Grammel et al. (2010). However, they found that the roadblocks were more apparent when the users were able to interact directly with the system. Participants were given a brief introduction to the Jigsaw system (Stasko, Görg, & Liu, 2008) before completing a set of investigative tasks drawn from the work of Kang et al., (2009). Qualitative coding of the interactions revealed four major roadblocks experienced by their participants: failure to choose appropriate views, failure to execute appropriate interactions, failure to interpret visualizations, and failure to match expectations and functionality. The users also had difficulty learning how to use the given VA system within a limited amount of time, which is a common shortcoming of experiments within a lab setting. The difficulties in interpreting the visualization support the view of Liu and Stasko (2010), while the difficulties in knowing and selecting proper interactions highlight possible interface design issues relating to Norman (2013)'s "gulf of execution" that requires future work.

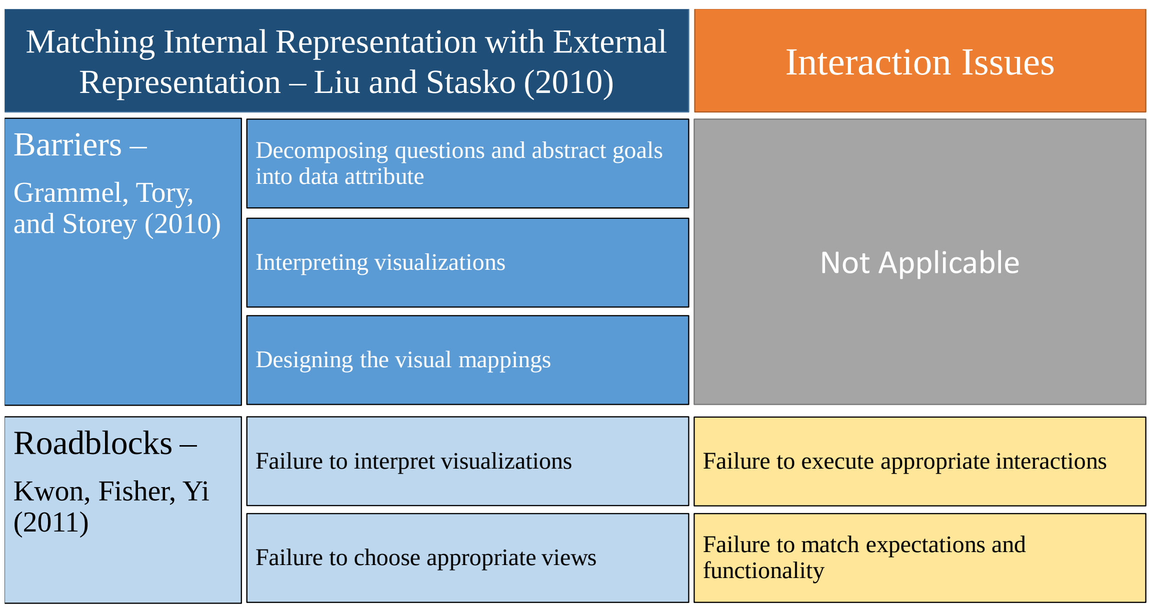


Figure 1 - Issues for Novice VA users

2.2 Relevant Issues in VA Interface Design

The following section discusses interface design issues highlighted by researchers in the Human-Computer Interaction (HCI) community that are relevant to VA systems. Specifically, traditional

Graphical User Interfaces (GUI) are contrasted against Natural User Interfaces (NUI) and multimodal systems.

2.2.1 GUIs and WIMP

The default modern desktop computer visual interface is a Graphical User Interface (GUI), which uses the desktop metaphor, WIMP (Windows, Icons, Menus, and Pointer), direct manipulation, and What You See Is What You Get (WYSIWYG) to guide users visually through their interactions. The desktop metaphor is an effort to recreate some of the affordances of a physical desktop within the interface, including the ability to place and move objects around and icons that match real-world objects. WIMP interfaces emphasize visual search to find the appropriate widget available within the desktop space and hierarchical menus. Direct manipulation, as proposed by Shneiderman (1981), moves from an emphasis on programmatic interactions to a fully user-controlled system, where the user directs visible objects on screen while the system responds predictably. This ties into the WYSIWYG paradigm, where users have a visual operation of their system, so they can see the current state of an object like a word processing document that matches its eventual output (W. Liu, 2010).

The use of WIMP interfaces for VA systems may exacerbate the challenges faced by novices. Lee et al. (2012) point out many of the known challenges with WIMP interfaces identified by the HCI field. A well-known drawback of WIMP interfaces is the tendency towards hierarchical menus and an abundance of buttons as the availability of new functionalities increases, causing a phenomenon referred to as “drowning in functionality.” This makes discovering appropriate functions especially difficult for novice users as the number of functions available grows. Novice users may have difficulty forming a useful mental model of the system and then may experience a large “gulf of execution” in understanding the capabilities of the system. Lee et al. (2012) also find WIMP to be inappropriate in many new contexts outside of an office setting, such as hand-held devices. They challenge the visualization field to expand interactions in new ways, concluding that WIMP is not appropriate for many of the tasks and environments in which VA systems are being used.

2.2.2 Natural User Interfaces (NUI) and Multimodality

Natural User Interfaces (NUIs) are seen as a progression in the field of HCI beyond GUIs to address the needs of users by integrating and synergizing more modalities. The availability of methods for users to interact with digital objects is then made more diverse, robust, and reflective of how humans interact in the physical world. Liu (2010) breaks down the characteristics that distinguish NUIs from GUIs, including new interaction types such as voice as well as an emphasis on capturing multiple signals from users, creating a multimodal interaction environment. Modalities such as hand gestures, body positioning, and facial expressions have all been successfully integrated into systems as input modalities because of progress in fields such as computer vision field (Ko et al., 2004). Early development of multimodal systems included the “Put-that-there” system (Bolt, 1980), where users used voice and gesture to control shapes projected onto a large screen. Users then gain flexibility in how they can interact with the system, with modalities offering the ability to communicate equivalent information in the manner most appropriate for the user at the time.

Lee et al. (2012) suggest the VA field move towards new interfaces within the Natural User Interfaces (NUIs) paradigm, where the emphasis is on developing interfaces as invisible to the users as possible “to minimize the gap between a person’s intent and the execution of intent” (Lee et al., 2012). This approach often calls for multimodal interactions, where the user can choose whichever modality or a mix of modalities that best fit their task. The focus then is not on managing functions in a limited visible space, but instead on mapping functions to their appropriate modality or multiple modalities so users can better interact. Cook and Thomas (2005) in their seminal work of the VA field *Illuminating the Path* proposed that multimodality is a necessary addition to VA systems, as the availability of many modalities can overcome the weakness in any of them individually, and called for more work into the benefits of multimodality for analytical reasoning. This is also a logical extension of the cognitive efficiency view and the interaction cost model of Lam (2008), as these new modalities are seen as an opportunity to lower barriers for users to interact with the system to meet their goals. In terms of Norman's "gulf of execution," these new modalities are expected to be helpful since they utilize a conceptual model already present in the user for how to interact in situations outside of the system, such as communicating using voice and gestures with other people. The challenge is then to map these actions and intentions to interactions with the system and perhaps more easily facilitate those encounters.

Multimodality goes beyond just offering different types of discreet interactions, as systems can combine these modalities in a synergistic manner. Bourguet (2006) presented the major ways modalities can be combined in a system including redundancy, complementarity, disambiguation, support, and modulation. Redundancy occurs when the same information is conveyed over two different modalities, such as capturing the sound of a user's voice and the user's lip movement visually, which then increases the accuracy of the signal recognition (Duchnowski, Meier, & Waibel, 1994). Jaimes and Sebe (2007), in their survey of multimodal systems, found the ability for systems to use redundancy to resolve ambiguities as a strong advantage over traditional unimodal interfaces. Complementary modalities each convey one portion of a message that then can be combined to form a complete message, allowing for flexibility as the user can select which modality is most appropriate for a given type of information signal. Mutual disambiguation occurs when the combination of two otherwise ambiguous signals results in an unambiguous message, such as the gesture and voice in Bolt (1980). Support is where one modality is secondary to another dominant main modality, such as integrating gestures or eye-movement for voice controls to assist in communication but never to overrule the main modality. Modulation is when one modality can alter the contents of another modality but is not a complete message on its own, such as facial expressions altering the meaning of words spoken. Modalities can then be related in any of these ways, given different system environments and different user tasks.

2.2.3 Challenges in Designing Multimodal Systems

Designers and developers of multimodal systems are presented with challenging technological problems as these new types of actions, and the combinations of them add uncertainty to each communication. Bourguet (2006) listed some of these challenges for system designers, including the need to process heterogeneous streams and dealing with uncertainty and recognition errors. For example, keyboard input is precise every time while user actions like gesture and voice are inherently contextual and probabilistic, therefore requiring a certain amount of real-time processing before taking action. NUIs may also have an unexpected effect on users despite aiming for naturalistic interactions.

Offering additional modalities may not fully address the "gulf of execution," as described by Norman (2013), that novice users experience with VA systems. Norman (2010) argues against the term "Natural User Interface" since these novel interfaces still require some level of cognitive effort as users learn how to interact and think through each interaction. Grafting modalities onto a digital system to reflect more real-world interactions may further confuse users. Examples include calculating the proportional momentum of objects in a digital space when using gestures, as well as the use of conversational language when interacting with traditional systems. The idea of what is natural is specific to the context in which the interaction takes place. Hinckley et al. (2010) find that a match between interface and task is what makes for more natural interaction, observing that users prefer to use their fingers to manipulate objects and pens to write when given a choice. These new modalities may only seem natural in the context of specific situations. They may cause an increased level of effort to execute compared to the traditional and therefore expected modality within the context of different systems, interactions, and goals.

2.2.4 Multimodality in VA

Drawing from the HCI’s movement towards NUIs, many VA researchers are moving towards multimodal interfaces that include more novel interactions for users. Keim et al. (2010) suggest that the average person’s increase in digital data such as emails, photos, and files creates an opportunity to support these “naïve analysts” as they deal with their growing data collections. They predict this will require a new set of interaction modalities to meet this new population, and this challenge must be met by new visual analytics applications. Endert, Bradel, and North (2013) see an opportunity to integrate novel modalities to allow the user to express more complex intentions, such as being able to manipulate multiple objects using multi-touch interfaces and gesture. Increasing the available modalities may also be a form of information recovery, as Blandford and Attfield (2010) point out the digitalization of an object inherently means some loss of modality (e.g., the feel or smell of the original).

More novel interactions have been created to better approximate the physical world, with the emerging field of Immersive Analytics integrating focused on creating engaging and embodied user experiences (Marriott et al., 2018). An example of this approach is the FiberClay system developed by Hurter et al., (2019), which allows users to manipulate and query 3D trajectories using user gestures and body positions. They explicitly set out to avoid any WIMP components in their system as they felt it distracted from the user’s ability to perform the tasks. Other systems have integrated more novel modalities such as smell (Patnaik, Batch, & Elmqvist, 2019) in order to offer more engaging interactions. This growth in multimodal interactions in VA has made evaluation more challenging for researchers, with the systems like VA2 (Blascheck et al., 2016) developed specifically to analyze and compare heterogeneous streams of interaction data coming from multimodal VA systems.

2.3 Natural Language and Voice Modalities

The following section summarizes previous work in developing Natural Language Interfaces

(NLI) and Voice User Interfaces (VUI) for VA systems and desktop environments.

2.3.1 Natural Language Interfaces (NLI) for VA

Natural language interfaces (NLIs) offer an important interaction modality as designers look for ways to improve the usability of their systems. The basic premise of an NLI is that it can process some form of semi-structured or unstructured requests from users and perform an appropriate action as a response. Multiple studies of novice users have found that they can articulate the questions they have about the data as well as their intents to a much better degree than they can interact with the system's existing tools (Aurisano et al., 2015; Grammel et al., 2010), so many researchers feel that a system's ability to handle more natural language may be the key to improved usability. A user being able to use their vocabulary and phrasing to interact with a VA system ameliorates part of the translation process as described by Liu and Stasko (2010), as well as addressing the interaction issues highlighted by Kwon et al. (2011). It is possible that novice users can communicate some component of their mental model of the data and then be shown a visual representation. If novice users are more able to articulate their desires than to directly utilize the functions available in a VA visual interface, then there may be an opportunity to close the gulf of execution by linking their language to interactions.

The earliest related work is the development of NLIs for querying databases, which has been around since at least the early 1970s, with LUNAR (Woods, Kaplan, & Nash-webber, 1972) as the most famous early example of this type of work. Some approaches limit the content and structure of queries in order to simplify the parsing process, such as Precis (Simitsis, Koutrika, & Ioannidis, 2008), which runs on keywords and domain-specific taxonomies. This is very constraining on users as it assumes that they have to have prior knowledge of the domain of the data as well as how to properly structure requests to the system. As more systems have increased the scope of requests allowed, developers have had to deal with more of the challenges of accepting natural language. This includes handling the terms of these requests, which requires an understanding of domain vocabulary and synonyms. The semantics of natural language is a more difficult problem, as drawing meaning from a request based on its structure includes cases such as negation and conditionals that transform its meaning and intention. Many approaches attempt to address some of these issues by mapping requests to SQL queries by receiving natural language and translating it into a formal structured language appropriate for data. NaLIR (Li & Jagadish, 2015) utilize the Stanford NLP Parser (De Marneffe, MacCartney, & Manning, 2006) to transform requests into a tree structure that can then be translated into SQL. They found through users studies their system to be effective enough that "even naive users are able to specify quite complex ad-hoc queries." Research in NLIs for databases address the questions concerning how users can ask questions about the data but do not capture many of the interaction issues inherent in a VA system as listed by Srinivasan and Stasko (2017) including providing input affordances (e.g., informing people what they can say or ask) and explaining system results.

VA researchers have attempted to integrate natural language into their systems, with early work by Cox et al. (2001) appending an NLI onto an existing VA system. They saw the key challenge to user adoption of VA systems as the level of detail necessary for a user to know about a system's capability before performing a task. The system was an extension of the NLIs for database technology of the time, where natural language is translated into a query, and the resulting dataset was plotted on a graph. Users were able to interact through a traditional GUI as well as a voice input system. They used domain-specific terms gathered from experts to disambiguate queries and interaction intents, which could create a barrier for novice users unfamiliar with the vocabulary.

An example of a modern approach to NLIs for VA systems is the Eviza system (Setlur et al., 2016), which allowed for both text and verbal queries to a VA system. They explicitly differentiate themselves from the work of Cox et al., (2001) by having their system deal with ambiguity by offering a single best guess at user's intents as opposed to requesting clarification, which they believed interrupts the user's flow. They compared their work against similar systems and found that others tended to return visualizations that are either static or very limited in interaction possibilities. They instead focused on facilitating a wider range of interactions but on an existing visualization, extending the work of Gao et al. (2015) on the DataTone system. The participants using the Eviza spent significantly less time on each task compared to users of the proprietary VA system Tableau. They found their system was preferred over the traditional mouse and keyboard interactions in a select few cases: the task required many clicks to complete, the users did not know how to complete the task, the user did not know where to find the function, or the user did not know the name of the variable where the target variable belonged (ex: South America found under the dimension “SubRegion”). One of the challenges faced by the system was recognizing long and complex queries that require complex semantic parsing. This is analogous to the challenge of verbose queries in Information Retrieval (Gupta & Bendersky, 2015). These findings may not translate to novice users, as the participants were already familiar with the system, so the voice modality may have just made the navigation of known functionalities easier versus making the system more accessible.

2.3.2 Voice User Interfaces (VUIs) for Desktop Environments

Although Voice User Interfaces (VUIs) offer a new opportunity for interaction, there are particular issues to using voice in a desktop environment. HCI research as early as Van Buskirk and LaLomia (1995) advise against using voice as the primary modality when there is a physical alternative since physical input of keyboard and mouse are considered more efficient in terms of task time. Many studies have taken this view of voice as secondary, and explore the usefulness of voice as a modality when there are situational impairments such as driving or limited hand available due to other activity or disability (Corbett & Weber, 2016).

Voice modality is particularly challenging in terms of learnability, a key aspect of the overall usability of technology. Grossman, Fitzmaurice, and Attar (2009) define learnability as “the ease with which new users can begin interaction and achieve maximal effective performance.” Chen (2006) breaks down the factors that compose learnability, including the availability and effectiveness of training/tutorials and the system’s ability to actively assist users in becoming proficient in use. The factor that is most relevant to voice as input is the degree of difficulty a user faces in the process of discovering system functionality, as interactions are inherently invisible. Yankelovich (1996) posits that voice interactions create two fundamental challenges to discoverability: users will assume the system can understand more than it is capable of, and users will be unaware of functionality that is available. White (2018) highlights this problem for virtual assistants, where new functions are challenging to discover, especially with the growth of functions through third-party extensions. This lack of discoverability due to the opacity of voice input leads to an inconsistent mental model of the system’s capabilities (Karsenty, 2002). Feng, Karat, and Sears (2004) found that listing available commands to users did not keep users from forgetting or missing commands in their voice system.

Researchers have tried different techniques to signal to the user what the system can and cannot do within the desktop environment. Often systems incorporate the “Say What You Can See” approach (Yankelovich, 1996) approach, where any voice interaction available should be clearly labeled and visible on the interface. Researchers such as Christian et al., (2004) and James and Roelands (2002) use this technique to signal to users how to interact with their systems. Corbett and Weber (2016)’s system “What Can I Say?” is meant to assist users using a voice system to discover functionality by offering voice requests examples on command. However, some users would forget the presence of the assistant system, and others could not recall the command to open it. Feng et al., (2004) and Hu et al., (2011) find that the initial stages of voice interactions are where the most intervention is needed to assist users, which is consistent with the findings of Corbett and Weber (2016) who find that learnability is most challenging within those initial encounters.

2.3.3 Voice Interfaces for VA

Voice has become more ubiquitous in HCI design, leveraging NLI technology and improved speech recognition to provide users the ability to interact outside of the traditional keyboard and mouse set up. Researchers consider two main opportunities when adding a voice system. The first is multi-tasking or hands-free interaction, which has been considered in a VA setting in terms of supporting other simultaneous modalities. Cook and Thomas (2005) believed that adding speech would allow for users to perform both visual-spatial tasks at the same time as verbal commands with little cognitive interference, citing the work of Kinsbourne and Hicks (1978) and Wickens (1984). The advantage that is more relevant to novice users is that voice systems can facilitate a form of communication that is more familiar and comfortable compared to a GUI system. Liu (2010) posited that voice is the "most natural flow, convenient and efficient way of information-sharing," and therefore the easiest form of communication for users. In their study of novice users, Kwon et al. (2011) observed that one participant verbalized every action during the introductory tutorial session, and then during the experimental task used more views for analysis than any other participant and had no noticeable interaction issues. Whether or not researchers directly cite Norman (2013) like Grammel et al. (2010), the developers of voice systems for VA are addressing the gulf of execution by providing what they believe is a modality that provides users an easier way to access functions.

Voice does present unique challenges as a modality for both the VA system designers and potential users. Bourguet (2006) emphasized how natural modalities like speech are inherently error-prone, as it gets difficult to separate signal from noise as well as maintaining an extensive vocabulary that matches the user's understanding of the system and the task. From the user's perspective, the functions available through voice are inherently invisible, so it is difficult for users to maintain a complete mental model of the system at any given time (White, 2018). Voice may not always be necessary to provide the most appropriate modality for an analytic conversation, as attempts using textual dialogue such as Tory and Setlur (2019) have been successful. Kassel and Rohs (2018) used text not only to converse but also to convey supplementary information about the data visibly, an approach coined as "data facts" by Srinivasan et al., (2019).

Additionally, most voice systems that offer a wide range of functions have to build a certain level of intelligence in order to support users in their goals. Interacting with a voice system is inherently probabilistic, and an increase in capability often comes with a need for assumptions and increased system influence over the interaction. Voice can be more limited to just addressing accessibility in simple tools such as text to speech, but offering more complex functionalities necessitates more system influence over the outcomes of interactions. This is an issue that has been brought up from the early development of intelligent systems, with no clear guidance on how much a system should infer from interactions. Bates (1990) frames this as an issue of influence and power and ultimately recommends keeping the user at the center of design as the best strategy.

Given the researcher is willing to address the challenges previously mentioned of voice as a modality, the mapping between what users say and how the system responds is critical to building successful systems. This directly affects the usability (Vassallo et al., 2010) and the user's willingness to engage with the system (Luger & Sellen, 2016). Molich and Nielsen (1990) stress that a system's success is predicated on its ability to match the "words, phrases, and concepts familiar to the user" (Molich & Nielsen, 1990). Language choice may be a particular problem for VA systems since it would be meet the criteria as a domain-specific voice assistant (Mennicken, Brillman, Thom, & Cramer, 2018), as both the data and the actions within the system are specialized domains as opposed to general-purpose conversation assistants such as Amazon's Alexa. The structure and content of communication are then essential design considerations. Shechtman and Horowitz (2003) categorize the different conversation styles between a user and a conversation system into three groups: task-oriented, communication-oriented, or relationship-oriented interaction. Though many developers have aimed for deeper and more communication-oriented systems, Luger and Sellen (2016) find that most users focus on simple tasks over more complex ones, as they often do not trust the system to operate correctly. Branigan et al. (2011) find that users tend to change and adapt their conversational style when they believe they are communicating with an intelligent system in comparison to communicating with another person, meaning grafting human-to-human communication structures may not be appropriate.

In order to respond appropriately to user requests, designers must determine the appropriate amount of inference the system is allowed to make about the intentions of the user during interaction. A system can be designed to perform simple operations called by explicit requests, or engage the user in a more conversational style interaction where the system has a more complex model of the user as well as the user's intentions. Given the user is within a task-oriented scenario (Shechtman & Horowitz, 2003), it is unclear if users expect or require a system to infer more from their behavior than an explicit command, or if they expect just an alternative interface to their default modality in order to fulfill their objectives. An example of a simple system is the work of Diesendruck and Zhao (2016), who create what they call a "grammar of graphics" or "ggspeak," where they broke apart statements into belonging to the session-level (ex: save, reset, quit) or graph-level class (ex: group by). They find that users tend to build iteratively when using their voice, "first defining a basic relation-ship, then proceeding to more complex features." (Diesendruck & Zhao, 2016). These actions are all explicitly called by the user and mapped to specific functionalities available in the system, thus minimal assumptions concerning the intent of the user beyond the content provided in the requests.

In contrast to simple command-driven voice interfaces, other developers focus on conversation voice interfaces for VA systems to help users navigate through their tasks. The main benefit of conversational voice systems over other approaches is that they "hide the complexity of visual analysis from the user to lower the engagement boundaries" (Kassel & Rohs, 2019). The work on the Articulate system (Aurisano et al., 2016) emphasized the role of gestures and conversation for interaction. Their previous work in understanding the effect of a conversational interface for a visual system involved a Wizard of Oz experiment, where the user used voice and gesture to interact with a system that was being controlled by an expert user in another room (Aurisano et al., 2015). They emphasized the importance of gestures in their system, citing the work of researchers such as Navarretta (2011) in multimodal expressions such as pointing at an object before asking a question. Kumar et al. (2017) acknowledged the possibility of other expressions of ambiguity resolutions such as eye gaze in building Articulate2, but choose not to integrate it into their study. They differentiated themselves from other systems such as DataTone (Gao et al., 2015) by allowing for two-way communication between the user and system, and Eviza (Setlur et al., 2016) by allowing for the creation, modification, and interaction with multiple visualizations. A significant limitation of their approach is the environment in which users interact with the system. They choose to offer two large 4K screens for improved visibility of objects in the system while removing the option of direct input through mouse and keyboard. Their experimental set up is not only unrealistic to most work environments available to users, and ignores the possibility that users may choose to utilize mouse and keyboard when given the option. This also fails to address why users would choose voice if they had the opportunity to interact traditionally with the system.

Developers of conversational voice systems for VA systems can leverage more of the features of bidirectional communication, such as requesting clarification in cases of ambiguity (Cox et al., 2001) or adjusting previous utterances to repair some ambiguity in Evizeon (Hoque, Setlur, Tory, & Dykeman, 2018). Orko (Srinivasan & Stasko, 2018) communicated its actions to its users while they interacted with network visualizations. The system was made with the intention of providing a multimodal environment that includes both natural language and touch-based interaction, combining the two modalities in a complementary manner. In a follow-up study, Saktheeswaran, Srinivasan, and Stasko (2020) confirmed that users preferred having multiple modalities available for VA since it gave them more room for complex expressions of their intents. The Orko voice system was fully automated, driven by a query parser that mapped utterances to intentions and data actions. They found that voice was generally used for search, filtering, and finding connections in the network. The touch interface was used for node selection and highlighting. In evaluating the voice system participants were "pleasantly surprised" by its ability to understand their queries, mostly due to negative previous experiences with voice systems. One domain expert participant found the voice system enjoyable to use and predicted it would be helpful for novice users. The program was run on a large touch screen device, which is quite different from the everyday environment of these users.

2.3.4 Summary

This chapter summarizes previous studies on how novice users interact with VA systems and how adding voice may affect the way users interact as they perform their tasks. Novice users have their own particular set of challenges when interacting with VA systems, with the main theme being the ability to reconcile their internal view of the data with the external representation. Their inability to translate their questions into appropriate system interactions is an example of Norman's "gulf of execution," where users are unable to properly build a mental model of the system to achieve their goals. Studies in HCI have found that the traditional WIMP interface design of VA systems may be a contributing factor to these difficulties, as functionalities become obfuscated in hierarchical menus or overwhelm the visual space with widgets. Natural language and voice have become increasingly prevalent in interface design in order to address these issues, and VA system designers have begun to integrate voice into their systems.

Chapter 3

Positioning this Study in the Literature

This chapter draws from previous literature to create a possible model of how novice users would utilize voice interfaces for VA systems. From that model, this research proposes two research questions that are unanswered in the current literature that will be examined in this study.

3.1 Modeling How Novice Users Interact with VA Systems

In building a model of novice interactions with VA systems, it is important to establish the level at which those interactions will be observed. If voice systems are meant to help users better communicate questions and reconcile their mental model with the data presented, then we could examine the interaction between the system and user at this cognitive level. Pike et al. (2009) advocate for more work in this area "[g]iven the close coupling between interaction and cognition" (Pike et al., 2009). However, it is difficult to observe and operationalize this type of cognitive interaction in practice. Other studies were able to operationalize changes in the users' mental model by collecting artifacts before and after interactive sessions. Zhang and Liu (2020) had users draw mental maps of topics before and after search sessions and measured the change in vocabulary. This approach may be applicable to understanding how users go through the sensemaking process interacting with VA systems, as mind maps collected during a task would likely change in a similar fashion when interacting with a dataset. However, this may be more difficult to operationalize when measuring the users' understanding of how to interact with a system as opposed to knowledge on a topic.

The literature presents voice systems as an opportunity to solve an interaction problem related to accessibility, as it can offer the ability to cut through a visual interface to access otherwise difficult to find functions. The model of interaction should then be at the function level in order to test this hypothesis. As mentioned earlier, the issue of navigating the visual interface is closely related to Norman (2013)'s "gulf of execution," where the user is trying to understand how to interact with an object in order to accomplish some goal. Designers rely on a set of tools in order to assist users in across this gap: signifiers, constraints, mappings, and a conceptual model. The first three are characteristics of the interface that signal to a user how to interact, while conceptual models are constructed by the users' previous experiences. In the case of novice users, they do not have an exact conceptual model of how to use a particular VA system but may have experience with other systems that are applicable. The accessibility of functions is then a "gulf of execution" problem where the application of signifiers, constraints, mappings, and the novice users' conceptual models are insufficient in accessing the appropriate functions. Figure 2 models this problem on an individual system function, where the designer uses the color green to signify the appropriate steps for the user to access the desired function.

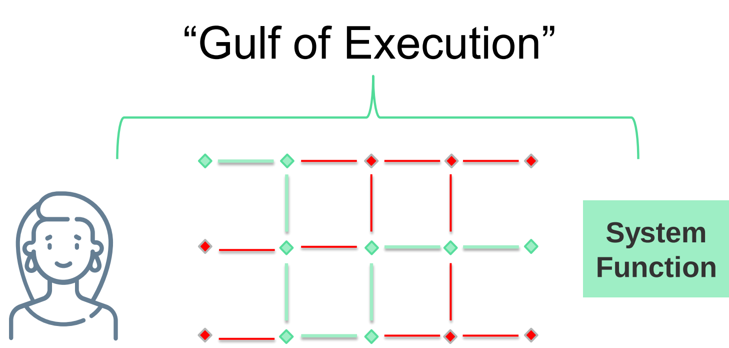


Figure 2 - "Gulf of Execution" in Accessing a Function

Given we model the difficulties novice users have in navigating the visual interface of VA systems, the next step is to model how those users would potentially use voice systems to solve this issue. In previous work like Orko (Srinivasan & Stasko 2018), the voice system was meant as a replacement to the visual interface, with the expectation that users would be better able to interact with the system. In applying this to the “gulf of execution” model, the voice system is meant to be a direct bridge between the user and the desired function. Using a voice system in this way is not a simple task, as covered in Section 2.3.3. Moving towards automated systems requires a large corpus of requests in order to properly model the content and structure of the users’ language and to map those requests to appropriate functions. Previous research has focused on this aspect of voice systems and has made progress on this front, but has relied on experimental designs that only offer the voice modality to users. This was deemed necessary in order to encourage enough requests for analysis at the expense of reflecting the environment in which most users would be interacting with VA systems.

If researchers are interested in addressing the “gulf of execution” in interacting with VA systems, then users should be allowed to interact with systems in the method they are most likely to encounter. This then requires allowing users to interact with VA systems using either the voice modality or keyboard and mouse. It cannot be assumed that offering a voice system would automatically mean users would choose to use it or that it would be easier to use in comparison to another modality. As expressed by Norman (2010), every modality has different affordances that lead to different challenges. Given the voice system is chosen by novice users, the “gulf of execution” problem would be assuaged if users were able to access different functions that would otherwise be difficult to access. An updated model of how users would access system functions using either the voice system or keyboard and mouse is presented in Figure 3.

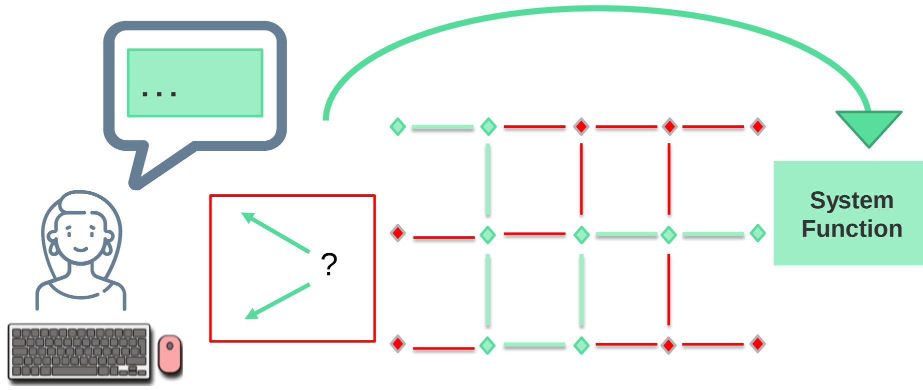


Figure 3 - Users Choosing Voice or Keyboard/Mouse for Accessing System Functions

3.2 Research Questions

The purpose of this study is to understand how novice users of VA systems would choose to utilize a voice modality within a traditional desktop computer environment when given a choice, and whether having the voice modality available changes what functions novice users choose when performing their tasks. If the interaction issues pertaining to the “gulf of execution” are better addressed by having access to a voice system, then users who use voice are expected to have accessed different functions than those who did not. This study is an extension of previous work such as Tory and Setlur (2019) that utilize a Wizard of Oz study design to observe how users would interact with a VA system that includes a voice modality component. Unlike previous studies, the users can choose to either use their keyboard and mouse or the voice system when available as they see fit. The users are novice users unfamiliar with the VA system before the study. Specifically, this dissertation addresses the following research questions:

RQ1: Under what circumstances do novice users of VA systems choose to use a voice system when given the option within a traditional desktop computer environment?

