With Respect to What?

Simultaneous Interaction with Dimension Reduction and Clustering Projections

John Wenskovitch, Michelle Dowling, Chris North
Discovery Analytics Center, Virginia Tech
Blacksburg, VA
{jw87,dowlingm,north}@cs.vt.edu

ABSTRACT

Direct manipulation interactions on projections are often incorporated in visual analytics applications. These interactions enable analysts to provide incremental feedback to the system in a semi-supervised manner, demonstrating relationships that the analyst wishes to find within the data. However, determining the precise intent of the analyst is a challenge. When an analyst interacts with a projection, the inherent ambiguity of some interactions leads to a variety of possible interpretations that the system could infer. Previous work has demonstrated the utility of clusters as an interaction target to address this "With Respect to What" problem in dimension-reduced projections. However, the introduction of clusters introduces interaction inference challenges as well. In this work, we discuss the interaction space for the simultaneous use of semi-supervised dimension reduction and clustering algorithms. Within this exploration, we highlight existing interaction challenges of such interactive analytical systems, describe the benefits and drawbacks of introducing clustering, and demonstrate a set of interactions from this space.

CCS CONCEPTS

• **Human-centered computing → Visualization**; **Visual analytics**; *Visualization design and evaluation methods.*

KEYWORDS

Dimension reduction, clustering, interaction, visual analytics

ACM Reference Format:
John Wenskovitch, Michelle Dowling, Chris North. 2020. With Respect to What?: Simultaneous Interaction with Dimension Reduction and Clustering Projections. In *25th International Conference on Intelligent User Interfaces (IUI '20), March 17–20, 2020, Cagliari, Italy.* ACM, New York, NY, USA, Article 4, 12 pages. https://doi.org/10.1145/3377325.3377516

1 INTRODUCTION

"With Respect to What" was first described as a usability issue with interactive projections by Self et al [57]. In their Andromeda system, analysts are presented with the two-dimensional output of a Weighted Multidimensional Scaling (WMDS) dimension reduction computation. By performing direct manipulation interactions on the observations in the projection, analysts communicate desired similarity relationships to the system. This triggers a learning routine that attempts to create such relationships in the projection by altering weights that are associated with each dimension. The process of inferring the intent of an analyst via such direct manipulation interactions and updating the visualization in response is termed "semantic interaction" [24, 26].

The usability issue that emerges from this interaction technique revolves around interpreting the analyst's intent appropriately. That is, when the analyst moves an observation to a new position, what is that movement in relation to? Is this relationship assumed or somehow explicitly communicated by the analyst (using additional interactions)? Possible interpretations for repositioning an observation in the projection include but are not limited to moving the observation away from the source, moving the observation towards a target, and moving the observation with respect to some other observation(s) within the projection. In other words, what did the analyst move, and *with respect to what?*

Such interactive projections are an increasingly popular feature in interactive visual analytics and human-centered machine learning applications [5, 7, 25, 46, 48, 49]. As a result, resolving this "With Respect to What" problem is increasingly important in order to accurately capture the intent of the analyst. In our previous work, we proposed a cluster membership solution to "with respect to what," utilizing interactive clustering reassignment to communicate similarity relationships in the projection [68, 69]. This use of clustering algorithms is a sensible choice, as implicit clusters often form naturally in dimension-reduced projections that display similarity relationships. Defining these clusters explicitly also enables explicit relationship communication to the system. Indeed, dimension reduction and clustering algorithms perform similar functions: dimension reduction algorithms simplify a dataset by reducing the number of dimensions to the most important existing or synthetic features, while clustering algorithms simplify a dataset by reducing the number of observations through grouping [67].

However, ambiguity in the interpretation of these interactions does still exist after explicit clustering has been introduced, as described in our motivating example in Section 3. In this work, our goal is to detail the interaction space for the simultaneous use of dimension reduction and clustering algorithms, particularly in interactive systems that feature a semantic interaction learning component. Specifically, we claim the following contributions:

(1) A discussion of the interaction space that exists when incorporating dimension reduction and clustering algorithms in the same projection interface, and a summary of the design factors that should be considered by visualization designers.

IUI '20, March 17–20, 2020, Cagliari, Italy
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7118-6/20/03…$15.00
https://doi.org/10.1145/3377325.3377516

(2) A system pipeline representation to demonstrate and detail these interactions.
(3) A discussion of additional implementation factors external to the "With Respect to What" problem that should be considered by visualization designers.

2 BACKGROUND

In this section, we provide three background components. In the first subsection, we briefly survey the use of dimension reduction and clustering algorithms in visual analytics systems, both independently and in combination. Following this, we describe the "With Respect to What" problem in the context of interactive dimension reduction (and occasionally interactive clustering) visual analytics systems, demonstrating the interactions supported by such systems. The section concludes with other works that have surveyed direct manipulation interactions in projections.

2.1 Dimension Reduction and Clustering in Visualization

2.1.1 Dimension Reduction. The goal of dimension reduction algorithms is to create a low-dimensional representation of high-dimensional data that preserves high-dimensional structures such as outliers and clusters [42]. Surveys of dimension reduction algorithms can be found in the literature [28, 29, 71]. While these representations can be of any number of dimensions smaller than the cardinality of the high-dimensional space, dimension reduction algorithms are most often used to reduce the dataset into a two-dimensional projection displayed as a scatterplot or node-link diagram. A number of dimension reduction techniques are prevalent in the visual analytics literature. In Andromeda and SIRIUS, WMDS is used to project a dataset into a two-dimensional representation [21]. Force-directed layout algorithms are also common [5, 25, 68], while other systems project data using Principal Component Analysis (PCA) [7] or t-distributed Stochastic Neighbor Embedding (t-SNE) [13].

2.1.2 Clustering. The goal of clustering algorithms is to group sets of observations so that observations in the same group are more similar to each other than to those in other groups. Surveys of clustering algorithms from various perspectives can also be found in the literature [14, 72]. Clustering algorithms can be classified into hierarchical and partitioning families, with the hierarchical family further split into divisive (top-down) and agglomerative (bottom-up) types [67]. The variety of methods for presenting clusters in visualization systems is nearly as broad as the variety of clustering algorithms themselves. Among others, the most common techniques are using color to denote cluster membership [1, 16, 30, 41], encoding clusters by position [15, 47], and enclosing groups of observations with distinct boundaries [47, 68]. It is also common to use dual-encoding [33] to reinforce cluster membership [13, 38].

2.1.3 Dimension Reduction and Clustering. The natural relationship between dimension reduction and clustering algorithms has long been recognized. Indeed, Ding and He proved that Principal Component Analysis (PCA) implicitly performs clustering as well as dimension reduction; the principal components are the continuous solutions to the discrete cluster membership indicators for k-means clusters [18]. Similarly, self-organizing maps are a dimension reduction technique that can be interpreted as a set of clusters [39].

Observing this relationship, a number of visual analytics systems include both dimension reduction and clustering algorithms. These algorithmic combinations come in a variety of visual representations, and the algorithms also process the data in a variety of ways. For example, iVisClustering [41] performs both dimension reduction and clustering on the high-dimensional data, implying that a change in the layout does not affect the clustering assignment. In contrast, both "Be the Data" [9] and Castor [68] performs clustering on the output of the dimension reduction algorithm, rendering low-dimensional clusters that are dependent on the positioning of observations in the projection. However, "Be the Data" uses color encoding to represent cluster membership, while Castor takes the distinct boundaries approach. Reversing this algorithmic order, Ding and Li create a system in which k-means clustering is used first to generate class labels, followed by Latent Dirichlet Allocation (LDA) dimension reduction for subspace selection [19].

2.2 "With Respect to What"

Given that the "With Respect to What" problem is defined by how the analyst interacts within the projection, it is first important to consider common types of interaction schemes. First, many of these interactions are considered visual to parametric interactions (V2PI), which is defined by Leman et al. [43] and explored by Hu et al [35]. At a high level, this paradigm considers interactions that are performed directly within a projection of the data. Given an interaction, parameters for the underlying projection model are *learned*, resulting in a new projection. As a result, analysts are able to remain within their cognitive zone [31], thereby enhancing analysts' efficiency in performing their analytic tasks.

