Tracking and Visualizing Dimension Space Coverage for Exploratory Data Analysis

by

Ali Sarvghad Batn Moghaddam
B.Sc., University Science Malaysia, 2006
M.Sc., University of Malaya, 2008

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the Department of Computer Science

Supervisory Committee

Dr. Melanie Tory, Supervisor
(Department of Computer Science)

Dr. Yvonne Coady, Departmental Member
(Department of Computer Science)

Dr. Valerie Irvine, Outside Member
(Faculty of Education)

ABSTRACT

In this dissertation, I investigate interactive visual history for collaborative exploratory data analysis (EDA). In particular, I examine use of analysis history for improving the awareness of the dimension space coverage [1] [2] [3] to better support data exploration. Commonly, interactive history tools facilitate data analysis by capturing and representing information about the analysis process. These tools can support a wide range of use-cases from simple undo and redo to complete reconstructions of the visualization pipeline. In the context of exploratory collaborative Visual Analytics (VA), history tools are commonly used for reviewing and reusing past states/actions and do not efficiently support other use-cases such as understanding the past analysis from the angle of dimension space coverage. However, such knowledge is essential for exploratory analysis which requires constant formulation of new questions about data. To carry out exploration, an analyst needs to understand "what has been done" versus "what is remaining" to explore. Lack of such insight can result in premature fixation on certain questions, compromising the coverage of the data set and breadth of exploration [80]. In addition, exploration of large data sets sometimes requires collaboration between a group of analysts who might be in different time/location settings. In this case, in addition to personal analysis history, each team member needs to understand what aspects of the problem his or her collaborators have explored. Such scenarios are common in domains such as science and business [34] where analysts explore large multi-dimensional data sets in search of relationships, patterns and trends. Currently, analysts typically rely on memory and/or externalization to keep track of investigated versus uninvestigated aspects of the problem. Although analysis history [4] mechanisms have the potential to assist analyst(s) with this problem, most common visual representations of history are geared towards reviewing & reusing the visualization pipeline or visualization states.

I started this research with an observational user study to gain a better understanding of analysts' history needs in the context of collaborative exploratory VA. This study

[1] In this dissertation, a *dimension* refers to a column in a tabular dataset where dimension's name is the column's header name. For instance, a financial dataset may include dimensions such as Sales, Profit and Inventory Cost.

[2] I define and use *dimension space* to refer to the set of all dimensions in a tabular dataset.

[3] I define *dimension space coverage* as the set of investigated data dimensions, either individually (e.g. a histogram showing distribution of Sales values) or collectively (e.g. a bar chart showing averages of Sales and Profit for different States).

[4] In the context of visual data analysis, history is usually comprised of recorded information about the analysis states/processes and/or outcomes. For more detailed description see Chapter 2.

showed that understanding the coverage of dimension space by using linear history [5] was cumbersome and inefficient. To address this problem, I investigated how alternate visual representations of analysis history could support this use-case. First, I designed and evaluated Footprint-I, a visual history tool that represented analysis from the angle of dimension space coverage (i.e. history of investigation of data dimensions; specifically, this approach revealed which dimensions had been previously investigated and in which combinations). I performed a user study that evaluated participants' ability to recall the scope of past analysis using my proposed design versus a linear representation of analysis history. I measured participants' task duration and accuracy in answering questions about a past exploratory VA session. Findings of this study showed that participants with access to dimension space coverage information were both faster and more accurate in understanding dimension space coverage information. Next, I studied the effects of providing coverage information on collaboration. To investigate this question, I designed and implemented Footprint-II, the next version of Footprint-I. In this version, I redesigned the representation of dimension space coverage to be more usable and scalable. I conducted a user study that measured the effects of presenting history from the angle of dimension space coverage on task coordination (tacit breakdown of a common task between collaborators). I asked each participant to assume the role of a business data analyst and continue a exploratory analysis work which was started by a collaborator. The results of this study showed that providing dimension space coverage information helped participants to focus on dimensions that were not investigated in the initial analysis, hence improving tacit task coordination. Finally, I investigated the effects of providing live dimension space coverage information on VA outcomes. To this end, I designed and implemented a standalone prototype VA tool with a visual history module. I used scented widgets [76] to incorporate real-time dimension space coverage information into the GUI widgets. Results of a user study showed that providing live dimension space coverage information increased the number of top-level findings. Moreover, it expanded the breadth of exploration (without compromising the depth) and helped analysts to formulate and ask more questions about their data.

[5] In the context of visual data analysis, linear representation of history is usually a comic-strip-like list of thumbnail images of visualization states.

Contents

List of Tables

List of Figures

Firstly, I would like to express my sincere gratitude to my advisor Dr. Melanie Tory for the continuous support of my Ph.D study and related research, for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor.

Besides my advisor, I would like to thank the rest of my thesis committee: Dr. Yvonne Coady and Dr. Valerie Irvine, and my external examiner Dr. Wesley Willet. Their insightful comments and constructive criticism that helped improve this dissertation.

I thank my fellow researchers in VisID for their ever present help and stimulating discussions.

Last but not the least, I would like to thank my family: my wife, my son and my parents. I would have never been able to end this journey without their unconditional love and support.

To my wife Narges:

The love of my life,
The pillar of all my success.
This journey was impossible without you by my side.

To my son, Iliya:

The light of my life,
the joy of my soul.

PUBLICATIONS

The materials presented in this thesis have been previously published in different venues. After each reference, I refer to chapters that present the material.

Journal Articles

- **Ali Sarvghad**, Melanie Tory, and Narges Mahyar "Visualizing Dimension Coverage to Support Exploratory Analysis". IEEE Transactions on Visualization and Computer Graphics (TVCG), September 2016 .

  Material from this publication appears in chapter 6

- Narges Mahyar, **Ali Sarvghad**, and Melanie Tory, "Note Taking in Co-located Collaborative Visual analytics: Analysis of an Observational Study". Information Visualization, vol. 11, no. 3, pp. 190-204, July 2012.

  Material from this publication appears in chapter 3

Conference Papers

- **Ali Sarvghad** and Melanie Tory. "Exploiting analysis history to support collaborative data analysis". Proceedings of the 41st Graphics Interface Conference. Canadian Information Processing Society, 2015.

  Material from this publication appears in chapter 5

- Narges Mahyar, **Ali Sarvghad**, Melanie Tory, and Tyler Weeres, "Observations of Record-Keeping in Co-located Collaborative Analysis". HICSS 2013, pp. 460-469, Jan. 2013.

  Material from this publication appears in chapter 3.

- Narges Mahyar, **Ali Sarvghad**, and Melanie Tory, "A closer look at note taking in the co-located collaborative visual analytics process". IEEE Visual Analytics Science and Technology (VAST10), pp. 171-178, 2010. [Selected for publication in the "Information Visualization" journal].

  Material from this publication appears in chapter 3.

Workshop Papers

- Narges Mahyar, **Ali Sarvghad**, Melanie Tory and Tyler Weeres "CoSpaces: Workspaces to Support Co-located Collaborative Visual Analytics". DEXIS 2011, Nov 2011.

  Material from this publication appears in chapter 3.

- **Ali Sarvghad**, Narges Mahyar, and Melanie Tory, "History Tools for Collaborative Visualization". Workshop on Collaborative Visualization on Interactive Surfaces (CoVis 2009), Oct. 2009.

  Material from this publication appears in chapter 2.

- Narges Mahyar, **Ali Sarvghad**, and Melanie Tory, "Roles of Notes in Co-located Collaborative Visualization". Workshop on Collaborative Visualization on Interactive Surfaces (CoVis 2009), Oct. 2009.

  Material from this publication appears in chapter 3.

Chapter 1

Introduction

In this thesis, I investigate analysis history (a.k.a. provenance) for supporting exploratory data analysis (EDA). In particular, I examine the effects of representing analysis history from the angle of dimension space coverage[1] on EDA. Prior research on *exploratory search*, (e.g. [54] [75]) defines EDA as an activity in which the user aims to gain an overview of the information space and engage in serendipitous discovery. Questions in this activity are open-ended and evolve as the exploration continues. EDA is in contrast to *search*, where the main objective is to find answers to specific questions. In this thesis, I will focus on supporting EDA.

EDA is inherently a breadth-oriented activity. The analyst "is to explore the data in as many ways as possible until a plausible *story* of the data emerges" [75]. Often while exploring the data, it is not clear where interesting results might lie and analysts need to cast a wide net and explore data from as many angles as possible. Therefore, analysts constantly formulate and answer new questions about their data. For example, to assess business performance, a data analyst may start by investigating Profit in combination with dimensions such as Region, Product Type, and Shipping Cost. Later, she may examine Profit with regards to Returned Goods to investigate how returned merchandise affect profit. To successfully create new questions or hypotheses that target uninvestigated aspects of the problem, she needs to be aware what has been investigated so far. [e.g., Did I examine the relationship between Profit, Customer Segment and Sales yet?]. In contrast, *search* requires focused querying of the information space to extract information that answers specific predefined questions.

The breadth of EDA is proportional to the size of the dimension space. A larger di-

[1] *Dimension space coverage* refers to the set of data dimensions that have been previously investigated, either individually or in combination with each other.

mension space would typically require greater breadth of exploration than a smaller space. To be able to efficiently browse the dimension space, an analyst needs to maintain an understanding of the breadth of the analysis [75] [46] [74]. However, maintaining a mental record of the breadth of the exploration can be costly for the analyst, especially when dealing with many dimensions. Factors such as limited short term memory and the recency effect (i.e. remembering recent items more clearly than those further in the past) [27] can impede recalling the breadth of exploration. Newness of data can also hinder exploration. Prior research suggest [76] [80] that analysts rely more on navigational assistance to explore an unfamiliar information space. Difficulties in conceptualizing the space as a whole and navigating within it can result in fixation on only a subset of the dimension space [80].

Another factor that can increase the complexity of EDA is collaboration. Sometimes the complexity of task, large volumes of data, and interdisciplinary problems require analysts to work together [37] [14]. In this case, effective browsing of dimension space requires support for *awareness*. In this dissertation, I refer to awareness in a general manner as "up to the moment understanding of a collaborator's analytical activities[2] and outcomes[3]". Prior research has shown the importance of providing awareness to support both collaborative activities and analysis outcomes. For instance, awareness of others' activities in a team can improve task coordination [35] [42] [71] and increase the number of findings [36] [53].

The following fictional scenario depicts how providing awareness can facilitate collaborative EDA: "Sam (in Asia), Ted (in Europe) and Nelly (in North America) are business data analysts for a large international retailer. Their current task is evaluate business performance by analyzing company's large multi-dimensional sales dataset. Sam, Ted and Nelly all use the company's cloud-based VA platform to carry out exploratory analysis. During analysis, Sam, Ted and Nelly constantly formulate and ask new questions by choosing among different combinations of data dimensions. For example, Nelly may ask "what is the relationship between *Sales*, *Region* and *Product Category*?", and after noticing unusually low sales for the West market, she may investigate "which *Product Types* in each *Product Category* are sold the most in *Region:West*?". Ideally, Nelly prefers to avoid asking duplicate questions that Sam and Ted have already asked (except to verify findings). In order for Nelly to achieve this, she needs to know the coverage of dimension space by Sam and Ted to decide what is left for her to concentrate on. The difference between their time/place set-

[2]This includes all the activities taken by analyst including (but not limited to) visualization and externalization. Creating a visualization can include steps such as data wrangling, script writing, data mapping, filtering, and spatial modeling. Externalization often includes activities such as note-taking and annotation.

[3]In this dissertation, I consider analysis outcomes to be an analyst's discoveries and observations such as findings and hypotheses.

tings makes direct communication very difficult but the visual data analysis platform that they use has a specific module that enables Nelly to review and understand Sam and Ted's analysis from the angle of dimension space coverage. Using a new dimension coverage tool, she discovers that *Inventory Cost* has not been looked into by Sam and Ted. Therefore she decides to explore this dimension further."

In the context of collaborative EDA, both exploration and collaboration can benefit from a mechanism that tracks and represents the breadth of exploration. Existing visualization *history* (a.k.a. provenance) tools partially fulfill this need by tracking and representing some information about a users' past data analysis activities. Depending on their architecture, these tools may capture workflows (e.g. user commands), visualization states (e.g. statistical charts), and externalizations (e.g. notes). Although these tools facilitate review and reuse of workflows, visualizations, and externalizations, they are rather limited in providing first-hand insight into the aspects investigated here: understanding breadth of the analysis and coverage of dimension space. In this dissertation, I design and investigate interactive visual representations of analysis history that provide first hand insight into the coverage of dimension space. I report the design and evaluation of this approach for supporting collaborative exploratory VA. In particular, I investigate the the effects of my proposed approach on coordination in a collaborative context, as well as the exploratory VA process and outcomes in a single user context.

1.1 Dissertation Problem

In this dissertation, I investigate how visualizing dimension space coverage information can support EDA. I examine how visualizing the breadth of dimension space coverage effects both EDA process and outcomes, and I investigate the value of dimension coverage in both individual and collaborative settings.

The following list introduces my high-level research questions:

RQ1: What is the state of the art of history tools for supporting collaborative EDA? To address this question, I performed a comprehensive literature review to understand the state of the art of history tools for VA. Based on the findings of this literature review, I identified common use-cases for history in the VA context (included in Chapter 2). I also noticed that captured history was most commonly represented in the form of a linear sequential list of captured processes and/or artefacts.

RQ2: What are the limitations of a linear history representation for collaborative EDA? To address this question, I conducted an observational user study. I designed CoSpaces, a tool for collocated collaborative exploratory VA (Chapter 3) that incorporated a typical linear representation of analysis history[4]. The results of this study showed the inadequacy of the linear representation of history in providing insight into the coverage of dimension space. I observed that after some time into the analysis, participants had difficulty in clearly remembering what was done versus what was left to analyze. Although the linear representation of history supported reviewing and reusing prior states, it fell short when users were trying to understand the coverage of dimension space.

RQ3: Does representing analysis history from a dimension-centric angle better support understanding the coverage of dimension space than a linear representation of history? I designed and implemented Footprint-I (Chapter 4), a history tool that was specifically designed to visually represent the coverage of dimension space. This interactive view provided first-hand insight into the coverage of dimension space. In a user study, I compared Footprint-I's dimension-centric history view to a linear representation. Participants answered questions about what dimensions were examined and in what combinations. Participants with access to the dimension-centric history view were twice as fast and more accurate in answering questions about dimension space coverage.

RQ4: How does dimension space coverage information influence task coordination? To answer this question, I designed and implemented my next prototype, Footprint-II (Chapter 5). Similar to Footprint-I, this history tool contained an interactive visualization of the coverage of dimension space. This view enabled users to understand "which data dimensions" had been examined and in "which combinations". I conducted a user study to evaluate the effects of my approach on collaboration. Findings of this study showed that dimension space coverage information improved task coordination. Participants with access to the dimension-centric history view focused more on univestigated aspects of the dataset. To measure coordination, I compared the overlap between each participant's analysis and an initial analysis done by a fic-

[4]Linear representation of history refers to a comic-strip-like list of recorded history information. In the context of VA, the list is usually comprised of visualization thumbnails augmented by textual information.

tional collaborator.

RQ5: How does providing live information about dimension space coverage influence EDA? To answer this question, I designed and implemented a self-contained visual data analysis tool with dimension space coverage information embedded in the interface widgets (Chapter 6). The results of a user study showed that this approach increased the number of top-level findings, expanded the breadth of exploration, and helped analysts to formulate more questions.

1.2 Dissertation Scope

This dissertation investigates how dimension space coverage information can support EDA. I will focus on collaborative EDA for tabular business data. In the following subsections, I will justify my choice of dissertation scope.

1.2.1 Why tabular business data?

All the research questions in this thesis are investigated in the business domain. I chose this domain for two reasons. At the beginning, this research was in collaboration with SAP Business Objects, an industrial partner interested in Business Intelligence (BI). Moreover, there are large sample business datasets publicly available that could be used in the user studies. Since business data is typically and traditionally stored in tabular format, I chose to focus on tabular data in this research.

1.2.2 Why collaboration?

A new generation of business intelligence systems has been emerging during the last few years to meet the new, sophisticated requirements of business users. The term *BI 2.0* has been coined to denote these systems. Collaboration is among the main characterizing trends of BI 2.0. In collaborative BI, analysis of large volumes of business data is carried out collaboratively across organizations [66] [34]. For this reason I chose to focus on small team collaborative EDA in research questions RQ1 to RQ4 . This type of collaborative analysis may happen in domains such as business and science [34].

To investigate each research question, I focused on the collaborative setting that best helped me to investigate that question. Collaboration can happen across varying time/place

settings (e.g. same time/same place, different time/different place) and investigating each research question across all time/place combinations was beyond the scope of this dissertation. Therefore, for each question, I picked a time/place setting that would best help me to investigate that question. For RQ2, I selected a collocated and synchronous collaborative setting. Collocated situations may represent the best case for collaboration, as users have all of the advantages of working synchronously together at the same place.

Furthermore, in collocated situations, researchers can more easily examine how team members collaborate in real-time. Direct interaction with all the group members is a great opportunity to understand their needs and challenges and it is much easier to conduct post study interviews in collocated studies. RQ3 and RQ4 were investigated in a different time/different place setting. In both cases, an analyst continued working on a problem that was investigated by his/her collaborators before. The main rationale behind selecting this time/place setting was to investigate history isolated from any other channels for providing information about the coverage of dimension space such as direct communication between team members. Unlike my other research questions, RQ5 was investigated in a single user setting. The main reason for this decision was to factor out effects that collaboration could possibly have on the analysis outcomes and consider exploration in isolation. Future work could investigate the extension of RQ5 to collaborative work.

1.3 Methodological Approach

To address my research questions, I used both qualitative and quantitative analytical methods. With the exception of RQ1, I performed controlled laboratory studies to investigate my research questions. All user studies consisted of the following steps 1) identifying a problem, 2) generating a hypothesis, 3) proposing a solution, 4) designing and implementing prototype tools, and 4) evaluating the solution with users. Table 1.1 summarizes the methodological approach used for each research question. I followed guidelines for conducting user studies to minimize bias and privacy. To avoid "positivity bias", I requited participants who had no prior familiarity with the experimenter or the project. In addition, I distributed participants evenly between experimental conditions based on gender (male/female) and education level (grad/undergrad). Following the guidelines for conducting user studies [51], I ensured that I had at least 10 participants for each tested condition.

For all the user studies, I closely followed guidelines provided by University of Victoria for conducting human research and obtained required approvals from Human Research Ethics Board of the university.

Table 1.1: Methodological approach used for each research question (RQ). To analyze collected video and audio data, I performed multi-pass open coding analysis. Text analysis refers to analyzing participants' notes. I used various statistical techniques (depending on the characteristics of data) for analyzing quantitative data. Quantitative data was gathered from coding of audio/video files and/or questionnaires. Qualitative data was gathered from Audio/video data, notes, interviews and questionnaires.

RQ	**Method**	**Data Collection**	**Evaluation**	**Data Analysis**
RQ1	Literature review	• Documents	• Qualitative	
RQ2	User study	• Video • Audio • Observations • Interview	• Qualitative	• Multi-pass open coding • Text analysis
RQ3	User study	• Video • Audio • Time to complete task • Task scores • Questionnaire • Interview	• Qualitative • Quantitative	• Multi-pass open coding • Statistical analysis
RQ4	User study	• Video • Audio • Software log • Participants' notes • Questionnaire • Interview	• Qualitative • Quantitative	• Multi-pass open coding • Text analysis • Statistical analysis
RQ5	User study	• Video • Audio • Software logs • Participants' notes • Questionnaire • Interview	• Qualitative • Quantitative	• Multi-pass open coding • Text analysis • Statistical analysis

1.4 Contributions

Following are the main contributions of this dissertation. Contributions (C) are listed under corresponding research questions (RQ):

RQ1 contribution:

- **C1: Identified most common history use-cases for collaborative data analysis**. Based on an extensive literature review, I complied a list of the most common use cases for history in the context of collaborative visual data analysis. Many of the use cases are extendable to non-collaborative situations.

RQ2 contributions:

- **C2: Demonstrated that users innately expected history to provide information about dimension space coverage.** I observed that users reviewed their work history to determine "what has been done" and "what else is left" for further investigation.
- **C3: Demonstrated the inadequacy of the linear history representation in providing information about the coverage of dimension space.** I observed that it was cumbersome for users to understand dimension space coverage using a linear representation of history.

RQ3 contribution:

- **C4: Demonstrated that representing history from the angle of dimension space coverage resulted in a faster and more accurate understanding of which dimensions had been explored and in which combinations.**

RQ4 contribution:

- **C5: Demonstrated that representing history from the angle of dimension space coverage can improve tacit coordination between collaborators.** I observed that participants tried to avoid asking duplicate questions that were already investigated by their collaborator and focused more on what the collaborator had not yet investigated.

RQ5 contributions:

- **C6: Demonstrated that using scented widgets to represent dimension coverage information increased the number of questions asked during exploratory analysis.**

- **C7: Demonstrated that using scented widgets to represent dimension space coverage information increased the number of top-level findings.**

- **C8: Demonstrated that representing dimension space coverage information resulted in a greater breadth of exploratory analysis.** Interestingly, this approach resulted in a greater breadth of analysis without compromising the depth.

Design contribution:

- **C9: Illustrated some viable visual representations of dimension space coverage information and how such information can be incorporated into visual data analysis tools.** This contribution is based on the iterative process of examining different visual representations for dimension space coverage through RQ2 to RQ5. In Footprint-I (Chapter 4) and Footprint-II (Chapter 5), I used circular and treemap layouts for visualizing dimension space coverage. In Chapter 6, I used scented widgets [76] to incorporate dimension space coverage information into the interface elements of a visual data analysis tool.

1.5 Outline

This dissertation is structured around the five main research questions:

Chapter 2: Literature Review

I introduce relevant background material related to history for visual data analysis, collaborative visualization and data exploration. This chapter also includes results of my initial literature review (RQ1).

Chapter 3: Investigating Limitations of the Linear History Representation (RQ2)

I report on the design and evaluation of a prototype tool that incorporates visual history and record-keeping for exploratory collaborative VA. Based on results of the user study,

I derive a list of actions and intentions in regard to the use of the visual record-keeping module. I also discuss the shortcomings of the linear visual representation of history for exploratory analysis.

Chapter 4: Understanding the Breadth of Exploration: Linear History versus Visualizing Dimension Space Coverage (RQ3)

I present Footprint-I, a visual history prototype designed to represent the coverage of dimension space. I report results of a user study demonstrating that a dimension-centric view of analysis history enabled people to more quickly and more accurately understand the investigation done by a previous analyst.

Chapter 5: Investigating the Effects of Providing Dimension Space Coverage Information on Task Coordination (RQ4)

In this chapter I introduce Footprint-II, the successor of Footprint-I. This prototype incorporates a different visual history for dimension space coverage. I show that visually representing the dimension space coverage information can improve tacit task coordination between collaborators. This approach enabled analysts to focus on questions not previously investigated by their collaborator.

Chapter 6: Supporting Exploratory Data Analysis via Scented Widgets for Dimension Coverage (RQ5)

In this chapter, I present Scented View, a visual representation of dimension coverage embedded in GUI widgets. My approach extends the concept of scented widgets [76] to reveal aspects of one's own analysis history, and offers a different perspective on one's past work than typical visualization history tools. Results of an empirical validation study showed that participants with access to embedded dimension space coverage information relied on this information when formulating questions, asked more questions about the data, generated more top-level findings, and showed greater breadth of their analysis without sacrificing depth.

Chapter 7: Discussion and Future Work

I discuss lessons learned, limitations, threats to validity and future directions.

Chapter 8: Summary and Contributions

I summarize the thesis and reflect on the contributions.

Chapter 2

Related Work

In the context of information visualization and visual analytics, history (a.k.a. provenance) refers to the process of capturing and representing information about the analysis processes and/or outcomes. I start this section by describing common history architectures and representations. I also report common history use-cases in the information visualization and VA context. Since my research investigates history in an exploratory collaborative context, I continue this section with a report on history tools for collaborative and exploratory data analysis.

2.1 History Models

Many researchers have mentioned advantages of history tools and their importance for data visualization and analysis [30] [33] [37] [44] [55]. These tools support data analysis in various ways, ranging from helping a single user to review his past analysis to providing support for collaborative and/or exploratory visual data analysis. In general, history tools achieve these goals by capturing and representing information about the analysis. Depending on their underlying history model (i.e. what is captured), representation (e.g. visual, textual), supported user operations (e.g. review, search, share) and architecture (e.g. stand alone, web-based, etc.), different history tools support different use-cases [30]. According to Heer's [30] survey of history tools, two main history models can be identified: 1) state-based, and 2) action-based. Generally, history tools with an underlying action-based model capture single or groups of user interactions/commands; these interactions typically result in a transformation of the system and/or visualization. In contrast, state-based history tools record information about the state of the system and/or visualization at specific times; these

records can be used to duplicate that system state at a later time. State-based history tools may also include analyst externalizations such as notes and annotations. History tools most commonly utilize node-link data structures to internally store captured information [30].

2.2 History Use-cases

Based on my initial literature survey, I identified some of the most common history use-cases in the context of information visualization and visual analytics:

- **Recall:** in line with previous researchers [30] [64] [65] [31] [39] [48], I identified Recall as one of the most common history use-cases. This is probably the most generic history use-case. In the context of information visualization, recall has been mainly used for remembering past visualization states and/or analytical steps. Part of my research investigates how history can be used to help an analyst recall the coverage of dimension space.

- **Exploration:** having a repository of history items enables data analysts to try alternating courses of analysis by revisiting a history item and trying a different possible path. This is specifically important for exploratory VA because "Insight often comes from comparing the results of multiple visualizations that are created during the data exploration process" [11]. In addition, a history module that captures and represents pipeline and/or workflow enables an analyst to explore alternating pipelines/workflows and try/compare different visual outcomes [17] [41] [61] [82] [7].

- **Validation:** [30] [32] [37] [55]: Correctness and admissibility of decisions/findings or appropriateness of a single visualization can be examined by using history items. For instance, analysts may review visualizations created in the course of an analysis process to double-check that their findings are correct, or they may revisit a particular visualization to ensure that it is the result of correct mapping and filtering of data. This might be even more helpful during shifts between different collaboration styles. Participants may need to corroborate the outcomes of individual work that will be continued later.

- **Memory aid/Externalization:** The limitation of humans' short-term memory is a known fact, and a history tool can act as an external memory aid [44]. Data analysts

can add important notes, observations, calculations etcetera to history items for future referral.

- **Correction/Recovery:** If data analysts find their current visualization undesirable for any reason, they can perform a selective undo/redo [23] [30] [32] [40] [57]. It is also possible to continue a visualization and analysis process from the last point in the history repository after a system failure.

- **Reporting / Storytelling / Presentation [30] [44]:** A history repository, wholly or partially, can be sent to peers or managers as a progress report, indication of the amount of work done, or formal report of findings. History items can be summarized and presented in a meeting situation. Presentation is similar to reporting, but typically occurs synchronously.

- **Coordination:** [24] [32] [40] [44] [55] History can help collaborators coordinate their effort by increasing awareness in situations such as loosely-coupled collaborative work or remote synchronous/asynchronous situations. Also, viewing a collaborator's analysis history can bring a person up-to-speed on the work done so far.

- **Training :** [44] Novice data analysts can learn from experts by reviewing the history of visualizations created and decisions made.

2.3 History representations

Although some researchers [30] [64] have worked on identifying the most common history use cases for visual data analysis, to the best of my knowledge, there are no comprehensive generic guidelines for visually representing the history based on the intended use-case(s). Most commonly, variations of a node-link graph are used to visually represent history. Depending on the underlying history model and captured information, nodes of the graph may represent data, actions or states, and connections show dependencies and/or precedence. VisTrails [7] [10] and CzSaw [41] are two examples of history tools that use a node-link graph structure to visually represent the history. Both of these tools aim to support the analysis process by incorporating action-based history modules. VisTrails (Figure 2.1) was designed mainly for spatial data sets and captures visualization pipelines, the path from data to a visual representation, in the form of user commands. The pipeline can be modified and re-applied to the same data to explore various visualizations; alternately, the pipeline can be applied to similar data sets.

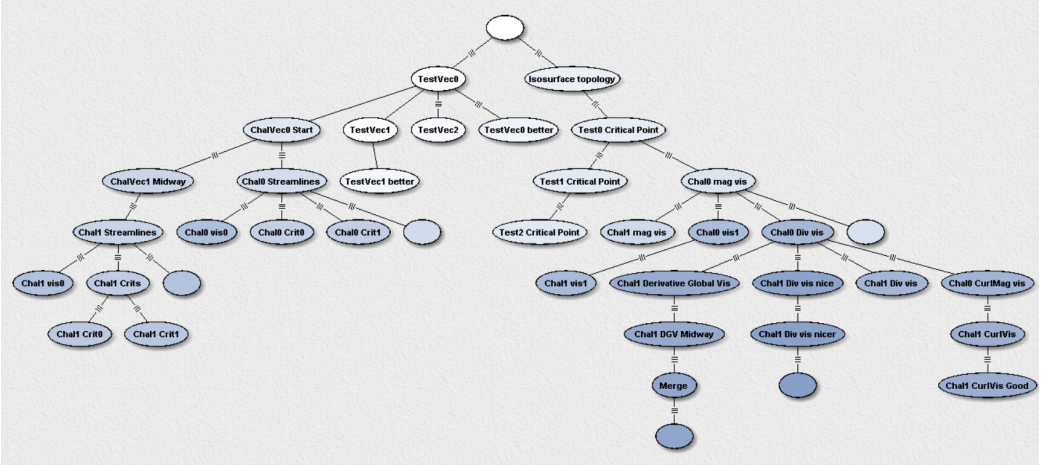


Figure 2.1: A snapshot of VisTrails' history management module. Each node represents a user command. [7]

Similarly, CzSaw (Figure 2.2) is a visual document analysis tool with a history module that captures user interactions and builds data-independent scripts. Scripts facilitate reuse by enabling the analyst to apply scripts on different document sets. Both systems use node-link graphs to visually represent recorded information. VisTrails' Builder View depicts an analysis pipeline as an acyclic node-link graph where each node represents a user command and links show dependencies and flow of the analysis. CzSaw uses a tree to visualize the captured script, where nodes represent variables and directed links represent dependencies. In general, action-based history tools capture history independent from data, and as result, are unable to provide rich insight into the history of data space exploration.

GRASPARC [8], ExPlates [39] and GraphTrails [21] are additional examples of data analysis tools that contain an action-based history module with a node-link graph representation. One exception to this trend is SensePath [60], a provenance tool that represents its action-based history using a list of icons and textual descriptions. Figure 2.3 shows a snapshot of Timeline View that "shows all captured sensemaking actions in temporal order" [60].

