Choosing Visualization Techniques for Multidimensional Data Projection Tasks: A Guideline with Examples

Ronak Etemadpour[1], Lars Linsen[2], Jose Gustavo Paiva[3], Christopher Crick[1], and Angus Graeme Forbes[4]

[1] Oklahoma State University, Stillwater, OK, USA.
[2] Jacobs University, Bremen, Germany.
[3] Federal University of Uberlandia, Uberlandia/MG, Brazil.
[4] University of Illinois at Chicago, Chicago, IL, USA.

Abstract. This paper presents a guideline for visualization designers who want to choose appropriate techniques for enhancing tasks involving multidimensional projection. Specifically, we adopt a user-centric approach in which we take user perception into consideration. Here, we focus on projection techniques that output 2D or 3D scatterplots that can then be used for a range of common data analysis tasks, which we categorize as pattern identification tasks, relation-seeking tasks, membership disambiguation tasks, or behavior comparison tasks. Our user-centric task categorization can be used to effectively guide the organization of multidimensional data projection layouts. Moreover, we present real-world examples that demonstrate effective choices made by visualization designers faced with complex datasets requiring dimensionality reduction.

Keywords: Multidimensional data analysis, task taxonomy, multidimensional data projection, user-centric evaluation.

1 Introduction

Visualization is a crucial step in the process of data analysis. Often, when analyzing multidimensional data, dimensionality reduction (DR) techniques are displayed in form of 2D or 3D scatterplots that project the multidimensional points onto a lower-dimensional visual space. Methods using different algorithms to generate scatterplots with particular point placements are the most common visual encoding (VE) techniques for the resulting lower-dimensional data. DR techniques, coupled with appropriate VEs, enable an understanding of the relations that exist within the higher-dimensional data by displaying them in such a way that makes it easier for users to discover meaningful patterns [36].

Data analysis tasks are primarily concerned with the detection of structures such as patterns, groups, and outliers. Within a multidimensional data set, data points can be grouped manually into classes or automatically into clusters. For

example, classes may be defined through manually labeling a collection of documents so that each document belongs to one topic within a set of topics, or by splitting an image collection into ten classes by assigning each image a particular theme from a set of ten themes. Clusters, on the other hand, are generated automatically using a clustering algorithm that may, for instance, identify groupings of similar points, or partition the data into dissimilar groups where each cluster contains similar items [25]. However, it may be difficult to see these clusters or classes when projected onto a lower-dimensional space. To make sense of this multidimensional data, it can be useful to know how the clusters or classes are defined and structured in the original multidimensional attribute space. However, multidimensional projection mappings are especially prone to distortion because projection methods may not necessarily preserve the spatial relations of the data. Thus, it is important to know how effective the scatterplots are at preserving segregation of the data [42]. Several studies evaluate the quality of projections with respect to preserving certain properties, thus guiding a user to select the most appropriate projection method for their task. Various numerical and visual methods have been introduced to quantify the accuracy of projection methods with respect to such properties [42, 46]. Recent studies [41] have shown that the quality of cluster separation by these measures was highly discrepant with user assessment of the cluster separation within the same data sets. Lewis et al. [24] believe that accurate evaluation of clustering quality is essential for data analysts, and they showed that such clustering evaluation skills are present in the general population. On the other hand, other studies have attempted to find a perception-based quality measure for scatterplots. They either evaluated users' performance on layouts generated by different projection techniques [14] and used eye-tracking while users are asked to perform typical analysis tasks for projected multidimensional data or allowed users to assess a series of scatterplots [2]. Other studies have investigated the perception of correlation in scatterplots from a psychological perspective; however these studies did not consider real-world data sets [34].

Because of the absence of a standard approach for evaluating multidimensional data projection, the results of these studies, and others like them, are difficult to compare. We present a taxonomy of visual analysis tasks for multidimensional data projection that we believe could be a useful means for evaluation. The idea of creating a task taxonomy has been recently explored by Brehmer and Munzner [7]. They contribute a multi-level typology of visualization tasks that augments existing taxonomies by filling a gap between low-level and high-level tasks. Specifically, they distinguish what the task inputs and outputs are, as well as why and how a visualization task is performed. In doing so, they more thoroughly organize the motivations for and methods of specific tasks for particular data analysis situations. Their task taxonomy is more general, and does not address multidimensional data projection in any detail. In this paper, we provide a taxonomy of visual analysis tasks related to multidimensional data projection. Our task taxonomy enables evaluation designers to investigate visu-

alization performance effectively on both synthetic and real-world data sets. The main contributions of the paper are:

- We provide a systematic user-centric taxonomy of visual tasks related to projected multidimensional data.
- We divide the projection-related tasks into different categories based on their impact on the analysis of multidimensional data. The categories we identify are relation-seeking, behavior comparison, membership disambiguation, and pattern identification tasks.
- We enable, via our task taxonomy, visualization designers to improve visualization tasks related to the analysis of multidimensional data.
- We present our taxonomy as a guideline for researchers in choosing visualization techniques for these tasks, and provide explicit examples.
- We adapt multilevel typology of abstract visualizations to multidimensional data projection tasks [7].

In the next section, we provide a brief review of existing task taxonomies for DR and VE techniques. In Section 3, we introduce our task taxonomy for multidimensional data projection by describing new sets of tasks related to typical analysis tasks, including *pattern identification*, such as detecting clusters, *behavior comparison*, such as comparing characteristics of subsets, *membership disambiguation*, such as counting the number of objects in a cluster, and *relation seeking*, such as correlating subsets to each other. We discuss the effects of our proposed tasks on the evaluation of scatterplots by providing some examples of how different tasks support decision making respective to human perception over multidimensional data projections. We also characterize our proposed tasks using the multi-level typology of abstract visualization tasks [7]. We applied Brehmer and Munzner's multi-level topology concept for describing two tasks as guidelines, while the three questions (WHY, WHAT, HOW) can be used to structure the description of all tasks.