RQ2: Does offering a voice modality change the functions chosen by novice users?

3.3 Significance of This Study

This dissertation proposes to build on previous work on studying novice users of VA systems and on using voice as an input modality for interaction. Novice users of VA systems build mental models of data but must also build a mental model of the system. Voice presents an opportunity for system designers to present a modality that can potentially close the “gulf of execution.”

The major distinction between this study and previous works is the emphasis on modality choice. As opposed to limiting users to just using a voice system, this study will analyze when a novice user would choose to use voice for interaction when the traditional alternative of keyboard and mouse is also available. The voice system is an addition to the traditional desktop set up and not a replacement, and better represents the likely environment a novice user would interact with a VA system. The tasks themselves will require users to build visualizations from scratch instead of interacting with premade visualizations, thus allowing the study to capture more interaction choices. This study will compare a VA system running in a traditional desktop environment to one with an additional voice system that focuses on providing the functionalities available to the user that would normally be accessed through keyboard/mouse actions. If the voice system does improve the level of support provided by the VA systems to users, then there should be a noticeable difference in functions used between the two groups. The strategies used in choosing voice, as well as difficulties encountered, will be informative for future VA system design.

3.4 Summary

Previous studies have presented the argument for how voice systems could assist novice users in interacting with VA systems. This study presents a model demonstrating how this could be possible but highlights points of concern that have previously not been tested. This dissertation proposes to offer an additional voice modality on top of a traditional desktop system and to evaluate the voice system based on its effectiveness in making functions more accessible.

Chapter 4

Methodology

4.1 Research through the Cognitive Viewpoint Framework for Information

This study aimed to understand the choices made by novice users as they interact with a VA system and proceed through two cognitive processes: the sensemaking of a data set and the creation of a usable mental model of a VA system. Both of these models stem from or are influenced by the Cognitive Viewpoint, which posits that all information is mediated through a person's cognitive model. Research should then be focused on "the variety of individual world models and knowledge structures that [underlie] the surface structures of the variables of interaction" (Ingwersen, 1992). Though neither of these cognitive processes can be observed directly, the interactions with the VA systems are expected to be related to the state of those models at a given point, as explained in Section 2.1.3. The data collection then focused on users as they interacted with the system, as well as their own explanations of their thought process.

4.2 Principles in Making Methodological Choices

It was clear from the beginning of this study that addressing these research questions would require observing participants interacting with a system. Studies such as Saraiya et al., (2006) on bioinformaticians were able to examine the actual use of VA systems for tasks in the participants' natural environment over time. However, VA systems with additional voice interfaces are not widely available, so conducting this type of study was not possible. Participants would have to be exposed to a prototype voice system for the first time. Their behaviors and responses would have to be collected with the assumption that they should reflect patterns in potential real-life situations. This meant a series of decisions concerning the experiment would have to be made to simulate how novice users would potentially use a voice system. The following principles were crafted to best address the research questions and provide guidance on the many necessary design decisions.

No Preferential Treatment of the Voice System

Given the emphasis in understanding the reasons for choosing the voice modality to interact with the VA system, it was important then not to influence the user into choosing the voice system. Any action that would overly emphasize or encourage the use of one modality over the other would have skewed the results of the study. This was especially important when dealing with novice users who would be particularly sensitive to this during training since they had no previous experience with the VA system. Design decisions then tried to ensure that every effort was made to offer all possible interactions to the participant without bias, with no indication that the voice system was in any way a focus of the study or preferential for use. The risk that a participant would choose not the use the voice system at all was acceptable as long as it was believed that it better reflected their attitudes and actions outside of this experiment.

Access to All Functions Available

It was expected that the addition of a voice modality to the VA system would increase the use of functions that would normally be difficult to access through the visual interface. If access to functions was going to be a key metric in comparing conditions, then it was important to allow participants to access as many functions as possible during the tasks. Limiting the functions available would affect the interactions possible and skew the results. If the participant were limited to only the small subset of functions demonstrated during training, it would have possibly diminished the participants’ need to find new functions through the interface or through the voice system. The reverse was also possible, where a designer could have purposely limited the functions available to those hidden deep within menus or widgets, therefore making using the voice system more attractive to the users. A wide range of functions would have to be provided by the system for the participants to explore and use for the tasks, and participants were allowed to discover and use all of the functions available.

4.3 Special Note on the Effect of COVID-19 on this Research

The original design of the study had participants perform the tasks within a controlled setting. A university UX laboratory served as the experimental setting, which would have allowed for control of the environment. Additional technical support, including the video recording of participants and potentially capturing eye-gaze data, would also be available, using a similar methodology and algorithms as in the work of Alam and Jianu (2017). Changes in research protocols were necessary due to the COVID-19 pandemic. At that point, two participants had already performed the experiment in the UX lab, but almost all in-person research was shut down afterward. The choice was either to change the protocols to run the study remotely or to pause the research for an indeterminate amount of time until it was determined that it was safe to conduct in-person research again. The research questions did not explicitly require in-person participation, so the study was adjusted to be conducted remotely. The new remote setup had to be tested again in order to verify the data collection methods and to ensure the experience was as smooth as possible for participants. The change from in-person to remote affected many aspects of the study, including system design, data collection, training materials, and participant recruitment. Any changes caused by this adjustment are mentioned in their respective sections.

4.4 Experiment and System Design Testing

Testing of the VA system, the voice system, and the data collection tools required many different iterations with the help of volunteers and colleagues. It was important to get enough testers who had little or no experience with the VA system to test the various configurations of the experiment and especially the voice system. Guidance for the wizard depended on documenting issues related to domain language, synonyms, the ambiguity of intent, and error handling. These sessions would sometimes be short to cover specific testing issues, such as the task instructions, with a few testers going through the complete experiment to offer feedback. There was no formal pilot testing planned for this study, and the plan was to rely on testing the task and system with these volunteers and iterating through changes. However, two users did participate in the study in-person before the necessary change in protocols. Their data is not included in this study since their interactions occurred in a very different environment. They did serve as an opportunity to test the training, tasks, and some of the data collection techniques under experimental conditions. Once it became necessary to change the experiment to be run remotely, a new set of testing was necessary. The lessons learned from these sections are referenced in their respective section in this chapter.

4.5 Longitudinal versus Single-Session Study

Given the decision to observe participants interact with a system, the frequency and duration of the study became the next issue for consideration. A longitudinal study could have been conducted, where participants would have been observed over the course of several sessions, such as Sedlmair et al. (2011), on the use of visual systems by employees at a large automotive company. This would have allowed for more training time as well as much more interaction with the VA system. The tasks could have also progressed to include larger and more complex datasets over time, reflecting the expected increase in the participants' ability to interact with the VA system. However, the focus of this study is on novice users as they interact with a VA system and explore its capabilities. In thinking about the "gulf of execution" issue, the first set of interactions are where the novice user is most likely to run into barriers as they gain a better understanding of how to interact with the system. Additionally, longitudinal studies require a larger commitment from participants, which affects the recruitment pool. Single sessions were found to be most common in VA system evaluation studies, so the experiment was designed to have novice users interact during a single interactive session.

4.6 Participant Setting and System Requirements

The switch from in-person to remote interaction meant any potential participants would have to have access to an appropriate computer environment to complete the experiment. Participants were required to meet minimal system requirements that would be common enough in the general population as not to be overly burdensome on their participation. This included a computer with a stable internet connection, Google Chrome, and a microphone. A stable internet connection was necessary to connect to the remote system. Google Chrome was necessary in order to use Google Remote Access, which facilitated the connection between systems. Each participant was given a unique code to access the system, which would expire once the connection was broken by the facilitator to ensure the security of the remote system. It also required users to be signed in to a Google email account, which could be used to verify the identity of participants. The remote experience in full-screen mode was found to be comparable during testing to interacting with the system in person. Other solutions required the installation of special software on both the participant and VA computers, which was considered overly burdensome and unnecessary.

Participants needed to have a microphone input into their computer in order to interact with the voice system and to be interviewed at the end of the experiment. There was also necessary communication at the beginning and end of each task, where the system would be reset and prepared with the appropriate data and materials for the next task. An open communication channel was necessary in case of technical difficulties during the task. Google Hangouts was chosen as the communication channel between the facilitator and the participants. Google Remote access did not have the ability to capture audio from the participants, so the additional communication channel was necessary. It is a free service that allows for communication through voice or text chat within a browser. Participants were told that the facilitator would be monitoring the text chat in case of any issues. It was understood that participants would be situated in their homes during the regional lockdown, so there would likely be environmental noise that would otherwise not be present during lab testing. Recording the participants' faces through webcam was deemed invasive and unnecessary in answering the research questions, so all participants were told to disable their webcams during the experiment.

4.7 Participants

4.7.1 Definition of Novice User

This study was meant to focus on voice systems as a possible tool to assist novice users better interact with a VA system, so the definition of novice users was paramount to this experiment. As discussed in Section 2.1.5, there are many aspects of the VA process that a user could be inexperienced in, including the system, dataset, visualizations, and analysis. Each of these facets creates its own barriers for novice users during the VA process. Because this research emphasized access to functions so heavily, the type of novice user tested here is specifically inexperienced with the particular VA system. Novices for this study pertained to new to a particular VA system because of the focus on the ability to find functions when the participant has limited previous experience with the interface. An effort was made to try and control for experience with the task domain, but the priority was always on recruiting participants new to Tableau and capturing their first interactions. The participants' level of experience in the other facets, such as business and data analysis, were self-reported in the pre-experiment survey.

4.7.2 Participants and Recruitment

The original intent of this study was to try and control for inexperience with the task domain by recruiting participants with business knowledge. Actual business professionals were found to be difficult to recruit and compensate within the resources of this study. Because this study was done using university resources, university students were more readily accessible. The original recruitment pool were students in the Masters of Business Science program at Rutgers University. The study had progressed to the point that two participants from the MBS program had completed the experiment in person in the UX lab. However, it became very difficult to recruit participants from this group once the COVID-19 protocols had shut down in-person learning and most in-person research. Ultimately it was necessary to change to the pool to include undergraduate students with business majors, but that recruitment group did not meet the 24 participant goal of this study. Finally, the pool was expanded to all undergraduate students, with 12 participants selected from business majors, while the rest were from other majors mostly related to computer and information science. This would create another layer of comparison where those students presumably with more business domain knowledge compared with the others.

The twenty-four Rutgers University undergraduate students who recruited were split into two experimental groups; one with access to the voice system and the other with just the baseline VA system. Participants were required to be fluent in English since that was the language used by the wizard and all of the experimental materials. The study required the users to have access to an internet-enabled computer with a microphone in order to access the remote system and communicate with the facilitator. Participants were compensated with a $40 gift card for their time. A sample recruitment letter can be found in Appendix A.

4.8 User Interface

4.8.1 Choice of VA System

The VA system for this study had to be robust enough to support a user interacting with a large dataset, from selecting data through examining details of visualizations. Building a VA system for scratch for this experiment, like the creators of Orko (Srinivasan & Stasko, 2018), is difficult, especially when trying to offer a large number of functions. The system would also have to be multimodal, with a traditional GUI accessible by keyboard/mouse with an additional voice interface. The choice was made to build a voice interface as an add-on to an existing full-featured VA system and to focus on integrating the two together.

Tableau Desktop was chosen as the VA system for this study since it met all of the criteria necessary. The system maps the interactions of users to the underlying VizQL query (Hanrahan, 2006), which describes both the data selected and the visual encoding. This allows for users to interact with data without any prior knowledge in a formal query language, as the system handles the mapping from drag-and-drop interaction to SQL query and visualization components. Tableau has a free academic licensing program and provides a robust library of training materials for new and experienced users. Tableau has been used as the basis for other user studies such as Mackinlay, Hanrahan, and Stolte (2007) and Tory and Setlur (2019). Version 2020.1 was used, which was the most recent update to the system as of the first participant's session.

4.8.2 Voice System Interface

Though voice systems have no requirement for a visual interface, providing feedback to the participant that they are engaging with the system was deemed necessary. The remote set up required one system focused on providing access to the VA system and recording interactions, while another served as the communication channel. The facilitator had could not control what was and was not heard from the participant during the experiment. Since the wizard could hear everything the participant was saying, it would have been possible to present the voice system as ubiquitous and always listening to the participant. However, this could have caused some issues of the ambiguity of intention, as a participant could be thinking out loud and be misunderstood as providing a request to the wizard. Communicating a clear request was important to the interactions, so creating a “Hey Viz” button to begin a request was thought to be the easiest solution.

The functionality of the “Hey Viz” button also served as a way to provide feedback on the current state of the system. Users were alerted that the system/wizard is listening because a “Listening” icon appeared on their screen after users clicked the “Hey Viz” button. While the wizard is handling the utterance, a “Processing” icon appeared on the screen. During this time, the participants were unable to view Tableau as the wizard controls the interface. This was facilitated by moving the Tableau window on to a different monitor than the one viewed by the participants. This meant that participants could not use the voice and keyboard/mouse interfaces simultaneously. This loss in functionality was necessary in order to maintain the illusion of an automated system, as the wizard’s movements would have indicated a human operator. When the wizard was able to perform a command, a “Complete” icon then appeared after the operation has been performed, and the window returned to the participants' view. If the wizard was unable to perform a command, an “Error” icon appeared. This functionality was built using AutoHotKey scripting.

4.9 Wizard Design

4.9.1 Wizard Abilities

The wizard was meant to provide a voice interface for a VA system comparable to previous studies. In choosing the most appropriate voice modality for this study, it is important to offer the best range of functionality available to ensure it is in-line with current development research as well as attractive for users to choose. Many different levels were considered, including a simple voice system that would be limited by a required request structure and literal references to objects within the visual interface. It was limited enough to actually be built for this study with limited human interventions and would have been similar to ggspeak (Diesendruck & Zhao, 2016). Upon further discussion, it was thought to be too simple and would require very particular training in order to use it. Instead, the experiment focused on representing a more advanced voice system, which pushes the limit on the current plausibility of what is possible in voice for VA systems. Tory and Setlur (2019) tested four different levels of wizard support within a VA environment, with the most capable system taking into account the analytical intent of users as well as the context in which they operate. The wizard of this study matched the same level of capabilities, which required wizards to exercise judgment on how to best serve users beyond explicit function requests, including understanding domain synonyms, semantics, and context. The wizard had a set of example utterances mapped to actions in Tableau in order to facilitate responses that are more standardized than otherwise. These examples were drawn from a collection of Tableau training materials to ensure that at least the basic commands made familiar to the user during training are available via voice, as well as examples from initial testing. Most of these commands are generalizable to basic VA system interactions, such as selecting subsets of data or grouping objects. The wizard was provided with guidance on how to map example user utterances to actions, as multiple utterances may map to the same actions due to common synonyms. Domain-specific synonyms to the tasks were also provided to the wizard. The wizard was trained in how to respond to requests in an appropriate manner through Tableau training videos and some written guidance, which was tested and adjusted during testing sessions. The wizard observed all of the participants' interaction in order to monitor the current state of the Tableau system, including selected objects in order to disambiguate specific references (e.g., "Make this red").

4.9.2 Wizard Interface

In supporting the wizard in responding to request, the system had to prioritize speed. The first attempt was using AutoHotKey to create a separate GUI specifically for the wizard that included taking control of the Tableau interface, returning control, and sending error messages. The GUI included buttons for other functions found to be common during testing, such as adding sheets or switching rows and columns. In testing sessions, the GUI was found to be slower than operating Tableau directly, as it became more difficult to find the appropriate button for the appropriate function. Instead, simple response buttons were mapped to keyboard keys, and the wizard acted directly on the Tableau interface by request. The functions included taking control of the Tableau interface, returning the interface with a “Complete” confirmation, and returning the interface with an “Error” notification.

4.9.3 Facilitator / Wizard Experimental Setting

The researcher served as both the facilitator and the wizard for this experiment. The COVID-19 lockdown, along with limited resources, made it very difficult to recruit participants for this study and even more difficult to recruit outside assistance in running the experiment. The researcher had previous experience using Tableau professionally, though in no way an expert on every single function available. The wizard is meant to simulate a reasonably plausible future system, so it was not necessary to have a master certified Tableau professional but instead someone reasonably knowledgeable enough to respond to requests in a way that a potential future system could. This may mean a loss in objectivity when examining the performance of the wizard during the experiment, but every effort was made to present what was done in a comprehensive manner.

The facilitator monitored the participants’ actions through two separate systems. System 1 was the computer with Tableau Desktop installed along with all of the data sets. Participants remotely accessed System 1 using Google Remote Desktop through their local Google Chrome browser. System 1 recorded the participant’s screen and created videos of the interaction sessions using Open Broadcaster Software (OBS). Communication between the facilitator and the participants was done through Google Hangouts on System 2, which was also recorded the sessions using OBS. Necessary communication included pre-experiment explanations, in-between task set-ups, and the exit interview post-experiment.

The wizard in the voice system experimental condition managed communications through System 2 and responses to requests through System 1. Those participants who had access to the voice system spoke their requests through Google Hangouts, which were monitored by the wizard through System 2. The wizard monitored all participant interactions with Tableau, and then took control by request through a simple macro mapped to shortcut keys. Testing over the internet highlighted the need for two separate systems in order to improve system reliability and to ensure the participants had enough computer resources on System 1 to interact with Tableau without a large amount of lag. The facilitator/wizard workspace is shown in Figure 4. A representation of all of the systems and communication channels is shown in Figure 5.

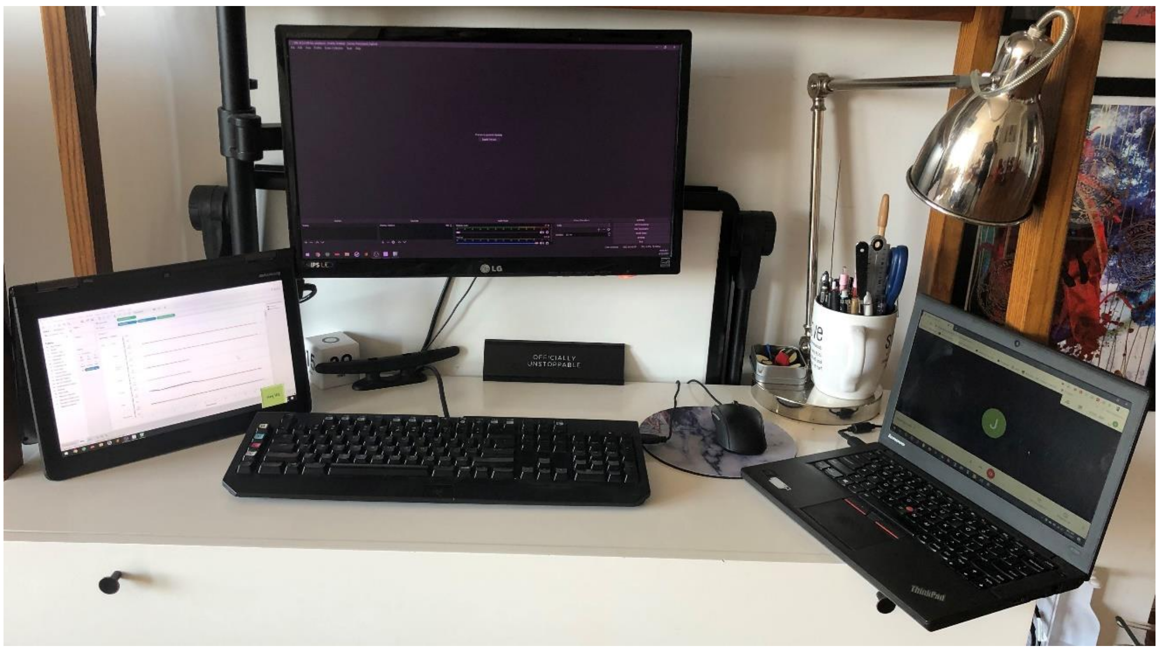


Figure 4 - Facilitator / Wizard Workspace.

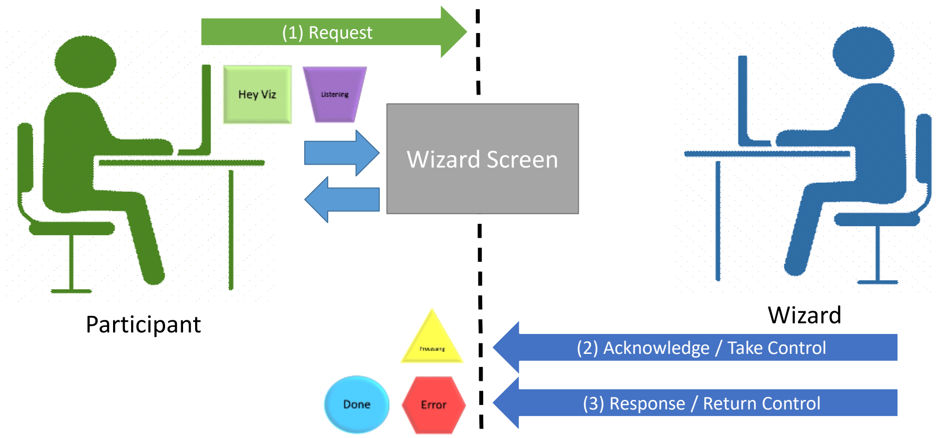


Figure 5 - System and Communication Channels

4.9.4 Wizard of Oz Guidelines

The Wizard of Oz experimental design is a very common method for testing systems that are potentially too costly to build or towards the edge of current technological capabilities. This involves having a human operator taking the place of an automated system. This is a common practice when developing systems that incorporate some form of human communication as a modality (Dahlbäck, Jönsson, & Ahrenberg, 1993). It is a frequent tool for testing voice interactions for VA systems and has been used previously in studies such as Aurisano et al. (2016), Kumar et al. (2016), and Tory and Setlur (2019). Tools such as (Srinivasan & Stasko, 2018) and FlowSense (Yu & Silva, 2019) are built using natural language parsers and mappings to both data commands and visualization requirements. A Wizard of Oz set up is for this study since the researcher questions concern the use of a potential voice system and not the technical aspects of its development.

Fraser and Gilbert (1991) provide common guidelines for the use of the Wizard of Oz techniques for simulating speech. Riek (2012) reviewed their guidance and found that the core first three tenets to be widely used in a review of Wizard of Oz studies, but the remaining points to be not widely adopted. Below is how each point pertains to this study.

A. Possible to simulate the future system, given human limitations – The development of voice modalities for VA systems is already underway, so the viability is certain for a future system to exist in some way. Systems like Orko (Srinivasan & Stasko, 2018) that are fully automated multimodal VA system that allows for natural language input demonstrate that these systems are viable and improving.

B. Possible to specify the future system's behavior – The wizard was provided guidance and examples on how to respond to user utterances, but ultimately had to exercise some judgment. The expectation is that future systems will be able to approximate the wizards' responses using increased data collection and probabilistic modeling.

C. Constrained Wizard recognition ability – The participant and the user are in separate locations, but due to the structure of the remote set up the wizard was able to hear and view everything done by participants.

D. Constrained Wizard production ability and Wizard errors – The wizard was expected to act in a predictable manner according to the guidance provided and only when given a request from the participant. The wizard was not allowed to be proactive and act on behalf of the participant until called upon.

E. Provided training to Wizards – The wizard was given guidance on how to react to example utterances. The same Tableau training materials that were seen by the participants were used as well as more advanced training materials, with additional examples of utterances to actions in Tableau. These materials were adjusted during testing to address issues that arose.

F. Possible to make the simulation convincing – The wizard made every effort to make the responses from the system seamless, aiming to keep delays at most between 30 to 60 seconds to stay in line with the wizard performance in Tory and Setlur (2019). Participants were told during training that the system's ability to recognize and react to commands depends on the complexity of their request. The system provided visual cues as to where in the process the system (wizard) is in responding to the request, which may reinforce the synthetic nature of the voice system.

4.10 Task Design

4.10.1 Training Design

The initial task T0 served as a training task and was a modified version of the Tableau tutorial "Getting Started" available on their website. Sections that train users on how to connect to data sources, export data, and build dashboards were removed since it is not relevant to this study. The chapter pertaining to trend lines was also be removed since it assumes some basic statistical knowledge of its users that this study does not. A list of sections and whether they were included in the study can be seen in Appendix F. The participants watched a training video during T0 and were provided a full transcript of the video along with appropriate screen captures. The video was broken apart into sections relating to the topic covered in order to be easier to navigate during training, which was found to be important to users during testing. They were also provided with an annotated screenshot of the Tableau interface and toolbar. The content of the training materials can be seen in Appendix G adjusted to fit the formatting of this dissertation. Participants in the training task were required to answer a question with a specific answer drawn from the training dataset. All tasks, including training, were considered complete once the user had submitted a written answer to the task prompt and then filled in the post-task survey.

For those users in the condition containing a voice modality, there was an additional section in the video and provided written materials on how to access the voice system using the "Hey Viz" button. The training emphasized that the voice system could access all of the functions covered in the training as well as all of the remaining functions available in Tableau. The video segment concerning the voice system was appended to the end of the Tableau training video and used the same training dataset. Examples were purposely drawn from the training material and the training dataset in order to provide a sense of continuity. A female voice not belonging to the researcher was used in the training video to avoid having the users connect the facilitator with the voice system. The training examples were mostly structured as “Show me …” requests, such as “Show me this graph over time” or “Show me Sales by Market as a Bubble Chart.” Mixed in were questions such as “Which Category had the most Sales in 2012?” and imperative statements such as “Show me this graph over time.” A mix of examples was given to encourage the users to form their own requests as they see fit. Testing sessions found that limiting the users to just using questions would make requests for certain functions more awkward (ex: “Can you delete this sheet?”) while not allowing for questions was overly constraining. Since this study is about maximizing the accessibility of functions, the choice was made to allow for any type of request. A reference sheet for the “Show Me” charting tool was provided since it would be the basis for how the voice system would build charts, so users would understand the necessary types of data for each chart before providing a request. Users were told that the voice system could handle a large range in type and complexity of requests, and would make every effort to answer whatever request was provided. Users were not told by the facilitator that the experiment was meant to focus on the voice system in order not to overly encourage its use, though all users were given written consent materials explaining the full scope of the study.

4.10.2 Experimental Task Type

The experimental tasks T1 and T2 were designed to encourage users to use the system to explore the data set utilizing whatever functions and modalities they have available and feel most appropriate. Keim et al. (2010) speak to the unusually challenging nature of evaluating VA systems “given the explorative nature of visual analytics, the wide range of user experience, the diversity of data sources and the actual tasks themselves” (Keim et al., 2010, p. 15). Selecting and operationalizing the study tasks and goals is particularly difficult since they can go beyond the typical Information Visualization (InfoVis) tasks of identifying correlations and outliers into much deeper work, often characterized as “connecting the dots” (Youn-ah Kang, Görg, & Stasko, 2011). Tory and Setlur (2019) find that highly structured tasks tend to bias the user’s choice of language when interacting with voice systems, so prefer open-ended task descriptions to incite user behavior at the cost of objective indicators of success. This study emphasized encouraging interaction over having an objectively correct answer, so the experimental tasks were made to be open-ended and exploratory in nature.

4.10.3 Task Length

The length of the task was intended to be long enough to elicit a large set of interactions while not being overly burdensome on the participants. If users are performing a cognitive task in performing these VA tasks and interacting with the system, then it is important to not harm their performance by overtaxing their energy. Testing found that two hours was more than adequate to finish the training and the tasks while not being overly long.

4.10.4 Task Domain

The experimental tasks were drawn from similar domains in order not to overly influence a difference in participants’ behavior. It also made controlling for the effect of domain expertise compared to inexperience with the system much more straightforward. Since the training materials produced by Tableau were all related to the business domain, it was sensible to have the experimental task be related to business as well. It seemed reasonable to assume those familiar with business tasks would also be more familiar with data analysis than the general public, further isolating the effect of the participants being novices with respect to the system versus novices to the task at hand. Participants’ prior experience with the task domain and other related topics was assessed in a pre-experiment survey.

4.10.5 Task Data Sets

The datasets for the tasks were required to be related to business topics and robust enough to elicit a large amount of interaction. The first task T1 used the same sales data set from T0 but did not provide step-by-step guidance on how to build the appropriate visualizations. The training dataset provided by Tableau was found appropriate for the experimental task during testing. Given the same dataset, participants were asked to provide recommendations to the company on how to allocate resources for the next year. The dataset included information on the sales of different types of goods, including sales price, costs, and delivery location. There was also an added benefit that participants felt more comfortable beginning their work since they had been introduced to the data during the training. Participants were required to provide at least one visualization that supports their argument, which was meant to encourage the participants to interact with the VA system with this purpose in mind. The task instructions given to participants can be found in Appendix C.

The aim of the second dataset was to stay in the business domain but to avoid having participants duplicate their work from the previous task. Another dataset involving profit and losses on goods shipped could have led participants to build the same visualizations, producing redundant interactions that would not further the aims of this study. Users were asked instead to examine data from a product coupon campaign over a two year period and to make recommendations on how to adjust the coupon strategy for the future. It was much larger and thus more challenging for users to interact with, which would encourage a new set of visualizations and interactions. At the same time, the task was in practice very similar to the Titanic dataset used in Tory and Setlur (2019), where participants explored aspects of an event that led to a binary outcome. In this case, instead of survivors of a disaster, the participants searched for reasons why coupons for a particular good were redeemed. The Titanic dataset was found to be appropriate for an exploratory visualization task, and similarly, this coupon dataset was found appropriate for this experiment during testing sessions.

4.11 Data Collection

The need to divide the experimental system into two parts to optimize computing and internet bandwidth meant that the data collection would also have to be split. System 1 had Tableau installed and where users logged into using Google Remote Access. The appropriate Google Form was loaded for each task in a Google Chrome tab to collect task answers and survey responses. The content of the pre-experiment and post-task questionnaires that each user filled out are shown in Appendix B and Appendix D, respectively. System 1 recorded videos of the interactions during the study using OBS. Each task resulted in a Tableau project containing the final visualizations that were saved on System 1. Recording keyboard strokes and mouse movement was found to slow down the system performance and to not be additionally informative to the screen recordings during testing sessions. Mouse movement was especially difficult to reconcile since users could move and resize windows in their desktop environment, which meant matching the action with the participant's target was especially difficult. The final system did not include these data collection methods and instead relied on the video screen recordings.