Bayesian Visual Analytics (BaVA) as defined by House et al. describes a probabilistic variation of V2PI [34]. However, the deterministic variations are more commonly seen across current visual analytic implementations, including observation-level interaction (OLI) [27]. OLI specifically defines interactions on projected observations of data in which relative pairwise distances between a subset of points is defined by the analyst. From these pairwise distances, new parameters for the underlying distance metric are learned, which in turn are used to produce an updated projection of the data. Note how all of these schemes support incremental formalism [59], which enables analysts to gradually concretize their hypothesis as they investigate the data.

These interaction schemes can be applied in a wide variety of applications. Each method has implications for how the "With Respect to What" problem can or should be solved. For example, using control points within a visualization is a common method for enabling interactive and iterative refinement of the projection [5, 17, 20, 25, 36, 45, 48, 50, 58, 66]. Control points often take the form of analyst-selected and manipulated points within the projection, but these control points can also be represented as anchors on the projection boundaries as well. In either case, the analyst is manipulating a given point with respect to the entire visualization. That is, this interaction is meant to have an effect on a global scale rather than performing local refinements. This concept is reflected in the fact that a single movement of one control point typically results in all other non-control points moving themselves in relation to the control point's new location.

Rather than use control points, some tools instead use manipulated points to describe desired pairwise distances in a projection [7, 21, 53, 55–57, 68]. As a result, the information that is communicated through this interaction is a desired set of distances expressed as relative pairwise distances between the moved points, which often reflect similarity/dissimilarity relationships in the data. Using these relative pairwise distances, the system *learns* a new distance metric, typically by updating the parameters of the chosen distance function. The new distance function is then applied to all projected data, not just the interacted points, to produce an updated visualization. The implied "With Respect to What" in such interactions is limited to the points the analyst interacted with; all other points are ignored until the data is reprojected.

There are still other types of interactions which address this "With Respect to What" problem. Podium [65] allows analysts to interactively alter the rank of any item in a table. Podium explicitly defines this interaction to be with respect to the other rows that changed ranks as a result of the interaction. Additionally, ReGroup [3] enables interactive cluster formations in which each additional item that is added to a cluster results in updating a list of suggested items to add to the cluster. Thus, this interaction is with respect to the existing items within the cluster. Andrienko et al. take yet another approach in which cluster definitions can be interactively altered by the analyst, such as merging or splitting clusters [4]. Such interactions are with respect to the involved clusters (i.e., the clusters that are being merged together or which cluster is being split). These examples demonstrate the variety of manners in which the "With Respect to What" problem can be addressed, indicating the vast design space present in this area.

2.3 Surveys of Projection Interactions

Interactive dimension reduction is currently a heavily-researched area of visual analytics, and as such, a number of recent surveys have been published that survey various aspects of the space. For example, Sacha et al. present a structured literature review of dimension reduction, with portions of their analysis discussing the interactive, semantic interaction-driven topics that are most relevant to this work [54]. Other surveys of the dimension reduction literature have been produced by van der Maaten et al. [63], Wismüller et al. [71], and Liu et al. [44].

Still other surveys focus on interactions that underlie the ideas of semantic interaction and interactive model manipulation. For example, Buja et al. [8] present a review of interaction techniques for high-dimensional data visualization, which von Landesberger et al. [64] construct an interaction taxonomy to track and analyze user interactions in visual analytics. Similarly, Brehmer and Munzner [6] present a thorough description of dimension reduction tasks, while Yi et al. [73] reduce relevant dimension reduction interactions to a set of low-level interactions.

3 MOTIVATING EXAMPLE AND DESIGN SPACE OVERVIEW

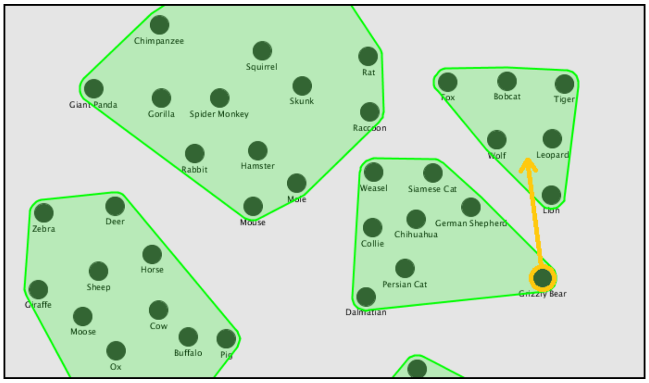

Figure 1: An analyst repositions the Grizzly Bear observation within the projection, indicated by the orange arrow.

To motivate our discussion of the dimension reduction and clustering interaction space, consider the example shown in Figure 1. In this example, an analyst is provided with a dimension-reduced projection of an animal dataset [40], positioned according to their attribute relationships with initially equal weights. A clustering algorithm then groups the observations into discrete categories, following the "Dimension Reduction Preprocessing for Clustering" pipeline described in our previous work [67]. After viewing the projection, the analyst chooses to reposition the Grizzly Bear observation, removing it from one cluster and placing it into another. With this simple interaction, the analyst could be trying to convey a number of possible intents to the system.

Perhaps the analyst is looking specifically at the relationships between the animals in the projection. For example, the analyst could be trying to convey a relationship about the starting position of the observation ("the Grizzly Bear is not similar to the other animals near the source") or a relationship about the ending position of the observation ("the Grizzly Bear is more similar to the other animals near the destination"). There is also the question of how many observations the analyst considers; the analyst could be trying to communicate a relationship with respect to the closest observation ("the Grizzly Bear is most similar to the Lion"), the closest *n* observations, all observations in a cluster, or all observations in the projection. These types of relationships would be best handled by the Dimension Reduction Model.

Alternatively, the analyst may have mapped some semantic meaning onto the cluster groupings in the projection, trying to communicate a membership assignment based on those groups ("the Grizzly Bear is a better fit in the Predators cluster than in the Pets cluster"). Such relationships could incorporate both the source and the target cluster, or perhaps a case where the target is irrelevant ("the Grizzly Bear appears to be an outlier in the Pets cluster and belongs elsewhere") or the source is irrelevant. These relationships would be best handled by the clustering algorithm.

The analyst may also be trying to communicate a relationship that includes both observations and clusters. In such cases, the relationship may be relevant to all observations within the cluster ("the Grizzly Bear is more similar to the observations in the target cluster than the source cluster"), or the precise positioning of the observation within the cluster may be important ("the Grizzly Bear

Table 1: A collection of example intents that an analyst may wish to communicate via repositioning an observation or a cluster in a projection of the Animals dataset [40]. Creating unambiguous interactions to support each of these potential intents remains an open challenge.

			Analyst Repositions	
			An Observation	**A Cluster**
With Respect to What	**Nearest 1**	Observation	The Grizzly Bear is most similar to the Polar Bear.	The Predators cluster shares few similarities with the Blue Whale observation.
		Cluster	The Grizzly Bear is similar to many of the other members of the Predators cluster.	The Predators cluster is dissimilar from the Large Herbivores cluster.
	Nearest *n*	Observations	The Grizzly Bear behaves similarly to animals such as Wolves, Leopards, and Lions.	The Predators cluster shares few similarities with the aquatic animals.
		Clusters	The Grizzly Bear is a predatory animal.	The Scavenging Predators cluster is similar to the small actively hunting and large actively hunting predators.
	Cluster	Single	The Grizzly Bear belongs in the Scavenging Predators cluster.	The Predators cluster are similar to the Grizzly Bear observation.
		Multiple	The Grizzly Bear is a predator.	The Scavenging Predators cluster is a subset of the overall Predators group.
	All of the	Observations	The Grizzly Bear is more similar to the predatory animals on the left than the herbivorous animals on the right.	The Scavenging Predators cluster is more similar to the other carnivorous animals on the left than the herbivorous animals on the right.
		Clusters	The Grizzly Bear is more similar to the predatory animal clusters on the left than the herbivorous animal clusters on the right.	The Scavenging Predators cluster is more similar to the carnivorous animals clusters on the left than the herbivorous animal clusters on the right.

belongs in the Predator cluster, but it is not similar to the small predator (Fox)").

