Semantic Interaction for Visual Analytics: Inferring Analytical Reasoning for Model Steering

Alexander Endert

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Computer Science and Applications

Christopher L. North, Chair
Doug A. Bowman
Scott C. Leman
Richard A. May
Francis Quek

July, 10, 2012
Blacksburg, VA

Keywords: user interaction, visual analytics, model steering, visualization

Semantic Interaction for Visual Analytics:
Inferring Analytical Reasoning for Model Steering

Alexander Endert

ABSTRACT

User interaction in visual analytic systems is critical to enabling visual data exploration. Through interacting with visualizations, users engage in sensemaking, a process of developing and understanding relationships within datasets through foraging and synthesis. For example, two-dimensional layouts of high-dimensional data can be generated by dimension reduction models, and provide users with an overview of the relationships between information. However, exploring such spatializations can require expertise with the internal mechanisms and parameters of these models.

The core contribution of this work is ***semantic interaction***, capable of steering such models without requiring expertise in dimension reduction models, but instead leveraging the domain expertise of the user. Semantic interaction infers the analytical reasoning of the user with model updates, steering the dimension reduction model for visual data exploration. As such, it is an approach to user interaction that leverages interactions designed for synthesis, and couples them with the underlying mathematical model to provide computational support for foraging. As a result, semantic interaction performs incremental model learning to enable synergy between the user's insights and the mathematical model. The contributions of this work are organized by providing a description of the principles of semantic interaction, providing design guidelines through the development of a visual analytic prototype, ForceSPIRE, and the evaluation of the impact of semantic interaction on the analytic process. The positive results of semantic interaction open a fundamentally new design space for designing user interactions in visual analytic systems.

This research was funded in part by the National Science Foundation, CCF-0937071 and CCF-0937133, the Institute for Critical Technology and Applied Science at Virginia Tech, and the National Geospatial-Intelligence Agency contract #HMI1582-05-1-2001.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Visual analytics is a science based on supporting sensemaking of large, complex datasets through interactive visual data exploration [69]. The success of such systems hinges on their ability to combine capabilities of statistical models, visualization, and human intuition – with the goal of supporting the user's analytic process. Through interacting with the system, users are able to explore possible connections, investigate hypotheses, and ultimately gain insight. This complex and personal process is referred to as sensemaking [57].

Sensemaking is composed of two primary parts – foraging and synthesis. Foraging refers to the stages of the process where users filter and gather collections of interesting or relevant information, while synthesis describes stages of the process where users create and test hypotheses about how foraged information may relate. In general, foraging lends itself to more computational support, while synthesis leverages human intuition for establishing relationships between information. Thus, a goal of visual analytics is to develop visualizations that are tightly coupled with mathematical models to provide computational support for the user – integrating foraging and synthesis.

Semantic interaction [25] is an approach that enables such coupling, where analytic interactions designed for synthesis in visualizations are also designed to steer the underlying computation responsible for foraging of relevant information. Semantic interaction focuses on enabling direct manipulation of spatializations, which are two-dimensional views of high-dimensional data such that similarity between information is represented by relative distances between data points (e.g., a cluster represents a collection of similar information) [66].

ForceSPIRE (shown in Figure 1) is a visual analytic prototype incorporating semantic interaction for analysis of text document collections represented in a spatialization [25]. Semantic interactions in ForceSPIRE include repositioning documents, highlighting text, searching, and annotating documents. When users perform semantic interactions in the course of their reasoning process, the system incrementally updates a keyword weighting scheme in accordance with the user's analytical reasoning (Table 1). The learned weighting scheme emphasizes relevant keyword entities within the dataset and adjusts the layout of the spatialization accordingly. Thus, the goal of ForceSPIRE is to automatically steer the spatialization based on the user's interaction with a subset of the information. Essentially, human and computer co-create the spatial layout.

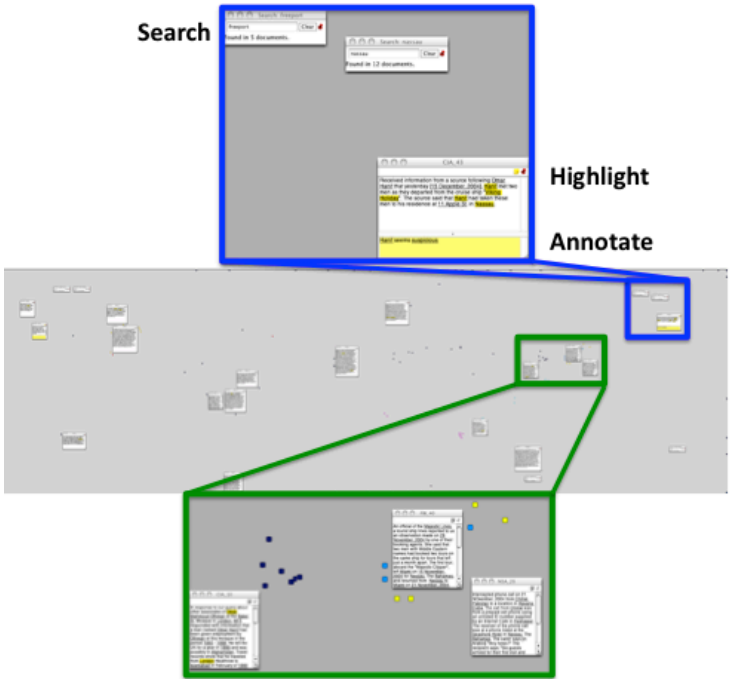


Figure 1 A scaled-down screenshot of ForceSPIRE taken on the large, high-resolution display used in this study (two zoomed in views shown). Users can search, highlight, annotate, and reposition documents spatially. Documents can be shown as minimized rectangles, as well as full detail windows.

1.1 Research Overview

Overall, the goal of this research is to *explore and analyze semantic interaction for spatializations on large, high-resolution displays*. The resulting insight can inform future research on the capabilities of the design space of semantic interaction, focused on combining the human sensemaking abilities with statistical models through usable and understandable (i.e., *semantic*) interactions. This goal can be broken down into the following research questions:

1. (**RQ1**) What semantic interactions do users perform during exploratory analysis of textual information within a spatialization?
    a. Given a spatialization where users can freely reposition text documents, what interactions occur within their analytic process?

b. What mappings exist between these interactions and the analytical reasoning?

2. (**RQ2**) How can visual analytic systems be designed to leverage semantic interaction?
    a. Defining semantic interaction.
    b. What changes must be made to the mathematical models responsible for the spatial layouts?
    c. How do these changes in design impact the current model for visualization (i.e., the visualization pipeline)?

3. (**RQ3**) How does semantic interaction impact the analytic process?
    a. How can incremental formalism extend to incremental model learning?
    b. How does semantic interaction create synergy between the insights of the user and the model's learned characteristics?

1.2 Organization

The structure of this document addresses the research questions (RQ1 – RQ3) as follows. Chapter 2 provides relevant information on related research. This includes defining how visual analytic tools have been designed to address two primary components of the sensemaking loop (*foraging* and *synthesis*) individually. Spatializations for sensemaking are also discussed, including both a technical discussion of how they can be generated via mathematical models, as well as their cognitive benefit to sensemaking. Also discussed is research on the importance of user interaction for information visualization, as well as large, high-resolution displays.

Chapter 3 addresses **RQ1**, primarily through the results of two studies, appearing in publications [5, 6, 21, 27]. The chapter focuses on the user studies exploring the methods

in which users manually leverage spatial layouts for sensemaking (including the interactions used in doing so). In Chapter 4, **RQ2** is addressed through discussing the design principles of semantic interaction and the design of the prototype, ForceSPIRE. This chapter is informed by previously published research (i.e., [25, 26, 28, 50]). An evaluation of semantic interaction in ForceSPIRE is presented in Chapter 5, addressing **RQ3**. This chapter has been published [24]. Based on these results, Chapter 6 discusses the design space created by semantic interaction. Finally, Chapter 7 summarizes the work and contributions.

Chapter 2

Background and Related Work

2.1 Foraging Tools

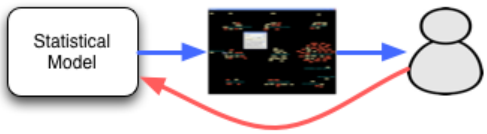


Figure 2. A model of interaction with foraging tools. Users interact directly with the statistical model (red), then gain insight through observing the change in the visualization (blue).

We categorize foraging tools by their ability to pass data through complex statistical models and visualize the computed structure of the dataset for the user to gain insight (Figure 2). Thus, users interact with these tools primarily through directly manipulating the parameters of the model used for computing the structure. As such, users are required to translate their domain expertise and semantics about the information to determine which (and by how much) to adjust these parameters. The following examples further describe this category of tools.

Visualizations such as IN-SPIRE's "Galaxy View" (shown in Figure 4) present users with a spatial layout of textual information where similar documents are proximally close to one another [74]. An algorithm creates the layout by mapping the high-dimensional collection of text documents down to a two-dimensional view. In these spatializations, the spatial metaphor is one from which users can infer meaning of the documents based on their location. The notion of distance between documents represents how similar the two documents are (i.e., more similar documents are placed closer together). For instance, a cluster of documents represents a group of similar documents, and documents placed between two clusters implies those documents are connected to both clusters. These views are beneficial as they allow users to visually gain a quick overview of the information, such as what key themes or groups exist within the dataset. The complex statistical models that compute similarity between documents are based on the structure within the data, such as term or entity frequency. In order to interactively change the view, users are required to directly adjust keyword weights, add or remove documents/keywords, or provide more information on how to parse the documents for keywords/entities upon import.

Similarly, an interactive visualization tool called iPCA uses Principal Component Analysis (PCA) to reduce high-dimensional data down to a two-dimensional plot, providing users with sliders and other visual controls for directly adjusting numerous parameters of the algorithm, such as individual eigenvalues, eigenvectors, and other components of PCA [39]. Through adjusting the parameters, the user can observe how the visualization changes. This allows users to gain insight into a dataset, given they have a thorough understanding of PCA, necessary to understand the implications behind the changes they are making to the model parameters.

Alsakran et al. presented a visualization system, STREAMIT, capable of spatially arranging text streams based on keyword similarity [3]. Again, users can interactively explore and adjust the spatial layout through directly changing the weight of keywords that they find important. In addition, STREAMIT allows for users to conduct a temporal investigation of how clusters change over time.

2.2 Synthesis Tools

Figure 3. A model of interaction with synthesis tools. Users manually create a spatial layout of the information to maintain and organize their insights about the data.

Synthesis tools focus on allowing users to organize and maintain their hypotheses and insight regarding the data in a spatial medium. In large part, this is done through presenting users with a flexible spatial workspace in which they can organize information through creating spatial structures, such as clusters, timelines, stories, etc. (Figure 3). In doing so, users externalize their thought processes (as well as their insights) into a spatial layout of the information.

For example, Analyst's Notebook provides users with a spatial workspace where information can be organized, and connections between specific pieces of information (e.g., entities, documents, events, etc.) can be created. Similarly, The Sandbox [75] enables users to create a series of cases (collections of information) which can be organized spatially within the workspace.

From previous studies, we found cognitive advantages associated with the manual creation of a spatial layout of the information [5]. By providing users a workspace in which to manually create spatial representations of the information, users were able to externalize their semantics of the information into the workspace. That is, they created spatial structures (e.g., clusters, timelines, etc.), and both the structures as well as the locations relative to remaining layout carried meaning to the users with regards to their sensemaking process. Marshall et al. have pointed out that allowing users to create such informal relationships within information is beneficial, as it does not require users to formalize these relationships [47].

From this related work, we believe a trend is emerging in how interaction is currently handled in many visual analytic systems where complex statistical models are used – users are required to *go outside of the metaphor*. That is, while the visual representation given to users is spatial, the methods of interaction require users to step outside of that metaphor and interact directly with the parameters of the statistical model using visual controls, toolbars, etc.

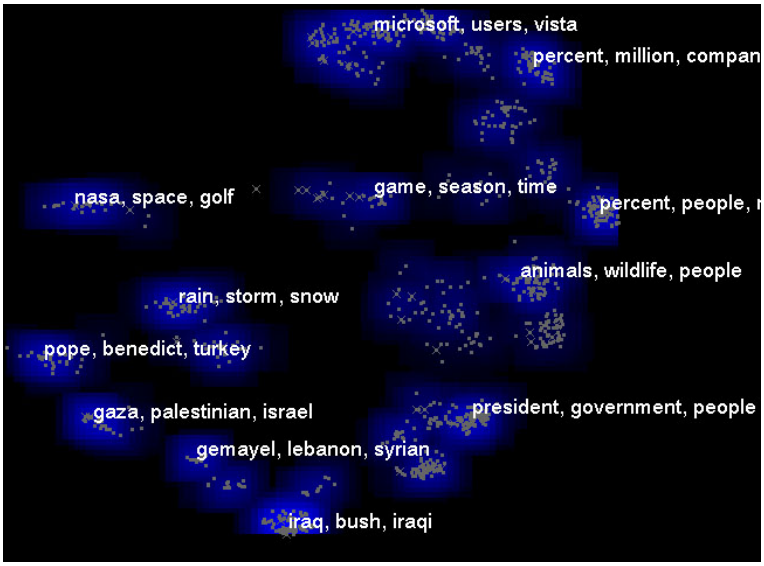


Figure 4. The IN-SPIRE Galaxy View showing a spatializtiation of documents represented as dots. Each cluster of dots represents a group of similar documents.

There has been some work in providing more easy to use interactions for updating statistical models. For example, relevance feedback has been used for content-based image retrieval, where users are able to move images towards or away from a single image in order to portray pair-wise similarity or dissimilarity [72]. From there, an image retrieval algorithm determines the features and dimensions shared between the images that the user has determined as being similar. We view this as one example where the interaction stays in the spatial metaphor of the visualization.

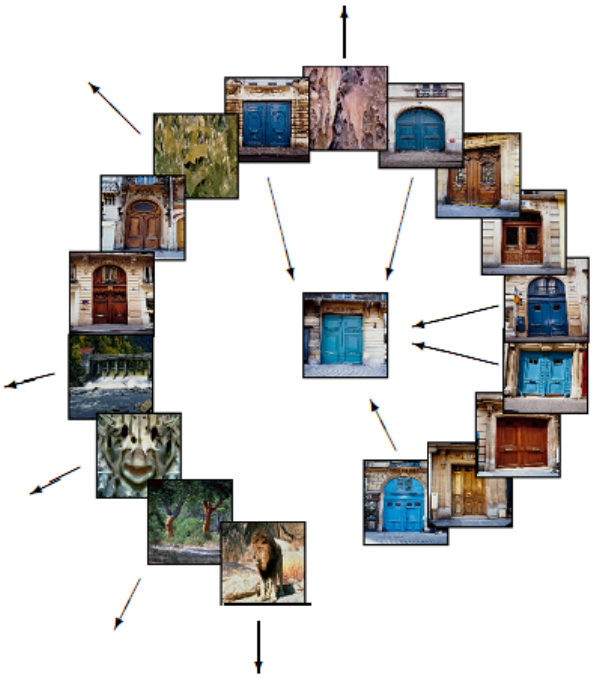


Figure 5. A relevance feedback model example. Users move images arranged in a circle towards or away from the image in the middle to signify similarity or dissimilarity. Figure originally published in [61]. Used under fair use guidelines.

Also, spatializations of document sets exist that allow users to place "points of interest" into the spatial layout. In VIBE, users are allowed to define multiple points of interest in the spatial layout that correspond to a series of keywords describing a subject matter of interest to the user [52]. Similarly, Dust & Magnet [79] allows users to place a series of "magnets" representing keywords into the space and observe how documents are attracted or repelled from the locations of these magnets. Through both of these systems, users can interact in the spatial metaphor through these placements of "nodes" representing keywords. However, the focus of semantic interaction is on interacting with data (i.e., documents), an important distinction discussed in the following section.

From the sensemaking loop presented by Pirolli and Card [57], we learn that in intelligence analysis, that analytic process consists not only of the information that is explicitly within the dataset being analyzed, but also the domain knowledge of the analyst performing the analysis. It is through this domain knowledge that analysts interact and explore the dataset to "make sense" of the information. Thus, we believe this interaction

(and the domain knowledge associated with it) is equally important as the raw data, and must be incorporated into the visualization by tightly coupling the model with the interaction.

From this body of work, we most notably come away with an understanding that

1) analysts fundamentally understand the spatial metaphor used in many spatial visualizations
2) many of these systems are constructed using complex mathematical algorithms to transform high-dimensional data to two dimensions, and
3) in most cases these algorithms can be controlled by analysts largely through visual controls (e.g., sliders, knobs, etc.) to directly adjust parameters of the algorithms, updating the spatial layout.

2.3 Spatializations and Clustering Algorithms

Several algorithms exist with a similar purpose of mathematically generating two-dimensional layouts from which users can interpret important information about a dataset. In general, algorithms group or organize data based on similarity, which is a function of the features of the dataset. Dimensionality reduction algorithms can provide a 2-d spatial visualization of the clustered data. For example, algorithms like self-organizing maps [66] or generative topographic mapping [41] provide a direct method of visualizing text data spatially, but do not provide explicit cluster membership information. A survey of clustering algorithms can be found in [77] and is outside the scope of this paper. The primary criteria upon which these models generate layouts are structure extracted from the dataset, such as term frequencies, temporal attributes [2], etc. from textual datasets [74].

However, we contend that in order for these algorithms to aid users during the entire process of sensemaking, their design needs to change from focusing primarily on the structure of the data to combining this structure with semantics derived from the user.

2.4 User Interaction in Visualization

One of the challenges for information visualization is to gain a deeper understanding of how users interact within visualization, and more importantly how these interactions are integrated into their analytic process [56] (e.g., creating "fluid interactions" that emphasize not impeding with or hindering the flow of the analytic process [19]). Yi et al. have addressed this lack of understanding by presenting an extensive categorization of user interactions available in popular exploratory visualization tools [78]. However, interaction in visualization has been shown to be inherently complicated to categorize [12].

Dou et al. have shown that through logging user interactions in a visualization of financial data, low-level analytical processes can be reconstructed [13, 17]. Most importantly, these results indicate that a detectable connection exists between the low-level user interaction and the analytic process of that user.

Our work described in this paper addresses a related topic area – analyzing the relationships between the spatial layouts users create while exploring a dataset, and investigating how the user interactions within that process can be correlated to the solution users generated. We discuss how these findings can extend to help enhance the effectiveness of clustering algorithms.

2.5 Large, High-Resolution Displays

The term "large, high-resolution display" can be defined a number of ways. First, it can be defined in terms of the technology being used (i.e., "larger than a traditional display"). This definition is subjective, as it implies that one's perception of the display itself is what defines it. As technology advances over time, and what may be considered a "traditional display" changes, displays that were once considered large (and high-resolution) by this definition may no longer be. A second way to define the term is in terms of quantity of data that it can visually represent, or perhaps more importantly, the ability to represent multiple views, scales, and "units" of data (e.g., documents, web-pages, etc.). However, this definition requires that we define what a unit of data is, which changes for each application or dataset.

While both of these definitions can be useful, we prefer to define the term "large, high-resolution display" as being a display that is *human scale* [7, 20]. By human scale, we mean that the display's size and resolution are closely matched to the sphere of perception and influence of the human body. In a practical sense, we use the label to describe displays whose combined size and resolution approach or exceed the visual acuity of the user. Displays at this scale afford the user the opportunity to trade virtual navigation for *physical navigation* (turning, leaning, moving around), thus allowing the user to exploit embodied human abilities such as spatial awareness, proprioception, and spatial memory [8]. Physical navigation can be exploited for interaction purposes, such as synchronizing cursor location and user focus based on chair rotation [23], adjusting the selection scale based on distance to the display [55], and others. Similarly, designing visual representations (or glyphs) for such displays can greatly impact performance [22, 80]. This tipping point is important because it heralds a change in user behavior, and thus presents on opportunity for how interactions can be designed for visualizations where the emphasis is on the spatial nature of the information.

For the purpose of this work, a large, high-resolution workstation, such as the one shown in Figure 6, is used. Users performing analysis on such workstations have been shown to

emphasize the spatial nature of the large virtual space [5, 27]. When performing tasks on such workspaces, users place information (i.e., documents) in specific locations, ultimately constructing a spatial layout. The added display space provided by such workstations allows users to emphasize the spatial relationships between information, and thus make for a great platform upon which to explore and design spatializations.

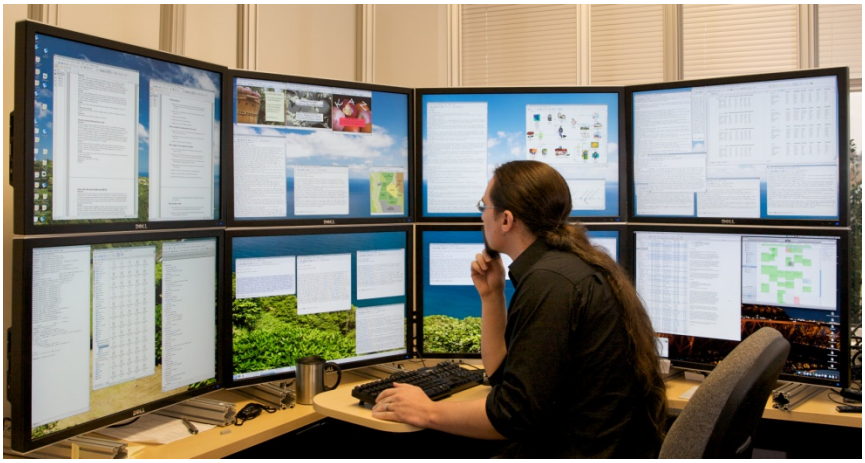


Figure 6 A Large, High-Resolution Workspace (33 MPixels) allows users to refer to information spatially. Photo by Christopher Andrews, 2010. Used with permission.

Chapter 3

Understanding Spatial Sensemaking

In order to understand how users interact with a spatialization for exploratory data analysis, we performed the following study.

In this chapter, we present the results of a user study exploring these two challenges. We asked users to perform a spatial sensemaking task on a large, high-resolution display using a prototype visualization, LightSPIRE (shown in Figure 8), that provides basic text analysis functionality (i.e., searching, highlighting, annotating, and document positioning). The tool provides only manual layout capabilities, with no algorithmic layout support. Users can manually spatially organize the documents however they desired to help them complete their analysis. We then analyzed the user-generated clusters in each user's spatial layout to better understand the semantics of their clusters, as well as analyzed their process in terms of the interactions.

Based on the findings of the user study, the contributions of this work are as follows. First, we discuss how the criteria on which users clustered information for sensemaking was not necessarily based only on structure found within the data. Users created clusters (and other spatial constructs) based on a combination of the structure within the data (e.g., the entities in the text), as well as their intuition and higher-level concepts. Second,

we analyze the user interaction during their analytic process to show how certain interactions indicate important discriminating features in the data. Third, we discuss how these findings can influence the design of statistical models created for interactive visual analytics.

3.1 Leveraging Space for Sensemaking

Visualizations exist that aid users in sensemaking by allowing them to manually organize information spatially. The cognitive benefits of allowing users to generate spatial layouts of information have been studied. For instance, Marshall and Rogers [48] found that users prefer to create implicit relationships between information by positioning related information closer together. They found that the ease, flexibility, and informality associated with creating these relationships spatially were important to users.

From Andrews et al., we learn that users externalize semantic information about a dataset into the layout and organization of documents [5]. The spatial layouts created represent specific meaning about each individual user's analysis. Therefore, user's findings from their analysis task were present in their spatial layouts. This study extends on this work by quantifying these relationships in terms of the clusters generated, as well as the interactions utilized during the processes.

Pirolli and Card present a model for sensemaking for intelligence analysis task [57]. This model (shown in Figure 7) illustrates the series of cognitive stages users proceed through when performing a sensemaking task. Most notably, we learn from their work that much of the success of sensemaking is based on the ability for humans to combine their domain expertise gained from previous experiences and the information from the dataset they are currently investigating. It is through this combination that they are successful in identifying complex relationships within the data and ultimately gaining insight.

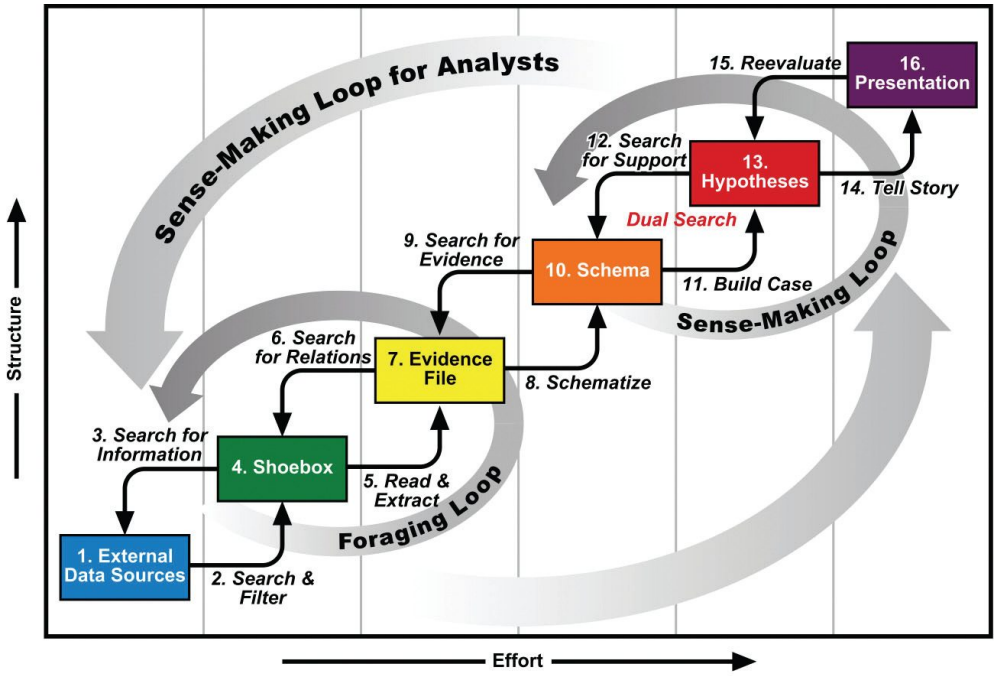


Figure 7 The sensemaking loop, modeling the cognitive stages for individual intelligence analysis (adapted from [57]).

3.2 Method

The purpose of this study is to analyze users' spatial clustering of information to aid with their sensemaking task. Participants in the study analyzed a textual dataset to understand and uncover a fictional terrorist activity using a simple spatial document organizational tool called LightSPIRE. We chose this task and dataset, as it is representative of intelligence analysis tasks that are largely focused on sensemaking.

This study explores two primary questions:

> *Analysis of Spatial Layout*. What structure exists within the user-generated clusters? That is, given the clusters created by the users, what structure can be algorithmically detected?

Analysis of Process. What can we learn from users' interactions during the analytic process that can help guide algorithms? What indicators of analytical reasoning can be derived from these interactions?

3.2.1 Equipment

Users of the study were given a spatial document organization tool, LightSPIRE, for their task. LightSPIRE (shown in Figure 8) provides a workspace where documents can be manually organized using basic, familiar interactions. The primary interaction afforded by LightSPIRE is movement of the documents. Users also have the choice to view documents using two levels of detail, full-text and filename only (documents could not be deleted). A search function allows users to query the dataset for text strings. Search hits are shown within the documents as permanent, green highlights. The documents that contain the current search query are shown in a darker red until another search is performed or the search is cleared. Users also have the ability to highlight text (in yellow) as they are reading the documents. LightSPIRE captures and logs all of these interactions for post-study analysis.

