An Examination of Grouping and Spatial Organization Tasks for High-Dimensional Data Exploration

John Wenskovitch, Member, IEEE, and Chris North

Abstract—How do analysts think about grouping and spatial operations? This overarching research question incorporates a number of points for investigation, including understanding how analysts begin to explore a dataset, the types of grouping/spatial structures created and the operations performed on them, the relationship between grouping and spatial structures, the decisions analysts make when exploring individual observations, and the role of external information. This work contributes the design and results of such a study, in which a group of participants are asked to organize the data contained within an unfamiliar quantitative dataset. We identify several overarching approaches taken by participants to design their organizational space, discuss the interactions performed by the participants, and propose design recommendations to improve the usability of future high-dimensional data exploration tools that make use of grouping (clustering) and spatial (dimension reduction) operations.

Index Terms—Clustering, dimension reduction, spatialization, grouping, cognitive study.

1 INTRODUCTION

Sensemaking refers to a cognitive process for acquiring, representing, and organizing information in order to address a task, solve a problem, or make a decision [46, 72]. Models with varying levels of information granularity have been proposed for approaching and solving sensemaking problems [66, 67, 72], and these models represent strategies for addressing sensemaking problems in a variety of domains. For example, Pirolli and Card’s Sensemaking Process [67] is designed for sensemaking problems faced by intelligence analysts. Despite the specific challenge addressed by each of these models, they all highlight the need to organize the data. Continuing the intelligence analyst example, they may work to understand the actors and motivations by grouping documents by location, by person, or by subplot.

A fundamental behavior in sensemaking is the act of grouping similar observations in order to understand their properties, effectively forming a cluster. This organizational strategy is true both in paper-based sensemaking tasks [25, 87] and in tasks performed on electronic displays [2, 30]. Clusters therefore have a natural connection to sensemaking. Clusters can also help to reduce clutter in a workspace, compressing similar observations into a group that requires less physical or screen space [57, 71]. Simplifying the workspace leads to further cognitive benefits, as humans struggle to think about more than a small number of observations or dimensions at one time [75]. Thus, using groups of items to perform analysis tasks can lead to improved memory and recall by providing a simplified method of understanding the data [19]. Previous research has shown that humans use a variety of organizational principles to cluster information [22], even when addressing the same task [2]. In order to identify clusters computationally, hundreds of clustering algorithms have been implemented, each with strengths and weaknesses [33].

Another technique for sensemaking, particularly relevant to multidimensional datasets, is to embed the high-dimensional data in a low-dimensional projection in such a way that the structure of the high-dimensional space (e.g., clusters and outliers) is maintained in the projection [41, 55, 81]. Often, these projections embody a “proximity = similarity” metaphor, in which the distance between observations represents the similarity of those observations. As a result, groups of similar observations form clusters within the projection. A number of interactive tools make use of such techniques for exploratory data analysis and sensemaking [9, 23, 29, 64, 75].

The intersection of these grouping and spatialization sensemaking techniques (and their corresponding clustering and dimension reduction algorithms) has become an area of interest for visualization research, including tools and techniques [20, 52, 59, 94], studies [10, 73], and surveys [32, 83]. However, the behavior of users when performing these tasks on a dataset has not been thoroughly explored, with the most similar study to this work performing a post-hoc analysis of created groups after the sensemaking task [30]. In contrast, this work presents a study in which participants have been asked to perform spatial and grouping operations within a single sensemaking task, with the goal of understanding the organizational strategies of the participants and the interplay between the operations.

In particular, we note the following contributions:

1. The design and execution of a study to understand the cognitive group and spatialization processes involved in organizing a dataset for a sensemaking task.
2. A discussion of organizational strategies and structures, grouping and spatial operations, decision making, and external knowledge displayed by study participants.
3. Design recommendations that are intended for both current and future tools, with the goal of better supporting user organizational processes for sensemaking in interactive visualization systems.

We identified three overarching organizational strategies demonstrated by participants, saw tight interplay between spatial and grouping operations during sensemaking, and noted the creation of organizational spaces that were more complex than those currently supported by interactive sensemaking tools. The results from this study can be used to inform the design of future tools for interactive visual data exploration.

2 BACKGROUND AND RELATED WORK

This work was primarily inspired by “The Semantics of Clustering” by Endert et al. [30]. In that study, participants were directed to perform a spatial sensemaking task in a large display space, organizing a collection of documents with only manual layout capabilities. After completing the task and participating in a semi-structured interview, the participants were asked to draw the final cluster structure of their organizational space. While this study provides a useful starting point for understanding the spatial and grouping behaviors of analysts, the post-hoc analysis of clusters misses information about the development of clusters and the interplay between grouping and spatial actions. Indeed, research has demonstrated that categories that emerge during the analysis process are often shifting and ad-hoc, evolving throughout the course of the analysis to represent the information uncovered [5, 44].
Our goal with the study discussed in this work is to investigate grouping and spatial behaviors throughout the analysis process.