The examples in the preceding paragraphs suggest two primary dimensions to consider when judging the intent of the interaction. First, the interaction could be applied to the observations, the clusters, or both. Second, the interaction could be applied to a variety of cardinalities: the nearest observation, the nearest *n* observations, all observations within a cluster, or all observations in the projection. These dimensions are summarized with respect to the Grizzly Bear observation and Predators cluster in Table 1 and are expanded upon in the following sections. However, it is also worth considering further aspects of the interaction and of the visualization itself. We discuss several additional dimensions of this interaction space in the next section.

4 INTERACTIONS ON OBSERVATIONS AND CLUSTERS

This section presents interactions and their potential interpretations when interacting with both dimension reduction and clustering algorithms. Given the discussion that follows, we summarize the following factors that should be considered by a designer when mapping the intent of an analyst to an interaction in this space:

- **Interaction Target:** An interaction could be applied to the observations, the clusters, or both.
- **Cardinality:** An interaction could be applied to a variety of cardinalities: the nearest observation or *n* observations, all observations within a cluster, or all observations in the projection.
- **With Respect To What:** Is the relationship relative to a target at the source or destination location, or both?
- **Level of Thinking:** When performing an interaction, is the analyst is thinking high- or low-dimensionally? In other words, is the analyst merely altering the projection, or are they considering all properties of a group of observations?
- **Visual Design:** Is the intent of the interaction influenced by the way that observations and clusters are encoded in the visualization?
- **Algorithm Order:** Is the DR or the clustering processed first? Or are they simultaneous?

In this section, we first consider the possible interpretations that result when an analyst repositions an observation in the projection. We begin by describing observation interactions that affect observations, before moving into interactions that affect clusters, and then turn to interactions that affect both. Subsequently, we consider the possible interpretations that result when an analyst repositions a cluster in the projection. This discussion begins by describing cluster movement interactions that communicate a similarity relationship to other observations or clusters, followed by a discussion for interactions unique to the cluster-to-cluster relationship. In some cases, we cite systems that demonstrate interaction properties that we present. In other cases, no such system currently exists, and so our presented interaction challenges are more speculative. We summarize some of the properties of these interactions in Table 2.

4.1 Observation Interactions with Respect to Observations

As detailed by the left column of the intents in Table 1 and summarized in the previous section, when an analyst repositions an observation, the system must determine what the analyst is moving the observation with respect to. The analyst might be repositioning the observation to move it away from something near the source, towards something near the target, or relative to any other observation in the projection. The analyst might also be repositioning the observation relative to the position of just a single observation or a collection of *n* observations.

Determining the other component(s) involved in the interaction is an additional challenge, particularly when differentiating

Table 2: A summary of interactions by cardinality, the importance of the interaction source, target, or both, and whether an analyst is typically thinking high-dimensionally, low-dimensionally, or both.

	Cardinality	Source/Target	High-D/Low-D
Observation–Observation Similarity			
Move observation towards another observation	1:1	T	LD
Move observation away from another observation	1:1	S	LD
Move observation towards several observations	1:*n*	T	LD
Move observation away from several observations	1:*n*	S	LD
Observation–Cluster Similarity			
Move observation towards a cluster	1:1	T	B
Move observation away from a cluster	1:1	S	B
Move observation towards several clusters	1:*n*	T	B
Move observation away from several clusters	1:*n*	S	B
Cluster–Observation Similarity			
Move cluster towards an observation	1:1	T	B
Move cluster away from an observation	1:1	S	B
Move cluster towards several observations	1:*n*	T	B
Move cluster away from several observations	1:*n*	S	B
Cluster–Cluster Similarity			
Move cluster towards another cluster	1:1	T	B
Move cluster away from another cluster	1:1	S	B
Move cluster towards several clusters	1:*n*	T	B
Move cluster away from several clusters	1:*n*	S	B
Observation Change in Membership			
Move observation into cluster	1:1	T	HD
Move observation out of cluster	1:1	S	HD
Move observation between clusters	1:*n*	B	HD
Move observation external to clusters	1:*n*	B	HD
Move observation within a cluster	1:1	B	HD
Cluster Change in Membership			
Move cluster into cluster	1:1	T	HD
Move cluster out of cluster	1:1	S	HD
Move cluster between clusters	1:*n*	B	HD
Move cluster external to clusters	1:*n*	B	HD
Move cluster within a cluster	1:1	B	HD
Join/Split Clusters			
Join Clusters	*n*:1	T	HD
Split Cluster	1:*n*	T	HD
Create/Remove Clusters			
Create Cluster	1	T	HD
Remove Cluster	1	S	HD

between movements with respect to 1, *n*, or all observations. The most straightforward solution to this challenge is to provide analysts with one or more selection mechanisms. For example, Andromeda implements a number of methods to permit the analyst to choose all elements necessary for the interaction. At the source of the interaction, the nearest neighbor is selected by default. At the target, all observations within a set radius of the target position of the interaction are selected. Following these default selections, the analyst is permitted to select or deselect any observation in the projection. An example of each of this selection mechanism is shown in Figure 2, in which an analyst has repositioned the Beaver observation closer to the whales, possibly signifying an interest in exploring aquatic-dwelling animal behavior. As the nearest neighbor to the source, the Polar Bear was automatically selected as part of the interaction, as were the Blue Whale and Humpback Whale in the target radius. After this observation was repositioned, the Wolf was also selected to denote dissimiliarity between land-dwelling and water-dwelling animals.

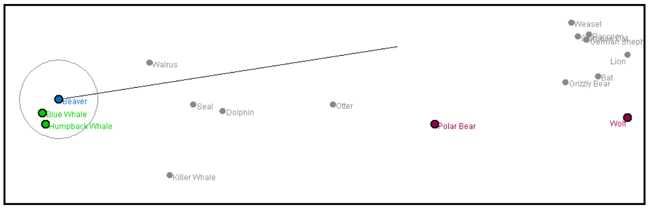

Figure 2: Selection interactions in Andromeda [55]: nearest neighbor selection at the source, radius selection at the target, and additional observation selection in other regions.

The further possibility exists that the analyst does not wish to alter any underlying models with the interaction they provide. Instead, they may be merely exploring the current projection. Endert et al. define these categories of exploration as exploratory and expressive: *exploratory* interactions provide an analyst with insight into the structure of the data, whereas *expressive* interactions communicate an intent to the system and effect underlying models [27]. For example, Castor treats interactions that do not cross cluster boundaries as exploratory, allowing analysts to investigate relationships between observations without affecting the underlying learning system [68]. This is generally true for drag interactions in other systems that incorporate force-directed layouts, such as ForceSPIRE [25] and StarSPIRE [5].

That said, StarSPIRE allows for explicit interactions by having the analyst overlap the boundaries of two documents. In this case, the system interprets this interaction as the analyst expressing not just document similarity, but their immediate, close relatedness. Thus, StarSPIRE uses this interaction to increase the weight associated with all shares entities between the two documents, resulting in an updated projection that includes new documents discovered through semantic interaction foraging [5].

4.2 Observation Interactions with Respect to Clusters

Repositioning an observation with respect to a cluster leads to a further set of challenges, primarily centered around the means by which cluster information is encoded in the projection. This is due to the fact that the visual encoding of clusters leads to different affordances for interaction. An important consideration when introducing both dimension reduction and clustering interactions in the same interface is determining the order of these algorithms. If the dimension reduction algorithm runs first, then the clustering algorithm is computed on the low-dimensional data. In contrast, if the clustering algorithm runs first, a layout needs to be constructed to appropriately project these clusters into the two-dimensional view necessary for display. It is also possible to perform both computations in the high-dimensional space at the expense of computational

Table 3: A collection of example interactions and intents that an analyst could communicate via reclassifying an observation with respect to a cluster in a projection that uses cluster boundaries.