2 Related Work

Many projection methods exist to generate 2D similarity-based layouts from a higher-dimensional space. The design goals include maintaining pairwise distances between points [6] as implemented in multidimensional scaling (MDS), maintaining distances within a cluster, or maintaining distances between clusters [47]. Principal component analysis (PCA) generates similarity layouts by reducing data to lower dimensional visual spaces [22]. Some projection methods, such as isometric feature mapping (Isomap), favor maintaining distances between clusters instead. Isomap is an MDS approach that has been introduced as an alternative to classical scaling capable of handling non-linear data sets. It replaces the original distances by geodesic distances computed on a graph to obtain a globally optimal solution to the distance preservation problem [47]. Least-Square Projection (LSP) computes an approximation of the coordinates of a set of projected points based on the coordinates of some samples as control

points. This subset of points is representative of the data distribution in the input space. LSP projects them to the target space with a precise MDS force-placement technique. It then builds a linear system from information given by the projected points and their neighborhoods [31]. The correlations of data points or clusters are not always known after they have been mapped from a higher-dimensional data space to 2D or 3D display space. Thus, several approaches evaluate the best views of multidimensional data sets. Sips et al. [42] provide measures for ranking scatterplots with classified and unclassified data. They propose two additional quantitative measures on class consistency: one based on the distance to the cluster centroids, and another based on the entropies of the spatial distributions of classes. They propose class consistency as a measure for choosing good views of a class structure in high-dimensional space. Tan et al. [44], Paulovich et al. [31], and Geng et al. [18] also evaluate the quality of layouts numerically. By ranking the perceptual complexity of the scatterplots, other studies investigate user perception by conducting user studies on scatterplots, finding that certain arrangements were more pleasing to most users [45]. However, these operational measures were not necessarily equivalent to the measures of user preference based on their qualitative perceptions. Sedlmair et al. [40] have discussed the influence of factors such as scale, point distance, shape, and position within and between clusters in qualitative evaluation of DR techniques. They examined over 800 plots in order to create a detailed taxonomy of factors to guide the design and the evaluation of cluster separation measures. They focused only on using scatterplot visualizations for cluster finding and verification. DimStiller [20] is a system to provide global guidance for navigating a data-table space through the process of choosing DR and VE techniques. This analysis tool captures useful analysis patterns for analysts who must deal with messy data sets. Rensink and Baldridge [34] explore the use of simple properties such as brightness to generate a set of scatterplots in order to test whether observers could discriminate pairs using these properties. They found that perception of correlations in a scatterplot is rapid, and that in order to limit visual attention to specific information it is more effective to group features together. Etemadpour et al. [17] postulate that cluster properties such as density, shape, orientation, and size influence perception when interpreting distances in scatterplots, and specifically, observe that the density of clusters is more influential than their size.

In general, little attention has been paid to providing details about low-level tasks that guide users to choose DR and VE techniques. However, both high-level goals and much more specific low-level tasks are important aspects of analytic activities. Amar et al. [3] presented a set of ten low-level analysis tasks that they found to be representative of questions that are needed to effectively facilitate analytic activity. Andrienko and Andrienko distinguish elementary tasks that address specific elements of a set and synoptic tasks that address entire sets or subsets, according to the level of analysis [4].

Brehmer and Munzer [7] emphasize three main questions, *why* the tasks are performed, *how* they are performed, and *what* are their inputs and outputs.

These questions encompass their concept of multi-level typology. They believe that "low-level characterization does not describe the user's context or motivation; nor does it take into account prior experience and background knowledge." Their typology relies on a more abstract categorization based on concepts, rather than a taxonomy of pre-existing objects or tasks. In contrast, we attempt to specify tasks at the lowest level that can provide details about multidimensional data projection. However, the general approach of Brehmer and Munzner can be easily adopted as a tool to put these low-level tasks in context, facilitating the evaluation of user experiences by evaluation designers. This approach provides essential information, such as motivation and user expertise, for field studies that examine visualization usage. Therefore, we show how our defined tasks can be described according to a typology of abstract tasks relating intents and techniques (how) to modes of goals and tasks (why).

We 1) categorize possible tasks performed when analyzing a specific multidimensional data visualization, and 2) formulate guidelines for analysts to assist in selecting appropriate projection techniques for performing specific visualization tasks on data sets.

3 Task Taxonomy for Multidimensional Data Projection

We define a list of tasks from studies of different projection techniques and their 2D layouts such as PCA [22], Isomap [47], LSP [31], Glimmer [21], and NJ tree [29], as well as the applications behind the data (e.g. document and image data). We explain some of these tasks in detail and provide examples of effective data representations for relevant visual analysis tasks. As explained in Section 2, how well groups of points can be distinguished by users in scatterplots defines visual class separability. Our cluster-level tasks also focus on how easily a grouping of related points in multidimensional space (e.g., clusters) can be detected by users when projected into lower-dimensional space. However, rather than only looking at visual class separability, we consider how effective users are performing meaningful tasks related to the perceived clusters.

Although other researchers have explored some of these tasks, we systematically list the full range of analytic tasks for multidimensional projection techniques appropriate for large data sets. Additionally, our organization of these tasks takes into consideration user perception. We divided the tasks into four categories according to the typical visualizations required to support them:

Pattern identification tasks: We examine trends, which are more obvious for lower-dimensional data than for projected higher-dimensional ones. Relevant issues include cluster/class preservation and separation.

Relation-seeking tasks: Relationships and similarities between different reference sets are considered.

Behavior comparison tasks: To compare characteristics of subsets (or clusters), we consider capturing different data behaviors (like asking the subjects to compare the point densities within clusters, where density is defined as the number of points per area).

Membership disambiguation tasks: Positional and distributional relationships within classes/clusters are particularly considered where objects occlude each other. Clutter and noise obscure the structure present in the data and make it hard for users to find patterns and relationships. Peng et al. [32] state that clutter reduction is a visualization-dependent task. Therefore, the DR and VE need to minimize the amount of confusing clutter. We believe that clutter can be measured by users using a wide variety of visualization techniques.

We now clarify these taxonomic categories by looking at common tasks found in the literature.

3.1 Pattern identification tasks

Multidimensional data sets may include hundreds or thousands of objects described by dozens or hundreds of attributes. Data characteristics regarding the distribution within multidimensional feature spaces vary for different application domains. For example, consider document data versus image data: text usually produces sparse spaces while imagery produces dense spaces. As Song et al. [43] state, traditional document representation like bag-of-words leads to sparse feature spaces with high dimensionality. This makes it difficult to achieve high classification accuracies. Figure 1 shows histograms of the distribution of the pairwise distances between four data objects after normalization to the interval [0,1]. The document data sets are referred to as CBR and KDViz[1]. The image data sets are referred to as Corel[2] and Medical[3]. The revealed histograms illustrate different characteristics for document data sets and image data sets. Both image data sets exhibit lower mean distance values and much wider variance (representative of a denser feature space) than the document data sets.