All communication between the participant and facilitator took place on System 2 through Google Hangouts, which allowed for both text chat and voice communications. Participants on their local system had one Google Chrome tab open to view System 1 and another to communicate through System 2. All utterances made by the user in the voice modality condition were recorded on System 2 using OBS and transcribed afterward for analysis. Users were asked questions during a semi-structured exit interview about their experiences using Tableau and performing the tasks. Users in the voice system conditions were explicitly asked about their choices of when to use the voice system and how they feel about the voice system.

Guiding questions for the semi-structured interview can be found in Appendix E.

4.12 Experiment Process

The following section describes the experiment from the participants' perspective step-by-step, from sign-up through closing and compensation. Twenty-four participants with little experience with the VA system Tableau were recruited and split into two experimental groups: one who had access to a simulated voice system, and another which did not. Both groups went through a training period and then attempted two exploratory data tasks. The experimental design details follow.

4.12.1 Participant Sign Up and Consent

This study was promoted through Rutgers University student email groups and online message boards. Participants were told that the study consisted of one session in which they would go through some training using a visualization system. They would then perform two tasks and answer some questions about their experiences. For completing the study, they would be compensated with a $40 Amazon gift card. Interested students who met the required criteria would then contact the researcher in order to schedule a session. An example recruitment email can be found in Appendix A. Upon confirming a date and time, users were given a consent form that explained the full scope and purpose of the research and then were required to confirm their understanding and voluntary participation through an online form. They were also sent a calendar invite via email that had a link to the Google Hangout at the appropriate time to begin the experiment.

4.12.2 Pre-Experiment Survey

Before beginning the experiment, participants filled out a pre-experiment survey in Google Forms sent to them as a link via email. The form collected basic demographic information, including age, gender, and college major. Participants also provided a Participant ID number that was provided to them via email upon signing up for the study. This was used in lieu of names or other personal information to connect the participants' data across data collection methods and tasks. This form also collected the participants' self-reported experience in domains related to the study, such as business analysis and other VA systems. All participants were required to have completed the survey before the experiment and were reminded of this the day prior to the session. The content of the survey can be seen in Appendix B.

4.12.3 Research Introduction and Training Task T0

Participants joined the facilitator in the Google Hangout chat and were reintroduced to the research study. Upon confirmation that they were prepared to move forward, they were sent a unique code to access the system via Google Remote Access. This code ensured the system remained secured once the participant had completed the study. The Google Hangout was left open throughout the experiment in cases where there was technical assistance needed and for voice communication for "Hey Viz" requests and the exit interview. An example Google Hangout screen from the facilitator's perspective on System 2 during the initial text chat with a participant is shown in Figure 6 with the name of the participant removed.

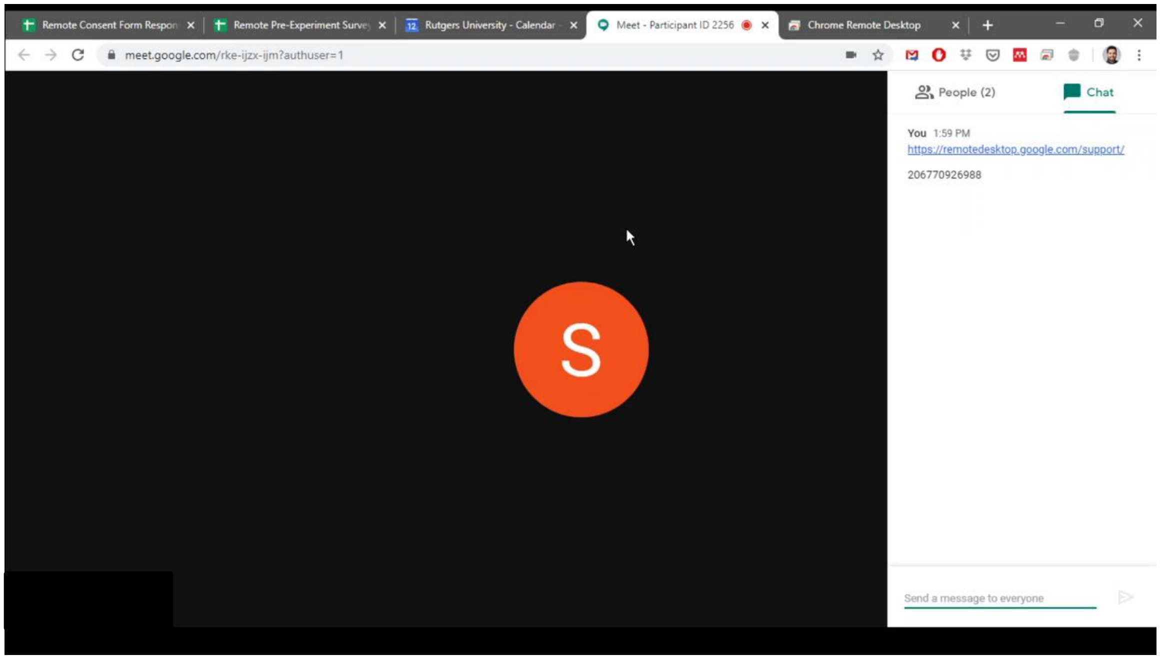


Figure 6 - Initial Google Hangout Text Chat

Participants were given a brief introduction to the remote system, which included a blank Tableau project with the training dataset loaded. Participants were then guided on how to adjust the settings in Google Remote access to view only the primary monitor, which had the Tableau system loaded and not secondary monitor, which the screen recording program OBS running. Having to view both 1920 x 1080 screens within a single Google Chrome tab would have made interaction almost impossible. This procedure was especially necessary for those participants who had access to the voice system as the secondary monitor served as the workspace for the wizard when responding to requests. Once the participant confirmed they were viewing only the primary monitor, the facilitator moved on with the introduction.

The primary monitor had a few necessary programs present for the training task. Tableau was loaded with the training dataset that matched the training video to allow participants to follow along. The appropriate training video was open using the VLC playback program, with those with access to “Hey Viz” having an additional chapter on how to access the voice system. All of the training materials and a description of the training task were open as PDFs, the content of which is in Appendix C and Appendix D, respectively. The final program open was Google Chrome, which had the appropriate Google Form loaded to collect the task answer and the post-task survey responses. Those who had access to the voice system were then told to run a necessary system check by clicking on the “Hey Viz” button and saying “Hey Viz testing” to make sure the voice system was running and responsive to their commands. In reality, the facilitator responded to the request by moving the Tableau screen off-screen briefly and then returning it with a “Complete” confirmation. This served as a way to reinforce the voice system as being automated as well as to remind those participants to keep their mic unmuted in order to use the voice system. Participants were then told they had complete and sole control over the system and to use the Google Hangout text chat if they required further assistance. They were instructed to use the training materials at their own pace and to answer the training task question and the post-task survey when they felt comfortable interacting with the Tableau system. An example opening screen from a training session can be found in Figure 7. This includes a blank Tableau project in full-screen mode with the training dataset loaded. The training video, training material PDFs, and Google Form are minimized here, but these windows could be moved around and resized at any time to fit the needs of the participant. The “Hey Viz” button is visible in the bottom right corner.

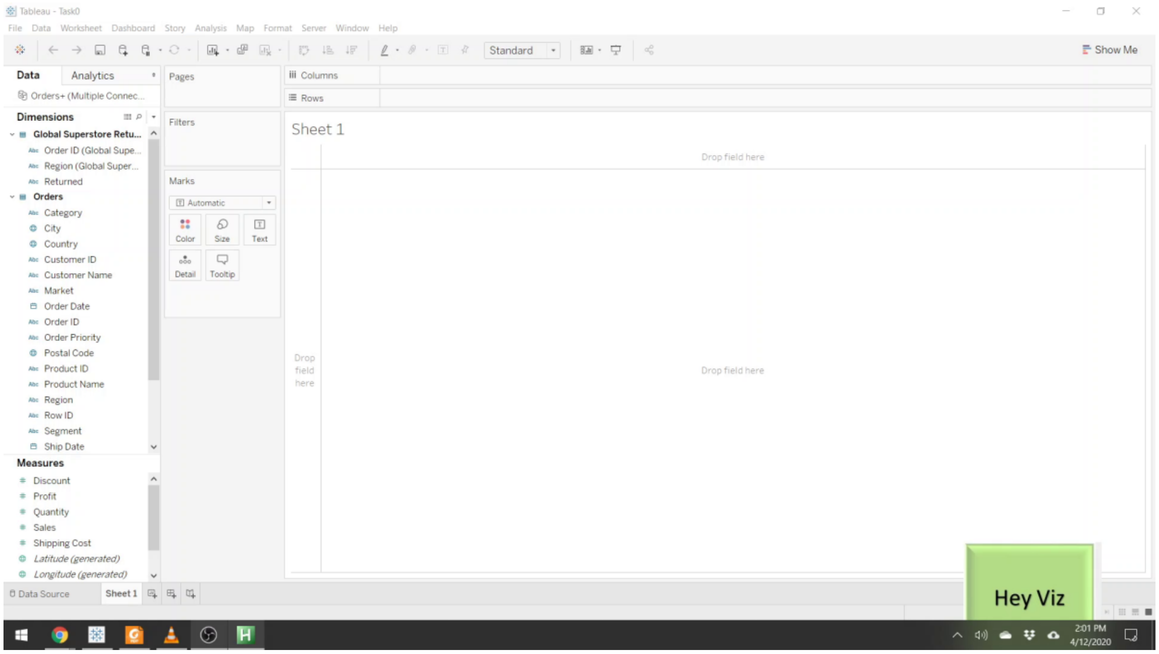


Figure 7 - Initial State of System 1 during Training

4.12.4 Experimental Tasks T1 and T2

The experimental tasks were similar in procedure to the training task but with different datasets and without access to the video training. At the beginning of each task, a blank Tableau workbook was opened that included the appropriate dataset but no pre-built visualizations. At the conclusion of each task, the participant had to submit an answer to the task question and then the post-task survey. Although the facilitator was monitoring the session the entire time, once the participant had submitted the required responses, the facilitator would purposely delay interacting with the participant for at least a minute. This was meant to give the illusion that the facilitator was not watching and was only monitoring some automated system of communication. The facilitator would then confirm with the participants that they had completed the task (“I just got the message that you submitted the survey. Are you ready for the next task?”) and inform the participant that they would temporarily retake control of the system to set up for the next task. The facilitator then saved and closed the Tableau workbook, opened a new Tableau blank workbook with the appropriate dataset, and opened the appropriate Google Form for the task answer and post-task survey responses. Participants had access to the full Tableau environment and could interact with the system in any way they felt appropriate, even in cases where functions fell outside of the training materials. No other programs were allowed to be opened or used during the sessions. Those in the voice system were again required to perform a test of the “Hey Viz” system and reminded to keep their microphone unmuted. The appropriate task description, along with the written training materials were opened, and the participants would then be told that they have control over the system again. The total task sessions lasted no more than two hours. The task materials are shown in Appendix C.

4.12.5 Post-Experiment Interview

After all of the tasks were completed, the participants were interviewed about their experience during the study. Appendix E shows the questions meant as guidance for the exit interview. They were meant to begin open-ended to collect the initial thoughts and reactions from participants about the system and the tasks. Topics were covered as they came up in discussion, but the facilitator made sure to cover the basic questions around functionality and interaction issues. Those who were provided access to the voice system were all asked about their experience using the system for reasons why they would or would not choose to use it. These interviews were recorded on System 2 using OBS and were later transcribed for analysis.

4.12.6 Closing Remarks and Compensation

After the participant had confirmed they had no further questions or comments about the study, they were then walked through the process of how they would be compensated. After the experiment, the facilitator would confirm that all of their materials were in order, including their consent form and pre-experiment survey. Once this was confirmed, the participant would be sent an email directly from Amazon with a redemption code for their $40 Amazon gift card. Participants were never told during the experiment that the voice system was a separate system from Tableau that was run by a human operator, but the full explanation was provided to them in the consent materials for them to keep at the completion of the study.

4.13 Summary

This chapter describes the methods used for this study, with an emphasis on implementing the voice system through a Wizard of Oz experimental setup. Users performed three tasks that involved using the visualization software Tableau, with the first being a training task and the other two exploratory tasks. Stratified random sampling split the business majors and the non-business majors into two experimental settings: a traditional desktop environment with a keyboard and mouse, and the same system with the addition of a simulated voice system. For this study, participants in the voice group were able to access both a simulated voice system or keyboard and mouse. Users could choose to use either modality in any given moment as they complete their exploratory task. A Wizard of Oz experimental design approximated the VA system's ability to understand and react to voice input. For this study, the general-purpose VA system Tableau Desktop Professional Edition v. 2020.1 was used. Interactions with and without access to the voice system were then compared. Users were also interviewed afterward to assess their attitudes toward the system and their interactions.

Chapter 5

Results

This chapter provides a detailed review of the data collected related to the research questions of this study using the methodology described previously, including sample demographics, interaction data, and hypothesis testing related to each research question.

5.1 Participant Demographics and Relevant Previous Experience

Participants answered a series of questions pertaining to their demographics and their previous experiences in relevant fields prior to the experiment. Half of the participants were business majors, while the rest majored in computer and engineering fields. Table 1 presents a breakdown of participants' demographics.

Age		
Mean	20.20	
Min	18.00	
Max	23.00	
Gender		
Male	11	46%
Female	13	54%
Native English Speaker		
Yes	22	92%
No	2	8%
Business Majors		
Finance and Accounting	5	42%
Business	4	33%
Supply Chain Managment	3	25%
Non-Business Majors		
Computer Science	7	58%
Engineering	5	42%

Table 1 – Participant Demographics

Mean participant responses pertaining to their previous experience by major can be found in Table 2 as well as the results of a one-way ANOVA. As expected, the business majors self-reported higher familiarity in fields related to business. The non-business majors reported higher familiarity with some visualization systems not mentioned in the survey. Python and MATLAB were the mentioned alternative visualizations tools during interviews. Excel was the area surveyed that almost all participants were familiar with across majors.

	Non-Business	Business	F	p-value
Data Analysis	2.83	3.17	0.58	0.45
Business Data Analysis*	1.92	3.33	9.60	0.01
Sales Data Analysis*	1.75	2.75	5.87	0.02
Marketing Research*	1.75	2.83	5.18	0.03
Microsoft Excel	3.75	4.17	0.82	0.37
Microsoft Power BI	1.42	1.92	1.66	0.21
Tableau	1.67	2.08	0.90	0.35
Other Visualization*	1.92	1.33	4.69	0.04

Table 2 – Pre-Experiment Survey by Major

5.2 Interaction Data Processing

Video recordings of the participants' sessions were transcribed to capture interactions with the Tableau system. An interaction was transcribed anytime a participant attempted to change the state of the Tableau system. These interactions included adding variables to a field, changing the graph type, adding a sheet, and scrolling. Attempts at changing the state of the Tableau system are actions that access a menu without performing an action. This included clicking on the top toolbar menu as well as right-clicking on objects within the interface, so these actions were collectively labeled as "clicks." The function used along with the target of that action, the sheet in which the action took place, and the state of the Tableau visualization in terms of rows, columns, and marks were recorded. Scrolling was recorded in terms of the number of seconds it was used continuously because a participant could scroll for a period of time. The other functions were discrete actions recorded and timestamped as they appeared in the video. A list of all of the functions recorded along with total counts can be seen in Appendix H.

5.3 Task Time

Task time is standardized here as the time between the users' first interaction with Tableau Desktop and their last during each task. Some participants took longer to look through the task materials provided than others, and others took longer to type their final responses to the task prompt. This study is concerned with the participants' interactions, so the time between the first and last interaction with Tableau Desktop was found to be the most appropriate basis of comparison. The distribution of task times is summarized in Table 3. The mixed ANOVA test conducted showed no significant difference between experimental groups, majors, or tasks.

	Mean	Median	F	p-value
Task 1	24.39	22.38	2.32	0.14
Task 2	19.89	18.03		
No Voice System	21.51	20.66	0.16	0.69
Voice System	22.77	21.01		
Non-Business	24.84	23.28	3.43	0.08
Business	19.44	17.21		

Table 3 – Distributions of Task Time (Minutes)

5.4 Use of the "Hey Viz" Voice System

The following section summarizes participants' use of the "Hey Viz" voice system by the number of requests and the type of requests given. The response time of the Wizard, who operated the system, is also reported.

5.4.1 Time Frames

The use of the "Hey Viz" system during the tasks was evenly spread out over the task time, with no apparent spikes or dips in use as a function of task time. Figure 8 shows the cumulative use of the "Hey Viz" system in terms of relative task time for the two experimental tasks. The graphs show a strong linear trend, with the R-squared of the best fit lines above 94%.

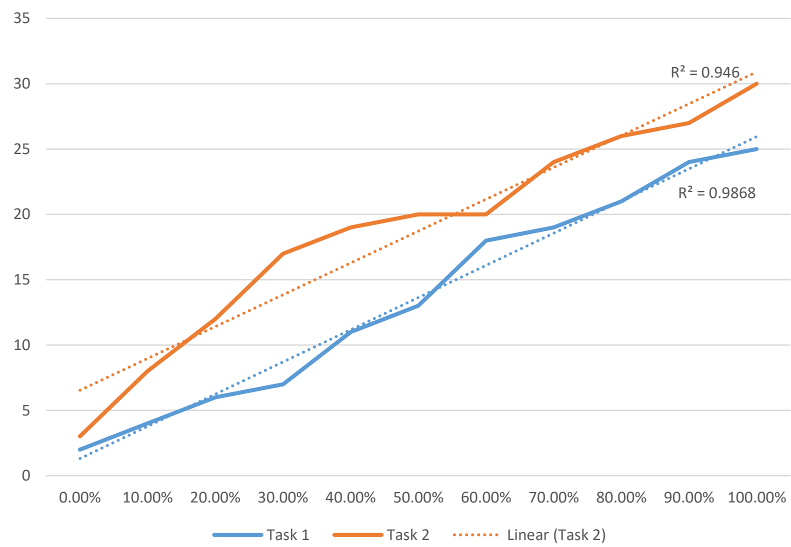


Figure 8 – Cumulative use of Hey Viz over relative Task Time

5.4.2 Requests

Participants provided 85 total queries to the "Hey Viz" voice system, with 55 queries occurring during the experimental tasks. The median amount of queries per user per task was 1. Three participants did not use the voice system at all during the experimental tasks, while one participant chose to use it a total of 27 times. The distribution of requests is shown in Figure 9.

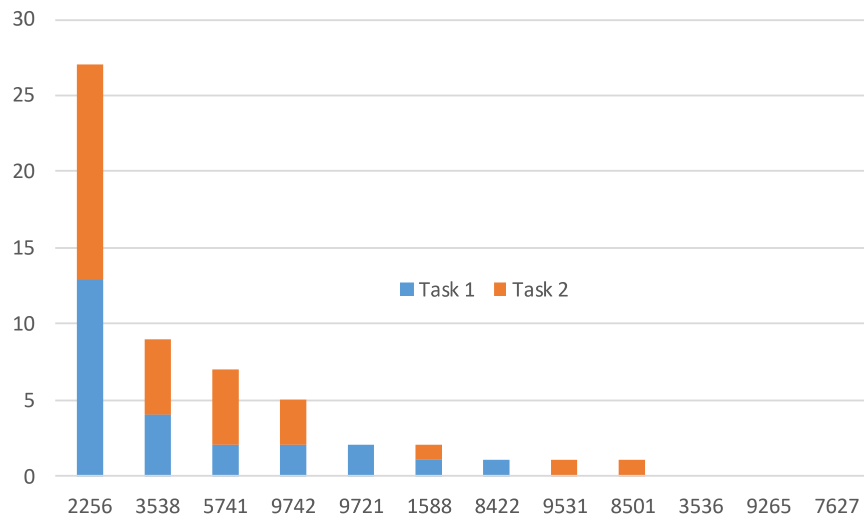


Figure 9 – Distribution of “Hey Viz” queries over Tasks by Participant ID

5.4.3 Types of Requests

Participant requests to the “Hey Viz” system had an average length of 7.41 words. 67% of requests during the experimental tasks took the form of “Show me…” followed by a list of variables, while the remaining requests took the form of imperative statements and one request in the form of a question. 82% of requests had explicit references to dimensions or measures provided to the participant. Setlur et al. (2016) developed query types for their natural language interface for a VA system based on the over 1,500 queries collected. The requests in this study mostly fit those categories, with some requests overlapping categories. This is a function of participants being able to requests multiple interactions and transformations of the data within one request. The count of requests by these categories can be seen in Table 4. The majority of requests involved comparing variables and displaying relationships, with an additional seven requests for comparing variables with additional filtering, sorting, or creating a custom metric.

Compare	26	47.27%
Filter	14	25.45%
Compare and Calc & Stats	4	7.27%
Calc & Stats	3	5.45%
Compare and Sort	2	3.64%
Analytics and Trends	2	3.64%
Formatting	2	3.64%
Filter and Calc & Stats	1	1.82%
Compare and Filter	1	1.82%
Total	**55**	

Table 4 – Type of Requests from Experimental Tasks

5.5 Explanations for Voice System Usage

Each participant was asked about their reasons for choosing or not choosing the voice system during the post-experiment exit interview. Those instances where a participant did choose to use the voice system were rare compared to the keyboard and mouse interactions, so recalling the particular reasons why it was utilized was not difficult for this group. Their responses are grouped here by their location within this study's model of how novice users would utilize a voice modality for a VA system, and specifically the following three aspects: modality preference, accessibility of functions, and mental model of the system.

5.5.1 Modality Preference

The first aspect of the interaction model that affected how participants used the voice system was the novelty of a voice modality with a desktop VA system. All participants stated that this was their first time using a voice system in a desktop environment. For some, the addition of a voice system meant an added convenience, while for others, it was strange and off-putting. Some expressed that the novelty of the voice system meant they were generally uncomfortable interacting with it. A few participants mentioned that they generally do not prefer voice as a modality, as one participant mentioned that they usually work in quiet spaces and that they generally do not use "voice-activated things." One participant never used the voice system due to their discomfort, and another stated although they used the system and found it useful that they were unlikely to use it in their daily work. On the other hand, some participants were more interested in the voice system, with one participant deciding to use the voice system because it was novel and thought "it would be cool to use it once." The most frequent term used by participants when describing their experience was convenient, as speaking to the system was viewed as much easier than having to use a keyboard and mouse in some situations. One participant described a general preference for using the voice system because it was so easy to speak, and "[i]f I was too lazy to look for it in the actual panel I would just click the voice button and then just say it." A participant described this preference in a situation where there were multiple items involved, and so "it's easier to just say four things tan to drag four things." Speaking to the system was similar to having communication with another human, and for these participants, it was "easier to ask someone to do it and then see it instead of doing it yourself."

Many participants appreciated the convenience of using the voice system in order to get a start on their task compared to using a keyboard and mouse. Forming their initial request to the system was easy, and this allowed participants to feel confident when beginning each of the experimental tasks. This was described by multiple participants as "setting up everything," getting "all the base stuff," "[giving] me a strong foundation," and "[putting] me in the right direction." They then used the keyboard and mouse to interact with the output provided by the voice system, and very rarely would use the voice system to further interact immediately afterward.

The preference for keyboard and mouse interactions was prevalent across participants and situations since the novelty of using a voice system also meant that participants tended to default to keyboard and mouse for most interactions. One participant focused entirely on the keyboard and mouse functions available and forgot about the availability of the voice system, which was discussed with regret afterward. Even when they knew that the voice system was available, most participants would focus on the traditional interactions because "I was already used to using the computer and the keyboard so I think I kept going back to what I already knew versus just trying new things." This tendency towards keyboard and mouse was especially true for searching for functionalities, as participants expressed a clear preference for exploring the visual interface to search and discover functions. All participants expressed that they had a particular function or outcome in mind before using the voice system, and did not consider prodding the voice system with vague requests or offering a range of requests to discover new functions. The system was easy to "[tell] her to do it for me," but only when the exact outcome was already known. One participant described this situation in terms of their own understanding of the task, and their lack of use was due to not knowing how to move forward, which led to "just doing my own thing" and clicking around the interface.

5.5.2 Accessibility of Functions

The interaction model for this study predicts that these novice users would choose to use the voice system in order to access functions that were difficult to find, and that reasoning was confirmed by every participant who used the "Hey Viz" system. Participants would use the voice system if they could not access a specific Tableau function in some cases, but not necessarily every time. The participants mostly used the traditional keyboard and mouse interactions, but in some situations when there seemed to be no alternative "it was more like I don't know how to do this, might as well try and ask the robot to do it for me." Participants described having a clear outcome in mind but a lack of knowledge on how to achieve it using the visual interface. Many connected their inability to find their desired functionality with their inexperience with the system, such as one participant who stated: "I used it because I didn't know how to use like I wasn't familiar with like how to use the interface yet so I thought it would be better if I used the voice thing." This lack of knowledge occurred most often when trying to transform data into ratios and other calculations, which was a common problem during the tasks across all participants.

Despite the difficulties many participants had in finding certain interactions, the interface of Tableau and the functionality provided were spoken about positively by almost all participants. Though there were mentions of being overwhelmed at times by the interface, most participants blamed their lack of experience with using Tableau for their trouble discovering desired functions and not the design of the interface itself. Using voice support was helpful when the participant had the desired outcome in mind. However, in cases where the participants were unsure of exactly what they wanted and struggled to find it using keyboard and mouse, all participants expressed a preference for continuing to explore the visual interface rather than possibly receiving an unhelpful output from the voice system.

5.5.3 Mental Model of the System

The final component of the interaction model was the mental models that the participants had for using the system. The model predicts that the users will use the voice system to access function, but this is dependent on them believing that the voice system could achieve these goals. Although this was the first time using a voice system in a desktop environment, all participants had previously used systems like Amazon's Alexa and Apple's Siri, which then affected their attitudes towards "Hey Viz." Those systems were seen as having limited capabilities in comparison to interacting using traditional modalities and require very specific structure and content in their requests. When asked to describe the capabilities of the system, one participant described the system was limited to the extent that "if I couldn't do it, I didn't think the voice could do it." Another stated that they would restrict requests to the voice system only to simple functions and concerning outputs that they could verify themselves through manual interaction. Despite being told during training that the system was capable of performing a wide range of functions and responding to a wide range of requests, participants still defaulted to their initial mental model of voice systems that assumed that they were not suitable for certain tasks.

The model for this study assumes that the participants are focused on accessing functions, but these participants valued the knowledge on how to access functions using keyboard and mouse almost as much as accessing them. Participants who spoke about the convenience of the voice system also expressed concern about overusing “Hey Viz” to the detriment of learning how to use the traditional interactions in Tableau. Thus working towards building a mental model of how to interact with the Tableau system was often preferred over using the voice system. This occurred even when participants knew they would be able to with “Hey Viz” to access the function. One participant felt they had used the system “too liberally” at one point and chose to refrain from further use even when facing difficulty accessing particular functions. Two participants mentioned that they would prefer a voice system that showed them how to access functions rather than completing them. In one situation, a participant used voice to create a forecast of sales but then became frustrated when trying to use a forecast in a different visualization and having no choice but to use the voice system. That participant expressed a strong attitude towards “building something myself” and felt forced to use the voice system in that circumstance.

5.6 Function Choices

5.6.1 Most Common Functions

A total of 91 functions were recorded from the experimental tasks. 90% of these functions were covered in the tutorial, while those functions outside of the scope of the tutorial comprised approximately one percent of total interactions. The functions were grouped into 45 categories, with the ten most frequently used shown in Table 5. Scrolling is excluded because it was recorded as a function of time and not recorded as discrete actions like the other functions. Using the “Hey Viz” system is excluded from function counts because it was not used as a discrete function in the same sense of adding a filter but instead resulted in other functions being performed.

Function	Task 1		Task 2		Total	
variable_add	303	19%	389	23%	692	21%
click	169	11%	261	15%	430	13%
variable_remove	121	8%	183	11%	304	9%
show_me_chart	137	9%	100	6%	237	7%
variable_move	97	6%	132	8%	229	7%
sheet_select	118	8%	104	6%	222	7%
filter	44	3%	67	4%	111	3%
sheet_add	45	3%	49	3%	94	3%
undo_redo	29	2%	52	3%	81	2%
variable_sort	20	1%	32	2%	52	2%
Other	239	15%	172	10%	411	13%
Grand Total	**1,561**		**1,713**		**3,274**	

Table 5 - Frequency of Ten Most Used Functions

5.6.2 Number of Functions Chosen

Participants used an average of 9.98 unique functions per task. There was no significant difference between tasks ($F(1, 22) = 0.00$, $p = 0.96$) or between experimental groups ($F(1, 22) = 0.56$, $p = 0.46$). The average total functions used per task was 60.79, there was also no significant difference between tasks ($F(1, 22) = 0.81$, $p = 0.38$) or experimental groups ($F(1, 22) = 0.73$, $p = 0.40$).

5.6.3 Differences in Functions Chosen

The number of instances a function was chosen per participant per task was compared between the two experimental groups using a mixed ANOVA statistical model. There were no functions that differed significantly between the two experimental groups. “variable_add” was used significantly less by the voice group only if the actions of the wizard were not taking into account; otherwise, there was no difference, as shown graphically in Figure 10. This indicates a shift in modality choice, but not an expansion or contraction in function usage. A full summary of the average use per task of the participants and the results of mixed ANOVA testing can be viewed in Appendix I.

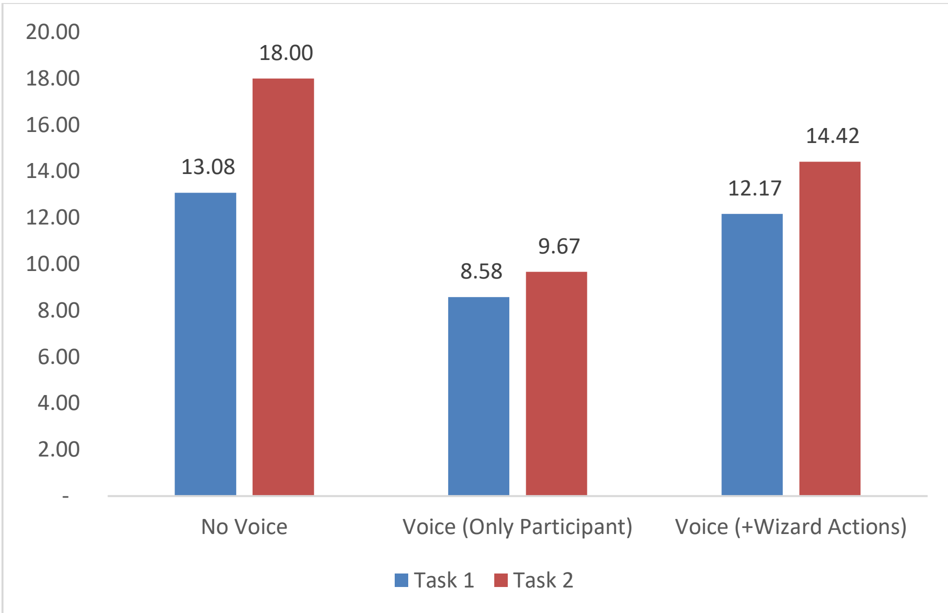


Figure 10 - Distribution of "variable_add" with and without Wizard Functions

The number “clicks” viewing underlying menu items was also not different between the two experimental groups. “Clicking” was recorded anytime a participant accessed a menu of an object to view options without actually selecting any. This occurred when right-clicking on variables or other items in the workspace, as well as when clicking the menu at the top of the Tableau interface. This behavior coincides with the strategy participants described when they were searching for particular functionalities in the visual interface. There was no significant difference between the experimental groups, meaning having the voice system available did not lower the amount of searching for functions done in the visual interface.