Moving an Observation	Intent Expressed by the Analyst
Into a cluster	The Grizzly Bear is a predator.
Out of a cluster	A Grizzly Bear is not a pet.
Between clusters	The Grizzly Bear is better classified as a hunting predator than a scavenging predator.
External to all clusters	The Grizzly Bear is more like the large cats than the wolves, though it belongs to neither group.
Internal to a cluster	The Grizzly Bear is a predator, and it is more like the large predators than the small predators.

efficiency. Our previous work [67] discusses tradeoffs with respect to algorithm order for such systems in more detail.

In the remainder of this subsection, we first consider clusters defined by an explicit border, as in the motivating example from Section 3 and Figure 1. After this discussion, we summarize these interactions with respect to color and cluster hierarchies.

4.2.1 Boundaries. Explicit cluster boundaries in a projection suggest that repositioning an observation into or out of a cluster is communicating a membership assignment to the system. Such an interaction then could be interpreted in a variety of ways: an observation is being repositioned into a cluster, out of a cluster, between clusters, separate from all clusters, or internal to a cluster. Regardless of the interaction performed, the observation is being repositioned with respect to some cluster at the source or target of the interaction. As a result, a distance between the repositioned observation and the source and/or target clusters is necessary to model the high-dimensional relationship between these entities. A collection of example observation reclassification interactions and their related intents are included in Table 3. Castor presents an example of an explicit cluster boundary system, treating any observation reposition that crosses a cluster boundary as an expressive interaction [68].

4.2.2 Color. If clusters are encoded by a mechanism other than boundaries, such as color, then the natural interactions afforded by the system will change. Color is often used to demonstrate cluster assignments in systems where items belonging to different clusters may be positioned nearby in a projection [1, 13, 38]. In other words, explicit cluster boundaries are more easily interpreted when cluster regions can be easily and accurately expressed by separate, non-overlapping regions. Simply repositioning an observation into a multi-colored grouping of observations will not always be sufficient to communicate a new cluster assignment. Instead, an alternative cluster reassignment mechanism would be preferred. For example, clicking on an observation could cycle through its possible colors and therefore its cluster assignment.

4.2.3 Cluster Hierarchies. If clustering is hierarchical, then the metric learning process becomes more complex. Such a system must to evaluate the intended position in the overall hierarchy at which an observation began and where it ended. To use the Animals dataset example, the Grizzly Bear could belong to a Large Predators cluster, a subset of the overall Predators cluster which in turn is a subset of a Carnivores cluster. If the Grizzly Bear is moved from its current position, then in addition to determining if the source of the interaction is important, the system must also determine which level of the hierarchy is the relevant part of the source. Therefore, the learning relationships also may depend upon not only the source and target cluster of the interaction, but also the parent clusters and their properties at each endpoint. A recursive computation of cluster properties and weights may be necessary to consider the full hierarchical structure.

4.3 Observation Interactions with Respect to Both Clusters and Observations

In addition to the prior examples, an analyst may also wish to communicate both position and membership information simultaneously via an interaction on an observation. Again, consider the interaction in the motivating example from Section 3. The analyst may wish to communicate that the Grizzly Bear belongs in the Predators cluster while also communicating that the Grizzly Bear is more similar to the large predators in the cluster (Lion, Tiger) than it is to the small predators (Fox, Bobcat). In such a case, factors from both of the previous two subsections must be considered.

4.4 Cluster Repositioning Interactions

Much like repositioning an observation with respect to another component of the visualization, relocating a cluster to a new position may need to consider the source and target positions of the cluster, as well as the positional relationships to other observations and clusters. The right column of Table 1 summarizes some cluster intents with respect to both observations and clusters.

However, a significant difference between observations and clusters is the space used by each interaction target in the projection – observations require a single point, while clusters require a broader space. As a result, a visualization designer should consider how to compute the location and value of a cluster in these interactions. Such computations could consider a simple centroid of the cluster, or potentially a weighted centroid based on the position of each observation within the cluster. Further, distances between a cluster and a second component of the visualization could be computed in a variety of ways, including single linkage, average linkage, and complete linkage [60]. Given that a cluster has a complex value that could be determined in a variety of ways, determining how to update underlying weights based upon the result of an interaction is also a complex decision.

There is further ambiguity with respect to drag interactions on clusters, particularly in the case where cluster membership is encoded by boundaries. Consider the case in which one cluster is dragged into another – is the analyst intending to join the clusters, or to express a hierarchical relationship between those clusters? It is also possible to use a drag interaction to reposition the boundary of a cluster without relocating the observations that it encloses. Such an interaction could be used by an analyst to correct for misclassifications, encapsulating additional observations within the cluster by performing an interaction to shift the boundary. This interaction should be interpreted by the system with a very different meaning, as the analyst is again performing an expressive interaction to reclassify observations.

4.5 Cluster-Specific Interactions

In addition to the repositioning relationships that can be implemented for clusters, a further set of cluster-specific interactions are possible to implement through ambiguous operations upon a projection. For example, consider the interaction in which an analyst drags one cluster towards another until their boundaries slightly overlap. One potential interpretation for this interaction is that the analyst is again providing a similarity relationship, indicating that these clusters are similar. On the other hand, the analyst could be intending that the clusters be merged into a single, larger cluster.

Such ambiguity within a single interaction computation is not limited to joining clusters. For example, consider a sequence of interactions in which an analyst drags some nodes to the left side of a cluster and others to the right. In the Castor approach [68], such an interaction is interpreted as exploratory, and as such is not handled by either interaction computation. However, the analyst could also be indicating an intent to split this cluster into two smaller clusters. In this case, both clustering and dimension reduction update computations are necessary: the clustering to create the new clusters and update assignments, and the dimension reduction to examine the dissimilarities between the groups that the analyst formed.

Further, an analyst could also attempt to create a new cluster by repositioning a set of observations into a single region of the projection. Again, both clustering and dimension reduction computations are necessary to judge this intention, as the analyst is creating a new cluster while also communicating similarity relationships amongst the collection of observations that are grouped together. Other interactions such as removing clusters, growing or shrinking the size of a cluster, and increasing or decreasing the importance of a cluster could be implemented through ambiguous interactions that must be interpreted to understand analyst intent.

5 AN INTERACTION-BASED PIPELINE REPRESENTATION

To represent and disambiguate the nuances of these interactions, we propose a pipeline representation that extends the work by Dowling et al [22]. In this pipeline representation, computations to create visualizations and respond to interactions are encapsulated within independent models. Each of these models includes an algorithmic component for each of those computations. Data flows clockwise through the pipeline representation, initially creating a visualization by transforming the initial data through the sequence of computations to produce a visualization. When an analyst performs an interaction, the computational flow continues clockwise through a sequence of computations that may or may not execute depending on the interaction. These computations provide some update to an underlying weight vector, which is then converted into a new visualization in response to the interaction through a followup execution of the visualization creation direction.

From this pipeline representation, we incorporate the following modifications for clarity of communicating an interaction:

- **Nomenclature**: We update the “forward” computations to instead be “projection” computations, and we update the “inverse” computations to instead be “interaction” computations, clarifying the purpose of each computation.

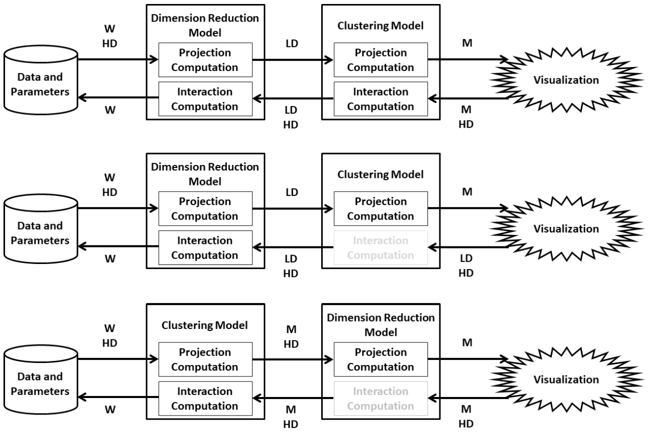

Figure 3: Three pipelines that demonstrate separate interactions that can be incorporated into interactive dimension reduction and clustering systems. In each of these pipelines, W=weights, HD=high-dimensional data, LD=low-dimensional data, and M=cluster membership.