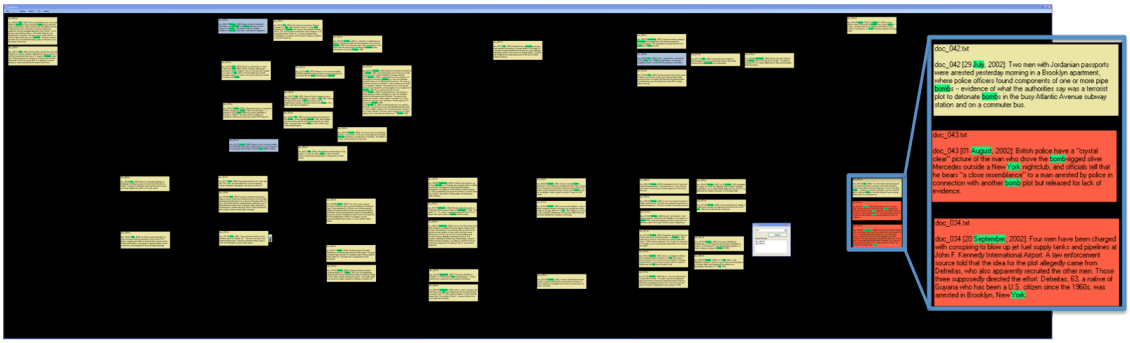


Figure 8. LightSPIRE, a large-display spatial workspace used in this study for organizing text documents. We observed users spatially analyzing a textual dataset, using common interactions such as searching, highlighting, and positioning documents.

The workstation used for this study is a large, high-resolution display (LHRD, constructed using ten 17" LCD monitors arranged in a 5x2 grid (total resolution: 6400 x 2048, or 13.1 megapixels), curved around the user to provide optimal access to all areas of the workspace [65]. The display is driven using a single workstation running Windows

XP, thus allowing familiar mouse and keyboard interaction with the workspace. Using LightSPIRE on this LHRD, users gain the ability to display the entire dataset in full-text if desired, as well as create an environment where spatial location of the information conveys meaning to the user [5]. Users were also given access to a whiteboard and a notepad for notes, although no users made use of these.

3.2.2 Dataset

The intelligence analysis training dataset used for this study consisted of 50 textual documents containing a hidden fictitious terrorist plot. The dataset includes a known ground truth, and includes a scoring rubric to assess the findings of each user. It also includes a list of "important" documents (22 out of 50) that are relevant to supporting the solution. Thus, we are able to draw conclusions on effectiveness of the solutions based on the scoring rubric, and analyze the interactions and spatial layouts based on which documents are important to the solution.

3.2.3 Procedure

Users were given practice with the workspace and LightSPIRE prior to beginning their analysis. During this time, all the functionality of LightSPIRE was shown to them, and they were able to ask any questions. Then, they were given instructions to analyze the dataset to uncover any suspicious activity, gathering as much information to support (and refute) their hypothesis as possible. No information was given regarding the important and unimportant documents. They were informed of the one-hour time limit for their analysis, after which they would be asked a series of questions about their solution. During this post-task questionnaire, the workspace would remain visible, but they would not be allowed to interact with it (other than looking at it and reading). This semi-structured interview provides users the ability to explain their solution in as much detail as possible, then goes on to ask details about relationships between people, places, and events to determine how well users could uncover these complex relationships during their investigation. Finally, we asked users to sketch (on a blank piece of paper) a

drawing to identify and label their clusters, and help us better understand the meaning of the layout they created. The entire duration of the study lasted approximately two hours.

3.2.4 Data Collected

LightSPIRE was designed to log all of the user interactions, including search terms, cursor movement and activities, document movement and positioning, and document opening and closing. From these logs, we can analyze the users' process at the interaction level. In addition, screenshots of the entire workspace were taken at 10-second intervals. The screenshots allow us to analyze the clusters and spatial layouts generated by the users. A description of the spatial layout was made through the sketch produced by each user at the end of the study, where clusters and other spatial constructs were clearly labeled. The entire study was video recorded primarily to capture the conversation between the user and the investigator, as well as capture any gestures made towards the workstation during the post-study questionnaire.

3.2.5 Participants

This study consisted of observing 15 users. The users were all male, undergraduate computer science students. While these participants had no prior training in intelligence analysis, the domain expertise required to correctly solve the dataset is basic intuition and reasoning. The participants were offered an opportunity to receive one of three monetary prizes of $50, $35, and $25 for the top three most accurate and complete solutions (based on the scoring rubric provided with the dataset) to provide motivation for their task.

3.3 Results

The results of this study are presented as follows. First, we analyze the final spatial layout the users created. We analyze the spatial layout produced by each user to gain a better

understanding of the structure of the user-generate clusters. Second, we analyze the user interactions during the users' processes of creating these layouts.

3.3.1 Analysis of Spatial Layout

The initial layout of the 50 documents was identical for each user. Each of the documents were minimized (showing only the filename), and arranged based on their filename (i.e., doc_01, doc_02, etc.) in the top left corner of the workspace. Each document was present only once in the workspace.

3.3.1.1 Primary Spatial Layout

What are users' overall layout strategies? The analysis of overall spatial layout reveals three distinct patterns of how users chose to spatially organize their information.

Topical Clustering. Nine out of the fifteen users in this study chose to organize their workspace based primarily on creating clusters of topically related documents. Figure 8 shows a representative example of a workspace organized by clustering. Users organized information into clusters to synthesize their hypotheses. For example, at times users labeled their clusters "Aryan activities" to represent a cluster that was focused around the documents within the dataset that relate to that information. However, users also created clusters labeled "junk" or "related but not big picture", indicating that clusters can also represent forms of insight about the dataset.

Temporal Clustering. Five of the fifteen users organized their workspace based on the temporal information in the documents. These users arranged their information from left to right based on the dates included with each document (see Figure 9). For this group, the users chose to place no relevant information on the y-axis of the workspace. When asked, one user replied that "[he] used the vertical dimension of the display to make room to fit documents if the dates overlapped". For example, one user outlined an area of the workspace and labeled it "August". We classify each one of such areas as a "cluster" for the purposes of this work.

Hybrid Clustering. One user generated a particularly interesting layout (shown in Figure 9). He started his investigation by organizing the documents based on a timeline on roughly the top half of the workspace. Then, he began investigation the relationships and interesting events within the dataset. As he found interesting terms or events, he pulled these documents out of the timeline and clustered them in the lower portion of the display. However, the documents retained their relative temporal positioning, as he took caution to only move the documents vertically, so as not to disturb the temporal left-to-right organization. As a result, we noticed this user balanced a tradeoff of maintaining temporal awareness of the documents, as well as gaining an understanding of the important events and topics within the dataset by establishing "rows" of related items.

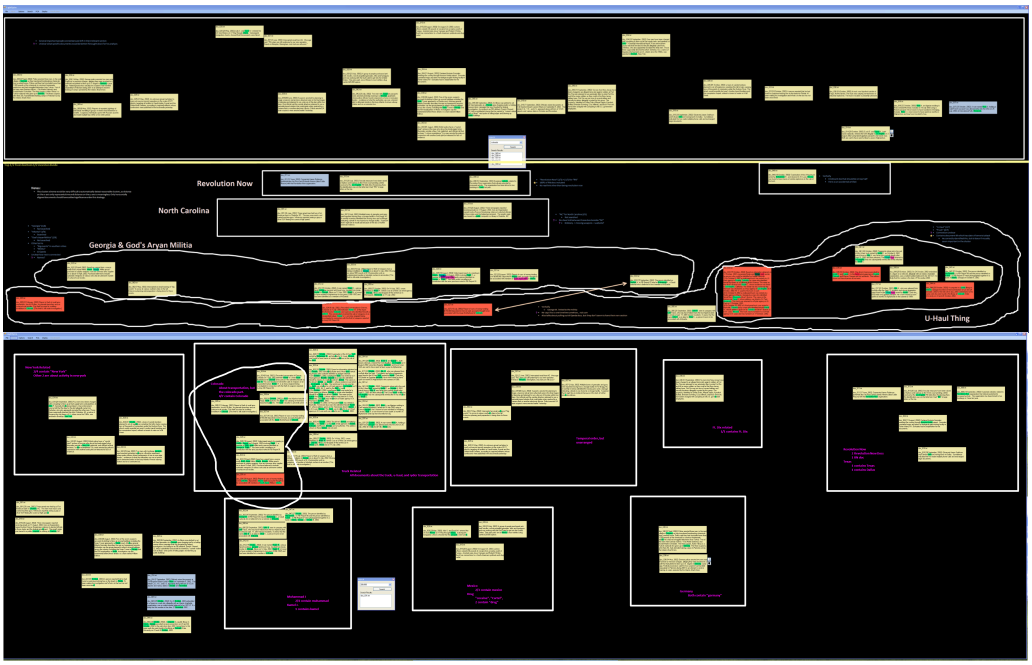


Figure 9 Annotated screenshots of two final layout states. The annotations (white frames and purple text) were added by the investigators based on the cluster boundaries and labels provided by the post-task interviews. (Left) shows an example of the Hybrid Clustering spatial layout, where the user organized the documents temporally from left to right, while the separation along the y-axis was used to organize topics of interest. (Right) is an example of the Topical Clustering layout, where the user chose to organize the documents in clusters based on topics important to the solution.

3.3.1.2 Cluster Structure

We analyze the raw clusters created by each user during the task, and later identified in their post-task interview. The 15 users created a total of 86 clusters. The number of documents contained in each cluster ranged from 1 to 25 documents, with a mean of 7.3 documents per cluster.

How do documents within a cluster relate to each other?

Intra-cluster Co-occurrence First, we analyze if one or more terms occurs in all the documents within a cluster. 26 of the 86 clusters (30%) had at least one term in common among all the documents in the given cluster. 10 out of the 15 users made these clusters containing common terms. 13 of these 26 clusters had a month as one of the common terms (these clusters belonged primarily to the *Temporal Clustering* users). As can be expected, for clusters of smaller sizes, there were more shared common terms. Only 5 of these 26 clusters contained more than two documents. For these 5 clusters with more than two documents, the number of common terms never exceeded four. Hence, for the remaining 70% of clusters the structure of the clusters is not based on any co-occurring terms in all the documents.

Transitivity An alternate but simplistic explanation of cluster structure is that pairs of documents within a cluster are related via terms that are common between them (i.e., clusters represented as connected graphs where nodes are documents and edges represent shared entities between the two documents). Therefore, any two documents within the cluster can be connected transitively via one or more other documents. We refer to such a cluster as a *transitive* cluster. For example, one user created a cluster with three documents in which one pair of documents did not have any words in common (shown in Figure 13). However, a pair of documents shared the term "Arrested*"* and another pair shared the terms "Cartels*"* and "Drug" and a transitive relationship between the documents in the cluster can be given by:

doc_39(Arrested)→ doc_15(Arrested, Cartel)→doc_28(Cartel, Drug)

Hence, while these three documents produce a connected graph, they do not share a common term between all three. The abstraction of a large corpus of text documents as a similarity network (the notion of similarity being induced by terms that are shared between document pairs) has been used by [36, 37] in a "Storytelling algorithm" to connect seemingly unrelated documents via a path referred to by the authors as a *story*. While the ordering of documents in the transitive relationship between two documents might bear some semantic meaning to the users, we do not account for ordering in our analysis (i.e., our graphs are undirected).

Based on this, 71 of the 86 clusters (83%) are transitive, excluding temporal information. We chose to exclude the temporal information of the documents for these connections, as the month names occur frequently throughout the dataset, creating large connected groups based on solely this information.

Transitive Terms We analyze the terms that cause links between documents within the cluster to determine which terms cause the transitivity. We call these terms *transitive terms*. Our goal is to understand the distributional properties of the transitive terms, and how often they occur within the cluster compared to occurring in the remaining dataset. The first statistic we look at is the proportion of documents in which the transitive term occurs. We observe that the proportion of documents with a transitive term within the user-generated cluster is 20% higher, on average, than the proportion of documents outside the cluster that contain the term ($t(2442) = -46.50$, $p < .0001$). Transitive terms have very low rates of occurrence outside of their clusters, and in some cases the only occurrences are within the single cluster.

3.3.1.3 User-Generated Cluster Labels

How do documents within a cluster relate to the cluster label? To determine if the labels can provide an indicator as to which terms within the cluster are important, we compare the user-generated cluster labels to the content of the documents within the cluster. For example, for a cluster named "Germany and Trucks", we extract the entities "Germany" and "Trucks". Then, we analyze the percentage of documents within the cluster that

contain the word “Germany” and the percentage containing “Trucks” (case insensitive and stemming). We report on the highest percentage of these, as we are not concerned with choosing an entity from the label that best represents the cluster of documents (addressed by work such as [60]). Rather, we present the results of how well the best-matched entity within a label matches the entities of the documents within a cluster.

The percentages of documents within each cluster that contain the best-matching entity from the label are shown in Figure 11. These results show that 12 of the 86 clusters (14%) can be characterized based on a single entity extracted from the user-generated label (i.e., 100% of the documents in the cluster contain the given entity). 10 of them are clusters of two or fewer documents (shown in Figure 11). Whereas 67 of the 86 clusters (78%) do not contain the given entity in more than 50% of the documents within the cluster.

Additionally, the users who chose temporal clustering to organize their workspace still showed inconsistencies between their temporal clusters and the documents actually contained in those clusters. For example, one user was very careful to maintain a relative ordering between documents based on the date included in each document. However, analyzing his layout, this ordering did not hold true for the majority of the layout. Another user chose to cluster the information through a broader temporal criteria (i.e., he clustered based on the months the documents occurred). However, 3 of his 5 clusters contained documents from months other than the month with which he labeled the cluster.

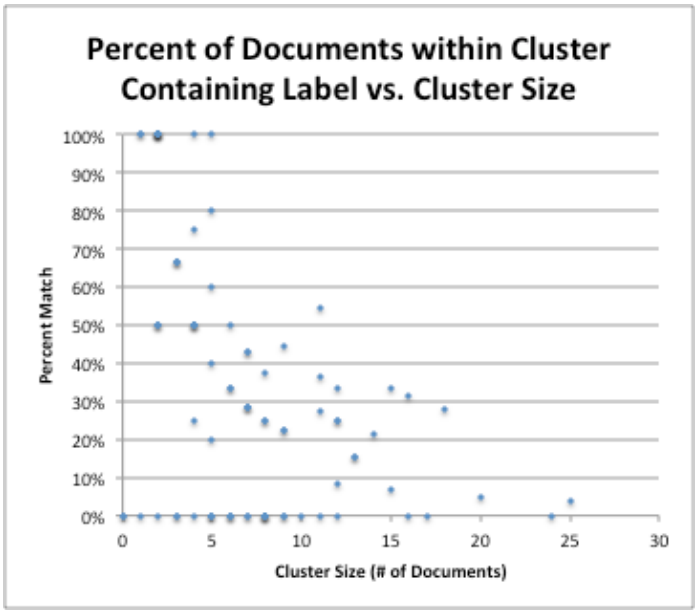


Figure 10 The size of a cluster compared to the percentage of documents within the cluster that contain the user-generated label. Notice that only clusters of 5 or fewer documents match 100%.

From these results, we confirm our hypothesis that users form clusters not solely based on entities or keywords within the data. Cluster labels such as “important people”, “unknown”, “events that have happened”, “random unrelated events”, “miscellaneous”, “terrorist activity timeline”, “big events in southern cities”, etc. indicate they are based on higher-level or process-oriented concepts. Further, we found that users struggled to answer what is the meaning of their clusters. This could be because clusters were created based on implicit and informal relationships perceived by the users (as described in [48]). Thus, asking users to formalize these relationships proved challenging.

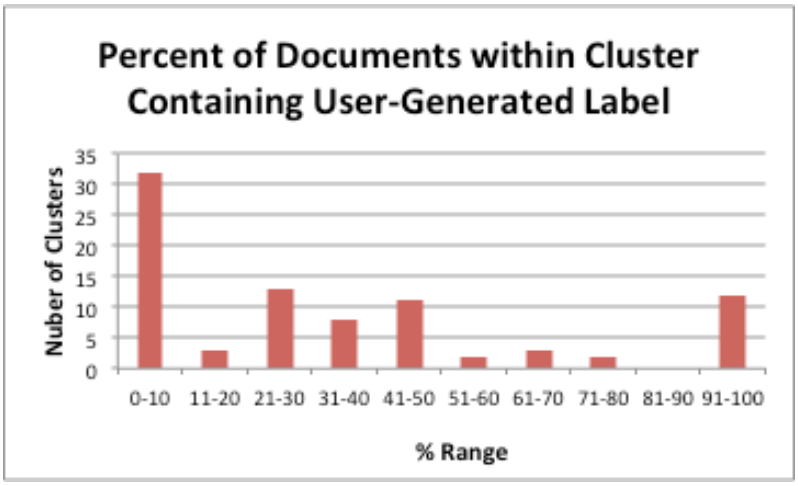


Figure 11 The distribution of the percentage of documents within each cluster that contain the cluster label keywords. Of the 86 user-generated clusters, 28 clusters did not have their label keywords present in any of their documents. 67 clusters do not contain the label keywords in more than 50% of the documents.

3.3.2 Analysis of Process

How can interactions provide effective discrimination of relevant structure? We analyze each user's analytic process in terms of the user interactions performed in LightSPIRE. Our goal is to gain a better understanding of how each interaction is used during the sensemaking process, and how models might exploit these interactions as a means for unobtrusively capturing information from the user about important discriminating features of the data.

3.3.2.1 Search

Search is a frequent operation in text analytics. Performing a search in LightSPIRE returns results visually within the layout. That is, documents containing the search result change color to red until the search is cleared. Even after the search is cleared (or another search is performed), the text matching the search query within the documents stay highlighted in a neon green. We divide the use of search into two categories: *constructive* and *awareness*. Constructive search indicates that the results of the search were used to create a cluster, whereas awareness search was performed to highlight where in the layout a term occurs.

Users performed a total of 2263 searches (broken down by user in Figure 12), 207 of which were constructive (9%), and 2056 of which were awareness (91%). A total of 326 unique terms were used in the search. Thus, many were repeated, as evidenced by the high number of awareness searches performed. Of these search terms, 222 contained a one word, 100 contained two, and only 4 were three words in length.

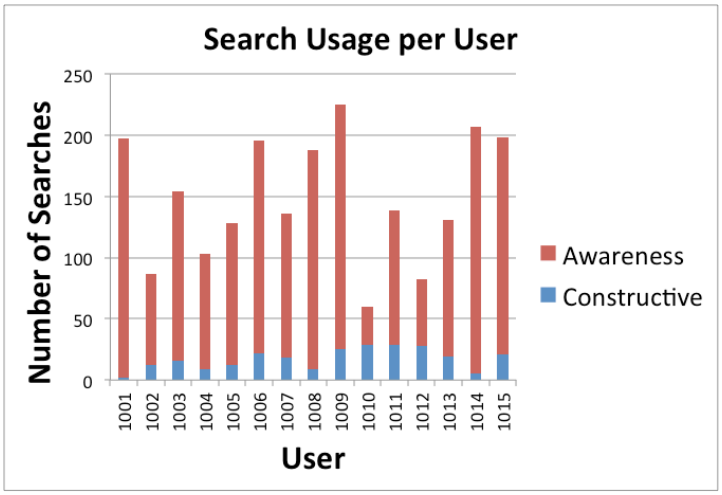


Figure 12 Users performed searches during their investigation for two reasons: constructing clusters (constructive), or to recall where the search term appears in the spatial layout (awareness).

A constructive search consisted of performing a search, then creating a cluster based on the documents in which the search term appeared. For example, one user found the term "u-haul" interesting while reading a document. He proceeded to search on this term, found that it appears in other documents, and dragged each of these documents to a location to "construct" a cluster.

This usage pattern for search might initially indicate that clusters are formed as a result of search terms, and therefore can be classified by a collection of entities. However, the structure of the clusters often changed during the investigation as the user gained more insight into the dataset. Clusters changed from their initial creation based on an entity (e.g., the "u-haul" cluster, containing only documents containing that entity), to a collection of documents whose connection or similarity is not based on that particular entity (e.g., the "transportation of suspicious material" cluster). This is evidence of incremental formalism [63].

Search can provide a good indicator as to what documents are important. We analyzed all the search hits (i.e., a document containing the search term is considered a search hit), and with the list of important documents provided with the dataset, found that the average number of times an important document was hit was higher than the non-important documents (Figure 14). The average number of times an important document was hit by each user is 14.2 times, compared to 6.9 times for non-important documents ($t(28) = 4.47$, $p < .0001$).

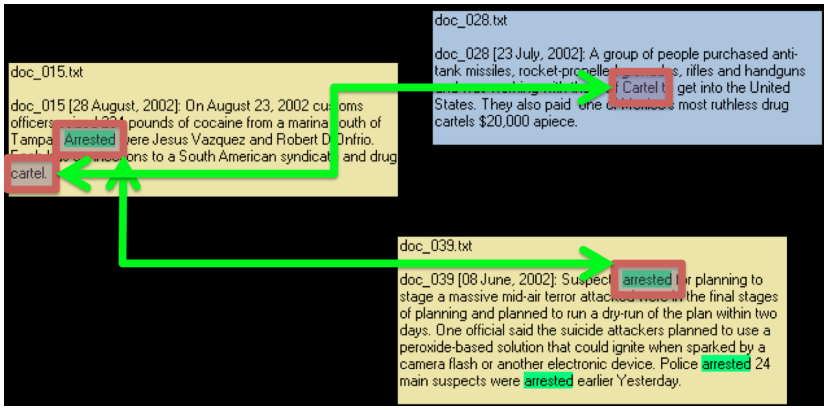


Figure 13 Example of a cluster that can be described by transitive relationships (shown by arrows). While a single term is not present in all three documents, we can form transitive connections between the documents via the terms "arrested" and "cartel".

3.3.2.2 Highlighting

Analysts frequently highlight information while reading. LightSPIRE allows for two types of highlighting. When users perform a search, the text within each document that contains the search term is highlighted green. Also, users can perform a standard yellow highlight of a phrase within a document using their cursor.

The design decision to create persistent highlights from search terms stemmed from the user feedback from a previous study [5], where the users mentioned that creating highlights within documents served as a means for not only marking important information within the documents, but also created non-uniform visual representations of these documents. That is, the highlights served as a way to transform the documents into visual glyphs, as the pattern of highlights within a document was meaningful to the user.

9 of the 15 users made use of standard highlighting, while 6 used only the highlights from search.

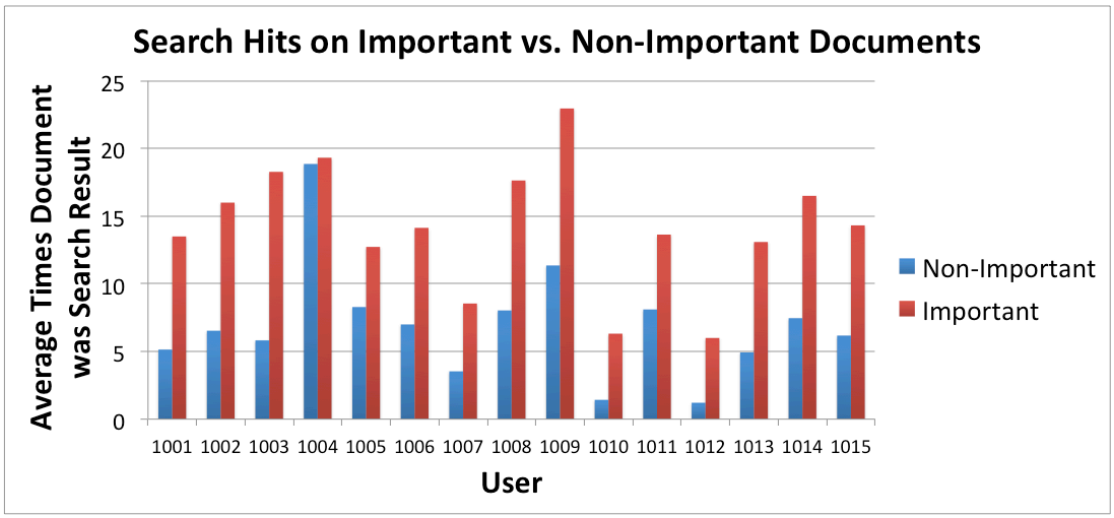


Figure 14 Comparison showing how often important documents were search result hits compared to non-important documents.

We found these two types of highlighting were used to indicate relevance at two different scales. Search terms were more concise indications of terms or entities that the user found interesting and relevant. This is evidenced by the analysis of search term length, showing that users searched mostly to find single words. In contrast, the standard form of highlighting was used to indicate broader portions of documents as important (e.g., sections or phrases). Users performed a total of 220 highlights, containing an average of 5 words per highlight (sometimes spanning entire sentences). One user even chose to perform a standard highlight spanning an entire document that he referred to numerous times, and wanted to "find [the document] more easily".

We analyze the standard highlights with respect to the cluster labels to determine if the labels match to the highlights. Only 9 of the 86 cluster labels contain entities that were highlighted by the users in the documents. This shows that while highlighting can indicate content relevant to the user, cluster structure is more complex.

3.3.2.3 Document Movement

Being a spatial workspace, one of the most predominant user interactions is the movement of documents to position (and re-position) documents throughout the analysis. As expected, users positioned documents within their workspace as a means of externalizing insights about the datasets [5]. However, in this study we are more interested in what information we can quantify about this interaction regarding the user's analytic reasoning.

The analysis of movement was performed based on the number of times a document was moved, and the average distance each document traveled per move (in number of pixels). Important documents were moved an average of 7.1 times, compared to 5.7 times on average for non-important documents ($t(28) = -1.63$, $p < .05$). While important documents were moved more frequently, their moves were more local, indicated by the average path length (in pixels) the document traveled each time it was moved. An important document traveled an average distance of 654 pixels per move, compared to an average of 792 pixels for non-important documents ($t(28) = -1.65$, $p < .05$). Thus, important documents were moved 25% more times, but 17% less distance per move.

Documents displayed in full detail versus the smaller, minimized views reveals a metric for discriminating between important and non-important documents. Given the added resolution and size of the display used, 12 out of 15 users chose to maintain all documents in full detail. The three who minimized some documents only did so for un-important documents.

While these metrics were statistically significant, the most notable difference between important and non-important documents in terms of movement were seen through the observations and post-task interviews. We observed that the important documents served as spatial landmarks for the users. That is, these documents anchored a concept to a specific location in the workspace, from which the remaining layout crystalized. The typical behavior observed for moving important documents was to perform one large movement to position the document in the workspace, with many future short movements

to refine the information within the cluster. In contrast, users quickly deemed non-important documents as irrelevant, placing them in such a cluster (e.g., "junk"). Other times, users did not refine the positioning of these documents within a cluster, but rather repositioned them into new clusters, often distant from the previous positions.

3.4 Discussion

The results of this study reveal new opportunities in the area of statistical models designed for co-creating spatial layouts. We initiate a challenge to statistics and data mining researchers to design models to support the interactive sensemaking process. First, designers can use the structure we analyzed from the layouts users created to design algorithms that better mimic users' clusters. For example, transitivity is a good metric in that it successfully extracted structure from the user-generated clusters. Therefore it could provide a good metric for use in spatial layout algorithms. In contrast, algorithms based on strict term co-occurrence between documents are not likely to coincide well with user's mental models. Algorithms can be designed to support the three layout strategies observed. To support incremental formalism, models can evolve from term co-occurrence to more complex metrics over time, such as transitivity.