The card-based approach used in this study was inspired by collaboration studies performed by Robinson [69] and Isenberg et al [42]. As noted by Robinson, the intention of this strategy was to permit understanding without constraints that may be imposed by software tools. For example, there are a variety of methods by which participants indicated groups during our study, including by stacking observations, by overlapping observations, by positioning observations so that cards were touching, and by positioning observations relatively close to each other. Similar studies also include affinity diagramming techniques for organization [62]. Other possible metaphors for indicating groups could be colored labels or drawing boundaries around collections of observations. Further, Sellen and Harper suggest that using physical artifacts in such studies can be useful for revealing how participants make use of their affordances to complete tasks, thereby providing interface and interaction design guidelines [76].

2.1 Sensemaking and Cognition

Sensemaking is an iterative process, building up an internal representation of information in order to achieve the user’s goal [72]. Distributed and embodied cognition share complementary roles in human sensemaking. Distributed cognition refers to the idea that external spaces can be used to extend and support cognitive reasoning. Analysts can thus use objects or symbols as a means of externally encoding relationships [70]. “Space to Think” demonstrated that space plays a meaningful role in sensemaking, providing a large high-resolution display grid to permit analysts to organize hypotheses and evidence spatially [2]. Spatial memory has also been leveraged by Robertson et al. for document arrangement in their Data Mountain system [68], and the role of spatial memory has been extended into 3D interfaces for information retrieval tasks by Cockburn and McKenzie [17].

Embodied cognition [88] focuses on the integration of the physical body and the environment with internal resources, reflecting how the body influences cognition. Embodied cognition allows analysts to offload cognition and create understanding within their workspace, allowing physical navigation to provide more meaning to locations [3]. “Space to Think” has been demonstrated to extend to more complex spaces which contain multiple displays and devices [15, 16, 39].

2.2 Dimension Reduction and Clustering Tools

Interactive dimension reduction (DR) and clustering are both active topics in visual analytics research. In many of these systems designed to support user sensemaking and exploratory data analysis, a learning module responds to incremental user feedback, structuring and displaying subsequent visual representations of the data in a way that reflects the goals of the analyst exploring the data [77]. One goal of this study is to better design future tools in this space by carefully examining user interactions and their motivations when applied to a dataset.

Interactions in high-dimensional data exploration systems can be organized into two broad categories: explanatory and expressive [31]. Explanatory interactions, or surface-level interactions, seek to understand the data without altering any underlying model. Such interactions are often used to support low-level tasks such as finding extrema or retrieving a value [1]. In contrast, expressive interactions communicate some intent from the analyst to the system, resulting in a model update. Parametric interactions are directly applied to model parameters, such as changing the weight on a dimension in Andromeda [75]. Observation-level interactions are applied to individual data items in a projection, which are then used to infer the analyst intent.

The semantic interaction work by Endert et al. [29] catalyzed much of the current research into expressive interactions in DR systems. Tools that resulted from this research direction can be divided into those that support quantitative and text data, though text data must be converted into a numeric representation for the algorithms to process. Quantitative tools such as Dis-Function [11] present similarity-based projections of the data, with a user interest weight vector applied to the dimensions of the dataset to influence the projection in response to user interactions. These learning techniques also support text data, as seen in systems like StarSPIRE [9, 82] and Cosmos [24]. One key change for the text case is creating a similarity computation based on a distance metric such as Cosine distance to resolve issues with term sparsity [78] or Gower distance to adjust for missing terms [37]. In contrast, quantitative tools often make use of Euclidean distance [23, 75, 85], though other distance metrics are also seen in the literature. With this study, we seek to both identify and understand which distances are important to the organizational structures created by the participants.

Interactive clustering supports similar data exploration processes, with analysts often seeking to find the clustering assignment that best suits their understanding or interpretation of the data [4, 56]. The Semantic Interaction paradigm can also be applied to interactive clustering, with a goal of understanding these analysts and adapting the clustering to suit their intent [14, 38, 79]. Using the principle of “Assignment Feedback” as coined by Dubey et al. [26], analysts are often afforded the ability to directly move observations between clusters [6, 18], thereby supplying constraints on future iterations of the cluster assignments. Beyond direct interaction with observations, systems can also support direct interaction with the clusters themselves, including operations such as merging and splitting clusters [8, 13, 40], removing clusters [6, 21, 38, 56], and expanding clusters [4, 6, 8, 21]. In this study, we identify the frequency of these cluster-oriented techniques and tie them to the exploratory intent of the user.

2.3 Cognitive Dimension Reduction and Clustering

Existing research in both the cognitive science and visualization communities has also explored the understanding of human factors in the realm of DR and clustering, with a particular focus on both perception and cognition of these techniques. With respect to the cognitive science field, researchers are actively studying the effects of both spatializations and groupings. Baylis and Driver [7] consider the role of proximity, finding that visual attention is not based solely on positional information. Similarly, Kramer and Johnson [48] consider the influence of Gestalt grouping principles of similarity, closure, and proximity in attention-based tasks, noting that the inclusion of distractors within groups yields a negative performance effect. Gillam [36] discusses the role of grouping on spatial layout in great detail, proposing new Gestalt grouping principles such as common region and connectedness. Others have examined specific areas in cognitive grouping, such as the work of Zadeh on fuzzy sets [91] and prototype theory [92], Fisher’s investigation of conceptual clustering [34, 35], and Duncan’s study of superimposition [27].