Identifying patterns in high-dimensional spaces and representing them using dimensionality reduction techniques, in order to reveal trends, is a challenge in many scientific and commercial applications. To identify outliers, trends and interesting patterns in data, one of the many objectives of data exploration is to find correlations in the data, thus uncovering hidden relationships in the data distribution and providing additional insights about the high-dimensional data [53]. Therefore, a list of questions are suggested that can reveal user's

[1] CBR comprises 680 documents, which include title, authors, abstract, and references from scientific papers in the four different subjects, leading to a data set with 680 objects and 1,423 dimensions. KDViz data has been generated from an Internet repository on the topics bibliographic coupling, co-citation analysis, milgrams, and information visualization, leading to 1,624 objects, 520 dimensions, and four highly unbalanced labels (http://vicg.icmc.usp.br/infovis2/DataSets).

[2] 1,000 photographs on ten different themes. Each image is represented by a 150-dimensional vector of SIFT descriptors (3UCI KDD Archive, http://kdd.ics.uci.edu).

[3] Each image is represented by 28 features, including Fourier descriptors and energies derived from histograms, as well as mean intensity and standard deviation computed from the images themselves. Hence, the data set contains 540 objects and 28 dimensions

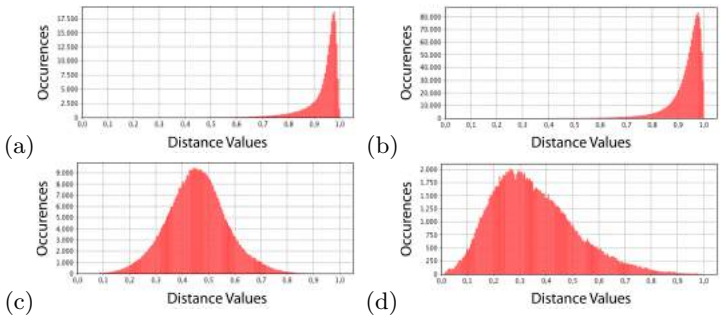

Fig. 1. Histograms of document data (top) and image data (bottom) exhibit characteristic distance distributions: (a) CBR. (b) KDViz. (c) Corel. (d) Medical.

perspective about local and global correlations with respect to features – for instance, those subsets of data which form relevant patterns (e.g. subsets of data within dense feature groups): 1) Estimate the number of outliers in the given layout; 2) Estimate the number of observed clusters; 3) Find the number of clusters in a selected region; 4) Find the number of subclusters in a given cluster; 5) Find a cluster with a specific characteristic (e.g., longish); 6) Find the specific characteristics (e.g., sparsity) of a cluster; 7) Determine the number of outliers in a given cluster.

If researchers aim to find the user's performance on class segregation, it is important to draw the user's attention to global project views. Thus, we suggest asking *Estimate the number of clusters in the given layout* to identify the informative aspects of the data.

Pattern identification tasks often favor clear segregation by class, which means that techniques which incorporate cluster enclosing surfaces can be helpful. In some situations, the labeled classes in each data set can be considered as ground truth. For such cases, Poco et al. [33] developed a 3D projection method by generalizing the LSP technique from a 2D to a 3D scheme. A non-convex hull (of each cluster) that is computed from a 3D Voronoi diagram of the cluster points is illustrated in Figure 4(a). This representation, when it is possible to construct, is both accurate and satisfying to users, compared to other techniques.

For situations in which a small set of representative instances from each class is available, or can be manually labeled from a large data set, Paiva et al [30] proposed a semi-supervised dimensionality reduction approach that employs the Partial Least Squares (PLS) [52] technique, producing reduced spaces that favors class segregation. PLS models relations between sets of variables by estimating a low dimensional latent space, that maximizes the separation between instances with different characteristics, resulting in a low dimensional latent space in which instances from the same class are clustered. The proposed methodology employs visualization techniques to show the similarity structure of the collection, in order to guide the user in selecting representative instances to train the PLS model, that can then be applied to a much larger data set very effectively. Figure 2

shows the LSP projection of Corel data set, with the original dimensionality (a) and after a PLS reduction to 10 dimensions (b). One can notice that the groups are more dense on the reduced space, highlighting the class separability. The methodology can also be used for situations in which the instances labels are not available. In this case, a clustering procedure is performed, and the cluster labels are then used to produce a PLS model. For data sets whose cluster structure reflects the class distribution, this methodology will create a reduced space that will favor class segregation.

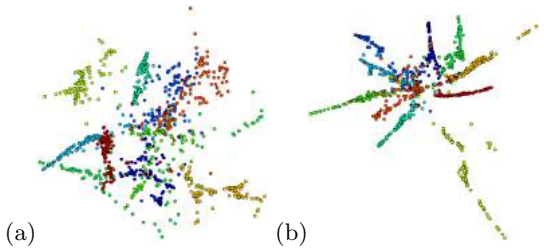

Fig. 2. Layouts for Corel data set obtained using LSP projection, using all 150 original attributes and using PLS reduced 10 attributes.

Also for situations in which a labeled instances set is available, Paiva et al. [28] proposed a visual classification methodology (VCM) that integrates point-based visualization techniques and automatic classification procedures to support control over the whole classification process by users. It yields visual support to classify evolving data sets by allowing user interference, via similarity based visualizations, during supervised classification in an integrated form, promoting users control over model building, application, evaluation and evolution. User insertion is made by the selection of instances to create a classification model, and this selection is performed using the layout, whose structure and point organization is able to guide the user towards a relevant selection. The created classification model can then be employed in the classification of any collection bearing the same feature space. Similarity layouts may represent, in these scenarios, a potential tool to explore the structure and relationship among instances and thus identify the representative ones of each class. That can be achieved, e.g., by analyzing class segregation or by determining outliers that could distort the classifier behavior. Additionally, the methodology allows, in situations in which a ground truth exists, a visual inspection of the classification results using the same visual strategy, in a tool named *Class Matching*, which provides an understanding of the classifiers behavior, and how the data set structure influence this behavior. Finally, model updates can be performed by selecting additional instances from a visualization layout, that offers the possibility of several model updating strategies. Figure 3 shows three layouts, using a NJ tree, of a subset

of the ETHZ[4] data set, containing 1,739 instances, with (a) representing the ground truth, (b) the result of a SVM classification on this data set, and (c) the corresponding class matching tree, exhibiting in red the misclassified instances. The training set used to build the SVM model contains 280 equally distributed instances. The layout provides several clues about the structure of this collection, as well as about the classifier behavior. Looking at (a), one can notice that the branches are usually homogeneous in terms of classes, as indicated by the circles colors. However, in (b) it is possible to see some heterogeneous branches, which coincide with most of the misclassified instances, indicating that the classifier is confuse about these instances. Moreover, it is possible to notice that class 6 instances are spread in four branches, which may indicate that this class is highly heterogeneous. The data set is originally unbalanced, and class 6 contains the highest number of instances, which may also cause instances from other classes to be classified as 6. By analyzing the confusion matrix, it is possible to notice that several instances from class 26 were classified as class 21 or 6. The layout shows that instances of these classes are positioned on the same branch, and it is possible that they share common attribute values, with similar content. The layout instances positions, allied with an adequate color coding, may facilitate the comprehension of the reasons by which the classifier took these decisions, as well as to indicate for which classes the classifier is deficient. Thus, users are capable to perform effective updates to refine the classification results.

