This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3216919

Out of the Plane: Flower vs. Star Glyphs to Support High-Dimensional Exploration in Two-Dimensional Embeddings

Christian van Onzenoodt, Pere-Pau Vázquez, and Timo Ropinski

Abstract—Exploring high-dimensional data is a common task in many scientific disciplines. To address this task, two-dimensional embeddings, such as tSNE and UMAP, are widely used. While these determine the 2D position of data items, effectively encoding the first two dimensions, suitable visual encodings can be employed to communicate higher-dimensional features. To investigate such encodings, we have evaluated two commonly used glyph types, namely flower glyphs and star glyphs. To evaluate their capabilities for communicating higher-dimensional features in two-dimensional embeddings, we ran a large set of crowd-sourced user studies using real-world data obtained from data.gov. During these studies, participants completed a broad set of relevant tasks derived from related research. This paper describes the evaluated glyph designs, details our tasks, and the quantitative study setup before discussing the results. Finally, we will present insights and provide guidance on the choice of glyph encodings when exploring high-dimensional data.

Index Terms—Glyph visualization, high-dimensional data visualization, two-dimensional embeddings.

1 Introduction

Analyzing high-dimensional data is a common task in many scientific disciplines. To make such data comprehensible to humans, they are often embedded in two or three dimensions. Although modern dimensionality reduction (DR) techniques can produce high-quality results and conserve several high-dimensional features, many relations are lost during the embedding process [1]. Even though DR is often referred to as a visualization technique in the machine learning literature [2], we argue that they are simply visual mapping techniques. The underlying visual encoding used in these embeddings has received relatively little attention. We argue that using appropriate visual encodings could enable the communication of additional dimensions beyond the two or three dimensions of the embedding space.

This paper explores the impact of visual encodings for two-dimensional embeddings by investigating glyph-based visualization techniques. We aim to discover how additional dimensions, beyond embedding dimensions, can be effectively communicated by employing an appropriate glyph encoding. For this investigation, we focus on flower and star glyphs, as they are the most commonly used multi-dimensional glyphs [3], [4], [5]. As illustrated in Figure 1, these glyphs employ area, i.e., through petal size in flower glyphs and segment size in star glyphs, as a visual channel to communicate attributes of interest. Thus, when using these glyphs to visualize two-dimensional embeddings, their position encodes the two primary dimensions while the glyphs communicate additional dimensions. Although recent visualization systems already use these glyphs to communicate high-dimensional data, it is still unknown how effective they are in doing so [5]. Relevant questions that desire an answer in this context are, for instance: How many additional dimensions, if any, can be communicated into these glyphs? Does the glyph's shape affect the encoding of the two primary dimensions? Are there individual strengths or weaknesses for these glyphs? Based on these and other questions, we have formulated the guiding hypothesis for our research: *Flower and star glyphs support the communication of additional dimensions in two-dimensional embeddings*.

To investigate this relatively high-level and broad hypothesis, we have formulated a set of specific hypotheses and identified different tasks that we see as indicative of our goals and included in our evaluation. In their seminal work on scatterplots, Sarikaya and Gleicher describe *browsing* tasks, which are relevant when users desire an overview of an unknown dataset [6]. In contrast to tasks that focus on individual objects, the nature of browsing tasks is to search for patterns within the data (e.g., clusters and correlations), as well as patterns like properties of objects in a specific neighborhood. We believe that two-dimensional embeddings are often a starting point for such a scenario, so we deem these browsing tasks a good selection for our evaluation. Thus, to investigate the glyph designs' impact on the communication of additional dimensions beyond the two embedding dimensions, we have tasked users to identify outliers and subclusters and investigate the correlations between dimensions. So, we are not looking at individual glyphs *in* the plane; instead, our goal is to find patterns in the additional dimensions - *out of the plane*. As these are the tasks chosen to help answer our guiding research question, we refer to them as *out of the plane* tasks. To further investigate if the glyph designs used to affect the communication of the two primary dimensions are something to be avoided, we have also tasked users with identifying clusters in the

- *Christian van Onzenoodt and Timo Ropinski are with the Visual Computing Group, Ulm University. E-mail: {christian.van-onzenoodt, timo.ropinski}@uni-ulm.de.*
- *Pere-Pau Vázquez is with the ViRVIG Group, UPC Barcelona. E-mail: pere.pau.vazquez@upc.edu.*

plane. During this task, we compare flower and star glyphs with dot symbols, as they are used in standard scatterplots. Finally, we have investigated the expressivity of flower and star glyphs wrt. to the number of additional dimensions.

While, as illustrated in Figure 2, both glyphs can be used to communicate different numbers of dimensions, there is naturally an upper limit to this, as readability will be affected. For the flower glyph, this upper limit is defined by the point where adding more petals to the glyph would lead to overlap between the glyphs. Perceptually, this limit could be reached earlier, specifically when observers can no longer reliably read individual values from an individual petal. To investigate these limits for our glyphs in real-world scenarios, we tasked users with estimating averages over selected regions in the two-dimensional embedding. As this task requires aggregation over several glyphs, we consider it was challenging enough to estimate a relevant upper limit.

We selected data from the U.S. Government's open data initiative for our studies to derive insights relevant to real-world scenarios. Therefore, we have scraped all data made available from data.gov and selected data sets with at least 13 decimal attributes and at least 100 data points for our studies, leaving us with 608 real-world data sets from different domains. As stimuli, we generated flower and star glyph visualizations based on the embedding of these data. To collect many responses to these stimuli, wrt. the selected tasks, we conducted crowd-sourced user studies through Amazon's Mechanical Turk platform, where we tasked a total of 912 participants.

Our results demonstrate that additional dimensions can be effectively encoded through the aforementioned glyphs without sacrificing positional encoding. We found that flower glyphs outperform star glyphs in tasks where high-dimensional data play a role, for example, in detecting high-dimensional outliers. While quantifying individual values is possible from the glyphs, this is associated with considerable uncertainty.

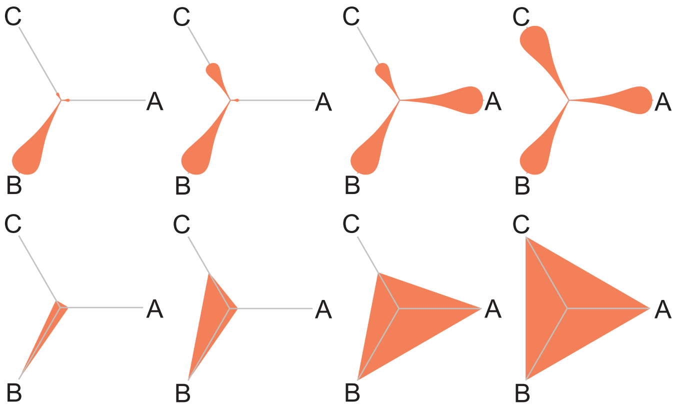

Fig. 1. Flower glyphs (*top row*) and star glyphs (*bottom row*) encoding different attribute values. Attribute A is chosen to be 10% for the first two columns and 100% for the remaining two, B always shows 100%, and C 10% for the first column, 50% for the two columns in the center, and finally 100% for the column on the right.

2 Related Work

Visualizing high-dimensional data is a common task when trying to find insights into unknown datasets. There are two main approaches: multi-dimensional visualization and dimensionality reduction techniques. The key difference is that the latter reduces the original data dimensions to a two- or three-dimensional set and uses common techniques, such as scatterplots, to visualize them.

Multidimensional Visualization Techniques. These techniques use many approaches that may spatially separate the dimensions, such as in small multiples or parallel coordinate plots, or try to present them in the same space, such as in glyph visualizations. Although related work found that such spatial separation can have advantages compared to clustering approaches [7], separating dimensions in space usually requires much room. Moreover, one of the main problems with such approaches is the difficulty in detecting patterns, such as outliers or clusters.

Therefore, a line of research in this area is finding appropriate techniques for reordering (e.g., [8], [9], [10], [11], [12]) such that the visual analysis is facilitated. Quality metrics can be calculated based on data alone or images generated. Some visual factors have been shown to have an essential influence on the perception of the data. Sedlmair et al., for example, analyzed and created a taxonomy of a large set of such factors in the context of cluster separation [13]. Crowdsourcing services, such as Mechanical Turk, have previously been shown to facilitate the successful outsourcing of tasks to large populations of users [14], and problems such as cluster detection can be addressed using this paradigm [15].

Two-Dimensional Embeddings. Two-dimensional embeddings can be used for various tasks, such as detecting clusters, correlations, or outliers in the data. Work by Sarikaya and Gleicher evaluated these tasks and categorized them into three different groups, namely *object-centric*, *browsing*, and *aggregate-level* [6]. Thus, the category *browsing* focuses on relevant tasks when exploring unknown datasets. Furthermore, perception-based studies of dimensionality reduction techniques have been carried out. For two-dimensional embeddings in the form of scatterplots, research has been devoted to determining whether projections provide separable clusters and improving them [16], [17]. Other researchers analyzed the quality of dimensionality reduction techniques [18].

Early work on perception suggests a ranking of visual channels based on the type of data [19]. For example, they found that, for quantitative data, visual attributes such as size and length can be measured and compared more accurately than color, saturation, and brightness. Then these findings were applied in the context of two-dimensional embeddings for optimized presentation [9], [20], [21], [22], [23], [24]. However, other works on data-driven quality measures do not consider human perception [25].