- **Edges:** We annotate the edges that connect the data, models, and visualizations to demonstrate the type of data that are updating during the upcoming computation.
- **Model Computations:** If a model projection or interaction computation does not need to execute, we gray out the computation to make it clear that the computation is not relevant to the interaction in question.

Several of these updated pipelines are provided in Figure 3. Each pipeline maps to a separate interaction. For example, consider the top pipeline in the figure. This pipeline demonstrates data flow that includes both a Dimension Reduction and a Clustering Model in the computational pipeline. In this representation, the Dimension Reduction Model is in charge of handling positional updates to observations, while the Clustering Model handles cluster membership updates within the projection. These models therefore take on varying levels of importance depending on the interactions that have been encoded into a system. Both the Projection Computation and the Interaction Computation of both models are activated in the pipeline, so both are relevant to the intended interaction. As the data flow is processed clockwise through the pipeline, the Interaction Computation of the Clustering Model executes before the the Interaction Computation of the Dimension Reduction Model. Using the Animals dataset again, an interaction that maps to such a pipeline is “Move the Grizzly Bear into the Predators cluster, but closer to the Lion and Leopard than the Fox.” Such an interaction demonstrates the importance precedence of the clustering reassignment, while also factoring in the target position of the observation that has been repositioned.

The middle pipeline is nearly identical to the one above it, but the Interaction Computation of the Clustering Model has been grayed out. As such, this figure indicates that no clustering computation exists within the interaction modeled by this pipeline; the interaction is instead purely focused on the Dimension Reduction model and the updated position of the observations within the projection. An interaction that maps to such a pipeline is “Move the Grizzly

Bear closer to the Leopard," only considering the spatialization of the projection and not any cluster membership assignments.

The bottom pipeline has now grayed out the Interaction Computation of the Dimension Reduction Model rather than that of the Clustering Model, indicating that no dimension reduction computation exists within the interaction modeled by this pipeline. Because the Clustering Model is the only relevant model to the interaction path, a corresponding interaction that maps to such a pipeline is "Move the Grizzly Bear into the Predators cluster," which only considers the cluster membership of the Grizzly Bear with no focus on the relationship between the Grizzly Bear observation and the position of other observations within the target cluster.

Additionally, the bottom two pipelines swap the computational order of the Dimension Reduction and Clustering Models. While this had no effect on the Interaction Computations, it could have if one of the interactions had not been grayed out. Instead, the primary difference in these pipelines is located in the Projection Computations. As noted in Section 2.1.3, running the Dimension Reduction Model before the Clustering Model in the projection direction shows clusters in the low-dimensional space. If the projection displays low-dimensional clusters, a cluster reassignment can be interpreted as informing the system that the high-dimensional interpretation of groups in the data does not match the low-dimensional classification. In such a pipeline, the analyst is most likely reasoning in low-dimensional space in updating the potentially incorrect clustering assignments from the reduced dataset. As a result, the high-dimensional data is not a focus when interpreting the interaction. Instead, updating the low-dimensional data weights to better reflect high-dimensional relationships is key to these interactions.

In contrast, running the Clustering Model before the Dimension Reduction Model in the projection direction shows that high-dimensional clusters are being projected into a low-dimensional projection. In such a pipeline, the analyst is most likely reasoning in the high-dimensional space, updating the relationships between observations in the high-dimensional space. Further, swapping the projection order of the Dimension Reduction and Clustering Models is demonstrated more thoroughly in the next section.

6 INTERACTION USE CASE

The interaction space for dimension reduction and clustering algorithms is certainly vast. In order to demonstrate one of these many interactions, we use Pollux [69] system, one of the few systems that exist at the interaction of dimension reduction, clustering, and semantic interaction. Among others, this system incorporates the following interaction from Table 3: if an observation is dragged across a cluster boundary, it is treated as a cluster reassignment without consideration for positioning (the bottom pipeline from Figure 3). Using the design factors from Section 4, we can express this interaction as follows:

- **Interaction Target:** The interaction is to an observation.
- **Cardinality:** The interaction is applied to a pair of clusters.
- **With Respect To What:** Both the source and target cluster are important to this interaction.
- **Level of Thinking:** The analyst is assigning a new cluster membership to the observation, considering many dimensions that constitute that decision.
- **Visual Design:** Clear cluster boundaries need to be crossed to perform the interaction.
- **Algorithm Order:** The Interaction Computation of the Clustering Model is the only one that executes.

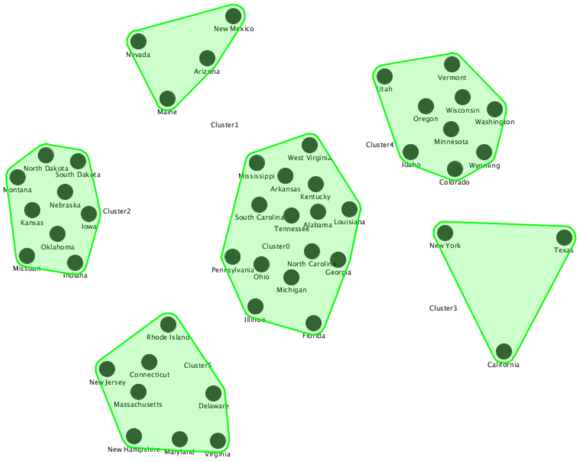

Figure 4: The initial projection of the Census dataset [23] in Pollux [69]. The nascent Midwest cluster is on the left.

In this use case, we make use of a U.S. Census dataset [23]. The United States Census Bureau defines the Midwest region as a collection of 12 states, ranging from Ohio in the east to the Dakotas in the west [62]. In the initial equally-weighted projection, 7 of the 12 states were already grouped appropriately in a cluster (see Figure 4). We needed to add five states to the nascent cluster and remove two others. To do so, we performed the following interactions:[1]

(1) Move Ohio into the cluster.
(2) Move Michigan into the cluster, an action which also corresponded to the system beginning to learn our intent and pulling Minnesota in automatically.
(3) Move Illinois into the cluster, which also automatically brought in Wisconsin (hoped for) as well as Arkansas and Kentucky (unintended). There are now four states to remove from the cluster.
(4) Remove Kentucky, an action which also resulted in the automatic removal of Montana.
(5) Remove Oklahoma.
(6) Remove Arkansas

The design of this interaction permits an analyst to clearly demonstrate their intent with respect to the clustering model. The visual feedback provided in the visualization immediately shows the analyst the result of their interaction. However, the design currently does not make clear whether that result comes from moving a state out of a cluster or into a cluster (or both). Additionally, there is no visual indication during the interaction that the positioning of the observation at the end of the interaction is irrelevant. The Pollux system could be improved by showing such visual cues, such as altering the color of both the source and target cluster boundary during the interaction to communicate their importance to the model update.

[1] A demonstration of several of these interactions is provided at youtu.be/1AYdfDYyatk.

7 DISCUSSION

In this section, we discuss several additional considerations that a system designer should factor in when designing a system that incorporates an interactive projection with ambiguous interactions. Many of these considerations are useful for such applications but are not directly applicable to the scope of the "With Respect to What" problem. In addition, we discuss bigger-picture considerations of this work, both with respect to the future of semantic interaction and to the relationship between HCI and ML. Further, we discuss the limitations of this work and present future research opportunities in this interaction design space.

7.1 Visualizing the Feedback

As an analyst performs interactions in a projection, some of the ambiguity can be removed by providing the analyst with visual feedback demonstrating how the system will interpret the interaction. Such features are similar to those seen in Explainable AI systems [32] as they reveal details of the underlying model state. Some tools with interactive projections have already implemented such feedback techniques. For example, StarSPIRE [5] includes a feature within the documents to highlight words judged to be important by the underlying models. This feature is used by analysts both to determine the overall importance of a document within the projections and to locate the important phrases and sections of a document [66]. In contrast, Andromeda [55] uses a dynamic-length slider to indicate the weights applied to dimensions. Related techniques seen in visual interfaces include changing the color of observations or colors (also seen in Andromeda) and drawing boundaries around tentatively-recognized clusters.