Second, the user interactions present in the spatial sensemaking process can be used to guide models during the analytic process for co-creation of the spatial layout. For example, algorithms can observe and incrementally respond to the process of users clustering data. When users perform sensemaking, they gain understanding of the data at a higher level. Models must be able to co-create clusters based on these higher-level concepts. These concepts are based not solely on term co-occurrence, transitivity, or other metrics, but incorporate the user's reasoning. The user interactions can serve as cues to help models understand these higher-level concepts.

For example, models can expand the "data" upon which these models calculate their similarity measures – broadening the scope of the distance metric. These models should incrementally adapt based on the interactions of the user throughout the analytic process. To do so, models must be based not solely on the *hard data* (i.e. the structure within the dataset), but also the user's reasoning derived from interaction (i.e., *soft data*). Soft data is defined as a captured and interpreted representation of a user's semantic knowledge regarding a dataset [25].

As evidenced by the results of this study, the user-generated layouts are often based on information that is *outside the scope of the hard data*. For instance, the user-generated cluster labels do not always map directly to a set of entities within the dataset, implying a need to add this information to the model. Cluster structure was not obvious until users identified and labeled the clusters, but was an important part of their sensemaking process. Knowing which of their three spatial strategies the user has chosen would help models understand the meaning of the clusters. Some soft data, such as search terms, can help distinguish between what hard data is relevant and not. Search terms can help indicate both what documents are important (based on being a more frequent search result), as well as which terms (or entities) to weight more heavily (indicated directly from the search terms). Document movement can be an indicator of not only similarity, but the pattern of movements can indicate the importance of the document. Users' cognitive similarity metrics are not limited to term co-occurrence or transitive relationships. This may indicate that users develop similarity based on higher-level concepts. Sometimes highlighted phrases were an indication of a user's reasoning, based on cluster labels.

In contrast to the results from Dou et al. and Chang et al. [13, 17] who successfully recovered reasoning from user's interactions, our measures indicate that doing so systematically yields lower probabilities. However, we have confirmed that analysts encode meaning into spatializations through complex spatial structures, using a rich set of cues. We can detect hints of meaning through these rich cues, such as the spatial layout and the interactions. All of them provided some benefit, but no single one gave an absolute indication of reasoning. Thus, a probabilistic approach that integrates all of them

is the most likely path for success. A tactful combination of the soft data can be exploited by clustering algorithms to help guide and enhance their outcome, incrementally during the course of interaction.

For example, an algorithm can exploit document movement in a spatial metaphor to learn and incrementally update similarity measures within a dataset. Observation-level Interaction [28] uses this form of soft data to couple the movement of data within a spatialization with updating parameters of popular clustering algorithms. In these models, users are given the ability to interact within the visualization, rather than directly with visual controls of parameters of the statistical model. While doing so, it is the responsibility of the model to update the parameters that correspond to the manipulation within the visualization. This is similar to the concept of metric learning, where models adjust the weighting of dimensions according to the user's input [76]. As another example, semantic interaction [25] exploits document movement, highlighting, annotating, and search to update the model and co-create a spatial layout. The system interprets the interaction and updates the layout incrementally.

3.5 Conclusion

In this paper we present the results of a study observing users analyzing a textual dataset spatially. We analyze the final layouts created by the users, and the captured user interactions performed while generating the clusters. We found how specific criteria within this process (including both the generated clusters and the interactions used) can indicate important and discriminating structure within the dataset.

Through analyzing the clusters created by the users, we found that only 15% of the 86 clusters contain at least one co-occurring term in all the documents within the cluster. Instead, we found that users tend to create clusters using transitive relationships between documents within a cluster. The challenge then, is determining which terms to use to

create these relationships. Many of the clusters users created are based on higher-level or process-level concepts during sensemaking. Thus, these concepts rarely relate directly to keywords, making simple term co-occurrence metrics less useful.

The interactions performed by the users (i.e., document movement, highlighting, searching) in spatially analyzing the dataset can provide indicators towards what structure within the dataset is important (or discriminating) to the user. For instance, important documents were returned as search results more frequently than non-important documents. Further, users' highlights sometimes indicated terms or phrases within a document that are important to the cluster definition.

This collection of interaction data, referred to as *soft data,* can be vital to unobtrusively gain an understanding of what aspects of a dataset a user finds important. As such, by incorporating both *hard data* (extracted directly from the dataset) and *soft data*, models can calculate more useful similarity metrics for users, and ultimately generate layouts from which users can gain insight.

Chapter 4

Semantic Interaction and ForceSPIRE

Figure 15. A model of semantic interaction. Users are able to interact directly in the spatial metaphor. The system updates the corresponding parameters of the statistical model based on the analytic reasoning of the users. Finally, the model updates the visualization based on the changes, thus unifying the synthesis and foraging stages of the sensemaking loop.

In the purest sense, semantic interaction refers to interaction occurring *within* a spatial visualization, with the added benefit that it is tightly coupled to the model calculating the spatial layout (Figure 15). Given the previous work of what interaction in visual analytic tools *is*, semantic interaction occupies a new design space for interaction. It merges the ability to change the statistical model while maintaining the flexibility and familiar methods for interacting within the metaphor of spatial visualizations. Users can benefit from semantic interactions in that they can interact within a metaphor which they are familiar with, performing interactions which are part of the spatial analytic process [5], without having to focus on formal updates to the model.

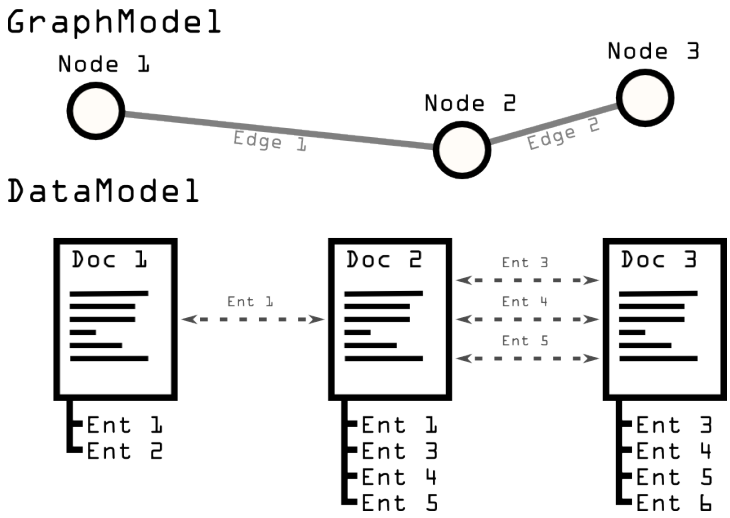


Figure 16. Overview of how nodes and edges in ForceSPIRE's force-directed layout are created from documents (Doc) and entities (Ent), respectively.

Semantic interaction leverages the cognitive connection formed between the user and the spatial layout. The following intelligence analysis scenario is representative of the strategies and interactions of analysts when performing an intelligence analysis task of textual documents in a spatial visualization, as previously found by Andrews et al. [5], and further motivates and explains the concept of semantic interaction:

During her analysis, an intelligence analyst finds a suspicious and interesting phrase within a document. While reading through the document, she highlights the phrase "suspicious individuals were spotted at the airport" in order to more easily recall this information later. After she finishes reading the document, she moves the document into the bottom right corner of her workspace, in the proximity of other documents related to an event at an airport. To remind herself of her hypothesis, she annotates the document with "might be related to Revolution Now terrorist group". Now, with the goal of further examining the events at the "airport", she searches for the term, continuing her investigation.

With semantic interaction, the system learns from these interactions and provides the analyst with visual feedback. Two documents relevant to the documents in the bottom

right corner begin to move closer to that cluster. She quickly reads through these, and notices that one of them seems related, and moves it into the cluster. It informs her that the "Revolution Now group is operating in airports", strengthening her insights. The other document talks about a "terrorist at an airport in Afghanistan". She moves this document away from the cluster, notifying the system that this recommendation was not relevant, as she is not currently investigating activities in Afghanistan. Through incremental learning and user feedback, the layout is co-created from the domain expertise of the user and the computed similarity of the system.

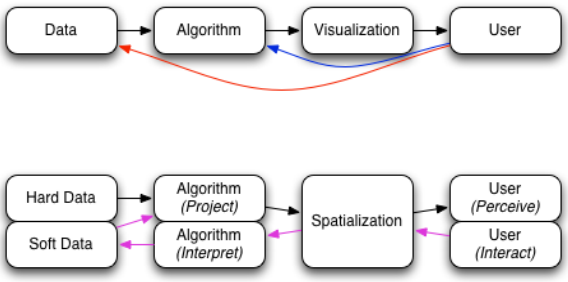


Figure 17. (top) The basic version of the "visualization pipeline". Interaction can be performed on directly the Algorithm (blue arrow) or the data (red arrow). (bottom) Our modified version of the pipeline for semantic interaction, where the user interacts within the spatial metaphor (purple arrow).

In addition to the three forms of semantic interaction in the scenario, Table 1 provides a list of various forms of semantic interaction, including how each can be used within the analytic process of investigating textual information spatially. We do not claim that this list is complete, but instead point out that each of these interactions can relate to a user's reasoning within the analytic process.

4.1 Designing for Semantic Interaction

In order for analysts to interact with information in a spatial metaphor, it must first be created. Following the model of the visualization pipeline [33], this creation calls for a series of mathematical transformations, turning raw data into a spatial layout – much the way many of the visualizations mentioned previously are constructed. However, these visualizations *fit* this model, as their user interactions are primarily focused on directly modifying the statistical model (as well as other attributes of the visualization or data transformation). Designing for semantic interaction requires a fundamentally different model for how tools integrate user interaction – one that can *capture the interaction*, *interpret the associated analytical reasoning*, and *update the appropriate mathematical parameters*.

Figure 17 illustrates this model, where the spatialization is treated a medium through which the user can perceive information and gain insight, as well as interact and perform his analysis. Through expanding the pipeline to accommodate for semantic interaction, it is a more appropriate match to the user's sensemaking process.

4.1.1 Capturing the Semantic Interaction

A non-trivial first step in the model is capturing the user interaction. Much research has been done in this area, primarily for the purpose of maintaining process history (e.g., [11], [64], [30], etc.). When considering how to capture interaction, one decision to be made is at what "level" to capture it. For example, GlassBox [15] captures interaction at a rudimentary level (i.e. mouse clicks and key strokes), while Graphical History [34] keeps track of a series of previous visualizations as a user changes the visualization during the exploration of the data.

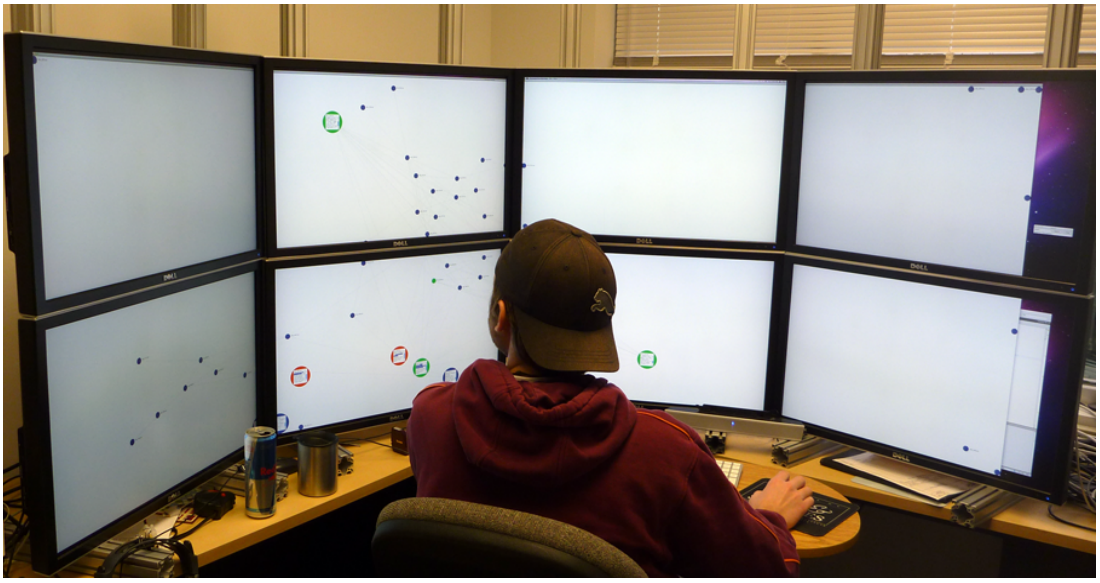


Figure 18. Using ForceSPIRE on a 32 megapixel large, high-resolution display. Photo by Alex Endert, 2012.

Semantic interaction is captured at a *data level*, as the interactions occur on the data, and within the spatial metaphor. Using the earlier analytic scenario, the interaction being captured would be:

- The highlighted **phrase**
- When the highlighting occurs (**timestamp**)
- The **color** chosen for the highlight
- The **document** in which the highlight occurs
- The new document **location**
- The text of the **annotation**

By capturing (and storing) the interaction history, we can interpret the analytical reasoning of the user. Thus, we not only capture the interaction, but also *use it.*

4.1.2 Interpreting the Associated Analytical Reasoning

In interpreting the interaction, the goal is for the system to determine the analytical reasoning associated with the interactions and update the model accordingly. From previous findings [5], we can associate analytical reasoning with forms of semantic interaction (see Table 1). It is essentially the model's task to determine *why*, in terms of the data, the interaction occurred. To answer this question, we do not propose that this model can accurately gauge user intent. Instead, the goal is to calculate, based on the data, what information is consistent with the captured interaction. For instance, we associate text highlighting with adding importance to the text being highlighted. We do not claim that we can associate the interaction of highlighting to the intuition that spurred the analyst to highlight the text, which is far more challenging, and arguably impossible.

Table 1. Forms of semantic interaction supported in ForceSPIRE. Each interaction corresponds to reasoning of users within the analytic process. Corresponding model updates are performed to steer the model based on the user's reasoning.

Semantic Interaction	***Associated Analytic Reasoning***	***Model Updates***
Document Movement	• Similarity/Dissimilarity • Create spatial construct (e.g. cluster, timeline, list, etc.) • Test hypothesis, see how document "fits" in region	• Similarity/Dissimilarity b/w documents • Up-weight shared entities, down-weight others
Text Highlighting	• Mark importance of phrase (collection of entities) • Augment visual appearance of document for reference	• Up-weight highlighted entities
Pinning Document to Location	• Give semantic meaning to space/layout	• Layout constraint of specific document
Annotation, "Sticky Note"	• Put semantic information in workspace, within document context	• Up-weight entities in note • Append entities to document and model
Level of Visual Detail (Document vs Icon)	• Change ease of visually referencing information (e.g. full detail = more important = easy to reference)	• (Full Document): "heavier node", increase node's friction • (Icon): "lighter node", less friction
Search Terms	• Expressive search for entity	• Up-weight entities contained in search • Add entities to model

We refer to the captured and interpreted interactions as *soft data*, in comparison to the *hard data* that is extracted from the raw textual information (e.g., term or entity frequency, titles, document length, etc.). We define soft data as the stored result of user interaction as interpreted by the system. In representing interaction as soft data, the algorithm can calculate and reconfigure the spatial layout accordingly. Figure 17 illustrates how our approach differs from the traditional visualization pipeline.

There has been previous work in capturing and interpreting reasoning from user interaction. For instance, Dou et al. [17] performed a study where financial analysts were asked analyze a dataset using WireVis, an interactive financial transaction visualization. The tool developers then analyzed the captured interaction, and assumptions were made about the reasoning of the analysts at specific points in the investigation. These results were compared to the analysts' self-recorded reasoning, and found to be accurate up to 82%. While our work has similar goals (i.e., interpreting the analytical reasoning associated with the analysts through an evaluation of the interaction) our model does so through tightly integrating the interaction with the underlying mathematical model. In doing so, the interpretation can be done algorithmically.

4.1.3 Updating the Underlying Model

Through metric learning of distance weights, the layout uses the soft data to update the underlying model. Depending on the algorithm used to compute the spatial layout, the precise parameters being updated will vary. In general, this will refer to weighting of a combination of dimensions that will help guide the model as to which dimensions the user finds important.

Figure 19. Moving the document shown by the arrow, ForceSPIRE adapts the layout accordingly. Documents sharing entities with the document being moved follow.

4.2 ForceSPIRE: System Overview

ForceSPIRE is a visual analytics prototype designed for specific forms of semantic interaction (document movement, text highlighting, search, and annotation) for interactively exploring textual data. The system has a single spatial view (shown in Figure 1), where a collection of documents is represented spatially based on similarity (i.e., documents closer together are more similar).

ForceSPIRE is designed for large, high-resolution displays (such as the one shown in Figure 18). As semantic interaction emphasizes the importance of context in which the interaction takes place (e.g., highlighting text in the context of the document), having the full detail text available in the context of the spatial layout is beneficial over having a single document viewer. Further, the physical presence of these displays creates an environment in which the virtual information (in this case the documents) can occupy persistent physical space. As a result, users are further immersed into the spatial metaphor, as they can point and quickly refer to information based on the physical locations.

4.2.1 Constructing the Spatial Metaphor

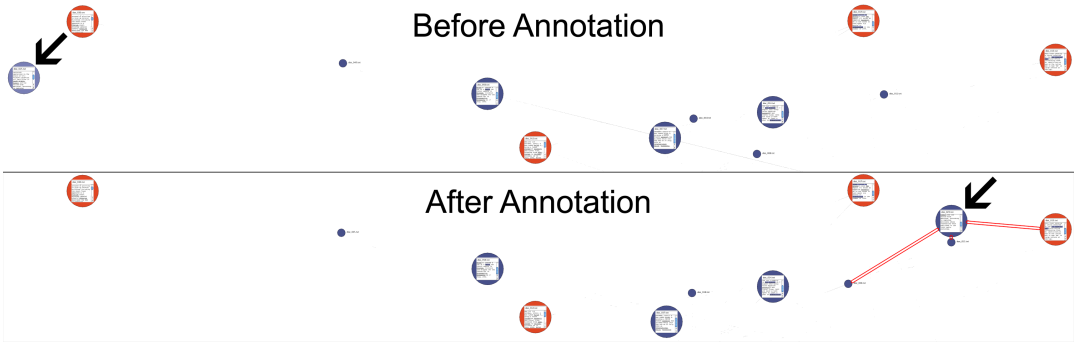


Figure 20. The Effect of adding an annotation ("these individuals may be related to Revolution Now") to the document shown with an arrow. As a result, the document becomes linked with other documents mentioning the terrorist organization "Revolution Now".

The spatial layout of the text documents is determined by a modified version a force-directed graph model [29]. This model functions on the principle of nodes with a mass connected by springs with varying strengths. Thus, each node has attributes of attraction and repulsion: nodes repel other nodes, and two nodes attract each other only when connected by a spring (edge). The optimal layout is then computed by iteratively calculating these forces until the lowest energy state of all the nodes is reached. A complete description of this algorithm can be found in [29].

We apply this model to textual information by treating ***documents*** as ***nodes*** (an overview is shown in Figure 16). The entire textual content of each document is parsed into a collection of entities (i.e., keywords). The number of entities corresponds to the ***mass*** of each document (heavier nodes do not move as fast as lighter nodes). A ***spring*** (or edge) represents one or more matching ***entities*** between two nodes. Therefore, the initial distance metric is a based on co-occurrence of terms between documents. For example, two documents containing the term "airport" will be connected by a spring. The strength of a spring (i.e. how close together it tries to place two nodes) is based on two factors: the number of entities two documents have in common, and the ***importance value*** associated with each shared entity (initially, importance values are created using a standard tfidf method [42]).

The resulting spatial layout is one where similarity between documents is represented by distance relative to other documents. ***Similarity*** in this system is defined by the strength of the spring between two documents. A stronger spring (and therefore a larger amount of shared entities) will pull two documents closer together, and thus represent two similar documents.

4.2.2 Semantic Interaction in ForceSPIRE

The semantic interactions in ForceSPIRE are: placing information at specific locations, highlighting, searching, and annotating in order to incrementally change the spatial layout to match their mental model. The primary parameters of the force-directed model that are being updated through this learning model are the importance values of the entities. This section describes the functionality of these features. The mathematics of how these analytic interactions update the underlying model are presented in Section 4.2.3.

Document Movement. The predominant interaction in a spatial workspace is positioning (and repositioning) documents. In previous work, we have demonstrated how users can perform both *exploratory* and *expressive* forms of this type of interaction [28]. In ForceSPIRE, we allow for the following exploratory interaction (i.e., interaction that allows users to explore the structure of the current model, but does not change it). Users are able to interactively explore the information by ***dragging*** a document within the workspace, ***pinning*** a document to a particular location (see Figure 19), as well as ***linking*** two documents. When dragging a document, the force-directed system responds by finding the lowest energy state of the remaining documents given the current location of the dragged document. Mathematically, this adds a constraint to the stress function being optimized (in this case the force-directed model). This allows users to explore the relationship of that document in comparison to the remaining documents.

In addition to the exploratory dragging of a document, users have the ability to ***pin*** a document. By pinning a document, users are able to incrementally add semantic meaning to locations in their workspace. By specifying key documents to user-defined locations, the layout of the remaining documents will adapt to these constraints. Thus, users can

explore how documents are positioned based on their similarity (or dissimilarity) to the pinned documents. For instance, if the layout places a document between two pinned documents, it may imply that the particular document holds a link between the two pinned documents, sharing entities that occur in both.

Finally, users can perform an expressive form of this interaction by ***linking*** two documents, performed by dragging one document onto another pinned document. In doing so, ForceSPIRE calculates the similarity between the documents, and increases the importance value of the entities shared between both documents. As a result, the layout will place more emphasis on the characteristics that make those two documents similar.

Highlighting. When ***highlighting a term***, ForceSPIRE creates an entity from the term (if not already one), and the importance value of that term is increased. Similarly, ***highlighting a phrase*** results in the phrase being first parsed for entities, then increasing the importance value of each of those entities. For example, Figure 22 shows the effect of highlighting the terms "Colorado" and "missiles" in the document pointed to with the arrow. As a result, the other documents containing that term are clustered more tightly.

Searching. When coming across a term of particular interest, analysts usually search on that term in order to find other occurrences. In a spatial workspace, this is of particular importance, because the answer to "where the term is also found" is not only given in terms of what documents, but also where in the layout those documents occur. The positions of documents containing the term are shown in context of the entire dataset, from which users can infer the importance of that term (as shown in Figure 21).

ForceSPIRE first creates an entity from the search term (unless it is already one), then increases the importance value of the search term. Figure 21 gives an example of how a search result appears in ForceSPIRE. Searching for the term "Atlanta", documents that contain the term are highlighted green, and links are drawn to show where the resulting documents are in relation to the current document.

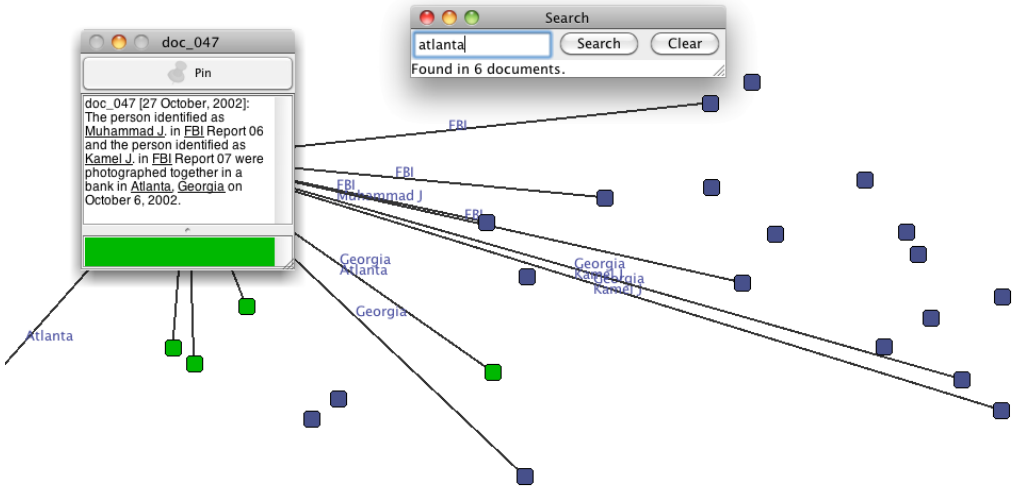


Figure 21. Searching for the term ”Atlanta”, documents containing the term highlight green within the context of the spatial layout. Additionally, the importance value of entity “Atlanta” is increased.

Annotation. Annotations (i.e., “sticky notes”) are also viewed as a form of semantic interaction, occurring within the analytic process, from which analytic reasoning can be inferred. When a user creates a note regarding a document, that semantic information should be added to the document. For example, if Document A refers to “Revolution Now” (a suspicious terrorist group), and Document B refers to “a group of suspicious individuals”, and the user has reason to believe these individuals are related to Revolution Now, adding a note to Document B stating “these individuals may be related to Revolution Now” is one way for the user to add semantic meaning to the document.

ForceSPIRE handles the addition of the note (shown in Figure 20) by 1) parsing the note for any currently existing entities, then 2) increasing the importance value of each, and 3) creating any new springs between other documents sharing these entities. In the example in Figure 20, edges are created between Document B and Document A (as well as any other documents that mention “Revolution Now”). Additionally, if the note contains any new entities not currently in the model, they are created, with the intent that any future entities that may match to that note can be linked at that time. ForceSPIRE also handles cases where notes are edited, with text added or removed from the note, by updating the entities associated with the document, and adjusting the importance values of these entities accordingly.

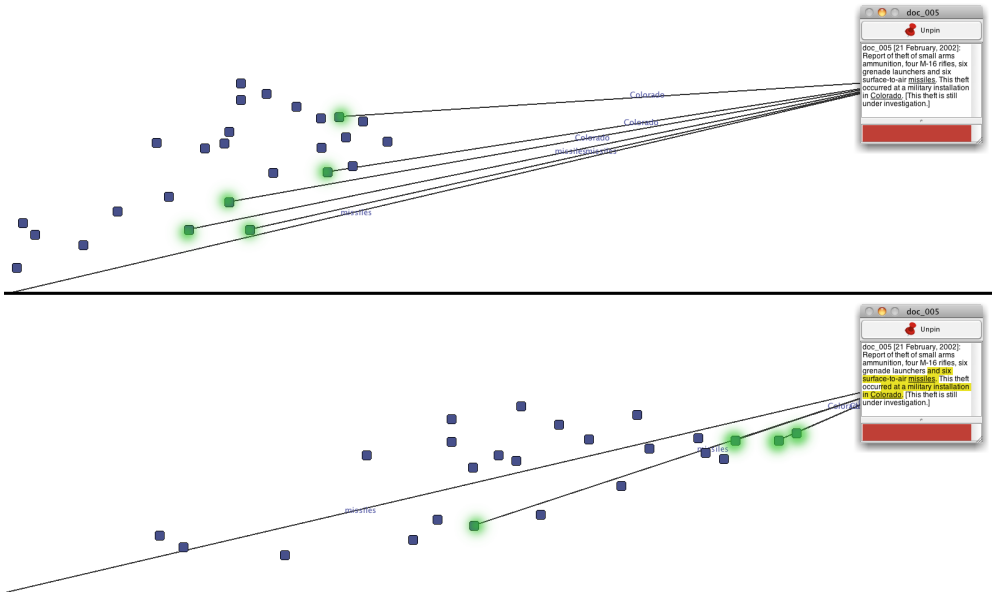


Figure 22. The effect of highlighting a phrase containing the entites "Colorado" and "missiles". Documents containing these entities move closer, as the increase in importance value increases the edge strength.