Glyph-Based Visualizations. Data projections come with the downside that data features can be lost. One way to preserve features within the data is to encode them in extra dimensions, e.g., by using glyphs instead of a simple dot-based encoding [5]. Glyph-based visualizations are a common visualization technique in which an individual visual object represents each data point. The dimensions of these data points are usually encoded in the visual attributes of these objects using properties such as size, shape, color, and orientation [3], [26], [27].

To the best of the authors' knowledge, little research has been carried out to explore to what extent the use of glyphs

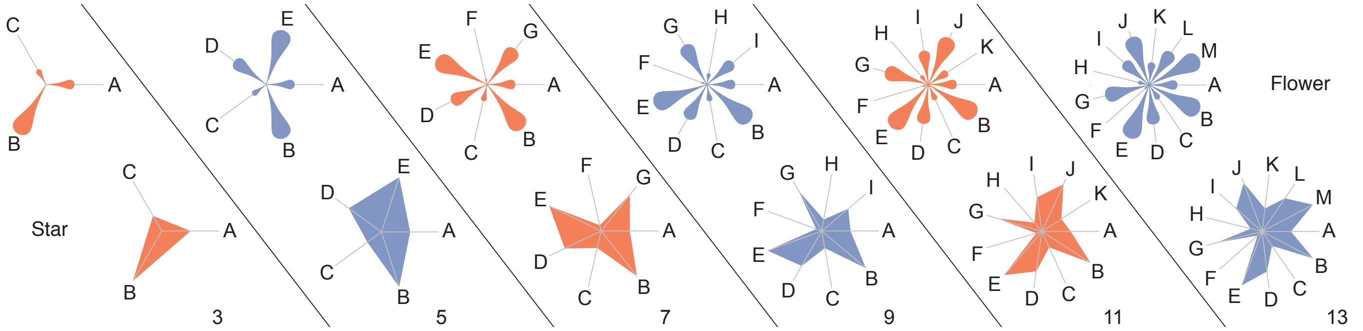

Fig. 2. Flower glyphs (*top row*) and star glyphs (*bottom row*) in both colors used (orange and blue). The glyphs encode different numbers of additional dimensions (three, five, seven, nine, eleven, and thirteen), as used in our study. In this example, dimensions are labeled A-M, whereby the respective attribute values are the same for all glyphs: $A: 50\%,\ B: 100\%,\ C: 30\%,\ D: 70\%,\ E: 100\%,\ F: 10\%,\ G: 80\%,\ H: 20\%,\ I: 60\%,\ J: 90\%,\ K: 40\%,\ L: 70\%,\ M: 100\%.$

in two-dimensional embeddings or scatterplot layouts is effective for visual communication [28]. More specifically, tasks like outlier detection and cluster counting have been explored using common spatial dimensions, but not when extra dimensions are encoded into the glyphs. Most studies in glyph perception relate to the ability to compare glyphs using similarity tasks, as in Lee et al. [29]. Fuchs et al. evaluated the perception of different glyphs [3]. They distinguished between position-based encodings (such as line graphs and bar charts) and angle-based encodings (such as star glyphs). They found that a radial layout is more effective for reading values at a specific location. Furthermore, there has been some work on designing star glyphs effectively [23], [30]. The works of Klippel et al. compared different orders of assignment of variables to glyph rays [30], [31]. Although they found some effects, the differences were less prominent than expected. There is also work by Miller et al., suggesting optimal ordering strategies depending on the task used [32]. While these papers are related to our investigations, we could not apply their findings to our work for multiple reasons. Their suggestion of creating glyphs with a distinctive spike seems easy to adopt. However, because our stimuli contain a larger number of glyphs, this is difficult since maximizing for a single spike within a specific array of variables might reduce the effect on other variables. Furthermore, when using tasks that focus on a subset of glyphs, an optimal ordering might differ depending on the subset. One last thing about reordering that we would like to mention is that an optimal order might be different depending on the task. Also, when used in a dashboard context, as shown by Kammer et al., there are usually external requirements for axis ordering and filtering [5]. As pointed out, a filtered subset might require a different ordering.

Other findings are targeted at other visual parameters, such as color, which we reserved for other aspects of the visualization. As a star glyph design guideline suggested, we added lines to our star glyphs supporting similarity judgmglyphs [23].

Keck et al. compares star glyphs and flower glyphs in tasks such as identifying extrema in regularly spaced arrangements [4]. However, they do not analyze other problems, such as cluster detection in two-dimensional embeddings. While star glyphs showed better performance in terms of solution time, participants preferred flower glyphs because of their novelty. Furthermore, Cao et al. proposed z-glyphs, a modified version of glyphs such as star graphs, optimized for outlier detection within a single chart [33].

Our work applies to high-dimensional data, visualized through a two-dimensional glyph-based embedding. There are two ways to use our approach. Firstly, by directly mapping multi-dimensional data to visual properties. If the data contains too many dimensions to apply to visual properties directly, dimensionality reduction could be applied before encoding the projected data into visual properties [2], [34]. So instead of reducing the data to two or three dimensions, we would reduce it to a higher number. In our work, we focus on the second approach.

3 General Methods

To investigate our guiding hypothesis, users had to complete the selected tasks using flower and star glyph visualizations. The following sections describe the general setup of these experiments, wrt. glyph encoding, stimuli generation, experimental design, and procedure.

3.1 Task Selection

Following our guiding hypothesis that flower and star glyphs support the communication of additional dimensions in two-dimensional embeddings, we had to decide on a set of tasks for our experiments. We identified the work by Sarikaya and Gleicher as offering a good categorization of tasks on scatterplots. This work describes three categories of tasks, namely *object-centric*, *browsing*, and *aggregate-level* [6]. As the name suggests, *object-centric* tasks are related to individual objects e.g., finding particular objects or reconciling object attributes with their spatial location. In contrast, *aggregate-level* tasks are related to tasks such as identifying the level of correlation, comparing numerosity, or understanding the relative distance between objects. So tasks related to a larger number of elements with a clear question in mind.

On the other hand, the *browsing* category describes tasks related to a set of elements but without a clear question in mind. Examples of such exploratory tasks could be to find patterns within the data, such as clusters and correlations, but also looking for unusual things (e.g., outliers). As this matches our research question of exploring high-dimensional data, we decided to use a set of tasks from this category.

We designed our studies to consist of three experiments. In the first experiment, we tried to find how many variables can be encoded into a single glyph. Subsequently, we ran an experiment to find if our glyph encodings affect positional cluster identification compared to a dot encoding. So, this second experiment, in particular, could be considered a *in plane* task since it focuses on the position within the plane. Finally, we carried out a third experiment using three tasks from the *browsing* category: outlier detection, correlation detection, and subcluster identification. Since these tasks could not be solved using the position within the plot, but by using the dimension encoded in the glyph, we consider these tasks to go *out of the plane*.

3.2 Glyph Encoding

We parameterized the two glyph types to generate visual stimuli for our experiments. As illustrated in Figure 1, in both flower and star glyphs, the extent of each axis represents the value of the encoded attribute. The full extent represents the maximum value, and the zero extents represent the minimum value for each attribute of the data point encoded by the glyph.

For star glyphs, each data dimension was encoded into a star-like shape around a center point. As said, the ray's extent represented this dimension's encoded value. Then each ray was connected to both adjacent rays, which means that the rays were not drawn independently (see Figure 1, bottom row). For the flower glyph, each pedal was drawn independently. As the encoded value grows, the pedal grows in two aspects: the length, just as for the star glyph, and the radius at the tip, resulting in a flower-like appearance.

The work of Fuchs et al. suggests that drawing the axis does support the ability to estimate values from star glyphs [23]. Therefore, we not only show the glyph itself but also depict an axis for each dimension. For the star glyph, observers could see the center of the glyph and the extent of the glyph. We support relative judgment by allowing observers to see the extent (maximum possible value). For the flower glyph, we did not need support to understand the center of the glyph because of the way this glyph was drawn. Therefore, we drew the flower on top of the axis, still allowing estimation relative to the maximum. Furthermore, this axis allowed drawing glyphs that encode small values for all dimensions. Without these axes, we would not draw anything in the extreme case with the minimum value for each dimension.

The number of dimensions of the data to visualize ranged from three to 13, as illustrated in Figure 2. Thus, each glyph encoded up to 13 dimensions beyond the two spatial dimensions. As star glyphs had to encode at least three dimensions, we also used this minimum for flower glyphs. We noticed that 13 dimensions mark an upper limit for the flower glyph before the petals start to touch each other, so we used this upper limit for both glyphs.

Klippel et al. suggested that the rays of star glyphs should be drawn using different colors to make them visually salient [30]. We still decided not to use color for the rays of our glyphs for two reasons. First, we needed to encode 13 dimensions per glyph, meaning we would need to find a set of 13 well distinguishable colors. However, finding 13 colors that ideally also work for people with color vision impairments is very complicated. We could still use an arrangement with only well-distinguishable colors next to each other. However, since our experiments did include tasks that require a target-distractor distinction, we decided to reserve color for this aspect.

