Predicting User Preferences of Dimensionality Reduction Embedding Quality

Morariu, Cristina; Bibal, Adrien; Cutura, Rene; Frenay, Benoit; Sedlmair, Michael

Published in:
IEEE Transactions on Visualization and Computer Graphics

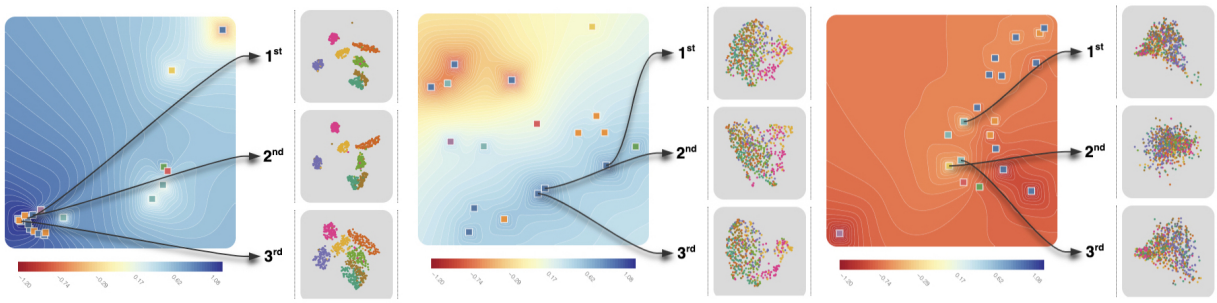
DOI:
10.1109/TVCG.2022.3209449

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Predicting User Preferences of Dimensionality Reduction Embedding Quality

Cristina Morariu, Adrien Bibal, Rene Cutura, Benoît Frénay, *Member, IEEE* and Michael Sedlmair, *Member, IEEE*



a) MNIST b) Flowers c) Paintings

Fig. 1: Overview of the best embeddings, as predicted by our method, for three of the datasets we have collected: MNIST, photos of flowers, and paintings. On the left of each sub-figure, we have an embedding of embeddings (called a *metamap*) where each square represents an embedding that was considered in our study and the contour color coding represents goodness of embedding (with dark blue tones meaning "good" and dark red tones meaning "bad"). On the right of the sub-figures, we have the top 3 best embeddings. The background of the metamap is visualized using a "goodness" score for the embeddings outputted from our model. The top 3 MNIST embeddings (blue background - high score) are of higher quality than the top 3 painting embeddings (red background - low score).

Abstract—A plethora of dimensionality reduction techniques have emerged over the past decades, leaving researchers and analysts with a wide variety of choices for reducing their data, all the more so given some techniques come with additional hyper-parametrization (e.g., t-SNE, UMAP, etc.). Recent studies are showing that people often use dimensionality reduction as a black-box regardless of the specific properties the method itself preserves. Hence, evaluating and comparing 2D embeddings is usually qualitatively decided, by setting embeddings side-by-side and letting human judgment decide which embedding is the best. In this work, we propose a quantitative way of evaluating embeddings, that nonetheless places human perception at the center. We run a comparative study, where we ask people to select "good" and "misleading" views between scatterplots of low-dimensional embeddings of image datasets, simulating the way people usually select embeddings. We use the study data as labels for a set of quality metrics for a supervised machine learning model whose purpose is to discover and quantify what exactly people are looking for when deciding between embeddings. With the model as a proxy for human judgments, we use it to rank embeddings on new datasets, explain why they are relevant, and quantify the degree of subjectivity when people select preferred embeddings.

Index Terms—Dimensionality reduction, Manifold learning, Human-centered computing.

1 Introduction

A wide-spread approach for data exploration is the use of dimensionality reduction (DR) techniques. DR is a process that projects high-dimensional data to a lower-dimensional space, such that the resulting embedding retains specific properties from the original data. An application of DR is in visualization, where users can create scatterplots based on two retained dimensions. DR methods are used in various domains ranging from biology and medical research to social sciences (e.g., [22, 26, 54]), and they are actively researched in both the machine learning (ML) and visualization (VIS) communities.

An extensive amount of techniques exists to produce such embeddings, such as principal component analysis (PCA) [7], multidimensional scaling (MDS) [23], isometric feature mapping (Isomap) [44], t-distributed stochastic neighborhood embedding (t-SNE) [46] and, more recently, uniform manifold approximation (UMAP) [32]. These methods can produce widely different results, all the more so given that some have hyper-parameters (e.g., t-SNE perplexity).

Evaluating the quality of these results is, however, the burden of users. In a typical process, a user generates a range of embeddings, visualizes them in scatterplots, and selects a suitable one from the line-up. Several attempts have been made to improve our understanding of what users look for when evaluating embeddings. Some studies focus on investigating whether human judgment is indeed reliable for evaluating embeddings [29], while others focus on defining the tasks that users perform when investigating embeddings [9]. Previous work also shows that people use DR as a black-box mechanism without necessarily understanding what the objective of the specific technique is [28, 29]. To consolidate the evaluation of embeddings quantitatively, both the ML and VIS communities proposed quality metrics that can be used to select the best embeddings automatically (these metrics are detailed in Section 2.1).

In this paper, we aim at bridging previous research on quality metrics for DR and scatterplot visualization, with the work done on understand-

- *Cristina Morariu and Adrien Bibal are co-first authors.*
- *Cristina Morariu, Rene Cutura and Michael Sedlmair are with the University of Stuttgart.Emails: {cristina.morariu, rene.cutura, michael.sedlmair}@visus.uni-stuttgart.de.*
- *Adrien Bibal is with the Université catholique de Louvain and the University of Namur. Email: adrien.bibal@uclouvain.be.*
- *Benoît Frénay is with the University of Namur. Email: benoit.frenay@unamur.be.*

Manuscript received 31 March 2022; revised 1 July 2022; accepted 8 August 2022. Date of publication 27 September 2022; date of current version 2 December 2022. This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2022.3209449, provided by the authors.
Digital Object Identifier no. 10.1109/TVCG.2022.3209449

ing human judgments of embedding quality. We evaluate to what extent metrics in the literature can quantify subjective user preferences. To this end, we gathered collections of images that we used to compute widely-used DR techniques (DRTs). Image datasets are used so that users have a glimpse of the high-dimensional space (via the image pixels) in the image thumbnails of the scatterplots. In total, 11 image collections were used, and 25 embeddings were computed for each, resulting from different hyper-parametrizations of 7 DRTs (PCA, MDS, Isomap, spectral embedding, Gaussian random projection, t-SNE with 10 different parameterizations, and UMAP also with 10 different parameterizations). Based on this data, we ran a 54-person user study to collect preferences on these embeddings. We then investigated in how far these human preferences can be formally expressed through existing quality metrics. Our aim is thus not to survey all DR methods, but rather to investigate whether quality metrics, or a combination thereof, can capture user preferences.

Our problem can be framed as the analysis of a supervised learning model, where a combination of quality metrics is used to predict human judgments. ML models are therefore used to compute how these metrics should be combined. The aim is to create and provide a model that can both predict embeddings users would most likely prefer, as well as to offer an explanation as to why they prefer them.

There are two main reasons for this choice. First, a supervised model will allow us to derive a composite metric based on user perception. The new metric can then be used to select embeddings that would generally be considered interesting. This is specifically important when many DRTs are considered, or for DRTs that have several non-trivial hyper-parameters to tune. Second, this approach will enable us to compare which quality metrics are important for expressing human preferences. In summary, our work makes the following contributions:

- a supervised framework that can be used to learn the relationship between a set of metrics and user preferences;
- the collection and analysis of data from a 54-participant user study on subjective preferences in DR embeddings;
- a quantitative analysis that (a) explains what users like when selecting DR embeddings, (b) sheds light on the feasibility of predicting preferences with quality metrics, and (c) allows us to better understand which ML and VIS metrics are important for that. To that end, we use three modeling approaches to combine quality metrics and to predict user preferences of embeddings, as well as an analysis on which approach performs best.

2 Background & Related Work

Our work considers the two main types of evaluation in dimensionality reduction (DR): the quantitative evaluation using visual and DR-specific quality metrics, and the qualitative evaluation based on human judgments. This section presents the latest work in these areas, as well as how our contributions build on top of this knowledge.