In the visualization community, Nonato and Aupetit examined the impact of distortions on analytical tasks performed by users when exploring projections of high-dimensional data [61]. Such distortions are common features of these projections, as the reduction from a high-dimensional space to a 2D (or occasionally 3D) space necessitates a loss of information. Similarly, Lewis et al. explore the reliability of human embedding evaluations, seeking to understand whether analysts agree on the quality of a projection as well as what types of embedding structures are favored by analysts [54]. In the realm of clustering, Lewis et al. also contribute a study that compares standard quality measures for clustering (e.g., Dunn’s Index [28] and Caliński-Harabasz [12], among others) to the interpretations generated by analysts who evaluate the clustering assignments [53]. Seldmaier et al. introduce a taxonomy of visual cluster separation factors, examining the density, isotropy, and clumpiness of clusters in scatterplots, finding that visual inspection with various DR and clustering techniques failed to match data classifications in more than half of the human inspection cases [74].

3 Experimental Design

The overarching question that motivates this research focuses on the organizational processes of analysts when performing exploratory data analysis. For the purpose of this study, we define an “analyst” as a person with expertise in managing complex data, more particularly high-dimensional quantitative data for the purpose of this study, as well as familiarity with basic data science techniques such as projections and clustering. We wish to understand the cognitive processes that underlie the approach that analysts take when trying to find insight within an
unfamiliar dataset. For example, when analysts are only afforded grouping and spatialization actions, how will they organize a collection of observations? Further, we wish to understand how analysts begin to explore a dataset, the types of group structures created and types of grouping operations performed, and the decisions that analysts make when exploring individual observations. This study was designed to investigate these components of the exploratory data analysis process.

3.1 Participants
We recruited 16 participants from statistics and computer science disciplines in an academic setting. These participants all self-reported some degree of experience with either data science or exploratory data analysis, with this experience ranging from taking a course that included a data science component to developing new methods and tools for data analysis. The participant cohort included three undergraduate students, eleven graduate students, and two faculty members.

Using a between-subjects design, we divided the participants into two equal groups, each of which performed their organizational tasks on a separate set of observations. The group that received the labeled dataset (participants referred to as L1–L8) received 17 index cards with an animal name and five dimensions that describe the animal (see Fig. 1 left). The second group received the same data but in abstract form (see Fig. 1 right; participants referred to as A1–A8), with all animal-related contextual information removed from the cards. In the original dataset, the attributes for the animals are normalized to a 0–100 range (the subset of animals and dimensions selected in the study dataset made the largest value on the cards 91). The study dataset is provided in the supplemental material. Participants were asked to ignore the large numbers on the cards; they were added for better video capture.

3.2 Dataset
The animals dataset provided to the participants is a reduced version of that created by Lampert et al. [51], selected because of its general knowledge applicability to all potential participants. From the initial dataset, we rounded all decimal values to the nearest integer, and then reduced the number of animals and the number of dimensions. We selected 17 animals with the foreknowledge that they could be naturally divided into three groups of five animals plus two outliers, but that alternate classifications and group assignments were possible. Five dimensions were selected to make the task challenging, but not overwhelmingly difficult, with a dimension selected to describe each of the three groups, plus two additional noise dimensions. One potential division of this dataset could be:

- **Predators** (described by Fierce): Bobcat, Grizzly Bear, Leopard, Polar Bear, Wolf
- **Aquatic** (described by Swims): Blue Whale, Killer Whale, Otter, Seal, Walrus
- **Large Herbivores** (described by Big): Cow, Deer, Giraffe, Hippopotamus, Moose
- **Outliers**: Bat, Squirrel

However, alternative natural groupings are possible. For example, the Polar Bear and Hippopotamus have Swims attributes that could place them in the Aquatic group, or the Killer Whale could be a Predator (and the Bat has a similar Fierce attribute). The animals could also be divided into two groups rather than three: Aquatic and Non-Aquatic, or Furry and Non-Furry.

3.3 Study Procedure
Noting the cognitive strain experienced by participants during a pilot study, we elected to limit the length of the study to one hour in order to minimize fatigue and frustration effects. Participants began the study by responding to four questions in a Google Forms survey, describing their familiarity with dimension reduction and clustering algorithms and answering two questions about exploratory data analysis.