Scrolling was recorded on a per-second basis and was not found to be significantly different between experimental groups. Participants did scroll more during Task 2 compared to Task 1, which could be related to the difference in data set sizes. The mean number of scrolls per task and experimental type, along with the results of a mixed ANOVA with the post hoc pairwise t-tests, are presented in Table 6.

	No Voice	Voice
Task 1	18.83	47.17
Task 2	52.08	94.92

	SS	DF1	DF2	MS	F	p-value
VOICE	15194.08	1	22	15194.08	3.68	0.07
Task	19683.00	1	22	19683.00	5.85	0.02
Interaction	630.75	1	22	630.75	0.19	0.66

	T	df	p-value	Cohen's d
Task	-2.46	23.0	0.02	-0.65

Table 6 - Mean Scrolling Per Task Per Experimental Group and Mixed ANOVA Model

5.6.4 Rare Filters and Functions

Though there were no differences found in the overall use of functions, there were some functions accessed through "Hey Viz" that were not accessed or rarely accessed by mouse and keyboard, as listed in Table 7. These functions pertained to transforming the underlying data of the visualizations, including creating ratios, percentages, and forecasts. Additionally, some voice participants were able to use more complex filtering compared to those with only a keyboard and mouse. These filters were mostly numerical sets (ex: Profit higher than $6000), with two other requests for filtering over multiple categories to remove Null values.

Function	Voice	Non-Voice
Number Filters	4	0
forecast	2	3
calculated_field	2	6
Complex Category Filters	2	0
table_calculation	1	0
percent_of	1	1

Table 7 – Rarely Used Functions and Filters

5.7 Post-Task Survey

Participants were asked after the completion of each task to answer some questions about their experiences. Average responses are shown for each question in

Figure 11. Six questions were significantly different per task, with a breakdown of the average responses and Cohen's d from post hoc analysis are presented in

Table 8. Most of the differences reported were due to the task, as participants reported less satisfaction and confidence in their work for Task 2 compared to Task 1. Whereas Task 1 was an extension of the training task the participants were familiar with, Task 2 used a larger data set and concerned coupon conversion rates instead of sales. Similarly, the tutorial and task instructions received for Task 2 lowered marks. Though this would indicate an increased level of difficulty encountered by participants while interacting with Tableau, judgments on the quality of the interface and the users' own ability to navigate it were consistent across tasks and experimental groups. However, there was an increased level of agreement with the statement, "I felt limited in what I could or could not do with the system" that was found to be significant only between the training task and Task 2.

System response time was the only question where those in the voice group responded significantly differently from those without access to that system, as the voice group reported a higher level of agreement with the statement "The system took longer than expected to respond," though on average still not enough to change from disagreeing to agreeing with the statement.

Responses per group and task and post hoc analysis are presented in Figure 12.

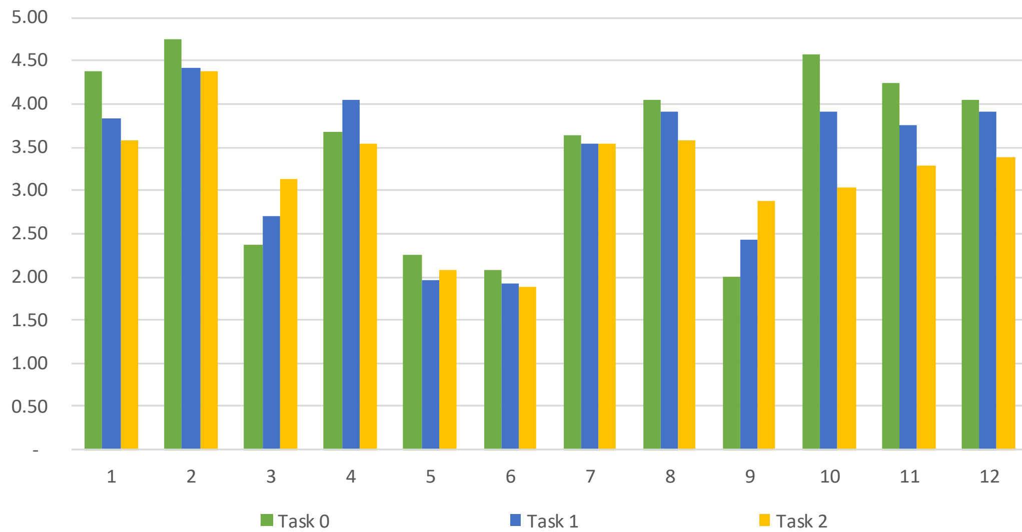
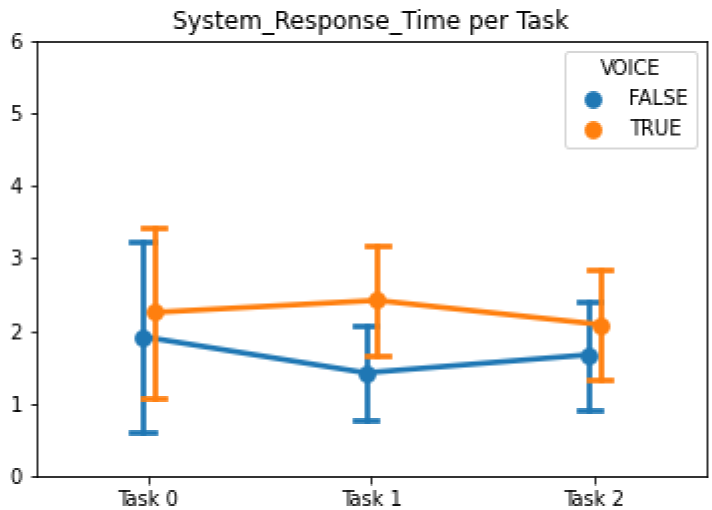


Figure 11 – Average Post-Task Survey Responses

		Task 0	Task 1	Task 2	T0 vs. T1	T0 vs. T2	T1 vs. T2
1	The tutorial provided the appropriate level of	4.38	3.83	3.58	0.77	0.90	-
2	I understood the task instructions.	4.75	4.42	4.38	0.56	0.50	-
9	I felt limited in what I could or could not do with the system.	2.00	2.42	2.88	-	(0.69)	-
10	I am confident that my answer to the task question is correct.	4.58	3.92	3.04	0.85	1.60	0.80
11	I am satisfied with the final product I produced for the task.	4.25	3.75	3.29	0.51	0.87	0.38
12	I feel confident using the system to complete another	4.04	3.92	3.38	-	0.60	0.47

Table 8 – Average Response to Post-Task Survey Questions that Differed Significantly Different per Task and Cohen's d Effect Sizes



		Voice	No Voice	Cohen's d
6	The system took longer than expected to respond.	2.25	1.96	(0.86)

Figure 12 – Average Response to Question concerning System Response Time by Group

5.8 Task Answers and Final Visualizations

Each participant was required to provide a written description of their answer to each of the task questions along with visualizations that support their argument. They were not required to specifically identify the relevant visualization but rather summarize their findings. There were no judgments on the accuracy or quality of the final products since there are no objective measures for their quality in this study (see Section 7.5 for further discussion). The average number of words for the task answers was approximately 89 words. The averages for the different groups per task are shown in **Error! Reference source not found.** A mixed ANOVA model found no s ignificant difference between tasks and experimental groups, as reported in

Table 9. A summary of the final visualizations is presented in Table 10, which also found no difference between the two experimental groups in terms of the number of graphs, the number of variables used, or in the usage of certain graph types.

	No Voice	Voice				
Task 1	84.50	87.58				
Task 2	74.58	110.08				
	SS	**DF1**	**DF2**	**MS**	**F**	**p-value**
Voice	4466.02	1	22	4466.02	0.70	0.41
Task	475.02	1	22	475.02	0.63	0.44
Interaction	3152.52	1	22	3152.52	4.20	0.05

Table 9 – Mean Number of Words per Task and Experimental Group and Mixed ANOVA

	T1	**T2**
Number of Graphs	2.17	2.38
Number of Variables	8.50	11.17
Frequency of Graph by Type	**T1**	**T2**
Bar	24	36
Table	10	16
Line	10	1
Map	4	0
Tree	0	2
Other	4	2
Total	**52**	**57**

Table 10 - Summary of Final Visualizations

5.9 Wizard Performance

5.9.1 Wizard Response and Performance

The average response time of the wizard to "Hey Viz" requests as measured from the time the participant clicked on the button until the participant had the Tableau system returned to their control was 22.60 seconds. The response time varied by request type, as simple functions ("Clear Sheet") required less action on behalf of the wizard compared to more complex requests ("Show me shipping costs per country relative to the amount of orders"). The requests were categorized using the groups provided by Setlur et al., (2016) in their study of a natural language system for VA. The average Wizard response time is listed per category in Table 11.

Request Type	Avg. Response (Seconds)	Number of Requests
Compare and Calc & Stats	34.50	4
Compare and Sort	30.50	2
Compare and Filter	29.00	1
Filter and Calc & Stats	28.00	1
Compare	23.19	26
Calc & Stats	21.67	3
Analytics and Trends	20.50	2
Filter	19.21	14
Formatting	4.50	2
Average Across Types	**22.60**	**55**

Table 11 – Average Response Time by Request Type

5.9.2 Reflection on Serving as the Wizard

There were two unexpected experiences from serving as the wizard of the study. The first was that given the amount of preparation, there was a lack of complicated requests to the voice system. The model for this study emphasized offering every function available in Tableau, so a lot of effort was made in gathering more obscure training examples from advanced Tableau tutorials and from volunteer testers. During testing, participants provided a wide range of examples, with many of them being purposefully complex and ambiguous in order to test edge cases for this system. This was important to forming many of the rules of how the wizard would respond to difficult requests. Formalizing how the wizard should handle a request to undo an action is an example. One could simply assume the participant's intent is equivalent to hitting the undo button, but this may not be the case if the last action was a request. In that case, the expectation was to undo all of the actions done by the voice system previously. In general, the volunteers who were testing the voice system offered more complex requests because they knew that this was meant to be a testing session for a future experiment. The actual requests during the experiment were much simpler in comparison, which made accuracy and speed much more critical. A complicated request may be expected to take a system longer to breakdown the language and respond to, but a short request like "Show me profit by year" is expected to have a quick turnaround. As the wizard, there was increased pressure to respond quickly to these simple requests. In the actual experiment, one feels much more pressure to perform even simple actions like selecting a variable in the side panel, and this can be detrimental to your mental and physical sharpness in performance. The wizard had to really focus on the words spoken by the participant and interpret the request according to the guidance. One did not want the participants not to choose the voice system simply because it was too slow, which would be a product of the wizard design and performance instead of a true measure of their modality preference.

The pressure to perform was not just in reacting quickly and accurately to requests but in trying to simulate being the automated voice system. Pretending to be a machine did not end with quickly clicking on buttons, but extended into creating a full experience for these participants. It began with the fake tests of the voice systems, where participants would click on the "Hey Viz" button and say "Hey Viz testing," which would result in a quick run of the Listening, Processing, Complete notifications. How an actual automated system would respond to requests had to always be kept in mind, though there are only a few examples where it affected the interaction directly. One participant requested during the training task that a sheet be renamed "crosstab," but perhaps due to a loss of train of thought literally said "rename sheet sheet sheet crosstab." The intent was clear, but in order to keep the illusion of an automated system, the wizard renamed the sheet "sheet sheet crosstab." That particular participant began making requests to the system almost immediately during training and used the voice system the most often throughout the experiment. A particularly challenging request involved filtering out a subset of regions from a list being displayed. They were being listed too quickly to write down, so it played out like an impromptu memory game. An actual automated system will likely not have as much of an issue keeping a list in its short-term memory the way a person does. It would have been helpful to have had the voice requests turned into text visible to the wizard so as not to rely so much on memory, but that would have added a barrier between the participant and the wizard that would have possibly caused more variation due to transcription issues.

The second issue that arose in serving as the wizard relating to monitoring this particular set of users in this study for a prolonged period of time. There was a general unease in being the wizard for those who had access to the voice system. Aside from the pressure to perform as an automated system, there were large periods of boredom and frustration during the experiments. The wizard had a responsibility to monitor and observe every interaction made by the participant, but watching someone use a system for up to two hours at a time requires a high level of endurance to keep focused. It did not feel like a passive observation, as the added pressure to respond quickly to a request at any moment added some anxiety to the work.

The relative lack of use of the system by these participants meant there were long periods of this intense and often boring observation time, broken apart by spurts of high pressure and frustration. The frustration was caused by situations where it was apparent to the wizard that the participant was looking for a certain function or in general need of assistance and yet did not use the voice system. Behavior such as clicking on different menus, undoing and redoing actions, and endless scrolling signaled to the wizard that the participant was struggling with something during the task. The wizard did not proactively engage with the participant at any point, but it was difficult observing participants who seemed in need of an intervention but not using a potential resource available.

Members of a research team are generally expected to behave objectively, but that did not mean that they make those feelings of frustration any less difficult to manage. It was fairly easy observing those participants who did not have access to the voice system, as the facilitator took a passive role and just made notes and monitored the text chat when necessary. It is possible that having the researcher serve as the wizard made these feelings much stronger than what would have been felt by a third-party. One could argue that anyone watching a two-hour interactive session with a high level of focus would likely struggle with some of these feelings, especially as in the case of three participants where the wizard was never used. These issues are little discussed in the literature reviewing the Wizard of Oz as a method, perhaps because it is usually assumed that the Wizard will be in high demand during the experiment. It would be beneficial to create better guidance in handling these situations as more studies explore given participants choices in interacting with systems instead of constraining them to use the wizard, such as perhaps recommending multiple wizards to relieve some of the strain. That was not done in this study due to resource restraints and may have caused some more variability in the wizard performance, but it could have helped the wizard's well-being a great deal.

5.10 Summary

This chapter describes the results of the study as they pertain to the circumstances under which novice users choose to use the voice system when interacting with a VA system and whether those users would choose different functions compared to a control group. A comparison of the functions chosen showed no significant difference between the two groups, although there were some functions that were used by the voice group rarely or never used by the control group. An analysis of the requests made by participants found that they did not tend to occur at any particular point during tasks and that most requests involved selecting and comparing variables from the data set. When viewing the feedback from those who had access to the voice system through the interaction model of this study, voice modality was convenient for some while unattractive to others. In some cases, participants did use the system for accessing difficult to find functions but mostly preferred to learn how to use the VA system using a keyboard and mouse. The post-task surveys revealed that having access to the voice system only significantly affected participants' feelings towards the response time of the system, and did not affect other factors such as their confidence in their final product or in using the system in the future.

Chapter 6

Conclusions

The main motivation of the work was to understand under what circumstances a novice user of a VA system would use a voice system, and if that choice would then lead to a different set of interactions. There were calls for an increase in the development of multimodal interfaces for VA systems, with the presumption that users would choose these new modalities to better address their task needs in terms of functionality and to close the "gulf of execution." Previous work focused on developing and testing new systems, with less consideration for whether or not users would actually choose these non-traditional interaction methods if given a choice. This study focused on novice users, who are a key demographic for developing new multimodal systems to improve their access to functionality, and on voice as the novel modality for VA systems.

6.1 RQ1: Under What Circumstances Do Novice Users of VA Systems Choose to Use a Voice System When Given the Option Within a Traditional Desktop Computer Environment?

The following use cases for the "Hey Viz" system emerged from an analysis of participants' responses during the post-experiment exit interviews. Participants used the voice system a limited amount times during the experiment, which meant they were able to better recollect their choices since those instances tended to stand out during their use of Tableau. Drawing from the discussion in Section 5.5 concerning the participants' attitudes towards using the voice system in relation to the interaction model of this study, the circumstances in which these novice users chose to use the "Hey Viz" voice system are summarized below.

6.1.1 Reasons Given by Participants for Using the Voice System

- **General Convenience**

The participants that used voice expressed satisfaction with the voice system and found it useful in interacting with Tableau. Speed was a key factor, as participants mentioned how quickly specific outcomes came about using voice versus manual interaction, such as selecting multiple categories because “it’s easier to just say four things than to drag four things.” A few mentioned being novice users of Tableau and how it is “easier to ask someone to do it and then see it instead of doing it yourself.” All participants mentioned being able to expect or picture an outcome before using the voice system and how it was easier in some cases to just speak about something versus manually creating a visualization, especially as a novice user.

- **Getting Started**

The voice system was often used to create a starting point for further analysis and manual interactions. Five of the participants who used the voice system said that they specifically preferred using it for initial data selection before manually building the visualization. Participants described this process as “setting up everything,” getting “all the base stuff,” “[giving] me a strong foundation,” and “[putting] me in the right direction.” This coincides with the content of the requests given to the system, as most of the requests including direct references to variables in the data set and the majority of requests being for visualizing comparisons. Participants were then more comfortable as a next step to interact using the keyboard and mouse once the voice system had provided the initial visualization, as opposed to creating a sequence of voice requests to further adjust their visualizations.

- **Being Unable to do Something Specific**

The other key circumstance in choosing the voice system was when participants were aware of a particular function but were unable to interact in a way with the system to create the desired outcomes. Participants were able to identify specific graphs or indicators but were unable to reconcile those with the visualization using their mouse and keyboard. The voice system in these cases was the fallback option to mouse and keyboard, as one participant phrased it "I don't know how to do this, might as well try and ask the robot to do it for me." Most participants struggled with transforming the underlying visualization data, such as creating averages or ratios. Four of the voice system participants mentioned how they wanted to transform the data values shown into percentages such as redemption as a percentage of total coupons redeemed or ratios such as profit to sales but did not know how to using keyboard and mouse. This was a common difficulty across all participants, as one participant in the non-voice group admitted to using a calculator offline after being unable to transform the data into percentages.

Participants were sometimes made aware of the capabilities of Tableau through navigating the visual interface but were unsure how to use a particular function. In one case, a participant found the forecast function under the "Data Analysis" tab but could not create the desired forecast data until they used the "Hey Viz" system. Afterward, the participant still could not access the function outside of the voice system and expressed a preference for the system to demonstrate how to access the functionality over just performing it. Participants also wanted to use more complex filtering, such as only showing categories by the top N percent or filtering out Null categories. They found the voice system was the most efficient way to perform these actions.

6.1.2 Reasons Given by Participants for Avoiding Using the Voice System

- **Believing the System Could Not Accomplish the Goal**

One mitigating factor in the use of the voice system was whether or not the participant believed that the voice system could accomplish the desired goal. Two participants went as far as to say that they doubted the system could perform certain functions when they themselves were having difficulty in doing so because "if I couldn't do it, I didn't think the voice could do it." Despite this being their first experience using Tableau, they still believed that their capacity to interact with the system exceeded that of the integrated voice system. Users would then sometimes not choose voice even when struggling to interact with Tableau.

Many participants discussed their previous experiences with voice systems and how those had set their expectations low for voice systems in general. Amazon's Alexa and Apple's Siri assistants were mentioned as examples of voice systems that the participants had used previously and were very limited in their capabilities. Participants expressed general surprise at the quality of the voice system once they had offered a request, which then exceeded their initial low expectations. One participant mentioned how they assumed that the system could only respond to simple requests because of negative experiences with Siri, and therefore was not willing to attempt more complex requests. One participant mentioned their lack of trust in voice systems meant that they only provided requests to "Hey Viz" that they verify themselves, even if though they understood the system could provide much more complex output. Participants' previous negative experiences with voice systems also affected how participants formed requests, as experiences with other systems had taught them "[y]ou have to tell it very specifically exactly what you're [going to] do." The lowered expectations of the voice system, therefore, affected the quantity and complexity of requests.

- **Negative Feelings Towards Using Voice Systems**

Participants appreciated the convenience of using the voice system in comparison to the keyboard and mouse but felt that using the voice system was incongruent with learning Tableau. Every participant who used the voice system reported it as being helpful for the task, especially in accessing functions in a convenient manner. However, that convenience often was spoken about in negative terms such as laziness. Participants were particularly adamant about not relying on the voice system to interact with Tableau and sometimes avoided voice as an unnecessary crutch during interactions. Participants mentioned that the risk of over-reliance is particularly high when users are novices and that the use of the voice system would be an impediment to learning the software. The fear of over-reliance on the voice system pushed many participants to navigate the visual interface by trial and error and to struggle with the task more than necessary.

- **Common Issues with Voice as a Modality**

Using voice in the desktop environment offered different challenges compared to traditional keyboard and mouse interactions in a WIMP environment. The three participants who abstained from using voice during the experimental tasks all mentioned that they had forgotten that the voice system was available and that they were more focused on the task and using the Tableau interface than on trying the “Hey Viz” system. Though all participants were introduced to the voice system during training and reminded of its existence before each task and a large green “Hey Viz” button was presented prominently, participants still found it difficult to remember the voice option was available. This was a problem even when participants could not find their desired functionality using mouse and keyboard. One participant described being unable to find a few different specific functions, including building ratios and drill-down data views, and regretted that they had not remembered to use the voice system.

Using voice in a desktop environment was unfamiliar and sometimes uncomfortable for many participants. None of the participants had previously used a voice system in a desktop environment, and all mentioned that they were more used to and more comfortable with using the keyboard and mouse. Some reported general unease when using their voice, while two participants said that they were very uncomfortable with speaking out loud to a system because they work in quiet spaces where they would be causing a disturbance. Both reported that they would almost never use a voice system for this type of work in the future regardless of its usefulness or effectiveness.

6.2 RQ2: Does Offering a Voice Modality Change the Functions Chosen by Novice Users?

A few measurements were taken into account in order to understand if the voice system caused a difference in function choice. Since the model predicts that offering the voice system would make functions more accessible, there could be a difference in the number of unique functions, total number of functions used, or in the frequency in which particular functions were chosen. A difference in unique functions chosen would be a clear indicator of accessibility, meaning if those who had access to the voice system used a broader set of functions, then one can come to the conclusion that the voice system increased accessibility for those participants. If there were a difference in the total number of functions used, one could argue that the system made accessing functions easier, and therefore participants were more likely to interact in comparison to the traditional modalities. Finally, if there were an increase in the use of one particular function, then that could indicate that the voice system made that particular function more accessible.

In examining these predictions made by a model that supports voice systems as a tool to increase access to functions in a VA system, this study finds none of these to be true. Some participants in this study did use the voice system to access functions that they otherwise could not, but the difference in overall function selection was not significant between the experimental groups. The only statistically significant difference in the choice of functions between the two groups was when looking at the "variable_add" function, but this difference was only present when not taking into account the actions of the wizard. This signals a shift in modality, but not an expansion of access as predicted by the model. Some function and filter choices by the voice users were rarer or non-existent among the non-voice group, but those were relatively few interactions over a large set. Being able to access these functions may have contributed to a participant's choice to use the voice system, but it is difficult to differentiate the two experimental groups by such a small set of actions. In looking at the interaction model, participants described instances where they had difficulty accessing certain functions in Tableau but still chose not to use the "Hey Viz" system for the reasons described in Section 6.1.2 but did not use the voice system as much as predicted. The argument that having access to the voice system allowed these users to access more functions and close the "gulf of execution" is not strongly supported by this study.

6.3 Summary

This section addresses the hypothesis that offering the participants access to the voice system affected their ability to access functions and therefore changed the functions they chose. The analysis takes into account unique functions chosen, the total number of functions chosen, and the difference in the number of times chosen a particular function is used and found no difference between offering the voice system or not. Functions that were infrequently chosen are also taken into consideration but collectively did not have a significant effect on the overall function choices of the participants when measured by frequency. Interviews with participants revealed that they chose to use the voice system when they felt it was convenient, to get started on analysis, or in some cases when they were unable to access a specific function otherwise. Participants often refrained from choosing voice because their previous experiences with voice systems had led them to believe all voice systems were not capable of addressing their task needs. They also felt using the voice system was incongruent with gaining mastery of the underlying VA system, as the convenience of using the voice system could lead to its use as a crutch.

Chapter 7

Limitations

The choices made in designing this experiment, as well as in modeling the underlying phenomena of interest, led to corresponding limitations to the conclusions found by this study and the validity of its results. This chapter discusses some of these challenges, possible effects on the results, and possible alternatives possible for future studies.

7.1 Model of How Novice Users Would Use Voice Systems for VA

The underlying model developed to guide this study on how novice users would choose the voice system over traditional keyboard and mouse actions relies on a few key assumptions that may have comprised the conclusions of this study. The first is that it models interaction at the function choice level when there are many relevant levels to this process, as outlined by Pike et al. (2009). Given that previous work such as Grammel et al. (2010) had shown that novice users had difficulty in forming proper questions, the choice of focusing on access to functions may be overly simplistic and inappropriate. Though other works such as Sun et al. (2014) highlight the need to assist novice users in accessing appropriate functionality, future work should return to this question on whether voice systems assist novice users at the cognitive level to achieve their goals. Measuring changes in mental models using concept maps like in Zhang and Liu (2020) may be more appropriate for this aim.

Another assumption made by this model is that sensemaking specifically concerning the underlying dataset is the main driver of interacting with a VA system. Each function choice was meant to reflect how well the system supported participants in this goal. Lack of access to the desired function is then considered a poor quality design, and accessing a broader range of functions would be considered better design. This underlies much of the HCI perspective discussed in this study of how voice systems could potentially assist novice users in navigating VA systems, and led to the emphasis on function choice in this study's methods. However, the participants had clear instances where they were unable to find a desired function in the GUI but did not choose to use the voice interface. In most cases, the participants knew that they could use the voice system to access these functions but preferred to continue searching or find alternative functions using a keyboard and mouse. It is inappropriate to state the tasks did not motivate the participants to interact since these participants did move forward with their analysis. The participants would sometimes pursue alternative options if they could not find a particular function. It is more applicable to state that these participants were often willing to use combinations of functions less desirable to accomplish the task instead of using the voice system to access their original aims. These participants expressed a stronger desire to learn how to navigate the visual interface over building the visualization in the way they wanted at the time. The model for this study focused on the differences in affordances and mental models between using the traditional keyboard and mouse input in comparison to a voice system, and never considered that there would be a separate general user preference for learning the traditional interactions. This unexpected behavior led to lower than expected use of the voice system, given what the model predicts under these circumstances. The model's emphasis on function access does not reflect the attitudes of the participants in this study, and perhaps the model should be rebalanced to include learning or some other measure of utility driving modality choices.

7.2 Experimental Setting

In moving the study from in-person to a remote study, the facilitator lost control over the environment in which participants would be interacting with the VA system. Though the remote

VA system was the same for all participants, the system used by participants to access that system was not standardized. The original system in the UX lab had a large high-resolution screen with a high-quality microphone to capture the participant's voice during the session. The participants in their homes may have had a wide range of setups, including laptop computers with smaller screens and built-in microphones. The size of the participant's local screen could have affected how they viewed and interacted with the visualization system. The quality of the microphone could have affected how the cadence, volume, and perhaps even the content of their requests to the system. Local internet speed could have affected response times and how the participants interacted with the system. The choice to split the experimental system into two parts where one served Tableau while the other captured voice utterances may have also affected the response times experienced by the participants, though the choice to not split the system could have caused more lag and increased response times as well.

The physical space in which participants used the VA system may have had an effect on their use of the voice system. Though none of the participants explicitly spoke about background noises, the recordings picked up on many environmental sounds during the session, including roommates, family members, pets, and outside traffic. There would have been very little background noises and other activities around them in the UX lab, which may have been more conducive to using the voice system. Though this meant that participants used the voice system less during this study, this may actually be a better representation of how users would be affected by their natural environment when using voice systems for VA. During the regional lockdown, many data workers have been stationed at home preforming their VA tasks, and going forward this may become the default setting for many users.

7.3 Participants Selection

The focus on novice users and the participants ultimately recruited are also a source of limitations to the conclusions of this study. By design, the study did chose novice users of VA systems as the target demographics, so it is expected that many of these findings may not extend to more experienced users. The results are not expected to extend to expert users who are already proficient with the system and have different expectations. This study was focused on the interface design issues brought up by the HCI community that make interacting with VA systems more difficult for novice users. However, there are different aspects to being a novice that are ignored in choosing to narrowly define novice users as those who have little to no experience with a particular VA system. Users may be unfamiliar with the dataset, the domain of the data, building visualizations, and analysis in general. Increased access to functions does not necessarily mean better outcomes for users if they do not have an idea of the questions they would like answered and the outcomes they desire.

This study tried to address some of the other issues of being a novice user through the recruitment of participants. In choosing the business domain, it was expected that not only would this make it possible to control for domain familiarity, but those familiar with business tasks were expected to be more experienced in data analysis and likely some visualization techniques. The expectation was that business majors could fill these criteria without having prior experience with Tableau in order to meet the needs of the study. It was understood that they would likely interact with the system differently than other majors in the humanities, but that was an acceptable limitation in order to address the research questions.

In practice, the recruitment pool did not have enough participants in the business program, so there was an extra variable to consider between business majors and non-business majors. The pre-survey self-evaluations showed a gap in domain knowledge, though no difference in experience in data analysis was reported. The lack of difference in function choices during the experiment between the two groups could mean that domain knowledge did not affect these participants to a significant degree in these tasks. This could be a function of the simplicity of the tasks, which was by design to be easy enough for novice users to use. The other possibility is that though the business majors and the non-business majors in this study did not interact with the system significantly differently, other groups of business professionals who also happen to be novices to Tableau may have approached the tasks differently therefore interacted differently with the systems. Though extending the results of a user study using university students is a prevalent problem, the reasons for choosing or not choosing the voice system are in line with other studies on the strengths and weaknesses of voice systems for VA, as discussed in Section 8.1.

7.4 Longitudinal versus Single-Session Study

The choice to have this study as a single-session experiment did align with the goal of examining how novice users interact with VA systems, but it only captures a small portion of the user's experience. The effects of learning are not captured in this study since it focuses so heavily on the initial interactions. Though the initial encounters are important in understanding how users expect to interact with the system, users over time would be more exposed to the system's capabilities and would likely change their interactions. It is possible that over time these users would become more comfortable with the voice system and choose to use it more often. At least one participant regretted not choosing to use the voice system in this particular session, and so may have chosen to use it if given another chance. It is also possible that participants could use the voice system even less as they gain experience interacting using the keyboard and mouse input.

An extended study would have also complicated the definition of novice users for this study, as one would expect that any group of users forced to use a system over a period of time would gain some expertise. It would have been very difficult to model and understand how access to the voice system supported the users as they transition from novice to more experienced users of not just a particular VA system but also in VA tasks in general. It is also important to note that the study would also be forcing these users to interact with Tableau over a period of time when there is a chance many of these users would prefer to go back to their known tools such as Excel after the initial session. Those coerced interactions would also affect the conclusions of that potential study.