4.2.3 Model Updates

A force-directed model is based on a set of nodes connected with springs of different strengths, which pull on each other until the entire graph reaches a lowest-stress state [29]. First, we parse the text of each dataset into T unique entities (i.e., dimensions). Then, a weighting vector is applied to each dimension. As such, each dimension (i.e., entity) in the dataset at the nth iteration has a weight given by

$$\underline{w_n} = (w_{n1}, \dots, w_{nT})$$

Each document is consists of a set of entities contained in it, and membership of an entity in a document is defined as

$$x_{it} = \begin{cases} 0, \text{entity } t \text{ is not in doc } i \\ 1, \text{ entity } t \text{ is in doc } i \end{cases}$$

The mass m of a document i determines the weight of that document in the model. Documents with more weight move more slowly, thus "anchoring" the spatial layout. The

mass of a document does not depend on the total number of entities in the document, as each entity only attributes to the mass of the document once. It is described by

$$m_i = \sum_{t=1}^{T} x_{it}\, w_t$$

The strength of a spring (i.e., edge) between two nodes (i.e., documents) can be defined by

$$K_{ij} = \sum_{t=1}^{T} x_{it} x_{jt} w_t$$

Therefore, a spring with a higher strength will attract the two nodes more closely. In the following sections, we discuss the semantic interactions in ForceSPIRE, and show how the underlying force-directed model is updated.

4.2.3.1 Document Movement

Users can manipulate the spatial layout directly in the spatialization by placing documents in locations based on the user's domain knowledge. These movements (i.e., *Observation-Level Interactions*) can be both *exploratory* and *expressive* [28], differentiated by how they adjust the underlying model. *Exploratory movements* do not change the weighting of keywords (or entities), but use the current weights to determine the position of the remaining documents given the user-defined location of the document being moved. These can be seen as "model constraints", as the user decides the placement of one or more documents, and the model produces the remaining layout based on these static locations. Additional models, and how they can be adapted to support observation-level interaction are discussed in Section 4.3.

With *expressive movements*, users are able to inform the system that the weighting vector should be updated to reflect a change in similarity between two (or more) documents. For example, when placing two documents closer together, the system determines the similarity between those two documents, and increases the weight on the corresponding

entities. As a result, a new layout is incrementally generated reflecting the new similarity weighting, where those two documents (as well as others sharing similar entities) are closer together.

In terms of model updates, *expressive movements* are the only ones that change the weight vector of the model. First, a set of entities $\mathcal{M} = \{m_1, \ldots, m_M\}$ that co-occur in both documents moved closer are calculated. Then, the weight of each entity, w_m, is described by:

$$w_{n+1,m} = w_{nm} + K,$$

where K is a constant by which the weight, w_m, is updated. This constant can be adjusted based on how quickly and aggressively the user wants to model to adjust. Each of the remaining entities is adjusted equally to ensure that a normalized total energy is maintained in the system. This step of the update is described by,

$$w_{n+1,t} = \max\left(0, w_{nt} - \frac{KR}{(T-R)}\right), t \notin \mathcal{M}$$

Once the updated weight vector is computed, the model updates the spring strengths and document masses, and the layout iterates until settling again.

Users also have the ability to pin documents to specific locations. These documents serve as spatial landmarks, in that they persist at that location, and the force-directed model treats them as layout constraints, organizing the remaining documents around them. Additionally, pinning allows ForceSPIRE to distinguish between exploratory and expressive movements. Dragging a document near a pinned document will briefly color both documents pink to alert the user of the expressive movement (if the user releases the document at this location). Thus, all other movements in the space are exploratory movements.

4.2.3.2 Text Highlighting

Users can highlight text segments directly in the full-detail document views. As a result, the system increases the importance of the terms highlighted, updating the underlying mathematical model, and ultimately the layout. The phrases highlighted are also parsed for entities using a more aggressive entity extraction algorithm, so as to add entities to the model that may have been initially missed. Users also have the ability to select specific entities to add, delete, or modify by using the Entity Viewer.

When highlighting a phrase of text within a document, ForceSPIRE creates a set of entities contained in the highlight. By using a more aggressive form of entity extraction on the highlighted phrase, some of the entities found in the highlight may be newly discovered. Those entities are added to the model before updating the weighting vector. Suppose the user highlights a phrase, composed of a set of entities $\mathcal{H} = \{h_1, \dots, h_H\}$. The updated weight, w_h, of an entity h based on a highlight, is described by:

$$w_{n+1,h} = w_{nh} + K, h \in \mathcal{H}$$

where K is a constant by which the weight of the highlighted entity, w_h, is updated. Each of the remaining entities is adjusted equally to ensure that a normalized total energy is maintained in the system. This step of the update is described by,

$$w_{n+1,t} = \max\left(0, w_{nt} - \frac{KH}{(T-H)}\right), t \notin \mathcal{H}$$

Once the updated weight vector is computed, the model updates the spring strengths and document masses, and the layout iterates until settling again.

4.2.3.3 Search

Search allows users to perform a standard text search within the dataset. As a result, documents containing the search term will be highlighted, and an edge between the search box and those documents will be created (multiple search boxes can exist). The

model is updated by increasing the weight of the entity searched for (and creating a new entity for the search term if it does not already exist).

Searching for a term creates a new node in the graph consisting of only the entity contained in the search. This node can be pinned to define an absolute location in the spatialization given the search term, or can be positioned by the model. The updated weight, w_s, of a search on entity s, is described by:

$$w_{n+1,s} = w_{ns} + K,$$

where K is a constant by which the weight of highlighted entity, w_s, is updated. Each of the remaining entities is adjusted equally to ensure that a normalized total energy is maintained in the system. This step of the update is described by,

$$w_{n+1,t} = \max\left(0, w_{nt} - \frac{K}{(T-1)}\right), t \neq s$$

Once the updated weight vector is computed, the model updates the spring strengths and document masses, and the layout iterates until settling again.

4.2.3.4 Annotation

An annotation can be individually added to each document. Through annotating a document, users can add "meta-information" to the document based on their domain expertise. For example, adding a note "relates to the events in Chicago" results in parsing the note for entities (i.e., "Chicago"), and adding them to the document, which creates edges to other documents containing "Chicago".

Each of these semantic interactions creates ***soft data***, a quantitative representation of captured user interaction within the context of the dataset. Figure 17 models how soft data is collected (i.e., captured and interpreted interactions within the context of the dataset), as well as how it is combined with the *hard data* to produce the spatial layout. As a result, the soft data steers the underlying force-directed model. Also, soft data serves

as a log of the entity weighting throughout the user's analytic process, and can be examined at any time to gain insight about their process.

The entities included in an annotation $\mathcal{A} = \{a_1, \ldots, a_A\}$ (which can include newly identified entities not previously in the dataset) update both the mass of a node, as well as the weighting vector as follows. The updated weight, w_a, of an entity a contained in an annotation is described by:

$$w_{n+1,a} = w_{na} + K, a \in \mathcal{A}$$

where K is a constant by which the weight w_a is updated. Each of the remaining entities is adjusted equally to ensure that a normalized total energy is maintained in the system. This step of the update is described by,

$$w_{n+1,t} = \max\left(0, w_{nt} - \frac{KA}{(T-A)}\right), t \notin \mathcal{A}$$

Once the updated weight vector is computed, the model updates the spring strengths and document masses, and the layout iterates until settling again.

4.2.3.5 Undo

When *undoing* an interaction using the standard "Control+Z" keyboard shortcut, a linear history of the interactions will be reversed, and the importance values of affected entities will be returned to their prior values (as well as document masses). As for the locations of the documents, the reverted importance values and document masses will be responsible for updating the layout. However, this does not guarantee that the layout will return to the exact previous view, and the user may find it necessary to perform small adjustments.

The model updates used in ForceSPIRE serve as an initial approach at how to couple semantic interactions with model updates. Other, more complex methods may exist, and we encourage further research in this area. Sensemaking is a complex exploratory process. As such, semantic interaction can enable analysts to explore their hypothesis in-situ, while the provenance of their insights is captured and stored. An open area of

research is what analyzing the soft data might reveal about the analytic process. For instance, if the importance values of entities converge on a small number of entities, specific biases might be revealed. Similarly, instances during the analysis when new hypotheses are being explored may be indicated by diverging importance values.

4.3 Observation-Level Interaction

> *This section contains mathematical algorithms and formulations developed by collaborators and co-authors in statistics, including Drs. Scotland Leman and Leanna House, Dipayan Maiti, and Chao Han. These mathematical contributions are included in this dissertation to show further examples of how one form of semantic interaction, observation movement in spatializations, can be applied to three dimension reduction models. The complete and discussion can be found in [28].*

In general, observation-level interaction refers to interactions, occurring within a spatialization, that enable users to interact directly with data points (i.e., observations). A spatialization in this context refers to a two-dimensional layout calculated from high-dimensional data where the metaphor of relative spatial proximity represents similarity between documents. That is, data points placed closer together are more similar. Observation-level interactions are therefore tightly coupled with the underlying mathematical models creating the layout, thus allowing the models to update parameters based on the interaction occurring. While numerous forms of interaction may exhibit these characteristics (e.g., moving clusters of documents, marking regions of interest within the spatialization, etc.), in this paper we will focus on one – movement of observations. From previous studies, we found that movement of observations (in those cases documents) closer together is one way for the user to externalize the analytical reasoning that those documents are somehow similar [5]. In this study, the spatial rearrangement of documents was an integral part of each intelligence analysts'

sensemaking process. Further, this study points out that users perform observation-level interaction in two ways, *exploratory* or *expressive*, based on the particular analytical reasoning associated with the interaction, and also how the system responds.

During an exploratory interaction, users utilize the algorithm to explore the data and the space. For example, through dragging one observation within the layout, users gain insight into the structure of the data by observing how other data reacts given the algorithm. While an observation is dragged through the layout, the algorithm adjusts the layout of the remaining data according to how the algorithm computes similarity. Thus, when the observation is dragged towards a cluster of data, similar data points attract, while dissimilar ones repel. Additional information such as a list of similar and dissimilar parameters can also be displayed. Through this process, users learn about a single observation, and how it relates to the other observations in the dataset.

An expressive interaction is different, in that it allows users to "tell" the model that the criteria (i.e. the parameters, weights) used for calculating the similarity need to be adjusted globally. For example, as a user reads two documents, she denotes they are similar by dragging them close together. If this were exploratory, the two documents would repel again. However, in an expressive form of this interaction, it is the responsibility of the underlying mathematical model to calculate and determine why these documents are similar, and update the model generating the spatial layout accordingly. Using the methods below, we illustrate how both expressive and exploratory forms of observation-level interaction are enabled through modifications made to three common statistical methods (PPCA, MDS, and GTM).

4.3.1 Methods Integrating Observation-level Interaction

A probabilistic model assumes a sampling distribution for the observed data and an uncertainty over the model parameters (e.g. PPCA and GTM discussed in Section 4.1 and 4.3 respectively). A deterministic method makes no such assumptions about the data or the parameters (e.g. Weighted MDS, discussed in Section 4.2). House et al. describe in detail the underpinnings of the probabilistic framework, termed as "Bayesian Visual

Analytics" (BaVA) [38]. The BaVA process begins with an initial display of the data. In turn the user may assess the display and decide if it matches her mental model of the data. If it does not, the user may convey her cognitive feedback $f^{(c)}$ by adjusting the locations of two observations to convey her mental model about the two observations. The user might also explore an alternative spatial location of an observation and see how the other observation responds to such an interaction. In short, iterations of user interaction and subsequent regeneration of the visualization are modeled as sequential updating of maximum a posteriori estimates of parameters. The deterministic version of the framework, termed as "Visual to Parametric Interaction" (V2PI), also starts with an initial display and upon obtaining a user feedback sequentially updates the parameters, but the updated values of the parameters are such that they minimize some measure of discrepancy between the expected configuration of the data under the user's reasoning and the original data [44].

For each of the models discussed in this paper, we present an overview of the model, describe the modifications made to allow observation-level interaction, and show a use case demonstrating how an end-user can interact with each model. Given that each of these models is designed for different types of data (varying in structure, size, and nature of the data), the example use cases below each use different datasets to match the intended use of the models with the use case. The use cases are performed in prototype visualizations to show a proof of concept, and we are actively working to incorporate these models into more fully featured tools.

4.3.1.1 PPCA

Overview

Principal Component Analysis (PCA) [40, 54, 71] is a common, deterministic method used to summarize data in a reduced dimensional form. The summary is a projection of a high-dimensional dataset in the directions with the largest variance. When only two directions are chosen, PCA may produce a spatial representation or map of the data that is easy to visualize. One problem with PCA is that important structures (e.g., clusters) in

data may not correlate with variance. Thus, PCA spatializations may mask information in the data that analysts may find useful.

Probabilistic PCA [70] is, simply, a probabilistic form of PCA. This means that PPCA is not a deterministic algorithm, but a statistical modeling approach (specifically, a factor modeling approach) that *estimates* low-dimensional representations of high-dimensional data. Let $\boldsymbol{d}=[d_1,\ldots,d_n]$ represents a $p\times n$ high-dimensional data matrix, where n represents the number of observations, p represents the number of columns, and d_i *(*for $i\in\{1,\ldots,n\}$*)* represents a $p\times 1$ vector for observation i. Also, let $\boldsymbol{r}=[r_1,\ldots,r_n]$ represent a low-dimensional analogy of $\boldsymbol{d}$, such that $\boldsymbol{r}$ is $q\times n$ and $q<p$. For our purposes, we set $q=2$. PPCA models $\boldsymbol{d}$ as a function of $\boldsymbol{r}$,

$$d_i|W,r_i,\mu,\sigma^2 = No(Wr_i+\mu, I_p\sigma^2)$$

where, No(.,.) represents the Multivariate Normal Distribution; μ represents a $p\times 1$ mean-vector of $\boldsymbol{d}$; $\boldsymbol{W}$ is a $p\times q$ transformation matrix known as the factor loadings of $\boldsymbol{d}$; $\boldsymbol{I}_p$ is a $p\times p$ identity matrix; and σ^2 represents the variance of each dimension in $\boldsymbol{d}$. By convention, PPCA models each r_i with a Multivariate Normal distribution centered at zero and with unit variance: $r_i\sim\text{No}(\mathbf{0}_2, \mathbf{I}_2)$. In turn, the conditional posterior distribution for r_i is $\text{No}(\eta,\Sigma_r)$, where

$$\eta = (W'W+I_2\sigma^2)^{-1}W'(d_i-\mu)$$

$$\Sigma_r = \left(W'W\sigma^{-2}+I_2\sigma^2\right)^{-1} \qquad (1)$$

A spatialization of data $\boldsymbol{d}$ that relies on PPCA plots the posterior expectation η. Similar to PCA, the coordinates η rely on the variability observed in $\boldsymbol{d}$. To see this, let $\boldsymbol{\Sigma}_d$ represent the marginal variance of d_i, ($\boldsymbol{\Sigma}_d=\text{V}[d_i|\boldsymbol{W},\mu,\sigma^2]$). Since $\boldsymbol{\Sigma}_\text{d}=\boldsymbol{W'W}+\boldsymbol{I}_2\sigma^2$, we can rewrite η as $\eta=\boldsymbol{\Sigma}_\text{d}^{-1}\boldsymbol{W}(d_i-\mu)$ which shows that the relationship between $\boldsymbol{\Sigma}_\text{d}$ and η is well defined.

The final step in PPCA is to estimate the model parameters, $\{\boldsymbol{W}, \mu, \sigma^2, \Sigma_d\}$. We take a Bayesian approach. We specify either reference or flat priors for each unknown (as suggested by [70] and use Maximum A Posteriori (MAP) estimators to assess (and plot)

η. For example, when we assign $\pi(\boldsymbol{\Sigma}_d) \propto 1$, the posterior distribution for $\boldsymbol{\Sigma}_d$ is an Inverse Wishart (IW) distribution,

$$\pi(\Sigma_d | d) \propto \mathrm{IW}(nS_d, p, n-p-1) \qquad (2)$$

Where $\boldsymbol{S}_d$ represents the empirical variance of $\boldsymbol{d}$. The MAP estimate of $\boldsymbol{\Sigma}_d$ is $\boldsymbol{S}_d$.

User Guided PPCA

To enable analysts to guide PPCA via the data visualization, we take advantage of the relationship between $\boldsymbol{\Sigma}_d$ and η. Namely, changes in $\boldsymbol{\Sigma}_d$ will effect η, and changes in η will effect $\boldsymbol{\Sigma}_d$, when we invert Equation (1).

After obtaining an initial PPCA display, the user adjusts the locations of two observations; i.e., adjusts two columns in η. If the two observations are moved close to one another, the analyst is conveying that in her mental map, the observations are more similar than what they appear in the display; and, if the observations are dragged apart, the analyst is conveying that the observations differ more than what they appear.

The challenge in BaVA is to parameterize the cognitive feedback and update the visualization [38]. First, we determine the dimensions of the data $\boldsymbol{d}$ for which the adjusted observations are similar and different. Second, we transform the adjustments to η into a hypothetical $p \times p$ variance matrix. We denote this matrix by $f^{(p)}$, as it is a quantified version of $f^{(c)}$. In $f^{(p)}$, the dimensions for which the adjusted observations are similar have small variances and the dimensions for which adjusted observations differ have large variances. Third, we consider the hypothetical variance $f^{(p)}$ to be a realization of a Wishart distribution that has an expectation equal to $\boldsymbol{\Sigma}_d$. Finally, we apply Bayesian sequential updating [67, 73] to adjust Equation (2) by the parametric feedback $f^{(p)}$,

$$\pi(\Sigma_d | d, f^{(p)}) = \mathrm{IW}(pS_d + \upsilon f^{(p)}, p, n+\upsilon-p-1)$$

where, υ is solved from a specification $\kappa (\kappa \in [0,1])$ made by the analyst that states how much weight to place on the feedback relative to the data. Namely, the updated MAP estimate for $\boldsymbol{\Sigma}_d$ is a weighted average of the empirical variance $\boldsymbol{S}_d$ and feedback $f^{(p)}$

$$MAP(\Sigma_d) = \frac{v}{v+n} f^{(p)} + \frac{n}{v+n} S_d$$

thus $\upsilon = n\kappa/(1-\kappa)$. Now, the PPCA projection of the data ***d*** that is based on MAP($\boldsymbol{\Sigma}_d$) will portray both information in the data and expert feedback.

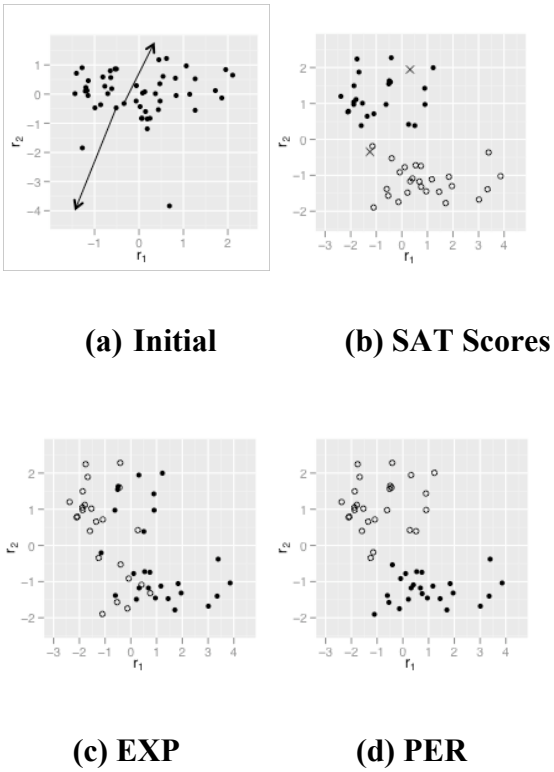
Example

A sensitive issue for taxpayers, parents, children, educators, and policy makers is whether an increase in money devoted to education will increase education quality. Money provides a means to buy modern textbooks, employ experienced teachers, and provide a variety of classes and/or extra curricular activities. Although, do the students who benefit from these high-priced resources actually improve academically?

In 1999, Dr. Deborah Guber compiled a dataset for pedagogical purposes to address this question [31]. Based on the following variables, the dataset summarizes the academic success, educational expenses, and other related variables in 1997 for each U.S. state: the average exam score on the Standard Aptitude Test (SAT); the average expenditure per pupil (EXP); the average number of faculty per pupil (FAC); the average salary for teachers (SAL); and the percentage of students taking the SAT (PER). To increase the complexity of the dataset slightly, we added two variables from the National Center for Education Statistics(www.http:nces.ed.gov): the number of high school graduates (HSG) and the average household income (INC). We hypothesize that states that spend more on education will cluster with states with high SAT averages.

To assess the hypothesis and explore the data, we implement the BaVA process using PPCA. Figure 1a), displays our initial view of the data. Notice that the visualization does not present any structure in the data. Analysts in the field of education, notice that two states with different expectations for SAT scores are displayed close to one another. Thus, we select the appropriate observations and drag them apart as an expressive interaction to obtain an updated view that is displayed in Figure 1b). There are two clusters in 1b). These clusters correspond with SAT scores above and below the national median.

Based on our hypothesis, we suspect that the clustering structure in SAT relates to EXP. However, when we re-plot 1b) and label the upper and lower EXP 50% quantiles in Figure 1c), EXP does not explain the clusters. Thus, we used a bi-plot to identify which variables explain the structure we see in Figure 1b). When we mark the observations above and below the empirical PER median in Figure 1d), we see that PER and SAT clearly relate to the formation of clusters in the dataset. Thus, further analyses of SAT and EXP must control for PER.



(a) Initial **(b) SAT Scores**

(c) EXP **(d) PER**

Figure 23 After injecting expert feedback into a), we obtain Figures b)-c). For frame of reference, we marked the two points moved to inject feedback by `x' in Figure b). The configuration of points in each graph are identical, but the observations are labeled differently. In Figure b), symbols `●' and `○' mark the upper and lower 50% quantiles for SAT scores respectively; in Figure c), symbols `●' and `○' mark the upper and lower 50% quantiles for EXP scores respectively; and in Figure d), symbols `●' and `○' mark the upper and lower 50% quantiles for the percentage of students taking the SAT (PER) respectively. Notice the clusters in each graph correspond with SAT and PER, but not EXP.

4.3.1.2 MDS

We extend our framework to another deterministic method, which forms the basis for a large number of visualization techniques: Multi-Dimensional Scaling (MDS).

Overview

All complex data visualizations are based on high-dimensional datasets, which contain features corresponding to dimensions, and the relative importance of such features through a set of weights (w_i). Classically weighted multidimensional scaling deals with mapping a high dimensional dataset $\boldsymbol{d}=[d_1,\ldots,d_n]$ into a low dimensional (in our case two-dimensional) space ***r,*** by preserving pairwise distances between observations in the low dimensional representation. Let ***w*** represent the p-vector of feature weights: $\boldsymbol{w}=\{w_1,\ldots,w_p\}$. Given a set of feature weights, the low dimensional spatial coordinates are found by solving:

$$\min_{r_1,\ldots,r_n} \sum_{i<j\leq n} \left\| \left\| r_i - r_j \right\| - \delta_{i,j}^{(w)} \right|,$$

where

$$\delta_{i,j}^{(w)} = \sum_{k=1}^{p} w_k \, \mathrm{dist}(d_{ik}, d_{jk})$$

such that $\sum_k w_k=1\sum_{\mathrm{d}} \mathrm{w_d} = 1$. dist() represents any distance function for measuring individual features in the high dimensional space. Because it is not possible to estimate weights and the set ***r*** simultaneously, we provide a uniform weighting of the space $w_i=1/p$ for our first iteration.

User Guided MDS

Once a visualization is generated, the user may either agree with the display and learn from certain aspects of the visualization, or disagree, based on their domain expertise. Hence, the user may wish to interact and rearrange a few of the observations in the visualization. Given a spatial interaction in the form of adjusting the relative position of a

set of points, we compute a set of feature weights, which are consistent with both, the users adjustment and the underlying mathematical model. These are computed by inverting the optimization, by fixing the locations of the adjusted points and finding an optimal set of weights, which are consistent with the visualization. Explicitly, we solve for ***w*** such that

$$\min_{w_1,\ldots,w_p} \sum_{\tilde{r}_i,\tilde{r}_j \in M} \left| \left\| \tilde{r}_i - \tilde{r}_j \right\| - \delta_{i,j}^{(w)} \right|$$

where $\sum_k w_k=1$ $\sum_{\mathrm{d}} \mathrm{w}_{\mathrm{d}} = 1$, and M the set of adjusted observations (r_i, r_j). It should be noted that computing the new weights is extremely fast, and is then followed by a full MDS step. Thus, the entire generation of a new view can be performed in real time, depending on the size of the dataset and the specific hardware used.

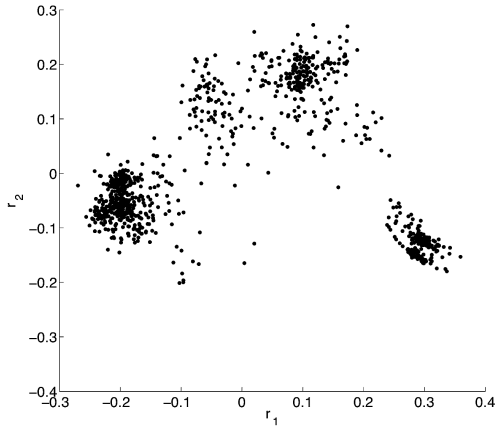


Figure 24 Visualization of the 1990 census dataset using classical MDS.

Example

Consider for example a visualization produced by a standard MDS technique. In this example we focus on the 1990 census dataset [9] under a Classical Metric Scaling (CMS) [62], using a Hamming distance (due to the categorical nature of the dataset) for measuring features in the high dimensional space. Figure 24 illustrates results obtained under a Classical Metric Scaling (CMS).

Given this visualization, a user may distinguish 3-5 main clusters, and inquire what they mean. We see two major ways a user can interact with the visualization, in order to explore the space, and learn about the underlying dataset. The first of these is by highlighting a subset of the data, based on some question the user seeks to answer, and then rearranging the visualization based on inconsistencies with their mental model (expressive interactions).

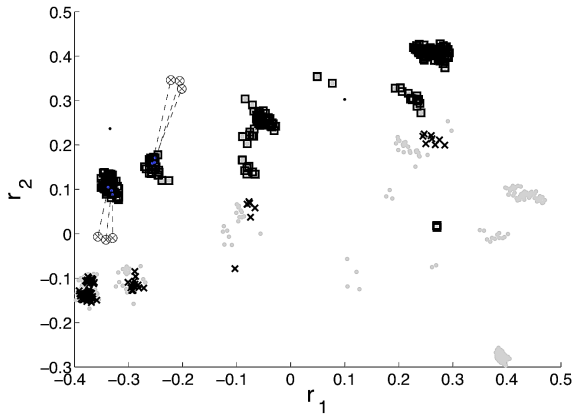


Figure 25 A user performing an observation-level interaction to learn what distinguishes two clusters.

The second approach is to hone in on visual structure, and move points in the visual space in order to learn what the structure relates to in terms of the feature space (Figure 25). Both of these interactions are nearly identical, however the motivation for the interactions will differ. We will illustrate both types of visual reasoning through an example based on the 1990 census dataset.