Following this, they were provided with a short description of the tasks they must complete and the goals of the study, after which they saw the dataset for the first time. In the task (referred to as Organization Task), participants were asked to organize the observations in any way that they wished, though they were limited to grouping and spatialization operations (i.e., “place two observations in the same group” and “place two observations some distance apart”). Participants were instructed to think aloud in order to better capture their organizational process [60, 90]. In addition to notes written in real-time by the proctor, this portion of the session was video recorded for later review. Finally, participants completed a second survey of open-ended questions addressing their thoughts related to their personal analysis process.

3.4 Research Questions
As previously noted, our overarching research question is to study the organizational processes of analysts when approaching a sensemaking task with an unfamiliar dataset. We divide our more specific research questions into four broad themes, which correspond to the subsections of the Results that follow in the next section:

1. Participant Analysis Process
   (A) How do analysts begin to evaluate an unfamiliar dataset, and what actions do they take during this initial evaluation?
   (B) What are the overarching analysis strategies of study participants throughout the complete process of transforming an unfamiliar dataset into an organized space?

2. Representations Created
   (A) What types of grouping and spatial structures are created during the full analysis process?
   (B) How do the individual dimensions appear within the organizational spaces created by participants?

3. Interactions with Representations
   (A) What types of grouping and spatial operations are performed during the full analysis process?
   (B) How do participants approach individual dimensions vs collections of dimensions when exploring the dataset?

4. The Effect of Domain Knowledge
   (A) What is the role of external information in the analysis process?
   (B) How do the layouts created by the abstract condition participants change when they are provided with labeled information?

4 Results
Using notes taken during the session and video recordings, we reviewed the actions of each participant throughout their analysis process. During this review, we were purposefully searching for and noting the frequency of some behaviors, such as the types of cluster interactions previously noted in Section 2.2. For the broader participant strategies that are described in this section, we used an open coding approach to describe these events, which we later synthesized into general themes.

4.1 Participant Analysis Process
In this section, we discuss the analysis process undertaken by the study participants, with foci on both how their analysis began and how their overall strategy developed.

4.1.1 Beginning the Analysis (RQ1A)
There were two primary methods by which participants approached the Organization Task. The most common strategy, the Grid Method, was to begin by laying out the full dataset on the table, often in a grid pattern, in order to inspect the full dataset simultaneously. Slightly less often, we saw the Stack Method, in which participants kept the
We observed a tight coupling between spatialization and grouping actions performed by analysts. Rather than adopting a purely group-first or layout-first mentality, participants in this study switched between the two frequently. This was even true in cases where grouping or spatialization actions accounted for a great majority of the overall interaction total. Further, we note that this complex relationship develops over time, where spatializations are used to drive grouping and groupings are used to drive spatializations. This results in complex organizational spaces that were produced by the participants in this study. We delve into these issues further in this section.

There were three main strategies that participants performed when approaching the task (see Fig. 3), which we noted share some similarities to the investigative strategies observed by Kang et al. [43] in their document-based study using Jigsaw. Though the switch from documents to quantitative data made the precise interactions and motivations different, the strategies that we note here show patterns that are similar. For example, the Divide and Conquer strategy that we describe next resembles their Overview, Filter, and Detail strategy, while our Bottom-Up strategy is comparable to their Build from Detail.

The most common strategy seen is the **Divide and Conquer Strategy**, demonstrated in the photo in Fig. 4. In the first pass through the data, participants selected a single dimension and separate the observations into a small number of groups. They then attempted to find meaning within these smaller groups, either by selecting another dimension to separate by or spatializing within the group. After structuring the individual groups, they turned their attention to the full space and attempt to organize the large groups, with occasional refinement within the groups.

An example of this strategy is displayed by Participant L4 in Fig. 5. In Panel A at 4:47, she has binned the animals by size, creating seven temporary groups that increase in size from left to right across groups and from bottom to top within groups. Panel B at 9:04 includes a dimension for Swims, forming a structure that approximates a scatter plot. At this point, she identifies three main groups: swimming animals, sort-of swimming animals, and non-swimming animals. In Panel C at 16:46, she has decided to separate the swimming animals group as she began to sort within each group by the Furry dimension. The global Swims dimension still persists from bottom to top, but the global Size dimension is now discrete across the two columns of groups. A local Size dimension was maintained vertically in the non-swimming animals...
Fig. 5. Five stages from the analysis produced by Participant L4. The participant used the Divide and Conquer strategy, dividing the dataset into groups to be individually analyzed and refined.

The Furry dimension was vertical in the larger swimming and sort-of swimming groups, but was horizontal in the non-swimming group (and was not clearly specified in the small swimming group). In Panel D at 21:03, the two groups of large animals were positioned closer together (though kept as separate groups) and the sort-of swimming animals group was flipped vertically because the Blue Whale, Moose, and Hippopotamus have similar Fierce and Solitary attributes. The vertical axis of the smaller animals group (joined into a single group) now has a Fierce dimension with the non-swimming animals, though the Swims dimension is still maintained at the top of the group. Finally in Panel E at 31:21, the Fierce axis was rotated within the small animals group, and the vertical axis has been replaced with a Solitary dimension.