Another common representation of history is a linear list representation (a.k.a. comic-strip). Similar to the node-link representation, list items represent captured information, but there are no explicit links between them. List items can be ordered based on different criteria such as chronological precedence or similarity. Heer's history bar for Tableau [30] is an example of a linear representation (Figure 2.4). In this example, list items (i.e. thumbnail images) contain both action and state information. Each item contains a thumbnail image, which is labeled by the user action that resulted in that state.

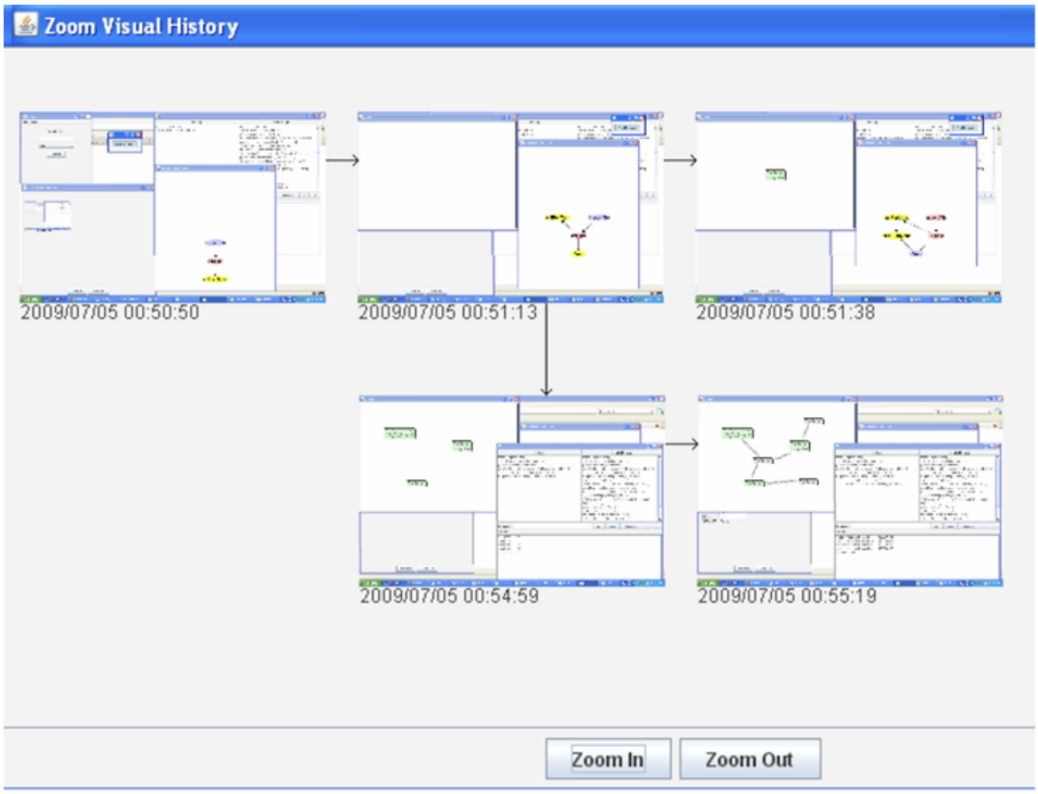


Figure 2.2: CzSaw visual history module. Each node shows the state of the visualization after a change is applied to the previous state. [41]

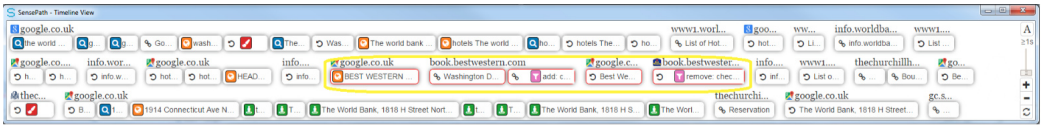


Figure 2.3: Timeline View within SensePath, showing all the actions taken by user [60].

Similarly, PivotSlice [81], an interactive visual data analysis tool for faceted browsing of network data, contains a visual history module that provides a chronologically ordered list of thumbnail images of previous states (Figure 2.5). The history module also stores a list of recently added or removed attributes on the right side of the History Panel (Figure 2.5), which can be reused via drag-and-drop interactions.

In addition to the aforementioned history representations and in a considerably smaller scale, other visual representations such as treemaps [19] and tag clouds have been used [68] [13] for representing the history.

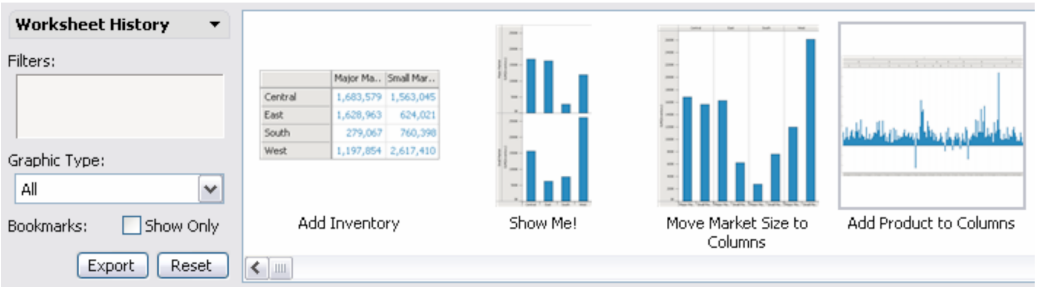


Figure 2.4: Visual history for Tableau. [30]

Figure 2.5: History Panel in PivotSlice [81].

2.4 History, Collaboration, and Exploration

In the collaborative context, prior history research has been mainly focused on suporting communication of analytical processes and outcomes between collaborators. Building *common-ground* (i.e. shared understanding of each others' work from different perspectives) that facilitates collaborative work is one of the foremost goals of collaborative VA tools. In recent years, history tools have been investigated as a means for facilitating common-ground construction across distributed data analysts. In synchronous collaborative VA, real-time shared views and instant-communication modalities can help in building common ground. For instance, CoMotion [15] enables sharing of personal views across the group. Similarly, Cambiera [35] enables an analyst to maintain an awareness of a collaborator's search queries and reviewed documents for co-located analysis of document collections. In an asynchronous context, history tools most commonly help build common-ground by capturing and sharing externalizations. CLIP [53], Sense.us [33], CommentSpace [77], and ManyEyes [72] are examples of information visualization and VA tools that use externalizations to support awareness. These tools allow discussions (often in a forum-like structure with posting, replying and tagging) to be weaved around visualizations and/or analysts' findings. The linkage between visualization(s) and externalization(s) enables each collaborator to review, understand and contribute to the on-going analysis built around a data view. Chen's [13] history tool aims to provide top-level awareness by grouping externalizations. Users can dynamically create different groupings of externalizations by changing similarity parameters (e.g. similarity of tags attached to notes).

AnalyticTrails [50] is a history tool that captures and communicates the analysis process. AnalyticTrails, built into a web-based visual data analysis tool, was designed to automatically record trails of analytic steps performed by the user. An analyst can share the recorded action trails with others (e.g. collaborators) or reuse them personally. For example, an analyst can reuse action trails on an updated version of a dataset.

2.5 Analysis History for Tracking the Breadth of EDA

In this section, I first introduce the notion of "information scent" introduced by Pirolli and Card [63] and how it can be used for navigating information spaces. Next, I will present the prior work in the EDA domain that uses analysis history as source for generating information scent.

"Information scent refers to the cues used by information foragers to make judgements related to the selection of information sources to pursue and consume" [62]. In Brunswick's Lens Model [9], *proximal* cues help in judging the value of *distal* objects. Researchers have applied the same model in the information space. Proximal cues (e.g. links on a web page) function as mediators for availability and value of distal information sources (e.g. web pages). "On the basis of these proximal cues, the user must make judgments about what is available and potential value of going after the distal content" [26].

In the context of EDA, prior research has used analysis history for generating information scent. Willett et el. [76] used history of **data** space coverage (breadth of data values exploration) generating proximal cues about uninvestigated values. They used scented widgets, an information visualization technique inspired by the notion of information scent, to incorporate coverage information into user interface elements. Other research tools have also implemented concept of scented widgets for providing information related to user's task at hand. Phosphor [6] superimposed a halo effect on recently used interface widgets to assist users in noticing changes that had taken place in the interface. Derthick [18] and Eick [25] introduced modified versions of slider controls that visually embedded information in the widget. Depending on the design, this information could be related to the data values in the dimension that the slider is bound to or values of a different dimension. For example [25], a slider that is bound to City (i.e. that allows users to pick a city name) could contain an embedded visualization showing the average number of frost-free days for each city.

In the context of EDA, scented widgets have been mainly used to enable analysts to understand the coverage of data space. Willet [76] used scented widgets to integrate users'

information seeking history into interface elements. Figure 2.6 shows a snapshot of re-implemented HomeFinder, a map-based housing search tool with dynamic query widgets for filtering the view. Information about prior house searches were embedded in the widgets to help people better understand which data values had been investigated by other users. In my research, I will examine use of information scent and scented widgets for visually assisting analysts to understand the coverage of dimension space.

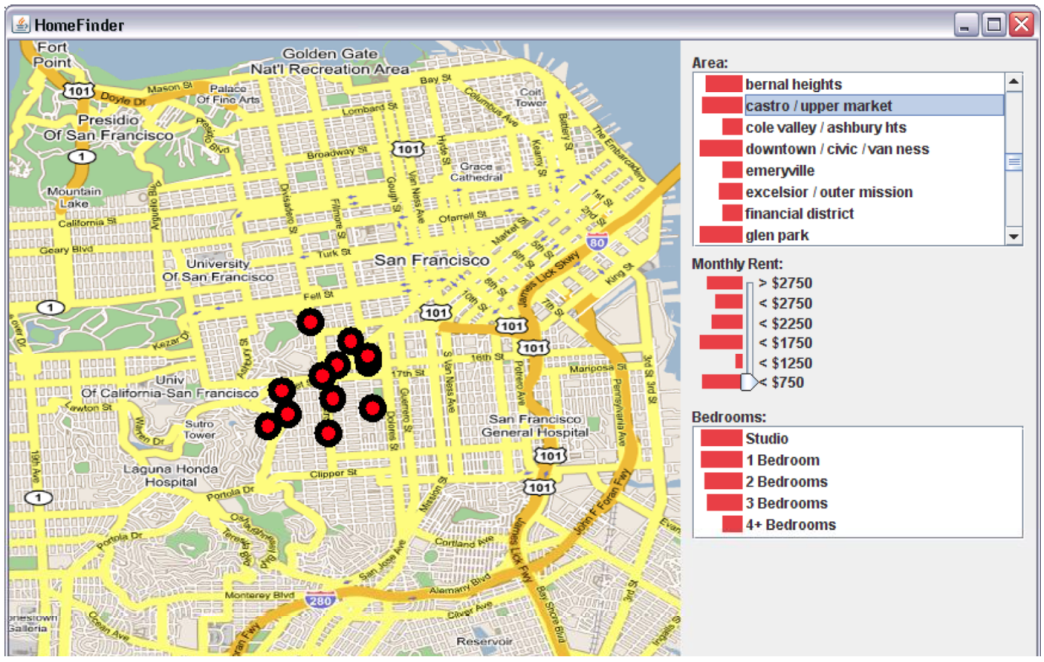


Figure 2.6: Re-implemented HomeFinder with scented widgets. Values for each data dimension (i.e. Area, Monthly Rent, Bedrooms) are populated into combo boxes and sliders. Users can make dynamic queries by filtering dimensions. Bars next to each data value show frequency of prior investigations. For example, bar charts in the Monthly Rent slider show that in comparison, fewer people looked at Monthly Rent$<$1250. [76]

In [73], Wattenberg hypothesized that providing visual cues into the past exploration of data space may encourage people to analyze uninvestigated dimensions. In their prototype tool (Figure 2.7), previously investigated time series items are in grey, in contrast to uninvestigated ones that are coloured. Although their design helps one to discover uninvestigated data, it falls short of fully exposing the coverage of dimension space. Moreover, they did not formally evaluate the idea.

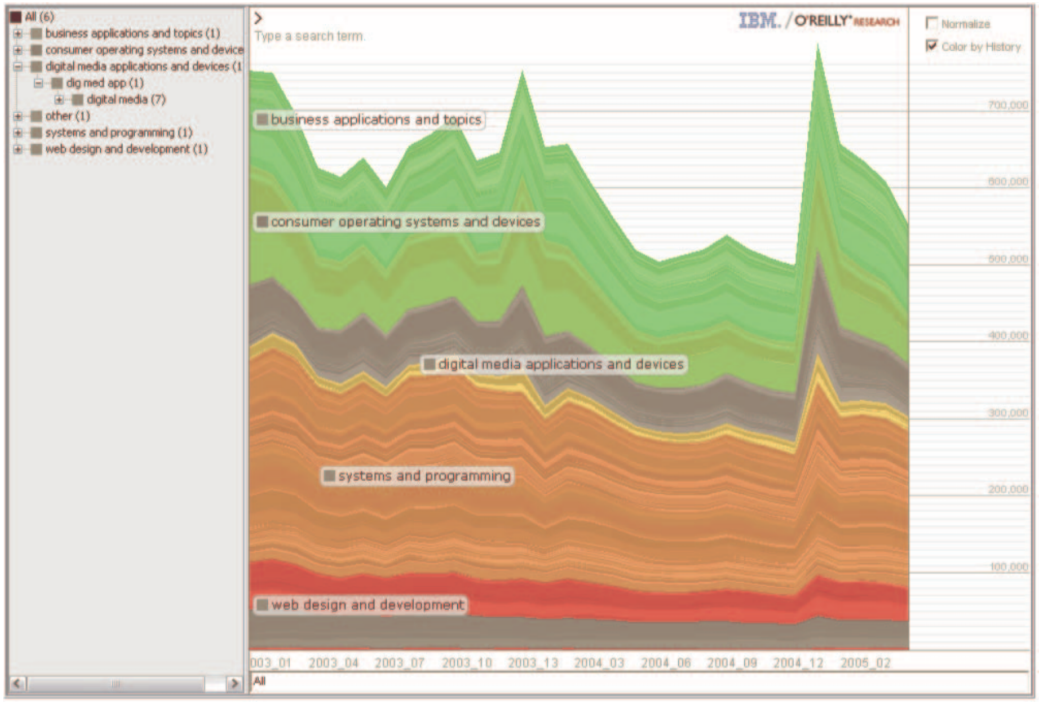


Figure 2.7: Gray series have been visited by users and Coloured items remain unexplored. Colour is used to provide information scent. [73]

2.6 Summary

Although prior research has investigated use of history tools for facilitating collaborative EDA, the focus has been mainly on communicating visualization states and/or externalizations. In the context of exploratory data analysis, history has been used for keeping track of analytical processes and visualization states. In this thesis, I investigate the use of history for visually representing dimension space coverage to communicate which dimensions have been explored and in which combinations. I will examine a number of history representations that visualize this information and investigate their effects on collaboration and exploration processes.

Chapter 3

Investigating Limitations of the Linear History Representation

In this chapter, I report my investigation of **RQ2: What are the limitations of current history practices for collaborative EDA?** Towards gaining a better understanding of users' history needs in an exploratory collaborative context, I designed and evaluated CoSpaces (Collaborative Spaces), a prototype tool for collocated collaborative Visual Analytics on interactive tabletops. [1] Following common history design practices in the VA context, this tool included a visual history module that represented analysis history as a list of thumbnail images of captured visualizations. To support externalization, the history module also included a note-taking component that allowed users to take notes. l will start this chapter with a description of CoSpaces' design and features. Next, I will describe the design and results of an observational user study that I conducted to observe how small groups of analysts used history while performing exploratory analysis. The most important finding of this study was that linear history representations cannot efficiently support understanding the coverage of dimension space, an important aspect for carrying out exploratory analysis. The main contributions of this chapter are described in Chapter 1 as C2 and C3.

3.1 Overview of CoSpaces

To investigate visual history for collaborative exploratory VA, I designed and implemented CoSpaces, a visual data analysis tool that contained a history module. Because there was

[1] I would like to acknowledge my colleagues: Tyler Weeres for his help during the system implementation and Narges Mahyar for her contribution to the design, observations and analysis of the results of this study.

no existing tool that incorporated current best practices for collocated collaborative work, record-keeping and tabular data visualization, I decided to build a tool rather than using an existing one that was not designed for this context. I designed CoSpaces for a large multi-touch tabletop display since such devices are thought to facilitate collocated collaborative work. The following subsections describe the primary features of this tool.

3.1.1 Worksheet

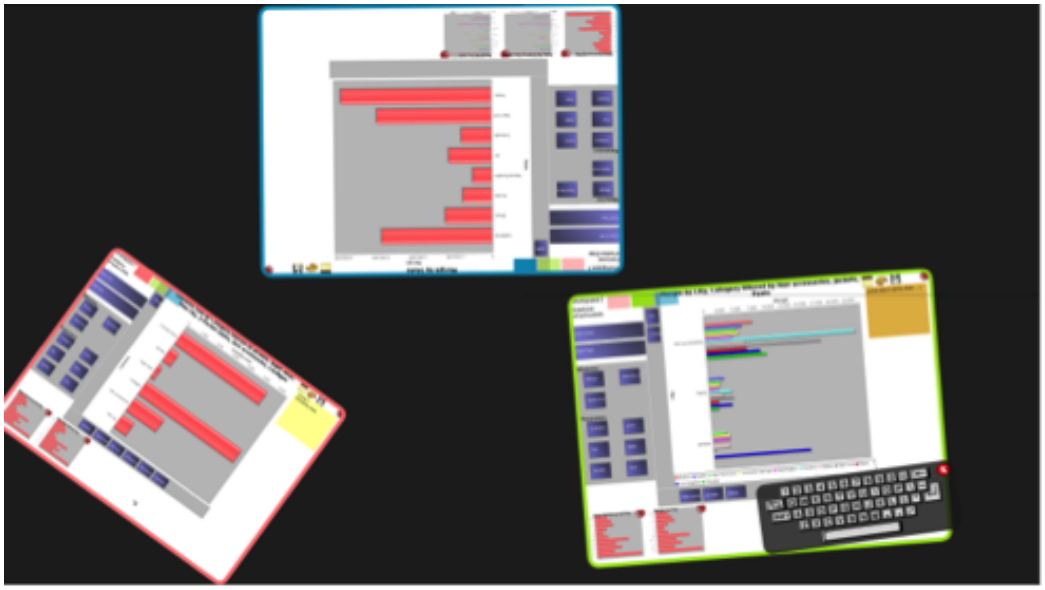


Figure 3.1: CoSpaces Interface. Dark background is the tabletop surface. There are three open Worksheets.

The CoSpaces interface is composed of Worksheets, as shown in Figure 3.1. The Worksheet was designed using the principle of "one space, many uses". Its design provides a team with the flexibility to work collectively on one or more Worksheets, or separately and simultaneously on multiple Worksheets. Each Worksheet defines a work territory, either personal or shared. Worksheets therefore enable both individual work territories and shared work territories, as advocated by Scott et al. [67]. Moreover, users may create several Worksheets, perhaps to investigate different data attributes and compare them side-by-side.

Personal versus shared Worksheets are identical as far as the system is concerned; ownership is define by the way in which they are used. This makes it easy for users to convert a personal space into a shared space or vice versa. Worksheets can also be moved and resized. Each Worksheet's relatively wide border is uniquely coloured with a bright distinc-

Figure 3.2: Worksheet Details: Analysis pane (A) for creating and modifying charts, Visualization pane (B), History pane (C), Notes pane (D), and Tabs (E) that provide a portal view to other worksheets.

tive colour. This enables users to easily distinguish Worksheets from each other. Sections of a Worksheet are shown in Figure 3.2.

3.1.2 History Module

The history module served to track and facilitate individual work. Analysts could review their previously created visualizations and reuse captured artifacts to perform analytical tasks such as chart comparison.

I designed the history interface such that it facilitated capture of both visual artifacts (i.e. charts) and users' externalizations (i.e. notes). Notes and visual snapshots were linked to the underlying analysis state so that the state could be easily reloaded by tapping on a note or dragging a thumbnail to the central area. I define an analysis-state as the information that is required to replicate a system state at a later time (i.e. mapping and filtering information plus the chart type).

A Worksheet automatically captures and saves a copy of the current analysis-state right

before a change, made by the user, has been applied. I use a simple heuristic inspired by the chunking rules of Heer et al. [30] to reduce history repository size. An analysis-state is saved only when a change in the current mapping of data takes place. In other words, adding or removing filters will not result in an automatic save. An analyst can externalize findings, hypotheses and so on using the notes pane. The importance of connecting externalized material to the visual representation of data has been previously recognized [13] [48]. Therefore, I automatically create a link between the current chart and the note.

As part of the analysis-state, I capture a thumbnail picture of the chart. Following the common design of history representations for VA, thumbnails are placed in the history pane in chronological order from oldest to newest (Figure 3.2C). The pane scrolls as the number of thumbnails grows. Notes are placed in the notes pane in chronological order, matching the chart thumbnails. The notes pane scrolls when the available space is exceeded. Moreover, since a note and its corresponding visualization are linked, an analyst can easily reload that specific visualization from the note.

3.1.3 Tab Portal Views

CoSpaces uses a tab metaphor to facilitate awareness of other users' analysis history plus sharing of artifacts. Coloured tabs at the top of each Worksheet (Figure 3.2E) are associated with other existing Worksheets. Each tab is colour-coded to match the border colour of the Worksheet that it links to. Tabs act as portals to view other Worksheets. Tapping on a tab replaces the local worksheet content with a view of another Worksheet. Tapping on the local Worksheet tab switches the view back. When another Worksheet's tab is selected, the contents of all panes are changed to reflect the remote information, including the current visualization as well as recorded items in the history and notes panes. The user may browse charts and notes to learn about another user's past analytical activities and interests. To prevent unintentional changes and interruption, a Worksheet's remote view is read-only and navigation in a remote view is not linked to the other Worksheet's local view. To share charts, one can select an item in the history pane of a remote view and copy it to the local Worksheet's history pane.

3.1.4 Implementation

CoSpaces is multi-touch application written in JAVA. The Multi-touch for Java (MT4J) open source API was used to provide multi-touch capabilities within the Java code. CoSpaces

uses the TUIO protocol to communicate with the finger and object tracking software, Community Core Vision (CCV). I used the JFreeChart chart library to create and display statistical charts of data. I have adopted prevalent multi-touch gestures for scaling, rotation and translation. User interface objects such as Worksheets and on-screen keyboards are scaled, rotated, and translated by using two or more fingers. I have also implemented target highlighting to facilitate chart creation. As users drag a data dimension into the Visualization area, sections of the Visualization pane highlight only if the selected data dimension may be dropped in that location.

3.2 Observational User Study

I observed pairs of participants working collaboratively on an analysis task using CoSpaces on an interactive tabletop. My goal was to observe how people use history in practice and what were strengths and shortcomings for supporting exploratory collaborative VA. Therefore, I focused primarily on users' actions that involved use of history items and notes. Details about this study including consent form, introduction to the task and system, tasks, questionnaire, and follow up interview questions can be found in Appendix A.

3.2.1 Pilot study

Prior to the user studies I run three pilot studies. The pilot study 1) ensured that the analysis tasks were clear and understandable, 2) enabled me to discover and fix bugs in the prototype tool, and 3) ensure instructions were adequate and unambiguous. During these pilots, I also determined ideal times for the different steps.

3.2.2 Participants

I recruited 10 pairs of computer science students (16 graduate students, 4 undergraduates; 15 male, 5 female) who were familiar with basic data analysis activities and basic statistical charts. Age ranged from 19 to 35 (average = 27). Pairs were not required to know each other beforehand. Participants were compensated with $20 each.

3.2.3 Tasks, Dataset and Procedure

Participants performed two tasks in which they could use system features freely and were not explicitly required to take notes or save charts. After a 20-minute introduction, they

started a training task (Task 1), which took about 30 minutes and focused on learning CoSpaces (details can be found in Appendix A). They could ask either of the two observers if they had any questions.

After Task 1, each group was given a short 5-minute break to rest and read Task 2. Task 2, which took almost 40 minutes, required exploring the dataset in search of any interesting findings that would indicate both strong and poor performance. The two tasks were followed by a questionnaire and a follow up interview that took almost 20 minutes.

The dataset used for this study included sales revenue, margin and quantity sold of clothing items in eight US states for three consecutive years, and consisted of 9 dimensions (i.e. columns) and 3273 rows. The sample data set was provided by our industry partner SAP Business Objects.

3.2.4 Apparatus

For this study, I used a rear-projected 70-inch (diagonal) tabletop with a resolution of 3840 x 2160. The tabletop used a rear mount infrared camera to detect a (practically) unlimited number of touches.

3.2.5 Data Capture and Analysis

My colleague[2] and I independently observed users' interactions. I also videotaped each session. 400 minutes of video data were collected (around 40 minutes for each session). Then we manually coded the video data using a two-pass analysis approach. We first analyzed videos together to identify a set of repeated actions on history items and notes. In the second pass, we coded each individual's activities using the defined set of actions. My coding and qualitative observations are based on Task 2, as Task 1 was only intended as practice.

3.3 Findings

I focus on qualitative observations and participants' comments from the interviews. However, for completeness, I also include quantitative results from the questionnaires and qualitative results from the observations and videos. In my dissertation, I will only report

[2]Please note that this research was result of collaboration with my colleague, Narges Mahyar. Thus any referral to "we" in this section acknowledges her work.

findings that are related to use of visual history by participants and will not include findings related to note-taking, collaboration style and analysis phases. A complete report of the findings can be found in [52].

Based on my observations, participants used history for two main purposes, Reuse and Review (Table 3.1).

Action	Description	Count
Reuse	Reloading a previously created chart from the history, either the local history or a collaborator's history	163
Review	Browsing the thumbnail images of the charts within the history, either the local history or a collaborator's history	141

Table 3.1: Primary actions on visual record-keeping and the frequency of each.

Both Reuse (Total= 163, Avg. = 16.3, StdDev.=15.58) and Review (Total=141, Avg.=14.1, StdDev.=4.9) actions were performed to achieve more than one goal. To infer user intentions, I relied on participants utterances, my observations, and action sequences. In the case of Reuse, participants reloaded charts for three main reasons 1) reinvestigate, 2) analysis and 3) support discussion. Based on my analysis, I identified 109 instances of reloading a chart from history for the purpose of reinvestigating the view or mapping/filtering of data. I also identified a total of 51 cases of reloading a chart for trying a new analytical path or drilling down. In three cases, participants reloaded charts while having a discussion with their collaborator, as evidence to support their reasoning.

Review happened to achieve two main goals 1) recall and 2) search. In 113 Review cases, participants only reviewed charts to recall the depth and breadth of analysis so far. I observed that participants would browse the charts in the history (in some cases reading the chart title (that included the mapped dimensions and filtering information) out aloud) to try to gain an understanding of the coverage of dimension space to determine what to explore. In 26 instances, participants reviewed history in search of a specific chart. In all cases, this action was a followed by reloading a chart from history. Figure 3.3 shows a breakdown of actions and intentions.

Interestingly, I also observed that participants frequently used history without direct physical interaction. On several occasions, I observed head movement, suggesting that a participant quickly glanced at the visible portion of the history pane where the most recent charts were placed. This quick review happened under various circumstances. For

Figure 3.3: Breakdown of Review and Reuse actions and identified primary user intentions for each.

instance, a quick review happened after almost every work interruption. It also often occurred before making a new chart. This could have helped participants to stay focused on the recent analysis path or to confirm that they had not made that chart already. Quick review could have also been performed for making non-detailed comparisons between the current visualization and charts in the history pane. Without eyetracking data, counts of these quick review actions would be unreliable. I therefore only counted and categorized concrete actions on history (when there was a clear physical direct interaction with the system), and do not have quantitative information for quick review actions. Nonetheless, the observation that these quick review actions occurred suggests that visible thumbnails of recent visualizations provide useful support for data analysis.

In the follow-up interview, 11 participants explicitly regarded history as one of the most useful features of the system. For instance one participant expressed that "the ability to save the charts is great" and another one said "reloading from history was really fast and efficient and worked well for me."

3.4 Discussion

To summarize, in this chapter I investigated RQ2: What are the limitations of linear history representation for collaborative EDA? Based on the findings of my observational user study, in the context of exploratory collaborative VA, participants mainly used history to review/recall past analysis and/or reuse created visualizations. Using the linear history, participants reloaded charts and tried a new avenue of analysis or drilled in by manipulating mapping and filtering of dimensions. An interesting observation was that participants used

history to understand the scope of exploration. More specifically, they reviewed history to gain a holistic understanding of the coverage of dimension space and decide on what to explore next. I noticed that this process was rather slow and cumbersome, especially when history was relatively long. In order to retrieve the information about completed work, participants had to browse the history and rely on their memory to record the extracted information. Although the linear history representation contained dimension space coverage information, it did not provide a first-hand overview.

The list of actions on history is based on my own observations and could be influenced by this study and tool design. For example, the limited history representation did not provide search and filter actions on recorded artefacts. The frequencies of actions that I observed are also undoubtedly related to particulars of the study and design of CoSpaces. I suspect that the actions and intentions themselves would be repeated in other VA situations, but that their distribution over time and their relative frequency could change. For instance, with a group of three or more participants, I speculate that there may be more instances of Review since it would be more difficult to keep track of what everyone is doing. Similarly, a more complex task might lead to the use of more reloads from history items in order to branch the analysis to a greater degree.

I also recognize that my inferences may not always be correct, and so these numbers should be taken as approximate. For example, some instances of the Reuse action could have been to replace a wrongly reloaded chart. In addition, the frequencies of actions and primary user intentions are influenced by the system design and the individuals. For example, I suspect that there would have been fewer search actions if the tool had a better search mechanism.

3.5 Conclusion

In this chapter, I addressed the second research question (RQ2): What are the limitations of linear history representation for collaborative EDA? I designed CoSpaces and evaluated visual record keeping for exploratory collaborative VA. I conducted a user study which showed that the linear history adequately supported reusing of captured states for branching the analysis. On the other hand, I observed that it was rather cumbersome for participants to gain an understanding of the coverage of dimension space by using this representation. Despite the observation that users innately refer to history to seek an understanding of what they've done so far, the representation did not provide first-hand information about the coverage of dimension space, making it hard for people to assess what was left to

investigate. Based on these findings, I suggested representing history from an angle that would provide top level and first-hand insight about the coverage of dimension space. In the next chapter, I further discuss this approach. More specifically, I address RQ3:

Does representing analysis history from a dimension-centric angle better support understanding the coverage of dimension space than a linear representation of history?