Figure 25 shows how the user might perform an observation-level interaction in order to learn what explains the clustering structure between the working/low income groups. To suggest the clusters could be moved further away from each other than they appear in the current visualization, the system reports back the weights, which explains the differences in the groups. For this example, the user learns that one of these clusters contains individuals that have a reliable mode of transportation to work (93% explained). The visualization could be updated based on this information, or the user could simply document this fact and proceed by explaining other areas of the spatialization. As always, there are an endless number of possibilities for learning about a high dimensional dataset

via visual expression/exploration. Another example of an exploratory interaction with MDS is demonstrated by Buja et al. in which users can constrain observations to specific spatial locations [10].

The user may wish to interact expressively and identify points in the space that pertain to high and low income groups. The user highlights individuals with incomes below 15K and over 60K, as shown by ■ and × in leftmost panel of Figure 26, respectively. Because of the close proximity of the highlighted groups in the main clusters, the user drags (denoted by ⊗) a few representative low and high-income individuals into sets of groups in each of the 3 main sub-clusters. The system reports back a set of weights, which explain how much a particular feature explains the arrangement of points suggested by the user. High weights relate to important features, while low weights suggest their corresponding features do not relate to the user's visual rearrangement. For our example, we learn not only that income level (29%), but also by their means of transportation to work (20%), whether or not they worked the full year (25%), and their level of education (10%) are related to the user's repositioning of points. Given this information, the system updates the visualization, as shown in center panel of Figure 26. We notice that in the resulting visualization, the income groups are clearly separated. The resulting visualization displays a much richer spatialization than simply showing clusters relating to the income groups. For example, we highlight individuals that actually worked in the right most panel of Figure 26, and notice these individuals are shown in distinct sub-clusters. 2 of the 4 clusters in which individuals work pertain to low-income groups, and the other 2 pertain to high-income groups (as illustrated by the ■ and × symbols).

4.3.1.3 GTM

Overview

Introduced by Bishop et al, [14] Generative Topographic Mapping (GTM) is a nonlinear latent variable modeling approach for high-dimensional data clustering and visualization. It is considered to be a probabilistic alternative for both the Self-Organizing Map (SOM)

algorithm [43] and Nonlinear PCA. Similar to PPCA, GTM estimates a latent variable $\boldsymbol{r}=[r_1,\ldots,r_n]$ ($q\times n$ matrix) that is a low-dimensional representation of high-dimensional data $\boldsymbol{d}=[d_1,\ldots,d_n]$ ($p\times n$ matrix such that $p>q$). However, unlike PPCA, the q-dimensional coordinates $\boldsymbol{r}$ in GTM map nonlinearly to a complex manifold $\boldsymbol{m}=[m_1,\ldots,m_n]$ that is embedded in the high-dimensional space. This manifold, ideally, characterizes important structure in data $\boldsymbol{d}$ and represents geometrically the expected value for $\boldsymbol{d}$ in the Gaussian model,

$$d_i : N(W\Phi(r_i), I_p\beta^{-1}) \tag{3}$$

To estimate a coordinate m_i, GTM takes a weighted average of J radial basis functions $\{\Phi_1(),\ldots,\Phi_J()\}$ ($\Phi_j()$ represents a radially symmetric Gaussian kernel) given r_i and parameters there in,

$$m_i = W\Phi(r_i), \tag{4}$$

$$\Phi_j(r_i) = \exp\left(-\frac{\|r_i-\mu_j\|^2}{2\sigma^2}\right), \tag{5}$$

where $\boldsymbol{W}$ is a $p\times J$ transformation matrix; $\boldsymbol{\Phi}(r_i)$ is a $J\times 1$ vector such that $\boldsymbol{\Phi}(r_i)=[\Phi_1(r_i),\Phi_2(r_i)\ldots,\Phi_J(r_i)]'$; and μ_j is a $q\times 1$vector that centers the basis functions. The center coordinates $\mu=[\mu_1,\ldots,\mu_J]$ cover the q-dimensional latent space uniformly. Model parameters are estimated using the EM algorithm [16].

One advantage of GTM is that, by construction, it lacks sensitivity to outliers. For tractability, the coordinates of each r_i are limited a priori to a finite set g of K possibilities, $r_i \in g=\{g_1,\ldots,g_K\}$ that covers the q-dimensional latent space uniformly. To decide which value for r_i generates d_i, GTM estimates the posterior probability, i.e., *responsibility*, that $r_i=g_k$. Given a prior probability that $r_i=g_k$ is 1/K for all k $\in\{1,\ldots,K\}$, let R_{ik} represent the posterior responsibility that latent variable r_i generates d_i, when $r_i=g_k$,

$$R_{ik} = \frac{\pi(d_i | r_i = g_k, W, \Phi())}{\sum_{l=1}^{K} \pi(d_i | r_i = g_l, W, \Phi())} , \quad (6)$$

In turn, GTM plots the posterior mode, expectation, or any quantile of r_i given specifications g and estimates for $\{R_{i1}, \ldots R_{iK}\}$.

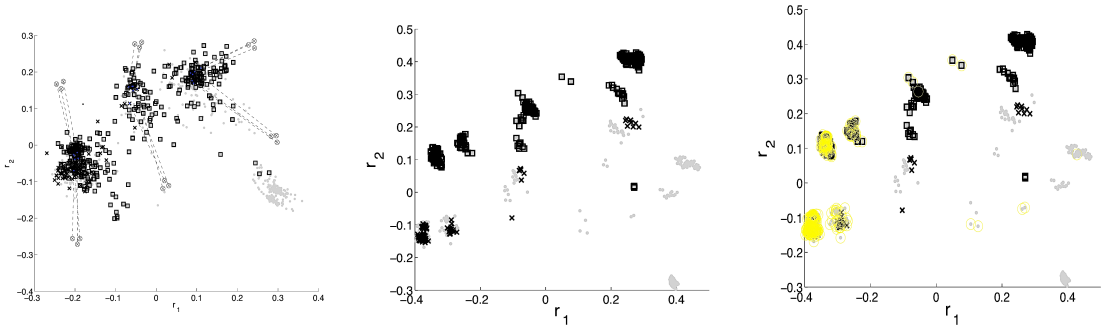


Figure 26 A sequence of visualizations derived through observation-level interaction with a modified MDS method. (Left) The user moves a set of points into new locations, communicating his intuition that there may be additional structure within each cluster. (Middle) The updated visualization showing new clusters. (Right) Highlighting showing the separation of income groups in the updated visualization.

User Guided GTM

GTM is a complex modeling approach that relies on many tunable parameters that are hard to interpret. User Guided GTM (ugGTM) will allow analysts to both take advantage of the benefits of GTM and guide the complicated GTM parameterization. Specifically, analysts may label, i.e., *tag* clusters, *tag* regions of the visualization space, and query differences in documents.

Here, we illustrate ugGTM within the context of an example. We have a collection of 54 abstracts from proposals funded by the National Institute for Health (NIH). After standard preprocessing, we apply a ranking system that we will call an Importance Index (ImpI), which is based on the Gini coefficient. ImpI considers both the frequency and uniqueness of words that are shared across documents and assigns a metric between 0 and 1. Entities that occur equally frequently in all the documents have ImpI=0 and entities that occur in only one document has ImpI=1. We selected the 1000 entities with the highest ImpI. One advantage of ImpI is that we can measure document similarity using Euclidean distance between proposals. Pairs of documents with small Euclidean distances

have comparable terms with similar frequency; and pairs of documents with large Euclidean distances have few, if any, words in common.

Table 2 Cluster tags (top 10 keywords) for NIH abstract groups.

Group A	tumors, brains, stem, treatments, patients, generations, drugs, ordering, controlling, therapeutics
Group B	stem, neuronal, brains, proteins, deliveries, regulations, neural, patients, differentiation, expression, treatments
Group C	stem, genetically, regulations, drugs, structurally, proteins, genomics, epigenetics, RNAs, complexities
Group D	Infections, treatments, tuberculosis, expression, patients, drugs, strains, resistance, vaccination, immunity
Shared by All Groups	cells, functionalization, diseases, developments, genes, cancerous , studying, researchers, proposing, mechanisms, specification

We apply GTM for J=16 and K=400 to obtain an initial display of the proposals, shown in Figure 27. Notice four clusters appear in Figure 27 that we labeled A, B, C, and D.

Tagging the Clusters and the Space. To understand the meaning of the clusters, we determine the words that both overlap the least within each cluster and have the highest ImpI's. Specifically, we apply k-means [46] to the low-dimensional data coordinates to determine cluster memberships. For each cluster we sum the ImpI vectors across the documents and rank the entities based on the ImpI sum. Entities ranked highest are those that 1) have importance in the corpus (as determined by the ImpI) and 2) have occurred most frequently. Given top rankings from each cluster, we delete those shared by all four clusters. Table 2 lists the unique key words that describe each cluster. Group A represents proposals that include brain related cancer studies and their clinical applications. Group B represents proposals related to human neural systems. Group C represents proposals that

address genomic and transcriptomic research problems. Group D represents proposals about infectious diseases, such as tuberculosis, and immunity.

As described previously in Equation (3), GTM characterizes high-dimensional data as random perturbations from a complex manifold $\boldsymbol{m}$; $E[d_i] = m_i$ for all $i \in [1,\ldots,n]$. To tag the visualization space, we select any spot, r^+, in the visualization and use Equation (4) to estimate its corresponding location on the manifold, m^+. The estimate m^+ will be a 1000×1 vector of ImpI's that we may use to rank the entities. We report the top ranked entities to tag the space. For example, in Figure 27, we pick up a spot r^+ (represented by a pink circle) that locates roughly at the center of cluster D. Several of the tagged top keywords overlap with the words describing cluster D.

Document-Based Query and Cluster Reorganization. It is common for users to assess documents by searching for keywords. However, keyword searching may be a tedious task and fail to reveal document clusters of interest. For example, keyword searches may identify documents with similar keywords, but used in different contexts; miss documents that contain combinations of the keywords; or prioritize words that have little relative importance for the user. In response to the challenges of keyword searching, many analysts rely on document matching. For document matching, entire documents can be used to identify which of the remaining documents in the corpus are most similar (to the chosen document). Hence such a matching algorithm is a document-based query of a corpus.

In our ugGTM, users may query documents in the corpus by dragging a document of interest directly in the visualization and watching how the remaining documents respond; e.g., similar documents will follow the document being dragged and dissimilar documents will repel. The behaviour of the documents is similar in spirit to Dust and Magnets (DnM) [79]. In DnM, analysts may drag or shake magnets that represent variables in the dataset and watch as relevant documents follow the magnets. However, a major difference between DnM and ugGTM is that when users drag documents (not variables) and watch how the remaining react, they are comparing documents based on

all of the variables in the dataset simultaneously. In turn, users may learn which variables are important for comparisons, based on tags within the visualization space.

The interaction is possible because ugGTM gives control to the users of some parameters in the model via the visualization. Let r^* represent the low-dimensional coordinates for a document that an analyst has chosen to drag. Given r^*, we add to the model described in Equations (3)-(6) by expanding sets g and $\mathbf{\Phi}$ so that $g=\{g_1,\ldots, g_K, g^*\}$ and $\mathbf{\Phi}=\{\Phi_1,\ldots,\Phi_J,\Phi^*\}$, where $\Phi^* = \exp\{-||r_i-\mu^*||^2/2\sigma^2\}$ and $g^*=\mu^*=r^*$. In turn, we assign the posterior responsibility (Equation 6) that r^* generates d^* via m^* to 1 (where, m^* is defined by Equation (4) so that the mapping between the low- and high- dimensional coordinates for the moving observations is deterministic.

To propagate the effect of moving r^* to the remaining visualization, we take a local regression approach [32] to characterize high-dimensional data $d_i \mid \{ r_i=g^*,m^*\}$ in that we scale d_i-m^* by the square-root of function V given scaled distance $\Delta_i = ||d^*-d_i||/c$ so that,

$$\pi(d_i|r_i = g^*, W, \Phi) = \left(\frac{\beta}{2\pi}\right)^{-p/2} \exp\{-\frac{\beta V(\Delta_i)}{2}\|d_i - m^*\|^2\}$$

where c is user-defined; e.g., $V(\Delta_i)=\Delta_i^2$ and c=0.5. In turn, both posterior responsibility estimates (Equation 6) and estimates for $\boldsymbol{m}$ (Equation 4) change. Let $m_i^{(c)}$ and $m_i^{(u)}$ represent the current and user-adjusted manifold estimates for observation i. We define the BaVA-GTM estimate for the manifold, $m_i^{(c+1)}$, by

$$m_i^{(c+1)} = \delta_i m_i^{(c)} + (1-\delta_i) m_i^{(u)},$$

where $\delta_i=||r_i-r^*||/b$ and $b=\max\{||r_1-r^*||,\ldots,||r_n-r^*||\}$ so that $\delta_i \in [0,1]$. This definition for $m_i^{(c+1)}$ controls the visualization so that only the regions of interest respond to user interactions; areas that are distant from the dragged observations do not change.

Parameters g^*, Φ^*, $V(\Delta_i)$, δ and $m^{(c+1)}$ in ugGTM work together in the following way. When a data point d_i is far from d^*, $V(\Delta_i)$ will be large and thus decrease the posterior responsibility (Equation 6) that $r_i=g^*$ generates d_i. Similarly, when d_i is near d^*, the corresponding responsibility will increase. Increases in the responsibility for $r_i=g^*$ will

cause the coordinates for r_i to gravitate toward r^*. Thus, analysts may specify constant c in our definition Δ_i, depending upon how many document matches they seek for the moving document. Also, the degree to which the observations gravitate toward r^* is determined by δ and $\boldsymbol{m}^{(c+1)}$. When the manifold shifts from $\boldsymbol{m}^{(c)}$ *to* $\boldsymbol{m}^{(c+1)}$, the meaning of the visualization space changes, as we demonstrate in our example.

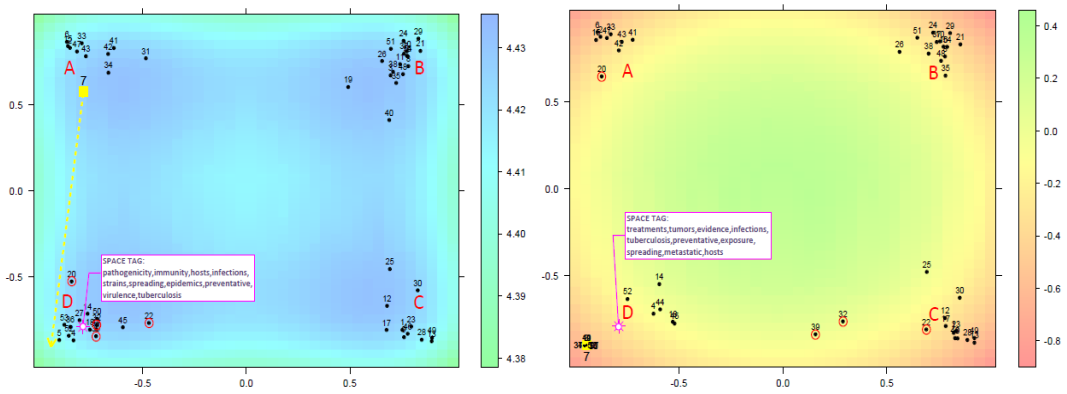


Figure 27 GTM display of the NIH abstracts. Black dots mark documents and labeled by their document ID. (Right) shows the updated view after performing an observation-level interaction on the shown points.

Example

For our NIH example, we apply ugGTM. We display an initial GTM view of the documents (the 54×1000 dataset) in Figure 27. Suppose a user identifies a specific document of interest, e.g., Doc 7 to investigate. A preliminary investigation might involve a sequence of non-spatial interactions, such as, searching of multiple keywords, reading all or part of the document etc. However, a comprehensive assessment of the document may require spatial interactions as well. The user might explore space tags across the screen and determine a more appropriate location for the document of interest. In this case, Doc 7 is closer to Group A, and is about developing new brain tumor therapies and tumor stem cell quiescence. The keywords this document shares with group A include tumors, brains, cancerous, therapeutics and chemotherapy. However, since Doc 7 relates to therapy developments for disease, it shares some keywords with Group D; e.g., treatments, strategies, patients, drugs, resistance, clinically.

As an exploratory spatial interaction, the user drags Doc 7 to the lower left corner of the display and watches how the remaining documents react. By repositioning Doc 7, the

user redefines the spatialization of the screen, i.e., modifies the space tag corresponding to a location. For example, when we tag the same coordinates r^+ (r^+ are the coordinates of the space tagged in Figure 27), we learn that the top keywords include treatments and tumors as well as those that were there earlier. Recall that ugGTM uses every variable in the dataset to compare documents. For this reason, documents that mention stem cells and other important keywords in Doc 7 follow Doc 7. As expected, many documents in Group D gravitate toward Doc 7. However, a few documents in Group B also followed. Future work will allow users to weight the keywords in Doc 7, if desired. Also Documents with ID 20, 22, 32 and 39 change locations. Important keywords for these documents include the following: Doc 20 discusses diagnosis of HIV infection in patients who live with limited access to therapeutic treatments; Doc 22 discusses expression characteristics of a drug-resistant gene; Docs 32 relates to varying yeast strains; and Doc 39 relates to Lymphocyte Homing. Docs 20 and 22 repelled against Doc 7 because the redefined-manifold down-weighted their important entities in the lower left corner and up-weighted the entity tumor. Thus, Doc 20 and Doc 22 shifted to Groups A and C respectively. Docs 32 and 39 are separated slightly from Group D and gravitated toward Group C because they have a few words in common with each group, but not enough to place them in either corner.

An interesting note about the updated manifold is the change in shape or magnification factor [68]. The color in the background is plotted based on the logarithm of the magnification factor evaluated on a fine grid that covers the visualization space. Due to the nonlinear mapping from r_i to m_i, equal distances in the visualization do not necessarily imply equal distances in the high-dimensional space. The magnification factor describes the rate of change between distance or area in the latent space and the corresponding distance or area on the manifold and can be interpreted as a description of how wiggly the manifold is. Overall, the magnification factor is lower in the updated view (right illustration in Figure 27) and the clusters formed are mainly in low magnification areas. This means the initial clusters are in flat, stable regions of the estimated manifold. Thus, observations in these clusters are closer to one another than observations shown initially.

4.3.2 Discussion

We present a comparison of key characteristics of the methods used in this paper in Table 3. Again, the purpose of this work is not to make a direct comparison of these three methods, but rather to present how to apply observation-level interaction to each method.

Mappings. The three methods discussed in the paper provide us with a spatialization of the data within the bounds of their algorithmic complexity. Points that are close in the higher dimensional space remain close to each other in the visualization in all the algorithms although the concept of proximity varies depending on the algorithm. As an artifact of the algorithms, in both PPCA and MDS, the high dimensional data is assumed to be a linear mapping of the visualized representation while GTM is a non-linear mapping of the same. Hence, the same dataset might provide widely disparate visualizations for different algorithms. Spatially this might translate to the fact that based on the algorithm, the user's spatial interaction might target different sets of observations. Each algorithm can potentially have its own set of diagnostics overlaid with the visualization that might aid the user in understanding the proximity of the data in the higher dimensions; e.g. visualizing the magnification factor along with the data in GTM indicates the level of distortion. The goal of the user is to obtain a view in multiple steps that matches with his mental model irrespective of the algorithm used to visualize the data. The specific steps that the user goes through should be immaterial in so far as the final visualization is concerned and all the algorithms discussed here have the flexibility to provide that.

PPCA relies on the assumption that a single linear projection exists that can reveal useful structure. MDS provides a two- dimensional representation of the observations via penalization of any distance distortion that happens in the two-dimensional representation using a stress function. However, the linear projection assumption may not hold for complex datasets or, the visualization based on minimizing stress in MDS might not reveal all the information in the data. In PPCA, using variance to select the direction in

which to project data makes sense for datasets with a global linear structure [70]; the projection will minimize the number of observations that overlap so that they are as visible as possible.

Table 3 Comparison of the methods used in this paper. Each method has different characteristics, that are more suited for different tasks and datasets.

	PPCA	**MDS**	**GTM**
Mapping Type	Linear	Linear	Non-linear
Method Characterisation	Variance	Similarity	Manifold
Distribution Assumption	Probabilistic	Deterministic	Probabilistic
Scalability (Observations)	★★★	★	★
Scalability (Dimensions)	★	★★★	★★
Conceptual Clarity	★★★	★★★	★
Running Time	★★	★★★	★
Outlier Robustness	★★	★★	★★★

★★★ = Good ★★ = Average ★ = Poor

However, variance estimates and hence PPCA visualizations are sensitive to outliers and it is not uncommon for PPCA to display one or two outliers and a cloud of occluded points. Under Euclidean distance, MDS is algorithmically the same as PPCA and will suffer from the same sensitivity to outliers. Assessing such a visualization and making appropriate adjustments would be, at best, challenging. Thus, a more complex methodology is often needed to summarize datasets, e.g., mixture PPCA or GTM. GTM being a topographic mapping places the outliers at one end of the screen or at a position that is distant from the region that has more structure. In our interactive framework, outliers can be brought closer to existing user defined clusters through redefining the principal components in PPCA, reweighting of the dimensions in MDS and constraining *responsibilities* in GTM; in all the cases the user's observation-level interaction initiates the parameter update.

Scalability. In terms of time complexity, GTM is O(KND) (K number of latent points, N number of observations, D data dimensionality), PPCA is O(qND) (q is the dimension of the latent space, usually equals 2) and MDS varies from O(qND) to $O(N^3)$. The effect of high dimensionality (i.e. the number of columns for every observation) on the run-time

will be similar for all three algorithms. The challenge in scalability (large N) is also of the same order for the three algorithms when Euclidean distance is used.

However in the design of a visual analytic system that incorporates user interaction in the framework, the choice of the algorithm should be based not only on the run-time of the algorithm but also on the cost incurred in converting the observation-level interaction or feedback to updated values of the parameters for the method. In PPCA, it is the cost of evaluating the feedback matrix $f^{(p)}$; in MDS it is the cost of obtaining optimal feature weights $\boldsymbol{w}$ based on pair-wise distances of the observations that the user has moved; and in GTM, it is the cost of computing distances between data points and reference vectors. Under such considerations, we think MDS provides the quickest and easiest two-dimensional visualization of the data, followed by PPCA and GTM.

We maintain a probabilistic framework in PPCA and GTM. Specifically for PPCA, computation is quick since the primary parameter of interest Σ_d has a posterior distribution and a conjugate feedback distribution, and MAP(Σ_d) can be computed without MCMC. Thus, analysts can explore the data in real time. GTM (although being most flexible in handling more complicated data occlusion issues that challenge MDS or PPCA) is based on an expectation-maximization algorithm and hence needs more run time to converge to the optimal parameter value.

Sensitivity. The methods described in this paper will respond based on the interaction performed (i.e., number of observations moved, distance the observations were moved, etc.). For example, moving a single observation will generally result in a less drastic change in the layout compared to a similar interaction performed on a cluster of observations. Thus, the sensitivity of the models in terms of responding to the user's intuition is dependent on how large the change or update is provided by the user's interaction, the size of the dataset, as well as if the data supports the suggested updated layout. The methods will attempt to find the "best fit" given the user feedback, but will maintain mathematical validity (i.e., users cannot force the layout if the data does not support it). The result is such that the system balances the user's intuition with the

structure of the data to reduce bias. The goal of these techniques is not to converge on a single structure or layout, but rather to allow exploration of many possible structures.

Interaction. The examples of how observation-level interaction can occur within spatializations in this paper show only one form of interaction available to users within spatializations – movement of individual observations. The methods are expandable to allow more complex interactions, such as moving clusters of observations, annotating a region of the spatialization, and other interactions used for communicating the intuition of the user to the system. In a fully implemented visual analytics system, these interactions may include queries, highlighting, and other interactions from which analytical reasoning of users can be interpreted.

Implementation. The prototype visualizations shown in this paper are intended to provide working examples of the modified methods. Through the use cases, we highlighted how an end-user might interact with such systems. We plan to integrate these methods into more fully functional visual analytics tools. That will allow us to perform a series of user studies to evaluate the usability and effectiveness of observation-level interaction in terms of providing insight to users, and supporting the sensemaking process.

4.3.3 Conclusion

We described how modifications of powerful statistical methods allow one form of semantic interaction, observation-level interaction. By interacting within the visualization through movement of observations, users are able to perform exploratory and expressive interactions. Thus, users are able to perform sensemaking tasks, such as hypothesis validation, directly within the spatial metaphor. By keeping the interaction at the observation level, users are not required to transform their sensemaking into a combination of statistical parameter updates.

In particular, we modified PPCA, MDS, and GTM using BaVA [38] and V2PI [44] approaches, so that users can focus on their spatial analysis of data rather than directly

updating statistical parameters of models. We present three examples (one for each modified method) that illustrate the effectiveness of these new models. Based on the positive results in this paper, as well as the lessons learned, coupling interaction with statistical models provides an opportunity to explore additional forms of spatial interaction for visual analytic applications.

4.4 Discussion

4.4.1 Unifying the Sensemaking Loop

With the fundamentally different role occupied by semantic interaction, we explore a new design space for interaction in visual analytic tools. With the addition of soft data, and a model capable of interpreting the user's analytical reasoning, we leverage interactions that are already occurring in the spatial analytic process to further aid users in their sensemaking process.

With semantic interaction, the amount of formalization between foraging and sensemaking on the part of the user is reduced. For instance, in moving a document, users can formulate a hypothesis based on that document, expecting similar documents to follow. ForceSPIRE attempts to update the layout based on the interaction, and gives the user feedback. Thus, the foraging stage occurs as a result of the hypothesis being formed through semantic interaction**.** By not forcing users to over-formalize their analytic reasoning too early in order to forage for the relevant information, semantic interaction creates a more seamless transition between foraging and synthesis, unifying the sensemaking loop.

4.4.2 Future Directions

Semantic interaction, as a concept, opens up many possibilities for further research, such as: what interactions to capture and store, which parameters of the model to update, how to store the soft data, and which models present a metaphor that can be extended upon.

In order to make more concrete claims regarding the usability and effectiveness of ForceSPIRE (and thus, of semantic interaction), a formal user study is needed. Our plan is to introduce ForceSPIRE to professional intelligence analysts and have them solve scenarios that model their daily task, such as one of the VAST datasets [58]. The observations and feedback from these users will provide ecological validity for semantic interaction.

4.5 Conclusion

In this paper we have discussed how the concept of semantic interaction leads to a new design space for interaction in spatializations of textual information. Semantic interactions occur directly within the spatial metaphor, support spatial cognition, and exploit spatial analytic interactions. We describe semantic interaction, discussing the three components required – capturing the interaction, interpreting the analytical reasoning, and updating the mathematical model. Further, we present ForceSPIRE, designed for semantic interaction with textual information, discussing its functionality and demonstrating how it can be used through a use case. Lastly, we discuss how semantic interaction has the opportunity to unify the sensemaking loop, creating a more seamless analytic process. In allowing users to interact within the spatial metaphor, they can remain more focused on their analysis of the data, without having to become experts in the underlying mathematical models of the system.

Chapter 5

Evaluating Semantic Interaction

Semantic interaction can provide computational support for sensemaking through model steering. In this section we present the results of a user study exploring ForceSPIRE's ability to address these challenges. How can ForceSPIRE systematically quantify the reasoning process of users by building and modifying an entity-weighting scheme? Additionally, we explore if the weighting scheme aided the system in adjusting the spatialization in accordance with the user's analytical reasoning. Finally, how did users interact with the system during their analytic process (i.e., were they focused on adjusting the weighting scheme, or focused on synthesizing information)?