3.3 Stimuli Generation

For our studies, we collected real-world datasets provided by the U.S. Government's open data initiative, *data.gov*. Through their data API, we selected and downloaded 20k data sets, which were available as CSV files for easy processing. From these data sets, we excluded those that used compression or were corrupt. The remaining data sets were analyzed per column using the Python Pandas analytics library to select those with at least 13 decimal attributes and 100 data points. Thus, we obtained a total of 608 data sets, which we kept together with their associated metadata.

We then computed the variance for each dimension depending on how many additional dimensions we wanted to visualize (three, five, seven, nine, eleven, or thirteen). We sorted the dimensions from the highest to the lowest variance and finally picked the desired number of dimensions, starting from the highest variance. We included 100 data points for each stimulus based on initial lab experiments. Therefore, we subsampled each dataset to construct our stimuli. By doing so, we ended up with a set of datasets containing 100 data points and between three and thirteen dimensions.

We then projected these attributes down to two dimensions using UMAP [1] to obtain positions. As we considered it the state-of-the-art approach, we chose UMAP over other dimensionality reduction techniques, such as tSNE [35]. For simplified further processing, we normalize the obtained positions and each of the additional data dimensions to lie in $[0.0, 1.0]$. These two positional dimensions were added to each dataset so that each dataset contained between five and 17 dimensions, two used for position, and the remaining encoded in the glyph.

As some tasks required us to differentiate between target and distractor points, we used hues from the ColorBrewer qualitative color palette [36] to encode this classification. To provide visual separability and prevent bias toward red, we chose orange for target and blue for distractor glyphs [24], [37], [38]. We only used orange for tasks without the need for distinction between target and distractor. Examples of stimuli can be seen in Figure 3.

We used Data-Driven Documents (D3) [39] to show our stimuli in a web browser. We placed the glyphs on a white canvas with 800×800 pixels with a diameter of 40 pixels. The glyphs themselves have been drawn as described in Subsection 3.2.

3.4 Hypotheses

As the title suggests, our general research question for this work could be formulated as follows: *Do flower and star glyphs support the communication of additional dimensions in two-dimensional embeddings?* To approach these questions, we formulate the following hypotheses, which are then evaluated using our selected tasks:

H1: Position Preservation. Using glyphs rather than dot encodings does not hinder the ability to decode 2D positions. Research found that sizes of elements within a scatterplot

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2022.3216919

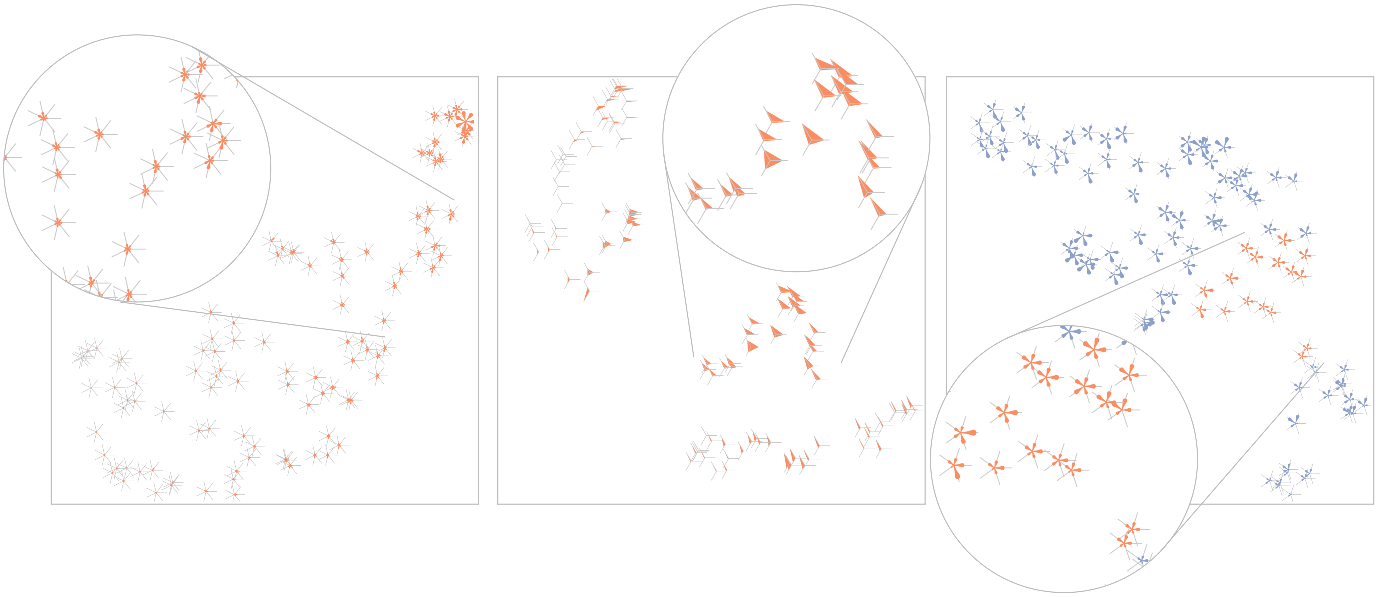

Fig. 3. Three example stimuli as used in our experiments. In our stimuli, color encodes whether a point is a target or distractor point, where orange is used for target points, and blue is used for distractors. On the *left*, we present an example from the subcluster task used in Experiment 3, encoding seven additional dimensions per glyph. The correct answer for this example would be that there *are* subclusters within the glyphs beyond the two-dimensional dimension. In the *middle*, we present an example from the cluster counting task used in Experiment 2, encoding three dimensions per glyph, forming three positional clusters. On the *right*, we present an example from the correlation task used in Experiment 3 with five dimensions per glyph. The correct answer in this example would be a correlation between dimensions *D* and *E*. As defined in Subsection 6.5, the combination of the remaining dimensions does not show a correlation.

have a stronger effect than shapes [40]. We, therefore, suspect that radial glyphs like the flower and star glyph do not hinder the ability to decode 2D position compared to dot encodings.

H2: Quantification. We suspect that observers can quantify individual values encoded in the glyphs.
Fuchs et al. found that radial layouts can be effective for reading values [3]. Also, while the Glyphboard application did not use a task related to reading values, using the application, we noticed that the glyphs allowed for it [5].

H3: Single Dimension Patterns. Due to the pattern-forming nature of these glyphs, we suspect that our glyph embeddings support the identification of high-dimensional data patterns, which are not encoded in the two embedding dimensions.
We took inspiration from the idea of the stick figure glyph [41]. Here, multi-dimensional data are encoded into connected lines, encoding values into the angle at which the lines are drawn, forming strong visual patterns. While, in contrast to the idea of the stick figure glyph, we are not uniformly distributing within the 2D domain, we still argue that our glyphs enable the identification of patterns. For example, if the value for a particular dimension is high, all glyphs seem to point in the direction encoding this dimension. Outliers that do not follow this are also perceived as strong outliers, breaking that pattern.

H4: Multi Dimension Patterns. Following H1, we suspect that glyph encodings enable the comparison of multiple values of glyphs and therefore support tasks like detecting correlations.
Our final hypothesis is somehow connected to H3, but goes beyond that. We suspect that it is possible to identify patterns within one extra dimension and patterns within multiple dimensions. Using parallel coordinates plots allows for identifying patterns like correlations between adjacent axes. While glyphs do not scale the same for the number of data points as parallel coordinates, we still suspect that glyphs allow for finding patterns like correlations between multiple dimensions.

3.5 Experimental Design and Procedure

Since our experiments involved many stimuli, we decided to divide the entire set of stimuli to ensure the participants' motivation. In previous crowd-sourced experiments, we found that participants tended to jump off studies if they exceeded around 20 minutes. Besides that, we think it is difficult to stay focused for longer, especially when completing repetitive tasks, as used in our experiments. Each task was therefore conducted as a within-subject design wrt. the glyphs and a between-subject design for the number of additional dimensions. So, each participant was confronted with both glyphs, each representing one of our additional number of dimensions conditions. We presented the stimuli in a randomized order to prevent learning and fatigue effects. For all experiments, we measured accuracy and response time.

Each experiment followed the same general procedure. We first presented a welcome screen showing an example visualization before introducing the glyphs using examples similar to those presented in Figure 2. Then, we showed an example stimulus, as shown in Figure 3. This stimulus was shown together with the glyph legend (top) and a response area (bottom), reassembling the user interface. We also displayed stimuli to introduce the concept of a target area if necessary. Subsequently, we presented the actual task using multiple examples drawn from different stimuli than those used in the study. The participants received instructions on responding to the stimuli before proceeding with the training phase. During this training phase, the participants got comfortable with the procedure and received feedback on their responses, which further helped them understand the task. In this training phase, we also provided feedback on participants' responses. If they did not respond correctly, we informed them about this, and the participants would select another answer until they did. For tasks involving a target area, we informed them if they did select a glyph outside the target area, saying they needed to select a glyph from within the target area. For other incorrect answers, we replied that they did not respond correctly and should try again. When they responded correctly, the participants were informed of the correct response and could see their response to understand why this answer was correct. After this training phase, we informed the participants that the study was about to start. The participants then completed all the stimuli in the study. Then, they completed a demographic questionnaire to complete the study. Supplemental material contains screenshots of each of the tasks and their introduction. With this design, we could keep each participant's time at 15 minutes. The effort of the participants was rewarded with a target rate of € 5 per hour.