2.1 DR Evaluation using Quality Metrics

Measuring the quality of embeddings is the work of two communities, and each brought quality measures that have distinct properties. These different quality metrics are presented in this section. We also include a formal description of the metrics in our supplemental material.

2.1.1 Measures from the Machine Learning Community

The machine learning (ML) community has defined several measures that can be used as objective functions within dimensionality reduction techniques (DRTs). A good example is *stress*, the well-known objective function of multidimensional scaling, which measures the preservation of pairwise distances between the instances in the high-dimensional (HD) and the low-dimensional (LD) spaces. Beyond that, the ML community has investigated metrics that seek to define and measure the quality of the DR process itself. The rationale for this choice is that metrics that are used in objective functions are constrained in their definition (e.g., being differentiable), constraints that may not be necessary if the sole purpose is to measure quality [25]. Examples of such measures are the local continuity meta-criterion (LCMC) [12], the measure of Trustworthiness and Continuity [48] and AUC$_{log}$RNX [24]. These measures typically check if the neighborhoods in the HD space are preserved in the LD embedding. For instance, LCMC computes, for each point, the average number of neighbors it has in common in HD and LD for a certain neighborhood size k. *Trustworthiness*, on the other hand, is defined by roughly summing the rank of all pairwise distances from a point i in the original HD data to its nearest neighbors in the LD embedding that are not among the k nearest neighbors of i in the original data. This metric seeks to measure whether one can trust what can be seen in the visualization. The measure of *continuity* is the exact opposite, as it tells how well the patterns from the original dataset are projected in the visualization. The continuity for a particular neighborhood size k is defined by the rank of all pairwise distances from the point i in the LD embedding to the nearest neighbors of i in the original HD data that are not among the k nearest neighbors of i in the LD embedding. While the previously mentioned approaches focus on a specific neighborhood size k, AUC$_{log}$RNX consider all neighborhood sizes, with a focus on smaller neighborhoods. In order to do so, AUC$_{log}$RNX considers, for each point, the number of neighbors in common in LD and HD for all neighborhood sizes with a logarithmic importance.

2.1.2 Measures from the Visualization Community

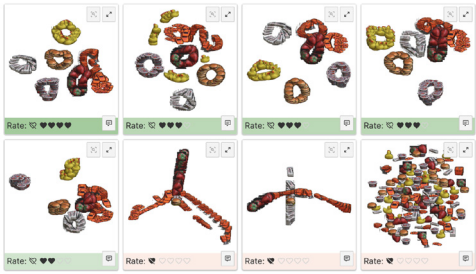
The other community that tackles measuring embedding quality is the visualization (VIS) community. Metrics from the VIS community generally focus on the quantification of visual patterns in the embeddings/scatterplots. A venerable example of such measures are the Scagnostics measures [52, 53], that quantify patterns such as Sparsity, Skewness, and Outlierness.

Recently, a substantial amount of work has focused on measuring class separability, that is, how well classes are separated in a DR embedding. Distance consistency (DSC), for instance, computes the number of instances that are closest to the centroid of their own class rather than another class. Alternatively, SepMe [1] provides an ensemble of separability metrics that use neighborhood graphs to assess how well classes are separated. These metrics are currently the best performing separability metrics in the literature. Other popular measures in this category are the average between-within clusters (ABW) [28], the hypothesis margin (HM) [20], the neighborhood hit (NH) [36] and the Calinski-Harabasz index (CAL) [11]. All these metrics measure the separability between clusters, albeit differently.

Similar to our goals, some works [1–3, 18, 27, 35, 39] focus on evaluating quality metrics against human perception, although with different use cases. Sedlmair and Aupetit [1, 40] examine perception of class separability in color-coded scatterplots, Pandey et al. [35] assess to what extent Scagnostics can be used as a proxy for human perception, and Lehmann et al. [27] evaluate whether Scagnostics can be used to filter perceptually interesting views for users. None of these works, however, focus, as we do, on recommending DR embeddings and explaining this recommendation using a large variety of quality metrics.

2.1.3 Accuracy and Interpretability Measures

The main difference between the measures designed in ML and those in VIS is their objective. ML metrics generally seek to measure how well the information is preserved when reducing the number of dimensions. In contrast, VIS metrics tend to focus on the presence of visual patterns that make it possible for users to grasp their visualizations and get insights about their data. Following the parallel of Bibal and Frénay [5] with supervised learning, the ML measures would be "accuracy" measures, while VIS measures would be "interpretability" measures. And, as in supervised learning, the two types of measures should be balanced to obtain results that would satisfy users [4, 5]. Indeed, accuracy measures are necessary because visualizations with well-separated clusters are not useful if they are not faithful to the high-dimensional space. Likewise, interpretability measures are also necessary as if readable patterns are not provided, nothing may be taken from the visualization.



(a) Image scatterplot view of the interface. This view is used so that users can see, through thumbnails, how the images from the dataset have been projected in 2D.

(b) Point scatterplot view of the interface. This view contains points instead of image thumbnails, with colors corresponding to class labels.

Fig. 2: Two views of the same trial from the experiment for collecting user preferences. Each view contains 8 embeddings of COIL-100 built by different DRTs. Black hearts correspond to the scores distributed among the embeddings. The use of the 2 views is analyzed further in Section 6.2

2.1.4 Combining the Different Quality Measures

One idea, which is the one followed by this paper, is to combine the two worlds by mathematically combining the metrics. For instance, Bibal and Frénay [5] formulated the linear combination of quality metrics as follows:

$$\begin{aligned}\text{combination} = (\alpha_1 * AM_1) + ... + (\alpha_i * AM_i) + ... + (\alpha_m * AM_m) \\ + (\beta_1 * IM_1) + ... + (\beta_j * IM_j) + ... + (\beta_u * IM_u),\end{aligned}$$

where AM (respectively IM) means accuracy metric (respectively interpretability metric). The different α and β, which are learned, represent the contribution of the metric to which they correspond.

Ensembles of metrics are also discussed in the quantitative survey of DR methods of Espadoto et al. [16]. The authors survey 44 DR methods and compute the average of several metrics (trustworthiness, continuity, neighborhood hit, normalized stress, Shepard goodness and local error) on 18 datasets in order to assess the global performance of individual DRT. Similarly, Nonato and Aupetit [34], as well as van der Maaten et al. [47], extensively review DRTs alongside quality metrics for DR, albeit without computing quality metrics on embeddings. We build on these works and go beyond by investigating learning the combination of measures that predict user choices.

2.1.5 Applications for Quality Metrics

Aside from the works mentioned above, the VIS community focuses on bridging the gap between quality metrics and human judgments by designing visual analytics (VA) systems that aid users in comparing [13] or selecting [14,21,31] embeddings. The insights derived from our contribution can be used as part of a VA system that recommends embeddings. Lehman et al. [27] also propose using specific quality metrics to filter out easily rejected embeddings, as scored by users. Wang et al. [50] use previously evaluated quality metrics of subjective class separability to propose a new DRT, which is implicitly optimized to model human perception of separability. Tian et al. [45] use accuracy metrics, alongside human judgments, to compare 2D and 3D embeddings. They note in their paper that each metric measures partial information captured by human judgment, which leads to the conjecture that combining metrics could lead to a more complete picture of the embedding quality.

2.2 Evaluation Driven by Human Judgments

Despite the existence of quality metrics, the burden in the evaluation of embeddings remains mainly on users. This section discusses DR research that collects and/or uses human judgments to assess quality.

2.2.1 Taxonomies for High-Level Tasks Related to DR

Brehmer et al. [9] aim to define what tasks users perform when they investigate embeddings. Following interviews, the authors introduce a characterization of tasks: *manifold tasks*, where users try to name the synthesized dimensions, and *cluster tasks*, where users verify, name, or match clusters with class names. These tasks have been considered in the selection of our datasets to ensure our study participants deal with different settings. This is important because our article focuses on an exploratory setup where users do not know in advance if clusters, outliers, or trends will be present. An alternative taxonomy of tasks, which partly overlaps with that of Brehmer et al. [9], is also proposed by Etemadpour et al. [17,18] for high dimensional embeddings. The authors proposed pattern identification, relation-seeking, behavior comparison, and membership disambiguation tasks. Another closely aligned work is the one of Sedlmair et al. [42], which proposes a cluster analysis taxonomy, one of the most important analysis tasks in the DR data exploration process.