7.5 Task Design

There were trade-offs in the design of the experimental tasks that were necessary given the participants of this study. The focus on novice users plus the limited amount of experiment time meant that the tasks and data sets chosen for this study were simplistic compared against many real-life VA scenarios. An increase in the complexity of the task would have likely led to different strategies and different interactions. However, more realistic scenarios risked being too complicated for novice users to accomplish within a limited timeframe, so this study chose to focus on simpler tasks that novices can accomplish. The intention of the task design was to elicit a significant amount of interactions from this user group for comparison and not necessarily to exactly match real cases of using VA systems.

Making the tasks exploratory without an objectively correct answer was to allow users a wide range of possible strategies and requests, but it also means that there is no way to say that having the voice system led to better analytical outcomes. There are ways that the final products could have been graded using a set of heuristics determined beforehand, which would have had to operationalize quality through some combination of content and aesthetics. The final answers and visualizations could have been rated by business professionals on their insight and overall quality, but that would still be a subjective measure of outcomes. Facets such as the quality of the writing and the organization of visualizations are too tied to factors outside of the experiment to be informative in differentiating the performance of the two systems. In general, judging the final product of a novice participant performing a task during such a short time period is unfair to the complicated nature of VA. This study collected actual data on the interactions used and provided by the VA system to these participants, and the tasks elicited enough engagement and interaction for participants that the aims of this study were met without a need for judging the final products.

The tasks were meant to motivate participants to interact with the Tableau system, but some expressed a stronger desire to learn how to interact with the Tableau system using a keyboard and mouse over some achieving optimum task output. This could be different if the final visualizations and recommendations had more intrinsic value to the participants, such as in connection to their workplaces. A competitive incentive instead of a flat compensation structure could have also affected their choices, as a participant would have a stronger disincentive to risk never accessing a particular function if it meant possibly losing money.

7.6 Training Design

The study may have underserved the voice system in an effort to not present it as the preferred method of interaction. It was important to not guide users towards using the voice system since this study wanted to understand how novice users would choose between modalities. However, it is possible that the design may have overly downplayed the emphasis on the voice system in comparison to the traditional keyboard and mouse interactions. For example, the portion of the training video is only three minutes in comparison to the sixteen minutes dedicated to the rest of the Tableau system. This was done in order to blend in the voice system as just another component of the Tableau system and not a separate piece, so the segment was meant to last about as long as the other segments covered in the video. The written materials reflected this same imbalance in training materials, as out of the twenty-four pages provided, only one was specifically for the voice system. Additionally, there was no requirement to ever use the voice system by participants even during the training task outside of confirming the “system” could hear the participant. Corbett and Weber (2016) found the initial training and exposure to the voice system to be critical in its usage, and by not requiring a certain amount of initial interaction, the participants may not have felt comfortable before getting into the experimental tasks. This study risked underselling the voice system in order to avoid influencing participants towards it, but participants may have required more training and exposure with the voice system in order to feel comfortable using it.

It is not expected that the novice users were able to learn all of the training materials within the task time, but given equal training, it is important to understand how these novice users would begin to interact with systems with or without a voice modality. Training time was limited, and it is difficult to gain substantial knowledge of how to use a system within such a short period of time. Srinivasan & Stasko (2020) acknowledge the difficulty in providing proper training for NLI for VA systems and suggest intermediary tasks that require using the NLI in order to get the user more accustomed to the system. These participants may have used the voice system more if given more training in its use, but the necessary amount of additional training they would require is unclear. Also, it is uncertain that additional exposure to training my lead to increased use of the voice system, as participants may instead choose to perform those same functions themselves using a keyboard and mouse.

7.7 VA System Choice

An unexpected consequence of choosing Tableau over building a custom testing system was the participants' eagerness to learn how to interact with this particular system. At times the value of learning to interact with Tableau was greater than exploring the voice system, and that may be because this particular user group was more interested in gaining skill in a popular tool for the workplace. This led many to keep exploring the interface for function even when they knew that the voice option was available, and then using alternatives in the GUI if necessary. Participants were not told during recruitment that they would be trained in Tableau, so there was no sorting effect that caused this study to attract users looking for training in that particular system. Though this was unexpected, it may actually better reflect the choices of users in potential real-life situations. Participants using a custom VA system in a lab user study have no incentive to learn the system outside of the construct of the experiment, whereas in real life, they may choose to learn how to navigate Tableau and better understand that part of the system over attempting an experimental voice system.

Choosing to use Tableau and not to build a VA system from scratch created some necessary trade-offs. Because it is a proprietary system, there was no easy way of connecting to Tableau in order to record the functions used and the state of the system. Instead features of the Tableau visualization had to be manually recorded afterward. A custom coding scheme was also necessary for mapping user actions to functions since there was no mapping provided. An impactful drawback of this choice was that it made integrating the voice system much more difficult. If this study used a VA system developed specifically for testing, then the voice interface could have been a seamless part of the experience. Aside from blending in with the rest of Tableau aesthetically, the voice system did not allow for options such as interacting with voice and keyboard or mouse at the same time or even parallel work where a request could be processed as the participant continued to interact. This was not possible since the wizard had to take full control over the Tableau interface in order to answer requests. Though the system could be considered multimodal because it offered different modalities, it is a very limited system that does not allow for the kind of simultaneous interaction found helpful in other systems like Orko (Srinivasan & Stasko, 2018), which allowed for multimodal interaction because it was fully automated. Badam et al. (2017), in their review of different modalities for VA systems, advocate for supporting mixed interactions to better support users. A future study involving offering the choice of using a fully automated system instead of using a human operator would better address the issue of multimodality in VA.

7.8 Voice System and Wizard Design

Limiting the multimodality of the system was one of many decisions concerning the design of the voice system that affected the conclusions of this study. The choice to have the voice system simulated using a Wizard of Oz experimental setup instead of fully automated likely affected the choices made by participants. A fully automated system may have been faster to perform the functions, but it may have also been more likely to commit errors. This would have been more realistic since any NLP system is inherently probabilistic, especially with difficult and complicated domains. The issues in developing this kind of system would have been in creating something that could handle a wide range of requests reasonably well so that the study conclusions relate to the modality choice of novice users and not in the development quality of some particular tool. Since the required system is thought to be plausible but still not in existence, the choice to have a human operator was made.

Having only one experimental voice system condition did not allow for controlling other factors that could have impacted the results of this study. The bias for or against using intelligent systems could have been controlled by having a separate experimental group have access to a text interface. This research was really concerned with voice systems as a modality, so focused the development on building and supporting that interface over adding an additional text interface. Similarly, there could have been multiple levels of functionality offered instead of just one very powerful option. Offering a simpler system that adheres to only a limited set of literal requests like ggspeak (Diesendruck & Zhao, 2016) could have made the capabilities of the system easier to understand for participants. It may have also assuaged some of the concerns of laziness on the part of participants since it would not be capable of complex tasks, and thus been a more attractive option compared to the wizard presented here. However, it is also possible that a simple system may have confirmed to participants that it was not able to meet their task needs based on their experience with other systems such as Alexa or Siri. There were limited resources to test multiple voice systems, which would require extra recruitment and training when the main question was about the choice of a voice system to access functions. The decision was made to have the voice system simulate the most advanced but possible future development in this area, but the system may have been presented as too powerful than what is believable with current technology. This would explain why participants expected the system to function at a similar level to Siri and Alexa when they were told that it could perform much more complicated tasks.

Aside from general capabilities, there were other decisions in how the wizard would respond to requests that could have affected this study. In order to disambiguate possible requests such as “Make this bigger,” the wizard monitored the state of the Tableau interface throughout the sessions. Because of the remote set up the wizard also listened in on every said by the participant even when not clicking the “Hey Viz” button. In the original in-person set up this was prevented, but this was not found to be possible in the remote configuration. Participants could have perhaps thought out load during their sessions, and this may have influenced the actions of the wizard, as this additional information may color the requests given. In this study, no participant spoke through their analytical process or made any utterances outside of the requests, so this was not a factor in how the wizard responded.

There was a concern that the participants would guess that the system is actually human-operated and that this would affect their interactions with the system. Those participants who had the voice system available were never directly asked if they believed the system was fully automated to avoid coloring their responses. However, in examining the interview transcripts, participants never mentioned there possibly being a human operator of the system. They did speak of the system in comparison to other automated systems such as Alexa or Siri and spoke about their dealings with the system in a similar way. Their general tone seemed in line with what one would expect when reviewing an automated system, but this belief was never confirmed.

7.9 Role of Researcher as Facilitator and Wizard

The researcher served every role during for this experiment, including facilitating the experiments and performing as the Wizard. This was due to the COVID-19 limitations of the time but it is important to acknowledge that this may have caused some bias in the experiment, data collection, and analysis. The researcher could have been motivated to either encourage or discourage the use of the voice system, or could have acted in a way that in general could have affected the choices made by the participants. The guidance created for the Wizard behavior allowed for a great deal of judgment on the part of the Wizard in how to respond to requests in order to ensure that participants would almost always get a response from the system. There were no automated quality controls on those responses. A bias in behavior could have then been ignored or unintentionally missed since the researcher would be examining their own performance. It is difficult researchers to be objective and critical of their own work, and having another party run the data analysis for the qualitative and quantitative data would have been appreciated. Ultimately the researcher is the one who has to present the results of this study and shoulders the responsibility to make every possible effort in presenting the results of the experiment as fully as possible.

7.10 Summary

This section presents some of the limitations of this study as they affect the interpretation of its conclusions. The participant selection and system design really limit how well these interpretations can extend to fully automated systems with actual potential users. Many other factors limit the conclusions reached, although using a popular commercial product in the participants' natural working environment did better reflect the circumstances in which a user would interact with these systems and as a byproduct led to decreased use of the voice system.

Chapter 8

Discussion and Future Work

8.1 Discussion of RQ1

Participants listed factors affecting their use of the voice system that are not just functions of their ability to interact with the interface but a reflection of their attitudes towards using voice systems in this environment. The motivating factors of convenience and increased access to functions are directly related to the structure and capabilities of the interface. Voice system becomes less convenient to use in comparison to using keyboard and mouse if the function required fewer clicks or movement, and less desirable if the desired functionality is more obvious to use in the visual interface. These reasons fall in line with the weaknesses identified by Lee et al., (2012) concerning the difficulty users have in using and accessing functions in a WIMP interface, and how voice presents an opportunity for users to cut through the visual interface to better access functions and close the “gulf of execution.” However, the need for participants to require specific functions in order to use voice is in contrast to their exploratory behavior within the visual interface. Given any uncertainty as to what they wanted to accomplish, participants expressed that they were not willing to attempt a voice request but would rather click on items in the visual space until they either develop a specific request or generate the desired outcome. Participants would rather be lost in the visual interface than to risk a potentially unexpected outcome from the voice system.

The factors that mitigated the use of the voice system were reflective of the negative attitudes participants had towards using voice systems. Many participants used popular systems such as Alexa or Siri and brought lowered expectations for the performance capability of the “Hey Viz” system, despite training materials expressing and demonstrating some of the capabilities. Although users had no experience with Tableau or with this particular voice system, they applied their previously developed mental model of voice systems and interacted with "Hey Viz" based on their past experiences. This then affected not just the type of requests to the systems but how those requests were structured, as shorter and very specific queries were thought to be necessary to achieve positive outcomes. The preference for shorter and simpler requests in order to minimize the risk of negative outcomes coincides with the findings of Luger and Sellen (2016). Srinivasan and Stasko (2018) found participants in testing their multimodal VA system Orko (Srinivasan & Stasko, 2018) had similarly low expectations for the system to be able to handle natural language queries despite being provided training examples. It is possible that improved training materials or a different training session structure may improve users' attitudes of the voice system, perhaps by directly contrasting the system against Alexa or Siri as being much more capable.

Improving participants' belief in the capability of the voice system does not address the negative feelings these participants had towards using the "Hey Viz," which they felt was incongruent with learning how to use Tableau. Though participants found the system useful, they also felt it was possible to overuse the voice system to the detriment of gaining the preferred ability to interact with Tableau using a keyboard and mouse. Even when presented a situation where they could not access the functions that they required, participants still chose to struggle with the visual interface because using the voice system was seen as a crutch and not truly gaining mastery over the Tableau. VA is about supporting users in the sensemaking process, but the users have to also build a mental model of the VA system itself as well as the data. These participants prioritized making sense of Tableau over perhaps developing a better understanding of the data set. This is not particular to voice, as Setlur et al. (2016) found similar negative attitudes towards automated systems when testing a text NLP system for Tableau. These intelligent assistants were seen as appendages to the main system and not as a helpful integrated component. This bias did not hold true for the "Show Me" visualization help tool built into

Tableau, even though it provides a large amount of assistance in organizing and encoding data fields.

Participants may have prioritized learning how to use Tableau over completing the experimental tasks because it is a well-known commercial system. To our knowledge, this is the only study of VA systems where the participants expressed a greater concern for learning the system than in optimally completing the tasks. If the underlying VA system were completely custom and made for this particular experiment, as is the case with many other studies, there would be little value in mastering navigating its interface outside of task completion. Because Tableau is a well-known commercial system, there was likely more interest and value in learning how to use the system since those skills are then transferable outside of the experiment. Many participants thought they might have to use it in a future work setting and expressed gratitude in being exposed to Tableau training for the first time.

The voice system was able to ameliorate the interaction issues for novice users of VA systems highlighted by Kwon et al. (2011) and to close the “gulf of execution” but only in cases where the novice users actually chose to use it. As discussed, some of the users were able to access functions more appropriate for their needs but often chose not to rely upon the voice system for assistance. The participants viewed themselves as novice users that had to gain mastery of the system through keyboard and mouse interactions without using the voice system as a crutch. Some participants insisted on only using the system for functions that they themselves can double-check. Others believed that their own ability to use the VA system could not be surpassed by an automated system, even as though this was their first experience with Tableau. These attitudes prevented these users from addressing their issues accessing functions in the VA system, and thus weakened the evidence for voice as a solution to these interaction concerns. A few participants mentioned that the voice system might be better suited for expert Tableau users that already know how to interact with the system but would like a more convenient way of doing so. This may not hold true in future work, but it does reflect the participants’ attitude that novice users should not rely on the voice system to solve their problems when interacting with the VA system.

Voice as a modality has its own specific challenges known in the HCI community that affected the rate of use by participants in this study. Multiple participants reported that they had forgotten that the system was available to use, even when reminded in-between tasks and having a prominent button present on their screen. This may also be a function of switching between modalities, which has been found to produce its own cognitive cost in studies such as Fintor, Stephan, and Koch (2018). This may be a counterbalance to the belief that speaking with a VA system would be cognitively easier. Though this study was more concerned with function accessibility over cognitive easing, this is certainly an issue that should be studied in the future. Other participants described their discomfort with using voice in a desktop setting, as it is not only a new experience but also potentially disruptive to others in their environment. Though no participant mentioned it, many of the recordings included background noises expected from being in a home environment (friends, family, cars, pets, etc.) that may also contribute to the participants' hesitancy to use voice since they were in noisy environments. Having the experiment take place remotely from participants' homes instead of an isolated lab highlights this issue, and future developers serious about pursuing this work will have to consider these work environments when building voice systems. Additionally, respondents reported a significantly higher level of agreement with the statement that the system responded longer than expected in the voice group. The average response time around 22 seconds between beginning a request and receiving an outcome may still be noticeably long and dissuade users from continuing to use the system.

8.2 Discussion of RQ2

Participants offered reasons why they did not make requests at a higher complexity and quantity, which means some were at least aware that they could have accessed more functionality but preferred searching through the interface themselves even at the risk of not finding it. These findings coincide with Luger and Sellen (2016) that found most users focus on simple tasks over more complex ones, as they often do not trust the system to operate correctly. Though voice systems present the opportunity for increased access to functionality, it is how users choose to use it that will ultimately determine if any new functions are used.

A potential opportunity for voice systems to make an impact on what functions a user can access may be to offer a "show me how" functionality, as multiple participants mentioned that they would appreciate if the voice system could demonstrate how to accomplish a certain function as opposed to just completing it. This feature would not be as affected by many of the mitigating factors for using voice listed by participants and would improve their knowledge of the system as well as their task outcomes. By using this feature, novice users can learn how to interact with Tableau using keyboard and mouse interactions, and then afterward may choose to use the voice system to perform the task without any risk of it feeling like a crutch. Instead of closing the "gulf of execution" by changing the modality of the function, this system would be used as a learning aid in support of the traditional desktop interactions that would still address the interaction issues of novice users. The Wizard of Oz experimental design of this study prevented this function from being possible because it would have likely signaled to participants that the system had a human operator. Future studies with a more automated voice system that allows the user to view the actions of the voice system could allow for "Show Me How" to be available. An extension of this approach could be to proactively guide users towards functions when their actions seem to indicate that they are searching for functions, but this would require a much more sophisticated user model.

8.3 Theoretical Implications

This study was based on two related models regarding how novice users would interact with the VA system driven by their need to develop a mental model of the system in service of the motivating task. Like many previous works evaluating VA systems, this study assumed that the main motivation of users to interact with a VA system is to build the appropriate mental model for their particular tasks. Novice users were then expected to build a mental model of the VA system for the sole purpose of building those models, and this would be done through training materials and hands-on experience. These two sensemaking processes would likely happen in parallel as the need to adjust the visualization would drive the user to learn more about the VA system. As a novice user became aware of a "gulf of execution," they would attempt to rectify that situation by likely using the voice system since previous studies had shown that they could verbalize their desires much easier than having to map them to interactions. Since task completion was the motivating factor, users would then choose the easier path.

Based on those two models, what is not expected then is that users would choose to put more effort into understanding the VA system at the potential cost of not being able to build their desired outcomes. These participants were given the opportunity to more easily interact with the system but still chose the more difficult path in navigating the VA interface. Using the evaluation cost framework proposed by Lam (2008), one would expect that the voice system would provide a much more efficient experience for users and lead to different outcomes, but that is not the case in this study. These users became aware of "gulf of execution," preventing them from particular interactions but still chose not to take advantage of an option to perform those same functions. The participants were more interested in sensemaking in terms of understanding how to use the Tableau system with the keyboard and mouse over optimizing some task output. The participants' function choices were still motivated by the task, as exploring the Tableau interface was always motivated by a particular outcome in mind from participants. What was unexpected then was that instead of using the voice system that they knew could accomplish the function, participants would find alternative approaches. Novice users may then necessitate an adjustment to the model for sensemaking in VA systems beyond task-driven mastery. Future models may incorporate the users' priorities in coming into an interactive session that may be driven by a particular task goal or by learning how to better interact with the system for future use, thus requiring different weights for these simultaneous and competing processes.

Though this study has many limitations due to its design, as discussed in Chapter 7, the underlying idea that offering users the choice to not use the novel system is important to future work. Across many domains, including VA, there is a movement to use new interactions to make systems easier to operate. But as explained by Norman (2010), there is no such thing as a natural modality, and every interaction requires some mental model to operate the system. Offering a choice to use the voice system also means that there may be a cost in switching between modalities, and that is an issue underrepresented in the field. The arguments for possible changes to this study further highlight how many different factors affect the use of a voice system, including the design of training and the capabilities of the system. This study aimed to understand when a user would choose to use a voice system and has gathered some important feedback for future work. Hopefully, the lasting impact of this work is not just to further work about voice as a modality but to encourage researchers to test whether their novel system would actually be chosen by their participants or if the data they are receiving is just a product of a synthetic situation. Assuming that a "natural" modality or any new design will mean lower cognitive effort on behalf of participants and therefore lead to a preference of use is too large of an assumption for a field dedicated to supporting users in their work.

8.4 Practical Implications and Future Directions

This study attempted to address the question of modality choice, which is key for any future development of multimodal VA systems. The traditional desktop environment is where most users will be interacting with data, and this study highlights some encouraging opportunities as well as some challenges in that domain. Participants appreciated the convenience of voice systems for functions that would otherwise require multiple interactions, as well as functions that were difficult to locate within the visual interface. These are items that can be operationalized so developers can be better prepared to answer those requests, such as mapping functions within menu hierarchies or through the analysis of clicking behavior.

If researchers are still interested in building voice applications for VA system for the purpose of helping novice users, they should probably then consider the systems most likely to be used and the environment that they will be used in. The focus then should move towards widely adopted systems like Tableau instead of custom solutions made for study. Workplace studies are difficult to conduct but are also necessary for understanding if and how these new modalities would be used. Longitudinal studies would also be necessary since this study was too brief to capture learning effects, and the usage of a VA system and a voice system are likely to change with exposure over time.

Researchers in the future can also highlight opportunities to utilize voice systems that do not make the user feel as negative about their choice. The most novel recommendation for future systems is the “Show Me How” function, which is technically feasible and addresses the mitigating factors participants had in using the voice system. Users would not only be assisted in accessing a particular function but once they understand how to perform a function would also likely be more willing to choose the voice system in the future if they feel it is more convenient to do so. This particular implementation could not show users how to interact with the system, but future interactions can be a better support tool for learning VA systems.

Developers need to be aware of users’ bias against using voice systems and should be prepared to present their solutions in a way that addresses these attitudes. A likely approach is to contrast their system against other simpler voice interfaces such as Amazon’s Alexa or Apple’s Siri that users feel are underpowered for their needs. This is a function of the prevailing technology of the time, with the expectation that future systems with increased capabilities will become more prevalent, and therefore, the users’ expectations will also change. This may happen soon with the rise of devices that are always listening and ubiquitous computing, but for now, users do not expect a very high-quality product from voice systems. If researchers would like to prove otherwise, they will have to deal with the many issues raised by this study.

Appendix A – Recruitment Message

My name is Jonathan Pulliza, and I am a doctoral candidate in the Department of Library and Information Science at RU School of Communication and Information (SC&I). My colleagues and I are conducting a study on how new users build charts, graphs, and other visualizations with a new visualization system. You are invited to participate in this study! **So you get to try something new and get paid for it!!!**

In order to participate in this study, you have to be an adult who is proficient in speaking and listening to English, you must be currently enrolled as a student at Rutgers University, and you have to be able to access the internet with a computer with a microphone.

This study consists of one laboratory session, in which you will be required to complete one training task and two search tasks using the system. You will interact with a visualization system to build visualizations to answer questions. Afterward, we will ask you to answer questions related to your experience and preferences.

The total time you will spend on this study will be no more than two hours, including a 10-minute exit-interview. You will receive a **$40 Amazon Gift Card** for your participation upon completion of the study. Participants will be required to perform the tasks during a designated timeslot with the researcher. Taking part in this study will help to advance our understanding of how to design future versions of visualization systems.

Here are the basic **requirements** to take part in this study:

- You must be at least 18 years old to participate.
- Proficiency in English is required.
- You must currently be a student at Rutgers University.
- You must have access to the internet using a computer with a microphone

Once you register for your participation, you will receive further instructions regarding how to go about taking part in the user study. Participation is purely voluntary. Choosing or declining to participate in this study will not affect any of your classes or grades at Rutgers. This study has been approved by the Rutgers Institutional Review Board (IRB Study #) and will be supervised by Dr. Nina Wacholder (ninwac@comminfo.rutgers.edu) at the School of Communication and Information. For more information about this study, please email Jonathan Pulliza at jonathan.pulliza@rutgers.edu. You can also contact Jonathan Pulliza to ask questions or get more information about the project.

Thank you for your interest! I look forward to hearing from you.

To sign-up*, please fill up this online link:

*Signing up for the study does not guarantee your participation in the study due to the limitation of availability. Once your signup has been confirmed, you will receive further instructions on how to proceed with the study.

Appendix B – Pre-Experiment Questionnaire

Please answer the following questions regarding your demographics and your previous experiences relevant to this study.

- Enter Your Participant ID Number
- Please Enter Your Age
- Please Select your Gender (Male / Female / Prefer not to say / Other)
- Are you a native English speaker? (Yes / No)
- What is your major?

For the following questions, please select a number between 1 and 5 that most represents your familiarity with the subject, with 1 being “Not Familiar At All” and 5 being “Very Familiar”

- Data Analysis
- Business Data Analysis
- Sales Data Analysis
- Marketing Research
- Microsoft Excel
- Microsoft Power BI
- Tableau
- Other Visualization Software

Appendix C – Task Instructions

Task 0:

For this task, we ask that you complete an introductory tutorial on how to use the Tableau system. Tableau is a visual analytics application to create interactive visualizations.

The tutorial video covers the basic functionality available to users in order to build visualizations. You may rewind and replay any portion of the tutorial. You may also follow along using the Tableau dataset already loaded. When you have completed the tutorial and feel comfortable using the system, please answer the following question:

What category had the most orders returned in 2015?

Task 1:

For this task, we ask that you take another look at the sales data you have already worked with during the Tableau tutorial. Below is a brief description of the source of the data, dataset provided, and the task. When you have completed the task, please provide a brief written description of your answer to the task question (max 500 words), referencing the visualization(s) you have built to support your argument.

Description of Task T1:

A member of the sales team has some questions regarding how to invest their resources for the next year. Using this dataset, please identify areas where you believe the company should invest in and/or divest from that would have the greatest impact in the next year and provide visualizations supporting your argument.

Description of the dataset:

The data is a collection of orders for a global shopping site for the years 2012-2015. Each order is described by the product, customer, order location, sales, and profit.

Variable	Variable Description
Order ID	Unique Order ID
Order Date	Date Order Submitted
Order Priority	Priority of Order (Low, Medium, High, Critical)
Category	Product Category
Sub-Category	Product Sub-Category
Product ID	Unique Product ID
Product Name	Product Name
Quantity	Quantity of Product Ordered
Market	Order Regional Market
Country	Order Country
Region	Order Region within Country
State	Order State
City	Order City
Postal Code	Order Postal Code
Segment	Segment of Order (Consumer, Corporate, Home Office)
Customer ID	Unique Customer ID
Customer Name	Customer Name
Sales	Amount of Sales in Dollars
Discount	Discount applied on sale as decimal
Returned	Product Returned (Yes/NULL)
Ship Date	Date of shipping
Ship Mode	Mode of shipping
Shipping Cost	Cost of shipping in dollars
Profit	Net profit on order in dollars

Task 2:

For this task, we ask that you take a look at a different dataset concerning coupons sent to potential customers. Below is a brief description of the source of the data, dataset provided, and the task. When you have completed the task, please provide a brief written description of your answer to the task question (max 500 words), referencing the visualization(s) you have built to support your argument.

Description of Task T2:

The marketing team would like to evaluate the success and failure of their campaigns during this period to improve their work going forward. The goal is to send out the right coupon for the right product to the right customer, so the customer redeems the coupon and buys the product. Using this dataset, please identify potential strategies the team should focus on in order to maximize the number of customers redeeming these coupons, and provide visualizations supporting your argument.

Description of the dataset:

The marketing team sent out coupons in a series of marketing campaigns for different products to potential customers from 2012 to 2013 and tracked which coupons were redeemed by which customers for which products. Customers were offered multiple coupons to different products in each mailing campaign, and customers may be a part of different campaigns over the period.

Variable	**Variable Description**
Campaign Id	Campaign ID Number
Campaign Start Date	Start Date of Campaign (MM/DD/YYYY)
Campaign End Date	End Date of Campaign (MM/DD/YYYY)
Coupon Id	Coupon ID Number
Customer Id	Customer ID Number
Coupon Redemption Status	Whether the Coupon was redeemed for the item (Redeemed, Not_Redeemed)
Customer Age Range	Age Range of Customer (18-25, 25-35, 36-45, 46-55, 56-70, 70+)
Customer Family Size	Size of Customer Family (1, 2, 3, 4, 5+)
Customer Income Bracket	Income Bracket of Customer on an increasing 1-12 scale
Customer Marital Status	Marital Status of Customer (Single, Married, Null)
Customer No Of Children	Number of Customer's Children (1, 2, 3+, Null)
Customer Rented	Whether the Customer rents their home (Rents, Owns)
Product Id	Product ID Number
Product Brand Id	Brand ID of Product
Product Brand Type	Brand Type (Established, Local)
Product Category	Category of Product (Bakery, Garden, Grocery, Meat, etc. …)

Appendix D – Post-Task Questionnaire

Enter Your Participant ID Number

Which task did you just complete? (Task 0 / Task 1 / Task 2)

Please provide your answer to the task below. See the task description for further instruction

Your Answer to the Task Question

For the following questions, please circle a number between 1 and 5 that most represents your agreement with the statement, with 1 being "Strongly Disagree" and 5 being "Strongly Agree"

1. The tutorial provided the appropriate level of information to complete the task.
2. I understood the task instructions.
3. The task was very difficult to complete.
4. I was able to access all of the functions without difficulty.
5. The system often did not react in the way I thought it would.
6. The system took longer than expected to respond.
7. I can easily navigate the visual interface of the system to find a function.
8. I understand how to interact with the system to complete a task.
9. I felt limited in what I could or could not do with the system.
10. I am confident that my answer to the task question is correct.
11. I am satisfied with the final product I produced for the task.
12. I feel confident using the system to complete another task.

Appendix E – Guide for Semi-Structured Exit Interview

1. [Warm-up] How was your overall experience of working on this study? Is there anything in particular that you want to talk about?
2. Describe how you went about building your visualization. Was there any strategy involved?
3. When you were interacting with the system, was there any unexpected system behavior?
4. What are some possible improvements to the system?
5. Did you feel there were some functionalities that were missing or you wish were available?
6. Is there anything else that you would like to share about your experience with the system?
7. Do you have any questions for me?

[For C1 Users]

8. Why did you [not] choose to use the voice modality available when you did?
    a. What were some system functions that you felt more comfortable using voice over keyboard and mouse or vice versa?
9. How did you feel about using your voice during the tasks? How did it affect your visualization process and the final product?
10. Can you talk about the differences you felt when you used the voice system versus keyboard and mouse?