3.6 Evaluation

We used different statistical tests for our analysis. We used the t-test and Wilcoxon signed-rank test for pairwise comparison, depending on the Levene and Shapiro-Wilk pretest. For the evaluation of our between-subject condition (number of dimensions per glyph), we used the Kruskal-Wallis test. Post hoc pairwise comparison was then performed using the Mann-Whitney rank test with Bonferroni correction applied. Because of the within-subject design, a comparison between the glyphs was made using Friedman's ANOVA and Nemenyi post hoc for pairwise comparison.

To ensure data quality (e.g., detect click-through), we had to exclude participants from our experiments. Depending on the experimental design, we used different exclusion criteria. Generally, however, participants needed to meet the criteria to outperform the chance level, e.g., for force choice experiments and did not show unusual fast response times.

4 Experiment 1: Number of Dimensions

One essential part of our guiding hypothesis was determining if flower and star glyphs could effectively encode additional dimensions beyond the two-dimensional plane. Furthermore, if so, up to what number of dimensions? To investigate this, we conducted the following experiment in which we varied the number of additional dimensions to be encoded.

Task Selection. For this experiment, we needed to balance task difficulty so that we would not end up with a trivial maximum number of dimensions. Among the *browsing* tasks described by Sarikaya and Gleicher, some tasks allow observers to explore the properties of data points in a given neighborhood to form aggregates [6]. As these tasks are not focused on a single data point or require assessing the entirety of data points, they allow for an adequate balance of task difficulty. Therefore, we tasked participants to *estimate the average value of one attribute from all glyphs within a given region* to determine how many dimensions can be communicated effectively. So participants needed to see which attribute was asked for in the example and then look at all glyphs within a region to estimate the average (mean) value for the given attribute. Figure 3 (*right*) shows an example stimulus from this experiment, where orange glyphs indicate the target area for the average estimate.

Stimuli Generation. To generate a target region needed for this task, we picked a random point from our data points. To avoid too sparse regions, we ensured that this point is adjacent to ten to twenty data points within a .2 radius of our normalized positions. If our sparseness criterion was not met, we discarded and picked a new point until it was met. The thus obtained data point, together with its neighbors in the .2 radius, was then used as a target area based on which the participants had to estimate the average value.

Experimental Design. We used 30 data sets for each dimension (three, five, seven, nine, eleven, and thirteen), and each data set was presented using flower and star glyphs. We confined ourselves to a maximum of thirteen dimensions, as we found that for flower glyphs, readability beyond that is hampered by petal overlap. By using both glyphs (2), the number of additional dimensions (6), and 30 datasets per number of dimensions, we generated a total of 360 stimuli for this task. As described in Subsection 3.5, we used a within-subject design wrt. glyph designs and a between-subject design for the number of additional dimensions.

Procedure. We followed the general procedure described in Subsection 3.5. We also indicated the additional dimension to estimate the average below the visualization area. We showed a slider ranging from 0 to 100 to obtain the average value of the replies. Once the participants were confident in their input, they confirmed their selection by pressing a button to continue to the next stimulus.

Evaluation. We recruited a total of 317 participants for this experiment. 52, 57, 55, 53, 49, 51 for three, five, seven, nine, eleven, and thirteen additional dimensions per glyph, respectively. We randomly excluded 23 people from a subset of groups to produce equally sized groups. Therefore, we present the results of 49 participants per condition for a total of 294 participants (102 female, 191 male, 1 did not report, $M_{age} = 33.89$, $SD = 9.42$).

We calculated accuracy using the absolute offset between the participants' responses and the actual value. Thus, all offset values presented in the following evaluation are offset in units from the original values of the glyphs (0 – 100).

When analyzing different numbers of additional dimensions encoded into the glyphs, we found significant effects on accuracy. As expected, we found a significant decrease in accuracy with an increasing number of dimensions (three ($Mdn = 17.31$, $IQR = 22.47$), five ($Mdn = 16.66$, $IQR = 21.91$), seven ($Mdn = 18.59$, $IQR = 24.23$), nine ($Mdn = 21.94$, $IQR = 28.79$), eleven ($Mdn = 21.39$, $IQR = 28.74$), thirteen ($Mdn = 22.53$, $IQR = 30.57$, $H(6) = 240.38$, $p < .001$)). During the post hoc analysis, we found two groups. Three, five, and seven additional dimensions form the first group. Each condition in this group significantly affected the remaining conditions (nine, eleven, and thirteen; $p < .001$) that formed the second

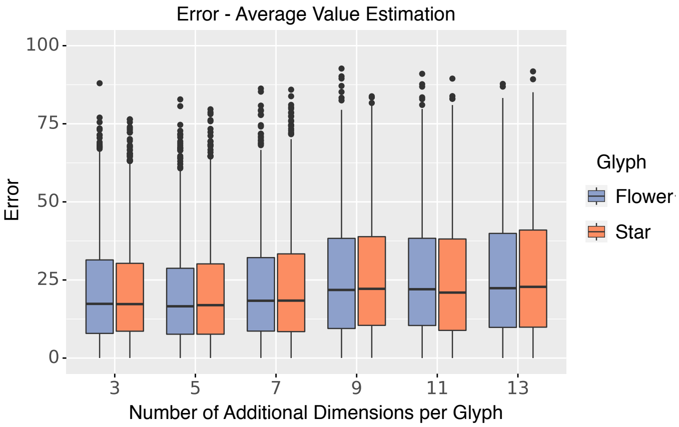

Fig. 4. Boxplot shows accuracy values for estimating the average value by condition and the number of additional dimensions. Accuracy is measured by calculating the absolute offset between the actual value of the target dimension and the estimate of the participants; therefore, lower values are better. As we can see, the accuracy decreases with an increasing number of dimensions. This finding is also consistent for both flower and star glyphs.

group. Within the first group, we found a significant effect between the conditions (three ↔ five ($p < .05$), three ↔ seven ($p < .05$), five ↔ seven ($p < .001$)). We did not find significant differences within the group with higher numbers of dimensions (nine, eleven, thirteen).

Figure 4 shows a boxplot of the accuracy per condition and the number of additional dimensions for this experiment. When comparing our glyphs directly, we could not find a significant effect. Not when comparing flowers vs. stars over all additional dimensions nor when comparing conditions directly (e.g., flowers using three additional dimensions vs. stars using three additional dimensions).

Results. We conclude that estimating the average value using glyphs is possible based on the results obtained, supporting our hypothesis H2. However, it comes with relatively large uncertainty, which grows with the number of encoded dimensions. Using seven dimensions per glyph seems to mark some threshold here since we found interesting significant effects between the lower dimensional group (three, five, and seven) and the rest (nine, eleven, and thirteen). We did not find a significant decrease in accuracy beyond seven dimensions per glyph. Since our experiments already included numerous variables, we decided to limit the remaining experiments to the first three conditions (three, five, and seven dimensions per glyph), keeping our experiments at a reasonable scale.

5 Experiment 2: Positional Encoding

Although Experiment 1 indicated that flower and star glyphs could communicate additional dimensions beyond position, we also wanted to evaluate the impact of these glyphs wrt. communicating these spatial dimensions. To investigate whether using these glyphs affects communicating the positional encoding, we compared the two glyphs against a baseline given by dots as used in standard scatterplots.

Task Selection. As with scatterplots, identifying clusters is essential when studying two-dimensional embeddings. Thus, we consider the results of a cluster counting task as a good indicator to compare the performance of glyphs and dots for in-the-plane tasks. Therefore, users had to *specify the number of positional clusters* in this experiment. Figure 3 (*middle*) shows an example of a stimulus as used in this experiment.

Stimuli Generation. To generate stimuli for this task, we use the approach described in Subsection 3.3. We used all data sets that contained at least 100 data points and at least 13 dimensions. We then subsample 100 data points from each data set so that each contains the same number of data points. These points are then projected to two dimensions using UMAP for the position.

Afterward, we used DBSCAN on these two positional dimensions (generated by UMAP) to compute class labels for each data point. If a data point was labeled as an outlier, we did run this process from the start again because we would rather not include outliers but also maintain a constant number of 100 data points for each stimulus. Finally, we used the class labels to compute the number of positional labels of the given dataset. DBSCAN hyperparameters for this experiment have been chosen based on internal piloting. Since this experiment aimed to determine whether our glyph encodings do impact cluster perception compared to standard scatterplot encodings, we argue that this approach is a suitable solution for this experiment.

Experimental Design. To evaluate participants' ability to detect positional clusters within the visualization, we decided to evaluate a range of one to five positional clusters. We decided to use four stimuli per number of positional clusters for a total of 20 datasets per condition. We used three, five, and seven additional dimensions encoded per glyph, in line with our findings from Experiment 1. These stimuli were presented using flowers, stars, and simple dots as a baseline. Thus, we used both glyphs, and dots (3), our number of different additional dimensions per glyph (3), different numbers of clusters (5), and repetitions of each of these conditions (4) to generate a total of 180 stimuli for this task.

Procedure. To input how many clusters of points the participants could detect, they had to respond using a drop-down menu below the visualization area. Once participants were comfortable with their choice, they could confirm and continue with the next stimulus by pressing a button.