2.2.2 Assessing User Preferences in DR

Lewis and van der Maaten [29] investigate whether human judgments are consistent by running a user study with groups of experts and novices. They offer the users little information regarding the original dataset and find out different users prefer different embeddings, inferring that user preferences are vastly subjective. They show that the more users have expertise, the more they are coherent in their judgment. Our study setup builds up on this one, as both studies focus on the real-life task of users selecting embeddings from a line-up. However, our goal is (i) to deepen the understanding about how users make their decisions and (ii) to model these for recommending embeddings.

Bibal and Frénay [4] also ran a user study collecting user preferences of t-SNE embeddings of the MNIST dataset. The objective of the authors is to study how cluster separability measures and their combination (using a modified Cox model) can predict user preferences. The study presented in this paper is larger in scale at all levels: more datasets, more DRTs (not only t-SNE), more quality metrics and different ways to frame the problem and to combine metrics. This enlargement in scope allows us to perform original analyses and to draw insightful conclusions. For instance, we can extract the quality metrics from the literature that can be used to predict user judgments, we can assess the importance of the accuracy of the DR process with respect to the visual quality for users, we can highlight the DR techniques that are both accurate and visually appealing for users, etc.

2.2.3 Selecting DR embeddings

Oftentimes, when new DR methods are introduced, a comparative study to other techniques is proposed as an evaluation. The embeddings get visualized in scatterplots and the reader assesses the line-up and decide for themselves which is the superior embedding. This can also be the case for the selection of hyper-parameter values inside a particular DRT. For instance, the authors of t-SNE invite users to try various hyper-parametrizations and select the embedding they prefer [46].

Wattenberg et al. [51] show that blindly trying hyper-parameters and selecting appealing embeddings has downfalls, in that it can mislead users on the faithfulness of the embedding. In our work, in order to avoid this issue, image thumbnails are provided in the scatterplot in order for users to check if the visualization reflects the HD space (characterized by the pixels of the projected images). Another issue is that user guidelines given by authors often are technique-specific. To overcome such issues, Sedlmair et al. [41] assess the best visualization methods to use during DR exploration, and provide guidelines on selecting DRTs using visualizations based on data collected in a user study.

Etemadpour et al. [18] designed a user study to assess which embeddings can best enhance users' abilities to detect clusters, outliers, or estimate density. The best embeddings were recommended based on the performance on different tasks. Not all embeddings perform well for all tasks, but in general, two techniques, Isomap and LSP, outperform the others. Recent popular DRTs, such as *t*-SNE or UMAP, were not included in this study.

3 User Study & Data Collection

The main idea behind our approach is to (a) generate a sample of DR embeddings from a set of datasets, (b) collect human preferences for them, and (c) calculate quality metrics for them in order to see how far they can predict human DR preferences. Here, we describe these three components in more detail and provide motivation for our design choices.

3.1 Motivation for Considering Image Datasets

We first select suitable datasets that allow users to provide quality judgments for different DR embeddings. We need users to consider both the patterns seen in the scatterplot, for example whether there are clusters forming, and the accuracy of the embedding visualizations, for example if the clusters make semantic sense.

Assessing preferences by only supplying minimal information about the original data can result in highly subjective and inconsistent judgments across participants [29]. It might not be possible to properly judge whether meaningful clusters appear or whether a manifold was adequately unrolled when users have limited or no access to the high-dimensional space [9].

To ensure users can process the HD data they are analyzing, we only use collections of images for our study (as the HD data are characterized by the pixels of the projected images). Under this setup, the embeddings visualized as scatterplots have each position encoded as the thumbnail of the image getting projected at this location. For example, in the case of the COIL-100 dataset, a collection of objects photographed from different angles, the scatterplot contains thumbnails of objects as shown in Figure 2a. By showing images as thumbnails, access to the HD attributes (the pixels) is given along with the projected 2D position in the visualization. We hypothesize that users can see that the visualization does not reflect the HD space if different image thumbnails are close together (forming false clusters or clusters of mixed content) or if similar image thumbnails are far from each other. The same pixel values considered by our users are also used as input to the DRTs and for calculating metrics when the HD space is required.

3.2 Selected Datasets & Criteria

We collected 11 image datasets (see Table 1) based on two criteria. First, we selected *datasets that implicitly cover different potential tasks*, because our user study is intentionally not framed around a specific task, like looking for clusters, but rather as an exploratory data analysis. For example, in the case of the MNIST digits dataset, the most common visualization task is matching class names (the digits) to various clusters formed. In contrast, for the Stanford face dataset, users would look for a manifold with semantic properties such as the lighting going from light to dark or the orientation of the figure changing from left to right.

We sought to collect *datasets of various complexity* on the premise that it is much easier to state a preference on embeddings from a dataset like MNIST, as opposed to a more complex dataset like the Paris Building dataset consisting of larger and more messy real-world photos where the potential number of analysis tasks also increases (e.g., a user could be assessing day-to-evening lighting changes or could group photos by buildings). As a consequence, we considered both datasets analyzed in literature and real-life photography collections. As part of our study, users were asked to score the difficulty in rating their preference after each trial. Then for each dataset, its empirical difficulty was aggregated from the user responses (see Table 1). While participants' answers show that our selected datasets vary both in terms of task type (see Section 3.6) and complexity, measured both by response time and self-reported dataset difficulty (see Table 1), we do not claim that our datasets cover the entire space of image datasets. We mainly aimed to collect a set of datasets that is representative enough to show that we can use metrics to model preferences.

3.3 Dimensionality Reduction Techniques

The dimensionality reduction techniques used to generate the embeddings are:

- principal component analysis (PCA) [7]
- multidimensional scaling (MDS) [23]
- isometric feature mapping (Isomap) [44]
- *t*-distributed stochastic neighborhood embedding (*t*-SNE) [46], with 10 perplexity values ranging from 5 to 100
- uniform manifold approximation (UMAP) [32], with 10 combinations of number of neighbors, ranging from 2 to 15, and maximum distances, from 0.1 to 0.8
- spectral embedding (SE) [33]
- Gaussian random projection (GRP) [6]

These seven DRTs selected for this paper are a representative set of what is popular in the literature [16]. Considering different hyper-parametrizations of the DRTs, 100 embeddings were initially generated for each dataset and, 25 embeddings for each dataset were uniformly sampled based on the metric space to be used in the user experiment. The metrics computed for each embedding and that we use to down-sample the initial 100 are described in Section 3.7. Last, we manually down-sampled embeddings that appeared very similar, e.g., rotated variants, or duplicates of one another. This resulted in 15 to 20 distinct embeddings per dataset. Finally, for each trial of the study, we selected 8 embeddings from the pool. We opted for uniformly sampling 8 embeddings from the pool rather than by DRT. On the one hand, this mean that PCA might not have been selected in each trial of the study, but on the other hand, it gave each hyper-parametrization and pattern a fair evaluation opportunity. A good example is UMAP, which produces some of the most preferred embeddings for specific hyper-parametrizations, and some of the worst otherwise (see Figure 3). Without equal chance for all hyper-parametrizations to be seen, we could by chance reach the conclusion that UMAP is a bad DRT overall if the less preferred embeddings were shown more often. We believe this sampling decision ensured that each pattern and hyper-parametrization is evaluated fairly. The DR hyper-parameter value ranges were selected from the original paper recommendations of what tends to work well [7, 32, 44, 46] and can be seen on the axes of Figure 3. We believe each DRT has a fair chance to produce good embeddings.

3.4 Visualizing the Embeddings

One issue with scatterplots is over-plotting. Since we are trying to model human preferences, it would not be fair to apply the metrics on points that are not seen by our participants. For this reason, we first created a set of scatterplots and measured the over-plotting of each point. When creating them, we made sure that the plotting order was random so that no sample with a particular property would be systematically occluded. We then removed from our embeddings datasets, from the HD datasets, and from the final scatterplots all points that would be invisible to the eye. This approach is inspired by Sedlmair et al. [40,41], who also remove occluded points when evaluating separability metrics based on human judgments.

3.5 User Preferences Dataset

This section describes the user experiment that has been set up to collect user preferences on embeddings.