Our results show that each user's weighting scheme was updated in accordance with that user's reasoning at specific times during the investigation, and that it provided the flexibility for this scheme to adapt to the dynamic process of each user. The updates to the spatialization based on semantic interaction provided support for each user's process, such as suggesting which documents to read next, incrementally determining the meaning of a cluster, and promoting the re-visiting of information. Users conducted their investigation by utilizing the semantic interactions to synthesize information, without focusing directly on the weighting scheme. The final spatializations generated in

ForceSPIRE were co-created by the user and the system, as evidenced by a mixture of user-defined document locations, and model-defined locations, and were representative of the findings of the user as evidenced by their debriefing. These positive results suggest that semantic interaction in ForceSPIRE provides meaningful computational support for sensemaking. As a result, users are able to focus on the synthesis of information in the spatialization, while the system provides computational foraging support suited to the user's analytic process.

5.1 Method

This user study investigates the following research questions about the capabilities and benefits of semantic interaction:

1. How well can semantic interaction systematically quantify analytical reasoning based on user interaction as a dynamic entity-weighting scheme?
2. How does the real-time modification of the weighting scheme and adjustment of the spatialization aid users' sensemaking?
3. What was the focus of users while exploring the dataset through semantic interaction? That is, were they focused on adjusting the weighting scheme, or synthesizing information?
4. How does the co-created spatialization map to the users' findings?

We hypothesize that the coupling between the semantic interactions and model updates will create a dynamic weighting scheme that appropriately captures the analytical reasoning of each user throughout his or her investigation. As a result, this weighting scheme will adjust the spatialization, aiding in the co-creation of the layout, where users need not develop the entire layout manually, but also not rely on solely algorithmic generation. During this process, this will help users by adjusting the layout while users read documents and synthesize the information, bringing related documents nearby. Also,

we hypothesize that the soft data captured during the analysis will be representative of the analytic product of each user, and therefore the co-created spatialization will be meaningful to the user. Throughout this process, we hypothesize that the users will remain focused on the synthesizing of information, rather than interacting to directly modify the weights of entities.

5.1.1 Equipment

For this study, we used a large, high-resolution display (shown in Figure 6). Such workstations allow users of ForceSPIRE to leverage the additional resolution to show many text documents at full detail, and the additional physical size to provide users with a more embodied analytic experience [4, 20]. This particular workstation is constructed using 8 30-inch displays, driven by a single node, providing a total workspace resolution of 10,240 x 3,200 pixels. The curvature allows easy access to all areas via physical navigation, such as chair rotation [23, 65].

The dataset used for this study is an analysis exercise called Atlantic Storm developed for the purpose of training and evaluating intelligence analysts, as well as analytic tools. The dataset consists of 111 text intelligence reports containing a fictitious terrorist plot. Using LingPipe [1]to extract keywords (i.e., entities) from these documents, 294 unique entities occurring more than once in the dataset were extracted (singletons were removed). The choice to use this dataset is based on the ability to have a realistic dataset, containing a known ground truth against which to compare the findings of the users, while requiring no detailed domain knowledge beyond English reading comprehension and creativity.

5.1.2 Data Collection and Analysis

ForceSPIRE has the ability to log the soft data used for semantic interaction. For the purpose of this study, this gives us a record of every interaction performed by the user, as well as how the system interpreted the interaction in context of the dataset. For example, when a user highlights a phrase, the soft data shows us when the highlight occurred, what

the text is, in which document, as well as what entities' weights changed, and what the new weights are.

The users were asked to provide us with verbal feedback throughout their process. In a post-study interview, subjects explained their findings, the resulting spatialization, and insights about their process that may have been missed during the think-aloud protocol. Video recordings and screenshots were also taken during each task for post-study analysis.

Table 4. Semantic interaction counts during each user's analysis.

	User						
Interaction	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Search	13	32	37	14	38	21	155
Highlight	47	58	12	10	5	0	132
Expressive Movement	45	76	47	62	26	27	283
Exploratory Movement	41	102	64	26	98	43	374
Annotation	3	40	3	0	0	0	46
Total	149	308	163	112	167	91	

5.1.3 Procedure

This study consisted of observing 6 computer science graduate students. The age of the participants ranged from 27 to 38, with an average age of 30.

Each participant was given a brief overview of ForceSPIRE, using a practice dataset, for the purpose of making each user familiar with how ForceSPIRE and the supported semantic interactions function. Upon informing us that they were comfortable, each user was given verbal instructions on their task. Each user was given the same initial view of the Atlantic Storm dataset in ForceSPIRE as a starting point. We informed the participants that they had a maximum of 90 minutes to analyze the dataset, after which they will be debriefed regarding their investigation. They were allowed to finish early if they felt they were finished before time expired.

5.2 Results

The success of a visual analytic tool hinges on the ability to provide support during the analytic process, as well as a meaningful representation of the user's findings. Thus, the results of this study are presented in terms of the analytic process and product. The analytic process describes how the semantic interactions within ForceSPIRE were used during the analysis, how the corresponding model and spatialization updates benefitted the users, and how the soft data mapped to each user's process. The analytic product details how the findings of each user are represented in the final spatialization, as well as the final weighting of keywords.

5.2.1 Analysis of Process

Each user's process was different, and thus utilized semantic interaction differently (Table 4). However, the analysis of each user's process reveals general usages of each semantic interaction. To address the research questions, we present the analysis of the processes of the users in terms of **usage** (how and when they used each semantic interaction), **reasoning** (what was their purpose for interacting, sensemaking or model steering), **impact on weighting scheme** (how the updated weighting scheme coincided with their reasoning process), and **impact on spatialization** (how did the updating spatialization benefit their analysis).

Table 5. Number of entities added via semantic interaction during each user's investigation. The majority of these new entities (92%) maintained a weight above 0 throughout their process.

	User					
	1	**2**	**3**	**4**	**5**	**6**
Entites Added	43	62	35	13	15	10
Weight > 0	38	54	35	13	14	10

5.2.1.1 Semantic Interaction Usage

Performing a spatial analysis of data focuses around rearranging documents and creating spatial constructs or clusters [4, 27]. As such, **pinning** and **document movement** (both *exploratory* and *expressive*), were the fundamental methods of exploring the dataset. Pinning documents to absolute positions in the spatialization was used to create "spatial landmarks". That is, users pinned a document to a specific location in the layout to create (and maintain) meaning of a specific region of the spatialization. Based on these landmarks, document movement was used to organize the spatialization based on the user's intuition. For example, User3 pinned a document mentioning "Nassau" in a specific location. From there, he placed other documents related to "Nassau" nearby, and also quickly re-acquired these documents when needed.

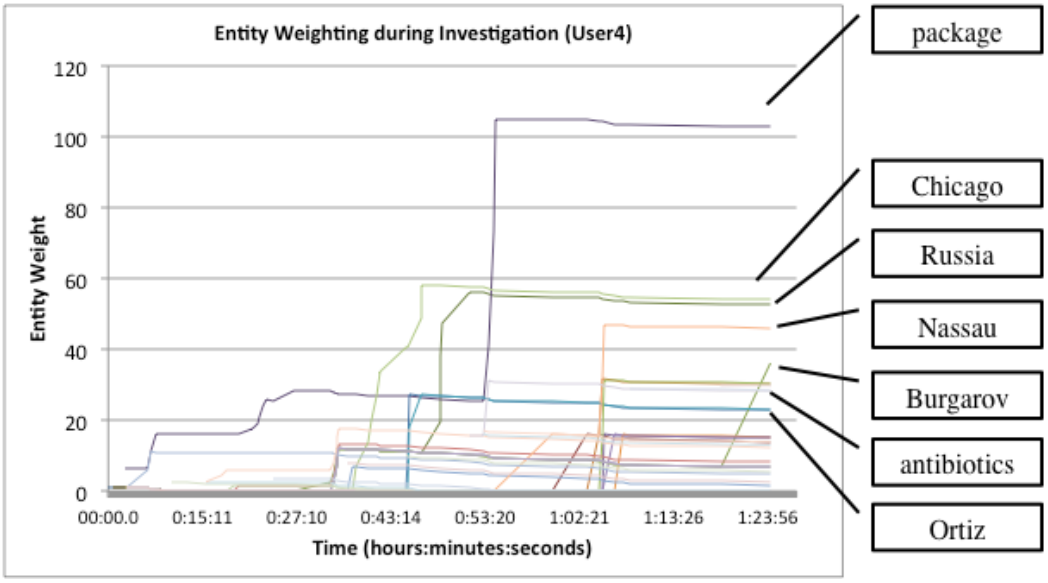


Figure 28 User4's entity weighting over the duration of his analytic process. Semantic interactions in ForceSPIRE adjusted the weights of entities to coincide with his investigation of multiple hypotheses. As a result, the layout adjusted incrementally with each interaction.

Highlighting was used mostly while reading a document to indicate terms or phrases that "stood out". These highlights were beneficial to users to produce visual and cognitive aids. The highlighted phrases (mostly single words and fragments of sentences) helped remind users of what information was important in a document when re-acquiring the

document later. User6 was the only user who did not perform any highlighting during his investigation, simply stating that he "did not feel a need to."

Search was used to find other documents containing a term of interest to the user. Generally, users performed a search on keywords for two reasons. First, the unique color assigned to each search provided a quick overview of where in the dataset the term occurred given the current spatial layout. Second, users treated the search window (of which more than one could be opened) as a means to "tag" the space. For example, all of the users commented that leaving multiple search windows open and pinned to specific locations was an effective way to recall the meaning of that specific position in the spatialization.

With the exception of User2, **annotation** was rarely used. User2 said that he enjoyed the ability to "add personalized notes to important documents." In his case, ForceSPIRE detected 23 entities in his annotations (that were not in the dataset), including entities such as "irrelevant", "suspicious", and "revisit" (extracted from a note stating "should revisit this later"). During his investigation, he also found it useful to track what documents he found important by scanning the workspace and seeing which documents had the yellow notes window visible. Thus, annotations can be helpful to some users, while others prefer to utilize other interactions to support their analysis.

5.2.1.2 Aiding the Sensemaking Process

The primary benefit for sensemaking provided by ForceSPIRE was aiding the user in adjusting the layout by bringing related information nearby. Each semantic interaction in ForceSPIRE is tightly coupled with the dynamic weighting scheme used by the force-directed model responsible for generating the spatial layout. As such, ForceSPIRE responds to each interaction via updating the spatialization as a result of the updated weighting scheme.

Each user's process involved multiple stages of the investigation, including exploring specific leads (e.g., a person, place, etc.) and hypotheses regarding the dataset. As such, it

is important for semantic interaction to allow a flexible entity weighting scheme to support exploring each of these aspects during different times of the investigation. For example, while a user investigates information regarding the entity "Atlanta", the weight of entities similar to (and including) "Atlanta" should increase. If the user chooses to investigate "weapons" at a later time, the weighting scheme should reflect this change. The challenge then, comes in supporting the rapid and fluid change of what is currently being investigated by a user through rapidly changing keyword weights, while maintaining a history of the previously emphasized keywords.

Table 6. Each user's top 5 entities, collected both from the user's debriefing (*user*), and based on the final entity weighting (*model*). Underlined entities indicate a match between the user and model. **Bold** entities were entities added to the model as a result of semantic interaction during the analytic process (i.e., missed by the initial entity extraction).

1 (*user*)	**1 (*model*)**	**2 (*user*)**	**2 (*model*)**	**3 (*user*)**	**3 (*model*)**	**4 (*user*)**	**4 (*model*)**	**5 (*user*)**	**5 (*model*)**	**6 (*user*)**	**6 (*model*)**
diamonds	**diamonds**	package	Nassau	explosives	Nassau	diamonds	package	Al Queda	Nassau	diamonds	**weapons**
scholarship	**weapons**	Hanif	Hanif	weapons	**students**	Nassau	Chicago	Caribbean	Miami	antibiotics	**diamonds**
jihad	**graduate**	antibiotics	Freeport	Nassau	**weapons**	Burgarov	Russia	Russia	Freeport	Ortiz	Nassau
weapons	Jamal	diamonds	Miami	students	**scholarship**	antibiotics	Nassau	Hanif	Apple St.	Bahamas	**graduate**
freeport	Nassau	Nassau	**package**	scholarship	Jamal	package	Burgarov	Odeh	**weapons**	Hijazi	Hijazi

For example, Figure 28 shows the temporal history of User4's keyword weighting during his analysis. The patterns and trends observed in User4's analysis of the soft data is also representative of the other users' history. One can see that approximately two minutes into the investigation, the entity "package" was created. The creation occurred while User4 read a document and found the phrase "carefully wrapped package" important, and thus highlighted it. The effect on the layout was that documents containing the entity "package" were brought closer. He did not immediately switch to reading those documents, instead continued to read the document while other related documents came nearby. This strategy was found in other users' processes also. "It was nice to see what

documents would come near while I was reading and highlighting", User1 told us after his investigation. He continued to tell us that he would notice other documents coming closer, but would "continue reading and highlighting until I finished that document, then decide where to go next depending on what's close by." Upon finishing reading the document, User4 pinned it, and chose the closest document to continue his investigation. This document was one related to "package", and important to the plot. He continued reading three more documents containing "package", highlighting other phrases that contained the term. As such, the term continued to increase in weight, and related documents continued to form more tightly around the one that was pinned.

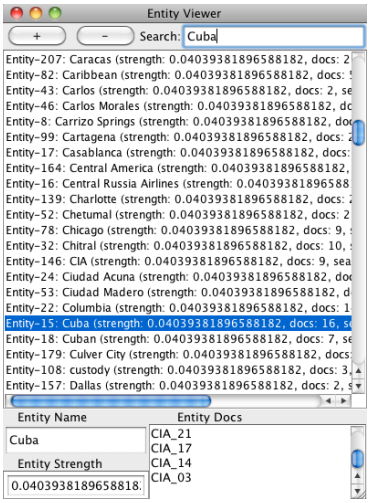


Figure 29 The "Entity Viewer" in ForceSPIRE allows users direct control over the weights of entities, adding entities, and removing them. With semantic interaction, this view was never needed.

Figure 28 also shows instances when User4 explored other potentially relevant information. For example, at 36 minutes into his analysis, he informed us that he wanted find out more information regarding events in "Chicago". This stemmed from reading a document that mentioned "Chicago". He highlighted the single word, and immediately pinned the document in a specific location, away from other documents. Then, he searched for the term "Chicago", and placed the search window next to the pinned document. Then, he opened and read some of the documents containing Chicago (they were highlighted teal from the search) that came closer. The first 2 documents he read, he

placed near the first pinned document. As a result, the weighting of "Chicago" increased, but so did the weights of other related entities, such as "Russia", "weapons", and "Panama City". This occurred because the documents he dragged near the pinned document containing "Chicago" were not similar based on only "Chicago", but those other entities as well. As a result, more documents came closer that did not contain strictly "Chicago", but were related. In this case, he read some of the documents containing "Russia" and "weapons" and moved them into another, separate location. Again, ForceSPIRE responded by moving the documents more similar to "Russia" and "weapons" into that location, rather than near the location regarding "Chicago". This benefitted the user as he noticed how some documents remained in the middle of the two areas, showing relevance to both topics. These were documents connecting these two, and important to the plot.

Towards the end of his investigation (approximately one hour into it), he focused on tying all the pieces of evidence he had collected together. He did so through exploring the relevance of "Bahamas" and "Nassau". He did so primarily through small, local movements within an area specific to each. He arranged the documents within the region to reflect a sequence of events related to transportation of a package. In addition to the weight increase of those entities, he discovered the relevance of some of the key persons involved in the suspicious activity (i.e., "Odeh", "Hanif", and others). He found an important document detailing how some of the weapons (which he investigated earlier) were possibly being transported by students funded through a suspicious scholarship fund. The history of his weighting scheme reflects each of these hypotheses and branches during his investigation, as indicated in Figure 28.

Table 7. Pinned and unpinned documents and search windows in each user's final spatialization.

		User					
		1	**2**	**3**	**4**	**5**	**6**
Docs.	***Pinned***	34	41	44	28	32	34
	Un-Pinned	77	70	67	83	79	77
	Detail	85	88	89	36	40	49
	Minimized	26	23	22	75	71	62
Search Windows	***Pinned***	9	19	31	2	16	5
	Un-Pinned	0	0	0	1	1	0

In addition to increasing the weighting of entities that were relevant to the investigation, semantic interaction also reduced the weight of entities that were not. ForceSPIRE does so through "decaying" the weights of previously emphasized keywords over time. That is, as other keywords are emphasized via various semantic interactions, weights of previously emphasized keywords will begin to decay. Therefore, if they are never investigated again, they will eventually return to their lower weight, but if they are revisited again later, they will increase in weight again. At times, this resulted in the weights of those entities going to zero (thus having no impact on the spatial layout). Across all users' processes, the average number of entities where this occurred at least once is 245 (out of the 294 unique entities initially extracted by ForceSPIRE). While these entities did not have an impact on the spatial layout when their weight was set to 0, subsequent semantic interactions continued to use these entities to measure similarity. As a result, entities that may not have been relevant during the early stages of an investigation and were relevant to a later hypothesis being explored, saw their weight increased. An example from User4 is the term "Nassau", which was not relevant to his investigation until approximately 54 minutes into the study, where the weight increased from zero (as shown in Figure 28). This happened as a result of him dragging one document close to another on the basis of both being about an event in the "Bahamas". ForceSPIRE interpreted this similarity, but also found these documents similar because of "Nassau", increasing the weight of the term and bringing those related documents nearby.

Semantic interaction aided two users from this study in **creating a "junk pile"** (i.e., a collection of documents that are not relevant to the main plot, and are thus placed in a location away from the relevant information). As these two users placed more information into the same cluster that they referred to as "junk", ForceSPIRE calculated the similarity between the documents being placed into this cluster and increased the weight of those entities. "Look! It's moving other junk into my junk pile for me" User1 remarked. However, he was sceptical of the system's ability to detect irrelevant documents, so he opened and read a few of them as they moved closer. Some, he agreed with being junk and left them in the junk cluster, while others he moved near other

pinned documents in the spatialization. By doing so, he continued to improve ForceSPIRE's ability to detect irrelevant documents. When asked about this experience after the study, he told us that the more he interacted with the layout (including his "junk pile"), the more pleased he became with the metrics for determining junk, and the "more [he] trusted it".

An important capability in ForceSPIRE is the **steering of the entity extraction algorithm** for generating additional entities during an investigation. Entities can be added to the system through semantic interaction, which was critical to the ability to capture and infer the users' reasoning processes. While the entity extraction algorithm in ForceSPIRE managed to extract 294 unique entities, each user found additional entities that were relevant to their analysis. Table 5 shows the number of entities created as a direct outcome of semantic interaction. Of these, most (92%) maintained a weight greater than zero throughout the investigation. This shows that not only was it important for users to steer the weighting of existing, extracted entities, but also to steer the entity extraction algorithms to generate additional entities. For example, User3 highlighted the phrase "he has students now in the USA", which was passed through a more aggressive entity extraction algorithm, and detected "students" as an entity. This entity was important to the user's findings, as well as highly weighted in the model (Table 6).

Pinning and un-pinning documents was used not only to place meaningful documents in absolute positions in the workspace, but also to check if the current weighting model would place the document in another region (or into another cluster). For example, three of the users commented that they un-pinned a document to see where it went after it had been pinned for a long time. They were interested in other possible topics it might relate to. If nothing particularly interesting was found, they returned the document to the previous location and pinned it again. However, often users found relationships between these documents and other clusters, and typically either left the document un-pinned, or pinned it in a different location from where it was pinned previously. For instance, User1 found that he had a document pinned from very early in his investigation that referred to the Freeport Star Hotel. When he un-pinned it, he saw it go near other documents about

the Bahamas and Nassau, which helped him make the connection about the events happening in that area.

In general, users emphasized the importance of observing the spatialization adjust incrementally. That is, to notice the change in relative distances between documents as a result of the highlighting they did while reading, searching, etc. Such exploration can be found in other tools, such as VIBE [52] or Dust&Magnet [79], where users can place "points of interest" corresponding to keywords in specific locations, and observe how the spatial layout adjusts given those keywords and locations.

Users did not treat the semantic interactions as a means to directly manipulate entities. That is, they interacted as a means to synthesize the information. For instance, based on their comments, highlighting was performed not to pass a phrase through a more aggressive entity extraction algorithms, but to emphasize a part of the text as being important, so as to be able to find it again more easily later. "Oh, that's important … [I] might need to come back to [it]", one user stated while highlighting a phrase. None of the users found the need to directly manipulate entities (e.g., adjust the weights, add, remove) via the "Entity Viewer" (shown in Figure 29). All users were shown this feature in the training, but none found it necessary to use during their actual investigations. These results evidence that semantic interaction properly coupled the semantic interactions with model updates, to the extent that users never felt the need to do so directly. This contrasts with the intended usage of other tools, such as IN-SPIRE [53], where model steering occurs via direct parameter manipulation on the part of the user. With ForceSPIRE, users were successfully able to focus on the synthesis necessary for sensemaking, while the parameter adjustments occurred systematically in accordance with their analytic reasoning.

5.2.2 Analysis of Product

At the conclusion of each trial, we asked the user to describe his findings (both in terms of what information is relevant to the suspicious plot, which all users found, and information that was not). ForceSPIRE remained visible during this debriefing, but we

asked users to not interact with the tool, but simply use the final layout as a means to help describe their findings. The analytic product with regards to this study refers to the final spatial layout in ForceSPIRE, as well as the user's debriefing after the study. We analyze this information in addition to the final weighting scheme (i.e., soft data) at the conclusion of the analysis. Compared to the known ground truth of the dataset, each user found the suspicious activities, with varying amounts of detail to support the findings. Thus, the analysis of the product here is with regards to ForceSPIRE's ability to create synergy between what the system and the user knows (rather than compared to the ground truth).

The final weighting schemes at the conclusion of each user's investigation served as a good approximation of the keywords relevant to their findings (see Table 6). This table shows the top 5 keywords from each user's debriefing and the top 5 entities based on entity weight, labeled "user" and "model" respectively. The entities from the debriefing were given to us by the user as part of their debriefing to represent their findings. The entity weights were obtained by taking the top 5 highest weighted entities in the final soft data state. The entities highlighted in bold are entities that were not initially extracted, but added to the model through the user's semantic interaction during the study.

These results reveal that 47% of the entities match directly from the user's findings to the highly weighted entities. In addition, 11 of the highly weighted entities were added to the model as a result of semantic interaction. Of these 11, 7 were important based on the debriefing of the user (i.e., they matched with the top 5 entities given to us by the user). Therefore, not only were these entities added as a result of semantic interaction, but they were also relevant to the user's findings.

In some cases, there is not a direct match between the entities obtained during the debriefing and the entity weighting. For example, User5's entities show no direct matches. However, a more sophisticated entity correlation algorithm may find connections between entities such as "Caribbean" and "Nassau", "Freeport", "Miami", and "Apple St." (an address in Nassau in this dataset). Thus, even when there are not direct matches, we find that the higher weighted entities provide a good estimate of

characteristics of the dataset users found important and relevant to their investigation. In fact, this indicates that the system was successfully able to interpret the user's reasoning within the context of the actual data. That is, the system identified keywords relevant to the user's process that the user did not think of or at least had used other words in place of. This suggests that the system fulfills the needs of incremental formalism and ill-defined user-generated clustering [27].

5.2.2.1 Spatialization Co-Creation

One of the goals of semantic interaction is to properly steer the underlying model to allow for co-creation of the spatialization between the user and the model. As such, we hypothesized that some documents (and search windows) would be pinned by the user to maintain their absolute position in the workspace, and others would be un-pinned so that the model could determine their position based on the current entity weighting. Table 7 shows the number of pinned and un-pinned documents and search windows in the final layout produced by each user.

In the final layouts, 32% of the documents were pinned. Based on the debriefing, the users informed us that some documents were pinned to maintain their absolute position in the workspace for the purpose placing meaning into the workspace. However, others were at times pinned for more detailed adjustments to cluster layout. For example, User2 commented that he would pin documents in a cluster so that he could create detailed spatial structures within a cluster (e.g., a timeline). ForceSPIRE currently does not well support such multi-scale spatial layouts, but feedback from users suggests doing so in the future.

The majority of search windows were pinned (98%). Users treated them as "tags" for their spatialization, as searches were performed on entities. The only two users who had one search window un-pinned (User4 and User5), explained that they preferred to have it "float" to get an idea of where the documents are that related to that term.

For example, User4's progression of the co-created spatialization can be seen in Figure 30. After 16 minutes, his layout was made up of two main clusters: one about "package" and one about the "Central Russian Airline". Then, as his investigation continued, he learned more about the dataset (e.g., a suspicious "fund", documents about "weapons", etc.). At the completion of his trial (83 minutes), he was aware of much more detail regarding the dataset, such as events happening in "Nassau" regarding the "package", a suspicious person named "Ortiz", and a unrelated plot in "Russia". Additionally, the layout placed a collection of documents along the far left, which the user told us were "junk". These results indicate that the spatialization was successfully co-created (based on the coinciding weights, shown in Figure 28), and maintained meaning for the user. As such, not only did the weights reflect the analytic reasoning of the user, but the spatialization seemed to successfully reflect the shared knowledge between the user and the system at each stage.