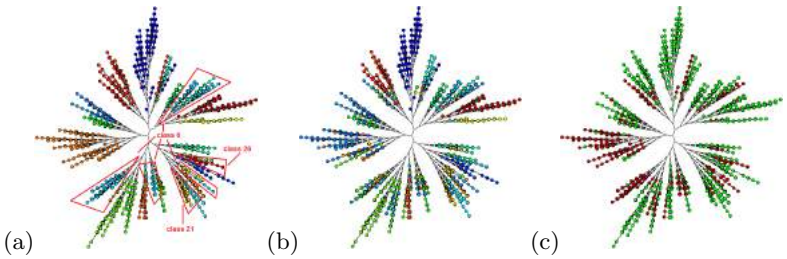

Fig. 3. NJ trees for ETHZ data set, showing (a) the ground truth ,(b) the results of a SVM classification, and (c) the corresponding class matching layout.

While this projection works well when the data's pre-assigned class structure accurately models the data's inherent organization, this is often not feasible. In many situations, analysts want to leverage human perception to identify "visual groupings" of points, and in this case a point cloud representation produces favorable results. For example, when grouping information is not available, a

[4] ETHZ represents a subset of the ETHZ dataset [13, 38], with 2019 photographs of different people captured in uncontrolled conditions. It is divided into 28 unbalanced groups, and each image is represented by a vector of 3963 descriptors, combining Gabor filters, Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP) and mean intensity.

point-based visualization as shown in Figure 4(b) is still applicable. Also, Glimmer [21], as a technique representative of force-directed placement MDS, does not favor class segregation when employed on the KDViz data set.Thus, color coding to separate nodes of different classes can be useful as shown in Figure 4(c). Therefore, if we have accurate class labels and good class separation, we suggest enclosing surfaces like nonconvex hulls. According to the eye-tracking study on Glimmer projection, the visual attention pattern is scattered and it is hard to identify any meaningful area of interest (AOIs) for Glimmer [17]. Hence, it is useful to differentiate classes when the projection doesn't reflect the class distribution at all.

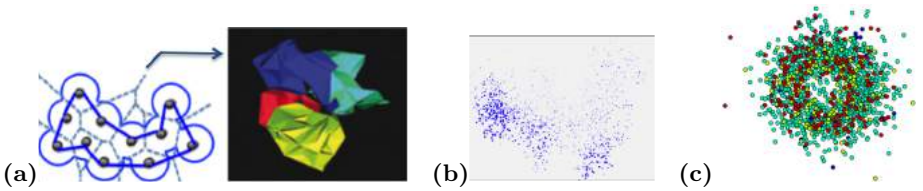

Fig. 4. Estimate the number of observed clusters: (a) Non-convex hulls computed from enclosing surfaces isodistant to cluster using LSP projection; (b) Point-based visualization using PCA projection taken from [37]; (c) The layout obtained with Glimmer projection on the KDViz data set. Circle color indicates instance class label.

3.2 Relation-seeking tasks

Relation-seeking tasks investigate the similarities and differences between subgroups which represent clusters or individual objects. Similarity layouts employ projection techniques to reducing data to lower-dimensional visual spaces, but in a different manner from that used in pattern identification. In this application, an analyst is interested in investigating whether a point (or object) is more similar to one cluster or to another, or whether a whole cluster is more similar to a second cluster or a third. We believe that relationship-seeking is a search task, Andrienko's visual task taxonomy model notwithstanding (in which search tasks are limited to lookup and comparison) [5]. In contrast, Zhang et al. [54] consider comparison and relationship-seeking to be compound tasks, containing at least two relationships, one being the data function and the other being relationships between values (or value sets) of a variable. Under this definition, we believe that finding similarities in projected high-dimensional data can be considered as a relation-seeking tasks. Users perform comparison tasks with respect to a given reference set, which can be a cluster or an individual object, and can undertake a similarity search by identifying a given cluster's neighbors. In such a search, the specified relationship is defined by a distance search within a high-dimensional data projection.

A list of potential tasks within the relation-seeking task category can be considered for multidimensional data visualization: 1) Identify the closest cluster to a given cluster; 2) Identify the most similar cluster to a given cluster; 3) Identify the closest cluster to a reference point; 4) Identify the most similar cluster to a given object; 5) Find k closest (most similar) clusters to the given cluster; 5) Find k closest (most similar) objects to the given cluster; 6) Find k closest (most similar) objects to the reference object; 7) Find the closest (most similar) cluster to a cluster with a specific characteristic (e.g., Find the closest cluster to the longish cluster); 8) Identify the cluster to which the reference set/sets belong; 9) Find the closest (most similar) cluster to the set of points with specific characteristics (e.g., points that have identical movement); 10) Find k closest (most similar) points to the set of points with specific characteristics; 11) Find the clusters that have hierarchical relations; 12) Find k similar objects within a cluster; 13) Find a cluster that is the parent of two reference sets.

Etemadpour et al. [15] investigated how domain-specific issues affect the outcome of the projection techniques. They used a number of similarity interpretation tasks to assess the layouts generated by projection techniques as perceived by their users. To show that projection performance is task-dependent, they generated layouts of high-dimensional data with five techniques representative of different projection approaches. To find a perception-based quality measure, they asked individuals to identify the closest cluster to a given cluster and object. Users also ranked the k nearest objects to a given object. As shown in Figure 5, the target cluster/object was shown in one color (red) and two other clusters in other colors (green and blue), from which the one closer to the target cluster/object should be identified.