Evaluation. We recruited 166 participants for this experiment, 54 for three additional dimensions, 53 for five additional dimensions, and 59 for the seven additional dimensions per glyph condition. In this experiment, participants had to decide between one of five possible responses (one to five positional clusters). We found a total of 39 participants who could not achieve a mean accuracy greater than 20%, meaning these participants were worse than the chance level. This exclusion criterion resulted in a rather large group of participants that we needed to exclude. We suspect this to happen because of the subjective understanding of what defines a cluster. While one might think of two groups of points that are close together as two clusters, others see the points as a single cluster. We found the same disagreement in our prestudy for Experiment 3, as outlined in Subsection 6.2.

Thus, we had to exclude nine participants from the three additional dimensions, 14 from the five additional dimensions, and 16 from the seven additional dimensions per glyph condition. To achieve equal groups between these conditions, we randomly excluded ten participants.

Therefore, we present the results of 39 participants per condition for a total of 117 participants (33 female, 84 male, $M_{age} = 36.33$, $SD = 9.73$). Figure 5 shows a boxplot of the results of the cluster counting task.

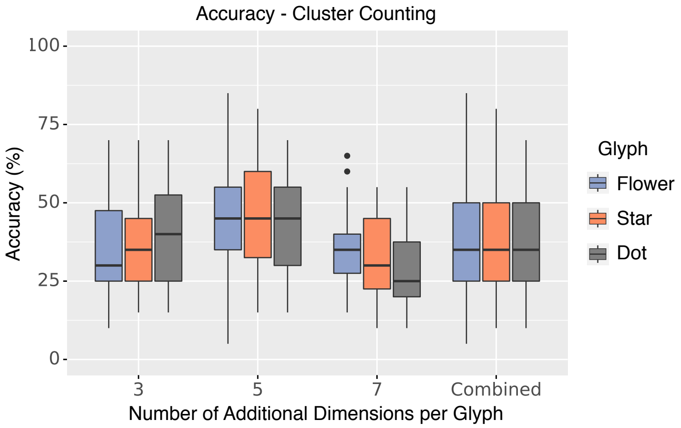

Fig. 5. Boxplot of accuracies per condition and number of additional dimensions for the cluster counting task. Accuracies are measured as the mean accuracy per participant for each condition; therefore, higher values are better. From this plot, we can see those different glyphs are on par in terms of accuracy compared to the baseline dot encoding.

Comparing the three visual encodings, we did not find a significant effect on accuracy between flowers, stars, and dots. However, we did find effects on response times between the glyphs (flowers ($Mdn = 4.86s$, $IQR = 4.45s$), stars ($Mdn = 4.8s$, $IQR = 4.07s$), dots ($Mdn = 4.23s$, $IQR = 3.36s$, $\chi^2(2) = 129.28, p < .001$)). During post hoc analysis, we found that dots allow for significantly faster responses compared to other encodings (flowers ↔ stars ($p = .08$), flowers ↔ dots ($p < .001$), stars ↔ dots ($p < .001$)).

When investigating different number of dimensions encoded into the glyphs, we found significant effects between three ($Mdn = 35.0\%$, $IQR = 25.0\%$), five ($Mdn = 45.0\%$, $IQR = 25.0\%$), and seven ($Mdn = 30.0\%$, $IQR = 15.0\%$, $\chi^2(2) = 31.93, p < .001$). Here, five additional dimensions are seemingly more accurate than the others (three ↔ five ($p < .001$), five ↔ nine ($p < .001$), three ↔ nine ($p = .052$)).

We also compared the glyphs under the individual dimension conditions. For conditions using three and five additional dimensions, we could not find a significant effect (three ($p = .33$), five ($p = .14$)). However, for seven dimensions per glyph, we did find a significant effect (flowers ($Mdn = 35.0\%$, $IQR = 12.5\%$), stars ($Mdn = 30.0\%$, $IQR = 22.5\%$), dots ($Mdn = 25.0\%$, $IQR = 17.5\%$, $\chi^2(2) = 11.75, p < .01$)). During post hoc analysis, we found a significant effect between the flower and the dot condition ($p < .01$) in favor of the flower glyph.

We also analyzed the response times between the glyphs for the individual number of dimension conditions, revealing the same effect as in the overall analysis. So for three, five, and seven additional dimensions we found a significant effect between the glyphs (three: $\chi^2(2) = 47.43$, five: $\chi^2(2) = 41.85$, seven: $\chi^2(2) = 44.49$, $p < .001$, each) and post hoc analysis revealed that dots allow for significantly faster responses, compared to the others ($p < .001$, each), while there was no effect between flowers and stars.

Results. From this task, we conclude that using glyphs in two-dimensional embeddings does not hinder perception compared to baseline dot encodings, supporting hypothesis H1. Response times have shown to be significantly faster when using dots compared to the glyph-based encodings, and flower glyphs could even outperform dot encodings in the seven additional dimensions condition. We suspect the faster response times happen due to the reduced visual clutter of dots compared to the glyphs. The higher accuracy of flowers when using more dimensions (and therefore more petals) could be due to the stronger visual appearance. Further investigation of this effect might be an interesting future research direction.

6 Experiment 3: Out of the Plane Tasks

In our main experiment, we wanted to investigate to what extent flower and star glyphs can facilitate the analysis of additional dimensions in two-dimensional embeddings. Therefore, we chose three representatives from the plane tasks, whereby decisions had to be made based on the data dimensions beyond the two embedding dimensions.

6.1 Task Selection

The three tasks included in this experiment are also based on the *browsing* tasks described by Sarikaya and Gleicher [6]. While we had to confine ourselves to a manageable number of tasks, we wanted at the same time to cover different varieties of interpreting glyphs. Thus, we identified the following three tasks, which span from single glyph readings to comparing groups of glyphs and detecting correlations among the additional dimensions.

Outlier Detection. During the outlier task, participants were asked to find an outlier based on the data encoded in the glyph. Hence, the data points stand out from the other glyphs, despite the two-dimensional position. So, one way to solve this is to find a glyph within the others that does not follow the general pattern of the other glyphs — going *out of the plane*.

Subcluster Detection. While dimensionality reduction techniques provide good separability when clustering data into larger groups, detecting subclusters within positional clusters is difficult. Therefore, we included this task to determine whether a subcluster can be found within a larger positional cluster. These subclusters are groups of data points that share some pattern in the additional dimensions encoded into the glyphs.

Correlation Detection. Here, observers are tasked to find correlations in the additional data dimensions. Since this is already a rather complex task, we decided to restrict the task to a target region, just as for the average value estimation task.

6.2 Stimuli Generation

All experiment stimuli were created using the same methods described in Subsection 3.3. Again, we limited the maximum number of additional dimensions to seven, in line with our findings from Experiment 1. In the following, we describe individual differences, particularly the computation of the target value (correct answer).

Outlier Detection. For the outlier task, we computed the Local Outlier Factor (LOF) [42] for all our data points, using the respective attributes presented in the stimuli (three, five, or seven). Of all the data points, thus marked as an outlier, we took the one with the largest negative outlier factor, i.e., the strongest outlier determined by LOF. Since almost all the datasets showed a larger number of outliers and the visual search process could take some time (which might frustrate participants), we decided to indicate a target area around the strongest outlier. To prevent this target outlier from constantly being in the center of this target region, we picked a random point within the radius of 0.2 around the outlier (within the normalized position as lying between 0 and 1). We used this random point as a new center for the target area. Thus, we could further ensure that there are between ten and twenty points within the target area. In cases where these conditions were not met, we selected a new random point and iterated until a region with the desired properties was found. If it was impossible to find a new center despite the fact that all points in this neighborhood had been tested, we rejected this dataset.

Subcluster Detection. For the stimuli used in the subcluster detection task, we followed a similar approach as used for the cluster counting task in Experiment 2. We computed class labels used for the position using DBSCAN. We only considered datasets where DBSCAN found a single cluster and ensured that the cluster lies in the additional dimensions rather than the spatial dimensions. Subsequently, we applied DBSCAN to the additional dimensions. Since this task requires visual inspection of the complete plot and is rather complex, we decided to use a binary forced-choice in this experiment: *Do the glyphs split into groups, yes, or no?*. We only picked datasets with no clustering or two clusters within the additional dimensions.

Correlation Detection. For the correlation detection task, we computed *Pearson correlation coefficient* for each combination of variables for glyphs within a marked region.

Stimuli Validation. We used UMAP to generate the positions of the glyphs, whereby one of the main features of UMAP is to build clusters. However, since our subcluster detection task focused on high-dimensional clusters, we had to ensure that our stimuli did not form strong visual clusters in 2D screen space. To mitigate this risk, we filtered the stimuli as follows.

First, we tried to use algorithms to check for positional clusters. However, choosing a good set of hyperparameters for these algorithms is challenging. Therefore, we used DBSCAN for every dataset, with every possible value of hyperparameters. This approach should allow us to find datasets without clusters in position, regardless of the choice of hyperparameters. Unfortunately, we ended up with an empty set using this approach. With a combination of large epsilons for the considered environment and a few samples required to define a set of points as a cluster, DBSCAN was ”seeing” clusters in all our datasets.