Appendix F – Tableau Video Tutorial Chapters

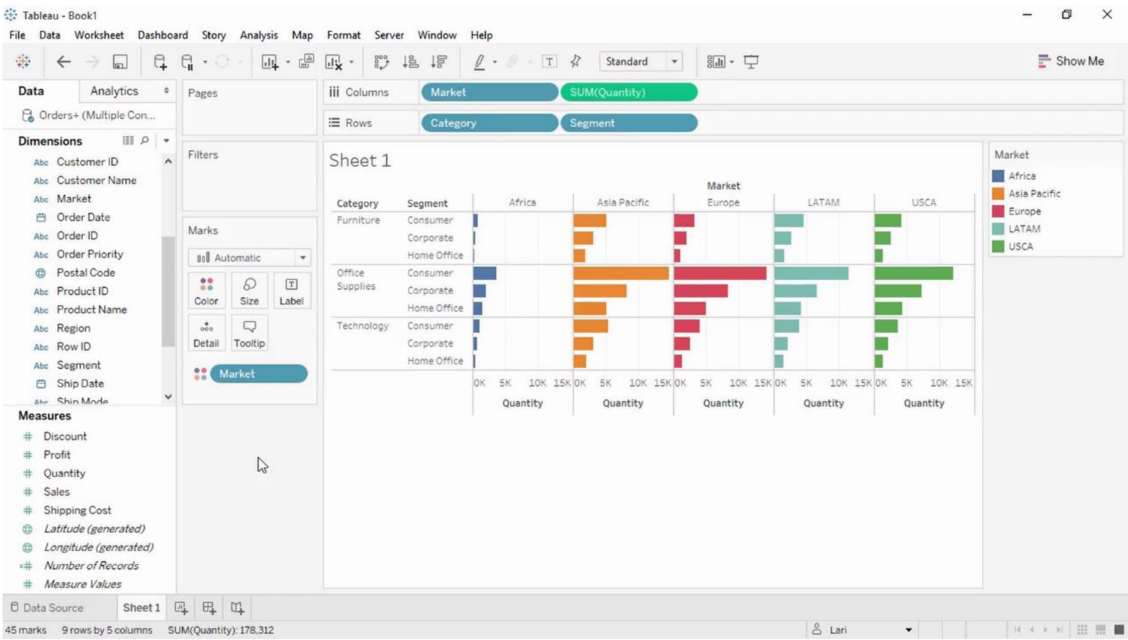
Chapter Number	Chapter Title	Included for Study
1	Connecting to Data	FALSE
2	Joins and Data Preparation	FALSE
3	Connecting Live versus Extracting	FALSE
4	Dimensions and Measures	TRUE
5	Building Views	TRUE
6	Quick Table Calculations	TRUE
7	Crosstab and Exporting Data	TRUE
8	Show Me	TRUE
9	Custom Territories	FALSE
10	Filters	TRUE
11	Bar Chart	TRUE
12	Hierarchies	TRUE
13	Sorting	TRUE
14	Grouping	TRUE
15	Working with Marks	TRUE
16	Trend Lines	FALSE
17	Dashboards	FALSE
18	Story Points	FALSE
19	Distributing Content	FALSE

Appendix G – Training Materials Provided to Participants

Dimensions and Measures

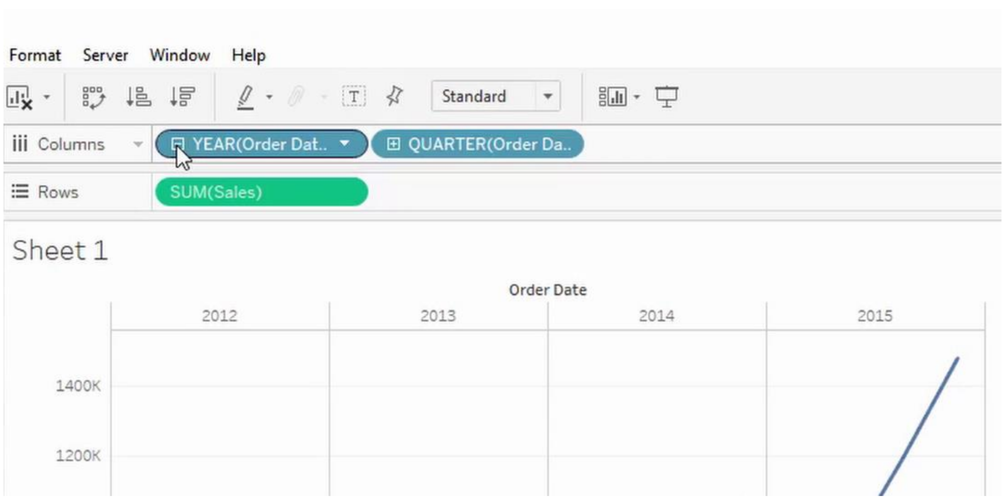
We're now connected to that data set. Let's see how easy it is to dive into our data. We simply drag the fields out, let's bring: Category to rows, Segment to rows, Quantity to columns, Market to columns, and let's bring Market to color, as well. It's that easy to create a visualization of how our Sales are looking per category, customer segment and market, in terms of number of items sold. We can quickly see that Africa is an emerging market for us.

You'll notice that I brought in those fields from this data pane here on the left. It's broken up into dimensions and measures that represent the column headers in the excel sheet. What are dimensions and measures? Dimensions are categorical fields, in this case, fields such as date, customer, and Category. These are fields that we want to slice and dice our numerical data by. Dimensions are often discrete. Discrete fields create labels in the chart and are color coded blue in the data pane and in the view. Measures, on the other hand, are our metrics. They are the numbers we want to analyze. Measures are often continuous. Continuous fields create axes in the chart and their pills are color coded green.

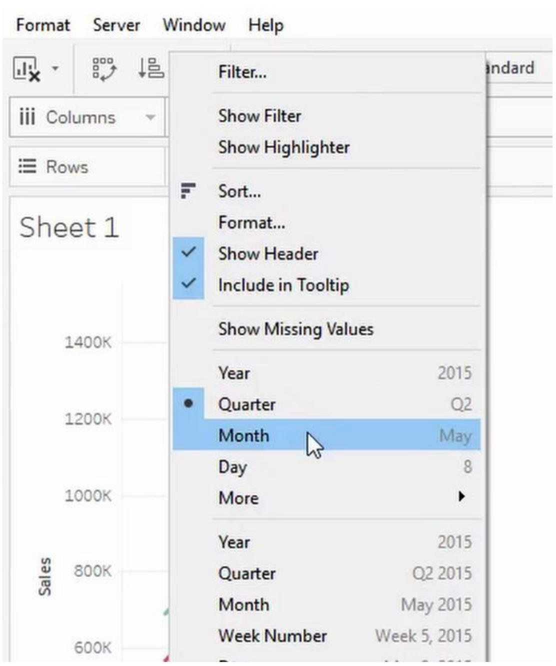


Building Views

Now, let's say we're interested in our total sales number. Let's place Sales in the view. We can see that Tableau queries the database and returns a single result giving us the sum of Sales. This company has done a little over 12 and a half million in sales. If we want to see this over time, we can drag Order Date to the top of the view. Tableau Desktop aggregates our dates at the year level. We can expand this with the plus (+) symbol on the pill. Now we see both quarters and years in the view.

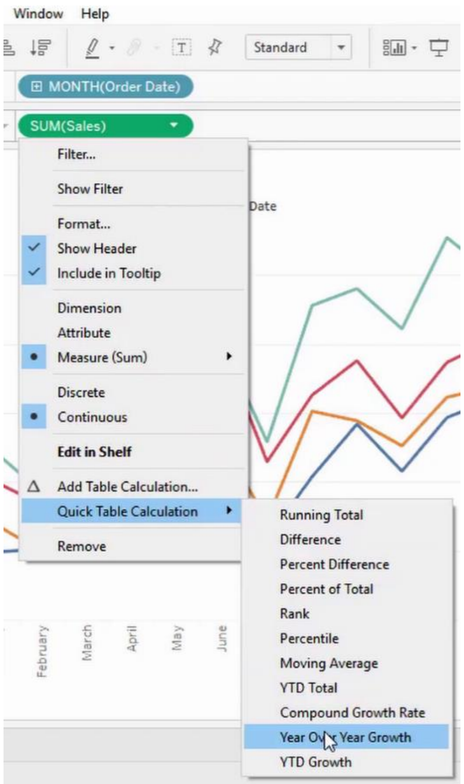


To see how all our Q1s are doing over the years, we can easily pivot the data so Quarter is in front of Year. Now we can compare how our growth looks by quarter across the years. Moving Year to Color shows us all the years on top of each other. If, instead of drilling down further, we want to change quarters to months, we can click on the pill to bring up the drop-down menu and change it. If looking at an average of sales is more useful than sum of sales, we can simply change that by using the dropdown menu and changing the aggregation to average. But let's undo that for now.

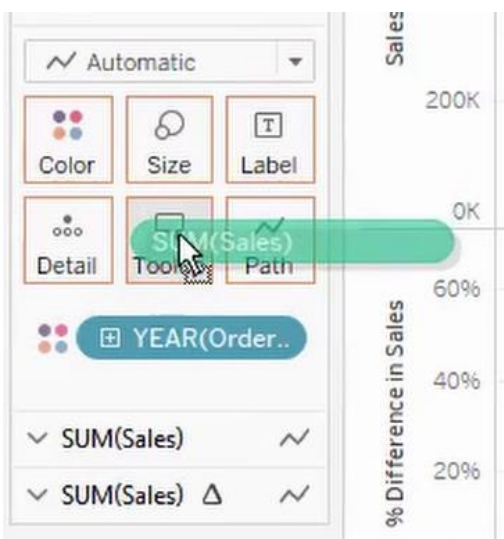


Quick Table Calculations

What about if we want to know about something like year over year growth? In Tableau Desktop, calculations like this are easy. Once again, clicking on the pill's dropdown brings up the menu, and now going to Quick Table Calculation, we can see common business calculations as single click options. Let's select "Year over Year Growth".

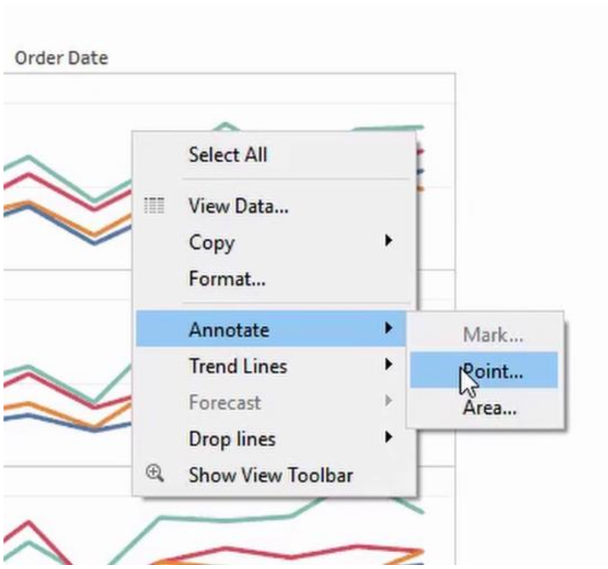


If we still want to see the original Sales, we can simply place it back into the visualization. Perhaps we want to have the Year over Year Growth values appear in a tooltip instead of a graph, we can simply move it to the Tooltip shelf. The tooltip provides additional information when we hover over marks in the view. For example, here in November of 2015, we see we're almost 50% up from the previous year.



Let's drag Category to the Rows shelf. We can now see which categories are doing well, and when they were doing well. We could even leave comments. For example, we see there's a yearly dip in sales in July, but we rebound in the fall. We can leave an annotation by right-clicking, selecting Annotate, and adding a point Annotation.

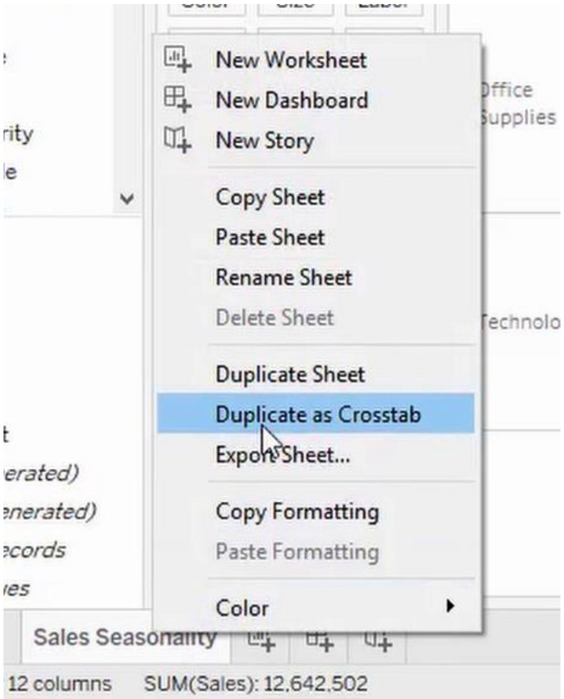
This is a useful view--if we wanted to easily share this, we could now right-click, copy the image, and quickly share it with other people in our organization.

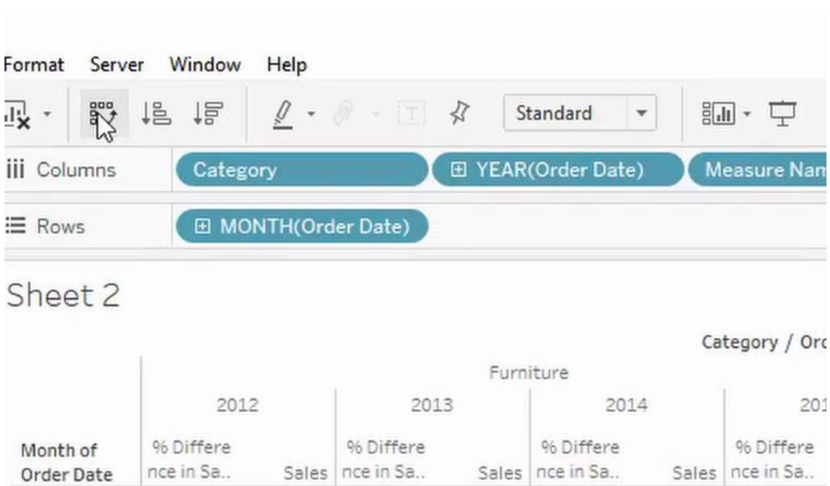


But for now we'll double click on the sheet tab and rename this "Sales Seasonality".

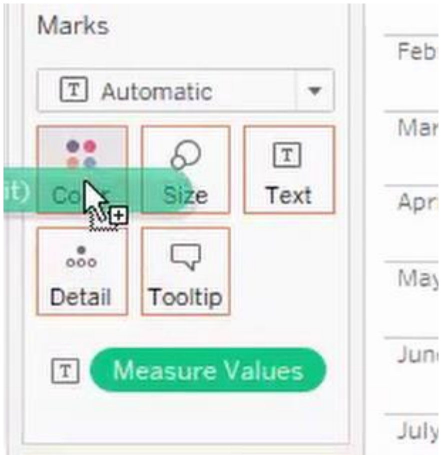
Crosstab and Exporting Data

What if we want the raw numbers behind this timeline? Tableau Desktop makes this very easy to do. We can simply right click on the tab and "Duplicate as a Crosstab". We can easily swap our axes and move Category to the Rows shelf. Let's make this fit a little better.

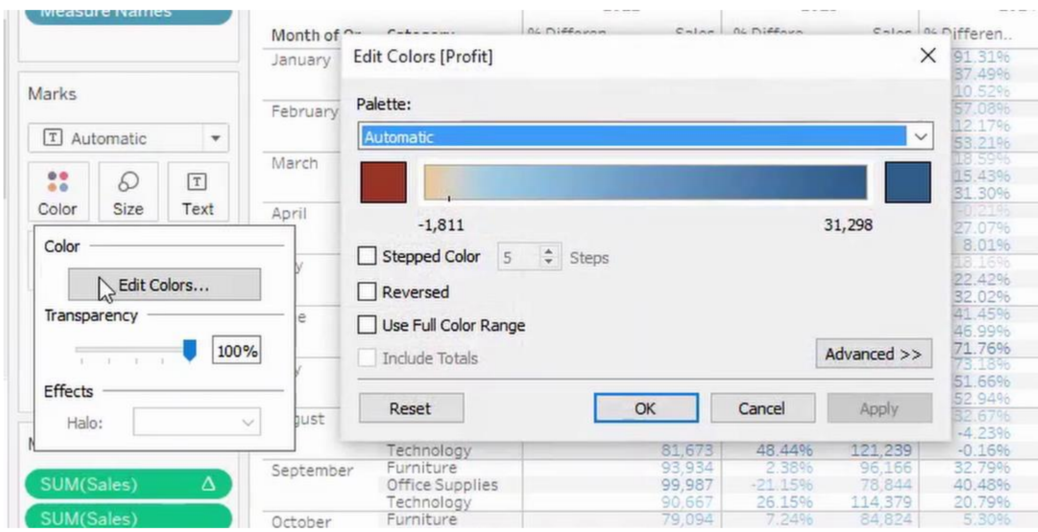




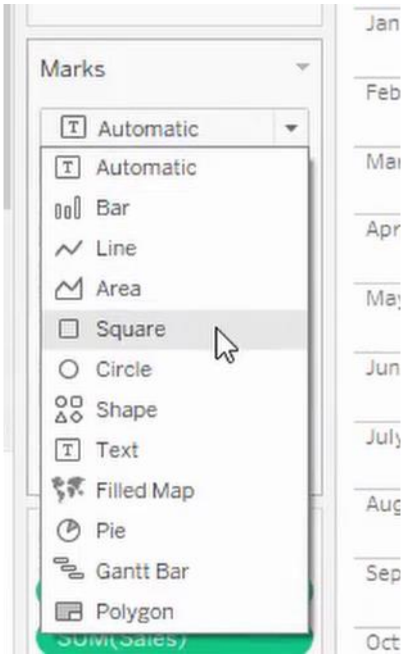
This looks nice, but I'm worried that profits for our Office Supplies weren't good during our sale and into the end of the year. Let's add profit to the crosstab and find out how we're doing. Adding Profits to color gives us a clearer understanding of overall trends.



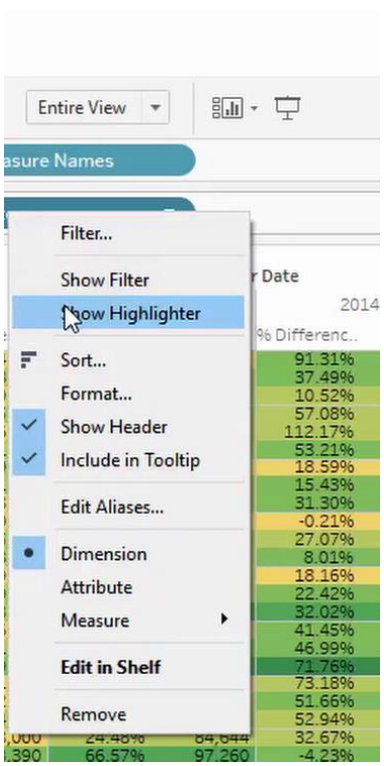
These colors are a bit pale, though, so let's edit how we display this. We'll click on color and click "Edit Colors". Here, we can choose from a wide variety of colors in the dropdown menu, I like green-gold, and we'll use stepped colors and make 6 of them.



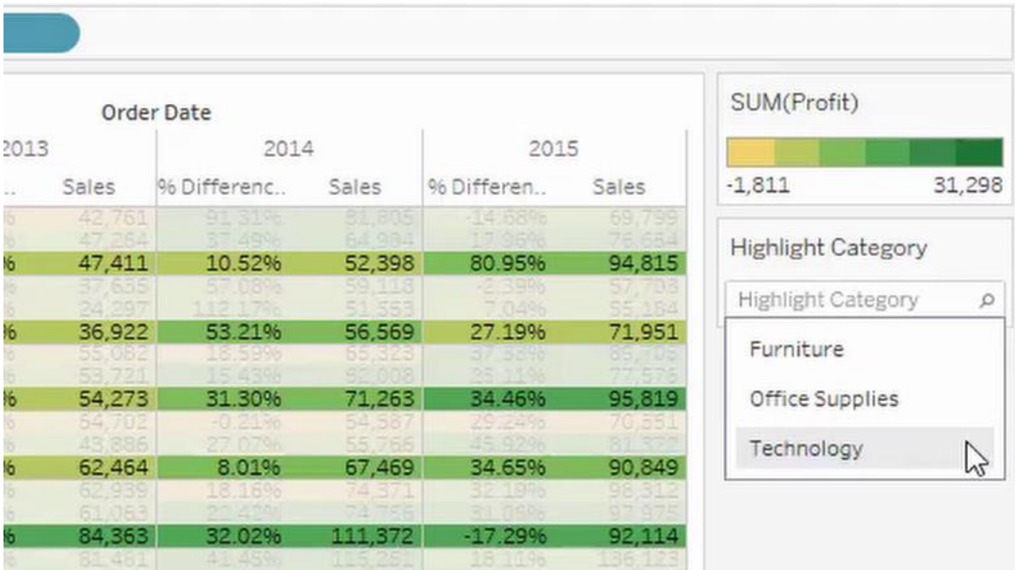
Let's change the mark type to square and turn on mark labels. Now we have a highlight table for profits.



We can right click on the Category pill and select Show Highlighter – if we select Office Supplies, we can see that the fall of 2015 is dark green, so our profits for those months are strong. Great!



Hovering over those categories in the highlighter, we can quickly see that although our fall profits are doing well in technology and Office supplies, furniture doesn't have that same dark green upswing in profits.



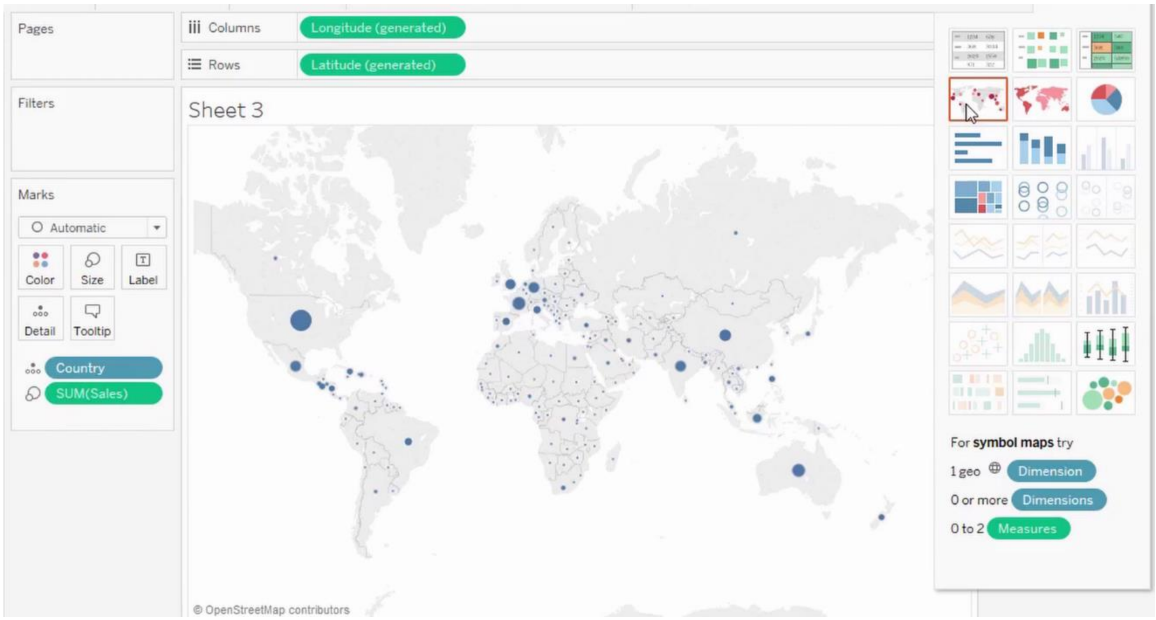
Is this happening across all stores? Let's find out! We'll double click on the sheet tab and rename this sheet "Crosstab" and create a new sheet.

Show Me

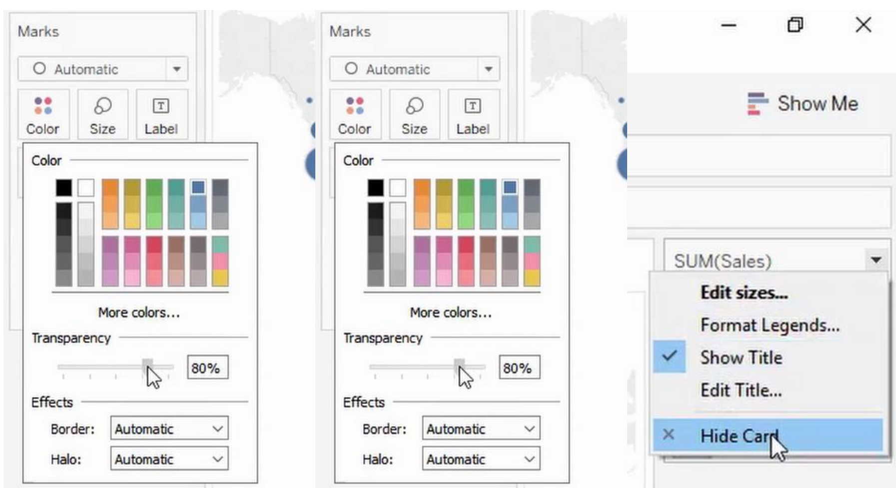
We know that furniture's profits are bad, and we think this may vary regionally. But we don't necessarily know the best way to view the data. Tableau Desktop provides a simple tool called "Show Me" to help in cases where we know the data we want to look at, but don't know how to create an effective view. "Show Me" contains a list of common chart types that can help you start your analysis.

Note: it's possible to build an enormous variety of charts in Tableau – Show Me is the one-click options, not a comprehensive list of possibilities.

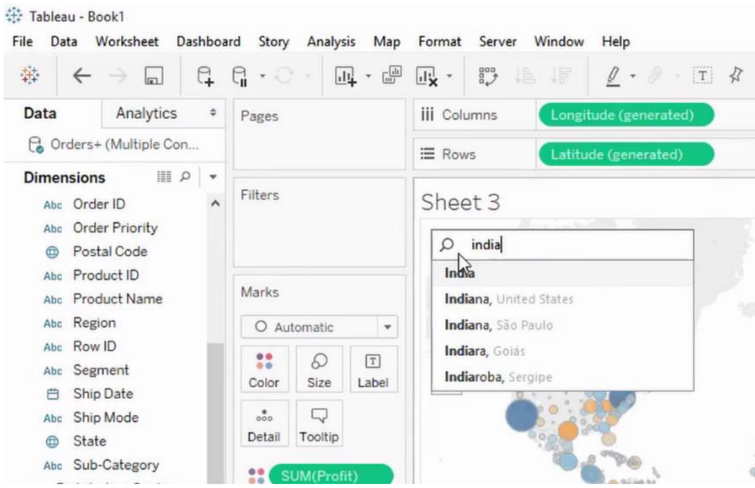
Let's see Show Me at work by selecting different dimensions and measures while holding down the control key. We're curious about our Sales, and how they're doing in different Countries. Notice how different chart types come available based on what measures and dimensions we've chosen. Symbol maps look like a good choice for these fields.

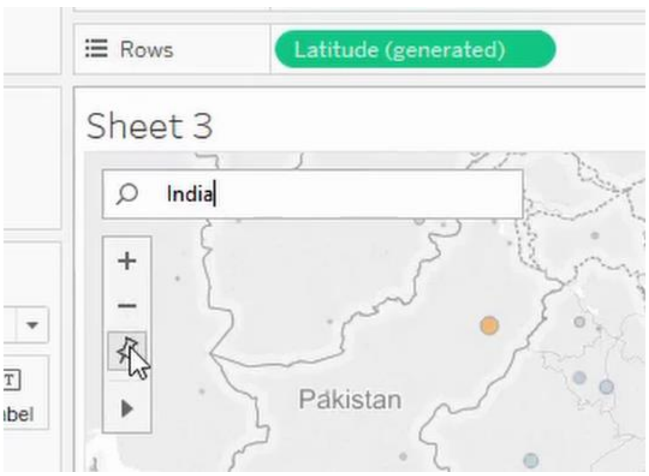


Let's also add State. We can increase the size of these dots by clicking on the size shelf, let's also adjust the transparency and add some borders. We'll hide the size legend, and let's color these states by Profits.



Note that we can do geographic search here – if we want to see how profits are doing in a certain location, we can navigate right to it. Let's unpin to zoom back out.

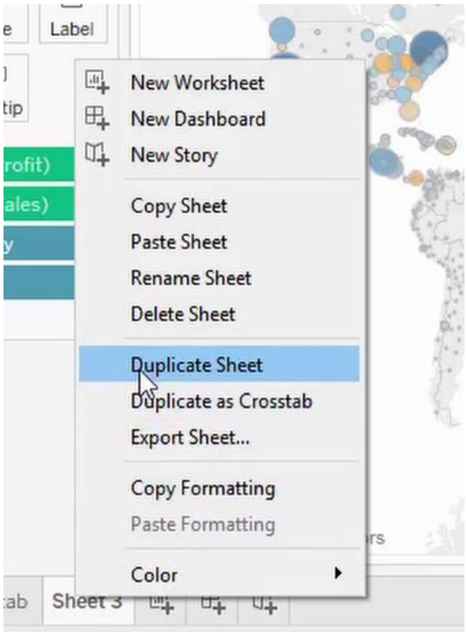




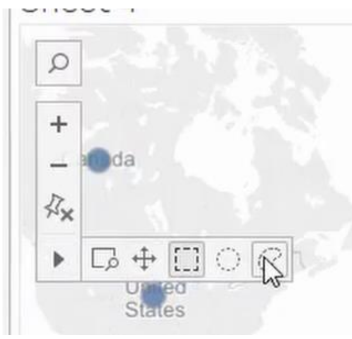
Custom Territories

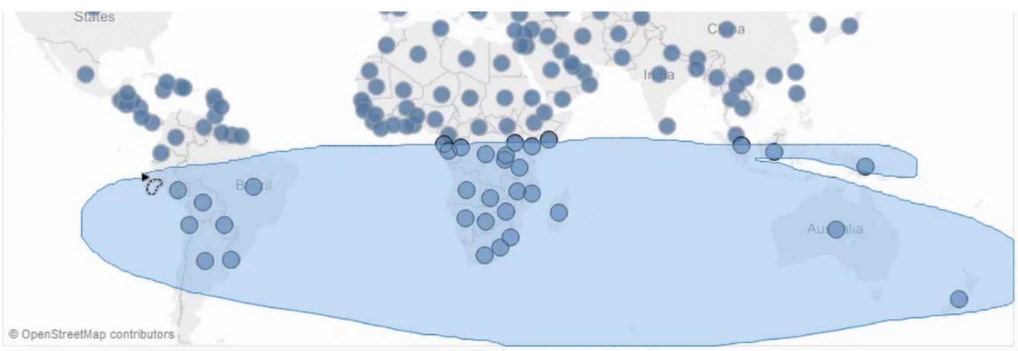
Now, we’re a global company, and there’s that dip in sales in July. Is that because of an action of ours, driven from headquarters, or is that a seasonal effect? We could tell by breaking up our sales over time by Hemisphere, but we don’t have that field in the data. However, we can create that custom territory ourselves, directly in the map.

Let’s right click and duplicate this sheet, so we can leave our original view intact. We can simplify the view, stripping out everything but Country.

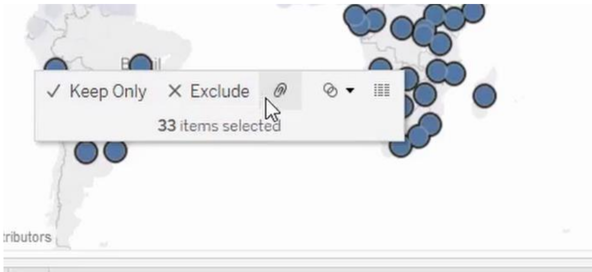


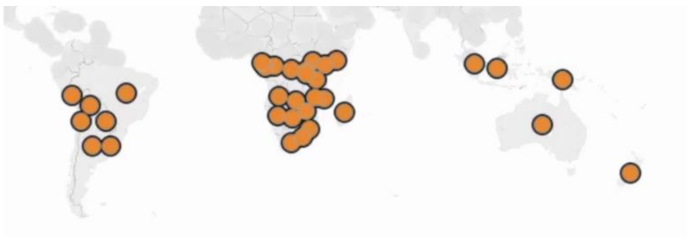
Next, we'll use the lasso select tool and lasso marks covering the approximate southern hemisphere—note this is very rough.



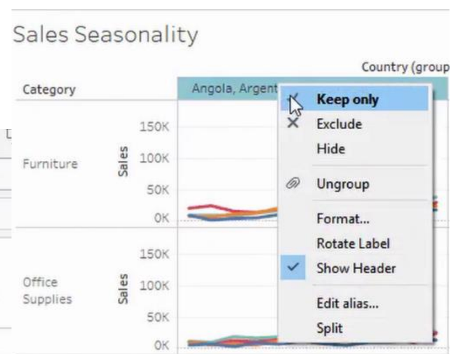


Clicking the paperclip icon in the tooltip creates a group for those countries, and we've made a new field in the data pane.