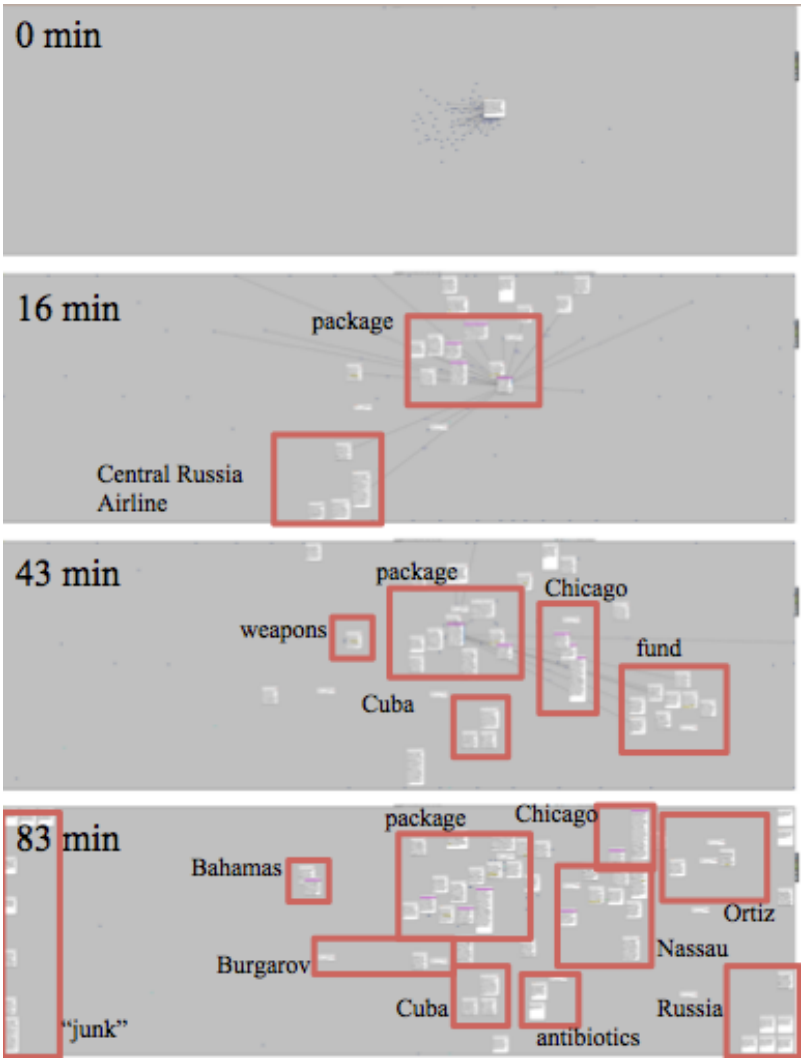
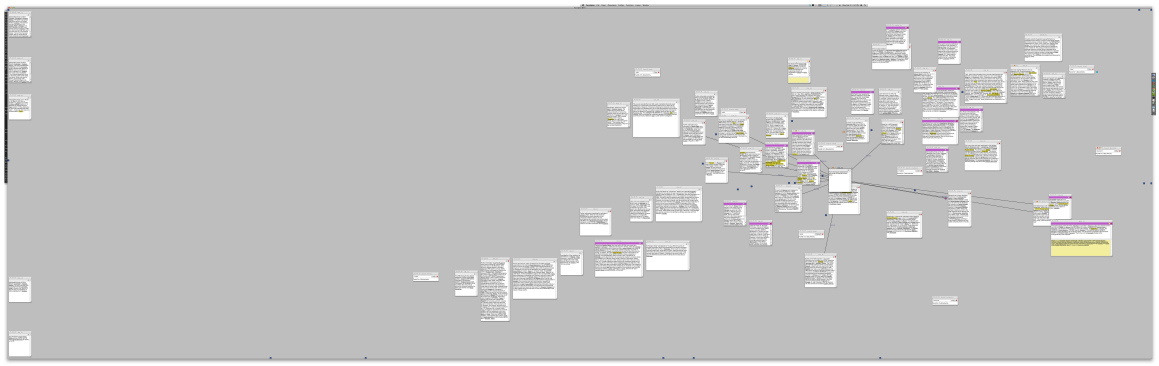
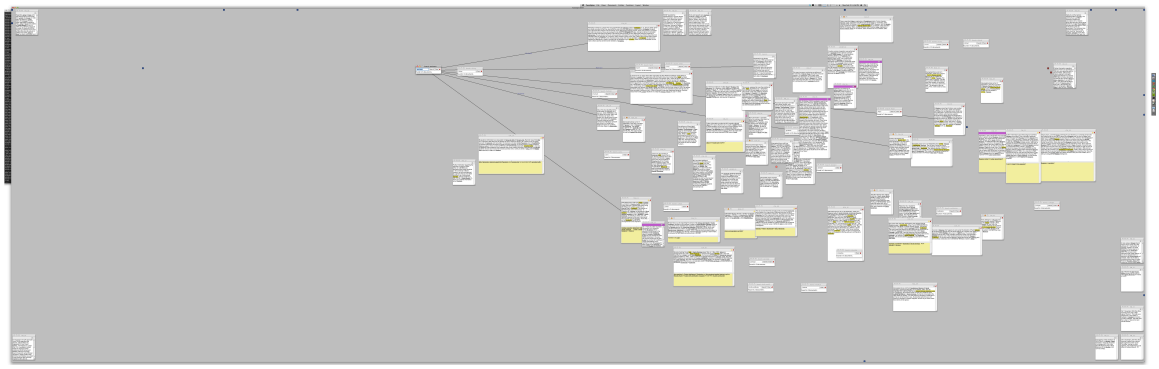


Figure 30 The progression of the spatialization over time for User04. The annotations (labels and red region boundaries) were added after the study. They represent the meaning of the regions of the space, as indicated by the user during his process.

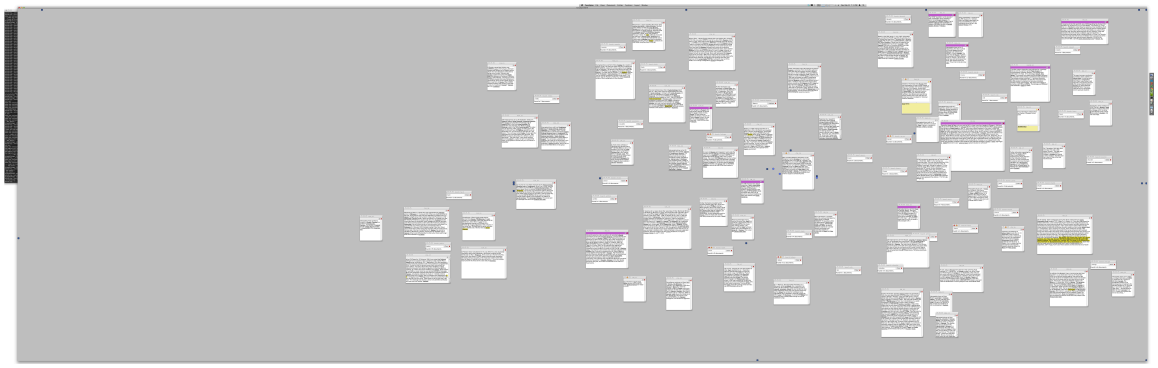
The final states of the all users are shown below:



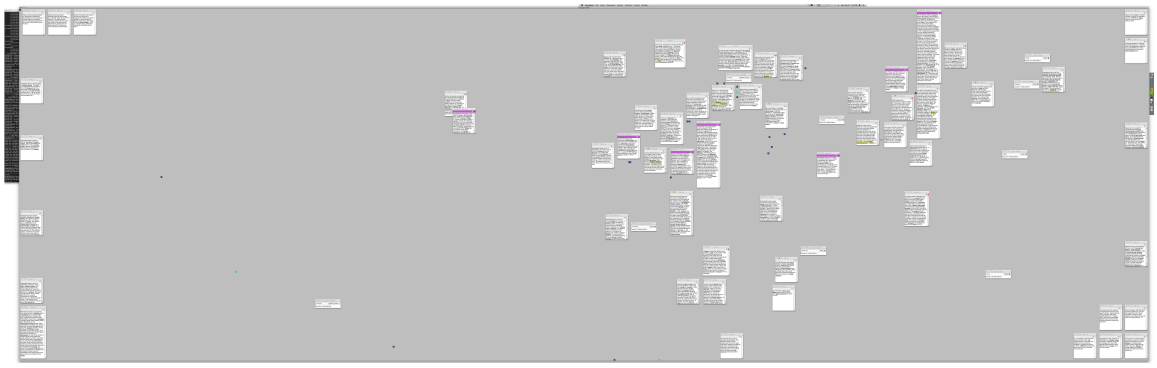
User1: Final Spatialization



User2: Final Spatialization

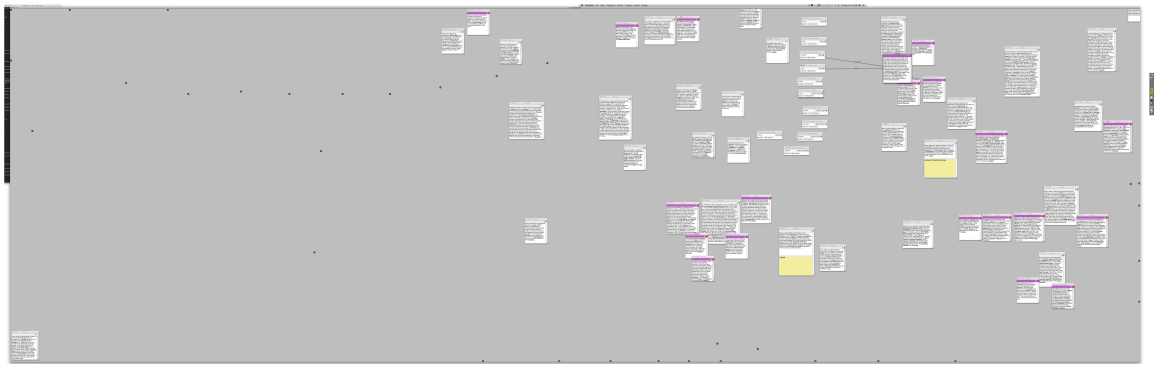


User3: Final Spatialization



User4: Final Spatialization

User5: Final Spatialization



User6: Final Spatialization

5.3 Discussion

Through this user study, we explored and validated the principles grounding semantic interaction. These principles include (fully described in [25]):

- capturing semantics from user interaction;
- shielding the users from direct parameter manipulation;
- incrementally co-creating of the spatialization through incremental model learning, to coincide with the incremental formalism [63] of the user.

5.3.1 Capturing Semantics from User Interaction

ForceSPIRE captured the semantics from users in terms of the entities contained in the dataset. The method for capturing these semantics is supported via changing the weighting of entities, as well as the creation of entities. In turn, the semantics are reflected in the steering of the weighting scheme of the model.

The ability to steer this model at multiple levels of detail was important because it coincided with the user's reasoning about different levels of detail. For example, document movement allowed users to provide more broad and informal feedback regarding important relationships between documents. Generally, this feedback was at the document level, where users would create clusters of documents without a formalized schema as to precisely why they are all similar. For example, User 2 told us he enjoyed the flexibility of moving documents because "[the system] will eventually figure out what [he] mean[s]". Similarly, User4 used expressive movements more heavily than exploratory movements. When asked about his preference, he informed us that "[expressive movements] tell the system something", and that the more he told the system, the "better chance the system will figure it out".

Other semantic interactions, such as search and highlighting, were used to provide more detailed insight into what characteristics of a dataset were relevant to the user. Similar to the results found by Endert et al. [27], highlights were typically short phrases containing

relevant entities, while searches were performed only on keywords. To the users, the highlighting was preformed to provide visual importance to portions of documents. However, the system interpreted these as indicators as to what information should be examined more closely, resulting in entity creation and up-weighting. As a result, ForceSPIRE did not only change the distribution of the weights, given the initially extracted entities, to attempt to produce the best fit given the user feedback, but also added entities when given feedback regarding relevant text within the dataset. As shown in Table 6, the entities added through semantic interaction were not only highly weighted in the soft data, but many also correlate to entities important to the users.

5.3.2 Shielding from Direct Model Steering

The users in this study treated their investigation not as steering a model, but rather synthesizing information. This is an important distinction, as it shows semantic interaction as providing a fundamentally different and richer method of interacting with visual analytic systems. As a result, semantic interaction enabled an analytic process similar to the effective spatial processes described in [4, 27, 63], where the informality of the spatial synthesis interactions were beneficial in supporting incremental formalism. However, unlike these examples where a majority of the spatialization is created manually, semantic interaction provides computational support. Therefore, while users realized an implicit ability for ForceSPIRE to learn about the characteristics important to them, the explicit use of the system was to synthesize information.

5.3.3 Incremental Model Learning and Formalism

Users develop insights into a dataset through interacting and exploring it. As such, users learn additional information as they proceed through their analysis. For example, when constructing spatial groups manually, the meaning of these clusters gradually changes as more information is learned [27]. A cluster that was created based on a single term of interest, may evolve to represent a more broad meaning, represented via a collection of terms.

ForceSPIRE was able to incrementally learn these insights, and translate them into representative entity weightings and ultimately helpful spatializations. For example, User4's cluster regarding "Nassau" transformed into a cluster also containing documents that did not include the term directly, but instead contained terms such as "Bahamas" and "Freeport Star Hotel". These were conceptually related events, and thus the user decided to group them. The structural relationship between these documents can be described as a "transitive relationship" [27] (i.e., there is a connecting document that contains both terms). One challenge in this form of relationship is determining which terms to use for the transitivity. ForceSPIRE incrementally learned these terms, through semantic interactions such as moving a document near a cluster (confirming membership) and moving a document towards a different cluster (disagreeing with transitive relationship).

5.4 Future Work

The concept of "undoing" a semantic interaction is not trivial, as not only was the spatial location of a document adjusted, but the coupled model updates changed the entity weighting scheme used to probabilistically determine the position of the other documents. The version of ForceSPIRE used in this study did not include an undo functionality. However, none of the users requested it. User3 made an erroneous move, where he accidentally dragged a document near the wrong cluster. "Ooops, oh well, [ForceSPIRE] will fix that later", he said (instead of asking how to undo), implying that the small number of erroneous interactions will be outweighed by the sum of the meaningful interactions over time. One approach in another version of ForceSPIRE is to store previous weighting schemes, so that when an undo occurs, the previous weights are restored, and the model updates the spatialization accordingly [25]. However, a "true undo" of a semantic interaction would be one from which the model learns about the user's analytical reasoning. As such, instead of restoring the weights of upweighted entities from the interaction, the system could lower the weights below the previous

weights to reflect the user's decision to discontinue the investigation of those topics. This is an open research area we plan to investigate.

The soft data captured during each user's process is used directly to steer the force-directed model calculating the spatialization. Based on feedback received from this study's users, we believe soft data has potential for additional benefits. Various types of biases are common pitfalls for analyses [35]. For example, confirmation bias can result in users accepting evidence to confirm a given hypothesis, and ignoring evidence that may refute it. Showing the history of weighting to users during their investigation might help illuminate some of these biases, making users more aware of the potential to explore other hypotheses. We are planning to incorporate a soft data graph view into the workspace in the future.

Soft data has the potential to aid a group of analysts with asynchronous collaboration. Each collaborator's soft data provides an approximation of both the process and the findings, which can serve as a baseline for the collaboration. For example, continuing an investigation started by another can be made easier by illuminating the process through observing the weighting changes. This gives the collaborator a better understanding of the hypotheses already explored. Additionally, when multiple analysts are asked to investigate the same dataset, the history of entity weighting for each user enables an overview of the group's collective investigation. One open challenge is how to effectively merge soft data from multiple users.

5.5 Conclusion

In this paper, we present results of a user study investigating the principles of semantic interaction for supporting sensemaking. Semantic interaction is an approach to user interaction with visualization that couples analytic interactions within a spatialization (e.g., document repositioning, text highlighting, search, annotations) with updates to the

underling model responsible for generating the spatial layout. As such, semantic interaction tactfully combines interactions enabling users to synthesize information with model updates to support computational foraging support for sensemaking.

The results indicate that the semantic interactions in ForceSPIRE provided the flexibility for users to investigate and explore data spatially. Semantic interaction provided the expressiveness required to update the model to coincide with the user's analytic reasoning. The captured soft data during the investigation supports the analytic product of each user, and was also able to adapt to the different points of emphasis and hypotheses during each investigation. Users regarded semantic interactions not as direct model steering, but as interactions for synthesis. Finally, ForceSPIRE updated the spatialization based on the semantic interactions of the user, and the final layouts were representative of each user's findings. Thus, the spatialization was successfully and incrementally co-created between the user and the system.

Given the positive results of this study we encourage further research in this area to advance the field of visual analytics through exploring the science of interaction.

Chapter 6

Semantic Interaction Design Space

There are many design decisions that factor into the successful implementation of semantic interaction. The following sections discuss some tradeoffs that exist when designing visual analytic systems with semantic interaction.

6.1 The Interaction-Feedback Loop

In semantic interaction, much emphasis is placed on providing a mechanism for the system to learn the domain expertise of the user through inferring a model weighting scheme. However, equally important is for the system to provide feedback about what analytical reasoning it has inferred (i.e., learned) from the user. The challenge comes in the way of determining the balance between providing this information *explicitly* (e.g., showing dimension weights, asking the user to confirm learned dimensions, etc.) versus *implicitly* (e.g., updating the spatial layout, etc.).

Providing explicit feedback for model steering has been previously studied. For example, Liu et al. describe how interactively adjusting the location of points within a spatialization enables the system to learn about dimensions of a dataset that correspond to the user's feedback [45]. Through performing this type of observation-level interaction, the users are given a set of weights that correspond to their newly generated

spatialization. As such, this work focuses on explicitly showing the user the dimensions that correspond to their interaction (i.e., the feedback from the system to the users).

Implicit feedback entails providing the feedback of the model through the visualization, rather than explicitly via the weighted dimensions. For example, ForceSPIRE provides an updated spatialization as a result of a semantic interaction. (an entity viewer window also exists, where dimension weighting can be directly adjusted). Similarly, previous work on observation-level interaction also uses the updated spatialization as a medium for communicating the learned domain knowledge [28].

However, *how can a system support a mixture between these two forms of feedback?* One can see that as the number of dimensions increase (and become more abstract), explicit feedback may not be effective or meaningful to the users. Further, the results of a user study of ForceSPIRE (where explicit feedback can be obtained by the entity viewer window) shows that users may not prefer, or need, this form of feedback [24]. Similarly, users may require some feedback to gauge what information the system is learning based on their interaction, and given the ability to provide more fine-grained model steering (e.g., steering at the entity-level, rather than at the document level).

One possibility is to maintain this feedback within the spatialization. That is, instead of providing a separate view for the explicit feedback, augmenting the spatialization to include this sort of information may be beneficial. For example, ForceSPIRE includes entity underlining within the text of a document to inform users of which keywords are entities in the model. However, this depth of information in could be increased, to highlighting words on a color ramp based on their weight. Then, if users find inconsistencies in the entity weighting scheme, adjustments can be made, and the bi-directional learning can continue.

6.2 Learning the Weighting Scheme

When inferring analytical reasoning in the form of a weighting scheme, a choice can be made as to how to translate an interaction to a set of dimensions and weight values. This learning can happen through cleanly inverting the mathematical model, creating a new heuristic by which to learn weights, or some combination of the two.

Examples of directly inverting the mathematical model can be seen in the modified dimension reduction models presented by Endert et al [28]. In these models, great care was taken to ensure that each interaction (i.e., newly positioned data point in the spatialization) corresponds to a learning of weights while maintaining a given amount of stress in the system. The weights are generated through a backwards-solving of the weights using the dimension reduction model, and using the user's newly positioned observations.

Using the same dimension reduction model (MDS), Liu et al. [45] performed the same MDS projection step, but inverted the observation-level interaction differently. Instead of directly inverting the projection used in MDS, they use a modified (biased) calculation of the weights for their learning step. This modification can be seen as a combination of a heuristic and a cleanly-inverted model for learning.

ForceSPIRE uses a model for learning weights that is more flexible, and thus does not adhere strictly to the inverted model (a force-directed model, in this case). For example, while performing an observation-level interaction in ForceSPIRE results in an emphasis of the similar characteristics between documents moved closer, the remaining weights are equally reduced for normalization of the global weight. As such, there is no direct inversion of the force-directed model, but instead the model is used to calculate the set of characteristics that correspond to the similarity, and the amount of emphasis those characteristics get (i.e., the increase of the weight of those entities) is via a constant.

The decision of inverting a mathematical projection model may be a good fit for systems where the semantic interactions are primarily observation-level interactions. However,

other forms of semantic interactions may not lend themselves to directly inverting a projection model (e.g., highlighting text, performing a search, etc.). One possibility for these forms of interaction is to create a forward model for each of these, by which the inversion can take place. For example highlighting can be automated given the weights of entities. Then, as users manually highlight (or change the highlighting that the system recommended), the system can invert the model used for highlighting to maintain mathematically valid visualizations. The fundamental principles of semantic interaction still apply to these interactions, as they generalize beyond spatializations and observation-level interactions.

6.3 Choice of Mathematical Model

To support semantic interaction, the underlying mathematical model must be appropriately tailored to couple with the user interaction. Previous work has shown how to modify popular dimension reduction models, such as PPCA, MDS, and GTM to allow the direct manipulation of the points within the spatialization (called *Observation-Level Interaction*) [28]. In ForceSPIRE, the focus is on steering a force-directed model for the purpose of analyzing text datasets.

Based on the feedback from evaluating semantic interaction in ForceSPIRE, two aspects of the mathematical model that are important to users are the incremental nature of how the model updates both the weight vector and the layout, and the probabilistic nature of how the model.

Incrementally updating the weight vector, and this the spatial layout, is beneficial to users, as it maps closely to how users incrementally gather insight about a dataset (i.e., incremental formalism [63]). For example, a minor adjustment to the location of a document within a cluster should not result in complete layout regeneration. Instead, the slight movement of a document within a cluster may reflect the user building a timeline, and thus only local adjustments should be made. For other, broader moves, such as repositioning a document from one cluster to another, the impact on the weight vector may be more severe, resulting in a greater impact on the positioning of other documents.

For this capability, we have found that deterministic models, by definition, are less likely to support these minor updates without requiring a broader layout update. Probabilistic models (such as a force-directed model) can achieve a low-stress state given multiple layout variations.

Table 8. Semantic interactions for directly modifying the spatialization can impact both the relative and absolute spatial positions of documents.

Relative	**Absolute**
on Document: • Move toward other Document • Move away from other Document • Add to Cluster • Remove from Cluster	*on Document:* • Pin to location • Pin to region
on Cluster: • Create/Delete Cluster • Preserve Cluster • Merge/Split Cluster	*on Cluster:* • Define location, region • Define shape

The updating of the spatial layout is equally important, as it provides the opportunity to show the user what has changed from one layout to the other. That is, it provides the user feedback on what the system has learned from their previous semantic interaction. Models that are incremental in nature (where the calculation of the lowest-stress configuration is incrementally obtain) more easily support this concept, as the user can observe the model achieving the state. For example, users can gain insight into both the characteristics of the model, as well as the weighting vector, through observing a force-directed model settling out.

Equally important is **determining the type of task or goal** that a user plans to accomplish within a spatialization. Given the type of interactions shown in Table 8, the style of model that each of these drives may differ. For example, defining the shape of a cluster may be better suited to a clustering model, rather than purely a dimension reduction model. Similarly, labelling a cluster (or modifying the label o a cluster) may lend itself to steering of a topic modelling algorithm. Developing such mixed-initiative

and mixed-metaphor systems can cover more adequately cover the spectrum of analytical reasoning that may be associated with various user interactions, as well as develop a more mature design space for semantic interaction.

6.4 Relative and Absolute Spatial Adjustments

Semantic interaction enables the flexibility to explore hypotheses spatially, such as exploring the relationships within the dataset through direct spatial adjustments. The spatial metaphor provides a rich medium for the interaction, as the spatial locations of documents are meaningful. The meaning is generated via a document's ***relative*** position to others, as well as its ***absolute*** position in the spatialization. Table 8 provides examples of how these two spatial characteristics can be modified. In providing users such flexibility, many challenges become apparent, such as the ones discussed below. With each of these design decisions exist tradeoffs, such as mathematical complexity, completeness of information, user friendliness, and user control.

When users interact within the spatialization, the goal of semantic interaction is to infer the characteristics of the dataset upon which to base the change of similarity between documents. To do so, one design decision to make is *how to capture and quantify the interaction*. Depending on the domain, task, and data for which the tool is being designed for, one of the following may provide more desired results.

Boolean similarity updates – updates to the similarity between documents are treated as being either "more similar", or "less similar" than in the previous view. The "amount" of change to the similarity is not taking into direct consideration. Such an implementation may lend itself more to hierarchical or categorical models where similarity is fundamentally based on membership to a topic or cluster. Drucker et al. present an example of using semantic interaction for such sorting and categorization tasks [18]. Their tool enables users to organize documents into folders, where each folder has the ability to suggest other documents that may fit into the folder based on the information currently contained in it. Semantic interaction aspects of this work are that the documents can be added or removed from these folders directly in the spatial "desktop" metaphor.

Scalar similarity updates – updates to the similarity between documents is calculated based on how much the distance between two documents changed from the previous view. That is, decreasing the distance between two documents by 40% will result in a less aggressive similarity update opposed to a decrease of 90%. For spatializations that carry a continuum of meaning throughout the space, scalar updates may be more appropriate. However, users may find it more meaningful to conceptualize the update of similarity through via Boolean similarity. For example, ForceSPIRE uses Boolean similarity updates while maintaining a continuous spatialization [25].

Chapter 7

Conclusion

Semantic interaction provides visual analytics with a new approach for user interaction in visual analytic applications. Instead of providing direct graphical interfaces based on Graphical User Interface principles (e.g., sliders, knobs, buttons, etc.), semantic interactions are grounded in principles embedded in the analytic process. The ability to create synergy between the user and the system through interaction is a familiar goal in human-computer interaction, however the complexity of mathematical models used in visualization have forced visual analytic system designers to re-think the role of interaction.

This work is driven by three fundamental research questions:

1. (**RQ1**) What semantic interactions do users perform during exploratory analysis of textual information within a spatialization?
    a. Given a spatialization where users can freely reposition text documents, what interactions occur within their analytic process?
    b. What mappings exist between these interactions and the analytical reasoning?

2. (**RQ2**) How can visual analytic systems be designed to leverage semantic interaction?
    a. Defining semantic interaction.
    b. What changes must be made to the mathematical models responsible for the spatial layouts?
    c. How do these changes in design impact the current model for visualization (i.e., the visualization pipeline)?

3. (**RQ3**) How does semantic interaction impact the analytic process?
    a. How can incremental formalism extend to incremental model learning?
    b. How does semantic interaction create synergy between the insights of the user and the model's learned characteristics?

7.1 Research Contributions

Understanding Semantic Interactions in Spatial Sensemaking

The studies presented in this work enabled us to gain an understanding for the effectiveness of analytic functionality in spatializations for sensemaking. Throughout the spatial exploratory process, interactions occur. We found that these interactions can serve as indictors for the analytical reasoning of the user, and thus can serve as characteristics from which the model can learn. This not only emphasizes the importance of interaction in the visual data exploration, but also opens questions such as : *how important (for the purpose of gaining insights) is interaction in information visualization compared to the visual representations themselves?*

Engineering and Technical Challenges when designing for Semantic Interaction

This work has shown that designing for semantic interaction requires all aspects of a system be designed to support these fluid forms of interaction. As semantic interaction is typed deeply with the visual representation and the underlying mathematical model, systems must be designed with this level of control and flexibility. This impacts the design of the user interface, back-end data repositories, and performance to maintain a real-time feedback between the user and the system. Further, the ability to treat the interaction as data in these systems opens opportunities for mining this information further - be it for collaboration or post-analysis evaluation.

Unifying the Sensemaking Process with Semantic Interaction

Exploratory data analysis hinges on the ability of users to explore information, forming and testing hypotheses continuously while sensemaking. Understanding and insight is generated as users can explore and validate their mental model of the information with the information itself (e.g., testing a relationship between information, exploring a hypothesized story, etc.). This process entails synthesizing currently known information/insights, while simultaneously foraging for new, relevant information (i.e., sensemaking). With semantic interaction, we provide the user with computational support for foraging, while he or she can focus on synthesis. Therefore, the system is tasked with a computationally intensive task, while the user can leverage the domain expertise and synthesizing of information – a task suited for humans.

As a result of this unification, this work presents an augmented version of the visualization pipeline, where the visualization is not an "output-only" view, but rather a medium for interaction and sensemaking.

Science of Interaction

User interaction is a critical component of any visualization intended to support exploratory data analysis – it empowers users with computational capabilities they would

otherwise not posses. As such, the design of interaction can have tremendous impact on the usability of a system, to the point where improper interaction design can inhibit the generation of insight. As the field of visual analytics continues to grow and mature, further research in the science of interaction can enable users to understand greater amounts of information in a shorter amount of time. While this science is in its early phases, focusing research on this area is important to generate technology that can aid in understanding the ever-growing amount of digital information being created.