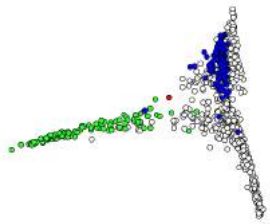

Fig. 5. Task: determine whether green or blue cluster is closer to red object in order to investigate PCA projection performance.

Node-link diagrams have been studied in detail in many graph drawing topics or graph visualization approaches, where a node is representing an entity that is connected to other nodes through lines (i.e., links). Although the node-link diagram is an intuitive way to visually represent relationships between entities for relatively small data sets [19], there may be too many lines crossing with each other that obscure relationships among entities when dealing with larger

data sets. In order to represent spatial distance visually in cases like these, a technique like the Force-Directed Placement approach [12] can be used to reveal connections and similarity magnitude between entities. This technique relies on iterative algorithms that model the data points as a system of particles attached to each other by springs. The length of the spring connecting two particles is given by the distance between their corresponding data points as shown in Figure 6. A spatial embedding is obtained with an iterative simulation of the spring forces acting on this hypothetical physical system, until it reaches an equilibrium state.

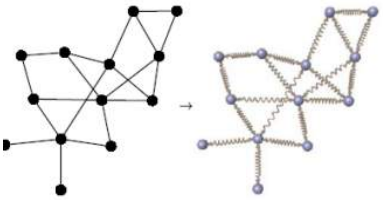

Fig. 6. The spring embedder model [11].

To *Find k closest objects to the reference object*, if the performance of a projection in terms of maintaining distances within a cluster is under investigation and the cluster structure is known, a combination of hull-based and point-based visualizations can be used. Schreck et al. [37] implemented an interactive system that combined these two visual presentations letting users choose the best visual representation of the projected data. They believed that such combined representations introduce visual redundancy; however, it can improve user's perception of the projection precision information depending on the application. Poco et al. [33] improved the performance of their 3D point representation when they combined standard point clouds with this user-guided process. Figure 7 demonstrates finding 3 closest objects to the red object within a cluster when the convex hull of the points is used.

Brehmer and Munzner's typology is intended to facilitate understanding of users' individual analytical strategies. We employ their multi-level code, used to label user behaviour, to enhance the evaluation of high-dimensional data projection. By utilizing the Brehmer and Munzner multi-level typology, we provide a systematic way of justifying the choice of a particular task through asking three main questions: Why, What and How. This multi-level typology of abstract visualization tasks fills the gap between low-level and high-level classification to describe user tasks in a useful way. This approach to analyzing visualization usage supports making precise comparisons of tasks between different visualization tools and across application domains [7]. For an effective design and evaluation of multidimensional data visualization tools, one should consider why and how our defined tasks should be conducted, and what are their potential inputs and

outputs. Meanwhile, sequences of tasks can be linked, so that the output of one task may serve as input to a subsequent task. We focused on *Find k closest clusters to the given cluster* in the relation-seeking category. We did not consider any specific projection technique because it can be changed based on the evaluator's motivation.

Find k closest cluster to the given cluster: **WHY:** The goal is to *Discover* k groups that are closest to a given cluster. A known target (given cluster) and the whole projection visualization are provided. If the location of a given cluster was known (or given by the examiner), then participants perform a *Lookup*. If the characteristic of the given cluster was given, the user can *Locate* the given cluster with specific characteristics (e.g., searching for a given cluster in which the elements are colored red). Then individuals search for k clusters that are in the neighborhood of the given cluster and list these groups. **WHAT:** The input for this task is a given cluster; this can be shown by the examiner or can be indicated by a particular characteristic like the color red. All other clusters in the entire visualization are also visible to the participants. The output is a list of k groups that are closest to the given cluster. **HOW:** Participants identify the k closest clusters to the given cluster. For example, they determine whether the green or blue cluster is closer to the red cluster. They provide a list of clusters that follow an ascending order, so that the distance of the first cluster in this list to the given cluster is shortest compared to the other clusters. *Select* refers to differentiating selected elements from the unselected remainder.

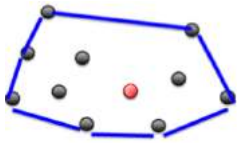

Fig. 7. Find 3 closest objects to the red object: Convex-hull of the point clusters.

Trees are a natural form for depicting hierarchical relations and can be used to *Find the clusters that have hierarchical relations*. A distinct category of 2D mapping employs tree layouts to convey similarity levels contained in a distance matrix. The algorithms to generate similarity layouts [9] are inspired by the well-known Neighbor-Joining (NJ) heuristic originally proposed to reconstruct phylogenetic trees. Similar points among members of the same subsets are placed at the ends of branches. The points nearer the root of the tree are less similar when compared with the points at the ends of branches. Similarity trees generate a hierarchy, creating a tree structure where interpretation is subject to organization of the branches; for example, mapping data setswith the NJ and LSP projections are compared in Figure 8. In this example, the INFOVIS04 data set is composed of documents published in a conference on information visualization, and its content is homogeneous. Using NJ, documents with a high degree of similarity are placed along the same branch. The branches circled in

the figure are examples of long branches without too many ramifications, and probably represent specific sub-topics inside the collection. LSP, on the other hand, has a tendency to create clusters in round clumps. This representation performs well for certain tasks, but is less useful for finding the closest clusters to selected objects [15].

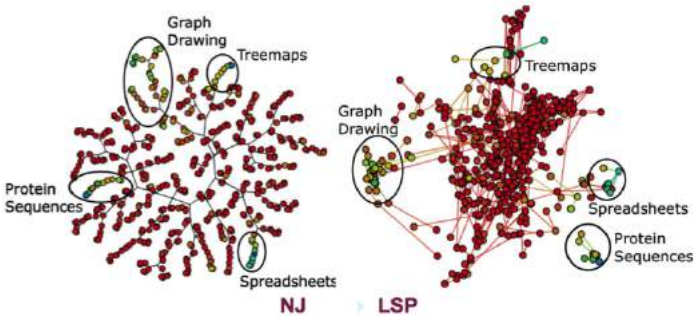

Fig. 8. Comparison of INFOVIS04 document data set map using Neighbor Joining and LSP projections: Four different topics of information visualization are identified by coloring points. Figure is taken from [9].

Authors in [8] introduced BubbleSets as a visualization technique for data that makes explicit use of grouping and clustering information. Members of the same set are in continuous and concave isocontour, while a primary semantic data relation is maintained with spatial organization. These delineated contours do not disrupt the primary layout, so they avoid layout adjustment techniques. This visualization technique is designed in order to facilitate depicting more than one data relationship in data sets that contain multiple relationships. Using this concept, we suggest contours around nodes belonging to the same set to *Find k similar objects within a cluster* in a projection technique. Figure 9 shows an example that uses the BubbleSets concept for an NJ heuristic projection. The points that are sharing the same contour are members of the same set. These boundaries are used to indicate the grouping clearly.