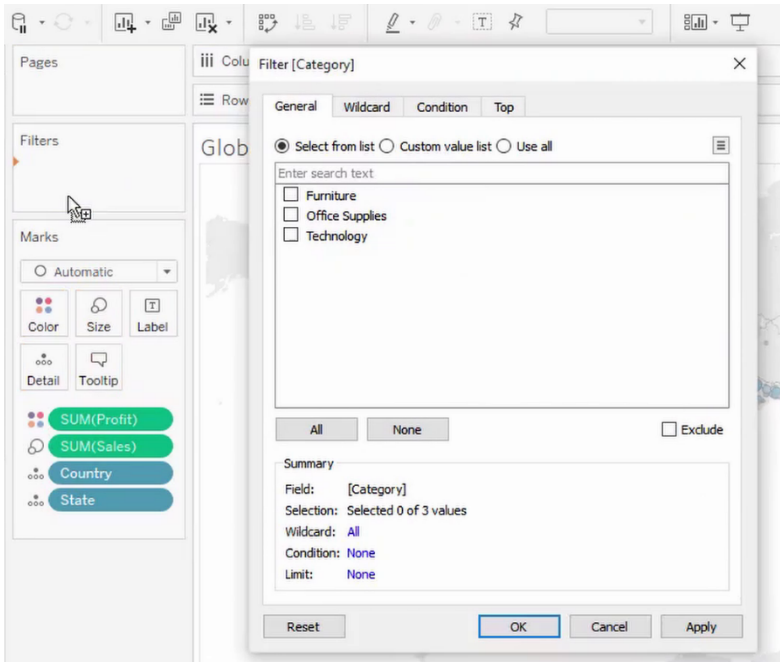
If we go back to our sales seasonality tab and add this new field to Columns…it looks like we have less revenue overall from the southern hemisphere, but if we keep only this column, there's no clear evidence of seasonality. Good to know!

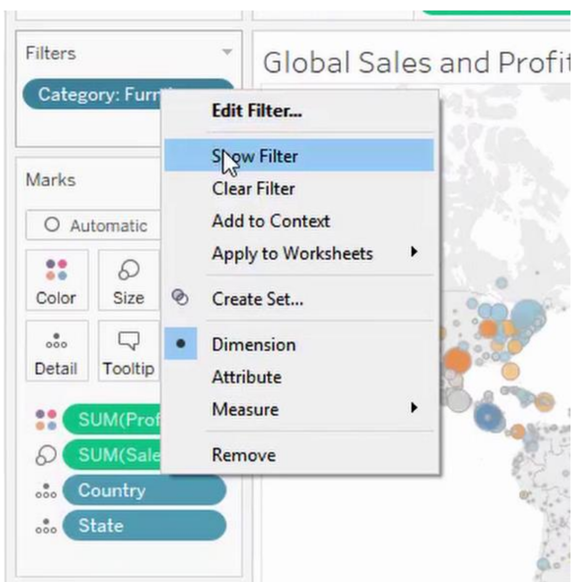


We can leave this avenue of analysis—and even delete this sheet and head back to our original map. We'll name it "Global Sales and Profits"

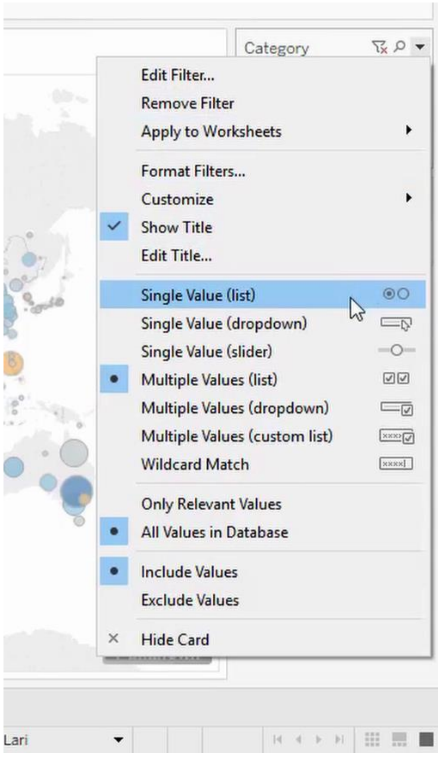
Filters

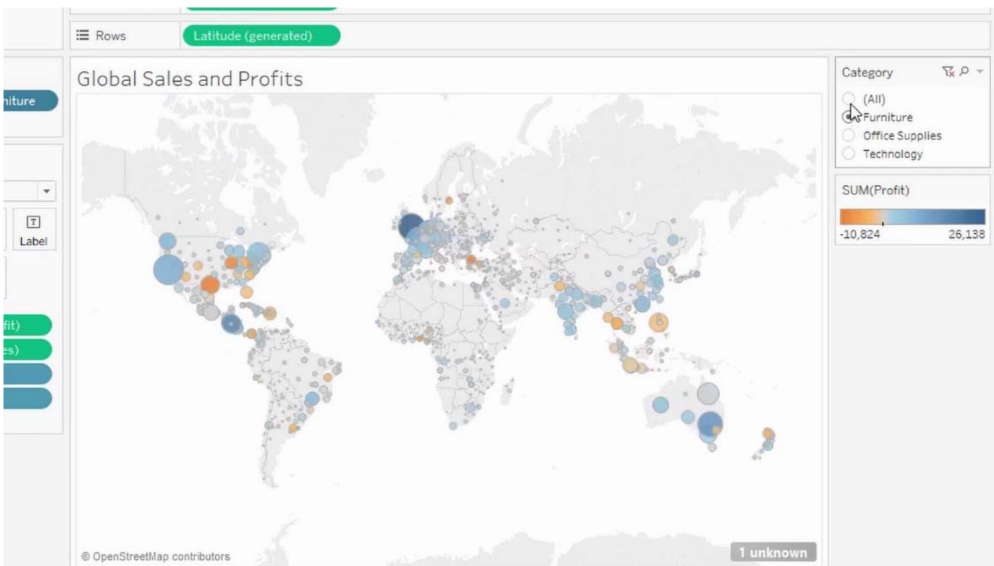
Earlier, we found that furniture had poor profits. To investigate this further, let's drag Category to the filters shelf. We'll choose Furniture. To make this an interactive filter, we'll right click the pill and select "Show Filter".





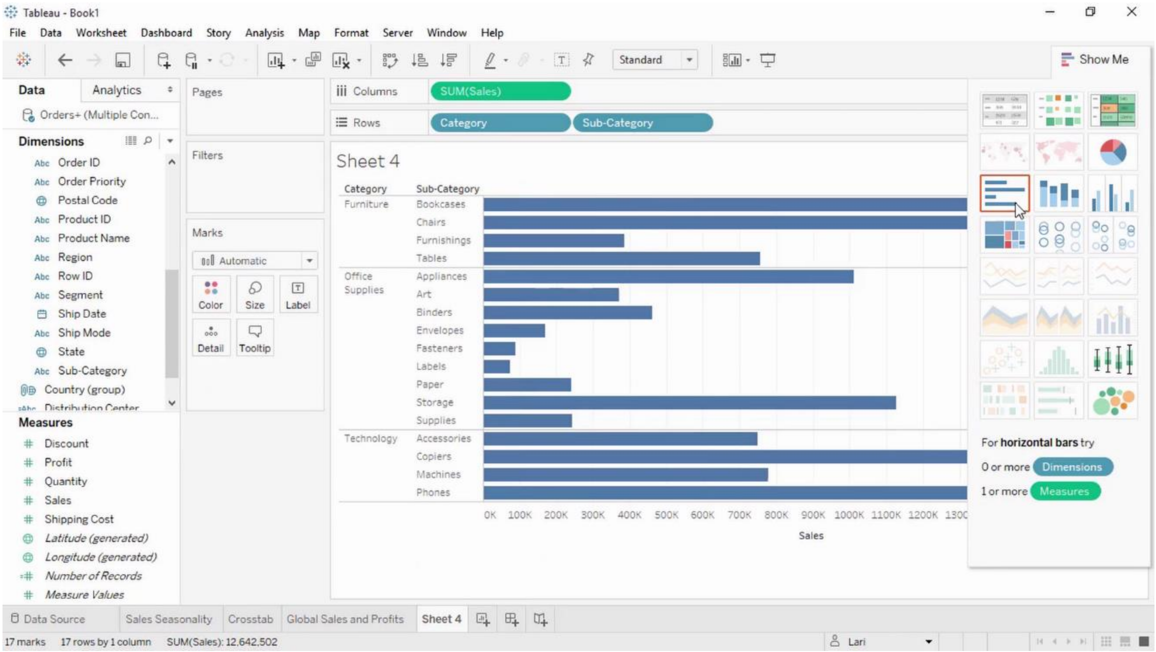
We can also modify filters by selecting their drop-down menu and choosing from a variety of options. Here we'll choose "Single Value List". Now anyone can easily choose the categories they're interested in, such as Furniture or Technology.





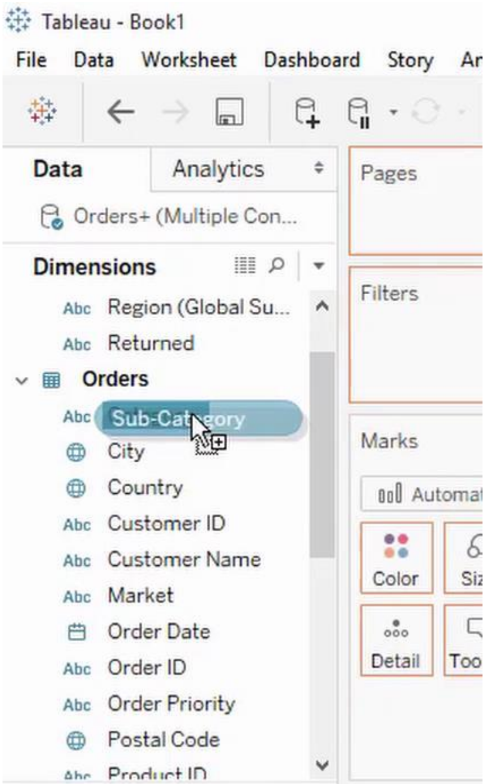
Bar Chart

So we know we have problems with furniture, but what types of furniture are doing poorly? Let's create a new sheet, and use Show Me to find out. Again, as we hold down control and select the variables we're interested in such as Category, Sub-Category and Sales, we see Show Me making various suggestions. We can click through a few charts to see which one looks best.

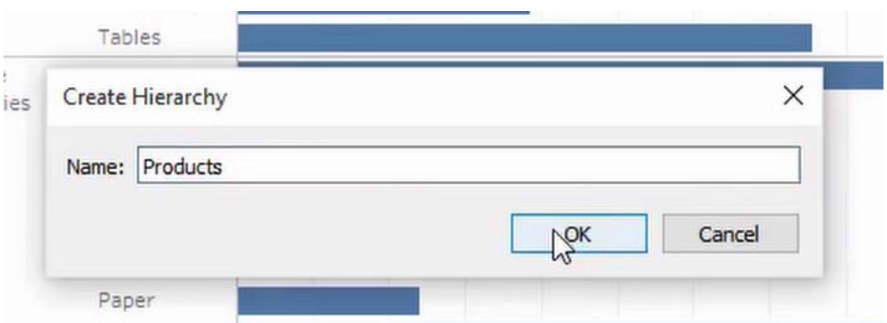


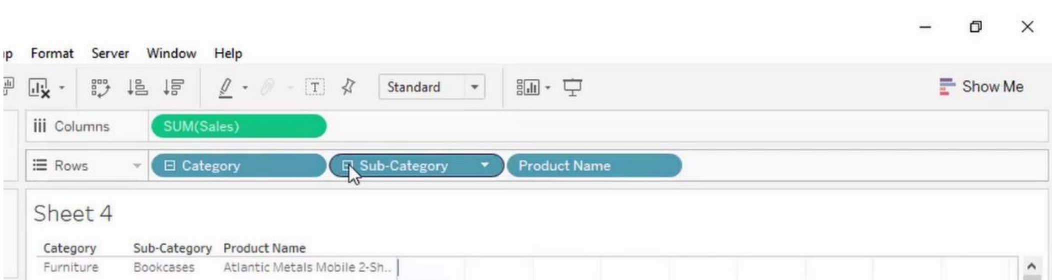
Hierarchies

There is a hierarchical nature between Category and Sub-Category in our data. In Tableau Desktop, we can create hierarchies by simply dragging and dropping fields on top of each other in the data pane. Let’s drag Sub-Category on top of Category and we’ll call this “Products”.



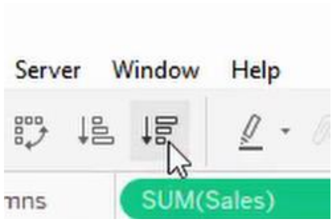
We can add Product Name to this hierarchy as well. Creating this hierarchy in Tableau Desktop only takes seconds and gives us full drill down capabilities.





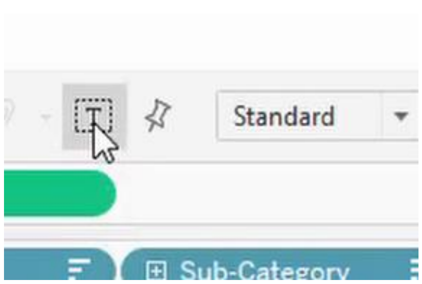
Sorting

To sort the three Categories by overall sales, we can click the appropriate sort button in the toolbar.

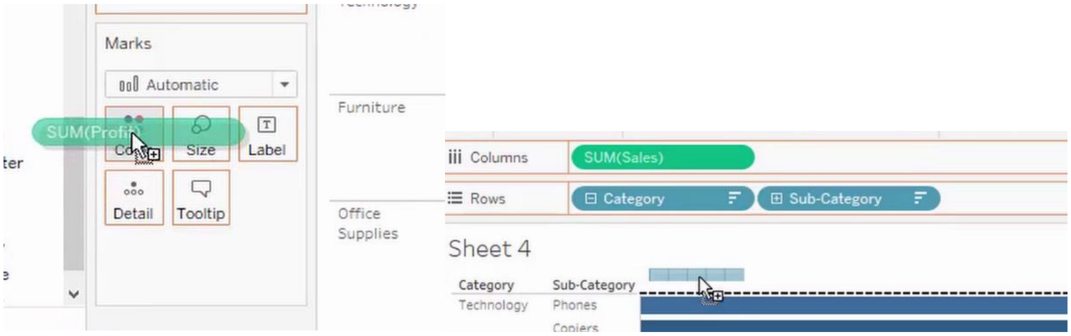


Now we see that technology has the most total sales. If we expand out to see Sub-Category, we see that those bars aren't sorted. Let's sort again, this time using a quick sort from the axis, like so – note that the order of categories stayed the same and we're only sorting the bars WITHIN each category.

We can see the actual sales values by clicking on the "T" button in the toolbar to turn on or off the mark labels.

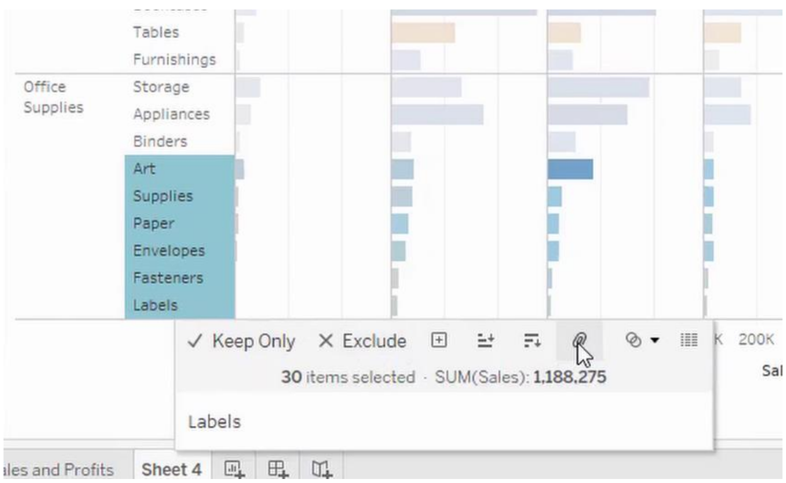


But again, how's profits? Let's place Profits on Color. We quickly see that Tables are doing poorly from a profitability standpoint, despite how good the sales looked. Is this happening across all our markets? Let's place Market here on the top. We quickly see that several markets seem to be having this same profitability problem when it comes to furniture.

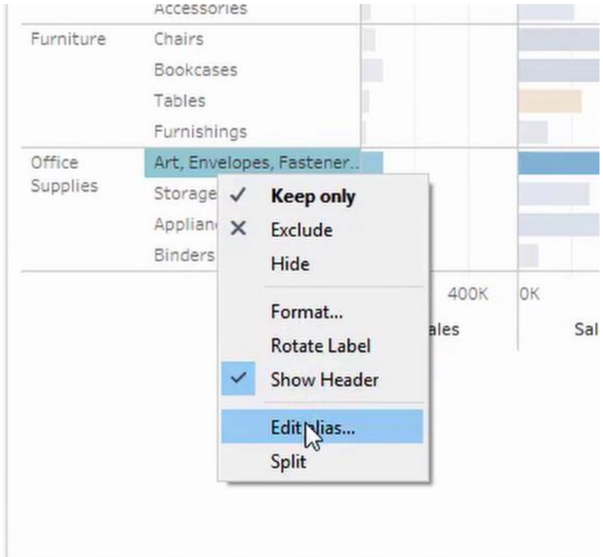


Grouping

In this view, it's useful to note that we can group items together. We see in Office Supplies that several items have very small sales. We can select the headers and group them using the paperclip icon.

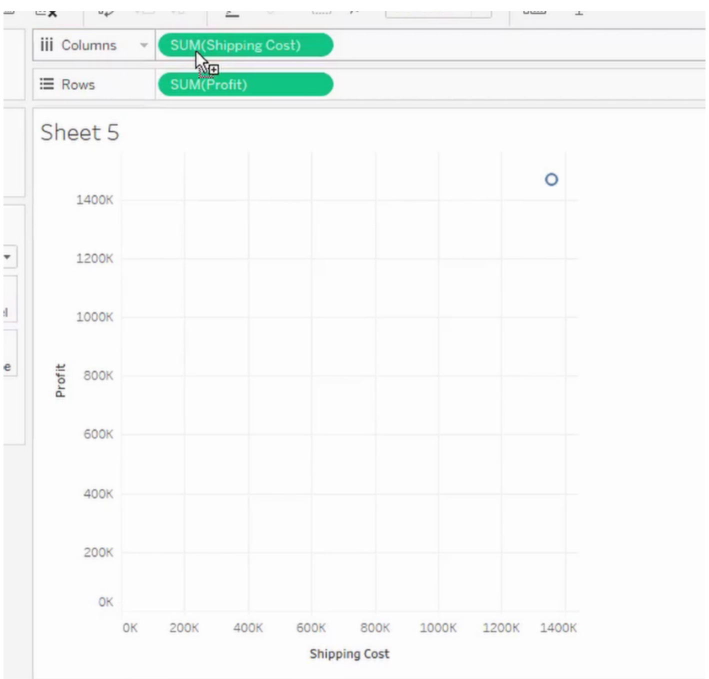


To rename that row, right click and select Edit Alias. Let's remove Market again and swap the axes. We can also right click on the header for columns and hide that label. Let's call this sheet, "Sales by Sub-Category" and create a new sheet.

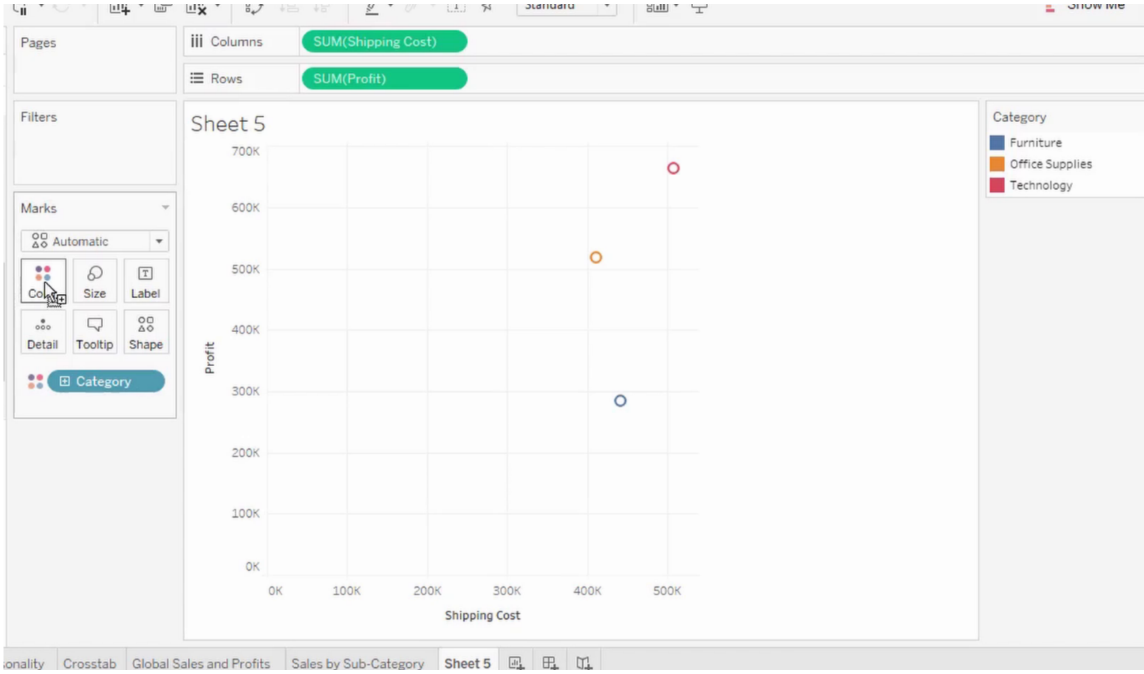


Working with Marks

We've seen that we have some profitability issues, and I have a hunch that this may be due to shipping costs eating into our profits. Let's take a look at our profits and shipping numbers. We'll place: Profits on the Rows shelf and Shipping Cost on the Columns shelf. Tableau makes a mark for the sum of profits and shipping cost.

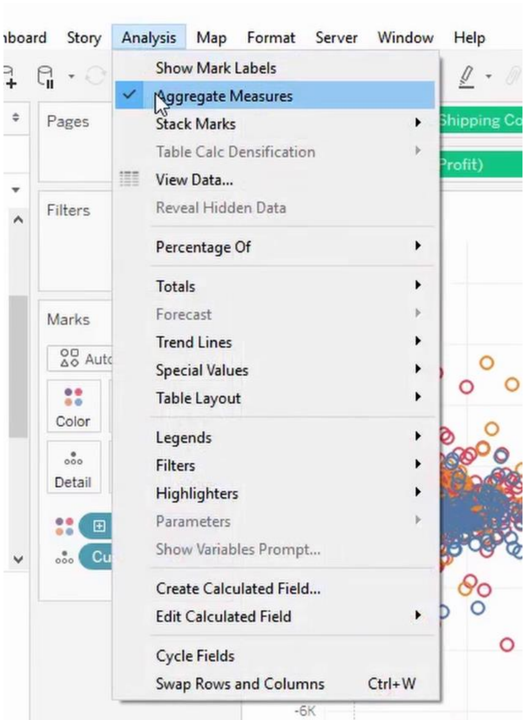


If we put Category on Color. That first mark is broken out by category and we wind up with 3 marks.

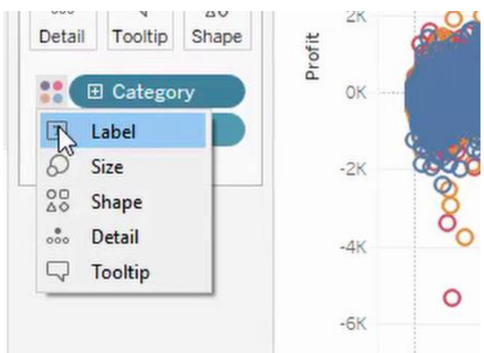


And if we add Customer ID onto Detail, Tableau makes a mark for each customer for each category. These marks represent the total shipping cost and profits for all transactions within a single category for each customer.

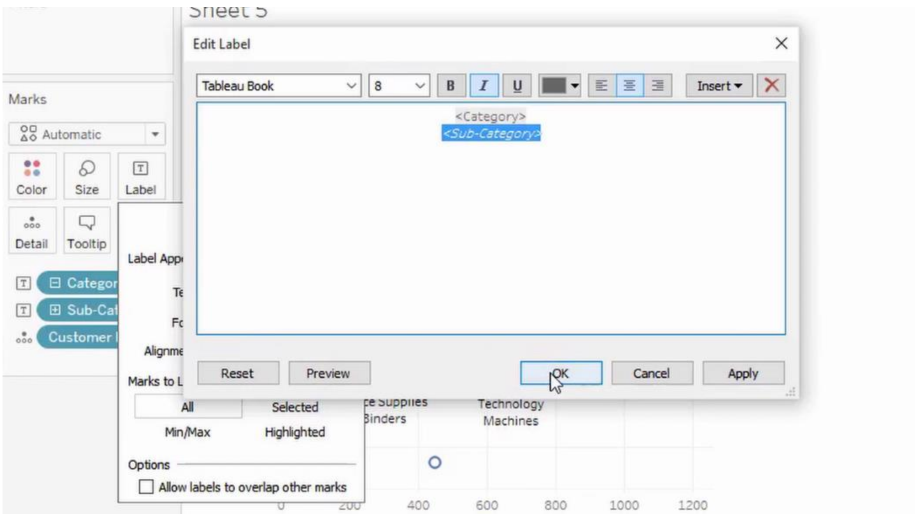
We could also fully disaggregate our data to plot each and every transaction at the record level.



We can assign fields on the Marks card to different roles. For instance, we can click on the color icon in front of Category and change it to Label.

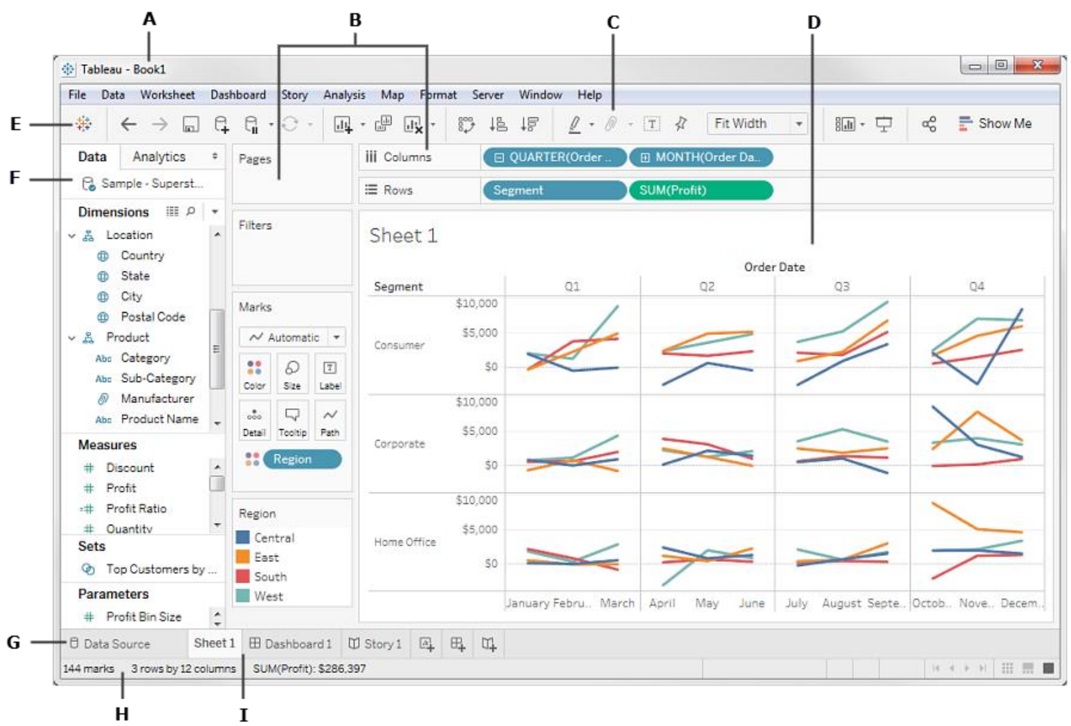


We can bring fields directly to the label shelf, such as Sub-Category. We can edit this label by clicking and then again by text and modifying as we see fit.



From here, we can see that we have a significant number of customers with low profits in various categories, so there's definitely something worth looking into. I wonder if those low profits orders were returned. We can bring "returned" to Size. It looks like the mark with the highest shipping cost was returned, but not the low profits orders.

Tableau Workspace Area



A. Workbook name. A workbook contains sheets.

B. Cards and shelves - Drag fields to the cards and shelves in the workspace to add data to your view.

C. Toolbar - Use the toolbar to access commands and analysis and navigation tools.

D. View - This is the canvas in the workspace where you create a visualization (also referred to as a "viz").

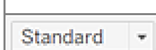
E. Click this icon to go to the Start page, where you can connect to data.

F. Side Bar - In a worksheet, the side bar area contains the Data pane and the Analytics pane.

G. Click this tab to go to the Data Source page and view your data.

H. Status bar - Displays information about the current view.

Tableau Toolbar Button Reference

BUTTON	DESCRIPTION
	**Undo**: Reverses the most recent action in the workbook. You can undo an unlimited number of times, back to the last time you opened the workbook, even after you have saved.
	**Redo**: Repeats the last action you reversed with the **Undo** button. You can redo an unlimited number of times.
	**Save**: In Tableau Desktop, saves the changes made to the workbook.
	**New Worksheet**: Creates a new blank worksheet, use the drop-down menu to create a new worksheet, dashboard, or story.
	**Duplicate**: Creates a new worksheet containing the same view as the current sheet.
	**Clear**: Clears the current worksheet. Use the drop-down menu to clear specific parts of the view such as filters, formatting, sizing, and axis ranges.
	**Swap**: Moves the fields on the **Rows** shelf to the **Columns** shelf and vice versa. The **Hide Empty Rows** and **Hide Empty Columns** settings are always swapped with this button.
	**Sort Ascending**: Applies a sort in ascending order of a selected field based on the measures in the view.
	**Sort Descending**: Applies a sort in descending order of a selected field based on the measures in the view.
	**Highlight**: Turn on highlighting for the selected sheet. Use the options on the drop-down menu to define how values are highlighted.
	**Group Members**: Creates a group by combining selected values. When multiple dimensions are selected, use the drop-down menu to specify whether to group on a specific dimension or across all dimensions.
	**Show Mark Labels**: Switches between showing and hiding mark labels for the current sheet.
	**Fix Axes**: switches between a locked axis that only shows a specific range and a dynamic axis that adjusts the range based on the minimum and maximum values in the view.
	**Fit**: Specifies how the view should be sized within the window. Select Standard, Fit Width, Fit Height, or Entire View. **Note**: This menu is not available in

BUTTON	DESCRIPTION
	geographic map views.
	The **Cell Size** commands have different effects depending on the type of visualization. To access the **Cell Size** menu in Tableau Desktop click **Format** > **Cell Size**.
	**Show/Hide Cards**: Shows and hides specific cards in a worksheet. Select each card that you want to hide or show on the drop-down menu.
	**Show Me**: Helps you choose a view type by highlighting view types that work best with the field types in your data. An orange outline shows around the recommended chart type that is the best match for your data.