7.2 Communication between Models

When incorporating both Dimension Reduction and Clustering Models in the same tool, there will often be a need to provide communication between the models. In other words, learning optimal weights for the new projection could depend upon the interaction computations of the Dimension Reduction and Clustering Models working together to determine the optimal configuration. This balances an updated clustering reassignment with precise low-dimensional coordinates of the interaction target. As such, the interaction computations of the Dimension Reduction and Clustering Models require some internal communication and negotiation to jointly determine the intent of the analyst.

For example, an ambiguous interaction discussed previous is in dragging an observation into a cluster – the analyst may simply be communicating a new cluster membership assignment for the observation in question, or they may also wish to convey the importance of the final position of the observation with respect to other observations within the cluster. In contrast, if the analyst drags one cluster fully into another, they may be demonstrating an intended hierarchical relationship between these two clusters. Both of these potential interactions are naturally handled by the Clustering Model, necessitating that the interaction computation of the Clustering Model know how to determine which of these analyst intents is best matched by the interaction. No Dimension Reduction Model influence is necessary for such an interaction.

Resolving the ambiguity in these interactions is certainly complex. As such, providing additional visual feedback to communicate the system's interpreted intent can help to resolve any resulting issues from these interactions. Techniques such as visual scent [10–12, 51, 70] can be used to convey the effects of an interaction to the analyst, so that the analyst has the ability to correct their interaction before the system begins to respond to the interaction.

7.3 Shared or Separate Weight Vectors

Much of the discussion in the sections above has glossed over the changes made to the system parameters after handling an interaction. In single-model systems like Andromeda, the weight update is straightforward as there is only one model learning weights. In multi-model systems such as StarSPIRE, the weight update becomes moderately more complex, as a relevance threshold needs to be learned in addition to set of term weights. Further increasing in complexity, SIRIUS [21] uses two separate weight vectors, one for the observation projection and one for the attribute projection, each of which are computed separately but with some dependency between them as interactions are performed.

Thinking more generally about a visualization system that incorporates both dimension reduction and clustering algorithms, a designer should consider whether each model should maintain its own weight vector or if a weight vector should be shared between models. The tasks supported and pipeline selected both play a large role in this decision. For example, Castor [68] follows the "Dimension Reduction Preprocessing for Clustering" projection pattern, with the cluster assignments naturally following from the low-dimension positions of the observations. In such a case, it is natural to support a shared weight vector between the models. In contrast, a system like iVisClustering [41] which computes dimension reduction and clustering separately on the high-dimensional data without interaction between the two (the "Independent Algorithms" pattern [67]) naturally supports separate weight vectors.

7.4 Parametric Interactions

In contrast to the interactions discussed in previous sections, Self et al. also defined a class of *parametric interactions* [57]. These parametric interactions provide explicit instructions to the system, bypassing the metric learning step necessary in the interaction computations and setting a precise value for a weight or other parameter. Though a different class of interactions, these interactions are still useful in visualization systems. Self et al. further identified a collection of low-level analytical tasks in Andromeda that are solved more efficiently by parametric interactions, particularly in interactions that focus on identifying values of a small number of dimensions [55]. In addition to the slider bars in Andromeda, a variety of techniques have been implemented to afford this functionality, including but not limited to Star Coordinates [37] and SpinBox widgets [2]. These parametric interaction techniques work equally well with projections of observations or attributes [21, 61].

7.5 Towards Resolving Semantic Interaction Ambiguity

Semantic interaction aims to improve the quality of user interactions by enabling an analyst to directly manipulate a projection rather than attempt to finesse the parameters of the underlying mathematical model(s) [22, 25, 26]. However, this work demonstrates that the variety of possible meanings and intents of an analyst's interactions can be difficult to capture in a single tool. In other words, interactions such as repositioning an observation are inherently ambiguous; this is the "With Respect to What" usability challenge [57]. Introducing clusters can make some interactions easier by introducing a hard target, but also introduces added ambiguity (e.g., has the analyst moved an observation into a cluster, or was their goal to move the observation closer to some of the observations within the cluster?).

Resolving this ambiguity is critical to the future of semantic interaction. As such, a number of techniques have been introduced to provide feedback to the analyst regarding how the system will interpret their interaction [35]. For example, Figure 2 displays the selection interactions in Andromeda, including nearest neighbor selection, radius selection, and additional observation selection. Pollux also limits the interaction space to reclassifying observations and manipulating their position within clusters [69]. However, limiting the interaction space can prevent analysts from learning more about their data from forbidden interactions.

To truly allow for free-form interactivity and data manipulation in systems, there is an inherent tradeoff between creating complex interactions that are precise but difficult for analysts to remember and perform and creating simple interactions that are ambiguous but easy for analysts. Precise interactions could include components such as a double-click to indicate the importance of the source of the target of the update, multitouch to denote the cardinality of the interaction, and presenting visual feedback to the analyst before the interaction is handled by the system [12, 51, 70]. More ambiguous interactions could learn from a small training set and/or the interaction history to match the intent of a user to the interactions that they perform, such as found in ActiveInk [52]. Such a training set could be generated by an elicitation study, understanding precisely how analysts wish to perform these interactions.

7.6 Complementary HCI and ML Perspectives

This work makes use of terminology from the machine learning community ("metric learning") as well as from the human-computer interaction community ("inferring user intent"). This choice is not accidental or unintentional, as the research presented here exists at the intersection of both fields: we include discussions of both the semi-supervised training of machine learning algorithms and design and interaction considerations for interactive visualization tools. Perhaps the clearest example of such symmetry within this work comes from the overlapping ideas of inferring the intent of a user and mapping that intent to a learned metric in the system.

From the HCI perspective, we provide discussions of design considerations, suggestions for responding to interaction ambiguity, and two representative interactive interfaces. For ML, we towards some of the underlying mathematics, reference a number of implementations of interactive systems that can learn from analyst interactions, and discussion learning issues within this design space. Both of these perspectives address separate but complementary facets of the same problem.

7.7 Limitations

Though we overview the challenges of the "With Respect to What" problem as it pertains to dimension reduction and clustering algorithms, we make no claim regarding the completeness of our survey of interactions. For example, an additional portion of the interaction that could be considered is its speed. Perhaps a quick interaction could be used to indicate a cluster reassignment only, whereas a slower interaction could be interpreted as more carefully positioning the final location of the observation, indicating a positional similarity interaction. Extending to a future extreme, a system could be designed with speech recognition support to permit a user to explain their intent in natural language while performing an interaction. The creativity of visualization designers is nearly boundless, and we fully expect that future designers can extend this work.

7.8 Future Work

This work is focused on a novel space at the intersection of dimension reduction, clustering, and semantic interaction. While a number of systems exist at the intersection of two of these three fields, only a few can be found at the intersection of all three [68, 69]. As a result, some of our discussion of potential interactions is speculative rather than demonstrative. We plan a future elicitation study to determine how analysts will naturally wish to perform these interactions, as well as future implementations that provide such interactions. Semantic interaction is a relatively new field that is still developing, and it presents a number of opportunities for future system and interaction development, including the potential for a generic toolkit or system for semantic interaction.

8 CONCLUSION

This work models the complexity and ambiguity inherent in the interaction space of dimension reduction and clustering algorithms in interactive projections. We framed this discussion in the context of the "With Respect to What" problem, a research challenge in visual analytics identified by Self et al [57]. Through our discussion, we identified several factors necessary to consider for such interactions: thinking in high- or low-dimensional space, interaction with observations or clusters, interaction with source and destination, and cardinality of interaction. We presented a new pipeline representation that incorporates both a projection direction to generate the visualization and an interaction direction to handle the interaction and interpret the intent of the analyst, and we demonstrated the utility of several of these interactions implemented in visual analytics tools. Finally, we discussed additional considerations related to the implementation of such systems, as well as supplementary interactions and visual metaphors that further assist in communication and exploration.

ACKNOWLEDGMENTS

The members of the Virginia Tech InfoVis and BaVA (Bayesian Visual Analytics) groups provided an excellent sounding board for the ideas contained herein. We also thank the reviewers for their insightful comments in improving the presentation of this work.