Chapter 4

Understanding the Breadth of Exploration: Linear History versus Visualizing Dimension Space Coverage

4.1 Introduction

In this chapter, I will investigate **RQ3: Does representing analysis history from a dimension-centric angle better support understanding the coverage of the dimension space than the linear representation of history?** Based on the findings of my observational user study (Chapter 3), in the context of EDA, analysts refer to history to build an understanding of dimension space coverage. Yet, the linear representation of captured analysis states does not efficiently support gathering this information. I observed that users had to browse and examine charts in history individually to collect dimension space coverage information. This model makes gathering a holistic view of the breadth of the dimension space coverage tedious and inefficient, especially as the analysis history grows. In addition, I observed that participants had to rely on their memory to recall dimension space coverage information at a later time. Based on these observations, I speculated that representing history from the angle of dimension space coverage (in addition to the linear representation) would facilitate understanding and answering questions about what another person investigated in a prior analysis.

In this chapter, I introduce Footprint-I, a prototype tool that provides quantitative, relational and temporal information about investigated dimensions through a number of interactive and coordinated views. Later, I report the design and outcomes of a user study

that compares Footprint-I against a linear representation of history. I hypothesized that Footprint-I users would be faster and more accurate in understanding a collaborator's analysis history from the angle of dimension space coverage.

In the rest of this chapter, I will first describe the design of Footprint-I. Next, I will report the design of the user-study and conclude with outcomes of the study.

4.2 Footprint-I

Footprint-I (Figure 4.1) is a prototype history tool specifically designed to visually represent the coverage of dimension space. The main design objective was to help an analyst quickly understand the breadth of dimension space coverage in a past analysis session done by a collaborator. Footprint-I provides temporal, relational and quantitative information about prior investigation of dimensions through three interactive and coordinated views: Dimension View, Timeline View, and List View, as shown in Figure 1. The following subsections describe each view in more detail.

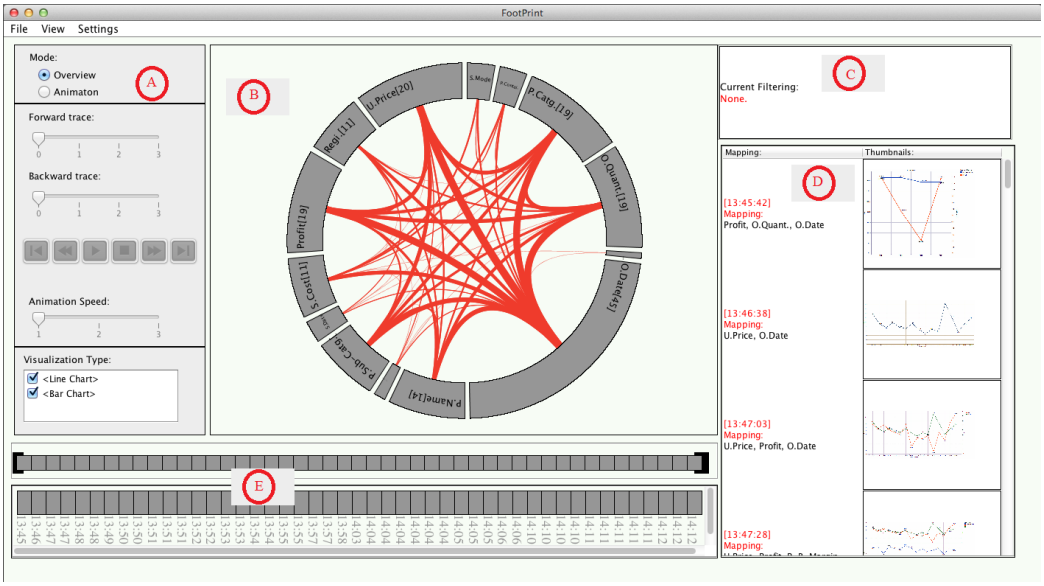


Figure 4.1: FootPrint: a prototype history tool for exploratory data analysis. The tool contains three views: Dimension View that provides quantitative and relational information about explored dimensions (B); List View that provides details about visualizations created (D); (E) Timeline View that delivers temporal information about analysis progress. A control panel (A) contains controls for various settings and the filtering panel (C) shows how the history views are filtered (if applicable).

4.2.1 Dimension View

The circular Dimension View (Figure 4.1B, Figure 4.2) is primarily designed to help a user gain quantitative and relational insight into the exploration of dimension space. It provides information about the investigated dimensions (i.e. dimensions that were mapped to visual variables to create visualizations) at different levels of granularity. In addition, it provides information regarding co-mapping of dimensions (i.e. combinations of dimensions that were investigated). I consider any two (or more) dimensions related if they are co-mapped in a visualization. For example, Profit and Year are related because the analyst created a bar chart depicting the trend of Profit over Years.

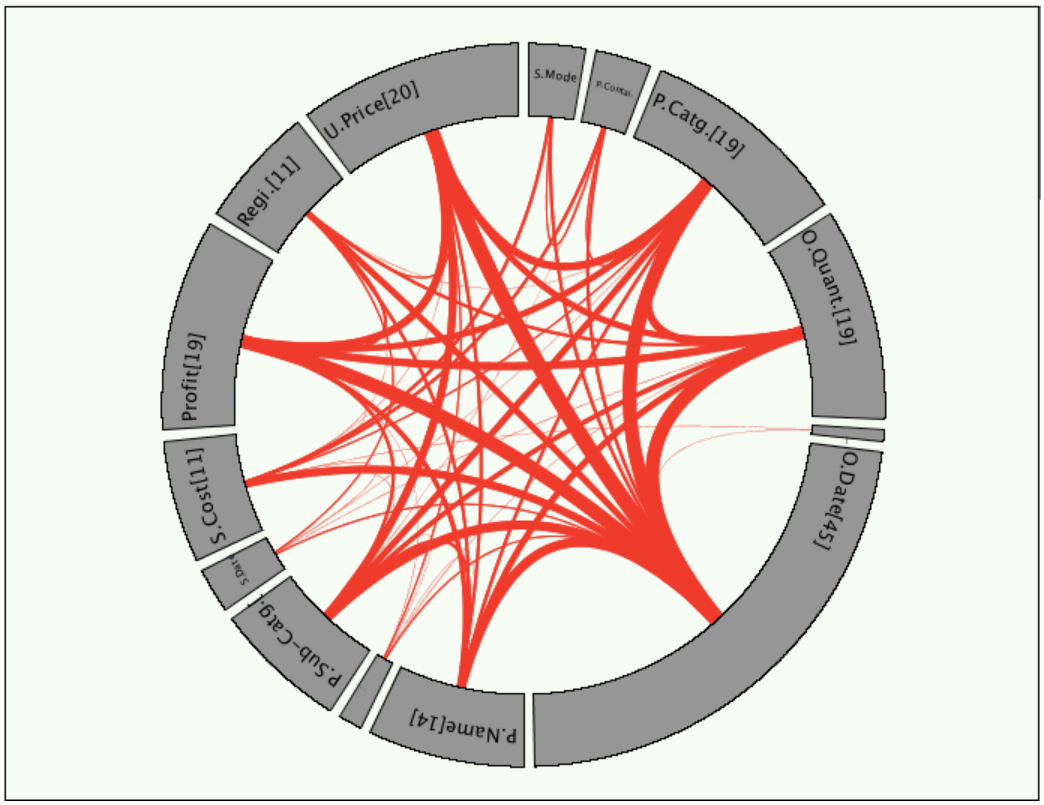


Figure 4.2: Initial state of Dimension View.

The Dimension View has a circular layout similar to Circos [45] that facilitates the display of relational and quantitative information. Each investigated dimension is represented as a curved rectangular segment, and collectively segments form a circle. Size of each segment is proportional to the total number of times the dimension is included in a visualization. Relationships are represented as curves connecting the segments. Width of a curve is relative to the total co-mapping frequency between pairs. I chose a circular layout

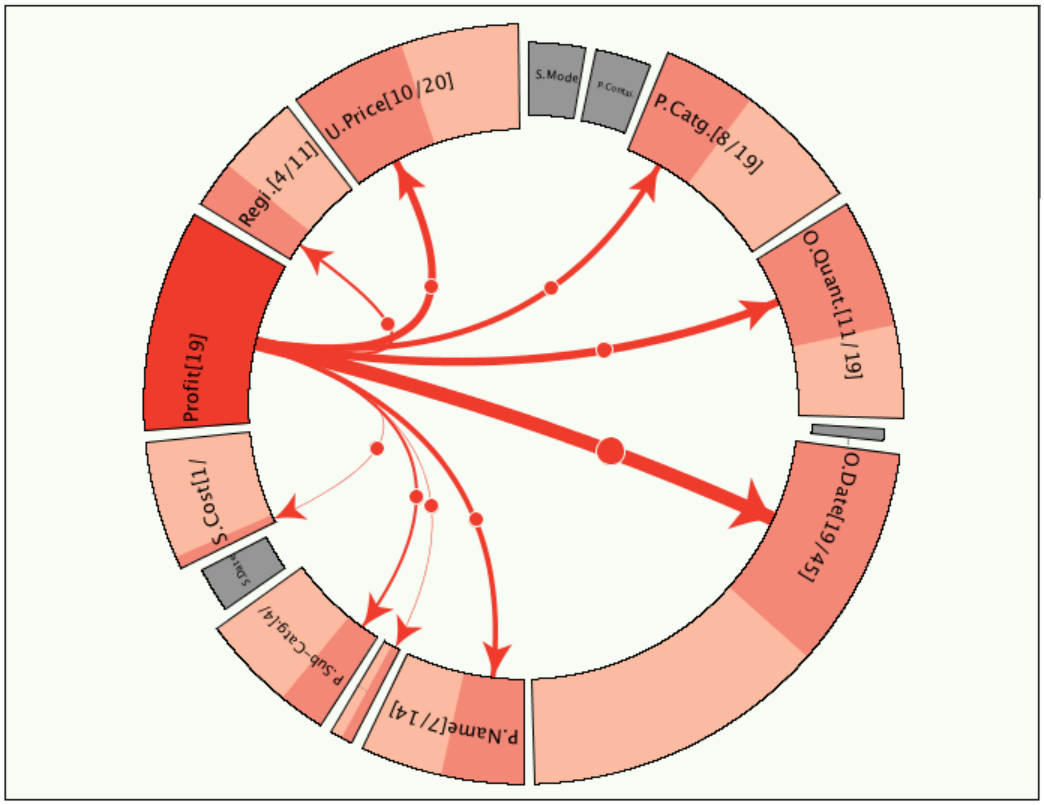


Figure 4.3: Transformation of Dimension View (from Figure 4.2) after selecting Profit.

for this view because it facilitates showing relationships between dimension pairs.

Initially, all the segments are rendered in grey (Figure 4.2). Labels on each segment display the name of the dimension represented by the segment and the total number of mapping instances for that dimension. Each segment can be selected to gain more information about a dimension. Figure 4.3 illustrates changes that took place in Dimension View after selecting a segment.

Changes in the sizes and the colours of the circle segments provide visual cues for fast processing of information. As shown in Figure 4.3, the selected segment (i.e. Profit) becomes wider and turns red in colour. Segments representing dimensions that have been mapped along with Profit also become wider and are coloured. Other segments remain small and grey. In the coloured segments, the two-tone colouring represents a proportion. The size of the darker portion represents the frequency with which that dimension has been co-mapped with the selected one (Profit). The two-shade design enables quick comparison of co-mapping information between dimensions. For example, a quick examination of Figure 4.3 shows that Profit was co-mapped with Unit Price (U.Price) about twice as often as with Region (Regi.). Labels of the segments provide exact co-mapping figures. Curved

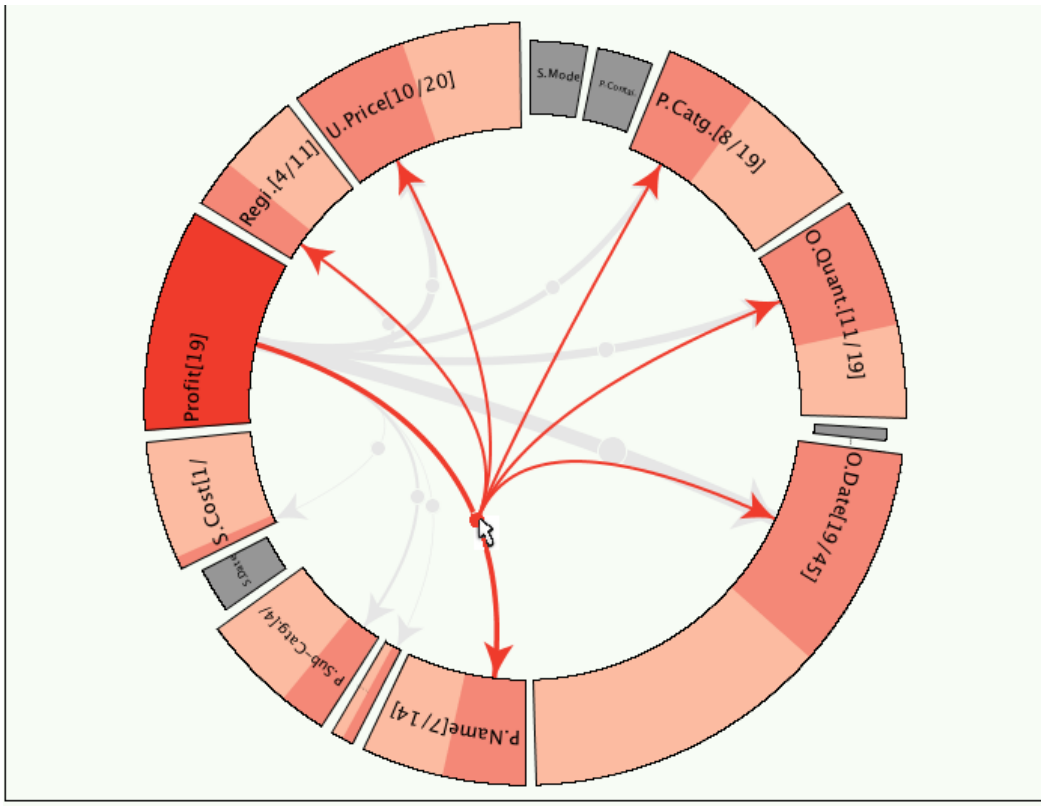


Figure 4.4: Dimension View after the user has placed the cursor over the InfoSpot between Profit and Product Name.

lines in Figure 4.2 are replaced with arrows in Figure 4.3 that travel from the selected dimension to segments with a co-mapped dimension. Thickness of each arrow is relative to the co-mapping frequency of the dimension pair that it connects. Selecting a segment in Dimension View also updates Timeline (section 4.2.2) and List (Section 4.2.3) views. Details of these changes are described in later sections. Each arrow has an InfoSpot (i.e. the filled circle on the arrow) that can be used to obtain more information about a pair. Figure 4.4 shows the result of placing the mouse cursor over the InfoSpot on the connection between Profit and Product Name (P.Name). New diverging arrows traveling from this InfoSpot to other dimensions indicate that Profit and Product Name were together investigated with these other dimensions. For example, based on Figure 4.4, the user can conclude that the Profit and Product Name pair has been investigated in conjunction with Region, Unit Price, Product Category, Order Quantity and Order Date. In other words, the analyst had created visualizations that mapped Profit and Product Name combined with each of these dimensions (or more than one). This InfoSpot design enables a user to quickly obtain an overview of the investigation of a dimension pair in relation to other dimensions.

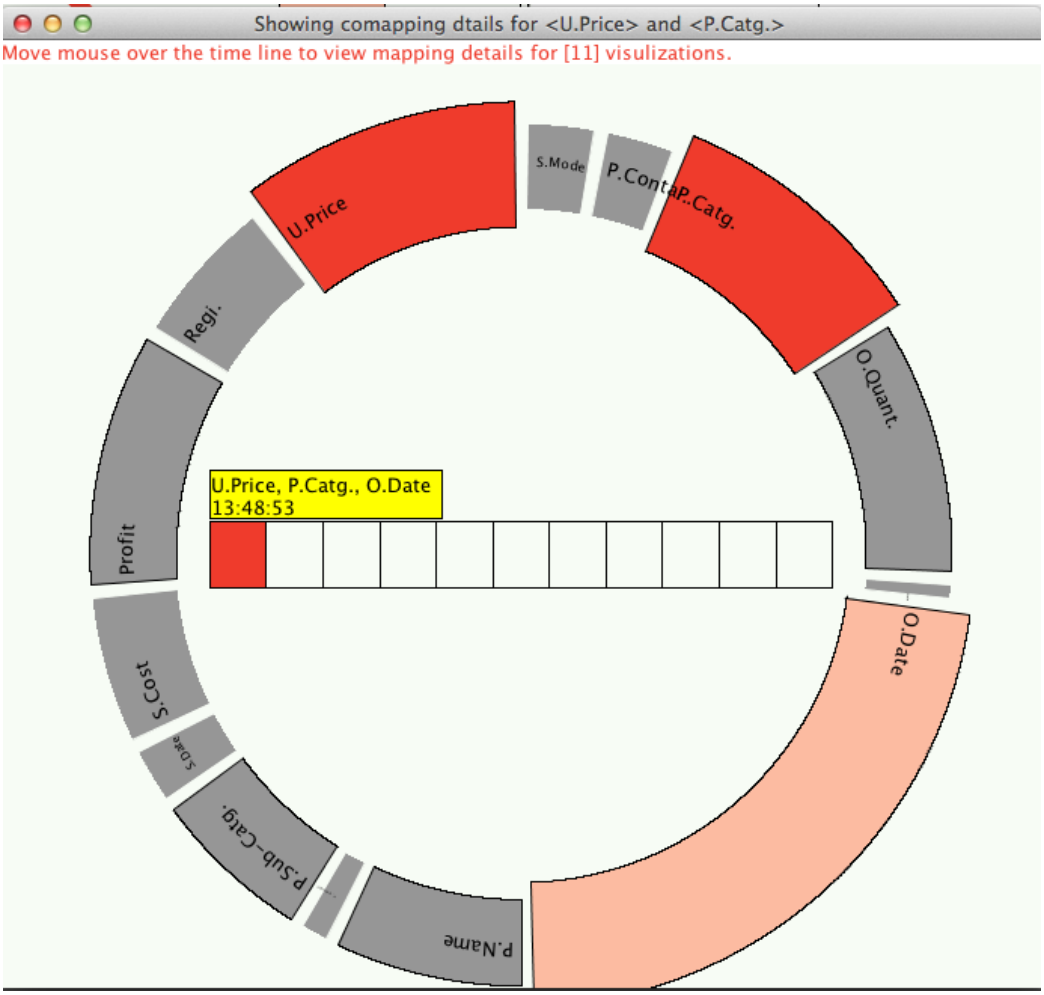


Figure 4.5: . Co-Mapping Details View. The user has opened this view by clicking on the InfoSpot for Profit and Product Name. This figure depicts the moment when the mouse cursor is placed on the first cell in the timeline. A tooltip-like label shows the exact mapping and time information for the visualization represented by the first cell. The circle shows additional dimensions involved in that visualization by highlighting them in light red. This implementation also enables comparisons between different dimension pairs by having multiple external windows open simultaneously.

Co-Mapping Details View

Clicking on an InfoSpot pops open an external window designed to provide detailed information regarding a dimension pair (the co-mapping details view). Figure 4.5 shows the co-mapping details view window that was opened after clicking on the highlighted InfoSpot in Figure 4.4. This view consists of a label on the top of the frame indicating the total number of visualizations created with this pair of dimensions (in the case of Figure 4.5 the number is seven), a replica of the circular Dimension View with segments representing the pair in red, and a timeline placed in the centre of the circle. Moving the mouse over each cell in the timeline shows details about a single visualization and related dimensions in the circle.

4.2.2 Timeline View

Timeline View (Figure 4.1E) provides temporal information about the analysis history. Cells in this view represent visualizations that were created by the analyst, in chronological order. Similar to the initial rendering of Dimension View, all the cells are initially rendered in grey. A user can click a time cell to gain detailed information about the visualization created at that specific timestamp. As illustrated in Figure 4.6, the clicked cell is rendered in red to depict selection, and information about the visualization is provided through changes to Dimension and List Views. The segments representing mapped dimensions become red and larger in size, with arrows connecting them. The List View shows a thumbnail image of the visualization that can be enlarged.

As shown in Figure 4.1E, two differently scaled versions of the Timeline were presented. At the top is an overview that would always fit the length of the panel it was placed in. In the bottom was a larger, scrollable version with timestamps of each cell shown.

4.2.3 List View

Initially, this view contains a temporally ordered list of thumbnail images of all the created visualizations. The first visualization is placed at the top of the list and the last visualization at the bottom. A label next to each thumbnail provide detailed information about creation time and mapping and filtering of dimensions. Clicking on a thumbnail opens an external window that contained the full size visualization. This view is the equivalent of the common linear representation of history.

User interactions with Dimension View and Timeline View updates the list. When a

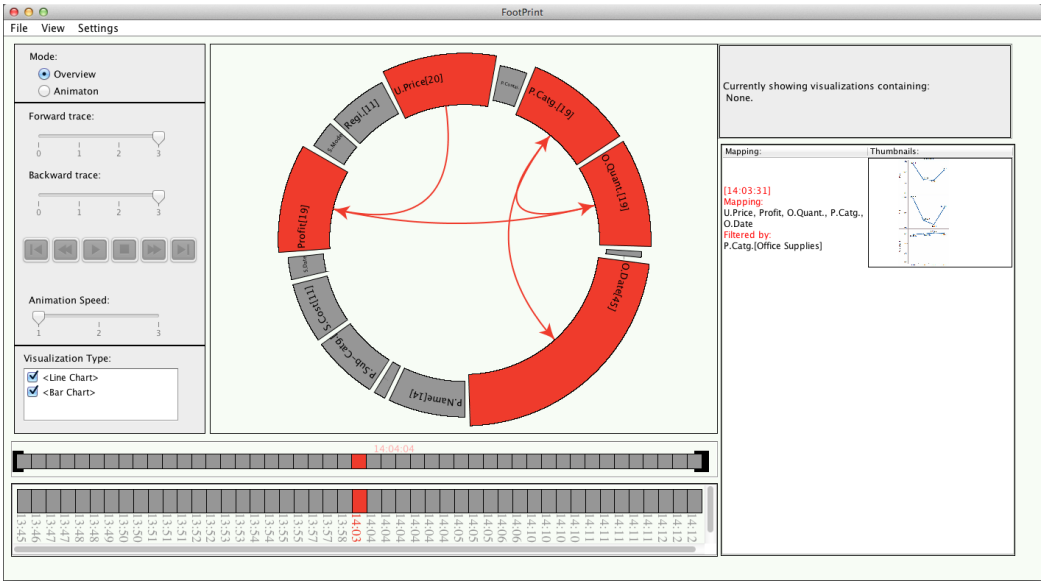


Figure 4.6: Selecting a specific timestamp in Timeline View.

segment in Dimension View is selected the List View is filtered to show only visualizations that include the selected dimension. Information about the current state of the list is provided in the label at the top of the List View (Figure 4.1C). As shown in Figure 4.1D, labels next to each visualization are also re-formatted and the selected dimension (i.e. Profit) is underlined, italic and bold. This design aims to facilitate fast and easy recognition of the selected dimension throughout the list.

4.3 Implementation

Footprint's history repository was comprised of visualizations created during an exploratory VA session along with timestamps (times of creation). Each visualization was manually processed to extract mapping, filtering, and type (e.g., bar, pie, or line charts) information. This data was stored in a Microsoft Excel spreadsheet file (I will refer to this spreadsheet as the history file) where each row contained data about a single visualization. Footprint-I was implemented in Java. All the shapes and layouts were computed and rendered by custom modules that used Java's native libraries. This enabled me to design visualizations and incorporate interactions that I believed would optimally fit the user's needs. Apache POI [1] was used to read the history file.

4.4 Evaluation

I designed and conducted a between-subjects experiment to assess how FootPrint-I's unique perspective into the coverage of dimension space compared to a baseline tool that used a linear representation of history states (described in section 4.4.2). I used time and accuracy as two metrics to assess participants' performance in terms of understanding the prior dimension space coverage.

4.4.1 Preparation of History

In order to prepare the history file, I asked a senior computer science PhD student with management background and visual data analysis experience to explore a sales data set using Tableau Public, a commercial VA tool based on Polaris [70]. The student was not involved in the design of Footprint-I. The task required her to investigate sales data in search of any interesting findings. I employed a think-aloud protocol and asked her to explicitly express questions (sub-problems) that she was investigating to solve the problem. During

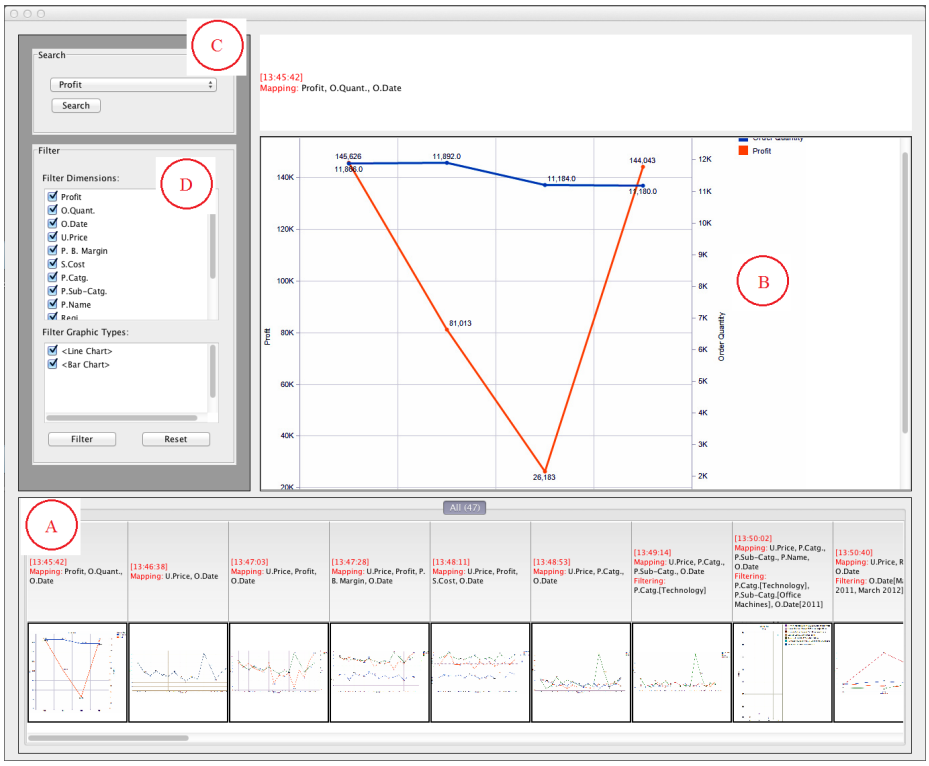


Figure 4.7: Baseline tool used for comparison in our study. Past states are represented using a chronological list of thumbnail images (A). Other panels show a detail view (B), and search (C) and filter (D) controls.

her analysis, she produced 47 visualizations. From these, I manually extracted dimensions mapped to visual variables, and any filtering of dimensions that returned a subset of values. This information was stored in a spreadsheet where each row contained information about a single chart. This information included timestamp (i.e. time of creation, extracted from videos), dimensions mapped, filtering of dimensions and chart type. This spreadsheet was then used in my user study as the analysis history file. Captured information closely resembled a *state-based* [30] history model in which each history artefact encapsulates parametrical values about specific visualization and/or application states.

4.4.2 Baseline History Tool

I compared Footprint-I against a baseline history tool (I will refer to this as the baseline tool) that I designed and implemented. The design of the baseline tool closely followed the design of a linear history representation and was inspired by a number of prior history tools [30] [68] [50] for VA that represented history as an interactive list of thumbnail images

(either as the default view or as an option).

Figure 4.7 shows the default view of the baseline tool after opening the history file. Thumbnail images of the charts, arranged in chronological order from left (older) to right (newer), were placed in a scrollable list on the bottom of the window (Figure 4.7A). Above each thumbnail was a label with detailed information about a chart including the timestamp and mapping and filtering of dimensions, identical to information available in Footprint's List View. Selecting a thumbnail would render the actual size image of the chart in the panel above the list (Figure 4.7B).

The baseline tool provided support for searching and filtering of the history repository based on dimensions. A user could perform a search by selecting a dimension from a drop down list (Figure 4.7C) that contained dimension names. Each search would return a subset of charts that contained the selected dimension as one of the mapped dimensions. The user could perform filtering by un-checking (i.e. excluding) the box next to a dimension's name on the filter list (Figure 4.7D). The result of a filtering operation was a subset of history items, excluding charts that have un-checked dimensions as part of their mapping.

4.4.3 Participants

I recruited 20 computer science students as my participants (14 graduate, 6 senior undergraduate, 11 male, 9 female, average age of 27.9). They were randomly assigned to use either Footprint-I or the baseline tool. All the participants were required to have a basic understanding of visual data analysis and prior experience with tools (e.g. Microsoft Excel) that enable constructing visualizations or statistical charts based on data. None of the participants had participated in my previous user studies. To minimize the effects of individuals differences on the outcomes of the tests, we tried to recruit participants with similar background (computer science) and close age range.

4.4.4 Procedure

At the beginning of each study session, I gave a verbal description of the scenario (i.e. role the participant was going to assume and the task to be performed). Following was a textual representation of the scenario given to participants:

> You are a business data analyst in a large international company. You are working collaboratively with another analyst in your company to explore sales data for the past 4 years and evaluate business performance. The geographical distribution of you and your collaborator does not allow for synchronous and collocated collaboration. As a result, work done by a collaborator is passed to another to be built upon. In order to efficiently continue your collaborator's work, you first need to review and understand what they have done. To achieve this, you will review the history of their analysis.

At this point, using a history file from a different dataset and analysis (so participants would not be exposed to the actual history before the main task), I explained what and how information about a collaborator's work was stored in the history file.

The initial introduction was followed by an introduction to the tool features (either Footprint-I or baseline tool). Afterwards, participants practiced using the tool by doing a short warm up task with the example history file. The warm up task for each system required working with all the main features of the system. I was present during the warm up session and participants could stop and ask questions about the system, task and history file. A list of the supported user interactions and their outcomes for each system was left with the participant to be referred to (if needed) during the actual review task.

After the warm up task was completed, participants were asked to read a short document that explained in detail the problem that was being collaboratively investigated. Later on, participants were given a booklet that contained questions about the collaborator's work (explained in Section 4.4.5). We asked our participants to think-aloud and verbalize all their thoughts while doing the task.

On average, the preparation procedure took 40 minutes for Footprint-I and 25 minutes for the baseline tool. The time difference was due to a much larger set of features in Footprint-I that required more time to explain. The time that users were given to practice and learn the systems was equal (˜10 minutes). For both tools, I gave the participant a printed list of the features and supported interactions that they could refer to during the analysis. None of the baseline tool users referred to this list, and five Footprint-I users made very brief reviews. At the end of each analysis session I interviewed the participant for their verbal feedback on tool, task and any other comments.

4.4.5 Task

The review task was comprised of close-ended multiple-choice questions. The following sections provide details about each part (for more information please see Appendix B).