Consequently, we conducted a crowd-sourced user study to facilitate stimuli filtering. For this study, we rendered all datasets using a simple dot encoding (similar to a scatterplot). We were aware that the perception of clusters is subjective, so we suspected disagreement among participants on this task. To account for this, we did not use a binary forced-choice question such as *Do the points form clusters?* but decided to use a triplet-based ranking approach. We presented our images in a way that allowed participants to rank three of the images from *clustered* to *unclustered*. By doing so, we suspected that we would be able to rank our datasets based on how clustered they are or how evenly distributed the points are. We could not find a stable ranking using this approach due to disagreement between the participants. These results, however, show that there is no strong agreement on whether points form clusters in image space, as we have already suspected. As a consequence of this finding, we argued that our datasets are suitable for our high-dimensional subcluster detection task. If participants cannot detect clusters in positions but can complete our subcluster task, the clustering must be communicated through the glyphs’ shapes rather than their position.

So to finally decide on a set of datasets for our study that also met our requirement of having the same number of stimuli for each condition, two domain experts carefully evaluated the generated stimuli. Based on their inter-observer results, we only included data sets where they could not spot positional clusters.

6.3 Experimental Designs

We used our glyphs (flowers and stars) for all tasks in this experiment. For the correlation detection task and the outlier detection task, we used 30 stimuli for each of the additional numbers of dimensions (three, five, and seven). Due to the more restrictive rule in generating stimuli for the subcluster task, as described above, we could only select 25 stimuli per number of additional dimensions per data point for the subcluster detection task.

We used our glyphs (2), three levels of additional dimensions (3), and 30 datasets for the outlier and correlation tasks to generate 180 stimuli. For the subcluster detection task, we generated 150 stimuli using our glyphs (2), three levels of additional dimensions (3), and 25 datasets.

6.4 Procedure

As in previous experiments, the tasks in this experiment followed the same general procedure as described in Subsection 3.5. In this section, we outline the individual differences for the respective tasks.

Outlier Detection. When detecting outliers, participants simply had to click on the outlier within the orange target region and confirm their choice by clicking on a button below the visualization.

Subcluster Detection. For subcluster detection, we presented all glyphs using the same orange color while participants had to decide whether the presented glyphs were divided into subclusters depending on their shape. Since the classes do not necessarily divide spatially, we decided not to use an interaction method based on selecting points within a region or the like. Instead, we decided that our task should follow a forced-choice setup, i.e., participants had to judge whether the presented glyphs divide into groups or not. Therefore, we also presented two radio buttons to respond to the stimuli.

Correlation Detection. As described in Subsection 6.2, we indicated a target area of glyphs from which the participants had to solve the task. Within these glyphs in the target area, participants had to decide if two target attributes correlated. The two asked attributes are shown below the visualization.

Participants could respond using two radio buttons, one for *correlation*, the other for *no correlation*.

6.5 Evaluation

In the following, we evaluate the results for all tasks.

Outlier Detection. We recruited a total of 151 participants for this task, 52 for three additional dimensions, 46 for five additional dimensions, and 52 for seven additional dimensions per glyph condition. Because participants either repeatedly chose points outside the target area, indicating that they did not understand the task, or they had mean response times below 200ms, indicating click-through, we had to exclude 19 participants. We randomly excluded four participants to achieve equally sized groups between these conditions. Consequently, we present the results of 41 participants per condition for a total of 123 participants (47 female, 75 male, 1 did not report, $M_{age} = 31.76$, $SD = 9.56$). Figure 6 shows a boxplot for accuracy per glyph and the number of additional dimensions for this task.

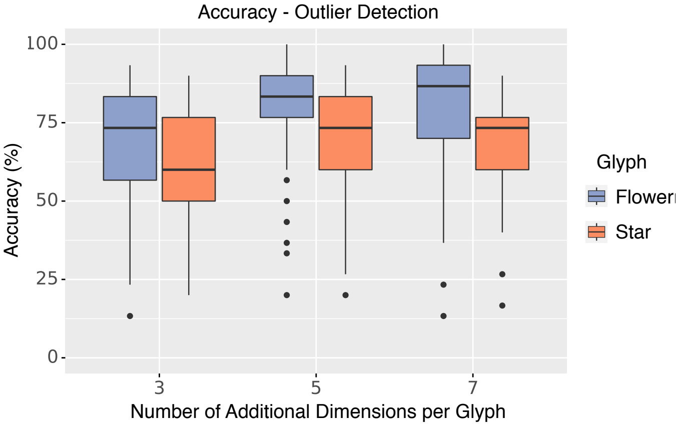

Fig. 6. Boxplot of accuracies per glyph and number of additional dimensions for the outlier detection task. Accuracies were measured as the mean accuracy per participant for each condition, so higher values are better. Here, we can see that flower glyphs show higher accuracies during this task across different numbers of dimensions per glyph. Furthermore, accuracy seems to increase with more dimensions per glyph.

We found that participants could identify outliers reliably, based on our real-world datasets ($Mdn = 76.67\%$, $IQR = 26.67\%$). When comparing our glyphs using t-test, we found a significantly higher accuracy when using the flower glyph ($Mdn = 83.33\%$, $IQR = 23.33\%$) compared to the star glyph ($Mdn = 73.33\%$, $IQR = 23.33\%$, $t(245) = 8.82$, $p < .001$, $r = .48$). However, by comparing the glyphs in terms of response times, we did not find a significant effect between flower ($Mdn = .8s$, $IQR = .6s$) and star glyphs ($Mdn = .8s$, $IQR = .58s$, $p = .23$, $r = .99$), indicating that none of the glyphs appears to be pre-attentive in the tested setups. We think that accuracy is the most important factor for this task, and we found that the response times in this task were generally low.

While analyzing the dimensions encoded in the glyph, we found an interesting effect of the increasing number of additional dimensions. Here accuracy even increased with additional dimensions (three: $Mdn = 66.67\%$, $IQR = 30\%$, five: $Mdn = 83.33\%$, $IQR = 22.5\%$, seven: $Mdn = 76.67\%$, $IQR = 25\%$, $H(3) = 20.16$, $p < .001$). During post hoc analysis, we found this to be significant between three and the remaining conditions ($p < .001$), while it was not significant between five and seven ($p = .52$).

As with accuracy, we found a similar effect on response times when the number of additional dimensions increases. Here, five additional dimensions showed the fastest response times ($Mdn = .83s$, $IQR = .62s$), followed by seven ($Mdn = .75s$, $IQR = .52s$), and three ($Mdn = .81s$, $IQR = .61s$). We found this effect to be significant ($H(3) = 53.04$, $p < .001$), and post hoc analysis revealed that this is the case between all conditions ($p < .01$ each).

Subcluster Detection. For this task, we recruited a total of 160 participants, 57 for three additional dimensions, 53 for five additional dimensions, and 50 for the seven additional dimensions per glyph condition. 17 participants could not achieve the chance level for this binary choice experiment and have therefore been excluded from this experiment. To achieve equal-sized groups between these conditions, we randomly excluded four participants. Therefore, we present the results of 46 participants per condition for a total of 138 participants (48 female, 88 male, 2 other, $M_{age} = 34.6$, $SD = 12.13$). The boxplot in Figure 7 shows the results obtained.

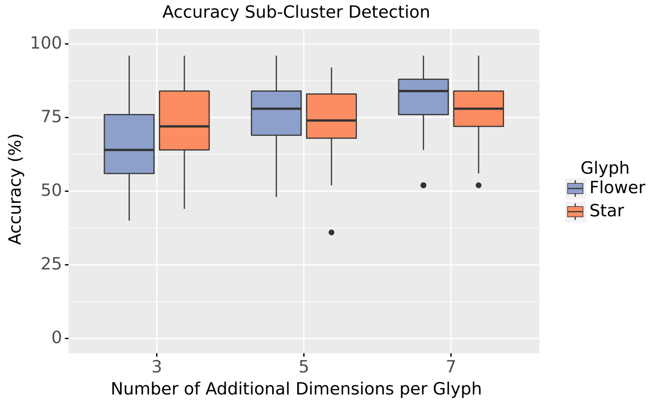

Fig. 7. Boxplot of accuracies per condition and number of dimensions for the subcluster detection task. Accuracies are measured as the mean accuracy per participant for each condition, so higher values are better. From this plot, we can see that the star glyphs seem consistent with accuracy as the number of dimensions increases, while the flower glyph seems to benefit from more dimensions.

While we did find significantly faster response times for the star glyph compared to the flower glyph, we did not find an effect between the two glyphs in terms of accuracy (flower: $Mdn = 76.0\%$, $IQR = 20.0\%$; stars: $Mdn = 76.0\%$, $IQR = 16.0\%$, $p = .76\%$).

When analyzing different numbers of dimensions encoded into the glyphs, we found a significant effect for this condition. We found an increasing accuracy for three ($Mdn = 68.0\%$, $IQR = 21.0\%$), five ($Mdn = 76.0\%$, $IQR = 16.0\%$) and seven ($Mdn = 80.0\%$, $IQR = 13.0\%$), with significant effects between these conditions ($H(3) = 24.01$, $p < .001$). Post hoc analysis found significant effects between each of these conditions (three ↔ five ($p < .01$), five ↔ seven ($p < .05$), three ↔ seven ($p < .001$)).