REFERENCES

[1] B. Alper, N. Riche, G. Ramos, and M. Czerwinski. 2011. Design Study of LineSets, a Novel Set Visualization Technique. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec 2011), 2259–2267. https://doi.org/10.1109/TVCG.2011.186

[2] J. Alsakran, Y. Chen, Y. Zhao, J. Yang, and D. Luo. 2011. STREAMIT: Dynamic visualization and interactive exploration of text streams. In *2011 IEEE Pacific Visualization Symposium*. 131–138. https://doi.org/10.1109/PACIFICVIS.2011.5742382

[3] Saleema Amershi, James Fogarty, and Daniel Weld. 2012. Regroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 21–30.

[4] G. Andrienko, N. Andrienko, S. Rinzivillo, M. Nanni, D. Pedreschi, and F. Giannotti. 2009. Interactive visual clustering of large collections of trajectories. In *2009 IEEE Symposium on Visual Analytics Science and Technology*. 3–10. https://doi.org/10.1109/VAST.2009.5332584

[5] L. Bradel, C. North, L. House, and S. Leman. 2014. Multi-model semantic interaction for text analytics. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 163–172. https://doi.org/10.1109/VAST.2014.7042492

[6] Matthew Brehmer, Michael Sedlmair, Stephen Ingram, and Tamara Munzner. 2014. Visualizing Dimensionally-Reduced Data: Interviews with Analysts and a Characterization of Task Sequences. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV '14)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/2669557.2669559

[7] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. 2012. Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 83–92. https://doi.org/10.1109/VAST.2012.6400486

[8] Andreas Buja, Dianne Cook, and Deborah F Swayne. 1996. Interactive high-dimensional data visualization. *Journal of computational and graphical statistics* 5, 1 (1996), 78–99.

[9] X. Chen, J. Z. Self, L. House, J. Wenskovitch, M. Sun, N. Wycoff, J. R. Evia, S. Leman, and C. North. 2018. Be the Data: Embodied Visual Analytics. *IEEE Transactions on Learning Technologies* 11, 1 (Jan 2018), 81–95. https://doi.org/10.1109/TLT.2017.2757481

[10] Ed H. Chi, Lichan Hong, Michelle Gumbrecht, and Stuart K. Card. 2005. ScentHighlights: Highlighting Conceptually-related Sentences During Reading. In *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '05)*. ACM, New York, NY, USA, 272–274. https://doi.org/10.1145/1040830.1040895

[11] E. H. Chi, L. Hong, J. Heiser, and S. K. Card. 2006. Scentindex: Conceptually Reorganizing Subject Indexes for Reading. In *2006 IEEE Symposium On Visual Analytics Science And Technology*. 159–166. https://doi.org/10.1109/VAST.2006.261418

[12] Ed H Chi, Lichan Hong, Julie Heiser, Stuart K Card, and Michelle Gumbrecht. 2007. ScentIndex and ScentHighlights: productive reading techniques for conceptually reorganizing subject indexes and highlighting passages. *Information Visualization* 6, 1 (2007), 32–47.

[13] J. Choo, C. Lee, C. K. Reddy, and H. Park. 2013. UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec 2013), 1992–2001. https://doi.org/10.1109/TVCG.2013.212

[14] Jason Chuang and Daniel J Hsu. 2014. Human-Centered Interactive Clustering for Data Analysis. *Conference on Neural Information Processing Systems (NIPS). Workshop on Human-Propelled Machine Learning* (2014).

[15] Jason Chuang, Christopher D. Manning, and Jeffrey Heer. 2012. Termite: Visualization Techniques for Assessing Textual Topic Models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI '12)*. ACM, New York, NY, USA, 74–77. https://doi.org/10.1145/2254556.2254572

[16] C. Collins, G. Penn, and S. Carpendale. 2009. Bubble Sets: Revealing Set Relations with Isocontours over Existing Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (Nov 2009), 1009–1016. https://doi.org/10.1109/TVCG.2009.122

[17] Marie Desjardins, James MacGlashan, and Julia Ferraioli. 2007. Interactive visual clustering. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 361–364.

[18] Chris Ding and Xiaofeng He. 2004. K-means Clustering via Principal Component Analysis. In *Proceedings of the Twenty-first International Conference on Machine Learning (ICML '04)*. ACM, New York, NY, USA, 29–. https://doi.org/10.1145/1015330.1015408

[19] Chris Ding and Tao Li. 2007. Adaptive Dimension Reduction Using Discriminant Analysis and K-means Clustering. In *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*. ACM, New York, NY, USA, 521–528. https://doi.org/10.1145/1273496.1273562

[20] E. P. dos Santos Amorim, E. V. Brazil, J. Daniels, P. Joia, L. G. Nonato, and M. C. Sousa. 2012. iLAMP: Exploring high-dimensional spacing through backward multidimensional projection. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 53–62. https://doi.org/10.1109/VAST.2012.6400489

[21] Michelle Dowling, John Wenskovitch, J.T. Fry, Scotland Leman, Leanna House, and Chris North. 2019. SIRIUS: Dual, Symmetric, Interactive Dimension Reductions. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (Jan 2019), 172–182. https://doi.org/10.1109/TVCG.2018.2865047

[22] Michelle Dowling, John Wenskovitch, Peter Hauck, Adam Binford, Nicholas Polys, and Chris North. 2018. A Bidirectional Pipeline for Semantic Interaction. In *Proceedings of the Workshop on Machine Learning from User Interaction for Visualization and Analytics (VIS 2018)*. 11.

[23] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[24] Alex Endert. 2014. Semantic Interaction for Visual Analytics: Toward Coupling Cognition and Computation. *Computer Graphics and Applications, IEEE* 34, 4 (July 2014), 8–15. https://doi.org/10.1109/MCG.2014.73

[25] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2879–2888.

[26] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic Interaction for Visual Text Analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 473–482. https://doi.org/10.1145/2207676.2207741

[27] A. Endert, C. Han, D. Maiti, L. House, S. Leman, and C. North. 2011. Observation-level interaction with statistical models for visual analytics. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 121–130. https://doi.org/10.1109/VAST.2011.6102449

[28] I K Fodor. 2002. *A Survey of Dimension Reduction Techniques*. https://doi.org/10.2172/15002155

[29] S. L. France and J. D. Carroll. 2011. Two-Way Multidimensional Scaling: A Review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 41, 5 (Sept 2011), 644–661. https://doi.org/10.1109/TSMCC.2010.2078502

[30] E. R. Gansner, Y. Hu, and S. Kobourov. 2010. GMap: Visualizing graphs and clusters as maps. In *2010 IEEE Pacific Visualization Symposium (PacificVis)*. 201–208. https://doi.org/10.1109/PACIFICVIS.2010.5429590

[31] Tera Marie Green, William Ribarsky, and Brian Fisher. 2009. Building and applying a human cognition model for visual analytics. *Information visualization* 8, 1 (2009), 1–13.

[32] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).

[33] Rex Hartson and Pardha S Pyla. 2012. *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier.

[34] Leanna House, Scotland Leman, and Chao Han. 2015. Bayesian visual analytics: Bava. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8, 1 (2015), 1–13.

[35] Xinran Hu, Lauren Bradel, Dipayan Maiti, Leanna House, and Chris North. 2013. Semantics of directly manipulating spatializations. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2052–2059.

[36] Paulo Joia, Danilo Coimbra, Jose A Cuminato, Fernando V Paulovich, and Luis G Nonato. 2011. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2563–2571.

[37] E. Kandogan. [n.d.]. Star Coordinate: A Multi-Dimensional Visualization Technique with Uniform Treatment of Dimensions. In *Proceedings of the IEEE Information Visualization Symposium*, Vol. 650. 22.

[38] M. Kim, K. Kang, D. Park, J. Choo, and N. Elmqvist. 2017. TopicLens: Efficient Multi-Level Visual Topic Exploration of Large-Scale Document Collections. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan 2017), 151–160. https://doi.org/10.1109/TVCG.2016.2598445

[39] T. Kohonen. 1990. The self-organizing map. *Proc. IEEE* 78, 9 (Sep 1990), 1464–1480. https://doi.org/10.1109/5.58325

[40] Christoph H Lampert, Hannes Nickisch, Stefan Harmeling, and Jens Weidmann. 2009. Animals with Attributes: A Dataset for Attribute Based Classification.