Part 1

For this part, the participant was asked to select dimensions that were explored (i.e. being ever mapped in a chart) from a list that contained names of all dimensions in the data set. This part was designed to evaluate participants' ability to understand the coverage of dimension space.

Part 2

Part 2 consisted of questions that tested participants' understanding of the co-mapping of dimensions (e.g. Did the analyst investigate the relationship between Unit Price, Order Date and Ship Cost?). This part tested participants' ability to understand the co-investigation (i.e."what was investigated with what").

Part 3

For this part the participant was asked to pick a diagram, amongst 4, that most closely depicted the temporal distribution of investigation of three dimensions, Profit, Unit Price and Ship Cost, over the course of the analysis session. This part was designed to evaluate a participants' ability to use history to form a high-level mental model of the analysis flow from beginning to end, based on the temporal distribution of most frequently mapped dimensions.

4.4.6 Data Capture

Participants were asked to record their answers to parts 1 to 3 within the paper-based task booklet. We audio recorded each session to capture participants' monologues while answering questions, and video captured the screen and logged users' interactions with the tool. We also video recorded the short interviews that followed the analysis task.

4.5 Findings

To measure accuracy, I checked answers against a key. To measure time, I used session videos and for each question extracted start and finish timestamps to calculate how much time a participant spent on that question. As part of the think-aloud protocol, I had asked participants to explicitly state the beginning/ending of each question using the phrase "I am starting question #" and "I am done with #". This enabled me to accurately extract time stamps.

4.5.1 Time Performance

Figure 4.8 illustrates the time (in seconds) spent by participants for performing parts 1, 2, and 3. As is evident from the figure, baseline tool users (BT) spent more time answering questions than Footprint-I (FP) users for all questions. The difference is substantial for parts 2 and 3, where Footprint-I users were nearly twice as fast. This corroborates my hypothesized benefits of Footprint-I in reducing the time required to acquire quantitative and relational information regarding dimension space coverage.

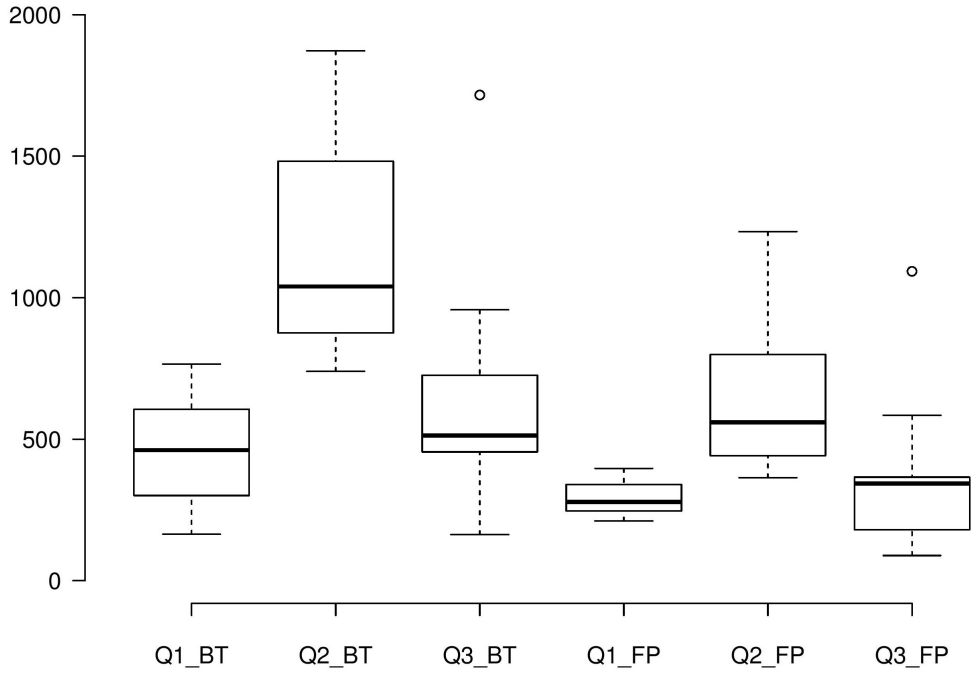


Figure 4.8: Boxplots showing time (in seconds) to complete parts 1, 2, and 3 with the baseline tool (BT) and FootPrint (FP).

After using a log transformation to improve the fit to a normal distribution, I analyzed the time results by 2 (Tool) x 3 (Question) ANOVA, with Tool as a between-subjects factor and Question as a within-subjects factor. Note that questions were not in random order, but I did not consider this important, as comparing between questions is not the purpose of our analysis. ANOVA showed a significant main effect of Tool ($F(1)$=9.4, p<0.004), demonstrating that Footprint-I was significantly faster than the baseline tool. There was no significant effect of Question and no significant interaction between Question and Tool.

To answer questions 1 and 2, a baseline tool user commonly performed searches and/or filters and reviewed the results to collect information required for answering a question. To answer the same question, a Footprint-I user typically used the Dimension View and InfoSpot. I believe the considerable time difference, especially for question 2, is due to the baseline tool's inability to provide direct co-investigation information about explored data dimensions. To answer question 3, all baseline tool users browsed the list of thumbnail images and checked the timestamp labels to gain an understanding of the temporal distribution of dimensions under investigation. In contrast, Footprint-I users used the Timeline View for the same purpose. The considerable time difference in performing question 3 indicates that the baseline tool's history representation was inadequate for providing information about the temporal distribution of dimensions in the analysis.

4.5.2 Accuracy

For each participant, we assigned an accuracy score based on the number of correctly answered questions. The maximum possible score was 11. As shown in Figure 4.9, Footprint-I users achieved higher overall scores than baseline tool users. As this data could not be transformed to match a normal distribution, I compared accuracy scores between Footprint-I and the baseline tool using the non-parametric Mann-Whitney test. Results showed that users of Footprint-I had significantly higher accuracy than users of the baseline tool (W=9.5, p<0.003).

4.6 Discussion

In this chapter I introduced Dimension View, a visual representation of analysis history from the angle of dimension space coverage. This view provided overview (i.e. what dimensions were investigated) and details-on-demand (i.e. co-investigation information) about the coverage of dimension space. I hypothesized that in comparison to a linear his-

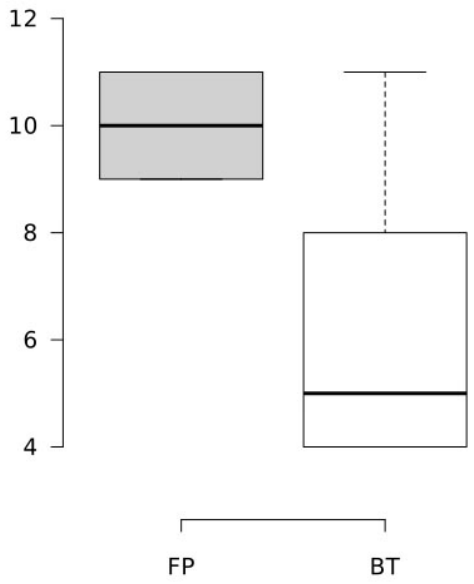


Figure 4.9: Boxplots showing total number of correctly answered questions for baseline (BT) and FootPrint(FP) tools.

tory, users would be faster and more accurate in understanding what another person investigated in a prior analysis. The results of a user study corroborated this hypothesis and showed that Dimension View helped users to be twice as fast and accurate in answering questions about the prior coverage of dimension space.

Similar to my observation in the CoSpaces user study (see chapter 3), acquiring dimension coverage information using the linear view required browsing/filtering the list of history items and extracting information manually. On the other hand, Dimension View provided this information first-hand. For example one the task questions (Question 2, part II, see Appendix B for details) asked participants to find dimensions that were co-investigated with "Unit Price + Order Date". Participants with access to Dimension View found the answer in two steps: 1) selecting Unit Price, and 2) placing the mouse pointer over the infoSpot on the connection between Unit Price and Order Date. Gathering similar information from the linear history required the following steps: 1) filtering the list to only keep charts with Unit Price and Order Date as mapped dimensions, 2) browsing the filtered set of seven charts and extracting co-mapping information. The latter process was more time consuming and error prone.

Understanding the coverage of dimension space in a collaborative EDA is can help in continuing a teammate's work, verifying analysis findings, and learning from an expert. I

anticipate that visualizing dimension space coverage will be even more useful for larger history repositories, because reviewing a linear history sequentially becomes very tedious in that case.

Although Dimension View outperformed the linear history in providing insight into the coverage of dimension space, the circular design seemed suboptimal. It had limited scalability and a large number of dimensions would result in very small segments with illegible name labels. In this design, I used the segment size to encode investigation frequency. Decoding frequency information would also be very difficult for many small segments. In addition, having too many dimensions would also result in too many connections in Dimension View, which would hinder a user's ability to understand co-mapping information. Similarly, Timeline View could become too long and tedious to browse. In the next version of my design (Chapter 5), I redesigned Dimension and List Views to address some of these problems.

4.7 Conclusion

To summarize, in this chapter I investigated RQ3: Does representing history from a dimension-centric angle better support understanding the prior coverage of dimension space in comparison to the linear representation of history? I introduced Footprint-I, a tool designed to support understanding analysis history from a dimension space coverage perspective. This approach provided a complementary way of reviewing analytic work, as compared to previous history visualization techniques. For example, it enabled analysts to answer questions such as "What dimensions were investigated, how often, and when?" and "What co-mappings of dimensions were most prevalent?". Answering these kinds of questions could be useful for reviewing analysis done so far and deciding how to proceed. This sort of review is particularly important for exploratory collaborative analysis, where one person needs to come up to speed on another's previous work in order to build upon it. Results of my user study suggest that the dimension space approach implemented in Footprint-I was both faster and more accurate in providing such information.

After positive results from this study, I asked the next natural question: **(RQ4) How does dimension space coverage information influence group coordination?** In the next Chapter, I will describe the effects of providing dimension space coverage on collaboration. In addition, I will describe the redesigned Dimension and List Views.

Chapter 5

Investigating the Effects of Providing Dimension Space Coverage Information on Task Coordination

In this chapter, I will investigate **RQ4: How does dimension space coverage information influence task coordination?** In chapter 4, I showed that visualizing the coverage of dimension space assisted analysts to more quickly and accurately understand what dimensions had been explored and in what combinations in a prior analysis session.

In this chapter, I will investigate the effects of visualizing analysis history from the angle of dimension space coverage on group coordination. Group coordination refers to "the way in which group members synchronize theirs actions in order to successfully complete the group task" [79]. Depending on the task and group dynamics, synchronization can happen in different forms, levels and temporal distributions [79]. For instance, a group whose task is to generate design ideas "may wish to avoid duplication of ideas in order to maximize the quantity of ideas produced" [79].I made an assumption in this work that a group of analysts working on the same data analysis task may wish to avoid duplicating analysis unless required for a reason (e.g. validation of findings).This seems reasonable in an exploratory analysis situation where the primary focus is finding as many trends and outliers as possible. I made this assumption based on prior research [20] [12] [29] in the asynchronous collaborative analysis field that suggested an awareness of "what has been covered" can direct current work towards "what has been left out". In addition, EDA is a breadth-oriented activity by nature and efficient exploration partly relies on the greater coverage of dimension space [75]. I speculated that representing dimension space coverage

would enable an analyst to focus on uninvestigated aspects of the exploratory task, hence improving group coordination.

To measure synchronization, I compared the overlap between the dimension space coverage of the past analysis and the coverage of the next person. I posited that representing the dimension space coverage would result in smaller overlaps between past and current work. Results of a user study corroborated this speculation. The main contribution of this chapter is described in Chapter 1 as C5.

5.1 Introduction

Group coordination is critical in distributed collaborative exploration of multidimensional data. To achieve synchronization and effectively coordinate their efforts, collaborating analysts need to understand what each person has done and what avenues of analysis remain uninvestigated. Two factors of time and explicitness determine how groups orchestrate efforts. Pre-plans (i.e. verbal or formal division of work before it starts), in-process planing (i.e. dividing work explicitly as it proceeds), tacit per-coordination (i.e. unspoken and implicit division of work before it starts), and in-process tacit coordination (i.e. implicit division of work as it proceeds) are four extremes of the continuum [79]. In a distributed collaborative exploratory VA setting, in-process tacit coordination is the most likely scenario to happen. In this case, collaborators are distributed across different times and places, making communication and planning rather difficult. In addition, the exploratory nature of the task and vague analysis goals inhibit pre-plans and explicit coordination. As a result, analysts need to rely on a different channel for understanding the depth and breadth of others' work. In this chapter, I introduce and evaluate a new visual representation for Dimension View. (For a description of Dimension View and its original design, see Chapter 4.) Similar to the old design, the new visual representation reveals which dimensions (i.e. attributes or variables in a tabular dataset) have been explored in past analysis and in which combinations. I will demonstrate that revealing this information can help collaborating analysts to better coordinate their work by improving their understanding of each other's activities in an exploratory collaborative VA context.

To investigate RQ4, I designed and implemented Footprint-II, a redesigned version of Footprint-I (see Chapter 4). Similar to Footprint-I, this prototype history tool visually represented analysis history from different angles. The redesign of Footprint-I was based on the feedback from the participants of a user study (see Chapter 4) as well as four visualization experts. In particular, experts commented that the circular design of Dimension View

in Footprint was not space-efficient and would not scale well. In addition, they suggested that the view should show high-degree relationships between investigated dimensions (i.e. co-mapping). Footprint participants complained about poor legibility of curved segments' labels and difficulty of tracing curves connecting segments (especially when number of curves increased). In addition, based on the feedback from the visualization experts, I added a new representation, Data View, that represented analysis history from the perspective of *data values* coverage. The next section describes Footprint-II.

5.2 Footprint-II

Footprint-II (Figure 5.1) was comprised of three main views: 1) Dimension View, 2) Sequence View, and 3) Data View. Subsections describe each view in more detail.

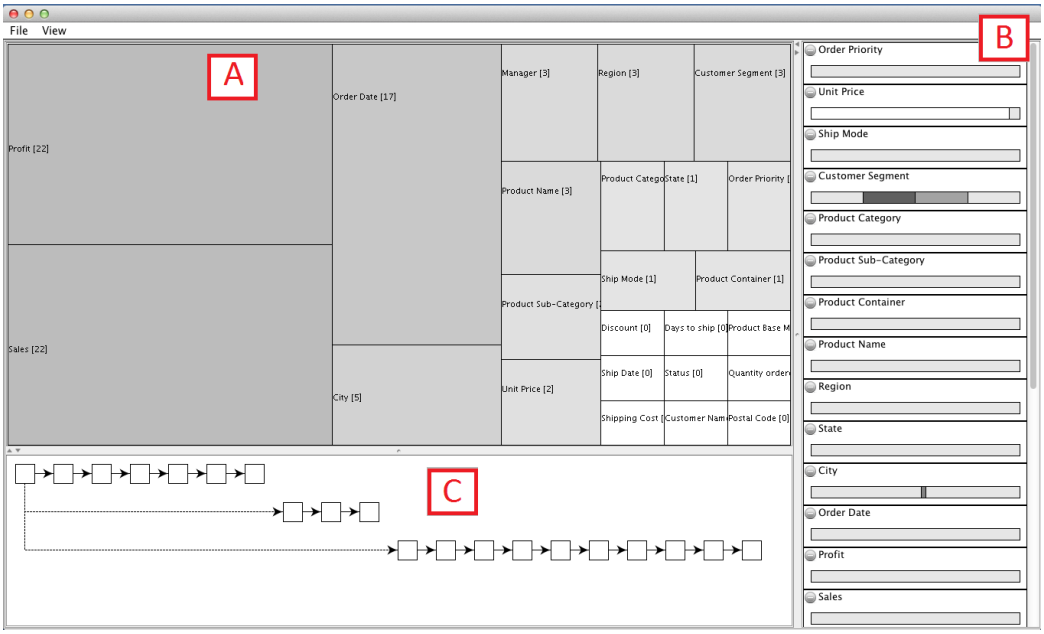


Figure 5.1: Footprint-II: Dimension View (A) and Data View (B) provide insight into the collaborator's coverage of dimension space and data space, respectively. Sequence View (C) depicts the branching and temporal progression of the analysis session.

5.2.1 Dimension View

At this point, I redesigned Dimension View (Figure 5.1A and Figure 5.2) as a treemap with a squarified layout, in which each cell represents a data dimension. Treemaps are mainly used for visualizing hierarchical quantitative data. Although my data is not hierarchical and

I only have two categories of *investigated* and *uninvestigated* dimensions, I found that the space-filling nature and scalability of the treemap made it a suitable choice for this view.

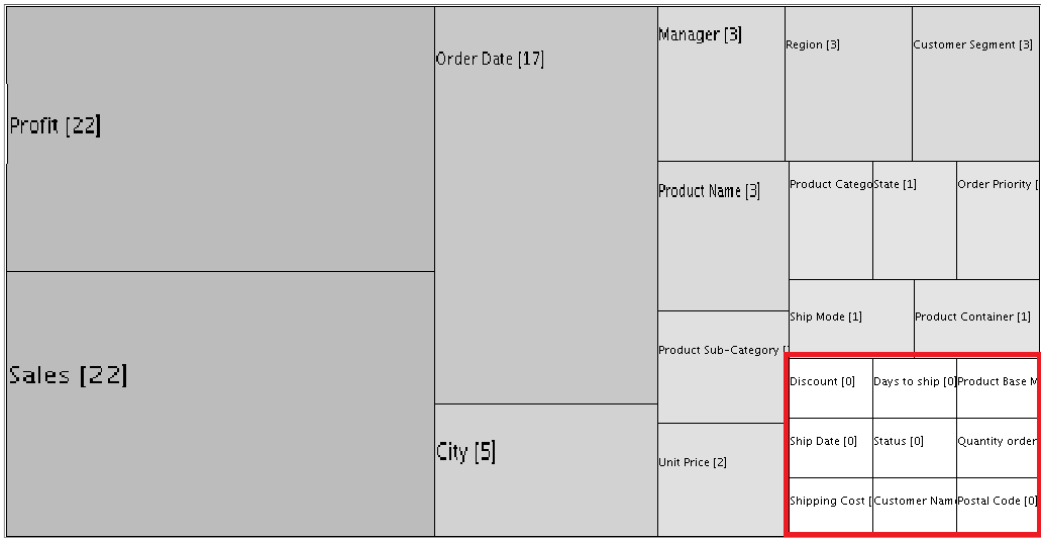


Figure 5.2: Initial rendering of Dimension View. Uninvestigated dimensions are grouped together and rendered with white background, shown enclosed within the red border (red border added for illustration purposes). Investigated dimensions have grey backgrounds. Greyscale and size redundantly encode mapping frequency.

Following the common design model in visualization, this view supported gaining an overview and investigating details on demand. The overview (Figure 5.2), which was also the initial rendering of the view, enabled users to instantly discover investigated / uninvestigated dimensions and the focus of prior work. I used redundant encoding with greyscale and size to convey this information. Size and greyscale represented the relative mapping frequency of a dimension: large dark grey rectangles represented the most frequent dimensions and small white rectangles represented dimensions that were never investigated. I refer to mapping as encoding a dimension in a visualization (in the previous analysis session) by an element such as position, size, shape or color. Each cell was also labeled with the dimension name and mapping frequency (e.g., City [5]). The redundant encoding using size and greyscale enabled users to gather top-level information very fast. For example, with a glance at Figure 5.2, a user could understand that the two dimensions at the left (Sales and Profit) have been the main focus in prior work.

Interacting with Dimension View enabled users to discover co-mapping dependencies. To accomplish this task, the user could click on a dimension. The selected dimension's background colour changed to orange and any dimensions that had been mapped in a visualization along with this dimension became blue. Other cells remained unaffected. Colour-

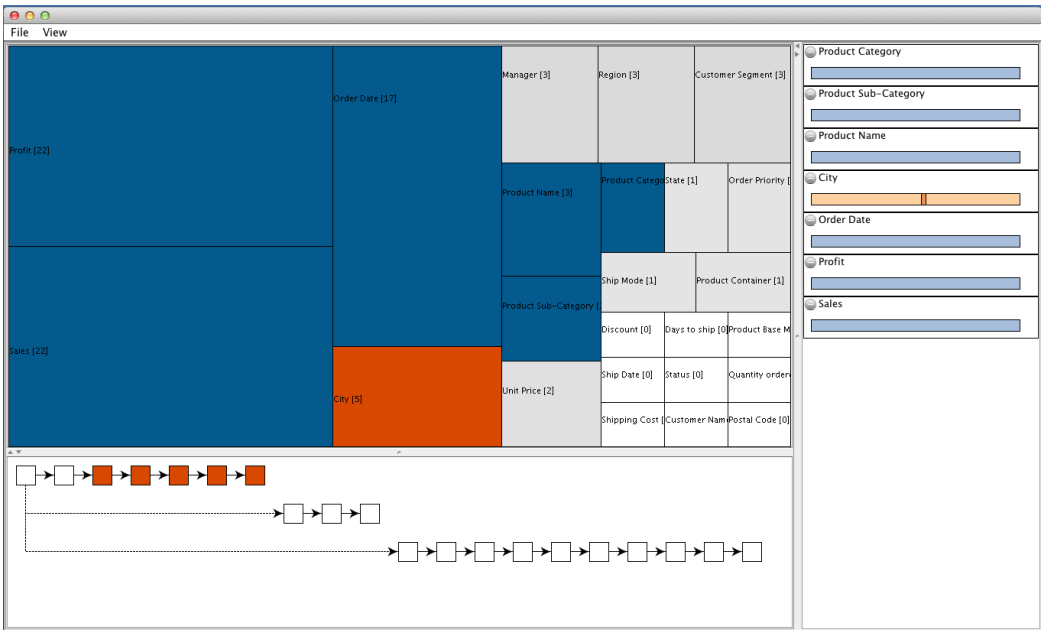


Figure 5.3: Revealing co-mapping information in Dimension View. Selecting City (background colour: orange) in the view showed that it had been considered with six other dimensions (background colour: blue), specifically Sales, Profit, Order Date, Product Name, Product Category and Product Sub-Category. City was not considered with Unit Price, Customer Segment, and several other dimensions.

coding assisted the user to immediately recognize related dimensions in the view (Figure 5.3). If required, the user could select multiple dimensions to investigate their co-mapping. This view was interlinked with the other two views: user interactions in Dimensions View propagated to Sequence and Data Views. These effects are explained in the respective sections below. The visual representation of Sequence View in Footprint-II was very similar to [68] and [50] which also used a similar graph structure for representing the analysis history.

5.2.2 Sequence View

Sequence view provided information regarding the temporal progression of the previous analysis. As shown in Figure 5.1C, a non-cyclic directed graph represented the analysis process. Each node in the graph represented a visualization (i.e., a chart) created by the analyst. Directed links depicted the progression of analysis over time from one visualization to the next. Each branch was indicative of a line of inquiry. In my design, reusing a previous state marked the beginning of a new line of inquiry and added a new branch to the

graph stemming from the revisited visualization. I changed the design of Sequence View from a linear list of thumbnail images of charts in Footprint-I to the graph design based on the feedback of visualization experts who reviewed Footprint-I. As part of their feedback, they had noted that the linear representation did not capture the branching which occurs in data exploration. Sequence View replaced List View (see 4.2.3) and Timeline View (see 4.2.2) in Footprint-I. In this view, temporal information was encoded using arrows that show the progression of analysis over time, and details about each chart were presented on demand by hovering the mouse cursor over a node.

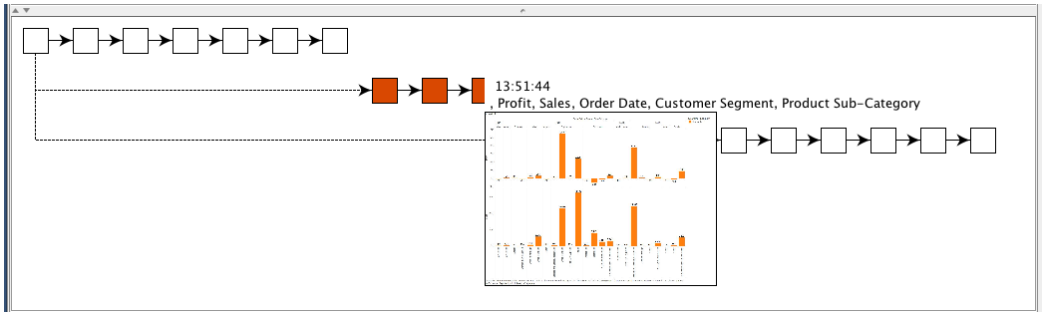


Figure 5.4: Hovering the mouse over a node in the Sequence View shows a thumbnail image of the visualization represented by the node and a list of dimensions included in that visualization.

This view supported a variety of user interactions. Hovering the mouse over a node in the graph would pop open a thumbnail view of the visualization represented by the node and information regarding the mapped dimensions (Figure 5.4). Double clicking on a node opened the full size chart in a separate window. The window came with a set of widgets that enabled a user to browse the entire history sequentially. Selecting one or more dimensions in the Dimension View triggered changes in the Sequence View. All the nodes that included the selected dimension(s) were highlighted in orange (Figure 5.4). This showed how investigation of the selected dimension(s) was distributed temporally over the analysis session. It also served as a visual search aid. By selecting a dimension, the user could easily find all the visualizations containing that dimension for closer inspection. Users could zoom in/out to change magnification of the view.

5.2.3 Data View

This view was added to Footprint-II based on the recommendations of visualization experts who reviewed Footprint. Data View provided information about the investigation of data

space. As shown in Figure 5.5, unique values in each dimension (e.g., all the unique city names under the dimension City), in ascending order from left to right, were represented by small segments within a bar. Each bar was labelled with the dimension's name (e.g., City). Darker shades indicated which data values were included in more charts within the previous analysis. A glance at this view would instantly guide user attention to darker regions, indicating which data values the previous analysis considered the most. This in turn could spark further investigation.

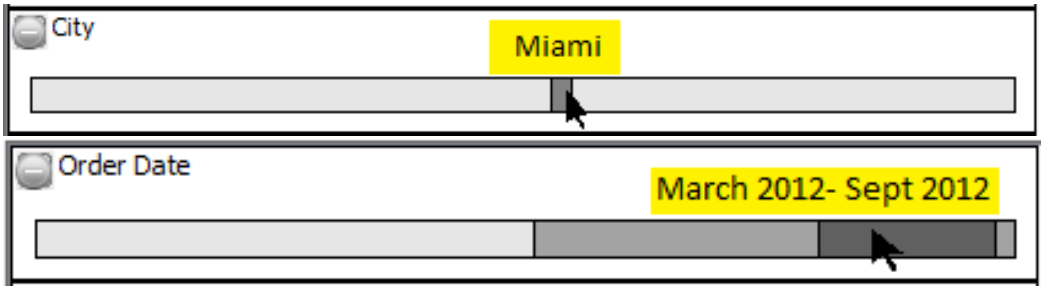


Figure 5.5: Two examples of mouse hover interaction in Data View. In the upper part, the user learns that the city Miami has been investigated more than other cities in the data set. In the bottom part, the user discovers that a specific date range (March to September 2012) has been investigated more than the earlier dates.

Hovering the mouse over a region showed the first and last values in the range within a tooltip (Figure 5.5). Double clicking on a region opened all the values in an external table.

5.3 Evaluation

I assessed the effects of providing analytic coverage information on collaboration. I hypothesized that providing this information would improve coordination by encouraging analysts to take an investigative path more divergent from the prior work, resulting in a better overall coverage of dimension space. In other words, I predicted that providing an analyst with dimension space coverage information that communicated what their previous collaborator did not do would result an analysis path more divergent from the prior work.

To test this hypothesis, I used a between-subjects design to compare Footprint-II to a baseline version containing only the Sequence View (Figure 5.4). The rationale behind this design of the baseline version was (1) to emulate current history tools, and (2) to control for design differences between tools, allowing me to directly assess the added value of a dimension centered perspective. The following subsections describe the study in more detail.

5.3.1 History Data

Participants continued a collaborator's analysis of a business data set consisting of sales data for four years from 2010 to 2013. It contained 25 data dimensions and 8400 records. This was the sample Superstore sales dataset provided by Tableau.

In order to prepare the history file containing the collaborator's work, I asked a business PhD student with considerable management background to analyze the same business data set using Tableau Public, a commercial visualization tool based on Polaris [70]. The task required him to explore sales data and try to find interesting patterns. I asked him to plan and carry out his analysis with two criteria in mind. First, he should intentionally neglect investigating some of the dimensions that could logically be investigated. For example, he should not investigate the possible relationships between returned goods and loss of profit though a rational aspect to consider. Second, some of his questions should remain at higher level of granularity to reserve potential for drilling down later. For example, if he observed an unexpected relationship between product packaging, type of delivery and cost of delivery, he should leave out looking into some of the package types. The rationale behind this design was to provide the next analyst (i.e. study participants) with equal opportunities of further investigating the prior work and / or exploring different dimensions. I also asked him to save all visualizations he created. He spent one hour on the task and created 27 charts.

For each chart, I manually extracted mapping and filtering information (i.e., dimensions mapped on the X and Y axes or other properties such as color and shape, and any filtering of dimensions that returned a subset of values) and the time of creation (from a video recording). This information was stored in a spreadsheet that we used as the analysis history file.

5.3.2 Participants

I recruited 20 business students as our participants (12 graduate, 8 senior undergraduate, 4 male, 16 female, average age of 25). I selected business students to ensure that our participants had necessary domain knowledge to investigate a finance-related problem. They were randomly assigned to use either Footprint-II or the baseline version. I only recruited only participants who reported having a strong understanding of business data analysis and a reasonable understanding and experience with tools that enable constructing statistical charts based on data (e.g., Microsoft Excel). None of the participants had participated in my previous user studies.

5.3.3 Physical Setup

We used a PC with two 21 inch monitors arranged side-by-side. An instance of Tableau Public software was open on one monitor and either Footprint-II or the baseline version of the prototype on the other monitor (Figure 5.6). The rationale behind this setting was to enable users to easily switch between the analysis task in Tableau and reviewing the history in the prototype. When asked at the end of the task, none of the participants found switching between two monitors distracting. Participants were also provided with pen and blank paper for taking notes.