3.3 Behavior comparison tasks

A third way in which high-dimensional data projections can display data items in lower-dimensional subspaces can provide insight into important data dimensions and details. Our taxonomy distinguishes the subsets of tasks used for behavior comparison: 1) Find the cluster with the largest (smallest) occupied visual area; 2) Find the cluster with the most (least) number of points or size; 3) Find densest (sparsest) cluster; 4) Given specific number of clusters (e.g. 5 clusters is given); 5) Rank the clusters by density; 6) Rank the clusters by their occupied visual area; 7) Rank the clusters by their size; 8) Compare density of two given clusters

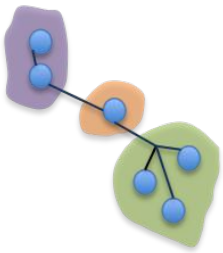

Fig. 9. NJ projection: geometric relationships, hierarchy and cluster perimeter are all clearly defined using BubbleSets concept.

with different or similar characteristics (e.g., density of a longish cluster vs. a roundish cluster); 9) Compare the size of two given clusters with different or similar characteristics; 10) Compare the visual area of two given clusters with different or similar characteristics.

Density is an important metric because it indicates stronger relationships between points within a cluster. Moreover, many studies [1, 39, 49] have indicated that representations of density can play an important role in visualization. Further, studies in psychophysics have shown that visual search can be affected by the variance in the number of objects within groups [10, 35, 48]. Authors in [41] named density as one of the Within-Cluster factors, namely, the ratio between count and size. This can range from sparse, with few data points and a large spread, to dense, with many points and a small spread. If the task is to *Compare density of two given clusters with different or similar characteristics* (i.e. different shapes), we suggest a point-based visualization. This allows users to easily see the point distribution within a cluster and the occupied visual space. Moreover, as investigated in [17], according to the Gestalt principle [23], the shape and orientation of a cluster should also influence decisions during visual analysis. For example, when two stretched clusters are aligned, they may be perceived as a continuation of one cluster or in other words, characteristics of the clusters influence the visual analysis from a perceptual view. Following these ideas, continuity and closure create the perception of a whole cluster. Figure 10 illustrates the density of a longish cluster versus a cluster that looks more roundish. In this example, cluster shape (e.g., whether a cluster appears to be round or elongated) has been examined, while density and size of the clusters were the same. In addition, 2D scatter plots are manually generated using synthetic clusters [17]. Cluster shape (in projected space) influences users' performance on various inference tasks.

Again by utilizing the Brehmer and Munzner multi-level typology, we provide an example that shows how our defined tasks can be fitted to this multi-level typology of abstract visualization tasks, in order to concisely describe our predefined tasks. *Find the cluster with the highest number of sub-clusters* in the behavior comparison category has been considered. Additionally, we did not

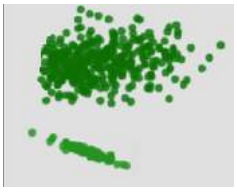

Fig. 10. Task: Compare the density of the longish cluster versus the roundish cluster. Scatter plots were generated with varying shapes, while holding density and size constant, in order to investigate the effect of cluster shape (in projected space) on a user's inferences and perceptions of the data.

consider any specific projection technique because it can be changed based on the evaluator's motivation.

Find the cluster with the highest number of sub-clusters: **WHY:** The purpose is to *Discover* a cluster with the highest number of sub-clusters. The cluster characteristic is not provided; therefore, the search target is unknown and *Explore* entails searching for the cluster with the highest number of sub groups. Once the search process is done, *Identify* returns the desired reference. **WHAT:** The input for this task is the entire visualization, including all clusters and their sub-groups. The output is the identity of a cluster with the largest number of sub-clusters. **HOW:** Individuals need to estimate the number of sub-clusters of each cluster. This involves counting sub-groups within successive clusters until the largest number is found. Therefore, they must *Derive* new data elements, then *Select* the desired cluster.

3.4 Membership disambiguation

It is desirable for the visual representation to avoid clutter, resolve ambiguity and handle noise. At times, "identifying overlaps" may indicate that the classes are not clearly separable, which suggests that the overriding task is one of pattern identification. However, too much data on too small an area of the display, such as a dense region of entangled clusters, diminishes the potential usefulness of the projections even if the projection consists of some clearly separated clusters. This is especially true when the user is exploring the data to: 1) Estimate the number of objects in a selection; 2) Find an object with specific characteristic (e.g. labeled point) within a cluster; 3) Count the number of objects in a given cluster; 5) Identify the objects that overlap in a selected area.

When *Finding an object with a specific characteristic within a cluster*, a visualization can favor good performance in preserving distances and relationships, but only at the expense of producing visual clutter. As an example, the PCA scatterplot of KDViz is too cluttered and distinguishing a specific object within a cluster is not an easy task (Figure 11).

To *Estimate the number of objects in a selection*, a target cluster/selection can be highlighted with a different color as shown in Figure 12.

Fig. 11. Find a purple object within the green cluster. Using a PCA projection employed on the KDViz data set, it is hard to distinguish the purple point.

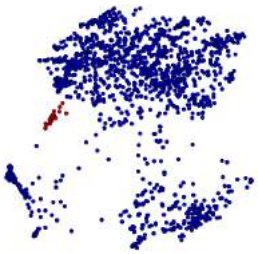

Fig. 12. Estimate the number of objects in a selection in LSP projection.

A recent study [16] showed that a density-based motion can enhance pattern detection and cluster ranking tasks for multidimensional data projections and also uncover hidden relationships in scatterplots.

3.5 Meta-projection

The tasks that are explained above can be used as given, or can be combined into multi-step macrotasks. We note that the tasks that we have provided may not cover all possible tasks of a given type, but they can be used as exemplars when defining new tasks. Sub-clusters of a given cluster or group of points can be considered as a meta-object. Meta-objects can create a meta-projection, and new tasks can be executed on this projection based on this process. In Figure 13(a), the task is: "*Find the closest cluster to the given cluster*". For instance, as apparent "Linear Square" is the closest sub-cluster to the "Information Visualization" sub-cluster and "Tree" is the closest sub-cluster to "Graph Drawing". Therefore, as shown in Figure 13(b) we can analyze the meta-projection to see that "Time Varying Filtering" is the closest cluster to the "Visualization" cluster and similarly "Visualization" is the closest cluster to "Data Mining". Using this meta-projection, we can get more insight into our data.