[41] Hanseung Lee, Jaeyeon Kihm, Jaegul Choo, John Stasko, and Haesun Park. 2012. iVisClustering: An Interactive Visual Document Clustering via Topic Modeling. *Computer Graphics Forum* 31, 3pt3 (2012), 1155–1164. https://doi.org/10.1111/j.1467-8659.2012.03108.x

[42] John A Lee and Michel Verleysen. 2007. *Nonlinear dimensionality reduction*. Springer Science & Business Media.

[43] Scotland C Leman, Leanna House, Dipayan Maiti, Alex Endert, and Chris North. 2013. Visual to parametric interaction (v2pi). *PloS one* 8, 3 (2013), e50474.

[44] S. Liu, D. Maljovec, B. Wang, P. Bremer, and V. Pascucci. 2017. Visualizing High-Dimensional Data: Advances in the Past Decade. *IEEE Transactions on Visualization and Computer Graphics* 23, 3 (March 2017), 1249–1268. https://doi.org/10.1109/TVCG.2016.2640960

[45] Gladys MH Mamani, Francisco M Fatore, Luis Gustavo Nonato, and Fernando Vieira Paulovich. 2013. User-driven Feature Space Transformation. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 291–299.

[46] G. M. H. Mamani, F. M. Fatore, L. G. Nonato, and F. V. Paulovich. 2013. User-driven Feature Space Transformation. *Computer Graphics Forum* 32, 3pt3 (2013), 291–299. https://doi.org/10.1111/cgf.12116

[47] Tauno Metsalu and Jaak Vilo. 2015. ClustVis: a web tool for visualizing clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic acids research* 43, W1 (2015), W566–W570.

[48] Vladimir Molchanov and Lars Linsen. 2014. Interactive Design of Multidimensional Data Projection Layout. In *EuroVis - Short Papers*, N. Elmqvist, M. Hlawitschka, and J. Kennedy (Eds.). The Eurographics Association. https://doi.org/10.2312/eurovisshort.20141152

[49] F.V. Paulovich, D.M. Eler, J. Poco, C.P. Botha, R. Minghim, and L.G. Nonato. 2011. Piecewise Laplacian-based Projection for Interactive Data Exploration and Organization. *Computer Graphics Forum* 30, 3 (2011), 1091–1100. https://doi.org/10.1111/j.1467-8659.2011.01958.x

[50] Fernando Vieira Paulovich, Danilo Medeiros Eler, Jorge Poco, Charl P Botha, Rosane Minghim, and Luis Gustavo Nonato. 2011. Piece wise laplacian-based projection for interactive data exploration and organization. In *Computer Graphics Forum*, Vol. 30. Wiley Online Library, 1091–1100.

[51] Peter Pirolli. 2003. A theory of information scent. *Human-computer interaction* 1 (2003), 213–217.

[52] Hugo Romat, Nathalie Henry Riche, Ken Hinckley, Bongshin Lee, Caroline Appert, Emmanuel Pietriga, and Christopher Collins. 2019. ActiveInk: (Th)Inking with Data. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 42, 13 pages. https://doi.org/10.1145/3290605.3300272

[53] Tuukka Ruotsalo, Jaakko Peltonen, Manuel Eugster, Dorota Głowacka, Ksenia Konyushkova, Kumaripaba Athukorala, Ilkka Kosunen, Aki Reijonen, Petri Myllymäki, Giulio Jacucci, et al. 2013. Directing exploratory search with interactive intent modeling. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*. ACM, 1759–1764.

[54] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim. 2017. Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (Jan 2017), 241–250. https://doi.org/10.1109/TVCG.2016.2598495

[55] Jessica Zeitz Self, Michelle Dowling, John Wenskovitch, Ian Crandell, Ming Wang, Leanna House, Scotland Leman, and Chris North. 2018. Observation-Level and Parametric Interaction for High-Dimensional Data Analysis. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2, Article 15 (June 2018), 36 pages. https://doi.org/10.1145/3158230

[56] Jessica Zeitz Self, Xinran Hu, Leanna House, Scotland Leman, and Chris North. 2016. Designing Usable Interactive Visual Analytics Tools for Dimension Reduction. In *CHI 2016 Workshop on Human-Centered Machine Learning (HCML)*. 7.

[57] Jessica Zeitz Self, Radha Krishnan Vinayagam, J. T. Fry, and Chris North. 2016. Bridging the Gap Between User Intention and Model Parameters for Human-in-the-loop Data Analytics. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics (HILDA '16)*. ACM, New York, NY, USA, Article 3, 6 pages. https://doi.org/10.1145/2939502.2939505

[58] John Sharko, Georges Grinstein, and Kenneth A Marx. 2008. Vectorized radviz and its application to multiple cluster datasets. *IEEE transactions on Visualization and Computer Graphics* 14, 6 (2008).

[59] Frank M Shipman and Catherine C Marshall. 1999. Formality considered harmful: Experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work (CSCW)* 8, 4 (1999), 333–352.

[60] Gabor J Szekely and Maria L Rizzo. 2005. Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method. *Journal of classification* 22, 2 (2005), 151–183.

[61] C. Turkay, P. Filzmoser, and H. Hauser. 2011. Brushing Dimensions - A Dual Visual Analysis Model for High-Dimensional Data. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec 2011), 2591–2599. https://doi.org/10.1109/TVCG.2011.178

[62] United States Census Bureau. 2016. Census Regions and Divisions of the United States. http://www2.census.gov/geo/pdfs/maps-data/maps/reference/us$_{r}$egdiv.pdf

[63] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. 2009. Dimensionality reduction: a comparative. *J Mach Learn Res* 10, 66-71 (2009), 13.

[64] Tatiana von Landesberger, Sebastian Fiebig, Sebastian Bremm, Arjan Kuijper, and Dieter W Fellner. 2014. Interaction taxonomy for tracking of user actions in visual analytics applications. In *Handbook of Human Centric Visualization*. Springer, 653–670.

[65] Emily Wall, Subhajit Das, Ravish Chawla, Bharath Kalidindi, Eli T Brown, and Alex Endert. 2018. Podium: Ranking data using mixed-initiative visual analytics. *IEEE transactions on visualization and computer graphics* 24, 1 (2018), 288–297.

[66] J. Wenskovitch, L. Bradel, M. Dowling, L. House, and C. North. 2018. The Effect of Semantic Interaction on Foraging in Text Analysis. In *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 13–24. https://doi.org/10.1109/VAST.2018.8802424

[67] John Wenskovitch, Ian Crandell, Naren Ramakrishnan, Leanna House, Scotland Leman, and Chris North. 2018. Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics Proceedings of the Visual Analytics Science and Technology 2017* 24, 01 (January 2018).

[68] John Wenskovitch and Chris North. 2017. Observation-Level Interaction with Clustering and Dimension Reduction Algorithms. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics (HILDA'17)*. ACM, New York, NY, USA, Article 14, 6 pages. https://doi.org/10.1145/3077257.3077259

[69] John Wenskovitch and Chris North. 2019. Pollux: Interactive Cluster-First Projections of High-Dimensional Data. In *2019 Symposium on Visualization in Data Science (VIS 2019)*. 9.

[70] W. Willett, J. Heer, and M. Agrawala. 2007. Scented Widgets: Improving Navigation Cues with Embedded Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov 2007), 1129–1136. https://doi.org/10.1109/TVCG.2007.70589

[71] Axel Wismüller, Michel Verleysen, Michael Aupetit, and John Aldo Lee. 2010. Recent Advances in Nonlinear Dimensionality Reduction, Manifold and Topological Learning.. In *ESANN*.

[72] Rui Xu and D. Wunsch. 2005. Survey of clustering algorithms. *IEEE Transactions on Neural Networks* 16, 3 (May 2005), 645–678. https://doi.org/10.1109/TNN.2005.845141

[73] J. S. Yi, Y. a. Kang, and J. Stasko. 2007. Toward a Deeper Understanding of the Role of Interaction in Information Visualization. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov 2007), 1224–1231. https://doi.org/10.1109/TVCG.2007.70515