The second most common strategy was the Incremental Layout Strategy. This strategy was almost exclusive to the labeled data condition (A2 once again being the exception). Participants considered each observation one at a time, adding them to a continually growing organizational space in the location which appeared most sensible. Each of these additions was often a grouping operation, but could also be a spatialization operation in some cases. Updates to the position of observations already positioned did occur, but were infrequent. As the participants continued to add data, the physical size of the utilized space increased. After all observations were added to the space, the participants began a more thorough refinement process.

A third strategy, not as common as the first two but still implemented by multiple participants, was the Bottom-Up Strategy. Participants began similarly to Divide and Conquer, laying out all of the observations to view simultaneously. However, their next step was to begin to build small groups of two or three similar observations, usually by only looking at a single dimension at a time but then considering others. After many of the observations had been placed in small groups, spatial relationships were created between the groups, often leading to the formation of larger groups.

Regardless of the approach strategy taken by participants, they all followed an incremental pattern that consisted of a period of organization followed by a period of reflection, after which the process repeated. Such incremental formalism has been demonstrated in previous studies and system use cases [9, 29, 43, 77], but took on a different form in this study. As noted previously and confirmed by the post-survey, participants almost universally approached the organization by considering a single dimension at a time. In doing so, they created an organizational space for a dimension, and then took a step back to consider the space. They then proceeded to a second dimension, introducing its effects into the space gradually by individual animal or group, and then examined the global changes made to the structure. This alternating pattern of sensemaking and synthesis segments connected the participants’ local interactions to their global understanding of their organizational space.

4.2 Representations Created

In this section, we examine the grouping and spatial structures that were created by the participants during the course of their exploration.

4.2.1 Grouping and Spatial Structures (RQ2A)

Both groups of participants were approximately equally likely to create hierarchies and cross-cutting groups in their organizational structures (4/8 labeled and 5/8 abstract). Many participants did create internal spatial structures within their groups, but only a small subset clearly delineated groups within groups or groups overlapping groups (for example, Participant A3 in Fig. 6). In many cases, these internal groups were formed by breaking up a larger supergroup, though occasionally two subgroups were joined together to form children of a larger parent group. These overlapping groups occasionally represented “fuzzy” or “soft” cluster assignments [91].

Similarly, both groups of participants were equally likely to create organizational structures in which the axes mapped to dimensions in the data. This is contrary to the properties of many dimension reduction algorithms, in which the axes have no direct mapping to the source data. Often, such constructions resulted from participants’ behavior in focusing on a single dimension at a time and organizing the observations on a spectrum along one or more dimensions (see Fig. 7). As axes
were seen to be important, we see benefits from tools that communicate the meaning of axes in a projection as seen in InterAxis [45] and AxiSketcher [50].

We also noted that both groups of participants identified outliers in the dataset that they were hesitant to place in any group. Near the beginning of their analysis, they referred to single-observation groups as groups (or the seeds of a group), but as they continued to structure observations in the space, they were more likely to refer to these observations as not fitting well with the others. The participants who created cross-cutting groups often had subgroups with just a single observation in their organizational structure, but they were clear to identify those as equally belonging to two or more of the broader groups (again see Fig. 6).

Spatial meaning internal to groups was also seen by both participant conditions. Often, the spatializations within the groups were designed to show differentiation within observations in the group (e.g., a size trend across the group), though occasionally the goal of the participants was to show relationships between members of the group and other parts of the space (e.g., an observation within the group that is quite similar to those in other groups). As a consequence of the second, participants in both groups were equally likely to create global spatial structures that spanned the entire structure or governed large portions of their organization.

Spatial operations were occasionally the driving force behind group creation. Participants in both the labeled and abstract conditions often identified collections of observations that were similar and positioned them close together spatially before identifying that collection as a group. Conversely, groups were often used to drive spatial operations as well, especially when participants were refining group memberships and identifying distances within groups.

Participants also reported that distances within groups were more important to their structure than distances between groups. This was partially a result of the cognitive difficulty in mentally computing a distance between groups of observations as opposed to computing a distance between a pair of observations. More meaningfully, participants reported that these fine-grained differences between observations within a group were more relevant to their understanding of group structure than were the differences between the groups themselves. In other words, it was enough for participants to say “these groups are different,” but they felt the need to incorporate more spatial detail when saying “this is why these observations within a group are different.”

4.2.2 Complex Spaces (RQ2B)

After considering several dimensions, participants began to create complex spaces. This was already seen through the hierarchical, cross-cutting set of groups created by Participant A3 (Fig. 6). Another example is seen within the structure created by Participant A4 (Fig. 8), in which the participant created spectra for two of the dimensions and influence regions for the three remaining dimensions. The spectra were not orthogonal, though the C dimension was aligned with the x-axis. These are represented by the solid arrows in Fig. 8. The three regions likewise overlapped in some places but not others, indicating portions of the space in which one dimension had a great deal of influence in determining how the participant structured the layout and groups. These are represented by the dashed arcs in the same figure.