Thus, in section 3, we saw examples of how appropriate visualization methods could be determined for specific tasks.

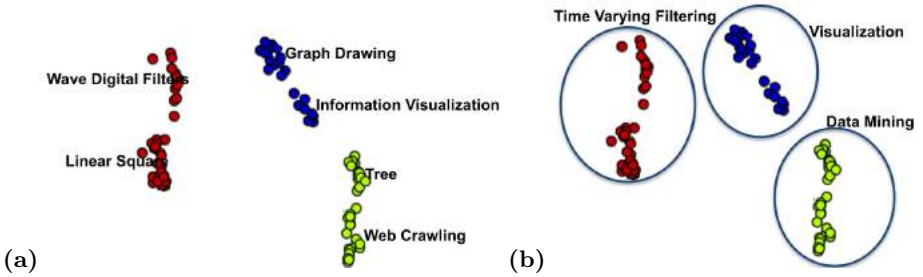

Fig. 13. A meta-projection: (a) sub-clusters; (b) clusters (meta-objects).

4 Conclusion

Our user-centric guideline supports precise comparisons across different multidimensional data projection techniques. However, it could be further extended by considering a wider range of application domains that could introduce new visualization scenarios, such as volumetric data sets with continuous scatterplots. The tasks we have defined are specific neither to a particular projection algorithm nor dataset. Although we delineate a number of example tasks within each of our taxonomic task classifications, they are not intended to be exhaustive. We believe that our guideline could easily incorporate additional tasks; in future work we plan to extend it with further user-centric tasks. We argue that projection methods are distinct in their characteristics in terms of both sparseness and distance distribution, and that the nature of the task (in taxonomic terms) should guide the visualization design. Our taxonomy can be used for examining projection layouts and scatterplots in order to analyze how users perceive multidimensional data in a variety of situations. We also incorporate recent findings about perception rules and cognitive processes as a valuable source of information for such analyses; our guideline can help in categorizing possible tasks when analyzing multidimensional data visualizations. These user-centric tasks could be used as a guideline for assessing when other scatterplot visualization techniques are appropriate, such as Star Coordinates [50], StretchPlots [26, 27], or even animations based on point cloud datasets [51]; future work will explore the application of our guideline to a wider range of existing techniques.

References

1. Ahuja, N. and Tuceryan, M. (1989). Extraction of early perceptual structure in dot patterns: Integrating region, boundary, and component gestalt. *Comput. Vision Graph. Image Process.*, 48(3):304–356.
2. Albuquerque, G., Eisemann, M., and Magnor, M. (2011). Perception-based visual quality measures. In *Proc. IEEE Symposium on Visual Analytics Science and Technology (VAST) 2011*, pages 13–20.

3. Amar, R., Eagan, J., and Stasko, J. (2005). Low-level components of analytic activity in information visualization. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, INFOVIS '05, pages 15–, Washington, DC, USA. IEEE Computer Society.
4. Andrienko, G., Andrienko, N., Bak, P., Keim, D., Kisilevich, S., and Wrobel, S. (2011). A conceptual framework and taxonomy of techniques for analyzing movement. *J. Vis. Lang. Comput.*, 22(3):213–232.
5. Andrienko, N. V., Andrienko, G. L., and Gatalsky, P. (2000). Visualization of spatio-temporal information in the internet. In *11th International Workshop on Database and Expert Systems Applications (DEXA'00), 6-8 September 2000, Greenwich, London, UK*, pages 577–585.
6. Borg, I. and Groenen, P. J. F. (2010). *Modern Multidimensional Scaling Theory and Applications*. Springer Series in Statistics. Springer, 2nd. edition edition.
7. Brehmer, M. and Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE Trans. Visualization and Computer Graphics (TVCG) (Proc. InfoVis)*, 19(12):2376–2385.
8. Collins, C., Penn, G., and Carpendale, S. (2009). Bubble sets: Revealing set relations with isocontours over existing visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1009–1016.
9. Cuadros, A. M., Paulovich, F. V., Minghim, R., and Telles, G. P. (2007). Point placement by phylogenetic trees and its application to visual analysis of document collections. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, pages 99–106. IEEE Computer Society.
10. Duncan, J. and Humphreys, G. (1989). Visual search and stimulus similarity. *Psychological Review*, 96:433–458.
11. Eades, P., Huang, W., and Hong, S. (2010). A force-directed method for large crossing angle graph drawing. *CoRR*, abs/1012.4559.
12. Eades, P. A. (1984). A heuristic for graph drawing. In *Congressus Numerantium*, volume 42, pages 149–160.
13. Ess, A., Leibe, B., Schindler, K., and van Gool, L. (2008). A mobile vision system for robust multi-person tracking. pages 1–8, Anchorage, AK, USA.
14. Etemadpour, R., Carlos da Motta, R., Paiva, J. G. d. S., Minghim, R., Ferreira, M. C., and Linsen, L. (2014a). Role of human perception in cluster-based visual analysis of multidimensional data projections. In *5^{th} International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 107–113, Lisbon, Portugal.
15. Etemadpour, R., Motta, R., de Souza Paiva, J. G., Minghim, R., de Oliveira, M. C. F., and Linsen, L. (2014b). Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 21(1):81–94.
16. Etemadpour, R., Murray, P., and Forbes, A. G. (2014c). Evaluating density-based motion for big data visual analytics. In *IEEE International Conference on Big Data*, pages 451–460, Washington, DC.
17. Etemadpour, R., Olk, B., and Linsen, L. (2014d). Eye-tracking investigation during visual analysis of projected multidimensional data with 2d scatterplots. In *5^{th} International Conference on Information Visualization Theory and Applications (IVAPP)*, pages 233–246, Lisbon, Portugal.
18. Geng, X., Zhan, D.-C., and Zhou, Z.-H. (2005). Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(6):1098–1107.