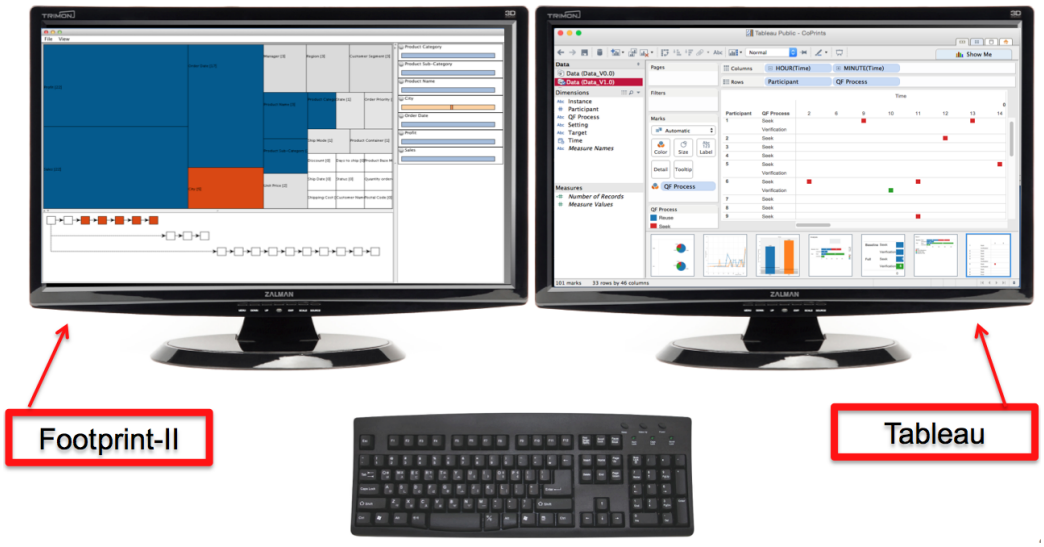


Figure 5.6: Physical setup of the study. Footprint-II is open on the left monitor and Tableau Public on the right monitor.

5.3.4 Procedure

At the beginning of each study session, we verbally described the task (5.3.5). The initial introduction was followed by introductions to the history prototype (either full or baseline version) and analysis tool (i.e. Tableau Public). Afterwards, participants practiced using the tools by doing short warm up tasks with the example history and data files. The warm up task for each system required working with all the main features of the system. An experimenter was present during the warm up session and participants could ask questions about the system, task and history file. A list of the supported user interactions and their outcomes for each system was left with the participant to be referred to (if needed) during

the actual review task. After the introductory part, participants were left alone and given one hour to perform the task.

5.3.5 Task

The analysis task required the participant to continue the exploratory analysis started by their collaborator. Following is the task description given to participants: "You are a business data analyst in a large international company. You are working collaboratively with other analysts in your company to explore sales data for the past 4 years and identify any possible strong and/or poor performance. Your collaborators are at different times/locations and work completed by others is passed around to be built upon. For your own analysis, you should explore the data and try to identify any interesting/unexpected patterns in the data with respect to business performance. In order to efficiently continue your collaborator's work, you first need to review and understand the prior work passed to you. This will also help you to keep the similar work minimized and investigate different plausible performance indicators. While doing your analysis, you can review the collaborator's work if required".

5.3.6 Data Capture

Participants were asked to think aloud. I video and audio recorded all the analysis sessions. In addition, we captured screen logs of both the history tool and the data analysis tool and logged user interactions with the history tool. Each analysis session was followed by a short interview. I also video recorded these interviews.

5.3.7 Data Analysis

In exploratory analysis, enabling collaborators to build an understanding of what work has been done will assist them in knowing where to allocate effort next [29], which implies better coordination [12] [20]. I hypothesized that this knowledge would lead to a better overall coverage of the dimension space between the participant and the collaborator. Therefore, I decided to measure the similarity between each participant's work and the initial analysis as an indicator of coordination. The more an analyst's work was different from the initial work, the greater cumulative coverage of problem space was achieved, which in turn indicated better coordination.

The data dimensions included in a chart gave strong clues as to the question being asked. Thus, if a chart created by the participant contained the same set of dimensions as a chart created by the initial analyst, it is likely they were investigating the same question(s). Likewise, charts containing some matching dimensions represented more similar investigative queries than charts contained completely different sets of dimensions. This observation formed the basis of my similarity analysis.

To compute similarity between a participant's analysis and the initial analysis, I first used videos and saved visualizations to identify all the unique questions that were asked by that participant. (I considered a question equivalent to a query that returns a subset of data, e.g. what is the relationship between Sales, Profit and Region?) Then using an alias assigned to each data dimension (e.g. Sales = A, Profit = B, Region = C), I converted each question into a set of letters (e.g. relationship between Sales, Profit and Region == {A, B, C}). Questions in the initial analysis were likewise transformed into sets of letters. Jaccard's Similarity Index (5.1) computes the similarity between two sets. We used a modified version of Jaccard's Index (Equation 5.1) to compute similarity scores (S) between each of the participant's questions and each of the initial analyst's questions.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{5.1}$$

My modifications to Jaccard's index took into consideration filtering of dimensions as well as exploring uninvestigated dimensions. My rationale for these changes were (1) investigating a completely new dimension was conceptually more different than investigating a filtered version of the same dimension and (2) investigating a dimension that nobody had considered before conceptually was more different than investigating new combinations of previously explored dimensions. When computing the sets intersection, if a dimension had different filtering in the two sets (e.g. City: [all cities] in set1 and City: [LA, NY] in set2), the intersection count was increased by 0.5 instead of 1.0. As a result, the similarity score decreased. After careful consideration, I also added a heuristic rule to give weights to dimensions. Dimensions investigated by the previous analyst carried a weight of 1.0 and previously uninvestigated dimensions carried a weight of 1.5. For example, when computing the union of two sets, each dimension was multiplied by its weight as follows: if set2={A, B, C} and set1={A, F, C} where F is the only previously uninvestigated dimension, then the union would be A:(1*1.0)+B:(1*1.0)+F(1*1.5) + C(1*1.0)= 4.5. The resulting score S is a value between 0.0 (no similarity) and 1.0 (identical).

Each question a participant asked should be compared to the most similar question

asked by the analyst. Therefore, for each question (i.e. a set) of each participant, I computed Jaccard's index in comparison to all questions (sets) of the original analyst; the maximum similarity found was assigned as that set's similarity score.

5.4 Results

I used individual questions as our unit of analysis. Figure 5.7 shows mean similarity score by condition. Since the data was normally distributed, I performed an independent-samples two-tail t-test to check whether there was a statistically significant difference between mean values of full version (I will refer to Footprint-II as full version in this section) (*mean*=0.33, *SD*=0.11) and baseline version (*mean*=0.58, *SD*=0.21) questions. The result ($t(339)=9.192$, $p<0.0091$) demonstrated that similarity scores for full version questions were significantly lower than for the baseline version.

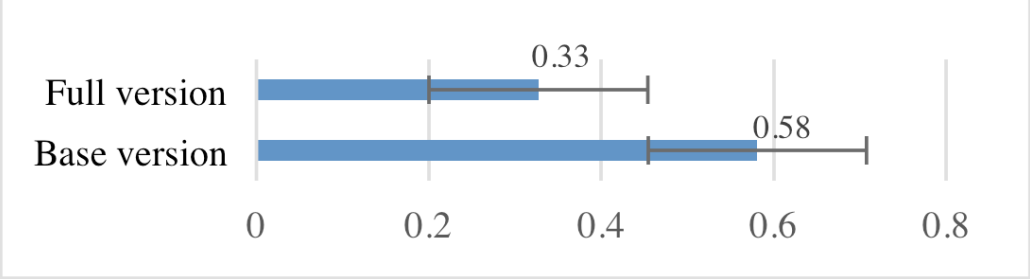


Figure 5.7: Average similarity scores for full and baseline version questions. Error bars show standard deviation. (1.0 = questions are identical to previous analysis; 0.0 = questions are completely different from previous analysis).

These results demonstrate that participants with access to Dimension View took analysis paths that were more divergent from the initial analysis than participants with only access to Sequence View. One contributor to this phenomenon was the number of unique uninvestigated dimensions considered by each participant. On average, full version users considered 19.6, and baseline version users considered 6.2 uninvestigated dimensions in their analyses. Results of a t-test ($t(18)=4.98$, $p<0.001$) showed a statistically significant difference between the groups. In addition, six out of ten full version users started their analysis by asking a question involving one or more of the uninvestigated dimensions. On the other hand, only one of the baseline version users did so. Interestingly, only full version users asked questions (total of 19) that were completely different from the questions asked in the initial analysis (i.e., $S=0.0$). Conversely, baseline version users asked more identical

questions (i.e., S=1.0): there were 20 of these for the baseline version and only 3 for the full version.

To summarize the quantitative results, full version users showed a greater tendency to focus on less explored aspects of the problem while baseline version users placed more effort on drilling in on previous questions. I argue that this was due to full version users' ability to more easily discover what had been focused on and what had been left out in the initial analysis. The Sequence View alone did not make it easy to acquire this information, corroborating the findings of the Footprint-I study (see Chapter 4). For baseline version users, gathering information about dimension coverage required multiple passes through items in the Sequence View, which presumably added cognitive costs.

In addition, baseline version users relied on external memory aids such as paper notes for recording their discoveries. There was a substantial difference between the number of notes taken by full version (total of three notes by three participants) and baseline version users (total of nine notes by nine participants).

Though this might be a result of personal preference and work style, closer inspection of notes taken by baseline version users revealed five instances of explicitly recording information about dimension space, similar to what was available in the Dimension View for full version users. These five baseline version users manually extracted dimension coverage information by tracing the history and recording different examined combinations; they recorded this information in their notes for later use. Figure 5.8 shows three examples of such notes.

I also reviewed captured videos of full version users to understand their Dimension View usage. Eight out of ten participants started their initial review through Dimension View (note that they were not instructed to do so and could use any view at any time). Statements made by participants suggested that they found Dimension View useful and intuitive for gaining an overview of the previous analysis. Following are a few examples from users' alouds that show how Dimension View helped to inform their understanding of prior work: "...[the initial analyst] didn't look at Returns...maybe I should look into this", "it seems that he focused greatly on Profit, Sales and Order Date...". During the analysis, users mostly referred to Dimension View to refresh their minds and avoid duplicating work. They also used Dimension View as a visual search aid. Selecting a dimension (or a combination) helped them to easily filter Sequence View. Subsequently they would look at the thumbnail view or open the full size view of the chart.

While Dimension View was clearly important, my findings showed that Data View was not used as much. Only half of the full version users (5 out of 10) referred to Data

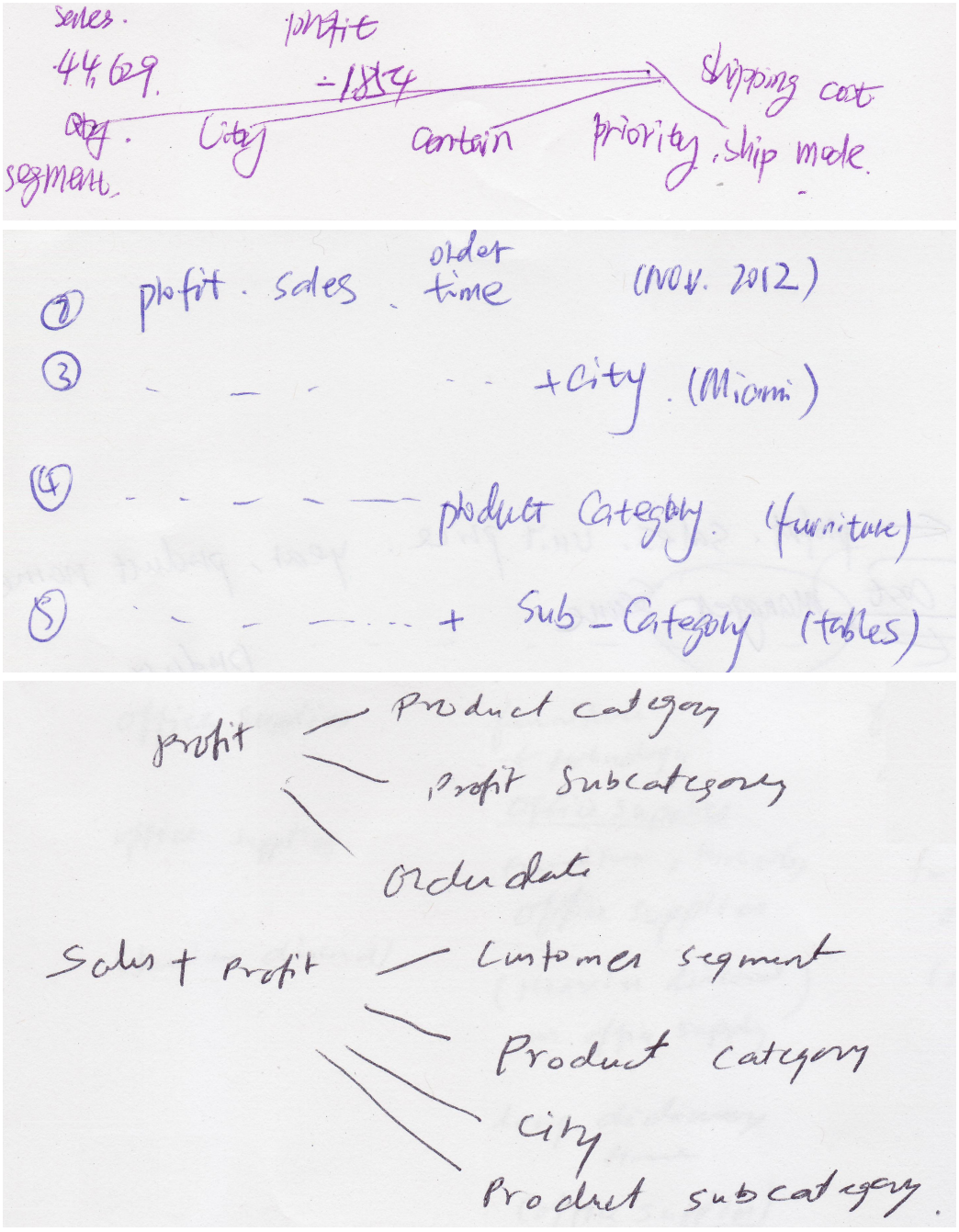


Figure 5.8: Three examples of how baseline version users used notes to record discovered coverage information. These are excerpts of notes taken. The bottom note shows that a participant recorded that "Profit" was considered with "Product Category", "Product Subcategory", and "Order Date". Later she noted that Sales+Profit was considered with "Customer Segment", "Product Category", "City", and "Product Subcategory".

View (total of 22 interactions, Avg = 2.2, SD=3.19). On the other hand, all of them used Dimension View (total of 78 interactions, Avg=8.9, SD = 2.6). Participants mainly used Data View during their initial analysis to gain an understanding of the focus of prior work in the data space. For example, after opening the darkest region in the Customer Segment in Data View to see the data points in an external table, one participant said "I noticed that in the customer segment, he [the initial analyst] was more focused on corporate than the others". Other participants also noticed this trend.

5.5 Discussion

My findings from the Footprint-II user study clearly demonstrated that collaborative visual data analysis can benefit from the addition of dimension coverage information. I found that full version users showed better coordination with the previous analyst through their focus on uninvestigated aspects of the problem. The similarity analysis showed that full version users asked questions that were more different from the ones asked by the initial analyst than baseline version users. They investigated the problem from new angles, meaning that overall there was a more comprehensive investigation of the problem. For example, the initial analyst did not investigate the 'Days to Ship' dimension (i.e., days from receiving to shipping of an order). Yet, inefficient order processing times could be responsible for overhead costs and loss of Profit. Six full version users, in contrast to only two baseline version users, examined this possibility. Based on my observations, I attribute the better coordination shown by full version users primarily to the presence of Dimension View: full version users reported that Dimension View helped them to easily and accurately identify underexplored aspects of the dataset. This knowledge in turn influenced the questions that they asked.

On the other hand, findings showed greater overlaps between the analysis of baseline version users and the prior analysis. This overlap represented duplicated work and a reduced overall coverage of the problem space, suggesting less well coordinated collaborative work. These users showed a greater inclination towards continuing the prior work and drilling down on questions. This was most likely due to the affordances of Sequence View. At the surface level, this view contained visualizations that each represent a question. Therefore, the immediate messages conveyed by this representation was questions. However, gaining an understanding of dimension space coverage required iteratively reviewing these questions (or manually generating notes) to build a mental map of which dimensions had been covered. This was a rather costly and cumbersome process. Therefore, as antic-

ipated by the principle of least cognitive effort [58], most baseline version users preferred the less costly action of 'picking a question' and drilling down on it rather than identifying the unexplored aspects of the problem and asking new questions. The description of task explicitly asked the participants to "keep the similar work minimized and investigate different plausible performance indicators". Although this may have influenced that participants natural behaviour and discouraged work similar to prior analysis, both Footprint-II and baseline users used the same task description, thus influenced similarly.

I made an assumption in this work that better overall coverage of the dimension space is an important aspect of good coordination. This seems reasonable in an exploratory analysis situation where the primary focus was finding as many trends and outliers as possible. I made this assumption based on prior research [20] [12] [29] in the asynchronous collaborative analysis field that suggested an awareness of "what has been covered" can direct current work towards "what has been left out". I believe my findings justified this assumption, in that participants reported that Dimension View made it easier for them to understand the prior work, particularly the coverage of dimension space. At the same time, I do not claim that in all exploratory analysis situations broad coverage of all possible questions (breadth of analysis) is the foremost goal.There may be situations in which drilling deeper on previous analysis (depth of analysis) is preferred or necessary. Nonetheless, I speculate that even drilling in on the prior work might benefit from Dimension View. While drilling in (i.e. keeping some dimensions the same while changing others and/or filtering), analysts can exploit Dimension View to discover what new combinations remain.

I posited that similarity between analyses could be used as a metric for quantifying group coordination. Based on prior research [73], good group coordination, in part, relies on the division of the task, and ideally, group members try to avoid duplication of efforts unless necessary. Therefore, I made the assumption that in an exploratory collaborative VA situation, overlap between participants' analyses can be used as a metric for assessing coordination. Other researchers have used different metrics for measuring coordination. In the context of synchronous collocated collaborative VA, Mahyar and Tory [53] use discourse analysis to measure coordination. In [71], Isenberg et al. investigate coordination of group interactions with shared visualizations. Yet their evaluation is limited to two use-case scenarios and do not include any metrics for quantifying coordination. Currently, there is a lack of standardized metrics and techniques for quantitatively assess coordination for collaborative exploratory VA. Future work is required to test and develop metrics for measuring group coordination.

Findings showed that there was a substantial difference between the total number of

references to Dimension View (69 times) and Data View (10 times) by the participants . Although I cannot truly isolate the value of Dimension View because it was in the same condition as Data View, it seems that participants found Dimension View to be much more useful. None of the full version users started their initial review of prior work by interacting with Data View. I believe the prominence of Dimension View relates to the specific task of our study, which required participants to find as many trends and outliers as possible. In this sort of exploration, gaining an understanding of "what has been investigated" in the dimension space comes before the same understanding of the data space (i.e. data values). Yet there might be other exploratory analysis situations where the reverse is true. In [76], the exploratory analysis task involved investigating a constant set of dimensions by manipulating the filtering of values. In such a case, being able to visually understand which data values were explored versus left out would be most valuable.

5.6 Conclusion

In this chapter, I investigated RQ4: How does dimension space coverage information influence task coordination? Based on the findings of a user study, I showed that in an exploratory collaborative VA context, representing analysis history from the angle of dimension space coverage can improve implicit division of exploratory data analysis task. Providing Dimension View assisted participants who used Footprint-II to better identify and focus on parts of dimension space that was not investigated in the initial analysis. The main contribution of this chapter is described in Chapter 1 as C5.

Up to this point of my research, my focus has been on supporting collaborative aspects of exploratory VA. After positively demonstrating that providing coverage information could improve collaboration, I shifted my focus to exploratory VA more generally. In particular, I asked the next question: **(RQ5) How does providing live information about dimension space coverage influence EDA?** In the next Chapter, I will describe the effects of providing dimension space coverage on ongoing analysis.

Chapter 6

Supporting Exploratory Data Analysis via Scented Widgets for Dimension Space Coverage

In Chapter 5, I demonstrated that in a collaborative EDA context, providing a dimension-coverage oriented representation of analysis history improved group coordination. In this chapter, I will investigate **RQ5: How does providing live information about dimension space coverage influence EDA?**

After investigating the effects of dimension space coverage views on collaboration, I shifted my focus to the exploratory aspect of collaborative EDA. In particular, I investigated if and how live presentation of dimension space coverage would affect the exploratory analysis. Based on my findings thus far, I hypothesized that providing live coverage information would increase the number of questions asked, increase the breadth of exploration, and increase the number of findings.

The rationale behind this speculation was based on my previous findings that Dimension View could increase awareness of "what was done" and assist in determining "what was left to do". Consequently, I posited that providing such awareness on the fly would expand the breadth of analysis by helping an analyst to quickly discover and focus on uninvestigated aspects of analysis. A greater coverage of problem space could in turn increase the number of findings. In this chapter, I address the design of a VA tool that uses the scented widgets [76] technique to represent dimension space coverage information, its evaluation through a user study, and the results. Unlike Footprint-I and Footprint-II, dimension coverage information was represented within a data analysis tool. Up to this

point, my main focus was examining if and how dimension-centric history better supported understanding depth and breadth of prior analysis than the linear history. After positive findings of Footprint I and II user studies, I used scented widgets to incorporate coverage information into the GUI elements of a prototype visual data analysis tool.

The main contribution of this study is the finding that providing a dimension space coverage view increases the number of top-level findings in exploratory visual data analysis. I also showed that providing this information can expand the breadth of exploratory analysis. Unlike my previous studies, this study was done in a single user setting. The main reason for this decision was to investigate the effects of visualizing dimension space coverage information on analysis outcomes in isolation from any other factor, such as collaboration, that could have possibly influenced the results. As part of my future work, I am planning to repeat the study in a collaborative setting.

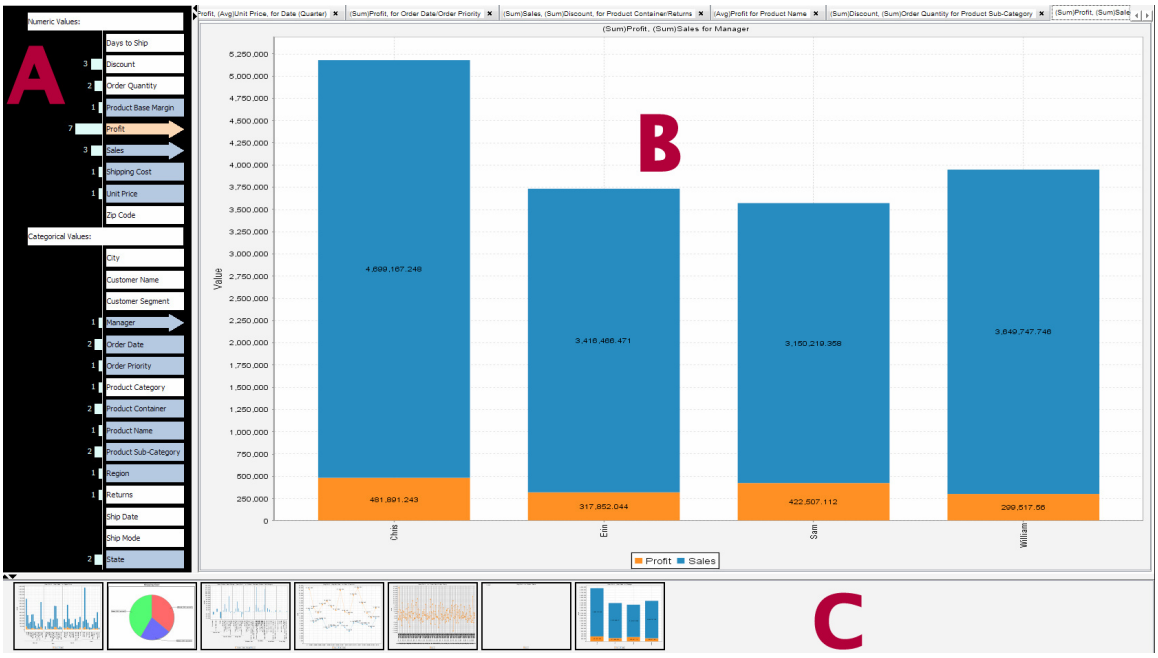


Figure 6.1: Visual analysis prototype that reveals dimension space coverage information. (A) Scented View reveals the dimension space coverage information using scented widgets. Bar charts to the left of each dimension name indicate how frequently each has been investigated. Co-investigated dimensions are revealed through colouring (blue) when one or more target dimensions are selected (orange). Arrows indicate dimensions included in the currently displayed chart. (B) Visualization panel. (C) Sequence View shows a chronologically-ordered list of created charts.

6.1 Introduction

In EDA, an analyst constantly formulates and evaluates new questions or hypotheses about data. However, selecting a data subset to explore can be quite difficult [5, 43]. In fact, Lam identified *deciding-what-to-explore-next* as one of three key data analysis challenges [47]. With current tools, analysts typically rely on memory to recall what questions they have asked and what they still need to do. However, factors such as limited short term memory and the recency effect (i.e. remembering recent items more clearly than those further in the past) [27] can impede recall. In other words, it can be difficult to maintain an awareness of dimension space coverage. In addition, unfamiliarity with the shape and structure of the data [5, 80], vague analysis goals [80], and insufficient domain or visualization knowledge [80] can hinder question formation, potentially leading to under-exploration of the problem. Moreover, Wongsuphasawat et al. [80] argue that typical interfaces for constructing visualizations (i.e. manually mapping and filtering dimensions, making data transformations, and selecting visual encodings) may encourage premature fixation on specific questions, promoting depth-first exploration at the expense of breadth. How then, can we help analysts to formulate questions and encourage them to go both broad and deep?

In the case of tabular data, an analytic question is comprised of a combination of data dimensions and can be reasonably characterized by that dimension set. For example, a business analyst might start by asking "what is the relationship between *Profit* and *State*?" and next she may filter state to California and examine, "what is the relationship between *Profit*, *State:California* and *Products*?" Throughout her analysis, she constantly formulates and evaluates questions, each containing a combination of dimensions. This suggests that revealing dimension space coverage information might help analysts recall what questions they have asked, and (perhaps more importantly), identify questions that have *not yet* been asked. To explore this idea, I extended scented widgets [76], a technique for embedding navigational cues into GUI widgets (see Figure 2.6). Specifically, I incorporated dimension space coverage information directly into the interface elements of an exploratory analysis prototype (Figure 6.1). The theoretical support behind this method stemmed from the notion of information scent (i.e. attention pointers that assist a person in navigating an information space) introduced by Pirolli and Card [63]. This approach enabled analysts to maintain an up-to-the-moment understanding of what data dimensions they have investigated and in what combinations.

Dimension coverage information is captured at the system level by visualization history modules that track and record visualization states (e.g., [30]). However, most vi-

sual representations of history provide very limited support for understanding dimension space coverage because they focus instead on representing past states and/or actions. Previously, I introduced alternative visualizations of history to explicitly reveal dimension space coverage information (see Chapters 4 and 5). In Chapter 5, I demonstrated that this perspective could improve asynchronous collaboration, where one analyst does some work and then "hands-off" the work to a collaborator who continues the analysis. Providing analysts with information about which dimensions were previously investigated by their colleague reduced the duplication of work. This finding suggested that dimension space coverage information might facilitate the flow of analysis in non-collaborative situations as well. Therefore, I investigated the effects of live dimension space coverage information on single-user EDA. I also explored how this information could be integrated within a visual data analysis tool in a subtle way; my previous stand-alone representations were space inefficient and were not integrated with an analysis tool. I hypothesized that visualizing the coverage of dimension space would:

H1: increase the number of formulated questions,

H2: increase the number of findings, and

H3: increase the breadth of exploration without sacrificing depth.

To evaluate my hypotheses, I built a prototype visual analysis tool (Figure 6.1) that used scented widgets to visualize dimension space coverage. I then conducted a between-subjects user study to compare my tool to a baseline version without the dimension space coverage information.

6.2 Incorporating Dimension Space Coverage Information into Visual History

I integrated dimension space coverage information directly into a visual data analysis tool by using scented widgets (for background on Scented Widgets, see Chapter 2, section 2.5). In [76], Willett et al. introduced guidelines for designing scented widgets. As one example, they used scented widgets to provide social navigation cues. They re-implemented HomeFinder [78], a map-based housing search tool. Information about prior house searches were embedded in the dynamic query widgets to help people better understand which data values had been investigated by other users (Figure 2.6).

This work has three main differences from my research: type of exploratory task, visualized information and context. In their case, the exploration only required filtering of data values. Each question about data included a fixed set of dimensions (Area, Rent, Number of Bedrooms etc.) and users only manipulated filtering. On the other hand, I consider exploratory analysis tasks that require investigating varying combinations of dimensions. For example, a business analyst might explore many performance indicators (e.g., Profit, Sales, Return on Investment etc.) in relation to other attributes. To support this type of analysis, I reveal *dimension* coverage information (embedded in dimension name widgets) rather than *data value* coverage (embedded in value name widgets), and extend scented widgets to capture *co-investigation* information (i.e. which dimensions were considered in combination). Finally, in my case, the context of analysis is single professional users as compared to online collaborative social data analysis. Most often, the main goal of the latter is to enjoy while the former is to discover and present [59]. More importantly, my research investigates the value and effect of seeing traces of one's own past analysis, rather than trails of investigation left by other unknown people.

In order to investigate how dimension space coverage information would influence exploratory analysis, I designed and implemented a history module that provided three distinctive representations of analysis history. The most important of these was *Scented View*, which used scented widgets to reveal dimension space coverage information. I also included two complementary history views: a data values coverage view (*Data View*) and a traditional linear list of past states (*Sequence View*). These views were embedded within a prototype tool for visual data analysis (Figure 6.1).

6.2.1 Scented View

Scented View reveals dimension space coverage information using scented widgets (Figure 6.1A). Previously, I introduced two visual designs for representing dimension space coverage information, a radial design (Chapter 4.1) and a Treemap design (Chapter 5). Both designs were implemented as standalone prototypes. They were space inefficient and not ideal for being directly integrated into a visual data analysis tool; therefore, a new design was necessary. Embedding information that is logically relevant to a GUI element (such as a textbox showing a dimension's name) makes information easy to discover and readily available in a more space efficient manner. This was my main rationale behind using a scented widget approach.

Following is a short usage scenario that motivates my solution: Sue is a business data

analyst who is exploring a sales data set for the first time. After loading the data, she examines the list of dimension names and decides to explore Profit. She first examines the profit trend over the past 4 years and notices a large drop in 2012. Based on this finding, Sue decides to investigate the relationship between Profit and Region in 2012. She notices that the company lost considerable profit in the west. Next, she examines Product Types in conjunction with Profit and Region (filtered to west and 2012). She continues down this analysis path for another 30 minutes. At this point, she wants to ensure she has considered all the factors that might have contributed to profit performance. She clicks on the Profit bar in Scented View (Figure 6.1A) and instantly discovers she has not yet examined Profit in relation to Ship Mode and Returns, two variables that could potentially impact profit. As her exploration proceeds, she continues to use Scented View to help her recall what has been done and what is left to do; in this way, Scented View assists her to devise new and meaningful questions.