7.2 Future Opportunities

7.2.1 User Interaction for Visual Analytics

Further research can be done, exploring the connection between user interaction, mental models, and mathematical models, such as:

- What input and feedback parameters for interactions are important to users?
- What tradeoffs exist between designing algorithms to minimize low-dimensional distances between information, versus maintaining absolute locations of information in the spatialization?
- How can interactions be designed to allow both entity weighting and document weighting, and what impact should those two weighting mechanisms have on the layout?
- How can we leverage large display, multi-touch interactions to enable users to implicitly or explicitly specify a wider range of parameters to inform the system of their analytical reasoning?
- How can we design effective means of providing visual feedback to users about what the system has learned from the user's interactions?

We are interested in the input (i.e., from the user to the system) and feedback (i.e., from the system to the user) parameters for semantic interactions in a spatialization. Generally, dimension reduction algorithms can be manipulated through directly controlling a collection of complex and confusing parameters. Our aim is to understand the analytical reasoning associated with traditional analytic interactions with textual information in order to couple those interactions with the parameter updates, shielding the users from doing so directly. We plan to investigate the analytical reasoning associated with users interacting in a spatial workspace through a user study. Through observing users manually positioning documents spatially as part of a sensemaking task, we will gain an understanding for the parameters users find important and necessary to convey their domain expertise to the system. We contend that many parameters and attributes are inherent in the quantification of "why" a user finds similarity between two documents. For example, relative distance from other documents may be a measure of the user's perceived similarity between those documents, or could convey the *certainty* of the user's conclusion about their similarity. Similarly, highlighting phrases of text may imply that phrase containing relevant information. It is important that we understand the relationship between these parameters to gain a better understanding of how users treat them, and ultimately give designers a better foundation upon which to base semantic interaction.

In a spatialization, the spatial positioning of information is one of the primary visual encodings for the information. Distances between documents can imply a similarity measure, while absolute locations of information can serve as a "landmark" for themes and concepts within the spatialization. These two encodings can inform the system of different, yet equally important information. For instance, generating clusters of documents can signal a similarity between the collection of documents, from which the statistical model can mine the important information creating the similarity. Instead, the absolute location of a document in the spatialization can inform the system that the user has linked a particular theme with the specific, persistent location. Therefore, the goal is to provide algorithms that not only aim to optimize the low-dimensional distances between information, but also attempt to maintain the persistence of the absolute locations provided by the user.

The input parameters for users will primarily control the "weighting" (or importance/relevance) of either entities or documents. Users have the ability to control the weights of entities through interactions such as searching, highlighting, etc. Changing the weights on entities will allow the system to update the spatial layout, emphasizing those entities. Weighting documents has a different effect. When users emphasize an important document, the weight of that document increases. As a result, that document becomes more persistent, and the location of the document is more resistant to change from the layout algorithm.

Similarity as a measure includes many parameters to help inform the system of the details of the similarity between information. For example, users can specify which information is being compared, their level of confidence, the specific entities causing the similarity, and more. However, forcing the user to specify each of these can be overbearing, especially in the early stages of the analysis. What we know from *incremental formalism* [63] is that users gain insight incrementally over the course of their investigation. That is, information to populate each of those parameters may not be known initially, instead only realized later in the analysis. It is important for the system to allow this incremental expressiveness in terms of parameter specification. This allows the analytic process to proceed, and enables users to provide the system with only the information they posses at their current stage.

One promising interaction modality for enabling users to specify these parameters is through the use of large display, multi-touch surface. Instead of a traditional mouse and keyboard interface, where parameters need to be specified by modifier keys, menus, pop-up dialog boxes, etc., multi-touch allows more modifiers through allowing more simultaneous points of input. We plan to explore how gestures can be designed to accommodate the specification of these parameters [51]. For example, while a user moves a document with his right hand, he can use his left hand to specify which documents he intends to compare the document to. Further, specific gestures can be used to specify parameters such as the level of granularity of the comparison (e.g., entity level, document level, etc.).

Another important aspect of this research task will be to explore effective means of providing visual feedback to users. In a spatialization, the challenge is to determine the correct balance between visual and mathematical feedback, as well as correctly determining the appropriate level of feedback. Visual feedback can be presented by updating the locations of information, changing the color of links between information, etc. to signify the system's response to the user's domain knowledge. We plan to experiment with different detail levels of the feedback, such as presenting feedback on the weighting of documents, down to the level of showing the weights of entities. It is also important to allow users to react to the feedback in case the system misinterpreted the analytical reasoning of the users. For example, feedback on the entities causing the similarity within a cluster can be represented using a tag cloud, where entities with strong weights are represented larger. Users can correct the similarity in this visual metaphor by changing the size of the entities, or removing them completely. The form of the visual feedback should correspond to the users understanding of the space. For example, document similarities could be expressed through proximity, while entity importance could be expressed through highlighting. One important outcome of this approach is that important information should become more visually salient to the user.

7.2.2 Flexible Visualization Framework

Another area of research is in exploring the opportunities for a flexible visualization framework to support semantic interaction. The concept of such a framework is that instead of commonly used multiple synchronized views, these flexible visualizations will be grounded in a single, integrated view. The primary points of this research include:

- How can different regions of a spatialization correspond to different algorithms to support multiple metaphors created by users in a spatialization?
- What possibilities for reusing soft data exist (i.e., streaming data, larger datasets, filtering relevant information from noise, etc.)?
- What information regarding the analytic process can be uncovered from analyzing the soft data after an investigation?

Users refer to information in different regions of spatializations with different contexts and metaphors [5, 59]. For example, while one region may be arranged as a cluster of documents pertaining to "explosives", another region may represent a timeline of events on how to acquire the explosives, and another region organized geographically to represent the areas possibly impacted by these events. This mixed-metaphor use of a spatialization poses challenges to layout and clustering models that are generally designed to compute a single model layout across the entire visualization. Additionally, the challenge of choosing which model best captures the user's domain knowledge based on the layout can be difficult.

We propose a framework to support the flexibility required for such a mixed-metaphor spatialization. First, we will make use of multiple simultaneous models being weighted based on the learned user feedback. Over the course of an investigation, the system will weigh each model based on the consistency between the model's computation and the user feedback. In doing so, the system does not have to rely on a single model to capture the semantic information from the user. Secondly, we propose the use of spatially aware models to address the challenge of using mixed (and multiple) metaphors throughout the spatialization. Our framework will employ mixture models to appropriately emphasize specific models for corresponding regions of the layout where those models create a better match [49]. Another approach is to first divide the spatializations into discrete regions (similar to the approach in [18]), then apply the appropriate model to each collection of documents within the region. This approach is a good match for the processes of users we have observed, which indicate that users first bin information based on high-level similarity, then later organize information within each bin [5].

An important and powerful characteristic of semantic interaction is the capture and use of soft data. In the purest sense, soft data is the interpreted result of a user interaction, stored in such a way so that the model can incorporate this information for future computation. Therefore, soft data can immediately benefit a system through allowing the model to compute on not only the raw data (i.e., the "hard data"), but also include the quantified form of the user's domain expertise. However, there are benefits of soft data beyond the immediate benefits to the model. Users can analyze soft data during or after their

investigation for provenance and history purposes. The collection of soft data can serve as a recollection aid to users needing to reconstruct their analytic process. Further, resulting soft data from one investigation can be reapplied to another dataset, as well as streaming data. Users can observe how their domain expertise from one dataset can extend to another.

Soft data can also be used to filter very large datasets, separating the relevant information from the "noise". We plan two approaches, a distance threshold and a weighting threshold. First, the computed layout can discard any documents that are beyond a defined distance from relevant and important documents (implying they do not share much of the relevant information with the important document). Also, a weighting threshold can be applied to streaming data to discard any documents that do not meet a defined lower threshold of weighting. This is calculated by determining the weight of the document by summing over the weights of all the entities included in the document. This research will explore these two approaches, as well as other possibilities for filtering using soft data.

Once soft data is stored, the question of how to apply it to models becomes complex. How should models handle the temporal aspect of soft data? How will soft data support multiple, competing hypotheses? How does a system handle the "undoing" of an interaction? We approach these challenges through maintaining soft data uniformly throughout the analysis. That is, soft data that is created earlier in the analysis receives equal weight to all other interactions. However, this is an open research area, as different stages of the investigation may afford different weighting. To support exploration of multiple hypotheses, a weight-fading scheme could be used to gradually diminish old insights and emphasize new ones. This is any additional parameter that we will need to fit naturally into the users sensemaking process. To support multiple hypotheses, we believe providing users feedback of their process by showing them trends in the soft data can help alert users of narrowing their focus too much, or broadening their investigation too much. This could be used to help overcome biases in sensemaking, such as "confirmation bias" in which analysts tend to overcommit to their first hypothesis [35]. The soft data can illustrate these scenarios by showing the weighting of entities. Narrow investigations

will correspond to a converging set of entity weights, while broad investigations will show a more diverging trend.

7.2.3 Evaluating Semantic Interaction

Additional evaluation can be performed to gain a better understanding of the impact of semantic interaction on the analytic process, and how well the soft data captured maps to the mental model of users.

Effects of Semantic Interaction on Analytic Process One approach is to observe users analyzing a textual dataset using a prototype, ForceSPIRE, which utilizes semantic interaction (and thus captures soft data throughout the analysis). To provide a realistic intelligence analysis task with a known ground truth to evaluate the performance of the users against, we plan to use one of the VAST Challenge Datasets [58]. One can collect both qualitative results in terms of the behavior and processes of the users, as well as quantitative results such as how close their findings match the known ground truth, and each of the user's soft data. As soft data gives us a quantified form of the user's domain knowledge and intuition regarding the dataset, we can analyze that information to establish a level of bias. If the goal of a successful visual analytic system is to balance human intuition with statistical models, the balance between hard data and soft data can provide one such measure. The questions to answer are what is the balance, and how much intuition is needed (and how much is too much, or too little)? The soft data will give an idea of how much the user has guided the model through weighting of entities and documents, and how much the user has constrained the model through pinning documents in specific locations.

Soft Data Analysis Another goal is to perform an in-depth analysis of the soft data captured during a user's analysis with the aim of detecting trends and patterns indicative of specific analytic behaviors. As exploratory data analysis consists of the situated generation and testing of a user's domain insight on the dataset, gaining insight into the amount of bias being introduced into the system is beneficial. For example, converging trends in the weighting of entities can indicate confirmation bias, where diverging

weights can represent an analysis involving multiple hypotheses. In particular, we may be able to quantify specific biases such as confirmation bias [35], and thus design real-time alerts to users that these specific biases are occurring during their analyses.

Capturing the User's Mental Model Also, it is worth investigating the effectiveness of semantic interaction to combine the intuition of users with the computational findings of algorithms. In essence, does semantic interaction enable users the functionality required to transform the layout into one that more closely matches their mental model? A goal of semantic interaction is to allow the model to better learn and match the mental model of the user. As the domain knowledge is being captured and integrated into the model through soft data, the aim is to have a closer connection between the layout and the mental model the user has of the information. We propose to use subjective measures of "closeness" to analyze if the incremental change of the layout creates less entropy between the two models (mental model and statistical model). Another way to assess the success of this goal is to ask one user to analyze the final layout and soft data of another user, and ask them to extract the findings and hypotheses. If the layout and model successfully captures the domain knowledge and insights of the user, our hypothesis is that another user could analyze this information and arrive at similar conclusions – as well as gain an understanding of the user's process and decisions.

7.2.4 Other Visual Representations and Interactions

Semantic interaction in this work was explored using primarily textual datasets visualized using spatializations. However, semantic interaction can generalize to other visual representations, and steer the underlying mathematics of a variety of visualizations. The fundamental principles can be applied to these visualizations by designing interactions that are not designed based on directly manipulating parameters, but instead adhere to the visual sensemaking process and reasoning of users given the visual representation. Further, these different visual representations may have unique styles of interactions that they afford, and thus other types of semantic interactions may exist for them which are not afforded in a spatial workspace.

7.2.5 Tackling Large Data

The amount of information is rapidly increasing in quantity and complexity. However, the computational foraging abilities of computers are extraordinary, and continue to advance. Visual analytics has embraced the necessary challenge of merging fields such as visualization, cognition, data mining, and others for the purpose of supporting sensemaking. Semantic interaction can contribute to this challenge by providing one approach for analyzing large datasets. Currently, semantic interaction focuses on the weighting of all characteristics of a dataset. However, the principles can be extended to steer not only the dimensions reduction model, but also a querying system that is tied to a database backend. As such, semantic interaction can control what information (filtered from a much larger, dynamic dataset) is shown to the user base on the inferred analytical reasoning.

7.3 And with that…

The work presented opens a new design space for user interaction, albeit not fully exploring all possibilities. Including the future opportunities above, there are many exciting possibilities when considering this fundamentally deeper, more integrated approach for user interaction – semantic interaction. This work has shown that user interaction in a spatial sensemaking environment can be captured, interpreted, and thus tightly coupled with the underlying model to provide a flexible yet expressive method for model steering. Through doing so, synergy is created between the system and user in terms of relevant or discriminating characteristics in the dataset. Semantic interaction amplifies the synthesis interactions embedded in the personal analytic process of users with computational foraging.

The field of visual analytics is at an exciting, yet critical point. From information visualization, we have learned many wonderful methods for visually representing

information. From these, users can leverage their visual system to recognize relationships and gain insight into datasets that would otherwise not be possible. Likewise, fields such as mathematics and data mining enable computers to extract meaningful relationships and structure from tremendously large and complex datasets. As such, it is the exciting time for visual analytics to grow from learning from human-computer interaction and the cognitive sciences in combining these abilities through exploring the science of interaction.

Bibliography

[1] *Alias-i. 2008. LingPipe 4.0.1*. 2008.

[2] Alonso, O., Gertz, M. and Baeza-Yates, R. Clustering and exploring search results using timeline constructions. *Proceedings of the 18th ACM conference on Information and knowledge management* (Hong Kong, China, 2009). ACM, 97-106.

[3] Alsakran, J., Chen, Y., Zhao, Y., Yang, J. and Luo, D. STREAMIT: Dynamic visualization and interactive exploration of text streams. *IEEE Pacific Visualization Symposium* (2011).

[4] Andrews, C. *Space to Think: Sensemaking and Large, High-Resolution Displays*. Virginia Tech, Blacksburg, 2011.

[5] Andrews, C., Endert, A. and North, C. Space to Think: Large, High-Resolution Displays for Sensemaking. *CHI* (2010), 55-64.

[6] Andrews, C., Endert, A. and North, C. VAST 2010 Challenge: Analyst's Workspace. *IEEE VAST Enxtended Abstracts (Contest Submission)* (2010).

[7] Andrews, C., Endert, A., Yost, B. and North, C. Information visualization on large, high-resolution displays: Issues, challenges, and opportunities. *Information Visualization*, 10, 4 (2011), 341-355.

[8] Ball, R., North, C. and A. Bowman, D. Move to improve: promoting physical navigation to increase user performance with large displays. *CHI 2007* (San Jose, California, USA, 2007). ACM.

[9] Blake, C. and Merz, C. J. *{UCI} Repository of machine learning databases*. 1998.

[10] Buja, A., Swayne, D. F., Littman, M., Dean, N., Hofmann, H. and Chen, L. Interactive Data Visualization with Multidimensional Scaling. *Journal of Computational and Graphical Statistics*, 17, 2 (2008), 444-472.

[11] Callahan, S. P., Freire, J., Santos, E., Scheidegger, C. E., Cl\, \#225, Silva, u. T. and Vo, H. T. VisTrails: visualization meets data management. *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (Chicago, IL, USA, 2006). ACM.

[12] Card, S. K., Mackinlay, J. D. and Shneiderman, B. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers Inc., 1999.

[13] Chang, R., Ghoniem, M., Kosara, R., Ribarsky, W., Yang, J., Suma, E., Ziemkiewicz, C., Kern, D. and Sudjianto, A. WireVis: Visualization of Categorical, Time-Varying Data From Financial Transactions. *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology* (2007). IEEE Computer Society, 155-162.

[14] Christopher, M. B. *GTM: The generative topographic mapping*. 1998.

[15] Cowley, P., Haack, J., Littlefield, R. and Hampson, E. Glass box: capturing, archiving, and retrieving workstation activities. *Proceedings of the 3rd ACM workshop on Continuous archival and retrival of personal experences* (Santa Barbara, California, USA, 2006). ACM, 13-18.

[16] Dempster, A. P., Laird, N. M. and Rubin, D. B. *Maximum likelihood from incomplete data via the EM algorithm*. 1977.

[17] Dou, W., Jeong, D. H., Stukes, F., Ribarsky, W., Lipford, H. R. and Chang, R. Recovering Reasoning Processes from User Interactions. *IEEE Computer Graphics and Applications*, 29(2009), 52-61.

[18] Drucker, S. M., Fisher, D. and Basu, S. Helping users sort faster with adaptive machine learning recommendations. *Proceedings of the 13th IFIP TC 13 international conference on Human-computer interaction - Volume Part III* (Lisbon, Portugal, 2011). Springer-Verlag, 187-203.

[19] Elmqvist, N., Moere, A. V., Jetter, H.-C., Cernea, D., Reiterer, H. and Jankun-Kelly, T. Fluid interaction for information visualization. *Information Visualization*, 10, 4 (2011), 327-340.

[20] Endert, A., Andrews, C., Bradel, L., Zeitz, J. and North, C. *Designing Large High-Resolution Display Workspaces*. 2012.

[21] Endert, A., Andrews, C., Fink, G. A. and North, C. Professional Analysts using a Large, High-Resolution Display. *IEEE VAST Extended Abstract* (2009).

[22] Endert, A., Andrews, C. and North, C. Visual Encodings that Support Physical Navigation on Large Displays. *Graphics Interface* (Virginia Tech, 2011).

[23] Endert, A., Fiaux, P., Chung, H., Stewart, M., Andrews, C. and North, C. ChairMouse: leveraging natural chair rotation for cursor navigation on large, high-resolution displays. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (Vancouver, BC, Canada, 2011). ACM, 571-580.

[24] Endert, A., Fiaux, P. and North, C. Semantic Interaction for Sensemaking: Inferring Analytical Reasoning for Model Steering. *IEEE Conference on Visual Analytics Science and Technology* (2012).

[25] Endert, A., Fiaux, P. and North, C. Semantic Interaction for Visual Text Analytics. *CHI* (2012).

[26] Endert, A., Fiaux, P. and North, C. Unifying the Sensemaking Loop with Semantic Interaction. *IEEE Workshop on Interactive Visual Text Analytics for Decision Making at VisWeek 2011* (Providence, RI, 2011).

[27] Endert, A., Fox, S., Maiti, D., Leman, S. C. and North, C. The Semantics of Clustering: Analysis of User-Generated Spatializations of Text Documents. *AVI* (2012).

[28] Endert, A., Han, C., Maiti, D., House, L., Leman, S. C. and North, C. Observation-level Interaction with Statistical Models for Visual Analytics. *IEEE VAST* (2011), 121-130.

[29] Fruchterman, T. M. J. and Reingold, E. M. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21, 11 (1991), 1129-1164.

[30] Gotz, D. *Interactive Visual Synthesis of Analytic Knowledge*. 2006.

[31] Guber, D. Getting What You Pay For: The Debate Over Equity in Public School Expenditures. *Journal of Statistics Education*, 7, 2 (1999).

[32] Hastie, T., Tibshirani, R. and Friedman, J. H. *The Elements of Statistical Learning*. Springer, 2003.

[33] Heer, J. *prefuse manual*. 2006.

[34] Heer, J., Mackinlay, J., Stolte, C. and Agrawala, M. Graphical Histories for Visualization: Supporting Analysis, Communication, and Evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 14, 6 (2008), 1189-1196.

[35] Heuer, R. *Psychology of Intelligence Analysis*, 1999.

[36] Hossain, M. S., Andrews, C., Ramakrishnan, N. and North, C. Helping Intelligence Analysis Make Connections. *Workshop on Scalable Integration of Analytics and Visualization* (San Fransisco, 2011).

[37] Hossain, M. S., Gresock, J., Edmonds, Y., Helm, R., Potts, M. and Ramakrishnan, N. Connecting the Dots between PubMed Abstracts. *PLOS One*, 7, 1 (2012), e29509.

[38] House, L., Leman, S. C. and Han, C. Bayesian Visual Analytics (BaVA). *In revision, Technical Report: FODAVA-10-02,* http://fodava.gatech.edu/node/34(2010).

[39] Jeong, D. H., Ziemkiewicz, C., Fisher, B., Ribarsky, W. and Chang, R. iPCA: An Interactive System for PCA-based Visual Analytics. *Computer Graphics Forum*, 28(2009), 767-774.

[40] Jolliffe, I. *Principal Component Analysis*. John Wiley and Sons, Ltd, 2002.

[41] Kaban, A. A Scalable Generative Topographic Mapping for Sparse Data Sequences. *International Conference on Information Technology: Coding and Computing (ITCC'05),* (2005).

[42] Karen A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, 28(1972), 11-21.

[43] Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V. and Saarela, A. Self Organization of a Massive Document Collection. *Transactions on Neural Networks*, 11, 3 (2000).

[44] Leman, S. C., House, L., Maiti, D., Endert, A. and North, C. *A Bi-directional Visualization Pipeline that Enables Visual to Parametric Interation (V2PI)*. NSF FODAVA Technical Report (FODAVA-10-41), 2011.

[45] Liu, J., Brown, E. T. and Chang, R. Find distance function, hide model inference. *Poster at IEEE Conference on Visual Analytics Science and Technology* (2011).

[46] MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297 (1967), 14.

[47] Marshall, C. C., Frank M. Shipman, I. and Coombs, J. H. VIKI: spatial hypertext supporting emergent structure. *Proceedings of the 1994 ACM European conference on Hypermedia technology* (Edinburgh, Scotland, 1994). ACM, 13-23.

[48] Marshall, C. C. and Rogers, R. A. Two years before the mist: experiences with Aquanet. *Proceedings of the ACM conference on Hypertext* (Milan, Italy, 1992). ACM, 53-62.

[49] McLachlan, G. J. and Basford, K. E. *Mixture models. Inference and applications to clustering*. Dekker, 1988.

[50] North, C., Chang, R., Endert, A., Dou, W., May, R., Pike, B. and Fink, G. Analytic provenance: process+interaction+insight. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems* (Vancouver, BC, Canada, 2011). ACM, 33-36.

[51] North, C., Dwyer, T., Lee, B., Fisher, D., Isenberg, P., Robertson, G. and Inkpen, K. Understanding Multi-touch Manipulation for Surface Computing. *Proceedings of the 12th IFIP TC 13 International Conference on Human-Computer Interaction: Part II* (Uppsala, Sweden, 2009). Springer-Verlag, 236-249.

[52] Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B. and Williams, J. G. Visualization of a document collection: the vibe system. *Inf. Process. Manage.*, 29, 1 (1993), 69-81.

[53] Pak Chung, W., Hetzler, B., Posse, C., Whiting, M., Havre, S., Cramer, N., Anuj, S., Singhal, M., Turner, A. and Thomas, J. *IN-SPIRE InfoVis 2004 Contest Entry*. 2004.

[54] Pearson, K. *On Lines and Planes of Closest Fit to Systems of Points in Space*. 1901.

[55] Peck, S. M., North, C. and Bowman, D. A multiscale interaction technique for large, high-resolution displays. *Proceedings of the 2009 IEEE Symposium on 3D User Interfaces* (2009). IEEE Computer Society, 31-38.

[56] Pike, W. A., Stasko, J., Chang, R. and O'Connell, T. A. The science of interaction. *Information Visualization*, 8, 4, 263-274.

[57] Pirolli, P. and Card, S. Sensemaking Processes of Intelligence Analysts and Possible Leverage Points as Identified Though Cognitive Task Analysis *Proceedings of the 2005 International Conference on Intelligence Analysis, McLean, Virginia*(2005), 6.

[58] Plaisant, C., Grinstein, G., Scholtz, J., Whiting, M., O'Connell, T., Laskowski, S., Chien, L., Tat, A., Wright, W., Gorg, C., Zhicheng, L., Parekh, N., Singhal, K. and Stasko, J. Evaluating Visual Analytics at the 2007 VAST Symposium Contest. *Computer Graphics and Applications, IEEE*, 28, 2 (2008), 12-21.

[59] Robinson, A. C. *Design for Synthesis in Geovisualization*. PhD thesis, Pennsylvannia State University, University Park, PA, 2008.

[60] Rose, S., Engel, D., Cramer, N. and Cowley, W. Automatic Keyword Extraction from Individual Documents. *Text Mining* (2010). John Wiley & Sons, Ltd, 1-20.

[61] Rüger, S. *Putting the User in the Loop: Visual Resource Discovery*. Springer Berlin / Heidelberg, 2006.

[62] Schiffman, S., Reynolds, L. and Young, F. *Introduction to Multidimensional Scaling: Theory, Methods, and Applications*. Academic Press, 1981.

[63] Shipman, F. and Marshall, C. Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems. *Comput. Supported Coop. Work*, 8, 4 (1999), 333-352.

[64] Shrinivasan, Y. B. and Wijk, J. J. v. Supporting the analytical reasoning process in information visualization. *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* (Florence, Italy, 2008). ACM.

[65] Shupp, L., Andrews, C., Dickey-Kurdziolek, M., Yost, B. and North, C. Shaping the Display of the Future: The Effects of Display Size and Curvature on User Performance and Insights. *Human–Computer Interaction*, 24, 1 (2009), 230 - 272.

[66] Skupin, A. A Cartographic Approach to Visualizing Conference Abstracts. *IEEE Computer Graphics and Applications*, 22(2002), 50-58.

[67] Spiegelhalter, D. and Lauritzen, S. Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(1990), 275-605.

[68] Svensen, J. F. M. *GTM: the generative topographical mapping*. Aston University, Birmingham, 1998.

[69] Thomas, J. J. and Cook, K. A. *Illuminating the path*. IEEE Computer Society, 2005.

[70] Tipping, M. E. and Bishop, C. M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society, SeriesB: Statistical Methodology*, 61(1999), 611-622.

[71] Torokhti, A. and Friedland, S. *Towards theory of generic Principal Component Analysis*.

[72] Torres, R. S., Silva, C. G., Medeiros, C. B. and Rocha, H. V. Visual structures for image browsing. *Proceedings of the twelfth international conference on Information and knowledge management* (New Orleans, LA, USA, 2003). ACM, 49-55.

[73] West, M. and Harrison, J. *Bayesian Forecasting and Dynamic Models (Springer Series in Statistics)*. Springer, 1997.

[74] Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A. and Crow, V. Visualizing the non-visual: spatial analysis and interaction with information for text documents. *Readings in information visualization: using vision to think* (1999). Morgan Kaufmann Publishers Inc., 442-450.

[75] Wright, W., Schroh, D., Proulx, P., Skaburskis, A. and Cort, B. The Sandbox for analysis: concepts and methods. *CHI '06* (New York, NY, 2006). ACM, 801--810.

[76] Xing, E. P., Ng, A. Y., Jordan, M. I. and Russell, S. Distance Metric Learning, with Application to Clustering with Side-information. *Advances in Neural Information Processing Systems 15* (2002). MIT Press.

[77] Xu, R. and Wunsch, D., II Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16, 3 (2005), 645-678.

[78] Yi, J. S., Kang, Y. a., Stasko, J. and Jacko, J. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13, 6 (2007), 1224-1231.

[79] Yi, J. S., Melton, R., Stasko, J. and Jacko, J. A. Dust & magnet: multivariate information visualization using a magnet metaphor. *Information Visualization*, 4, 4 (2005), 239-256.

[80] Yost, B., Haciahmetoglu, Y. and North, C. Beyond visual acuity: the perceptual scalability of information visualizations for large displays. *CHI 2007* (San Jose, California, USA, 2007). ACM, 101-110.