19. Henry, N. and Fekete, J. (2006). Matrixexplorer: a dual-representation system to explore social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12:677–684.
20. Ingram, S., Munzner, T., Irvine, V., Tory, M., Bergner, S., and Mller, T. (2010). Dimstiller: Workflows for dimensional analysis and reduction. In *IEEE VAST*, pages 3–10. IEEE.
21. Ingram, S., Munzner, T., and Olano, M. (2009). Glimmer: Multilevel mds on the gpu. *IEEE Transactions on Visualization and Computer Graphics*, 15(2):249–261.
22. Jolliffe, I. T. (1986). *Pincipal Component Analysis*. Springer-Verlag.
23. Koffka, K. (1935). Principles of Gestalt Psychology. . *Lund Humphries, London.*
24. Lewis, J. M. and Ackerman, M. (2012). Human cluster evaluation and formal quality measures: A comparative study. pages 1870–1875. 34th Annual Conference of the Cognitive Science Society.
25. Müller, E., Günnemann, S., Assent, I., and Seidl, T. (2009). Evaluating clustering in subspace projections of high dimensional data. *PVLDB*, 2(1):1270–1281.
26. Murray, P. and Forbes, A. G. (2014a). StretchPlot: Interactive visualization of multi-dimensional trajectory data. In *Proc. of IEEE Visual Analytics Science and Technology (VAST)*, pages 261–262, Paris, France.
27. Murray, P. and Forbes, A. G. (2014b). Interactively exploring geotemporal relationships in demographic data via stretch projections. In *Proc. of the ACM SIGSPATIAL International Workshop on Interacting with Maps (MapInteract)*, pages 29–35, Dallas, Texas.
28. Paiva, J., Schwartz, W., Pedrini, H., and Minghim, R. (2015). An approach to supporting incremental visual data classification. *Visualization and Computer Graphics, IEEE Transactions on*, 21(1):4–17.
29. Paiva, J. G. S., C., L. F., Pedrini, H., Telles, G. P., and Minghim, R. (2011). Improved similarity trees and their application to visual data classification. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2459–2468.
30. Paiva, J. G. S., Schwartz, W. R., Pedrini, H., and Minghim, R. (2012). Semi-supervised dimensionality reduction based on partial least squares for visual analysis of high dimensional data. *Computer Graphics Forum*, 31(3pt4):1345–1354.
31. Paulovich, F. V., Nonato, L. G., Minghim, R., and Levkowitz, H. (2008). Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Transactions on Visualization and Computer Graphics*, 14(3):564–575.
32. Peng, W., Ward, M. O., and Rundensteiner, E. A. (2004). Clutter reduction in multi-dimensional data visualization using dimension reordering. In Ward, M. O. and Munzner, T., editors, *INFOVIS*, pages 89–96. IEEE Computer Society.
33. Poco, J., Etemadpour, R., Paulovich, F. V., Long, T. V., Rosenthal, P., de Oliveira, M. C. F., Linsen, L., and Minghim, R. (2011). A framework for exploring multidimensional data with 3d projections. *Comput. Graph. Forum*, 30(3):1111–1120.
34. Rensink, R. A. and Baldridge, G. (2010). The perception of correlation in scatterplots. *Comput. Graph. Forum*, 29(3):1203–1210.
35. Rosenholtz, R., Twarog, N. R., Schinkel-Bielefeld, N., and Wattenberg, M. (2009). An intuitive model of perceptual grouping for hci design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1331–1340, New York, NY, USA. ACM.
36. Samet, H. (2005). *Foundations of Multidimensional and Metric Data Structures (The Morgan Kaufmann Series in Computer Graphics and Geometric Modeling)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

37. Schreck, T., von Landesberger, T., and Bremm, S. (2010). Techniques for precision-based visual analysis of projected data. In Park, J., Hao, M. C., Wong, P. C., and Chen, C., editors, *VDA*, volume 7530 of *SPIE Proceedings*, page 75300. SPIE.
38. Schwartz, W. R. and Davis, L. S. (2009). Learning discriminative appearance-based models using partial least squares. Rio de Janeiro, Brazil.
39. Sears, A. (1995). Aide: A step toward metric-based interface development tools. In *Proceedings of the 8th Annual ACM Symposium on User Interface and Software Technology*, UIST '95, pages 101–110, New York, NY, USA. ACM.
40. Sedlmair, M., Brehmer, M., Ingram, S., and Munzner, T. (2012a). Dimensionality reduction in the wild: Gaps and guidance - ubc computer science technical report tr-2012-03. Technical report, The University of British Columbia.
41. Sedlmair, M., Tatu, A., Munzner, T., and Tory, M. (2012b). A taxonomy of visual cluster separation factors. *Comp. Graph. Forum*, 31(3pt4):1335–1344.
42. Sips, M., Neubert, B., Lewis, J. P., and Hanrahan, P. (2009). Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum (Proc. EuroVis 2009)*, 28(3):831–838.
43. Song, Y., Zhou, D., Huang, J., Councill, I. G., Zha, H., and Giles, C. L. (2006). Boosting the feature space: Text categorization for unstructured data on the web. In *the Sixth IEEE international Conference on Data Mining, (ICDM 2006)*. IEEE.
44. Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison-Wesley Longman, Boston, MA, USA.
45. Tatu, A., Bak, P., Bertini, E., Keim, D. A., and Schneidewind, J. (2010). Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '10)*, pages 49–56.
46. Tatu, A., Theisel, H., Magnor, M., Eisemann, M., Keim, D., Schneidewind, J., and et al. (2009). Combining automated analysis and visualization techniques for effective exploration of high-dimensional data.
47. Tenembaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric faramework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.
48. Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. *The Experimental Psychology, Human perception and performance*, 8(2):194–214.
49. Tullis, T. S. (1988). A system for evaluating screen formats: Research and application. *Hartson, H. Rex, and Hix, Hartson, Advances in Human-Computer Interaction*, 2:214–286.
50. Van Long, T. and Linsen, L. (2011). Visualizing high density clusters in multidimensional data using optimized star coordinates. *Comput. Stat.*, 26(4):655–678.
51. Villegas, J., Etemadpour, R., and Forbes, A. G. (2015). Evaluating the perception of different matching strategies for time-coherent animations. In *Human Vision and Electronic Imaging XX (HVEI), volume 9394 of Proceedings of SPIE-IS&T Electronic Imaging*, pages 939412–1–13. San Francisco, California.
52. Wold, H. (2004). *Partial Least Squares*. John Wiley & Sons, Inc.
53. Zhang, X., Pan, F., and Wang, W. (2008). Care: Finding local linear correlations in high dimensional data. *2014 IEEE 30th International Conference on Data Engineering*, 0:130–139.
54. Zhang, Y., Passmore, P. J., and Bayford, R. H. (2009). Visualization of multidimensional and multimodal tomographic medical imaging data, a case study. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1900):3121–3148.