This runs counter to the common method of creating dimensionally-reduced projections, in which the entire space is governed by a single weight vector (for example, [9, 23, 75, 84]). Creating clusters that contain independent, internal weight vectors, as well as maintaining a global space, presents one solution to this challenge. For the example presented by A4, the low B region could be defined as a cluster that still maintains the influence of low A and high C. This finding presents opportunities for the introduction of subspace clustering techniques to support such complex spaces, as seen in previous works analyzing high-dimensional data [49, 63].

Further, these complex spaces are not limited to areas of attribute influence. Participants also created complex structures of hierarchical, cross-cutting groups. For example, Participant A6 created several complex sets of groups at various points in her analysis, two of which are provided in Fig. 9. This participant’s bottom-up process of connecting smaller clusters into larger groupings also reflects Gillam’s connectedness findings for cognitive grouping [36].

4.3 Interactions with Representations

Having established the created structures, we now examine the operations performed on these structures during the course of analysis.

4.3.1 Grouping and Spatial Operations (RQ3A)

We recorded instances in which participants performed four different types of grouping operations: create, remove, join, and split. These are differentiated in Fig. 10. The most common method by which participants in both study conditions approached the organizational task was to start with large groups and then subdivide. As a result, splitting an existing group into smaller groups was the most common grouping operation performed. A majority of participants also joined groups together at some point in their analysis, usually when considering a dimension for the first time and noticing new similarities among the observations. Only two participants in the abstract condition (A2 and A6) performed operations to create an entirely new group from observations previously in several other groups. None of the participants in the abstract condition removed a group and allocated its members into several other groups. In contrast, five of the eight participants in the labeled condition created and three of the eight removed a group.

Spatial operations were occasionally the driving force behind group creation. Participants in both the labeled and abstract conditions often identified collections of observations that were similar and positioned them close together spatially before identifying that collection as a group. Conversely, groups were often used to drive spatial operations as well, especially when participants were refining group memberships and identifying distances within groups.

Participants also reported that distances within groups were more important to their structure than distances between groups. This was partially a result of the cognitive difficulty in mentally computing a distance between groups of observations as opposed to computing a distance between a pair of observations. More meaningfully, participants reported that these fine-grained differences between observations within a group were more relevant to their understanding of group structure than were the differences between the groups themselves. In other words, it was enough for participants to say “these groups are different,” but they felt the need to incorporate more spatial detail when saying “this is why these observations within a group are different.”
4.3.2 Decision Making (RQ3B)

With the exception of A2, all participants in both groups spent the majority of their analysis considering only a single dimension at a time, confirming the observation seen in previous studies that analysts struggle to think high-dimensionally [3, 75]. Additionally, participants frequently processed attributes by either making binary decisions (e.g., divide observations by an attribute value greater than or less than 50), or alternatively by creating a small number of bins to discretely group observations by a single dimension. The participants commented that both the binary decisions and the binning operations were intentionally made so that they could focus their attention on subsets of the observations rather than the entire collection.

Two of the participants in the abstract condition created features from combinations of provided features while exploring the data, in both cases to reduce the amount of information that they were trying to cognitively process. Participant A5 computed the median of all five dimensions in order to perform an initial grouping, while A4 computed the difference between the final two variables she had not yet considered. Two of the participants from the labeled condition also created features, but these came from domain knowledge of the labeled animals instead. Participant L5 introduced both flying and speed features into her layout, while the other created groups that incorporated the likelihood of finding these animals in a zoo. This last participant, L1, also reported that she focused primarily on the animal labels, and did not consider the attribute values until making final refinements to the space. Kopanas et al. found similar data transformations and feature interpretations in their data mining study [47].

4.4 The Effect of Domain Knowledge

We now look at the differences between the labeled and abstract groups, with a particular focus on the role of external information.

4.4.1 External Knowledge (RQ4A)

Participants who received the labeled dataset often made use of their external knowledge about the animals that was not contained on the cards. These behaviors in response to external and domain knowledge reinforce the “Data is Personal” work by Peck et al. [65], particularly that personal experience drives decisions. Similar experience-driven behaviors have been seen in the cognitive psychology field, as evidenced by the experiments of Zemel et al [93].

External knowledge influenced participants’ organizational spaces in three ways. First, introducing external knowledge allowed the participants to begin forming groups of animals prior to closely examining the data. For example, Participant L1 created a number of groups without even considering the data, including features corresponding to environment, diet, and the probability of locating the animal in Canada. Likewise, Participant L5 introduced a flight group for the Bat into her organization, keeping it separate from all other animals as a result without considering the attribute values on the card.

Second, external knowledge was used to structure and spatialize within groups. For example, Participant L5 introduced a speed dimension into her organization, which permitted her to break up the land-dwelling (and non-flying) animals into groups based on how quickly...
they moved. The final result of this introduction was a group that contained the Deer along with the Wolf, Leopard, and Bobcat. Occasionally, incorrect domain knowledge could also affect their organization, as seen when Participant L3 used external knowledge about the diet of animals to create transition groups. For example, the Hippopotamus was used as a transition animal between Aquatic and non-Aquatic animals, not because of its mid-range Swims score, but because of the fact that she felt this animal was likely to consume fish while spending time in water as well as plants when spending time on land.