Table 1: List of the datasets used in our experiment. For each dataset, we provide the name, description, the proportion of difficulty ratings given by users (easy - green, medium - amber, hard - red), the percentage of disagreement in user rankings for projections in that dataset (% DA), and the median response time in minutes (Median RT).

Dataset Name	Description	Difficulty (as scored by users)	% DA	Median RT
COIL-100	Objects photographed from different angles (128 x 128)	83.0% 7.0% 10.0%	11%	1.38
MNIST	Handwritten digits (28 x 28)	52.0% 30.0% 18.0%	13%	2.93
Fashion MNIST	Images of clothes (28 x 28)	57.0% 26.0% 17.0%	20%	2.33
Stanford Faces	Bust from different angles/light conditions (50 x 50)	35.0% 35.0% 29.0%	19%	2.04
Yale Faces	14 people displaying happy, neutral or sad faces (320 x 243)	12.0% 40.0% 48.0%	20%	2.45
Flowers	Photos of 6 different species of flowers (500 x 500)	22.0% 42.0% 36.0%	20%	2.29
Caltech plants	Photos of 6 different species of plants (320 x 243)	24.0% 40.0% 36.0%	18%	1.90
Caltech vehicles	Photos of 6 different types of vehicles (320 x 243)	15.0% 44.0% 41.0%	22%	2.11
Caltech instruments	Photos of 6 different types of instruments (320 x 243)	10.0% 41.0% 48.0%	21%	3.35
Paris Buildings	Photos of buildings in Paris (1024 x 768)	29.0% 44.0% 27.0%	14%	1.95
Oxford Buildings	Photos of attractions in Oxford (1024 x 768)	12.0% 27.0% 61.0%	24%	2.95

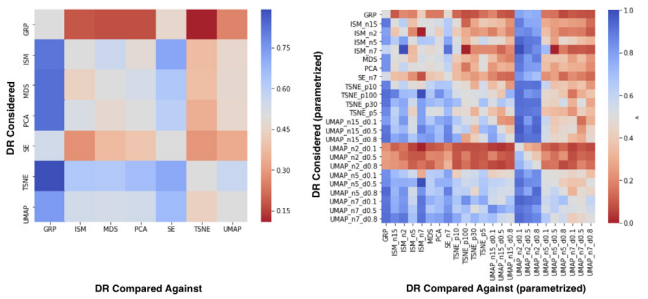


Fig. 3: User aggregated preferences of DRTs, overall (first) and hyper-parametrized (second). A score higher than 0.5, depicted in blue, means that more than 50% of the users preferred the DRT specified in the row (DR Considered: y-axis) over the one specified in the column (DR Compared Against: x-axis). Scores lower than 50% are encoded in red.

Participants In total, 54 users participated in our study, out of which 4 had finished a Ph.D., 38 had a master's degree and the remainder 12 had a bachelor's degree. We reached our users by advertising the study within the university network of the co-authors. Participation was voluntary and unpaid. We asked participants for their domain expertise in ML, VIS and dimensionality reduction, and over 85% of our user base reported familiarity with all these concepts. Seven participants reported no prior knowledge of dimensionality reduction, and one reported no prior knowledge of ML or VIS. An overview of what we asked to all participants can be seen in the supplemental material and in the data we released.

Study Procedure We conducted a web-based user study that took place online and on various display sizes (the minimum size was 700 x 500). The study began with a page explaining the subject of the study and its duration (40 to 60 minutes). Users were then presented a consent form, an explanation of what DR is and of the user interface, and a questionnaire to collect demographic and experience data. The study then proceeded with the trials. At the end, participants were asked about the overall difficulty of the setup and any other feedback.

Trial Setup Our study consisted of trials in which users had to rate embeddings. The stimuli in each trial were the embeddings generated by applying DRTs to the aforementioned datasets. Participants were not given a particular visualization task like finding clusters, but were rather asked to select the most appealing embeddings as part of an exploratory analysis task. No time limit was given for any section of the study, including the trials, as exploratory tasks are usually open-ended.

Eight embeddings of the same dataset were shown per trial. They were randomly selected from the total embeddings available for a dataset, placed on a 2-by-4 grid in random order, and shown as scatterplots of images on a white background. The views were connected by brushing and linking. At the beginning of each trial, participants received 15 points to distribute across the 8 embeddings. A higher number of points assigned to an embedding means that the participant preferred this embedding more. One embedding could receive a total of 4 points. A user could mark an embedding as bad by clicking a dislike button, rather than distributing any point to it. Participants could sort the grid by preference. This interaction is visualized in the supplemental material. Sorting enabled them to focus on a local comparison of embeddings with their better and worse direct neighbors. The sorting mechanism and the restricted number of points per trials were designed to force users into deciding which embeddings they liked more. Our intention was to avoid a situation where a user would award every embedding an equal number of points. Clicking on a dislike button was implemented to ensure the fact that the user actively decides something is disliked, rather than not rating something because of running out of points. A rated and sorted example of a trial, with both assigned liked and disliked examples, is presented in Figure 2. As Figures 2a and 2b show, the embeddings can either be visualized as non color-coded image-thumbnails or as color-coded points. The default view is the non-color coded thumbnail view, but the user can click a radio button to see the secondary color-coded dot view. In Section 6.2, we analyze how often users made use of this secondary view and the implication of including it. The user study is available at `https://kix2mix2.github.io/DumbleDR/public/index.html` and was tested on Firefox and Chrome. Screenshots of the entire study are also available in the supplementary material. Upon completion of a trial, participants were asked to score the difficulty of the trial and whether they would like to score another dataset. Users could complete up to 11 trials, each trial testing one dataset. The datasets across trials appeared in random order.

Descriptive Results We first analyze the degree of consensus between users when it comes to preferences. Previous work [29] shows that there is a high degree of subjectivity when it comes to users recording preferences of DR embeddings. Users' ability to select good quality embeddings is called into question. In our study, however, we report that while there were disagreements in ratings, the majority converged towards well-defined preferences. For each pairwise comparison between two embeddings, the best case scenario is that all judgments are in agreement, i.e. 0% of disagreement. The worst case is that opinions are evenly split when comparing the embeddings, i.e. half of the judgments are in disagreement with the other half (50% of disagreement).

On average, 18.5% of the ratings are in disagreement with the majority. A breakdown of disagreement in conjunction with the difficulty of the dataset as scored by users can be seen in Table 1. Datasets perceived as harder incur a higher percentage of disagreements. The same applies the other way around, where easy datasets such as MNIST have a low percentage of disagreements. In terms of response time per task, we observe a difference in the median response time dependent on the difficulty: easy trials took 1.69 minutes to answer, medium tasks 2.4 minutes, and hard ones 3.1 minutes. The average response time by dataset is also reported in Table 1. It can be observed that easy and medium scores are correlated to shorter response times.

From the ratings awarded in each trial by each user, a preference matrix is calculated by counting how many times an embedding is scored higher than another one. The results are aggregated to assess if particular DRTs are preferred. In Figure 3, the user preferences are aggregated on a DRT level. The heatmap encodes often users agree that one DRT (row-wise) is better than another (column-wise). The bluer the cell, the more people agree that the DRT in the row is better than the one in the column. Overall, a hierarchy can be observed: GRP and SE $\leq$ PCA and MDS $\leq$ Isomap $\leq$ t-SNE $\leq$ UMAP, where $DR_i \leq DR_j$ means that the visualizations generated by DR_j are more often preferred to the ones generated by DR_i. These observed preferences are already very much inline with the metrics-based results from Espadoto et al. [16], which indicates that a correlation could exist between the metrics and user preferences.

3.6 Qualitative Feedback from the Users

In our user study, users had the option to give three types of feedback: for each embedding in particular, after each trial where the users compared embeddings stemming from a particular dataset, and at the end of the entire study. The supplemental material provides screenshots from the study for each type of recorded user feedback.