When analyzing response times, we did not find a clear trend with an increasing number of additional dimensions. Here, using three dimensions per glyph ($Mdn = 3.83s$,

$IQR = 3.47s$) showed the fastest response times, followed by seven ($Mdn = 3.94s$, $IQR = 4.56s$) and five ($Mdn = 4.16s$, $IQR = 3.92s$; $H(3) = 42.64$, $p < .001$) additional dimensions. Using post hoc analysis, we found these effects to be significant between all conditions ($p < .01$ each).

When analyzing our glyphs as the number of dimensions increases, we found a significant effect on accuracy for the flower glyph (three: $Mdn = 64.0\%$, $IQR = 20.0\%$; five: $Mdn = 78.0\%$, $IQR = 15.0\%$; seven: $Mdn = 84.0\%$, $IQR = 14.0\%$; $H(3) = 26.29$, $p < .001$). Post hoc, we found significant effects between all these conditions ($p < .05$ each). However, we could not find a significant difference between the number of dimensions when using the star glyphs.

Finally, we compared our glyphs for each number of dimensions. Here we found significant effects of the glyphs for each number of additional dimensions. For three additional dimensions, star glyphs ($Mdn = 72.0\%$, $IQR = 20.0\%$) showed higher accuracy compared to the flower glyph ($Mdn = 63.0\%$, $IQR = 20.0\%$; $t(91) = -3.48$, $p < .01$, $r = .34$). However, for all remaining dimensions, the flower glyph showed higher accuracies (five / flowers ($Mdn = 78.0\%$, $IQR = 15.0\%$) ↔ five / stars ($Mdn = 74.0\%$, $IQR = 15.0\%$; $p = .13$) and seven / flowers ($Mdn = 84.0\%$, $IQR = 12.0\%$) ↔ seven / stars ($Mdn = 78.0\%$, $IQR = 12.0\%$; $t(91) = 2.34$, $p < .05$, $r = .24$)), the effect being significant between flowers and stars in the condition of seven additional dimensions.

Correlation Detection. For the correlation detection task, we recruited 118 participants, 39, 42, and 37, for conditions of three, five, and seven additional dimensions, respectively. In this experiment, we excluded eight participants who could not exceed 30% of correct responses for this task. Although this experiment was designed as a binary choice experiment, we did not use the 50% chance level due to the way we analyzed the data. Because this task is already complex, we decided to only use cases with strong correlation or cases where there is no correlation. Therefore, we adjusted this threshold level for this task. Furthermore, we had to randomly exclude two participants from achieving groups of equal size between the conditions. Therefore, we present the results of 36 participants per condition for a total of 108 participants (36 female, 71 male, $M_{age} = 34.16$, $SD = 7.53$). The boxplot in Figure 8 shows the results obtained.

As described in Subsection 6.2, we computed the degree of correlation using *Pearson correlation coefficient*. Here, the correlation was quantified from -1 (strong negative correlation) to 1 (strong positive correlation), where 0 means no correlation. Participants were tasked with a 2-alternative forced choice, either *Correlation* or *No Correlation*.

As a first step into the evaluation, we focused on responses to stimuli with clear correlations (either negative or positive). We defined the correlation values calculated by *Pearson correlation coefficient* greater than .8 or smaller than $-.8$ as strong correlations. Here, we already found a rather low accuracy with high uncertainty ($Mdn = 66.67\%$, $IQR = 50.0\%$).

When comparing our glyphs, we found a higher accuracy using the flower glyph ($Mdn = 71.43\%$, $IQR = 50.0\%$), compared to the star glyph ($Mdn = 66.67\%$, $IQR = 50.0\%$, however, this effect was not significant ($p = .9$). Although we found a decrease in precision during an increasing number of dimensions (three ($Mdn = 80.0\%$, $IQR = 50.0\%$), five ($Mdn = 73.21\%$, $IQR = 42.86\%$), seven ($Mdn = 66.67\%$, $IQR = 35.71\%$; $p = .19$)), this effect was also not significant. However, during the analysis of an increasing number of dimensions, we found a significant effect on response times three ($Mdn = 2.88s$, $IQR = 6.33s$), five ($Mdn = 2.43s$, $IQR = 3.32s$), and seven ($Mdn = 2.4s$, $IQR = 3.93s$; $H(3) = 12.82$, $p = .01$), however, without showing a trend. Post hoc analysis confirmed this interesting finding that the condition of three additional dimensions is significantly slower than the remaining conditions (three ↔ five ($p < .001$), three ↔ seven ($p < .01$), and five ↔ seven ($p = .8$)).

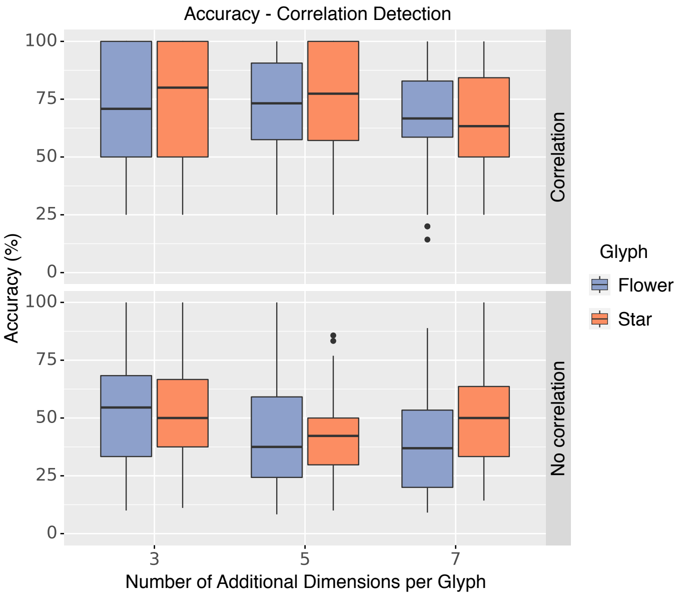

Fig. 8. Boxplot of accuracies per condition and number of dimensions for the correlation detection task. Here we distinguished between detecting true correlations and detecting if there is no correlation. Accuracies were measured as the mean accuracy per participant for each condition, so higher values are better. From these boxplots, we can see that participants can reliably detect correlations. However, participants seem unable to detect cases without correlation reliably.

Furthermore, we analyzed whether our participants could detect cases without correlation. Therefore, we considered stimuli with a correlation index between $-.2$ and .2. Here, we found the accuracy at the chance level ($Mdn = 45.45\%$, $IQR = 32.79\%$). Again, we could not find a significant effect on accuracy or response time when trying to detect that the glyphs do not show a correlation.

Unlike correlation detection stimuli, we found a significant effect on accuracy with an increasing number of dimensions. Here, the condition of three additional dimensions showed the highest accuracy (three ($Mdn = 50.0\%$, $IQR = 30.02\%$), five ($Mdn = 40.0\%$, $IQR = 31.18\%$), seven ($Mdn = 42.86\%$, $IQR = 35.11\%$; $H(3) = 8.2$, $p = .05$)), also supported by post hoc analysis (three ↔ five ($p < .01$), three ↔ seven ($p < .05$), five ↔ seven ($p = .47$)). As with strong correlations, we found the same effect on response times (three ($Mdn = 3.53s$, $IQR = 5.82s$), five ($Mdn = 2.78s$, $IQR = 3.79s$), seven ($Mdn = 2.5s$, $IQR = 4.37s$; $H(3) = 51.9$, $p < .001$)). Post hoc also confirmed the same effect of larger response times for the three dimensions per glyph condition (three ↔ five and three ↔ seven ($p < .001$, each), five ↔ seven ($p = .33$)).

We found the effect of higher accuracy for the perception of correlation versus the perception of cases without correlation to be significant (correlation ($median: 66.67\%$, $iqr: 50.0\%$), without correlation ($median: 45.45\%$, $iqr: 32.79\%$; $df(397) = 2981.0$, $p < .001$, $r = .99$)).

6.6 Results

From the *out of the plane* tasks, we summarize the following:
Outlier Detection. We found that flower glyphs support outlier detection in two-dimensional embeddings in this task. Even increasing the number of dimensions does not affect the accuracy or response time in outlier detection, possibly due to a stronger pattern effect, supporting our hypothesis H3.
Subcluster Detection. We argue that glyphs enable subcluster detection in additional dimensions based on our results. As with the outlier task, we again found that the accuracy even increased with an increasing number of additional dimensions for both flower and star glyphs. However, as the number of dimensions increases, the flower glyph shows higher accuracy than the star glyph, confirming our finding from the outlier task and further supporting H3.
Correlation Detection. Despite their high uncertainty, we believe that glyphs can be used to discover correlations in two-dimensional embeddings. We could not find a significant effect of individual glyphs during this task. Flower glyphs showed slightly higher accuracy, but increasing the number of dimensions affected this task's accuracy. Although this effect was not significant for correlation detection and thus rejecting H4, we still found a trend.

One interesting observation we would like to point out is that while using a low number of additional dimensions generally showed the highest accuracy; it also showed the slowest response times.

7 Discussion & Implications

This section discusses the observations made for different conditions and wrt. our hypotheses. Based on these observations, we further formulate our lessons learned.