Third, external knowledge about the animals was used as confirmation, checking to see if the groups the participants created were sensible after performing a dimension-specific organizational step. Participant L2, for example, frequently questioned his positioning of the Bat throughout his iterative organizational steps. Despite the attributes on the cards supporting his decisions, he continued to second-guess its position and occasionally to modify it based on his external knowledge. Participant L4 also reported scanning the names of the animals in the groups regularly, searching for refinements that could be made to the structure internal to groups based solely on those labels.

4.4.2 “The Big Reveal” (RQ4B)

Participants who received the abstract dataset were informed of its animal dataset source after completing their organizational space. The animals cards were overlaid above the abstract cards, and participants were asked to comment on the structure and relationships that were now apparent after seeing the labeled data, effectively replicating the confirmation use from the previous subsection.

In general, participants reported being satisfied with the layout that they created in the abstract data. Many had an Aquatic group of animals due to the high-C score, and they quickly picked up on this relationship. Similarly, participants who prioritized the D dimension saw groups of predators in their organization. The Bat often was a frustration point for participants, expressing dissatisfaction with its position in the space similar to the reaction seen by participant L2.

When asked what they would change in their space given this new information, most participants reported that they would increase the density of the groups in their space, moving groups of similar animals closer together after understanding their relationship to each other. This was often seen in any Aquatic groups, Predator groups, and among the Whales. Interestingly, this behavior was not necessarily true with the animals to create transition groups. For example, the Hippopotamus was used as a transition animal between Aquatic and non-Aquatic animals, not because of its mid-range Swims score, but because of the fact that she felt this animal was likely to consume fish while spending time in water as well as plants when spending time on land.

The incorporation of external knowledge by the labeled condition participants, and the changes made to the layouts of the abstract condition participants once the labeled information was provided, both support the utility of semantic interaction in interactive systems. The goal of semantic interaction [29] is to permit the analyst to focus their exploration on the observations themselves, rather than carefully examining data values and fine-tuning the parameters of underlying models. Because the labeled group participants often brought in external knowledge, they were able to add additional detail to their organizational structures that was not contained on the index cards. The updates made by the abstract condition participants after the presentation of labels show that their spaces would have been differently structured if they had access to this information throughout the study.

5 DISCUSSION

The results collected from this study have yielded valuable information towards how humans approach exploring and organizing data. Such information can be used to guide the design of interactive visual analytics systems in the future.

5.1 Post-Survey

All participants provided responses to the survey that followed the Organization Task, which addressed their interpretation of their strategy, easy and difficult parts of the analysis, and their thoughts on the usefulness and meaningfulness of grouping and spatialization actions.

Approaching the Task: Participants reported approaching the task by considering a single dimension at a time, confirming observations from the study. The difficulty that the participants experienced when attempting to think high-dimensionally suggests the need for computational support in similar organizational tasks. Each dimension was used as a transition animal between Aquatic and non-Aquatic animals, not because of its mid-range Swims score, but because of the fact that she felt this animal was likely to consume fish while spending time in water as well as plants when spending time on land.

Number of Interactions: When asked whether they thought they performed more grouping or spatialization interactions, participants gave a variety of responses. Among those who believed they performed more grouping operations, they noted that their overarching strategy was to isolate groups within the data in order to make future processing simpler. Those who believed that they performed more spatialization interactions generally reported that they were careful when refining the organization in later stages, leading to that majority. Some participants reported that they couldn’t determine which class was the majority.

Operation Difficulty: Participants generally reported that the easiest part of the Organization Task was the beginning or the end of the analysis. Some reported that selecting a starting point, picking the initial groups, or creating the initial spatial structure was easiest. Others noted that making final refinements within and between groups at the end of the analysis was easiest. Most participants reported the mid-stages of organization to be the most challenging, needing to balance updates to existing groupings of data when examining a new dimension with maintaining existing spatial relationships. Participant A1 did report that the amount of data seen after laying out all of the cards initially was overwhelming, but that did not stop her from arbitrarily selecting a starting dimension for analysis.

Operation Usefulness: Participants were also evenly split between considering the grouping or the spatialization actions to be more useful or more meaningful. Those who felt more positively about the groupings mentioned summarizing the big picture and making sense of the overall space visually, while those who felt more positively about the spatializations thought that these actions made them more careful in their analysis. Both groups also mentioned that their operation preference for this question impacted the other. Those who preferred grouping actions noted that it made spatialization actions easier to perform, while those who preferred spatialization actions noted that it made the task of creating meaningful groups easier.