In total, 100 comments on **the embedding level** were provided, out of the 3713 embedding judgments that were made. These comments can be grouped into the following categories:

- The **embeddings appear random** or wrong, 39 related comments from 11 distinct participants, e.g., *"no pattern visible"*, *"only visible pattern is background color"*.
- **Whether more hearts would have been needed** for proper assessment, 9 comments from 2 participants.
- **Presence of outliers**, 4 comments from 3 participants, e.g., *"strange outliers"*, *"outliers are maybe not meaningful"*.
- **Presence of manifolds**, 21 comments from 7 participants, e.g., *"just seems to order by lightness of the images."*, *"Sad people are on the left and happy ones are on the right."*
- **Whether clusters are formed or classes are separable**, 25 comments from 9 participants, e.g., *"clustered by people, not emotions"*, *"We can identify 4 clusters. One for each angle, and one for each lighting condition."*

At the **trial level**, comments were provided for 129 out of the 365 trials. People gave information about their ratings or ranking strategy. These comments aligned with the tasks introduced by Brehmer et al. [9], where people mentioned manifold tasks (e.g., *"If the face directions change continuously and consistent from left to right/top to bottom"*, *"manifold of flowers"*) and cluster tasks (e.g., *"class separability"*, *"Looked for very tight groupings when images were very similar, e.g., tightly grouped portraits with a dark background."*, *"I found only vague ordering with regards to overall color or lightness, and was dissatisfied."*). By manually coding each of these comments, we grouped the comments into 66 clustering tasks, 54 manifold tasks, and 9 trials with no patterns. At the end of the study, 25 out of the 54 participants gave us a **study level feedback**. Of these, 9 reported a positive experience, 10 provided us with potential study improvements (e.g., being able to access instructions again later on, or going back to redo ratings), 3 people reported the experience was *"mentally taxing"*, had a *"high cognitive load"*, or was *"confusing"*.

All data from our experiments, as well as the qualitative comments at all 3 levels, are available online (at `https://cloud.visus.uni-stuttgart.de/index.php/s/2tCMwl92LjISQ5a`).

Table 2: List of measures used in our analysis. If the metric is said to be applied on LD, then it only measures the quality of (or check patterns in) the visualization. These measures capture how interpretable the visualization is. If the metric is applied on HD to LD, it measures the accuracy of the DRT.

Metric Name	Type	Applied on
Outlying [52, 53]	Scagnostics	LD
Skewed [52, 53]	Scagnostics	LD
Clumpy [52, 53]	Scagnostics	LD
Sparse [52, 53]	Scagnostics	LD
Striated [52, 53]	Scagnostics	LD
Convex [52, 53]	Scagnostics	LD
Skinny [52, 53]	Scagnostics	LD
Stringy [52, 53]	Scagnostics	LD
Monotonic [52, 53]	Scagnostics	LD
ABW [28]	Cluster separability	LD
CAL [11]	Cluster separability	LD
DSC [43]	Cluster separability	LD
HM [20]	Cluster separability	LD
NH [36]	Cluster separability	LD
SC [37]	Cluster separability	LD
CC [19]	Correlation btw distances	HD to LD
NMS [23]	Stress	HD to LD
CCA [15]	Stress	HD to LD
NLM [38]	Stress	HD to LD
LCMC [12]	Small neighborhoods	HD to LD
T&C [48]	Small neighborhoods	HD to LD
NeRV [49]	Small neighborhoods	HD to LD
AUC_{log}RNX [24]	All neighborhoods	HD to LD

Even though our study was set with an exploratory task in mind, some users reported qualitatively undertaking a varied set of tasks. Taking into account these qualitative comments and our quantitative results, we are confident that our set of metrics is able to capture a well-rounded spectrum of DR tasks.

3.7 Quality Metrics Dataset

To predict user preferences, we gathered metrics from different communities that measure various aspects of visualizations, such as accuracy of representation or presence of observable patterns. Among the metrics in Table 2 that have not been presented in Section 2.1, one can find the silhouette coefficient (SC), the correlation coefficient (CC), the non-metric stress (NMS), the nonlinear mapping stress (NLM), the curvilinear component analysis (CCA) and the neighbor retrieval visualizer (NeRV). SC [37] is a classic metric in clustering that measures how clusters are separated from each other, versus how instances inside a same cluster are grouped together. This metric is similar to ABW, but diverges in its mathematical definition. CC [19] is a metric that computes the correlation between the vector of all pairwise distances in the original dataset and the corresponding vector of pairwise distances in the visualization. NMS [23], CCA [15] and NLM [38] are three stress measures that are considered in our study. Stress measures have in common that they measure how well pairwise distances in the high-dimensional space are preserved in the low-dimensional space. Each of the three measures have their particularities. For instance, NMS [23], as a non-metric measure, does not compare pairwise distances directly, but their ranking. Finally, NeRV [49] is a metric based on information retrieval, as it translates the concepts of precision and recall to a measure similar to the Trustworthiness and Continuity. NeRV redefines the distances in the original dataset and in the visualization as probabilities, like t-SNE. It also contains a perplexity hyper-parameter that defines the σ parameter of the distribution used to represent the size of the neighborhood to consider for each element in the visualization. Based on preliminary experiments, NeRV perplexity has been fixed at 5, as a single value should be chosen. Otherwise, the same NeRV metric would be duplicated several times in our set of metrics, with small differences due to the different perplexity values. The separability metrics are all highly correlated (see the supplementary material). Between pairs of highly correlated measures (more than 95%), only the most popular

one in each pair was kept. In consequence, the metrics dropped from further analysis were: $SepMe_{mvf}$, $SepMe_{mvt}$, Continuity, NH, and CC. Additionally we removed ABW and CAL, as they were low variance features.

4 Modeling User Preferences

In this section, we propose three ways to predict users' quality judgments of DR embeddings by using combinations of quality metrics. In order to do that, we model our data with incremental levels of detail:

- The first model classifies "good" and "bad" embeddings, as decided by users' consensus. In this case, only whether the heart of the embedding was crossed out or not is taken into account.
- The second model linearly learns which embeddings are preferred by users, by answering the question "Would embedding A be preferred over embedding B on average?".
- The third model provides a ranking of the embeddings, implicitly answering the same preference task as Model 2. In this case, we examine whether a nonlinear combination of the metrics can further improve the performance.

From all three models, we extract the most important metrics that help to predict user preferences. Despite disagreements in the data, no data and no participants were discarded from the training process. Hence, the models were trained on noisy annotations, as the same embedding may have conflicting annotations.

4.1 Modeling Setup

The evaluation of our models is operated on a leave-one-group-out basis. This is a cross-validation setup where the data is split into distinct groupings and a model is trained on the collected preferences related to all groups but one. The remaining group is used as a test set. The process is repeated for all combinations of groups. This is a special case of k-fold cross validation where the k folds correspond to well identified groups of the dataset. Throughout this section, we use the datasets in Table 1 as our different groups. We call this procedure leave-one-dataset-out (LODO). This setup allows us to check if our models generalize to unseen datasets.

Given the variety of datasets used and their different degrees of complexity (see Table 1), it is expected that all our models slightly vary in performance from dataset to dataset. Furthermore, computing a prediction score for each group also enables us to build a measure of prediction uncertainty on unseen data, by calculating the confidence interval over all test dataset results.

4.2 Model 1: Classifying Good and Bad Embeddings

Model 1 is set up to learn the distinction between "good" and "bad" embeddings. Each embedding was scored by multiple users, either with a set amount of points (i.e. hearts in the UI) or by crossing out the embedding (i.e. a crossed out heart in the UI). To aggregate all these potentially distinct scores across users for each embedding, we selected the median of the scores. We selected the median as opposed to mean, as this is an unbiased non-parametric estimator that is less susceptible to outlier annotations. Then, we classified an embedding as good (binarized to 1) if the aggregate score was at least one heart, and as bad (binarized to 0) otherwise. We also considered an alternative scenario where only embeddings scored with at least 3 hearts count as good, discarded embeddings whose median was 1 or 2, and rated the remainder as bad. Figure 2 shows these two categories with scatterplots highlighted either in green (good) or in red (bad). The data was fed to a boosted tree ensemble and evaluated on a LODO basis to determine the prediction performance for each dataset. We have selected this model as boosted trees are state-of-the-art models in supervised learning for tabular data like our dataset of metrics.

The area under the receiver operator curve (AUC) metric was optimized in the LODO setup. This AUC analysis resulted in the predictive performance of 89.81% with a confidence interval (CI) of $\pm$ 6.70%. In the 3-heart setup, the performance raises slightly to 90.2% ($\pm$ 5.9%). In terms of feature importance computed with SHAPley values [30], Scagnostics [52, 53] features such as Sparsity, Skinny and Outlying are the most important ones. For the majority of the embeddings, low Sparsity and high Skinniness increase the chances of an embedding to be disliked by participants. The embeddings selected by users as bad tend to be random (see example in the last position on the grid of Figure 2) or skinny embeddings (see example in the second and third to last places on the grid of Figure 2), where the 2D visualizations have no apparent meaning.