"Show Me" Charting Reference

Below is a table of all of the chart available using the "Show Me" chart tool, as well as the required data types. "Show Me" provides quick access to a limited number of chart types, and there are much more than can be created using the Tableau System.

"Show Me" Chart	Requirements
Text Tables	1+ Dimensions or 1+ Measures
Heat Map	1+ Dimensions, 1 or 2 Measures
Highlight Tables	1+ Dimensions, 1 Measure
Symbol Maps	1 Geo Dimension, 0+ Other Dimensions, 0 - 2 Measures
Maps	1 Geo Dimension, 0+ Other Dimensions, 0/1 Measures
Pie Chart	1+ Dimensions, 1 or 2 Measures
Horizontal Bar	0+ Dimensions, 1+ Measures
Stacked Bar	1+ Dimensions, 1+ Measures
Side-by-Side Bars	1+ Dimensions, 1+ Measures
Treemaps	1+ Dimensions, 1 or 2 Measures
Circle Views	1+ Dimensions, 1+ Measures
Side-by-Side Circles	1+ Dimensions, 1+ Measures
Lines (continuous)	1 Date, 0+ Dimensions, 1+ Measures
Lines (discrete)	1 Date, 0+ Dimensions, 1+ Measures
Dual Lines	1 Date, 0+ Dimensions, 2 Measures
Area Charts (continuous)	1 Date, 0+ Dimensions, 1+ Measures

"Show Me" Chart	Requirements
Area Charts (discrete)	1 Date, 0+ Dimensions, 1+ Measures
Dual Combination	1 Date, 0+ Dimensions, 2 Measures
Scatter Plots	0+ Dimensions, 2 - 4 measures
Histogram	1 Measure (not available for all measures such as Longitude/Latitude)
Box-and-Whisker	0+ Dimensions, 1+Measures (disaggregated) or 1+Dimensions and 1+Measure
Gantt	1 date, 1+ Dimensions, 0 - 2 Measures
Bullet Graphs	0+ Dimensions, 2 Measures
Packet Bubbles	1+ Dimensions, 1 or 2 Measures

Tableau "Hey Viz"

In addition to the desktop interface, you can also interact with the system using your voice and the "Hey Viz" button located in the bottom right of the screen.

Below are some examples of requests you can make to the system using "Hey Viz". All of the workspace and functions available in Tableau are accessible using the voice system, including color, size, text, as well as all of the filters and functions covered previously in the tutorial. The system can respond to a wide range of complex requests to fit your needs, though this may increase the amount of time needed to process. When making requests, it is important to note that the system will default to whatever visualization Tableau has set as default unless otherwise specified, and as long as the chart chosen is a chart in "Show Me".

The system will make every attempt to answer your request. In cases where the system returns an error, you will see "Error" pop in the corner. The most common error is that the system could not understand what you have requested, or that you have requested a particular chart that does not match the data type you have provided.

Example Requests

Show me **Sales** by **Category**

Show me **Profit** by **Market** and **Category**

Which **Category** had the most **Sales** in **2012**?

Show me this graph **over time**

Show me **quarterly Sales**

Just show **2013**

Filter on **Categories**

Which **Countries** have the lowest **Profit**?

Show me **Sales** by **Market** as a **Bubble Chart**

Show me **monthly Sales**

Move Market to **Columns**

Move Market to **Filter**

Duplicate this sheet as a **crosstab**

Show me India on this map

Create a **new sheet**

Change **Furniture** to **Yellow**

Appendix H - Interaction Annotations

The table below presents all of the functions annotated from the experimental sessions, broken down by their total usage and their usage by each experimental group. Functions are grouped into categories that are also compared.

Function		Total	Voice	Non-Voice
scroll		**2,647**	**1,720**	**927**
	scroll_down	1,100	708	392
	scroll_left	357	256	101
	scroll_right	636	422	214
	scroll_up	554	334	220
variable_add		**974**	**472**	**502**
	color_add	79	40	39
	column_add	346	154	192
	detail_add	38	19	19
	label_add	23	5	18
	row_add	418	212	206
	shape_add	1	1	-
	size_add	21	4	17
	text_add	45	37	8
	tooltip_add	3	-	3
click		**610**	**261**	**349**
	axis_click	4	1	3

Function	Total	Voice	Non-Voice
data_click	74	51	23
menu_click	93	37	56
sheet_click	6	4	2
title_click	1	1	-
variable_click	432	167	265
sheet	**489**	**263**	**226**
sheet_add	83	49	34
sheet_clear	6	4	2
sheet_duplicate	36	17	19
sheet_duplicate_crosstab	9	7	2
sheet_remove	29	14	15
sheet_rename	57	33	24
sheet_select	269	139	130
variable_remove	**403**	**176**	**227**
color_remove	44	25	19
column_remove	159	72	87
detail_remove	19	12	7
label_remove	11	3	8
row_remove	148	56	92
shape_remove	1	1	-
size_remove	11	4	7
text_remove	7	3	4
tooltip_remove	3	-	3
variable_move	**305**	**117**	**188**

Function	Total	Voice	Non-Voice
show_me_chart	**265**	**114**	**151**
filter	**155**	**83**	**72**
filter_add	62	42	20
filter_change	11	7	4
filter_remove	14	7	7
filter_select	57	20	37
filter_show	11	7	4
show_me_toggle	**146**	**73**	**73**
undo_redo	**114**	**45**	**69**
redo	1	1	-
undo	113	44	69
hey_viz_click	**85**	**85**	**-**
zoom	**77**	**4**	**73**
location_show	4	1	3
zoom_in	25	1	24
zoom_out	48	2	46
variable_sort	**69**	**28**	**41**
variable_measure_change	**48**	**22**	**26**
object_adjust	**44**	**41**	**3**
object_hide	34	33	1
object_show	1	1	-
object_size_bigger	4	3	1
object_size_small	1	-	1
title_hide	4	4	-

Function	Total	Voice	Non-Voice
label_toggle	**25**	**14**	**11**
date_group	**25**	**20**	**5**
date_group_change	1	1	-
date_group_deeper	18	13	5
date_group_shallower	6	6	-
color_edit	**22**	**13**	**9**
group	**20**	**13**	**7**
group_add	4	3	1
group_make	12	7	5
group_rename	4	3	1
data_source	**20**	**2**	**18**
axis_swap	**18**	**5**	**13**
window_fit	**15**	**6**	**9**
quick_calculation	**14**	**6**	**8**
quick_calculation_add	10	6	4
quick_calculation_edit	2	-	2
quick_calculation_remove	2	-	2
lasso_select	**12**	**10**	**2**
title_edit	**10**	**-**	**10**
mark_change	**9**	**3**	**6**
highlight	**9**	**6**	**3**
highlight_add	3	2	1
highlight_select	6	4	2
caption_edit	**8**	**-**	**8**

Function	Total	Voice	Non-Voice
calculated_field	**8**	**5**	**3**
calculated_field_add	5	2	3
calculated_field_edit	1	1	-
calculated_field_remove	2	2	-
forecast_add	**5**	**2**	**3**
annotation_add	**4**	**4**	**-**
story	**4**	**-**	**4**
story_add	2	-	2
story_remove	2	-	2
aggregation	**3**	**3**	**-**
measure_aggregate	1	1	-
measure_disaggregate	2	2	-
data	**2**	**-**	**2**
data_select	1	-	1
data_view	1	-	1
variable_describe	**2**	**-**	**2**
label_edit	**2**	**2**	**-**
percent_of	**2**	**1**	**1**
variable_rename	**1**	**1**	**-**
variable_edit	**1**	**-**	**1**
table_calculation_add	**1**	**1**	**-**

Appendix I – Mixed ANOVA Summary for Top 10 Functions

Function: variable_add

	No Voice	Voice				
Task 1	13.08	12.17				
Task 2	18.00	14.42				
	SS	**DF1**	**DF2**	**MS**	**F**	**p-value**
VOICE	60.75	1	22	60.75	0.64	0.43
Task	154.10	1	22	154.08	1.97	0.17
Interaction	21.33	1	22	21.33	0.27	0.61

Function: click

	No Voice	Voice				
Task 1	8.67	5.42				
Task 2	11.83	9.92				
	SS	**DF1**	**DF2**	**MS**	**F**	**p-value**
VOICE	80.08	1	22	80.08	0.42	0.53
Task	176.33	1	22	176.33	3.62	0.07
Interaction	5.33	1	22	5.33	5.33	0.11

Function: variable_remove

	No Voice	Voice
Task 1	5.83	4.25
Task 2	8.83	6.42

	SS	DF1	DF2	MS	F	p-value
VOICE	48.00	1	22	48.00	1.02	0.32
Task	80.08	1	22	80.08	2.36	0.14
Interaction	2.08	1	22	2.08	0.06	0.81

Function: show_me_chart

	No Voice	Voice
Task 1	6.75	4.67
Task 2	4.17	4.17

	SS	DF1	DF2	MS	F	p-value
VOICE	13.02	1	22	13.02	0.20	0.66
Task	28.52	1	22	28.52	1.12	0.30
Interaction	13.02	1	22	13.02	0.51	0.48

Function: variable_move

	No Voice	Voice				
Task 1	4.50	3.59				
Task 2	8.08	2.92				
	SS	**DF1**	**DF2**	**MS**	**F**	**p-value**
VOICE	111.02	1	22	111.02	1.82	0.19
Task	25.52	1	22	25.52	1.31	0.26
Interaction	54.19	1	22	54.19	2.79	0.11

Function: sheet_select

	No Voice	Voice				
Task 1	3.58	6.25				
Task 2	5.33	3.33				
	SS	**DF1**	**DF2**	**MS**	**F**	**p-value**
VOICE	1.33	1	22	1.33	0.02	0.90
Task	4.08	1	22	4.08	0.08	0.08
Interaction	65.33	1	22	65.33	1.24	0.28

Function: filter

	No Voice	Voice				
Task 1	0.83	2.83				
Task 2	3.67	1.92				
	SS	**DF1**	**DF2**	**MS**	**F**	**p-value**
VOICE	0.19	1	22	1.33	0.01	0.90
Task	11.02	1	22	4.08	0.82	0.37
Interaction	42.19	1	22	65.33	3.14	0.09

Function: sheet_add

	No Voice	Voice				
Task 1	1.67	2.08				
Task 2	2.17	1.92				
	SS	**DF1**	**DF2**	**MS**	**F**	**p-value**
VOICE	0.08	1	22	0.08	0.01	0.91
Task	0.33	1	22	0.33	0.08	0.79
Interaction	1.33	1	22	1.33	0.30	0.59

Function: undo_redo

	No Voice	Voice
Task 1	1.25	1.17
Task 2	3.08	1.25

	SS	DF1	DF2	MS	F	p-value
VOICE	11.02	1	22	11.02	0.92	0.35
Task	11.02	1	22	11.02	1.29	0.27
Interaction	9.19	1	22	9.19	1.07	0.31

Function: variable_sort

	No Voice	Voice
Task 1	1.17	0.50
Task 2	1.83	0.83

	SS	DF1	DF2	MS	F	p-value
VOICE	8.33	1	22	8.33	0.58	0.45
Task	3.00	1	22	3.00	1.23	0.28
Interaction	0.33	1	22	0.33	0.14	0.72

BIBLIOGRAPHY

Alam, S. S., & Jianu, R. (2017). Analyzing Eye-Tracking Information in Visualization and Data Space: From Where on the Screen to What on the Screen. *IEEE Transactions on Visualization and Computer Graphics*, *23*(5), 1492–1505. https://doi.org/10.1109/TVCG.2016.2535340

Andrews, C., & North, C. (2012). Analyst's Workspace: An embodied sensemaking environment for large, high-resolution displays. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 123–131). IEEE. https://doi.org/10.1109/VAST.2012.6400559

Aurisano, J., Kumar, A., Gonzales, A., Reda, K., Leigh, J., Di Eugenio, B., & Johnson, A. (2015). "Show me data": Observational study of a conversational interface in visual data exploration. In *IEEE VIS* (Vol. 15, p. 1). Retrieved from https://pdfs.semanticscholar.org/9a16/0416121331115f6e78774d607f232d3c2d40.pdf

Aurisano, J., Kumar, A., Gonzalez, A., Leigh, J., Di Eugenio, B., & Johnson, A. (2016). Articulate2: Toward a Conversational Interface for Visual Data Exploration. In *IEEE VIS 2016*.

Badam, S. K., Srinivasan, A., Elmqvist, N., Stasko, J., Karthik Badam, S., Srinivasan, A., … Stasko, J. (2017). Affordances of input modalities for visual data exploration in immersive environments. In *Workshop on Immersive Analytics*, *IEEE Vis*. Retrieved from http://www.visbox.com/products/cave/

Bates, M. J. (1990). Where should the person stop and the information search interface start? *Information Processing and Management*, *26*(5), 575–591. https://doi.org/10.1016/0306-4573(90)90103-9

Blandford, A., & Attfield, S. (2010). *Interacting with Information*. *Synthesis Lectures on Human-Centered Informatics* (Vol. 3). https://doi.org/10.2200/S00227ED1V01Y200911HCI006

Blascheck, T., John, M., Kurzhals, K., Koch, S., & Ertl, T. (2016). VA 2 : A Visual Analytics Approach for Evaluating Visual Analytics Applications. *IEEE Transactions on Visualization and Computer Graphics*, *22*(1), 61–70. https://doi.org/10.1109/TVCG.2015.2467871

Bolt, R. A. (1980). “Put-that-there”: Voice and gesture at the graphics interface. *ACM Computer Graphics*. https://doi.org/10.1145/800250.807503

Bourguet, M.-L. (2006). An Overview of Multimodal Interaction Techniques and Applications. In *Encyclopedia of Human Computer Interaction* (pp. 451–456). IGI Global. https://doi.org/10.4018/978-1-59140-562-7.ch068

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*. https://doi.org/10.1016/j.cognition.2011.05.011

Chen, F. (2006). *Designing human interface in speech technology*. Springer Science & Business Media.

Christian, K., Kules, B., Shneiderman, B., & Youssef, A. (2004). A comparison of voice controlled and mouse controlled web browsing. https://doi.org/10.1145/354324.354345

Cook, K. A., & Thomas, J. J. (2005). Illuminating the Path: The Research and Development Agenda for Visual Analytics. Retrieved from https://www.hsdl.org/?abstract&did=485291

Corbett, E., & Weber, A. (2016). What can I say? Addressing User Experience Challenges of a Mobile Voice User Interface for Accessibility. *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*, 72–82. https://doi.org/10.1145/2935334.2935386

Cox, K., Grinter, R. E., Hibino, S. L., Jagadeesan, L. J., & Mantilla, D. (2001). A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology*. https://doi.org/10.1023/A:1011368926479

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies — why and how.

Knowledge-Based Systems, *6*(4), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

De Liddo, A., Sándor, Á., & Buckingham Shum, S. (2012). Contested Collective Intelligence: Rationale, Technologies, and a Human-Machine Annotation Study. *Computer Supported Cooperative Work (CSCW)*, *21*(4–5), 417–448. https://doi.org/10.1007/s10606-011-9155-x

De Marneffe, M. C., MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*.

Diesendruck, M., & Zhao, H. (2016). *Data Visualization By Voice: Syntax, Parsing, and Recognition Techniques*. Retrieved from https://github.com/diesendruck/

Dimara, E., & Perin, C. (2020). What is Interaction for Data Visualization? *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 119–129. https://doi.org/10.1109/TVCG.2019.2934283

Duchnowski, P., Meier, U., & Waibel, A. (1994). See me, hear me: Integrating automatic speech recognition and lip-reading. In *Third International Conference on Spoken Language Processing*.

Ebert, D., Gaither, K., & Gilpin, C. (2008). *Enabling science discoveries through visual exploration*. Washington, D.C.

Endert, A., Bradel, L., & North, C. (2013). Beyond control panels: Direct manipulation for visual analytics. *IEEE Computer Graphics and Applications*, *33*(4), 6–13. https://doi.org/10.1109/MCG.2013.53

Endert, A., Fiaux, P., & North, C. (2012). Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 473–482). https://doi.org/10.1145/2207676.2207741

Endert, A., Hossain, M. S., Ramakrishnan, N., North, C., Fiaux, P., & Andrews, C. (2014). The human is the loop: new directions for visual analytics. *Journal of Intelligent Information Systems*, *43*(3), 411–435. https://doi.org/10.1007/s10844-014-0304-9

Feng, J., Karat, C.-M., & Sears, A. (2004). How productivity improves in hands-free continuous dictation tasks: lessons learned from a longitudinal study. *Interacting with Computers*, *17*(3), 265–289.

Fintor, E., Stephan, D. N., & Koch, I. (2018). Emerging features of modality mappings in task switching: modality compatibility requires variability at the level of both stimulus and response modality. *Psychological Research*, *82*(1), 121–133. https://doi.org/10.1007/s00426-017-0875-5

Fraser, N. M., & Gilbert, G. N. (1991). Simulating speech systems. *Computer Speech & Language*, *5*(1), 81–99. https://doi.org/10.1016/0885-2308(91)90019-M

Gao, T. . T., Dontcheva, M. . M. M. ., Adar, E. . E. E. ., Liu, Z. . Z., & Karahalios, K. . K. G. (2015). Datatone: Managing ambiguity in natural language interfaces for data visualization. *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology - UIST '15*, 489–500. https://doi.org/10.1145/2807442.2807478

Grammel, L., Tory, M., & Storey, M.-A. A. (2010). How information visualization novices construct visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *16*(6), 943–952. https://doi.org/10.1109/TVCG.2010.164

Grossman, T., Fitzmaurice, G., & Attar, R. (2009). A survey of software learnability: metrics, methodologies and guidelines. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2009)*. https://doi.org/10.1145/1518701.1518803

Gupta, M., & Bendersky, M. (2015). Information Retrieval with Verbose Queries. *Foundations and Trends® in Information Retrieval*. https://doi.org/10.1561/1500000050

Hanrahan, P. (2006). VizQL: A language for query, analysis and visualization. In *Proceedings of the ACM SIGMOD International Conference on Management of Data* (p. 721). New York, New York, USA: ACM Press. https://doi.org/10.1145/1142473.1142560

Hinckley, K., Yatani, K., Pahud, M., Coddington, N., Rodenhouse, J., Wilson, A., … Buxton, B. (2010). Pen + Touch = New Tools. In *Proceedings of the 23nd annual ACM symposium on*

User interface software and technology. Retrieved from http://iis-lab.org/paper/UIST2010_penplustouch.pdf

Hoque, E., Setlur, V., Tory, M., & Dykeman, I. (2018). Applying Pragmatics Principles for Interaction with Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2017.2744684

Hu, R., Zhu, S., Feng, J., & Sears, A. (2011). Use of speech technology in real life environment. In *International Conference on Universal Access in Human-Computer Interaction* (pp. 62–71).

Hurter, C., Riche, N. H., Drucker, S. M., Cordeil, M., Alligier, R., & Vuillemot, R. (2019). FiberClay: Sculpting Three Dimensional Trajectories to Reveal Structural Insights. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 704–714. https://doi.org/10.1109/TVCG.2018.2865191

Ingwersen, P. E. R. (1992). *Information Retrieval Interaction*. Taylor Graham.

Jaimes, A., & Sebe, N. (2007). Multimodal human-computer interaction: A survey. *Computer Vision and Image Understanding*. https://doi.org/10.1016/j.cviu.2006.10.019

James, F., & Roelands, J. (2002). Voice over Workplace (VoWP): voice navigation in a complex business GUI. In *Proceedings of the fifth international ACM conference on Assistive technologies* (pp. 197–204).

Kang, Y., Gorg, C., & Stasko, J. (2009). Evaluating visual analytics systems for investigative analysis: Deriving design principles from a case study. In *2009 IEEE Symposium on Visual Analytics Science and Technology* (pp. 139–146). IEEE. https://doi.org/10.1109/VAST.2009.5333878

Karsenty, L. (2002). Shifting the design philosophy of spoken natural language dialogue: From invisible to transparent systems. *International Journal of Speech Technology*. https://doi.org/10.1023/A:1015472130944

Kassel, J.-F., & Rohs, M. (2018). Valletto: A Multimodal Interface for Ubiquitous Visual

Analytics. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (p. LBW005).

Kassel, J.-F., & Rohs, M. (2019). Talk to Me Intelligibly: Investigating An Answer Space to Match the User's Language in Visual Analysis. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (pp. 1517–1529).

Keim, D. a, Mansmann, F., Thomas, J., & Keim, D. a. (2010). Visual analytics: how much visualization and how much analytics? *ACM SIGKDD Explorations Newsletter*, *11*(2), 5–8. https://doi.org/10.1145/1809400.1809403

Keim, D., Jörn, K., Ellis, G., & Mansmann, F. (2010). *Mastering the Information Age Solving Problems with Visual Analytics*. *Mastering the Information Age Solving Problems with Visual Analytics*. https://doi.org/10.1016/j.procs.2011.12.035

Kinsbourne, M., & Hicks, R. E. (1978). Functional cerebral space: A model for overflow, transfer and interference effects in human performance. *Attention and Performance VII*, 345–362.

Ko, T., Demirdjian, D., & Darrell, T. (2004). Untethered gesture acquisition and recognition for a multimodal conversational system. https://doi.org/10.1145/958432.958461

Kumar, A., Aurisano, J., Di Eugenio, B., Johnson, A., Gonzalez, A., & Leigh, J. (2016). Towards a Dialogue System that Supports Rich Visualizations of Data. In *Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2016)* (pp. 13–15). Association for Computational Linguistics. Retrieved from http://www.tableau.com/

Kumar, A., Di Eugenio, B., Aurisano, J., Johnson Abeer Alsaiari, A., Flowers, N., Gonzalez, A., … Leigh, J. (2017). Towards Multimodal Coreference Resolution for Exploratory Data Visualization Dialogue: Context-Based Annotation and Gesture Identification *. In *SemDial 2017, The 21st Workshop on the Semantics and Pragmatics of Dialogue*. Retrieved from http://www.tableau.com

Kwon, B. chul, Fisher, B., & Yi, J. S. (2011). Visual analytic roadblocks for novice investigators. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)* (pp. 3–11).

IEEE. https://doi.org/10.1109/VAST.2011.6102435

Lam, H. (2008). A framework of interaction costs in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, *14*(6).

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, *11*(1), 65–100.

Lee, B., Isenberg, P., Riche, N. H., & Carpendale, S. (2012). Beyond Mouse and Keyboard: Expanding Design Considerations for Information Visualization Interactions. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2689–2698. https://doi.org/10.1109/TVCG.2012.204

Li, F., & Jagadish, H. V. (2015). Constructing an interactive natural language interface for relational databases. *Proceedings of the VLDB Endowment*. https://doi.org/10.14778/2735461.2735468

Liu, W. (2010). Natural user interface-next mainstream product user interface. In *2010 IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design 1* (Vol. 1, pp. 203–205).

Liu, Z., & Stasko, J. (2010). Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE Transactions on Visualization & Computer Graphics*, (6), 999–1008.

Luger, E., & Sellen, A. (2016). "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. https://doi.org/10.1145/2858036.2858288

Mackinlay, J. D., Hanrahan, P., & Stolte, C. (2007). Show Me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1137–1144. https://doi.org/10.1109/TVCG.2007.70594

Marriott, K., Schreiber, F., Dwyer, T., Klein, K., Riche, N. H., Itoh, T., … Thomas, B. H. (Eds.).

(2018). *Immersive Analytics* (Vol. 11190). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-01388-2

Mennicken, S., Brillman, R., Thom, J., & Cramer, H. (2018). Challenges and Methods in Design of Domain-Specific Voice Assistants. *2018 AAAI Spring Symposium Series*, 431–435. Retrieved from https://www.aaai.org/ocs/index.php/SSS/SSS18/paper/view/17575/15465

Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM*. https://doi.org/10.1145/77481.77486

Morton, K., Balazinska, M., Grossman, D., & Mackinlay, J. (2014). Support the data enthusiast: Challenges for next-generation data-analysis systems. *Proceedings of the VLDB Endowment*, *7*(6), 453–456. https://doi.org/10.14778/2732279.2732282

Navarretta, C. (2011). Anaphora and gestures in multimodal communication. In *Proceedings of the 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2011), Faro, Portugal, Edicoes Colibri* (pp. 171–181).

Norman, D. (2013). *The Design of Everyday Things: Revised & Expanded Edition*. *The Design of Everyday Things*. https://doi.org/10.15358/9783800648108

Norman, D. A. (2010). The way I see it: Natural user interfaces are not natural. *Interactions*, *17*(3), 6. https://doi.org/10.1145/1744161.1744163

Patnaik, B., Batch, A., & Elmqvist, N. (2019). Information olfactation: Harnessing scent to convey data. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2018.2865237

Pike, W. A., Stasko, J., Chang, R., & O'connell, T. A. (2009). The science of interaction. *Information Visualization*, *8*(4), 263–274.

Pirolli, P., & 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). https://doi.org/10.1007/s13398-014-0173-7.2

Riek, L. D. (2012). Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119–136. https://doi.org/10.5898/JHRI.1.1.Riek

Russell, D. M., Stefik, M. J., Pirolli, P., & Card, S. K. (1993). The cost structure of sensemaking. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '93*, 269–276. https://doi.org/10.1145/169059.169209

Saktheeswaran, A., Srinivasan, A., & Stasko, J. (2020). Touch? Speech? or Touch and Speech? Investigating Multimodal Interaction for Visual Network Exploration and Analysis. *IEEE Transactions on Visualization and Computer Graphics*, *26*(6), 2168–2179. https://doi.org/10.1109/TVCG.2020.2970512

Saraiya, P., North, C., Lam, V., & Duca, K. A. (2006). An insight-based longitudinal study of visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, *12*(6), 1511–1522. https://doi.org/10.1109/TVCG.2006.85

Scaife, M., & Rogers, Y. (1996). External cognition: how do graphical representations work? *International Journal of Human-Computer Studies*, *45*(2), 185–213.

Sedlmair, M., Isenberg, P., Baur, D., & Butz, A. (2011). Information visualization evaluation in large companies: Challenges, experiences and recommendations. *Information Visualization*, *10*(3), 248–266. https://doi.org/10.1177/1473871611413099

Setlur, V., Battersby, S. E., Tory, M., Gossweiler, R., & Chang, A. X. (2016). Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (pp. 365–377).

Shechtman, N., & Horowitz, L. M. (2003). Media inequality in conversation: how people behave differently when interacting with computers and people. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 281–288).

Shneiderman, B. (1981). Direct manipulation: A step beyond programming languages. In *ACM SIGSOC Bulletin* (Vol. 13, p. 143).

Shneiderman, B. (2003). The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *The Craft of Information Visualization*, 364–371. https://doi.org/10.1016/B978-155860915-0/50046-9

Simitsis, A., Koutrika, G., & Ioannidis, Y. (2008). Précis: From unstructured keywords as queries to structured databases as answers. *VLDB Journal*. https://doi.org/10.1007/s00778-007-0075-9

Srinivasan, A., Drucker, S. M., Endert, A., & Stasko, J. (2019). Augmenting visualizations with interactive data facts to facilitate interpretation and communication. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 672–681. https://doi.org/10.1109/TVCG.2018.2865145

Srinivasan, A., & Stasko, J. (2017). Natural Language Interfaces for Data Analysis with Visualization: Considering What Has and Could Be Asked. In B. Kozlíková, T. Schreck, & T. Wischgoll (Eds.), *Eurographics Conference on Visualization (EuroVis) 2017* (pp. 55–59). https://doi.org/10.2312/eurovisshort.20171133

Srinivasan, A., & Stasko, J. (2018). Orko: Facilitating Multimodal Interaction for Visual Exploration and Analysis of Networks. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2017.2745219

Srinivasan, A., & Stasko, J. (2020). How to ask what to say?: Strategies for evaluating natural language interfaces for data visualization. *IEEE Computer Graphics and Applications*, *40*(4), 96–103. https://doi.org/10.1109/MCG.2020.2986902

Stasko, J., Görg, C., & Liu, Z. (2008). Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*, *7*(2), 118–132. https://doi.org/10.1057/palgrave.ivs.9500180

Sun, Y., Leigh, J., Andrew, A., Barbara, J., Eugenio, D., & Peterka, T. (2014). Articulate: Creating Meaningful Visualizations from Natural Language. In *Innovative Approaches of Data Visualization and Visual Analytics* (pp. 218–235). IGI Global. Retrieved from http://www.indigo.lib.uic.edu:8080/bitstream/handle/10027/9226/Sun_Yiwen.pdf?sequence=1&isAllowed=y

Thomas, J. J., & Cook, K. A. (2006). A visual analytics agenda. *Computer Graphics and Applications, IEEE*, *26*(1), 10–13. https://doi.org/10.1109/MCG.2006.5

Tory, M., & Setlur, V. (2019). Do What I Mean, Not What I Say! Design Considerations for Supporting Intent and Context in Analytical Conversation. In *2019 IEEE Conference on Visual Analytics Science and Technology, VAST 2019 - Proceedings*. https://doi.org/10.1109/VAST47406.2019.8986918

Van Buskirk, R., & LaLomia, M. (1995). A comparison of speech and mouse/keyboard GUI navigation. In *Conference companion on Human factors in computing systems* (p. 96).

Vassallo, G., Pilato, G., Augello, A., & Gaglio, S. (2010). Phase Coherence in Conceptual Spaces for Conversational Agents. In *Semantic Computing*. https://doi.org/10.1002/9780470588222.ch18

Viegas, F. B., Wattenberg, M., Van Ham, F., Kriss, J., & McKeon, M. (2007). Many Eyes: A site for visualization at internet scale. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2007.70577

White, R. W. (2018). Skill discovery in virtual assistants. *Communications of the ACM*, *61*(11), 106–113.

Wickens, C. D. (1984). The Structure of Attentional Resources. In *Attention and Performance VIII*.

Woods, W., Kaplan, R. M., & Nash-webber, B. L. (1972). The Lunar Science Natural Language Information System: Final Report. *ResearchGate*.

Yankelovich, N. (1996). How do users know what to say? *Interactions*, *3*(6), 32–43.

Youn-ah Kang, Görg, C., & Stasko, J. (2011). How Can Visual Analytics Assist Investigative Analysis? Design Implications from an Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, *17*(5), 570–583. https://doi.org/10.1109/TVCG.2010.84

Yu, B., & Silva, C. T. (2019). FlowSense: A Natural Language Interface for Visual Data Exploration within a Dataflow System. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/tvcg.2019.2934668

Zhang, Y., & Liu, C. (2020). Users' Knowledge Use and Change during Information Searching Process: A Perspective of Vocabulary Usage. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (Vol. 10, pp. 47–56). New York, NY, USA: ACM. https://doi.org/10.1145/3383583.3398532

Ziemkiewicz, C., & Kosara, R. (2008). The shaping of information by visual metaphors. In *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2008.171