I designed Scented View to support two primary tasks: 1) understanding which dimensions have been investigated versus which have been left out, and 2) understanding which combinations of dimensions have been examined (i.e. co-investigation relationships). I also aimed to support the secondary task of understanding frequencies of use. For instance, Figure 6.1A shows a greater focus on Profit (investigated in 7 charts) than Sales (investigated in 3 charts). From a visualization design perspective, the first primary task required distinguishing between investigated and uninvestigated dimensions (a categorical concept). Since it is not advisable to alter shape and/or spatial position of interface elements (a rectangular textbox cannot suddenly become triangular or move to some other part of the GUI), these two channels could not be used to encode the information. Therefore, similar to [76], for every investigated dimension, I placed a bar to the left of the textbox that contained the dimension's name. The length of the bar encoded the magnitude of investigation; this used position encoding, the most powerful visual encoding channel [16, 59], to encode the magnitude of investigation.

The second task required discovery of relationships between dimensions (i.e. dimensions investigated together within the same chart). Similar to the first task, and because of the limitations imposed by working with GUI elements, containment, grouping and proximity could not be used for encoding relationships. I considered drawing lines between the labels to visualize connections (e.g. lines traveling from sales to Profit, State and City to show co-investigation) but this design would add clutter and visual obstruction to the dimension panel. Therefore I opted to use colour hue (blue and orange) to encode dependencies and rely on interaction to reveal this information on demand. Figure 6.1A shows

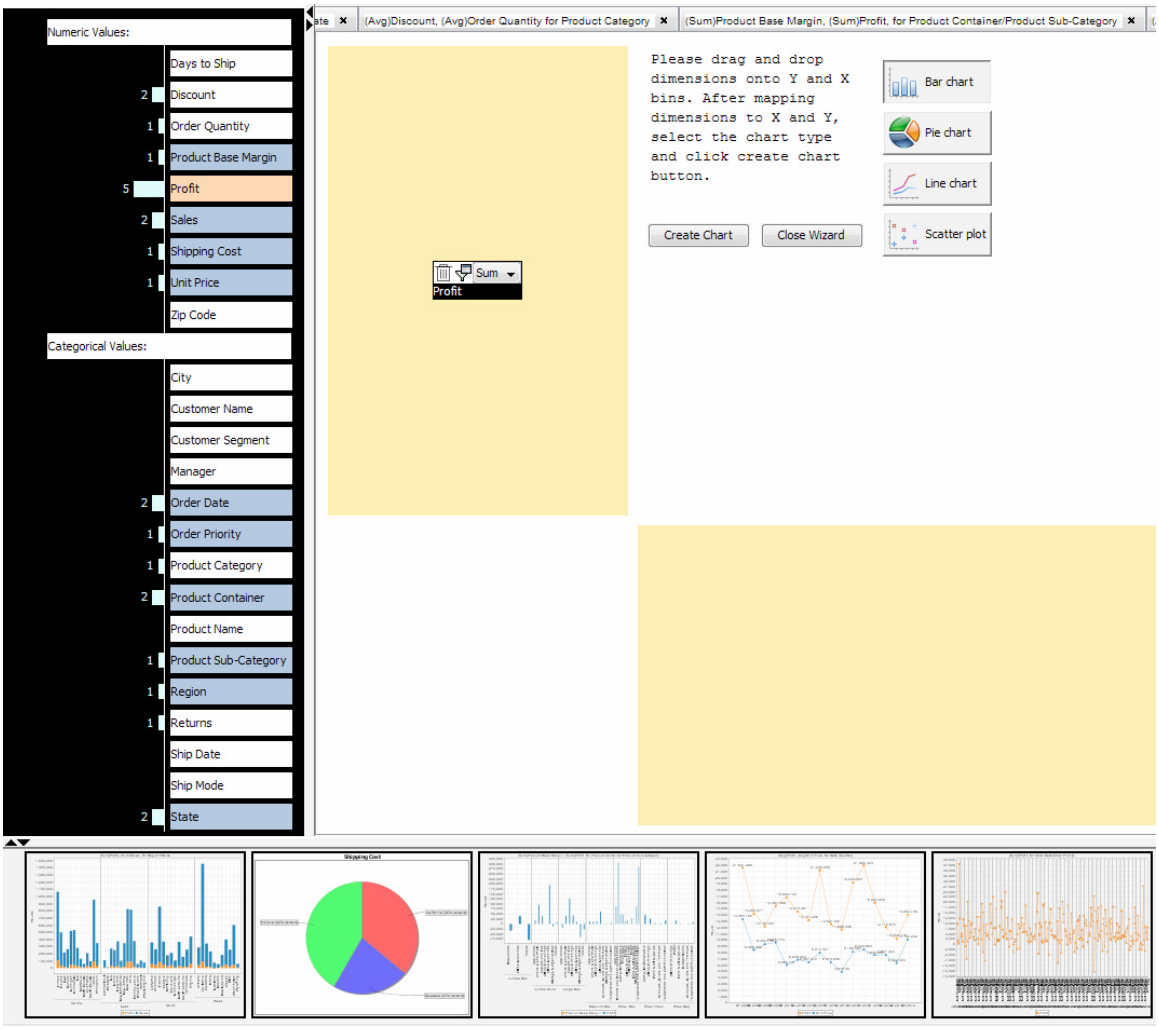


Figure 6.2: Automatic presentation of co-investigation information at the time of creating a new chart. As the user drags and drops dimensions into X or Y bins, the view changes to show prior co-investigation of the dimensions. In this example, as soon as the user dropped Profit in the Y bin, the colours changed to highlight dimensions that were previously co-investigated with Profit.

how this information was conveyed. When the user selected a dimension (by clicking on the name label), the background of the dimension textbox changed to orange and the background of any other co-investigated dimension(s) became blue. A user could select more than one dimension from the list to investigate higher-order relationships. While this visual encoding choice needed to be learned, it had several advantages: 1) the selected item(s) were clearly indicated by orange colour, making it easy to add to or change the selection, 2) selected items (orange) and their co-investigated items (blue) clearly stood out from the list, and 3) the encoding only minimally changed the standard appearance of interface widgets. This interaction approach and colour scheme were also easily learned and understood by participants in my earlier user study (albeit within a very different visual design. see Chapter 5).

Dimension coverage information for individual dimensions was constantly present in the user interface. Co-investigation information, however, was not continuously present; it became available in two different ways. One way to attain this information was by selecting a dimension (clicking its name in the list). This approach required the user's explicit interaction with the view. For example, in Figure 6.1A, the user has selected Profit; all dimensions ever considered with Profit have changed to blue. In the second approach, the system automatically represented co-investigation information while the user actively created a new chart. For example, in Figure 6.2, the user has just mapped Profit to the Y axis; at this point, ten other dimension widgets turn blue to remind the user about which other dimensions have been previously co-investigated with Profit. The rationale behind this design was to support two critical use cases involving co-investigation information. In the first use case, the user intentionally paused analysis to review and recall co-investigation information. In the second use case, automatic presentation of this information could help the user avoid duplicating earlier work and assist them to formulate novel questions on the fly. Selecting dimensions by either approach (direct selection via mouse click or by the system during chart creation) also filtered the content in Sequence View (discussed below). Filters were removed after the user deselected dimension(s) or after a new chart was created.

6.2.2 Sequence View

Sequence View showed past visualization states in a linear list format (Figure 6.1C). This view mirrored the linear representation approach to visualization history (e.g., [30,52,81]). In this approach, thumbnail images of past charts are ordered chronologically, labeled with

information such as chart name. My aim with Sequence View was to help users to quickly review and reuse past states. I included this view as a representative of typical history designs, because I felt it was complementary to my new views, and to enable a comparison to an appropriate baseline in my study. Thumbnail image size was proportional to the height of the Sequence View panel. If desired, a user could maximize the Sequence View panel to the entire width and height of the window to browse and compare large images side by side. Hovering the mouse over a thumbnail image showed a tooltip with detail information about the chart (e.g., (SUM)Sales, (SUM)Profit for Region, City, Product Category).

I changed the design of Sequence View from a acyclic-graph in Footprint-II to a linear list of thumbnails. This was mainly due to feedback from the Footprint-II evaluation: participants reported that they were more interested in seeing an overview of visualization states rather than the branching of exploration.

6.2.3 Data View

Data View represented the prior coverage of data values for each dimension (Figure 6.3). Although my primary focus was to reveal information about the coverage of dimensions, I added this view based on the recommendation of several visualization experts who gave feedback on Footprint-I. This information was only available on demand by hovering the mouse over a dimension's label on the list for 2 seconds. Hovering the mouse popped open a tooltip-like pane that contained a visualization of the data value coverage for that dimension. The visual encoding depended on the dimension type (quantitative versus ordinal/nominal). I used a tag cloud for ordinal and nominal values where font size encoded the frequency of prior investigation. Figure 6.3 (bottom) shows that for Product Categories, the analyst had focused more on "Binders and Binder Accessories" than on "Appliances". For quantitative dimensions (e.g. Sales), I used a bar to encode values within the dimension and colour saturation to convey information about the magnitude of investigation. Darker shades indicated data values that were included in more charts (i.e. other values were filtered out). For example, in Figure 6.3 (top), all values of the Profit dimension are represented in the bar, ordered smallest to largest from left to right. In this specific example, the analyst focused most on values ranging from approximately 43 to 441.

I designed Data View to be on-demand for two main reasons. First, based on the findings of the Footprint-II user study, participants mainly used Dimension View even though both dimension and data coverage information were constantly present (5.5). Second, the GUI elements related to data were already used to represent dimension coverage in-

formation. Embedding data coverage information into the same elements would result in clutter and interference. Therefore, I decided to give priority to using scented widgets for dimension coverage rather than data coverage.

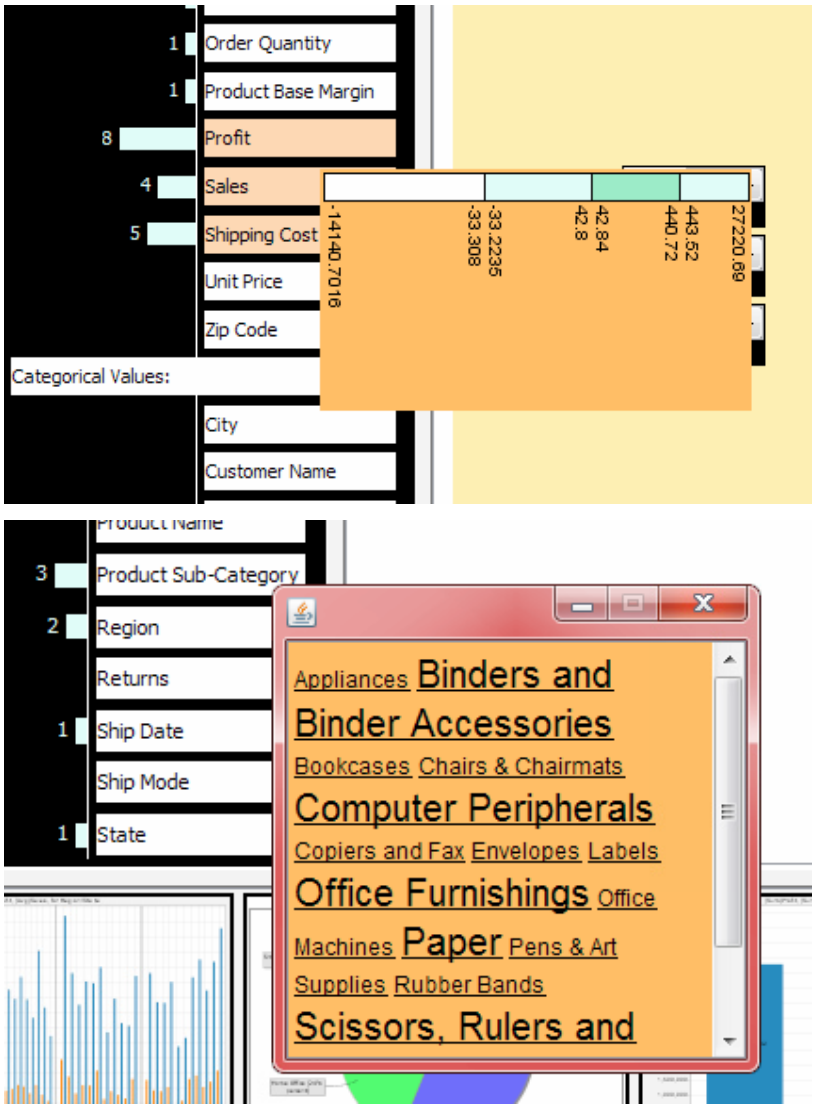


Figure 6.3: Two examples of Data View, showing pop-ups after hovering the mouse on Profit (top) and Product Sub-Category (bottom). In the top, the range of [42.84, 440.72] has been investigated more than the other values. In the bottom, some Product Sub-Categories have been investigated more than others.

6.3 Prototype Implementation

I implemented my three history views within a visual data analysis prototype that enabled a user to create a variety of basic statistical charts from tabular data. The key feature of my prototype was the use of scented widgets for live presentation of dimension space coverage within interface elements. The prototype was built in JAVA and used the JFreechart [2] API for chart creation. While the available chart types were not elaborate, and the tool's functionality was less comprehensive than commercial tools, the prototype provided a sufficient analysis environment in which to investigate the idea of embedded dimension space coverage information.

Figure 6.1 shows a screen shot of the prototype. After connecting to the data source, a list of data dimensions was created and placed in Scented View (Figure 6.1A). Dimensions were divided into categorical and numeric. Dimension names were ordered alphabetically. My chart creation design was loosely based on the shelf approach in Polaris [69] and Tableau [3]: the user could select dimensions from the list and drag and drop them into vertical and horizontal shelves in the chart pane (Figure 6.2) that respectively represented X and Y axes. After a dimension was mapped to an axis, if required, the user could filter values and apply simple statistical operations such as sum, average and standard deviation. Next, the user would select a chart type amongst the available options: bar, stacked bar, line, pie or scatter. To enable analysts to create complicated charts, they could map multiple dimensions on each axis. Each new chart was placed in a new tab in the charts pane (Figure 6.1B). A user could switch between the charts by clicking on the respective tab at the top of the panel. Each tab contained the title of the chart (e.g. [Avg.] Profit, [Avg.] Sales, [Avg.] Shipping Cost, City) to help identify it. Selecting a tab caused the shape of the dimensions that were involved in the corresponding chart to change to arrows (Figure 6.1A). The rationale behind this design was to visually assist a user to quickly identify the dimensions plotted in the currently selected chart.

6.4 Evaluation - Method

I conducted a controlled between subjects laboratory experiment to evaluate how access to dimension space coverage information would influence the analysis process and its outcomes. Specifically, I tested the three hypotheses described in the introduction, namely that dimension space coverage information would cause participants to ask more questions (H1), produce more findings (H2), and increase the breadth of their analysis without

sacrificing depth (H3). See appendix D for more details about the study materials. I compared the full prototype described in 6.3 to a baseline version that was identical in design and functionality except that dimension and data value coverage information was removed. Specifically, in the baseline version 1) there were no bars next to investigated dimensions' labels to show the frequency of investigation, 2) interaction with the list of dimensions provided no insight into the co-investigation of dimensions through colour-coding, and 3) Data View was removed. Similar to the full version, baseline version users could filter Sequence View by selecting dimension(s) from the dimension list. The background of the selected dimension(s) became orange but the co-investigated dimensions (if any) remained unchanged. Baseline version users could review history sequentially (by looking at visualizations one by one) or selectively (by filtering to show only visualizations with certain dimensions). I chose this experimental design, comparing the identical tool with features enabled versus removed, so that I could conclude that any difference in performance was caused by the additional dimension space coverage information.

6.4.1 Participants

I recruited 20 business students (12 graduate, 8 senior undergraduate, 4 male, 16 female, average age of 25). I selected business students to ensure that participants had the necessary domain knowledge to investigate a finance-related problem. I only recruited participants who reported having an understanding of business data analysis and experience with creating different types of statistical charts such as bar, line and scatter plots. To minimize the effects of gender on the outcomes (considering the much larger population of female participants), I randomly assigned eight female participants and two male participants to each condition (full or baseline). None of the participants had participated in my previous user studies.

6.4.2 Procedure

I began with an introduction (approximately 15 minutes) to the task, data and tool (either full or baseline version). After the introduction, participants practiced (approximately 30 minutes) using the tools by doing short warm up tasks with an example sales dataset different from the dataset used for the actual task. The practice task required working with all the main features of the system. In particular, participants practiced how to create charts by dragging and dropping dimensions onto X & Y shelves, how to filter data and perform statistical operations, how to use Scented and Data Views to obtain coverage information

(full version only), and how to review and reuse work history using Sequence View. An experimenter was present during the practice session and participants could ask questions about the software, task and history file. A list of the supported user interactions and their outcomes was left with the participant to be referred to (if needed) during the actual analysis task. After the introductory part, participants were left alone and given one hour to perform the main task. Pen and paper were provided in case participants wanted to take notes. Participants were told that there were no constraints on what they could record in their notes. The analysis session was followed by a short interview and a questionnaire.

6.4.3 Task

The open-ended exploratory analysis task required participants to evaluate the business performance of an online retailer using a sales dataset and identify any positive or negative performance indicators. The task was based on typical SWOT (Strength, Weakness, Opportunity and Threats) analysis tasks that help large organizations to evaluate their business venture. All of the participants reported that they were familiar with this type of analysis, as it is taught in business and commerce programs. Following are the instructions given to participants: "You are a business data analyst in a large online retailer. You should explore the performance data for the past 4 years and identify trends/outliers in the data that are indicative of strong and/or poor performance." I used the Superstore sales dataset provided by Tableau Public. This dataset contains sales information for four years and has 24 data dimensions and 8400 records. No specific directions or restrictions were imposed to influence or direct the participant's focus. Each participant was given 60 minutes for the task (exclusive of introduction and practice time).

6.4.4 Data Capture

Participants were asked to think aloud. I video and audio recorded all the sessions and interviews. The video camera was pointed at the screen to capture the screen contents as well as user actions that could not be logged by the system (e.g., tracing the dimension list with one's finger). Each participant's analysis work (i.e., charts created) were recorded by the system. Both base and full versions of the prototype automatically logged user interactions with the tool (e.g., selecting a dimension or reloading a chart). In addition, full version users gave their assessment of the scented widgets by answering a Likert style questionnaire.

6.5 Evaluation - Data Analysis and Findings

The following subsections describe the data analysis and findings related to each hypothesis, followed by interview and questionnaire results. I used a combination of both qualitative and quantitative techniques in order to best assess each hypothesis.

6.5.1 H1: Effect on the Number of Questions Asked

H1 speculated that providing dimension space coverage information would increase the number of questions asked. To evaluate this hypothesis, I first identified and categorized instances of questions through multi-pass open video coding. I qualitatively analyzed transcribed alouds to identify and count the number of questions asked by each participant.

Following Liu and Heer [49], I considered a question as "an indication of desire to examine certain aspects of the data." A question did not need to end with a question mark. Following is an example question from the study: "I want to look at Sales and Profit for Product Categories". I only considered utterances that I could confidently identify as a question and did not take into consideration vague and incomplete ones such as "...now I'm gonna look at [didn't continue]". Although very rare in both conditions, I also excluded questions that were logically invalid. For instance, one participant investigated the relationship between Sales and Container Type (types of containers used for shipping items to customers, e.g. large box): logically, the shipping container should have no impact on Sales. On the other hand, I did consider valid questions that were inconclusive and did not result in a particular finding. This decision was made based on the fact that a relevant question about data can and should be asked even if it does not yield any results. Each identified question was timestamped for future fast retrieval.

Table 6.1 shows the number of valid questions for each condition and Figure 6.4 shows the distribution of questions per condition. The result of a two-tail independent t-test ($p < 0.0001$, $t = 31.623$, $df = 9$) showed that the full prototype group asked more valid questions on average.

Table 6.1: Total number of valid questions for each condition.

Condition	Count of valid questions	Average	StdDv
Baseline	94	9.7	3.3
Full	187	18.7	6.6

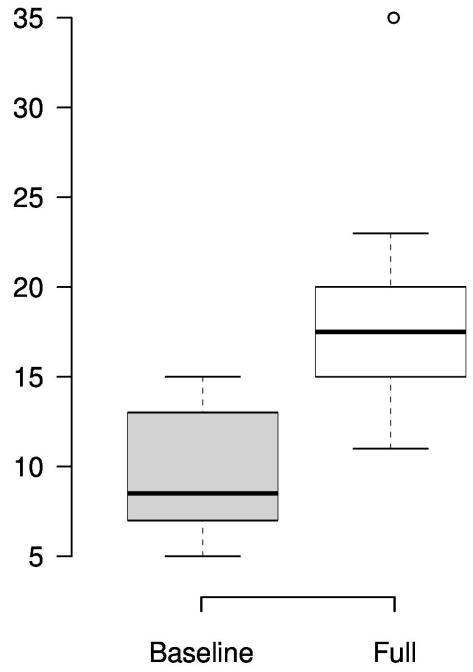


Figure 6.4: Boxplots showing count of valid questions asked by participants in each condition.

Next, I investigated why and how providing live coverage information (the difference between conditions) resulted in an increase in the number of questions by full version users. Using each question's time stamp, I reanalyzed videos and identified utterances right before the formation of the question (if any). After extensive multi-pass analyses, I identified two types of utterances related to the question formation process (Generative & Recollective). Generative utterances represented question formation activities that did not necessarily require the analyst to remember work so far (e.g., "Let's start with Profit and Sale for Regions"). On the other hand, Recollective utterances were indicative of a need to remember prior work (e.g., "let me see... what can I analyze more"). I identified a total of 101 Recollective (full=51, baseline=50) and 64 Generative (full=33, baseline=31) utterances. Interestingly, these numbers were very similar across conditions even though the total number of questions differed, likely because not all questions were preceded by intelligible utterances. Since Generative utterances marked questions that were not reliant on remembering history, I focused my further analysis on Recollective utterances.

Next, for each Recollective utterance, I further analyzed the videos to understand if and how participants interacted with the history views. I identified the interplay between

participant and tool (if any) and what part of the GUI was targeted. To detect the GUI target, I relied on three sources: 1) user interaction with the tool (e.g., clicking a dimension's name in Scented View or the Dimension List, browsing charts in Sequence View, opening Data View), 2) physical gestures (e.g., tracing the list of the dimensions with a finger or a pen and reading the dimension names aloud), and 3) participants' alouds (e.g., "oh it [Scented View] says I missed Returns"). Table 6.2 shows some typical examples of Recollective utterances, the targeted interface widget, and the user interaction.

Table 6.2: Examples of Recollective utterances for each condition. This table also shows the interface target that participants were interacting with at the time of producing utterances and the interaction itself.

Condition	Utterance	GUI Target	Interaction
Baseline	"did I do Order Quantity?"	Sequence View	Browsing
Baseline	"what else [dimension] I have here?"	List of dimensions	Tracing list with finger
Baseline	"let me see, what can I analyze more"	Sequence View	Browsing
Full	"what next? oh, didn't check [Product] containers"	Scented View	Tracing list with mouse pointer
Full	"lets go back and see. with Profit, I have looked at Sales and City"	Scented View	Clicking
Full	"did I consider Days to Ship with Product Category?"	Scented View	Clicking

Figure 6.5 shows the breakdown of Recollective utterances based on condition and GUI target. The total number of identified Recollective attempts were almost equal, 50 and 51 for baseline and full conditions respectively. As shown in the figure, full version users relied heavily on Scented View to recall prior work while formulating new questions (82% of cases). To achieve the same goal, baseline version users relied mostly on the Sequence View (62% of cases). Interestingly, I found that in the other 38% of cases, baseline version

users referred to the list of dimensions (even though it did not contain any dimension space coverage information) and tried to recall what had been done by looking or tracing through the list. This suggests that users intrinsically expect this view to help them to remember their prior analysis. The extensive use of Scented View for recalling past work by full version users corroborates the speculation that the availability of dimension space coverage information helped people to formulate questions.

Figure 6.5: GUI targets that participants interacted with while making Recollective utterances. Full version users frequently referred to Scented View whereas baseline version users relied on Sequence View and the list of dimensions.

6.5.2 H2: Effect on the Number of Findings

H2 posited that revealing dimension coverage information would result in more findings. Following Liu and Heer [49], I define a finding as one of the following:

- **Observation: "a piece of information about the data that can be obtained from a single state of the visualization system".** For example, "I see that lots of round-tables are sold in Texas".
- **Generalization: "a piece of information acquired from multiple visualization states".** For example, "In the south, sales of furniture are higher than any other product category".

This definition excludes any common sense or intuitive conclusions that participants made about the data, in order to isolate findings to only those that were supported by the

analysis tool. To collect findings, I used a multi-pass open-coding approach to qualitatively analyze participants' alouds and their notes. I manually transcribed all participants' alouds from the video recordings, using Transana [4] for video analysis. Later, using the transcribed data, I identified and counted findings for each participant. Each finding was time-stamped to enable cross-referencing with video. I also reviewed participants' notes and extracted all the recorded findings. I only considered relevant and correct findings and excluded all the vague and incomplete instances, such as "...it seems that Sales are [mumbling something]". I also ignored findings that were based on an invalid statistical function or an incorrect interpretation of data. For example, if using sum of values instead of average resulted in an incorrect interpretation of data, that finding was ignored. Table 6.3 shows examples of extracted findings for each condition.

Two independent researchers [1] coded the transcribed utterances. First, both coders analyzed 4 randomly selected experimental sessions (2 full and 2 baseline sessions). Next, each coder independently analyzed 8 of the remaining sessions (4 full plus 4 baseline, assigned randomly). For the sessions analyzed by both coders, I included only those findings where both coders were in agreement. Inter-coder reliability was 0.89, calculated using Krippendorf's alpha.

Table 6.3: Examples of participants' findings for each condition.

Condition	Finding
Baseline	"...in product categories, technology shows a strong performance..."
Baseline	"...Kansas and New Mexico have biggest negative profit..."
Baseline	"...in customer segments, consumer [segment] had the steadiest performance in all 4 years..."
Baseline	"...home-office and small business profit and sales are increasing..."
Full	"...office supplies in west and furniture in east have poor sales..."
Full	"...spikes in order processing times in Q3 and Q4 in 2010..."
Full	"...in west, California has the strongest performance..."
Full	"...profit has gone down from 2009 to 2010..."

[1] Narges Mahyar and Ali Sarvghad

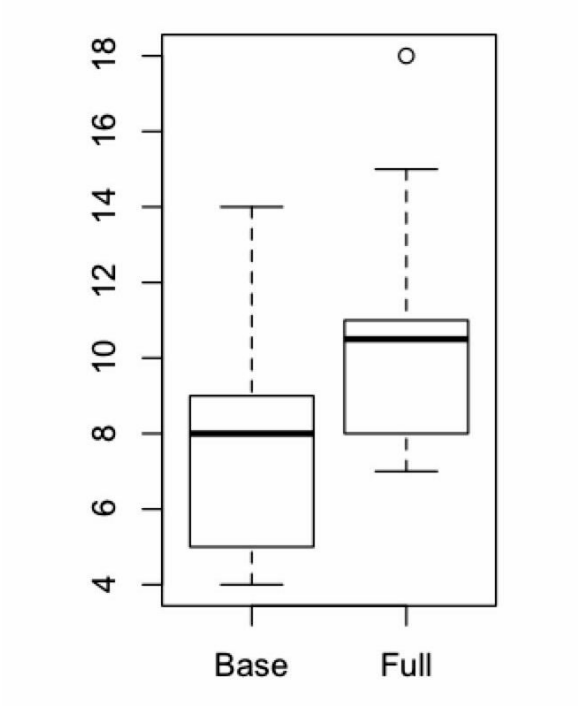


Figure 6.6: Total count of findings by participants in each condition.

Figure 6.6 shows the total count of findings by participants in each condition. Because the data could not be transformed to fit a normal distribution, I analyzed the results using the non-parametric Mann-Whitney test. Although full version users asked more questions on average, the Mann-Whitney test showed that this difference was only marginally significant ($w = 24.5$, $p < 0.055$, Cohen's d=0.419).

6.5.3 H3: Effect on the Breadth of Analysis

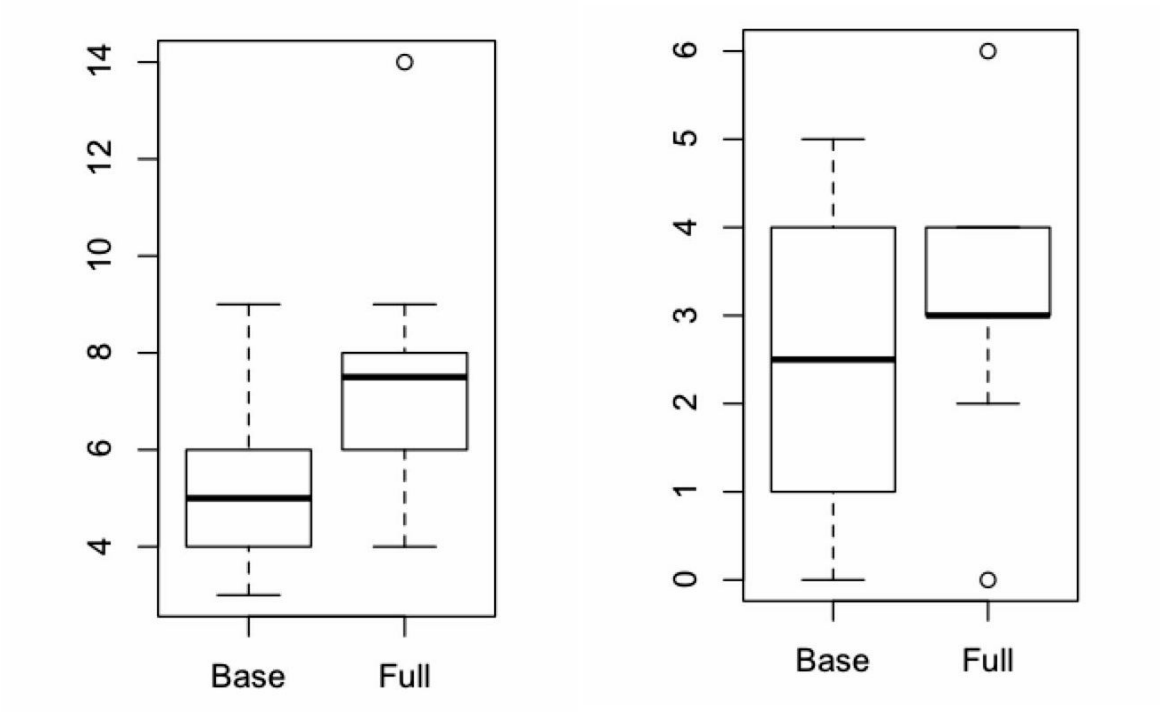
H3 posited that dimension space coverage information would increase the breadth of analysis without sacrificing depth. To evaluate H3, I conducted another multi-pass qualitative analysis to investigate the *process* leading up to participants' findings (findings were those identified in the analysis of H2). Using the timestamps associated with findings and questions, I found the question corresponding to each finding (if any). Looking more closely at questions and findings enabled us to categorize findings into two categories of *top-level* and *drill-down*. Top-level findings are the result of starting a new line of inquiry. Drill-down findings are the result of drilling in on top-level findings. For example, a participant examined "what is the relationship between Profit, Returns (i.e. returned merchandise) and Regions". She observed, "West loses lots of profit because of returns". Triggered by this

finding, she formulated the next question as "Which Product Category in Region [filter: West] has biggest Returns". Consequently, she discovered that "lots of furniture was returned in west". In this example, the former and latter findings are top-level and drill-down findings respectively. Each finding was only considered either top-level or drill-down. In the previous example, if the drill-down finding had in turn triggered a further investigation to discover what furniture items were returned the most, I would still have considered it as a drill-down finding.