4.3 Model 2: Linear Preference Learning

For Model 2, we re-defined the problem as a linear preference learning problem. To do that, for each pair (v_i, v_j) of visualizations in a dataset, the percentage of time v_i is preferred over v_j is considered. For instance, 90% means that 90% of the time, when v_i and v_j were presented in the same trial to users, v_i received a larger number of hearts than v_j. Because the comparisons are aggregated to get percentages, the number of instances becomes 2268 for this dataset. The goal is to linearly reconstruct the preferences between visualizations based on the percentage of time a particular visualization has been preferred to another visualization. The advantages of linear models are their robustness to overfitting and their interpretability.

Bradley-Terry models (BTm) [8] are used as linear preference learning models. BTm linearly combines features to derive probabilities of being preferred:

$$P(v_i > v_j) = \frac{e^{w_0 + w_1 * m_{1,i} + \ldots + w_{23} * m_{23,i}}}{e^{w_0 + w_1 * m_{1,i} + \ldots + w_{23} * m_{23,i}} + e^{w_0 + w_1 * m_{1,j} + \ldots + w_{23} * m_{23,j}}},$$

where w_0, w_1, ..., w_{23} are 24 weights to learn, and $m_{k,i}$ (respectively $m_{k,j}$) are the k^{th} metric evaluated on the visualization v_i (respectively v_j). We trained the BTm with a Lasso penalty in order to encourage sparsity among weights.

The metrics that have been selected by the sparse BTm are, by order of importance, AUC_{log}RNX, NLM, Monotonic, Skewed, Sparse and DSC. The accuracy of the BTm is 62.30%, with a 95% CI of $\pm$ 3.91%. The accuracy is obtained by counting the number of time the model is right when it says $v_i > v_j$, over the total number of predictions. To obtain accuracy on data that have not been used for training, the LODO strategy has been used. The final accuracy is the mean of the test accuracy scores of the 11 involved datasets. This way, the reported final accuracy offers some guarantees on the use of the presented sparse linear model on new datasets. If only the data where users strongly agree on good and bad visualizations (at least 80.05% of agreement) is used, the accuracy becomes 65.93% $\pm$ 4.51%. The λ balancing the importance given to the error and the Lasso penalty was 0.021.

4.4 Model 3: Nonlinear Ranking of Embeddings

In our final setup (Model 3), like for Model 2, we output a measure of how good each embedding is. This makes it possible to answer the question "By how much is embedding A better than embedding B?". This measure acts as a popularity score and can also be used to compare if the embeddings generated for some datasets have a higher quality than for other datasets. We call this measure the ranking score. The ranking score is visualized in the teaser Figure 1, which will be discussed in detail in Section 5.

As opposed to Model 2, we chose for Model 3 a nonlinear model to exploit more complex relationships among the metrics and to potentially increase our performance. We used a boosted tree ensemble [10] again, in order to find out whether a nonlinear combination of our features, unlike the one in Equation 2.1.4, can lead to better results.

Model 3 is fed with embedding lists that are sorted according to the hearts awarded by the participants. The model then learns to rank embeddings from the 3713 ones, but sorted into 458 groups of 8 embeddings, as initially ranked by our participants. Model 3's objective is to create a ranking for a new, unseen, dataset of embeddings. This set of embeddings can be of any length, not just 8 embeddings, and the model learns to minimize the number of incorrect pairwise comparisons, as described by the LambdaMART algorithm [10]. The LODO error is calculated the same way as for Model 2. Overall, the accuracy is 70.03%, with a confidence interval (CI) of $\pm$4.40%. When the LODO error is calculated only for comparisons where there was a strong agreement, such as 80% agreement, the accuracy increases to 78.09%, $\pm$6.51%.

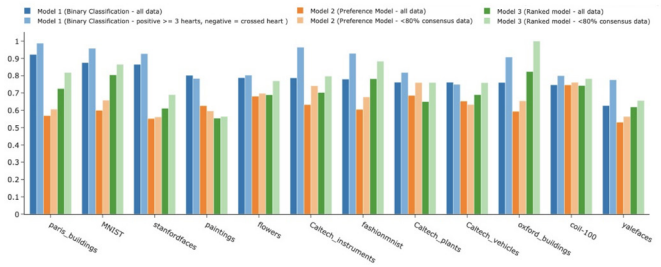


Fig. 4: Performance of the 3 experiments evaluated for each dataset. For the first experiment, we report the AUC on all data where a positive label is any positive "heart" ranking (dark blue), as well as when only a score greater or equal to three hearts counts as positive and a crossed heart counts as negative (light blue). For the two other experiments, we report the results on the entire dataset (dark orange for Model 2, and dark green for Model 3), as well as on data with more than 80% consensus (light orange for Model 2, and light green for Model 3).

5 Results & Discussion

This section discusses facets ranging from the performance of our setup to the generalization capacity of our models.

5.1 Performance on Unseen Datasets

In terms of generalization to other datasets, the LODO generalization performance of Model 1 ($\sim$90%) is our best result, which is expected given that a coarser classification between "good" and "bad" is the easier task. Models 2 and 3 achieve a LODO accuracy of 62.30% and 78.09% respectively on the more challenging preference ranking task. The 95% confidence intervals around these expected results are across all three models between 4 and 6.51%. These expected values alongside their CIs are a good indicator of how our trained models would perform on new unseen image datasets. As a further breakdown of our LODO performance, Figure 4 displays the accuracy for each dataset. Unsurprisingly, the model performs better on datasets that are rated as easier and with less disagreements between user annotations (see Table 1).

Some may wonder if the error could be further reduced. We offer two possible explanations for the cases of lower accuracy: 1) the metrics set that we have selected does not provide sufficient coverage over the space of user preferences, and 2) that it is difficult to mimic users' preferences with quality metrics given that users themselves disagree with each other.

5.2 Interpreting the Ranking Scores

Although with Model 3, we generate ranked predictions for each dataset, the top n embeddings are not equal in quality across datasets. This is reflective of reality: people might be more pleased with the results of DR on some datasets, but not on others. Similar to people's preferences, the ranking scores computed by Model 3 also vary from dataset to dataset. This can be seen in Figure 1, which shows the metamap of three datasets, MNIST, Flower photography, and ART UK paintings, and their corresponding top three embeddings according to Model 3.

Defined by Cutura et al. [14], metamaps are embeddings of embeddings. They are primarily used to find similar embeddings, encoded by the distance in between points in the metamap. However, another use of a metamap is to collect the most different embeddings in order to get different views of the same data. Indeed, as embeddings are clustered by similarity in the metamap, users can explore the different embeddings that can be produced. To do that, a user would compute hundreds of embeddings for the same dataset, produce the metamap, and consider the embeddings that are most distant from one another. We produced the metamaps for this use case by taking all the embeddings generated by our datasets in Table 1, computing the metrics in Table 2 for each of the embeddings, and finally applying UMAP on the metrics (which are the 23 features) describing each embedding (which are the instances) to obtain the metamaps. The left-hand side of each subfigure from Figure 1 represents a metamap, where each point symbolizes a particular embedding used in our study and rated by our study participants. The separability metrics took as labels the dataset associated with the embedding. We followed the same procedures as the ones described in Section 3.

The colors in the metamaps from Figure 1 represent the ranking scores of the visualizations: from dark blue for "good" visualizations, according to Model 3, to dark red for "bad" ones. The color interpolation is performed using inverse distance weighting. It can be seen that the top 3 embeddings of MNIST are good according to Model 3 (they are in a blue region and, so, have a high Model 3's score), barely good for Flowers (they are in a clearer blue region) and somewhat bad for Paintings (they are in a red region and, so, have a low Model 3's score). The information encoded in the metamap contours can be used to deduct that many embeddings from MNIST could be considered of good quality despite being outside of the top 3. For the paintings dataset, however, only few embeddings are good (there is a large red area, and few amber regions), and Model 3 helps to find the slightly better embeddings (in the amber area). The flower dataset, in the middle of Figure 1 contains both good and bad embeddings. A low ranking can also be interpreted as a result where subjectivity and disagreement are higher among humans. To further explore metamaps and top embeddings for the datasets used in our experiments, we have created a tool called DumbleDR (available at `https://renecutura.eu/dumbledr/`). The tool is described in the supplemental material.