A related observation while running the study is the role of terminology, particularly with the grouping operation. There were a number of times when participants were clearly separating the observations into piles, but they were somewhat hesitant to define this organization as a “cluster” or a “group.” Frequently, we found ourselves using a variety of terms when inquiring about the structures that the participants were creating (e.g., “Do you consider this a group, or is it a cluster, or an organizational construct, or a bin, or a collection, or . . .?”). To the participants, the terms “cluster” and “group” had a different, deeper semantic meaning than a simple “pile” of observations. In order to be classified as a “cluster” or “group,” participants often wanted to perform multiple iterations of analysis to ensure that more than just a single property defined the collection of observations.

5.2 Recommendations for Tool Design

Table 1 collects findings that were uncovered while running this study, each of which is described in further detail in the preceding sections. Each of these findings is accompanied by a design recommendation for interactive visualization systems that support these interactions in the study domain (high-dimensional quantitative data). Existing systems in this area (e.g., Castor [84] and Pollux [85]) begin to address some of these interaction and design issues, but do not fully address them all. For example, neither system makes an effort to map an important dimension to an axis, support is not provided for efficient cluster splitting, and both use a single weight vector to express the
Table 1. A summary of the main findings uncovered by this study and corresponding design recommendations for interactive data exploration tools.

<table>
<thead>
<tr>
<th>Finding</th>
<th>Design Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A tight coupling was seen between grouping and spatialization actions,</td>
<td>Systems should be designed so that analysts can perform such operations at any time</td>
</tr>
<tr>
<td>frequently switching between these operations. Participants formed</td>
<td>during the analysis process, as opposed to creating sequential phases. Algorithms</td>
</tr>
<tr>
<td>groups in order to make future spatializations easier, and also formed</td>
<td>should be selected or designed to support these highly incremental feedback loops.</td>
</tr>
<tr>
<td>spatializations to make future groupings either.</td>
<td></td>
</tr>
<tr>
<td>The axes often had meaning in the organizational spaces created by</td>
<td>Aligning an important property (e.g., the first principal component or the dimension</td>
</tr>
<tr>
<td>participants, with one or more dimensions frequently parallel or</td>
<td>with the highest weight) to an axis can boost the interpretability of the projection.</td>
</tr>
<tr>
<td>orthogonal to the front of the table.</td>
<td>Alternatively, the meaning of the axes should be clearly communicated.</td>
</tr>
<tr>
<td>A common trend was to progress from large groups to small groups,</td>
<td>While many types of cluster operations (e.g., joining, creating, removing) should be</td>
</tr>
<tr>
<td>splitting groups rather than joining them.</td>
<td>supported, an efficient interaction for splitting clusters should be a design priority.</td>
</tr>
<tr>
<td>Participants reported that distances within groups were more</td>
<td>Algorithms that favor local structures (e.g., t-SNE, subspace clustering) may be better</td>
</tr>
<tr>
<td>important to their structure than distances between groups.</td>
<td>representations of an analyst’s understanding of data than those which do not.</td>
</tr>
<tr>
<td>To further reduce the complexity of the data, participants often</td>
<td>Potential means of designing such an efficient cluster-splitting interaction include</td>
</tr>
<tr>
<td>binned the observations into smaller groups or separated the</td>
<td>automatically binning observations or separating them by an analyst-specified threshold.</td>
</tr>
<tr>
<td>observations with a binary decision.</td>
<td></td>
</tr>
<tr>
<td>Participants primarily explored one dimension at a time, expressing</td>
<td>Computational support is key for efficiently communicating multidimensional informa-</td>
</tr>
<tr>
<td>frustration with trying to consider all dimensions at once.</td>
<td>tion to the analyst.</td>
</tr>
<tr>
<td>Participants often created complex spaces that combined dimension</td>
<td>Tools should support complex spaces (e.g., subspace clustering and other techniques</td>
</tr>
<tr>
<td>spectra with dimension regions of influence.</td>
<td>that favor local structures) rather than using a single weight vector to express the</td>
</tr>
<tr>
<td>Participants in the labeled group often brought their external</td>
<td>full space.</td>
</tr>
<tr>
<td>knowledge into their organizational structure, adding additional animal</td>
<td>Systems should provide interactions for annotation or other notes, permitting analysts</td>
</tr>
<tr>
<td>properties that were not provided.</td>
<td>to inject information into the system.</td>
</tr>
</tbody>
</table>

To further support the cognitive side of human-AI collaboration and co-learning [86].

6 Conclusion
In this work, we experiment with a labeled and abstract set of data to examine how analysts approach and organize an unfamiliar dataset. We wish to understand the cognitive processes that underlie the approach that analysts take when trying to find insight in data. We found that participants used groups to create spatial structures as well as spatial structures to form groups. Participants created hierarchies and cross-cutting groups in their organizational structures, and frequently approached the task by creating large groups and subdividing them to refine additional structure. The complex spaces created by participants hint towards structures that should be supported in interactive applications. We summarize a list of main findings and corresponding system design recommendations in Table 1.

Acknowledgments
The authors wish to thank those who participated in this study, as well as the reviewers for their insightful comments.

References