Conceptually, top-level and drill-down findings can be considered to represent breadth and depth of exploratory analysis. Top-level findings involved investigating a new aspect of the problem (i.e. a new line of inquiry). For a top-level finding, an analyst created a new question with a focus different from the previous question (e.g. shifting focus from Profit to Sales). Therefore, a greater number of top-level findings suggests a larger breadth of analysis. On the other hand, drill-down findings involved continued investigation of the same problem. Though there may have been some changes in dimensions or filtering, a question resulting in a drill-down finding essentially followed the same analysis path as the preceding question. As a result, more drill-down findings suggests a greater depth of analysis.

Figure 6.7 depicts the count of top-level and drill-down findings in the two conditions. I performed separate Mann-Whitney tests to compare full and baseline groups in terms of their top-level and drill-down findings. I found that participants who used the full prototype produced more top-level findings than those who used the baseline (w=21, $p < 0.030$, Cohen's d=1.05). Full version users also produced slightly more drill-down findings on average, but this difference was only marginally significant (w=42.5, $p < 0.059$, Cohen's d=0.36).

As a second metric of breadth, I also examined the number of dimensions that each participant considered in their analysis (see Figure 6.8). Using the questions identified for evaluating H1, I counted the number of dimensions considered by each participant. Over the same period of time (60 minutes per participant), full version users considered an average of 16.6 dimensions (SD= 2.7), versus 13.6 (SD=3.1) for baseline version users. A two-tail Welch Two-Sample t-test showed a statistically significant difference between the averages ($t(16.3) = 2.43$, $p < 0.027$). In line with my analysis of top-level findings, this result shows that full version users exhibited a greater breadth of exploration.



(a) Top-level

(b) Drill-down

Figure 6.7: Count of top-level and drill-down findings by participants in each condition.

6.5.4 Questionnaire & interview Results

At the end of the study session, full version users filled out a questionnaire that elicited information on their experience. The questionnaire consisted of two parts. The first part asked participants what activities they found Scented, Data and Sequence Views most useful for. Figure 6.9 summarizes participants' responses.

The second part of the questionnaire asked participants to rate the overall usefulness and understandability of each view using a five-point Likert scale. All participants strongly agreed or agreed that the views were understandable. As depicted in Figure 6.10, 7 out of 10 participants were uncertain about the overall usefulness of Data View. On the other hand, participants all agreed or strongly agreed that dimension space coverage information (i.e. Scented View) was useful.

In the interviews, all full version users reported that they relied on Scented View to recall prior work. In particular, they stated that they found the dimension space coverage and co-investigation information very useful. Following are a few examples of participants' comments about Scented View: "I definitely used this [Scented View] a lot, it was quite nice to see what variables I did", "It's nice to see what I have done", "I found the changing

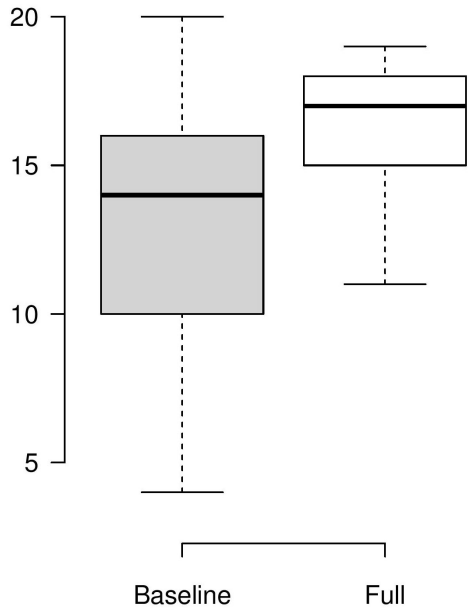


Figure 6.8: Dimensions considered by Full and Baseline tool users.

colours [i.e. co-investigation information] very useful", "colour coding helped me to [see] what I wanted". Although all the full version users valued dimension space coverage information, six out of ten participants reported they did not use the quantitative magnitude information (encoded as bar length). For example, one participant said "I think bars helped me to see what variables I did/did not [do], but I didn't really read the numbers". None of the participants reported that they used Data View for carrying out their analysis task. One participant said "it did pop open a few times, but I didn't really check it" and another participant said "this [pointing at Scented View] was enough for me".

6.6 Discussion

H1 was strongly supported: I found a statistically significant difference between conditions for the number of questions asked, with full version participants asking almost twice as many questions on average than baseline participants. My analysis of Recollective utterances showed that when users needed to understand what they had already done in order to formulate a new question, full version users relied on Scented View and baseline version users relied on Sequence View.

Sequence view is inherently limited in providing first-hand insight into the coverage of

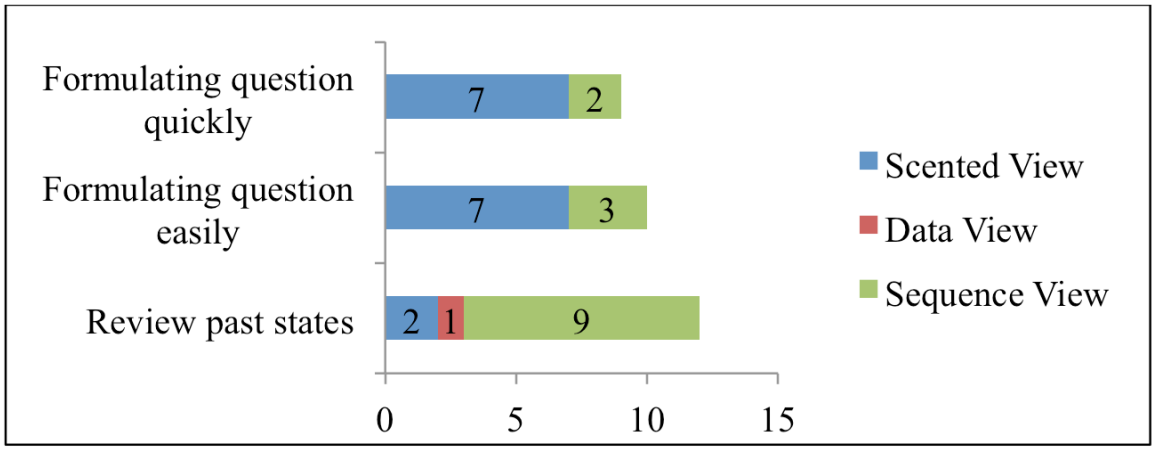


Figure 6.9: Tasks for which different history components were considered useful. Sequence View was most useful for reviewing past states. Scented View was considered helpful for formulating questions quickly and easily. Data view was not considered very useful.

dimension space. I observed that acquiring coverage information from Sequence View consisted of many steps, starting with filtering or browsing the list to find target visualizations, investigating individual visualizations to extract coverage information, and remembering which dimensions were included in these visualizations (note: no one recorded this coverage information on paper). Fewer steps were required for Full version users to acquire the same information because dimension coverage information was constantly present in the interface. Recent research [47] indicates that executing physical sequences is one of the main contributors to the overall cost of interacting with a visualization tool. The recall task was undoubtedly costlier for baseline version users, which is likely one of the key reasons why they formulated fewer questions overall.

H2 was not supported: while there was a trend towards more overall findings for full version participants, this difference was only marginally significant. I speculate that there was insufficient statistical power in my experiment and that more participants might have revealed a significant difference here. However, full version participants did produce significantly more top-level findings (full avg.=7.6, baseline avg.=5.2).

My results showed strong support for H3. Full version users demonstrated greater breadth in their analysis by identifying significantly more top-level findings than baseline users. They also investigated significantly more dimensions. These results indicate that full version users followed more lines of inquiry overall. I also note that this analysis breadth was not at the expense of depth. Drill-down findings represented additional details within the same line of inquiry (i.e. depth). I found no significant difference between full and

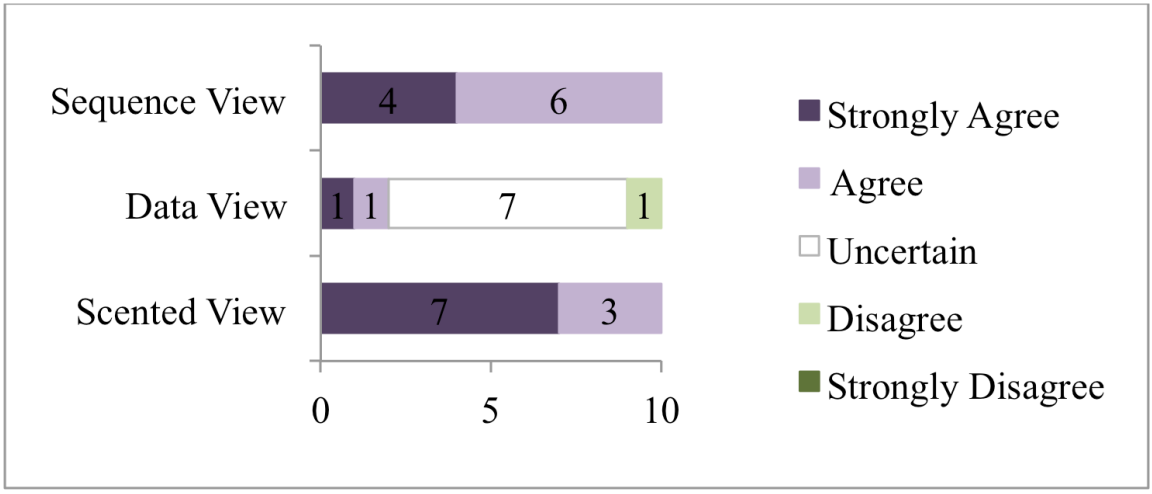


Figure 6.10: Rated overall usefulness of different history views. Participants were asked to agree or disagree with a statement that each view was useful.

baseline users in terms of drill-down findings (if anything, there was a trend towards full version users finding more of these as well).

I attribute the observed greater breadth of analysis by full version users to the availability of Scented View. This view facilitated discovery of "what has not been done", and consequently enabled the analyst to identify new relevant avenues of analysis. Note that I am not arguing that the breadth of analysis is more important than the depth (nor vise-versa). Yet, prior research has suggested that under some circumstances such as exploring new data, it is beneficial to encourage increased breadth, as it can reduce the likelihood of empty results [28] and prevent premature fixation on a single aspect of data [80]. In line with Wattenberg and Kriss [73], I speculate that encouraging breadth may be most useful when exploring a new dataset; the benefits may diminish as analysts become more familiar with the shape and structure of their data and clarify their analysis goals.

Questionnaire results indicated that full version users found Sequence View most suitable for reviewing past states. On the other hand, they found Scented View to be more helpful when forming questions. Interestingly, baseline users referred to the list of dimensions in 19 out of 50 (38%) attempts to recall analysis (Figure 6.5), even though this view did not contain any dimension coverage information in the baseline condition. This suggests that people may intuitively expect this view to help them in recalling dimension coverage information, and that my choice to embed this information within the list using scented widgets was appropriate.

Interestingly, results of the interviews with full version users showed they all found

information about "what has been investigated and in what combinations" more useful than the frequency information (i.e. bar length and the number). This suggests that the Scented View could be further compacted by reducing the bar to a smaller mark that is either present or absent.

Another interesting finding relates to the relative value of dimension coverage information as compared to data values coverage. While Scented View was highly used and considered very useful, the opposite was true for Data View. None of the participants reported using Data View to help them formulate questions (Figure 6.9). In addition, only two out of ten participants agreed or strongly agreed that this view was useful (Figure 6.10). These results suggest that dimension coverage information may be much more important than data values coverage. Although I could not experimentally isolate the value of Data View because it was in the same condition as Scented View, ratings and qualitative observations suggested that participants found Scented View to be much more useful. It is of course possible that the value of Scented View was related to the specific task in my study or to the view's visual prominence in the interface. In [76], the exploratory analysis task involved investigating a constant set of dimensions by manipulating the filtering of values. In such a case, being able to visually understand which data values were explored versus left out might be more important to formulating new questions. Nonetheless, evidence here suggests that dimension coverage information is more useful than data values coverage in the more general case.

6.7 Conclusion

I examined the value of providing dimension space coverage information to support exploratory visual data analysis of unseen data. I illustrated how this information can be incorporated into the interface of a visual analysis tool by using scented widgets. My results demonstrated that this approach could increase the number of questions asked about data and expand the breadth of analysis without sacrificing depth. In addition, providing dimension space coverage information increased the number of top-level findings.

In the next chapter I will summarize the main findings and contributions of this thesis and concisely discuss their implications. I will also highlight threats to validity that may impact my findings. Lastly, I will discuss future research directions to further investigate and/or consolidate outcomes of this thesis.

Chapter 7

Discussion and Future Work

In this chapter I further discuss contributions of this research and their applicability in other domains. I elaborate on the findings of each user study. I will also address threats to validity for each study such as the generalizability of my results and their limitations. Finally, I point out research directions derived from this research. This thesis was an initial step to representing history of exploratory VA from the angle of dimension space coverage. There are still many open questions that can be considered as future work. I hope the contributions of this thesis motivate more research to further investigate this field.

7.1 Summary of Studies

I started my research with an investigation of visual history in the context of collaborative exploratory data analysis (Chapter 3). Along with a co-investigator[1], I designed CoSpaces (3), a prototype for visual data analysis on large interactive surfaces. The prototype was designed based on the available guidelines and frameworks for designing collaborative VA tools at the time. It contained a record-keeping module that supported externalization along with a linear history representation of visualization states.

Next, I performed an observational user study with CoSpaces , in which I asked participants to explore a sales data set to evaluate business performance. Based on the findings of this study, participants mainly used history to review/recall past analysis and/or reuse created visualizations. Using the linear history, participants reloaded charts and tried a new avenue of analysis or drilled in by manipulating mapping and filtering of dimensions. An interesting observation was that participants used history to understand the scope of explo-

[1] Narges Mahyar

ration. More specifically, they reviewed history to gain a holistic understanding of "what was done so far" and decide on "what to explore next".

I noticed that this process was rather slow and cumbersome, especially when history was relatively long. In order to retrieve the information about prior coverage of dimension space, participants had to browse the history and rely on their memory to record the extracted information. Although the linear history contained dimension space coverage information (i.e. in the case of tabular data, mapping of data dimensions), it did not provide a first-hand overview or details on-demand.

Observations in the CoSpaces user study motivated me to further investigate how to support understanding the coverage of dimension space. I decided to focus on asynchronous collaborative exploratory VA because direct communication between collaborators is often very difficult or not possible. Therefore, the collaborators have to rely on other channels of communication such as sharing the analysis history. In addition, this type of collaborative work setting is easier to emulate in the lab environment. I designed and implemented Footprint-I, a history tool with ability to represent history of data exploration from the angle of dimension space coverage (Chapter 4). It provided first-hand information about investigation of data dimensions. Based on the results of a user study, in comparison to a linear history model, Footprint-I users were twice as fast and more accurate in answering questions about the dimensions considered in the prior analysis.

Positive findings of the Footprint-I user study motivated me to investigate the effects of providing dimension-centric history view on collaboration. I designed and implemented the next version of my visual history prototype, named Footprint-II (Chapter 5). I conducted a user study that asked participants to continue exploratory analysis started by a collaborator. Findings of this study clearly demonstrated that collaborative visual data analysis can benefit from visually representing the history of dimension space coverage. Participants with access to dimension space coverage information showed better coordination with the previous analyst through their focus on uninvestigated aspects of the problem. The similarity analysis showed that these participants asked questions that were more different from the ones asked in the initial analyst. For example, the initial analyst did not investigate the Days to Ship dimension (i.e., days from receiving to shipping of an order). Yet, inefficient order processing times could be responsible for overhead costs and loss of Profit. Visualizing dimension space coverage helped six out of ten participants with access to this information to examine this possibility. Based on my findings, dimension-centric approach for representing history of EDA can significantly assists analysts in recognizing and focusing on uninvestigated dimensions.

In chapter 6, I investigated the effects of providing dimension space coverage on exploratory data analysis. I used Scented Widgets to incorporate dimension space coverage information into the GUI widgets of a VA prototype tool. Based on the results of a controlled user study, I found that this approach helped participants to: 1) formulate and ask more questions, 2) demonstrate a greater breadth in their analysis, and 3) reveal more top-level findings. Further analysis of the study data showed that these benefits could be attributed to the presence of the dimension space coverage in the GUI. I observed that acquiring coverage information from the linear history consisted of many steps, starting with filtering or browsing the list to find target visualizations, investigating individual visualizations for details (i.e. mapping and filtering of dimensions), and memorizing acquired information. Fewer steps were required for users with access to dimension space coverage information to acquire the same information because coverage information was constantly present in the interface. My analysis of the number of findings showed that there was a trend towards more overall findings for participants with access to dimension space coverage information, although the difference was only marginally significant. Results of a questionnaire indicated that participants found the linear history model most suitable for reviewing past states. On the other hand, they found dimension coverage information (in Scented View) to be more helpful when forming questions.

I speculate that providing dimension space coverage is most useful when analysts explore a multivariate tabular data that they have not seen before. Exploration can benefit from dimension space coverage information because it often starts with vague analytical goals and analysts browse data to serendipitously come across interesting findings [80]. Therefore, assisting analysts in remembering the coverage of dimension space provides support for browsing and breadth of exploration. Unfamiliarity with shape and structure of data can impede exploration as well. Providing dimension space coverage information can help analysts to compensate for the their unfamiliarity with data. In line with previous research [73], I posit that this benefit could gradually diminish as the analyst becomes familiar with data.

In my research, I only focused on positive effects of providing dimension space coverage. Yet, further research is required to assess possible negative impacts. For instance, might visualizing dimension space coverage in a collaborative context lead to groupthink [38]? This might be a stronger possibility in a situation when less-experienced analysts work with seasoned colleagues. Viewing their analysis trail my bias the less-experienced to follow their seniors. Is it also possible that, in the same context, providing dimension space coverage information could impede drilling-in, preventing the analysis from going

deep? Further research is required to investigate these questions and other possible negative effects that dimension space coverage information may have on exploratory analysis as well as collaboration.

7.2 Threats to Validity

It is well known that the choice of methods and study settings can directly affect the results and their validity. In section 1.3, I justified the choice of methods and discussed approaches I took to offset possible flaws. Nonetheless, every single choice in designing and evaluating a lab study can influence the results. As expected, results of the studies conducted in this thesis are subject to some caveats. In this section, I describe the limitations of this thesis and address the empirical validity of the results by describing threats to construct validity, internal validity, external validity and reliability [22] [56].

7.2.1 Construct Validity

Construct validity "focuses on whether the theoretical constructs are interpreted and measured correctly" [22]. This threat to validity is mostly applicable the notion of "dimension coverage" and its measurement. I considered a dimension "covered" if it was investigated individually or in relation with other dimension(s) in a visualization (e.g. scatter plot). Although other researchers [80] have similarly considered mapping of dimensions in visitations as an indication of coverage, this notion of does not take into consideration factors such visualization type (e.g. bar chart versus treemap) or the exploratory dead-ends.

Footprint-II user study (Chapter 5), I used an adapted version of Jaccard's similarity index to measure overlap between analyses and quantify task coordination. I hypothesized that a similarity score computed based on the number of similar and different questions formulated by two analysts would indicate the degree of divergence of their works. Although I carefully adopted Jaccard's index (see Chapter 4 for details), it might not be optimally tuned for measuring analysis similarity. In addition, this metric only quantified similarity in terms of dimension-combinations and not the analytical focus. For example, two analysts A and B created charts showing Sales in different States across different Genders and analyst A was interested in both male and female population but analyst B was only interested in Female half. Yet, in both cases the tool considered Gender as covered.

In the scented widgets user study, I categorized analysts' findings as top-level and drill-down (see Chapter 6 for details). Although I referred to existing literature on EDA to define

what constitutes a finding in general, my definition and categorization of top and drill-down findings may require refinement. It should also be mentioned that my definition of finding does not capture anything about the importance or value of the finding. I quantified the breadth of analysis using two metrics: the number of top-level findings and the number of dimensions considered by the analyst. Although I paid careful attention in choosing, adapting and/or developing these metrics, they are not perfect and may not represent the breadth of analysis optimally.

7.2.2 Internal Validity

Internal validity deals with "the degree to which results of a study permit you to make strong inferences about causal relations" [56]. I aimed to carefully control differences between conditions (e.g. in the last study (Chapter 6), the only difference was the presence or absence of explicit dimension space coverage information; otherwise the systems were identical). But, there are still some issues that present possible threats to internal validity. In the CoSpaces user study (Chapter 3), two coders coded the data together; however, independent coding may have led to higher reliability. Therefore, learning from the this study, in the scented widgets user study (Chapter 6), coding was done by two independent coders. I also measured inter-coder reliability to ensure validity of coders' interpretations. For the Likert style questionnaires in all the studies, I closely followed accepted standards to alleviate acquiescence bias. For instance, I tried to avoid simple Yes/No questions (where possible); this should provide a response format that is less prone to the bias and also provide more detailed data.

7.2.3 External Validity

External validity "refers to how confident you can be that the findings of your study will hold up upon replication, and how confidently you can predict both the range over which your findings will hold and the limits beyond which they will not hold" [56]. Here I elaborate on issues that deals with the generality of results as well as their practicality and validity.

All the user studies were carried on in a controlled lab environment. Controlled laboratory user studies provide control over variables, enable careful measurements, and facilitate establishing cause and effect relationships. Therefore, I decided to examine my research questions using this method of evaluation. Yet, these should be followed by field studies that would explore how the findings hold up in real-world situations.

Because it is extremely hard to find real data analysts, I used students as my participants. For the CoSpaces (Chapter 3) and Footprint-I (Chapter 4) user studies, I recruited computer science students, and I used business students for Footprint-II (Chapter 5) and scented widgets (Chapter 6) user studies. While I paid careful attention to hiring only graduate or senior undergraduate student participants with data analysis knowledge, familiarity with creating statistical charts, etc., it is not clear to what extent the student participants behaved like real analysts. I also chose tasks that were carefully modeled after real domain tasks and consulted a business PhD student in the process of designing them. However, students can not replace professional business data analysts and contrived tasks cannot replace real work. These differences limit the value of a lab study as a model situation.

Although the main focus of the thesis was representing analysis history from the angle of dimension space coverage, the selected application domain might be a limitation to generalizability. In all the user studies, I chose to focus on the business domain. I suspect that the effects and benefits of representing analysis history from the angle of dimension space coverage will be similar for exploratory analysis of tabular data in other domains, but it is possible that I observed some peculiarities unique to business.

Another limitation to generalizability of the results is the prototype tools that I designed and implemented for the user studies. For all the studies, I paid special attention to the design and followed existing guidelines. For CoSpaces, I closely followed the guidelines for designing collaborative VA tools. I also iteratively improved the design of dimension space coverage in each prototyping iteration (Footprint-I, Footprint-II, ScentedView) using users' and experts' feedback as well as existing data visualization techniques and guidelines. However, the results of the studies might still have been influenced by the design of the prototype tools. Future user studies are needed to examine design decisions made in these tools and the effects of representing history from the angle of dimension space coverage on exploratory data analysis process and outcomes.

7.2.4 Reliability

Reliability "focuses on whether the study yields the same results if other researchers replicate it" [22]. I provided detailed descriptions for each study to be replicable by other researchers. Yet, it is possible that other researchers might find other interesting patterns in the data or code the data differently. For example, in the fourth user study, I developed a coding scheme for identifying findings based in a definition of what constitutes a finding. Yet, defining a finding differently may result in different analysis outcomes for

a similar study. Overall, this thesis initiates a research direction to representing analysis history from the angle of dimension space coverage. While there are certain limitations for each phase of this thesis, the main findings in the last study reflect what I found in earlier studies. In addition, results of each study corroborated many findings reported in previous work. Nonetheless further studies are required to better understand the reliability of these findings.

7.3 Future Work

This research is an initial step towards understanding the effects of representing analysis history from the angle of dimension space coverage. However, there remain many promising avenues for future research. All the limitations discussed above call upon further research. In addition, future research can go beyond this list to investigate additional aspects. In this section, I point out some of these opportunities.

One the most important issues is the notion of "dimension space" coverage. As of yet, there is no formal definition of what constitutes coverage. In my work, I considered presence of a dimension in a "question" as its investigation and coverage. This is based on the assumption that formulated questions summarize analyst's analytical intentions and interests in data. However, this definition falls short in accurately capturing the actual investigation of a dimension's values. For instance, based on my current definition of coverage, dimension **Gender** is *covered* when it appears in a visualization that investigates average income of men (Gender filtered to exclude females) in California. In this case, dimension Gender has only been used as a filter to exclude half of the data and the analyst did not really explore or compare values on this dimension. One possible solution is to present coverage information based on more than one factor. In my work, data view revealed the coverage of values for each investigated dimension. Combination of dimension and data views provided more comprehensive information about the coverage of a dimension.

Another important issue relating to the notion of coverage is visual representation design. Based on the guidelines for designing visualizations (e.g. [59]), some visual encodings better convey magnitude (e.g. position on common axis) while some other are better for identity information (e.g. hue). Should a dimension represented using less-effective visual encoding (e.g. using saturation for visualizing categories of birds) be considered "covered" since inefficient visual encoding failed to reveal underlying information?

In addition to mapping and filtering of dimensions, there are other techniques that could be used for understanding analyst's focus and attention. For example, eye tracking could

be used to collect information about analyst's focus. This technique can complement mapping/filtering information for more accurate estimation of coverage. Future research is required for investigating and defining the notion of "coverage" in the exploratory data analysis context.

Currently, the use of proximal cues in my work is limited to providing information about what is available to explore. Further research is required to determine how proximal cues could be improved to provide information about the value of distal sources. For example, text of a hyperlink on a webpage provides cues on both the existence of a related document (the link itself) and the possible content (text of the link). One possible solution for adding value cues to my current design is linking scented view with recommendation. Selecting an uninvestigated dimension could be complemented by recommending a list of views relating to that dimension that is worth further investigation. Another possibility is to show small histograms showing distribution of values for dimensions. Distribution of data values can function as proximal cues for investigating distal information source.

Since the primary focus of my research was to examine the speculated value of providing the history of the coverage of dimension space, I did not fully explore the rather immense possible design space. I did attempt to make good design choices based on current perceptual knowledge and iterative development; however, future research could further explore this design space and may be able to improve upon my representations. In addition, all different designs of dimension space coverage in this thesis were specific to tabular data. Extension of the data-centric history idea to other types of data that do not have this discrete tabular nature is not obvious. However, my work does demonstrate that showing people a summary of their past work can help them to decide what to do next. This finding should have wider generalizability and implications beyond one specific data type. Figuring out how to apply this idea to other data, however, would require completely rethinking the design and the content to show. I would like to explore such extensions in future work.

Another important issue that calls for further investigation is the effect of Dimension View on analysts information seeking behaviour. Based on my findings, dimension coverage information resulted in a greater breadth of exploration without com- promising the depth. Yet, providing dimension coverage information at the surface level may bias analyst towards breadth-oriented work. On the other hand, having data coverage information at the GUI level may bias analyst towards drilling-in. Future research is required to investigate if and how dimension or data coverage information should be represented to prevent biases. One possible solution is to enable users to customize the coverage view (i.e. what

information is placed within GUI controls) based on their exploratory goals and needs.

I would also like to examine the approach of providing live dimension and data space coverage in a collaborative setting. In particular, would like to extend ScentedView with capabilities to distinguish among the activities of multiple users in a collaborative team. For example, if four people are all exploring the same dataset, it could be helpful to see what aspects each person has worked on. When working simultaneously, a dimension centric view might serve as a helpful awareness mechanism, and indicate which parts of the data are being neglected and might be worthy of exploration. Such extensions would need to consider how to present the contributions of different users without cluttering the interface.

Coverage information can also be augmented with information that shows the findings from a line of inquiry. This information help collaborators to discover dead-ends. Design of Scented View has no chronological property and and does not provide information about the precedence of dimensions investigation. Yet, recency might be important, especially in a collaborative context where this information provides awareness of the chronology each persons work. Another important aspect is control over privacy and level of sharing. In my current design, coverage information is shared in full. Yet, there should be mechanisms to provide control over level of sharing or identity of the analyst.

Chapter 8

Summary and Contributions

In this chapter, I concisely review the main contributions and implications of each stage of my research.

I started this research by building an understanding of the role and use-cases of history in the process of collaborative EDA. Based on a literature review, I complied a list of the most common use cases for history in this context. Many of the use cases are also extendable to non-collaborative situations. I discovered that prior research has been more focused on investigating analysis history for reusing/communicating visualization states and/or actions/commands. My main contribution at this stage was:

- **C1: Identifying the most common history use-cases for collaborative data analysis**.