5.3 Accuracy versus Visualization Metrics

All three models show that metrics from both the VIS and the ML communities are important. In addition to Scagnostics and cluster separability measures from the VIS community for detecting bad embeddings, our models also rely on accuracy measures to find accurate embeddings among the ones that contain readable patterns. This is quantitatively observed by a drop in performance when accuracy measures are removed from the training of our models. This drop is roughly 2% across all three models. This systematic decrease logically stems from the fact that users do pay attention, to some extent, to the semantics inside visualizations, in addition to looking for readable patterns. All in all, users pay more attention to the visual disposition of points in the visualization instead of the accuracy of the embedding. A more comprehensive breakdown of which metrics are better within the VIS and ML categories is available in the supplemental material.

5.4 Performance of DRTs

A bias spanning from the selection of image collections is that linear techniques such as PCA get rated down. Given the fact that images lie on a nonlinear manifold in the HD space, it makes sense that linear DR methods such as PCA underperform in comparison to UMAP and t-SNE. To evaluate the generalization to new DRT, a leave-one-dimensionality reduction-out (LODRO) AUC is calculated for Model 1. Rather than splitting by dataset during our cross-validation, as in LODO, we train to detect "good" and "bad" embeddings by considering all DRT but one. The LODRO procedure allows us to check if our analysis applies to unseen DRTs. Overall, our LODRO AUC to new DRTs is settling at 63.03%, with a confidence interval of $\pm$ 3.1% (see the supplementary material for results per DRT).

GRP and MDS have the worst generalization performance. These methods might generate very different patterns than the other DRTs. Users in our study graded the embeddings resulting from GRP, SE and some UMAP configurations as universally bad across all datasets (see Figure 3) and, have even commented about how these embeddings appear to be random. However, visualizations that appear to be random to the human eye have in fact a very different quality according to quality metrics, meaning that bad embeddings are not all bad in the same way. The LODRO strategy cannot be easily applied to Models 2 and 3, since, in these setups, we require more DRT, and more than 20 total embeddings per dataset in order to achieve significant results.

6 Limitations & Future Work

As all research, our work comes with a set of limitations, which specifically attain to the modeling approach (Section 4) based on inherently imperfect human subject data (Section 3).

6.1 On the Existence of Misleading Embeddings

A concern is that users can select appealing embeddings that are wrong with respect to the HD data. Based on the availability to information regarding the HD space (image thumbnails), we are confident that if any such "false positives" existed, they would have been caught and marked as bad. Our different models show that the majority of embeddings flagged as bad by participants can be detected using Scagnostics and separability measures. Given that no accuracy metric is needed for spotting bad embeddings, it rises the question of whether embeddings where meaningful clusters are formed in the visualization when these clusters do not exist in the HD space are actually possible.

6.2 On the Use of Labels and Color-coded Scatterplots

In our experiments, we provided an additional color-coded version of the visualizations, as seen in Figure 2b. We asked the participants to consider it as further information, but not to form their preferences. One may argue that the colors could have biased their preferences (e.g., by looking for separable, color-coded clusters). To shed light into this, we recorded every click users made in our study and, we analyzed how users made their decisions. About 17% of all rating decisions were made from the color-coded view. Moreover, only 4.10% of the time was spent in the color-coded view. In fact, for 50% of all trials, users did not use the color-coded view at all. Future work should be considered to examine if color-codes are beneficial or detrimental.

6.3 On the Limited Breadth of Dataset Types

A weakness of our study is that we only use image-based datasets. We did so as images give a natural anchor into the HD space, which was essential for our purpose. We speculate that our analyses can also be performed on other types of data because (1) users maintain their preferences for different datasets, and, (2) that the metrics applied on different dataset types generate a similarly distributed metric dataset.

To have some intuitions about these claims, we performed a qualitative analysis on a tabular dataset. Instead of images described by pixels, the tabular data is formed of visualizations described by quality metrics. We then passed these meta-embeddings through Model 3 and assessed the top 3 recommendations. These can be seen in Figure 5, which shows on the right-hand side the top 3 meta-embeddings (or metamaps) and their respective hyper-parametrizations. As with the results on the other dataset, Model 3 selects UMAP for the best embeddings when a certain hyper-parametrization is chosen. The best metamaps show that the embeddings are loosely clustered in accordance to the dataset, rather than DR technique, with the implication being that the same DRTs do not produce similar embeddings across datasets. This dataset would be difficult to test-drive via a user study because of the knowledge needed in DR techniques and quality metrics, but the authors of this paper were familiar enough with the dataset and the quality metrics to be able to make qualitative judgments on how good the produced embeddings are. The authors selected their preferred embeddings without knowing which algorithm generated them and their decision aligned with one of the top 3 best embeddings. The metamaps in Figure 1 are produced using UMAP with 7 neighbors and a minimum distance of 0.8. We invite our readers to further inspect this data in our DumbleDR tool (available at `https://renecutura.eu/dumbledr/`).

With this preliminary example in mind, we believe an interesting research direction would be to look at extending our study with tabular and text data datasets such as the ones used in Espadoto et al. [16]. Such studies would necessitate adequate LD representations of the HD tabular or text space, similar to the image thumbnails for image datasets or expert users acquainted with a particular dataset. Once the datasets and visualization methods are ironed out, our evaluation framework can be followed. An interesting line of analysis would be to check if noisiness in user preferences varies depending on the data type and how it is conveyed to the user.

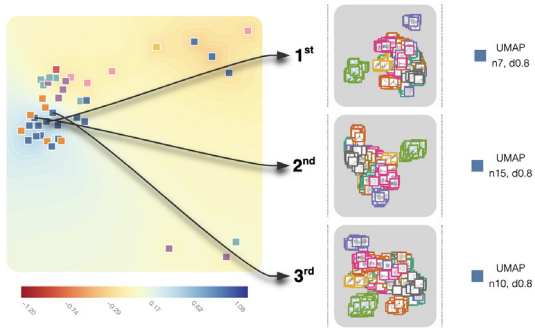


Fig. 5: Top 3 embeddings given by our tool on the set of metamaps. The ranking is provided by Model 3 and shows that UMAP with some particular hyper-parametrizations offers visualizations of good quality.

6.4 Extending to Other Metrics and DRTs

One can argue that new DRTs and quality metrics can be invented in the future. While this is true, one contribution of this paper is to present a framework on the use of quality metrics to predict user preferences. This means that new metrics can be plugged into our framework so that a new combination is automatically learned and analyzed without needing additional user feedback. Similarly, the combination can be re-trained on embeddings produced by new DRT, which would require a new user evaluation of these embeddings.

6.5 Predicting Behavior when Comparing Embeddings

Future work can consist of using the characteristics of users in our models to derive a different combination of metrics per user profile. The BTm model could be used to analyze how user characteristics influence their comparisons of embeddings. While BTm was used in this paper to predict the preferences based on features of the compared objects (the embeddings), BTm can also be used to predict the preferences based on the features of the users that stated their preferences. Another kind of behavior that can be modeled is when different tasks are performed. One can consider our framework as a basis to build models that would make it possible to understand what users consider for each task (finding clusters, outliers, trends, etc.), and how similar the tasks are when user preferences are modeled. One key question can be: does the importance of accuracy metrics change for each considered task?

7 Conclusion

This paper proposes a framework and an application of this framework to assess the quality of DR visualizations using metrics from the ML and VIS communities. We intended to open the black box of how users make up their preferences. We implemented three ML models to predict human preferences and examine to what extent metrics from both communities are used. The final model (Model 3) achieves 78.09% accuracy on ranking embeddings by user preference. Furthermore, Model 3 was implemented in a tool, called DumbleDR, to demonstrate the capabilities of our technique to highlight top quality embeddings.

In all three models, Scagnostics (in particular Sparsity, Skewed and Skinny) and separability measures (in particular DSC) have a large impact for predicting user choices. These metrics were able to easily discriminate between visualizations deemed good or bad by users. It seems that accuracy metrics from the ML community (in particular AUC_{log}RNX) are secondary, but they make it possible to discriminate between accurate and misleading visualizations with readable patterns.

Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project A08).

REFERENCES

[1] M. Aupetit and M. Sedlmair. SepMe: 2002 new visual separation measures. In *Proc. IEEE Pacific Vis. Symp.*, pp. 1–8, 2016.

[2] J. Bernard, M. Hutter, M. Zeppelzauer, M. Sedlmair, and T. Munzner. ProSeCo: Visual analysis of class separation measures and dataset characteristics. *Computers & Graph.*, 96:48–60, 2021.

[3] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Trans. Vis. Comput. Graph. (TVCG)*, 17(12):2203–2212, 2011.

[4] A. Bibal and B. Frénay. Learning interpretability for visualizations using adapted Cox models through a user experiment. In *NIPS Workshop Interpr. Mach. Learn. in Complex Syst.*, 2016.

[5] A. Bibal and B. Frénay. Measuring quality and interpretability of dimensionality reduction visualizations. In *ICLR Workshop Safe Mach. Learn.*, 2019.

[6] E. Bingham and H. Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proc. ACM Int. Conf. Knowledge Discovery and Data Mining (SIGKDD)*, pp. 245–250, 2001.

[7] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, New-York, NY, USA, 2006.

[8] R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[9] M. Brehmer, M. Sedlmair, S. Ingram, and T. Munzner. Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences. In *Proc. Workshop Beyond Time and Errors: Novel Evaluation Methods for Vis. (BELIV)*, pp. 1–8, 2014.

[10] C. J. Burges. From RankNet to LambdaRank to LambdaMART: An overview. Technical Report MSR-TR-2010-8, Microsoft Research, 2010.

[11] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Commun. in Stat. - Theory and Methods*, 3(1):1–27, 1974.

[12] L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association (JASA)*, 104(485):209–219, 2009.

[13] R. Cutura, M. Aupetit, J.-D. Fekete, and M. Sedlmair. Comparing and exploring high-dimensional data with dimensionality reduction algorithms and matrix visualizations. In *Intl. Conf. on Adv. Vis. Interfaces (AVI)*, pp. 1–9, 2020.

[14] R. Cutura, S. Holzer, M. Aupetit, and M. Sedlmair. VisCoDeR: A tool for visually comparing dimensionality reduction algorithms. In *Euro. Symp. on Artif. Neural Netw. (ESANN)*, pp. 105–110, 2018.

[15] P. Demartines and J. Hérault. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. on Neural Netw. and Learning Systems (TNNLS)*, 8(1):148–154, 1997.

[16] M. Espadoto, R. M. Martins, A. Kerren, N. S. Hirata, and A. C. Telea. Towards a quantitative survey of dimension reduction techniques. *IEEE Trans. Vis. Comput. Graph. (TVCG)*, 27(3):2153–2173, 2019.

[17] R. Etemadpour, L. Linsen, J. G. Paiva, C. Crick, and A. G. Forbes. Choosing visualization techniques for multidimensional data projection tasks: A guideline with examples. In *Int. J. Conf. Comp. Vis., Imaging and Comp. Graph.*, pp. 166–186, 2015.

[18] R. Etemadpour, R. Motta, J. G. de Souza Paiva, R. Minghim, M. C. F. De Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Trans. Vis. Comput. Graph. (TVCG)*, 21(1):81–94, 2014.

[19] X. Geng, D.-C. Zhan, and Z.-H. Zhou. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Trans. on Systems, Man, and Cybern., Part B (Cybernetics)*, 35(6):1098–1107, 2005.

[20] R. Gilad-Bachrach, A. Navot, and N. Tishby. Margin based feature selection - theory and algorithms. In *Proc. Int. Conf. Mach. Learn.*, p. 43, 2004.

[21] S. Ingram, T. Munzner, V. Irvine, M. Tory, S. Bergner, and T. Möller. DimStiller: Workflows for dimensional analysis and reduction. In *Proc. IEEE Symp. Vis. Anal. Sci. Technol.*, pp. 3–10, 2010.

[22] A. Koch, R. Imhoff, R. Dotsch, C. Unkelbach, and H. Alves. The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology*, 110(5):675–709, 2016.

[23] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[24] J. A. Lee, D. H. Peluffo-Ordóñez, and M. Verleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261, 2015.

[25] J. A. Lee and M. Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recog. Letters*, 31(14):2248–2257, 2010.

[26] W. Lee, D. Lee, Y. Lee, and Y. Pawitan. Sparse canonical covariance analysis for high-throughput data. *Statistical Appl. in Genetics and Molecular Biology*, 10(1):1–24, 2011.

[27] D. J. Lehmann, S. Hundt, and H. Theisel. A study on quality metrics vs. human perception: Can visual measures help us to filter visualizations of interest? *Inf. Technol.*, 57(1):11–21, 2015.

[28] J. Lewis, M. Ackerman, and V. de Sa. Human cluster evaluation and formal quality measures: A comparative study. In *Proc. Annu. Meet. Cognitive Science Society*, pp. 1870–1875, 2012.

[29] J. Lewis, L. Van der Maaten, and V. de Sa. A behavioral investigation of dimensionality reduction. In *Proc. of the Annual Meeting of the Cognitive Science Society*, 2012.

[30] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Mach. Intell.*, 2(1):2522–5839, 2020.

[31] R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea. Visual analysis of dimensionality reduction quality for parameterized projections. *Computers & Graph.*, 41:26–42, 2014.

[32] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426*, 2018.

[33] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Conf. Neural Inf. Process. Syst. (NIPS)*, pp. 849–856, 2001.

[34] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Trans. Vis. Comput. Graph. (TVCG)*, 25(8):2650–2673, 2018.

[35] A. V. Pandey, J. Krause, C. Felix, J. Boy, and E. Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *SIGCHI Conf. Human Factors Comput. Syst.*, pp. 3659–3669, 2016.

[36] F. V. Paulovich, L. G. Nonato, R. Minghim, and H. Levkowitz. Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping. *IEEE Trans. Vis. Comput. Graph. (TVCG)*, 14(3):564–575, 2008.

[37] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Comp. and Appl. Math.*, 20:53–65, 1987.

[38] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. on Comp.*, 100(5):401–409, 1969.

[39] P. Schader, R. Beckmann, L. Graner, and J. Bernard. LayoutExOmizer: Interactive exploration and optimization of 2d data layouts. In *Vis., Modeling, and Vis.*, pp. 99–107, 2021.

[40] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Comput. Graph. Forum*, 34(3):201–210, 2015.

[41] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Trans. Vis. Comput. Graph. (TVCG)*, 19(12):2634–2643, 2013.

[42] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Comput. Graph. Forum*, 31(3pt4):1335–1344, 2012.

[43] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Comput. Graph. Forum*, 28(3):831–838, 2009.

[44] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[45] Z. Tian, X. Zhai, G. van Steenpaal, L. Yu, E. Dimara, M. Espadoto, and A. Telea. Quantitative and qualitative comparison of 2D and 3D projection techniques for high-dimensional data. *Information*, 12(6):239, 2021.

[46] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal Mach. Learn. Research (JMLR)*, 9(Nov):2579–2605, 2008.

[47] L. Van Der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: A comparative review. *Journal Mach. Learn. Research (JMLR)*, 10(66-71):13, 2009.

[48] J. Venna and S. Kaski. Local multidimensional scaling. *Neural Netw.*, 19(6-7):889–899, 2006.

[49] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal Mach. Learn. Research (JMLR)*, 11(2), 2010.

[50] Y. Wang, K. Feng, X. Chu, J. Zhang, C.-W. Fu, M. Sedlmair, X. Yu, and B. Chen. A perception-driven approach to supervised dimensionality reduction for visualization. *IEEE Trans. Vis. Comput. Graph. (TVCG)*, 24(5):1828–1840, 2017.

[51] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-SNE effectively. *Distill*, 2016.

[52] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proc. IEEE Inf. Vis. Symp.*, pp. 157–164, 2005.

[53] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Trans. Vis. Comput. Graph. (TVCG)*, 12(6):1363–1372, 2006.

[54] W. Xu, X. Jiang, X. Hu, and G. Li. Visualization of genetic disease-phenotype similarities by multiple maps t-SNE with Laplacian regularization. *BMC Medical Genomics*, 7(2):1–9, 2014.