The first step of my research provided information about the state of the art about history in the context of EDA. Next, to gain a better understanding of possible shortcomings, I designed CoSpaces, a collaborative EDA prototype for a tabletop display. CoSpaces contained a history module which, following the existing guidelines and frameworks at the time, represented history as a linear list of visualization state thumbnails. I conducted a user study which showed that the linear representation of history adequately supported reusing captured states for branching the analysis. I also observed that users reviewed their work history to determine "what has been done" and "what else was left" for further investigation. It was rather cumbersome to gain this knowledge by using the linear history. Despite the observation that users innately refer to history for gaining this understanding, the representation of captured analysis history did not provide first-hand information about the coverage of dimension space. The main contributions of this stage were:

- **C2: Identifying that users innately expected history to provide information about dimension space coverage.**

- **C3: Revealing that the linear history representation is inadequate for providing top-level information about the coverage of dimension space.**

Based on findings of the CoSpaces evaluation, I suggested representing history from the angle of dimension space coverage. I designed and implemented Footprint-I (Chapter 4), a tool designed to support understanding analysis history from a dimension space coverage perspective. Results of a user study suggested that providing dimension space coverage information enabled people to be faster and more accurate in answering questions about dimensions used in a prior analysis. My research contribution at this stage was:

- **C4: Showing that representing analysis history from the angle of dimension space coverage assisted analysts to be faster and more accurate in understanding the breadth of a prior exploratory analysis session.**

After positive results from the Footprint-I study, I investigated if and how providing dimension space coverage information could affect group coordination. Based on the findings of a user study of Footprint-II (Chapter 5), I showed that in an exploratory collaborative VA context, representing dimension space coverage could improve implicit task coordination. It assisted participants who used Footprint-II to better identify and focus on parts of dimension space that were not investigated in the initial analysis and tried to avoid asking duplicate questions that were already investigated. The main contribution of this stage of my research was:

- **C5: Representing history from the angle of dimension space coverage can improve tacit task coordination between collaborators.**

I then shifted my focus to investigate how providing live information about the coverage of dimension space would influence exploratory process and outcomes. At this point, I illustrated how this information can be incorporated into the interface of a visual analysis tool by using scented widgets. My results demonstrated that this approach could increase the number of questions asked about data and expand the breadth of analysis without sacrificing depth. In addition, providing dimension space coverage information increased the number of top-level findings. The main contributions of this phase of my research were:

- **C6: Using scented widgets to represent dimension space coverage information increased the number of questions asked by analysts.**

- **C7: Using scented widgets to represent dimension space coverage information increased the number of top-level findings.**

- **C8: Using scented widgets to represent dimension space coverage information resulted in a greater breadth of exploratory analysis without loss of depth.**

Overall this thesis revealed the value of representing the history of exploratory analysis from the angle of dimension space coverage. I contributed an understanding of how this approach could facilitate collaborative and single-user EDA. In addition to contributions that directly correspond to different phases of my thesis, I examined a number of possible approaches and designs for representing dimension space coverage. More precisely:

- **C9: I illustrated some viable visual representations of dimension space coverage information and how such information can be incorporated into visual data analysis tools.**

This contribution is based on the iterative process of examining different visual representations for dimension space coverage through RQ2 to RQ5. In Footprint-I and Footprint-II, I used circular and treemap layouts for visualizing dimension space coverage. In Chapter 6, I used scented widgets to incorporate dimension space coverage information directly into the interface elements of an analysis tool.

Overall, this dissertation demonstrates that representing the history of collaborative EDA from the perspective of dimension space coverage is beneficial for both exploration and collaboration. In a collaborative context, showing dimension space coverage can help the next analyst to better identify missing investigative work and decide what to do next. In a single user context, incorporating dimension space coverage into GUI elements can expand the breadth of analysis and improve analysis outcomes. These results demonstrate the value of tracking and representing dimension coverage information for EDA and suggest that designers should consider incorporating this feature into EDA tools.

Appendices

Appendix A

Materials for the CoSpaces Study

A.1 Consent Form

University of Victoria, Department of Computer Science

Participant Consent Form

Participant ID:

Human Factors in Visualization

You are being invited to participate in a study entitled [Collaborative visualization around large displays] that is being conducted by Melanie Tory, Ali Sarvghad, and Narges Mahyar. Melanie Tory is a faculty member in the department of Computer Science at the University of Victoria and you may contact her if you have further questions by email at mtory@cs.uvic.ca or by phone at (250) 472-5798. This research is being funded by NSERC.

The purpose of this research project is to investigate how people use information, and how different visual representations of data affect how people perform tasks such as data analysis and decision making. Research of this type is important because it allows us to design better data displays to allow more effective, efficient, and enjoyable analysis of data in a variety of applications..

If you agree to voluntarily participate in this research, your participation will include:
o Filling out a background questionnaire that asks about your experience with computer technology and data analysis applications, as well as personal characteristics such as age and gender.
o Completing computer-based tasks.
o Participating in a verbal interview.
o Filling out a questionnaire about the computer-based tasks and tools you experienced.
o Being video and audio-taped.
o Being watched by live observers.
The research session take place at [Uvic ECS 654].

There are no known or anticipated risks to you by participating in this research.

Your participation in this research must be completely voluntary. If you do decide to participate, you may withdraw at any time without any consequences or any explanation. If you do withdraw, we will ask whether we may use your data for data analysis. If you decline, your data will be destroyed.

Your confidentiality and the confidentiality of the data will be protected by identifying data only with a participant number rather than your name, password-protecting computer files, and storing video and audio tapes in a locked office. Because this is a group study, confidentiality cannot be fully guaranteed since other participants in your group may know your identity. Confidentiality may not be guaranteed because the nature or size of the sample from which participants are drawn may make it possible to identify individual participants.

It is anticipated that the results of this study will be shared with others in the following ways:
o Published articles
o Conference presentations
o Video publications
o Theses
o Internet project descriptions

Data from this study will be disposed of within 5 years. Electronic data will be erased, paper copies will be shredded, and video/audio tapes will be recorded over or physically destroyed.

In addition to being able to contact the researcher at the above phone numbers, you may verify the ethical approval of this study, or raise any concerns you might have, by contacting the Associate Vice-President, Research at the University of Victoria (250-472-4545).

Your signature below indicates that you understand the above conditions of participation in this study and that you have had the opportunity to have your questions answered by the researchers.

Name of Participant _____Signature_____Date_____

Email address:_____Gender: F M Age: _____Degree:_____

A copy of this consent will be left with you, and a copy will be taken by the researcher.

A.2 Introduction

We have implemented an application for co-located collaborative visual analysis. The application that you will see is a prototype that is in the early stages of development and now we want to get feedback from users before the design is too far along.

The objective of this study is to evaluate our design and implementation decisions such as our proposed workspace, note taking and visualization history mechanisms. We also aim to evaluate usability of the system for co-located work, and to understand how you use any saved information.

I want to stress that we really want your honest feedback about the strengths and weaknesses of this program. Knowing what works and what doesn't work will help us to improve the next version.

Functionalities to be explained to the participants:
1- How to create a new workspace (also multiple workspaces)
2- How to create a chart (selecting chart type and mapping and filtering)
3- Talk about the data
4- How to map variables, how to undo mapping, target highlighting
5- How to filter data, how to undo filtering
6- How to clear a chart
7- How to take a note, text wrapping, explain colour coding of the notes and scrolling
8- How to review a note
9- How to reuse history, scrolling of history, reloading
10- How to use tabs, view other's current chart, reload from their history

Interactions to be explained to the participants:
11- Rotate
12- Zoom
13- Pan
14- Drag and drop

Explain where to touch to perform zoom and pan on keyboard and workspace

If you have questions or require assistance, please ask the observer. However, we would like you to first attempt to complete each task with the help of your partner.

Note that this study includes two tasks. In the first task you have some focused questions which will help you become familiar with the system as well as the dataset. The second task is an open-ended scenario that requires making a decision and creating a final report.

A.3 Task

Group:
Participant ID:
Date:
Task 1
Partcipant 1: NY Please follow all steps that are listed. We encourage you to make comments and discuss things with your partner aloud, as this will help us uncover problems with the prototype.

Suppose you are a team of managers in a retail chain, and you want to plan out future marketing campaigns. You have a spreadsheet of sales data for three years and you teamed up to get a reasonable overview of the company's sales across regions, products, and time. Please complete the following tasks:

1. Find out the trend in sales revenue over 2003.
a. Create a workspace.
b. Select line chart from the chart type menu.
c. Drag and drop sales revenue and quarter to yellow and blue highlighted areas and filter it based on 2003.
d. The current chart will appear in the history bar when you create a new chart.

2. Create a bar chart to see quantity of sales over category in NY.

3. How does the 2003 margin compare to previous years (2001 & 2002)?
a. Create a bar chart showing margin over year and compare their values.

4. What are the trends in the timing of purchases of Jackets and shirt-waist?

a. Create a bar chart that shows Quantity sold over year. Filter Lines to see data for Jackets.
b. Record your findings as a note.
c. Remove Jackets and filter based on shirt-waist.
d. Take a note.

5. Compare quantity of sales of different categories in NY and CA.
a. You created a chart for NY (Your team member created the same chart for California).
b. Create a new workspace, click on your partner's tab to view his/her workspace. From the history bar, reload the chart created for Quantity sold of products in CA.
c. Bring the new workspace next to yours and compare your NY chart to your team member's CA chart.
d. Take a note of your findings.

Task 2

Assume you are a new financial data analyst of a company that sells clothing to customers in the US over the Internet. Following is a list of the most popular product lines:

o Sweat t shirt
o Shirt waist
o Accessories
o Dresses
o Sweaters
o Outwear

You should divide this task with your partner. You will look at the first three items (underlined) and your team member will look at the rest. Analyze the sales data (Sales Revenue, Margin, Quantity sold) of the last 3 years over time, state and cities for these three items.

At the end, you and your team member should discuss and share your findings to be prepared to report back to your CEO.

A.4 Questionnaire

1- Do you feel that you successfully completed all the tasks on the task sheet?
Yes No

2- In relation to other software I have used, I found this prototype to be:
Very difficult to use 1 2 3 4 5 6 Very easy to use

3- The menu items were well organized and functions were easy to find.
Strongly disagree 1 2 3 4 5 6 Strongly agree

4- I immediately understood the function of each menu item.
Strongly disagree 1 , 2 3 4 5 6 Strongly agree

5- I found navigating around the prototype screen to be:
Very difficult 1 2 3 4 5 6 Very easy

6- It was easy and intuitive to reuse saved charts.
Strongly disagree 1 2 3 4 5 6 Strongly agree

7- The note taking mechanism was easy to use.
Strongly disagree 1 2 3 4 5 6 Strongly agree

8-I found my recorded information (charts and notes) useful.
Strongly disagree 1 2 3 4 5 6 Strongly agree

9-I found my team member's recorded information (charts and notes) useful.
Strongly disagree 1 2 3 4 5 6 Strongly agree

10- My overall impression of the prototype is:
Very negative 1 2 3 4 5 6 Very positive

11- In your opinion, what are the most useful features of the prototype?

Comments (please write down the application's problems. Your constructive suggestions

are also appreciated):

A.5 Follow up Interview

1. At what point did you get stuck? What was the reason? What are the most confusing features of the prototype?

2. What are the most useful features?

3. How did you find history items? Did they help you in any way?

4. How about note taking? Did you find it useful?

5. What are your suggestions to improve the application?

6. What other comments do you have on this study?

Appendix B

Materials for the Footprint-I Study

B.1 Consent Form

A consent form similar to: A.1.

B.2 Introduction

Introduction to the context of the study.

Introduction the Footprint-I and its functionalities.

A 20 minutes warm up task to let participant practice using Footprint-I. Participants answered a set of questions about similar in structure to the actual study questions but from the history of a different analysis. Participant was allowed to ask any questions from the present observer.

B.3 Task

Participant ID:
Date:
Task
Please answer the following questions 1 to 3. At the start of each question please say “I am starting question #”. Similarly, at the end of each question say “Question # is finished”.

Question1:

Complete the table below with information about data dimensions. Select Yes if a dimension was investigated or NO if it was not.

No.	Dimension Name	Was it investigated?
1	Product Name	Yes/No
2	Product Category	Yes/No
3	Product Sub-Category	Yes/No
4	Product Container	Yes/No
a5	Product Base Margin	Yes/No
6	Customer Name	Yes/No
7	Customer Segment	Yes/No
8	Ship Date	Yes/No
9	Ship Mode	Yes/No
10	Order Priority	Yes/No
11	Order Date	Yes/No
12	Order Quantity	Yes/No
13	Region	Yes/No
14	State	Yes/No
15	City	Yes/No
16	Zip Code	Yes/No
17	Profit	Yes/No
18	Sales	Yes/No
19	Unit Price	Yes/No
20	Discount	Yes/No
21	Shipping Cost	Yes/No
22	Manager	Yes/No

Question2:

I) Choose the most accurate order based on the overall mapping frequency of dimensions:

a) Order Date > Order Quantity > Unit Price
b) Order Date > Profit >Unit Price
c) Unit Price > Profit > Order Quantity
d) Order Date > Profit > Product Sub-Category
e) Order Quantity > Profit > Product Category

II) Which of the following questions did the analyst ask about Unit Price and Order Date?

a) Relationships between Unit Price, Order Date and Ship Cost
b) Relationships between Unit Price, Order Date and Profit
c) Relationships between Unit Price, Order Date and Product Container
d) Relationships between Unit Price, Order Date and Product Name

III) Which of the following questions about Profit seems to be most important to the analyst based on the frequency of co-mapping?

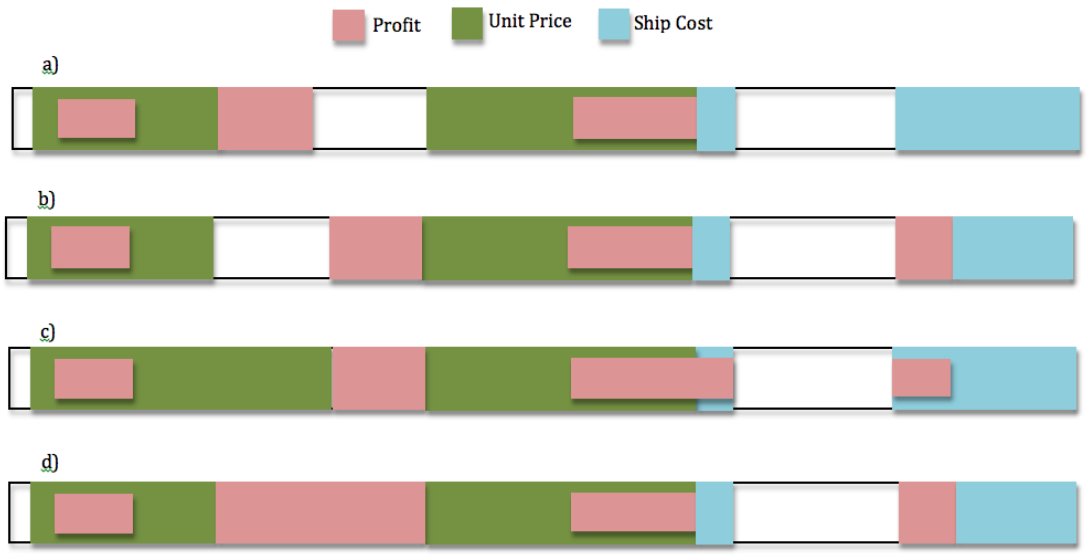
a) Relationship between Profit and Ship Cost
b) Relationship between Profit and Unit Price
c) Relationship between Profit and Order Quantity
d) Relationship between Profit and Product Category

IV) Which relationships did the analyst investigate when analyzing Ship Cost?

a) Ship Cost, Order Date, Unit Price
b) Ship Cost, Order Date, Profit
c) Ship Cost, Ship Mode, Product Container
d) Ship Cost, Order Date, Ship Date
e) Ship Cost, Product Container, Product Category

Question 3:

Which of the following is the closet approximation of temporal distribution of Unit Price, Profit and Ship Cost during analysis? (Time runs left to right.)



Appendix C

Materials for the Footprint-II Study

C.1 Consent Form

A consent form similar to: A.1.

C.2 Introduction

Introduction to the context of the study.

Introduction to Footprint-II and its functionalities.

A 20 minutes warm up task to let participant practice using Footprint-II. They used a dataset different from study's dataset. Participant was allowed to ask any questions from the present observer. Participants also practised "think aloud" protocol during this phase.

C.3 Task

Participant ID:
Date:
Task
You are a business data analyst in a large international company. You are working collaboratively with other analysts in your company to explore sales data for the past 4 years and identify any possible strong and/or poor performance. Your collaborators are at different times/locations and work completed by others is passed around to be built upon. For your own analysis, you should explore the data and try to identify any interesting/unexpected patterns in the data with respect to business performance. In order to efficiently continue your collaborators' work, you first need to review and understand the prior work passed to you. This will also help you to keep the similar work minimized and investigate different plausible performance indicators. While doing your analysis, you can review the collaborators' work if required.

C.4 Follow up Interview

1. At what point did you get stuck? What was the reason? What are the most confusing features of the prototype?

2. What are the most useful features?

3. How did you find dimension view? Did they help you in any way?

4. How did you find sequence view? Did they help you in any way?

5. What are your suggestions to improve the history views?

6. What other comments do you have on this study?

Appendix D

Materials for the Scented View Study

D.1 Consent Form

A consent from similar to: A.1.

D.2 Introduction

Introduction to the context of the study.

Introduction to the prototype and different history views.

A 20 minutes warm up task to let participant practice using prototype. They used a dataset different from study's dataset. Participant was allowed to ask any questions from the present observer. Participants also practised "think aloud" protocol during this phase. During the warm up task, participants were given a set of instructions on how to create different chart types. Instructions remained with the participant until end of the study for possible further referral.

D.3 Task

Participant ID:
Date:
Task
You are a business data analyst in a large online retailer. You should explore the performance data for the past 4 years and identify trends/outliers in the data that are indicative of strong and/or poor performance.

D.4 Questionnaire

D.5 Follow up Interview

1. Did you use any history views during your analysis?

2. Which view(s) did you find most useful?

3. Why did you find this (those) view(s) useful?

4. Can you use your analysis and show an example of how you used this (those vies) in practice?

5. Why you did not find other view (s) as useful? (if any)

6. What are your suggestions to improve the history views?

7. What other comments do you have on this study?

1 **Dimension View** was most useful for (select all that apply):
- ☐ Reviewing and understanding past work
- ☐ Formulating a new question easily
- ☐ Formulating a new question quickly
- ☐ Other (Specify)----------

2 **Data View** was most useful for (select all that apply):
- ☐ Reviewing and understanding past work
- ☐ Formulating a new question easily
- ☐ Formulating a new question quickly
- ☐ Other (Specify)----------

3 **Sequence View** was most useful for (select all that apply):
- ☐ Reviewing and understanding past work
- ☐ Formulating a new question easily
- ☐ Formulating a new question quickly
- ☐ Other (Specify)----------

	Questions	Strongly Agree	Agree	Uncertain	Disagree	Strongly Disagree
Dimension View						
1	The concept of Dimension View was easy to understand					
2	Overall, it was useful to have access to Dimension View					
Data View						
3	The concept of Data View was easy to understand					
4	Overall, it was useful to have access to Data View					
Sequence View						
5	The concept of Sequence View was easy to understand					
6	Overall, it was useful to have access to Sequence View					

Bibliography

[1] Apache poi. https://poi.apache.org. Accessed: 2014-03-17.

[2] JFreeChart. http://www.jfree.org/jfreechart. Accessed: 2014-04-18.

[3] Tableau. http://www.tableau.com. Accessed: 2015-03-28.

[4] Transana. http://www.transana.org. Accessed: 2016-01-12.

[5] James Abello, Frank Van Ham, and Neeraj Krishnan. Ask-graphview: A large scale graph visualization system. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):669–676, 2006.

[6] Patrick Baudisch, Desney Tan, Maxime Collomb, Dan Robbins, Ken Hinckley, Maneesh Agrawala, Shengdong Zhao, and Gonzalo Ramos. Phosphor: explaining transitions in the user interface using afterglow effects. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, pages 169–178. ACM, 2006.

[7] Louis Bavoil, Steven P Callahan, Patricia J Crossno, Juliana Freire, Carlos E Scheidegger, Claudio T Silva, and Huy T Vo. Vistrails: Enabling interactive multiple-view visualizations. In *Proceedings IEEE Conference on Visualization*, pages 135–142. IEEE, 2005.

[8] Ken Brodlie, Andrew Poon, Helen Wright, Lesley Brankin, Greg Banecki, and Alan Gay. Grasparc-a problem solving environment integrating computation and visualization. In *Proceedings IEEE Conference on Visualization*, pages 102–109. IEEE, 1993.

[9] Egon Brunswik. *Perception and the representative design of psychological experiments*. Univ of California Press, 2003.

[10] Steven P Callahan, Juliana Freire, Emanuele Santos, Carlos E Scheidegger, Cl textbackslash 'a udio T Silva, and Huy T Vo. Vistrails: visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 745–747. ACM, 2006.

[11] Steven P Callahan, Juliana Freire, Emanuele Santos, Carlos E Scheidegger, Claudio T Silva, and Huy T Vo. Managing the evolution of dataflows with vistrails. In *Proceedings. 22nd International Conference on Data Engineering Workshops*, pages 71–71. IEEE, 2006.

[12] John M Carroll, Mary Beth Rosson, Gregorio Convertino, and Craig H Ganoe. Awareness and teamwork in computer-supported collaborations. *Interacting with computers*, 18(1):21–46, 2006.

[13] Yang Chen, Jamal Alsakran, Scott Barlowe, Jing Yang, and Ye Zhao. Supporting effective common ground construction in asynchronous collaborative visual analytics. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 101–110. IEEE, 2011.

[14] George Chin Jr, Olga A Kuchar, and Katherine E Wolf. Exploring the analytical processes of intelligence analysts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 11–20. ACM, 2009.

[15] Mei C Chuah and Steven F Roth. Visualizing common ground. In *Information Visualization, 2003. IV 2003. Proceedings. Seventh International Conference on*, pages 365–372. IEEE, 2003.

[16] William S Cleveland and Robert McGill. An experiment in graphical perception. *International Journal of Man-Machine Studies*, 25(5):491–500, 1986.

[17] Susan B Davidson and Juliana Freire. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1345–1350. ACM, 2008.

[18] Mark Derthick, James Harrison, Andrew Moore, and Steven F Roth. Efficient multiobject dynamic query histograms. In *IEEE Symposium on Information Visualization*, pages 84–91. IEEE, 1999.

[19] Wenwen Dou, Dong Hyun Jeong, Felesia Stukes, William Ribarsky, Heather Richter Lipford, and Remco Chang. Recovering reasoning processes from user interactions. *IEEE Computer Graphics and Applications*, (3):52–61, 2009.

[20] Paul Dourish and Victoria Bellotti. Awareness and coordination in shared workspaces. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*, pages 107–114. ACM, 1992.

[21] Cody Dunne, Nathalie Henry Riche, Bongshin Lee, Ronald Metoyer, and George Robertson. Graphtrail: Analyzing large multivariate, heterogeneous networks while supporting exploration history. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1663–1672. ACM, 2012.

[22] Steve Easterbrook, Janice Singer, Margaret-Anne Storey, and Daniela Damian. Selecting empirical methods for software engineering research. In *Guide to advanced empirical software engineering*, pages 285–311. Springer, 2008.

[23] W Keith Edwards, Takeo Igarashi, Anthony LaMarca, and Elizabeth D Mynatt. A temporal model for multi-level undo and redo. In *Proceedings of the 13th annual ACM symposium on User interface software and technology*, pages 31–40. ACM, 2000.

[24] W Keith Edwards and Elizabeth D Mynatt. Timewarp: techniques for autonomous collaboration. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 218–225. ACM, 1997.

[25] Stephen G Eick. Data visualization sliders. In *Proceedings of the ACM symposium on User Interface Software and Technology*, pages 119–120. ACM, 1994.

[26] George W Furnas. Effective view navigation. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 367–374. ACM, 1997.

[27] Murray Glanzer and Anita R Cunitz. Two storage mechanisms in free recall. *Journal of verbal learning and verbal behavior*, 5(4):351–360, 1966.

[28] Marti Hearst. *Search user interfaces*. Cambridge University Press, 2009.

[29] Jeffrey Heer and Maneesh Agrawala. Design considerations for collaborative visual analytics. *Information visualization*, 7(1):49–62, 2008.

[30] Jeffrey Heer, Jock D Mackinlay, Chris Stolte, and Maneesh Agrawala. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1189–1196, 2008.

[31] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *Queue*, 10(2):30, 2012.

[32] Jeffrey Heer, Frank Van Ham, Sheelagh Carpendale, Chris Weaver, and Petra Isenberg. Creation and collaboration: Engaging new audiences for information visualization. In *Information Visualization*, pages 92–133. Springer, 2008.

[33] Jeffrey Heer, Fernanda B Viegas, and Martin Wattenberg. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1029–1038. ACM, 2007.

[34] Jeffrey Michael Heer. *Supporting asynchronous collaboration for interactive visualization*. ProQuest, 2008.

[35] Petra Isenberg and Danyel Fisher. Collaborative brushing and linking for co-located visual analytics of document collections. In *Computer Graphics Forum*, volume 28, pages 1031–1038. Wiley Online Library, 2009.

[36] Petra Isenberg, Danyel Fisher, Sharoda A Paul, Meredith Ringel Morris, Kori Inkpen, and Mary Czerwinski. Co-located collaborative visual analytics around a tabletop display. *Visualization and Computer Graphics, IEEE Transactions on*, 18(5):689–702, 2012.

[37] Petra Isenberg, Anthony Tang, and Sheelagh Carpendale. An exploratory study of visual information analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1217–1226. ACM, 2008.

[38] Irving L Janis. Groupthink. *Psychology today*, 5(6):43–46, 1971.

[39] Waqas Javed and Niklas Elmqvist. Explates: spatializing interactive analysis to scaffold visual exploration. In *Computer Graphics Forum*, volume 32, pages 441–450, 2013.

[40] Greg Johnson and T Todd Elvins. Introduction to collaborative visualization. *ACM Siggraph Computer Graphics*, 32(2):8–11, 1998.

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

[42] KyungTae Kim, Waqas Javed, Cary Williams, Niklas Elmqvist, and Pourang Irani. Hugin: A framework for awareness and coordination in mixed-presence collaborative information visualization. In *ACM International Conference on Interactive Tabletops and Surfaces*, pages 231–240. ACM, 2010.

[43] Alfred Kobsa. An empirical comparison of three commercial information visualization systems. In *IEEE Symposium on Information Visualization*, pages 123–130. IEEE, 2001.

[44] Anita Komlodi and Wayne G Lutters. Collaborative use of individual search histories. *Interacting with Computers*, 20(1):184–198, 2008.

[45] Martin Krzywinski, Jacqueline Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: an information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645, 2009.

[46] Barbara H Kwasnik. A descriptive study of the functional components of browsing. In *Proceedings of the IFIP TC2/WG2. 7 Working conference on Engineering for Human Computer Interaction*, page 191, 1992.

[47] Heidi Lam. A framework of interaction costs in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1149–1156, 2008.

[48] Heather Richter Lipford, Felesia Stukes, Wenwen Dou, Matthew E Hawkins, and Remco Chang. Helping users recall their reasoning process. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 187–194. IEEE, 2010.

[49] Zhicheng Liu and Jeffrey Heer. The effects of interactive latency on exploratory visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2122–2131, 2014.

[50] Jie Lu, Zhen Wen, Shimei Pan, and Jennifer Lai. Analytic trails: supporting provenance, collaboration, and reuse for visual data analysis by business users. In *Human-Computer Interaction–INTERACT 2011*, pages 256–273. Springer, 2011.

[51] Ritch Macefield. How to specify the participant group size for usability studies: a practitioner's guide. *Journal of Usability Studies*, 5(1):34–45, 2009.

[52] Narges Mahyar, Ali Sarvghad, Melanie Tory, and Tyler Weeres. Observations of record-keeping in co-located collaborative analysis. In *Proceedings of HICSS*, 2013.

[53] Narges Mahyar and Melanie Tory. Supporting communication and coordination in collaborative sensemaking. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1633–1642, 2014.

[54] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[55] Gloria Mark and Alfred Kobsa. The effects of collaboration and system transparency on cive usage: an empirical study and model. *Presence: Teleoperators and Virtual Environments*, 14(1):60–80, 2005.

[56] E Mcgrath. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000 (2nd ed)*, pages 152–169. Citeseer, 1995.

[57] Chii Meng, Motohiro Yasue, Atsumi Imamiya, and Xiaoyang Mao. Visualizing histories for selective undo and redo. In *Computer Human Interaction, 1998. Proceedings. 3rd Asia Pacific*, pages 459–464. IEEE, 1998.

[58] Bisgard Munk, Timme, and Kristian Morkk. Folksonomy, the power law and the significance of the least effort. *Knowledge organization*, 34(1):16–33, 2007.

[59] Tamara Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[60] Phong H Nguyen, Kai Xu, Ashley Wheat, BL Wong, Simon Attfield, and Bob Fields. Sensepath: Understanding the sensemaking process through analytic provenance. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):41–50, 2016.

[61] Steven G Parker and Christopher R Johnson. Scirun: a scientific programming environment for computational steering. In *Proceedings of the 1995 ACM/IEEE conference on Supercomputing*, page 52. ACM, 1995.

[62] Peter Pirolli. The use of proximal information scent to forage for distal content on the world wide web. *Adaptive Perspectives on Human-Technology Interaction: Methods*

and Models for Cognitive Engineering and Human-Computer Interaction, pages 247–266, 2006.

[63] Peter Pirolli and Stuart Card. Information foraging. *Psychological review*, 106(4):643, 1999.

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

[65] Eric D Ragan, John R Goodall, and Albert Tung. Evaluating how level of detail of visual history affects process memory. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, pages 2711–2720. ACM, 2015.

[66] Stefano Rizzi. Collaborative business intelligence. In *Business Intelligence*, pages 186–205. Springer, 2012.

[67] Stacey D Scott, Karen D Grant, and Regan L Mandryk. System guidelines for co-located, collaborative work on a tabletop display. In *ECSCW 2003*, pages 159–178. Springer, 2003.

[68] Yedendra B Shrinivasan, David Gotz, and Jie Lu. Connecting the dots in visual analysis. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 123–130. IEEE, 2009.

[69] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, 2002.

[70] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional databases. *Communications of the ACM*, 51(11):75–84, 2008.

[71] Matthew Tobiasz, Petra Isenberg, and Sheelagh Carpendale. Lark: Coordinating co-located collaboration with information visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1065–1072, 2009.

[72] Fernanda B Viegas, Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. Manyeyes: a site for visualization at internet scale. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1121–1128, 2007.

[73] Martin Wattenberg and Jesse Kriss. Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):549–557, 2006.

[74] Ryen W White, Mikhail Bilenko, and Silviu Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 159–166. ACM, 2007.

[75] Ryen W White and Resa A Roth. Exploratory search: Beyond the query-response paradigm. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 1(1):1–98, 2009.

[76] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, 2007.

[77] Wesley Willett, Jeffrey Heer, Joseph Hellerstein, and Maneesh Agrawala. Commentspace: structured support for collaborative visual analysis. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 3131–3140. ACM, 2011.

[78] Christopher Williamson and Ben Shneiderman. The dynamic homefinder: Evaluating dynamic queries in a real-estate information exploration system. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 338–346. ACM, 1992.

[79] Gwen M Wittenbaum, Sandra I Vaughan, and Garold Strasser. Coordination in task-performing groups. In *Theory and research on small groups*, pages 177–204. Springer, 2002.

[80] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, 2016.

[81] Jian Zhao, Christopher M Collins, Fanny Chevalier, and Ranjith Balakrishnan. Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2080–2089, 2013.

[82] Yong Zhao, Michael Wilde, and Ian Foster. Applying the virtual data provenance model. In *Provenance and Annotation of Data*, pages 148–161. Springer, 2006.