7.1 Condition Observations

While our investigations did focus on the feasibility of glyphs for communicating additional attributes in two-dimensional embeddings and the influence of the increasing number of dimensions encoded per glyph, we also investigated the individual strengths of flower and star glyphs per task.
Number of Dimensions. We aim to explore how many additional dimensions can be communicated using glyphs in a two-dimensional embedding in Experiment 1. We discovered that seven dimensions per glyph mark an interesting point beyond which accuracy decreases.

From Experiment 2, we can conclude that, as expected, the number of additional dimensions does not affect accuracy during the 2D cluster counting task since the appearance of the glyph itself does not influence the position estimation. This claim is supported by our evaluation, where we could not find a difference between the glyphs and the baseline encoding using dots.

During Experiment 3, we found that an increasing number of dimensions per glyph increased the accuracy of the outlier and subcluster detection tasks. This result initially puzzled us, but we believe that the number of emerging patterns increases with the number of dimensions, supporting this effect. In any case, more experiments should be carried out to assess this. On the correlation task, on the other hand, we see the expected drop in performance as the number of additional dimensions increases.
Glyph Shape. To approach whether our glyphs can be used to support high-dimensional exploration, our first question was to find out if the glyphs break the strong visual cue of positional encoding. Therefore, Experiment 2 compared our glyphs to a baseline dot encoding in a position-based task. From this experiment, we can conclude that the flower and star glyphs are on par with the baseline, suggesting that we go *out of the plane*.

Experiment 3 consisted of three tasks focusing on the glyph's additional dimensions. Our results suggest that glyph encoding does enable one to solve these common tasks. The flower glyph appears to be discernible visually. It performed well in both outlier and subcluster detection tasks, especially as the number of additional dimensions increased. We suspect this happens since the flower glyph offers strong visual saliency for individual high values because of how the petals are drawn. On the contrary, the star glyph takes a narrow shape in these cases. Figure 1, first column, shows an example of this effect. Furthermore, the star glyph has the drawback of overdrawing the region below when all the encoded dimensions are set to large values, leading to a larger amount of overdraw throughout the visualization.

7.2 Hypotheses Observations

In the following, we present our findings wrt. our hypotheses.
Position Preservation. We suspected that our glyph designs do not hinder the ability to decode 2D position compared to simple dot encodings. For the results of Experiment 2, we found partial support for H1. While we found significantly higher accuracy for the flower glyph when using seven dimensions encoded into the glyph, we could not find an effect on accuracy for the remaining conditions. We also found significantly faster response times for the dot encodings, probably due to less visual clutter using this simpler encoding. However, since we had to limit the parameters used in this experiment regarding the size or opacity of the dots, we think there might be a need for a larger user study to investigate this further.
Quantification. We would argue that our results support our hypothesis H2. However, this comes with a noticeable uncertainty growing with the number of encoded dimensions (around 20% for up to seven dimensions). One explanation for this is that neither of the glyphs scales in area linearly with the values. Although star glyphs suffer from the issue of different arrangements of the enclosed dimensions (as visualized in Figure 1), the petals of the flower glyph do not grow linearly compared to the encoded value. Using different radii for the leaves' ends to offset this effect could be an interesting research direction for optimizing value estimates.

Single Dimension Pattern. Due to the nature of the visual appearance of our glyphs, we suspected the pattern effect to be strong, as seen in Figure 3 on the left. The outlier and subcluster tasks' results suggest support for our hypothesis H3, while this effect is seemingly even stronger for the flower glyph. We suspect this is the case because of two visual properties of this glyph. First, as previously stated, the petals of the flower glyphs are independent of each other, emphasizing high values. Second, while the lengths of the petals grow linearly to the encoded value, the area does not. This behavior emphasizes higher values and builds a stronger visual pattern than the star glyph.

Multi Dimension Patterns. In Experiment 3, we used a correlation detection task to see whether glyph encodings allow for the comparison of multiple glyph values. We found that participants could not reliably recognize these patterns for this fairly complex task, rejecting H4. However, we discovered a hint that glyphs may support this task in circumstances when the stimulus features a strong correlation with a lower number of additional dimensions encoded into the glyph.

7.3 Lessons Learned

Based on our investigations of glyph-based encodings for high-dimensional data, we would like to distill some implications to inform glyph encoding in the investigated two-dimensional embeddings. Even though our investigations are limited, they nicely demonstrate that those individual glyphs are beneficial for specific tasks.

Our findings suggest that glyph encodings do not hinder the ability to decode the 2D position. We also found that flower glyphs showed promising results for two of our *out of the plane* tasks, namely outlier detection and subcluster detection. Thus, we derive the following implications from our experiments.

- Glyph encodings could be a viable choice when trying to find patterns like outlier and subcluster in two-dimensional embeddings.
- We suggest using flower glyphs for such embeddings since they perform on par for decoding 2D position and value estimating but outperform the star glyph for the pattern-related tasks.

7.4 Limitations & Future Work

We have already completed a relatively large series of studies with 912 participants divided into three experiments with five tasks and a diverse set of real-world datasets. However, to keep our user studies manageable, we had to restrict our investigations to various areas, such as glyph types and sizes.

However, we do need to establish some limitations. Firstly, we limited the number of glyphs or data points per stimulus. We also limited ourselves to a fixed size for the glyphs and the canvas, or in other words, a fixed relation between glyph size and canvas size. With this limitation and the relatively small number of 100 data points per stimulus, our goal was to limit the amount of visual clutter within the plot while still not sacrificing real-world transferability. To further limit our studies, we decided to limit Experiments 2 & 3 wrt. the number of additional dimensions per glyph, based on our findings from the first experiment.

Another aspect that we would like to point out is the factor of overplotting. As we already mentioned, our glyphs are not prone to overplotting in the same way because the star glyph overdraws the complete area below the glyph when encoding large values. Moreover, for Experiment 2, we tried to investigate if our glyphs still allow for decoding the 2D position but limited ourselves to a single size for the dot-based encoding. We are aware that these factors can potentially strongly influence task performance. However, since we did decide to use real-world datasets, we could not control this effect without altering the data. Investigating the influence of overlap on complex glyphs might be an interesting direction for further research.

In the future, we would like to investigate whether our findings can be applied to a broader range of conditions, such as a larger number of data points or different sizes of glyphs. We would also like to look at other glyphs, such as, for instance, the sunburst glyph.

8 Conclusions

Obtaining insights based on unknown high-dimensional datasets can be challenging. While dimensionality reduction techniques are a popular tool for visualizing such data sets, as they preserve high-dimensional features during the projection, many relations are lost during such a projection. This limitation opens up the need for appropriate visual encodings for additional dimensions beyond the two dimensions of the embedding. Therefore, we investigated the capabilities of glyph visualizations to visualize high-dimensional data within two-dimensional embeddings. Although glyphs are often used to communicate high-dimensional data, their value in the context of two-dimensional embeddings is largely unexplored. In a series of user studies involving five relevant tasks, we have investigated two commonly used glyphs for encoding individual attributes: flower glyphs and star glyphs.

Our findings suggest that glyph encodings support high-dimensional exploration without sacrificing positional encoding. We recommend using flower glyphs, rather than star glyphs, for tasks involving pattern detection, such as outlier and subcluster detection. Although quantifying values from the glyphs seems possible, it comes with relatively large uncertainty. We further found that while an increase in the number of encoded dimensions affects accuracy, pattern-related tasks like outlier and subcluster detection can even benefit from this, possibly due to a strong pattern effect.

References

[1] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[2] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[3] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg, "Evaluation of alternative glyph designs for time series data in a small multiple setting," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2013, pp. 3237–3246.

[4] M. Keck, D. Kammer, T. Gründer, T. Thom, M. Kleinsteuber, A. Maasch, and R. Groh, "Towards glyph-based visualizations for big data clustering," in *Proceedings of the 10th international symposium on visual information communication and interaction*. ACM, 2017, pp. 129–136.

[5] D. Kammer, M. Keck, T. Gründer, A. Maasch, T. Thom, M. Kleinsteuber, and R. Groh, "Glyphboard: Visual exploration of high-dimensional data combining glyphs with dimensionality reduction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 4, pp. 1661–1671, 2020.

[6] A. Sarikaya and M. Gleicher, "Scatterplots: Tasks, data, and designs," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 402–412, 2017.

[7] C. Frisson, S. Dupont, W. Yvart, N. Riche, X. Siebert, and T. Dutoit, "Audiometro: Directing search for sound designers through content-based cues," in *Proceedings of the 9th Audio Mostly: A Conference on Interaction With Sound*, 2014, pp. 1–8.

[8] J. Yang, W. Peng, M. O. Ward, and E. A. Rundensteiner, "Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets," in *IEEE Symposium on Information Visualization 2003 (IEEE Cat. No. 03TH8714)*. IEEE, 2003, pp. 105–112.

[9] W. Peng, M. O. Ward, and E. A. Rundensteiner, "Clutter reduction in multi-dimensional data visualization using dimension reordering," in *IEEE Symposium on Information Visualization*. IEEE, 2004, pp. 89–96.

[10] L. Wilkinson, A. Anand, and R. Grossman, "Graph-theoretic scagnostics," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.* IEEE, 2005, pp. 157–164.

[11] G. Albuquerque, M. Eisemann, D. J. Lehmann, H. Theisel, and M. Magnor, "Improving the visual analysis of high-dimensional datasets using quality measures," in *2010 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2010, pp. 19–26.

[12] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister *et al.*, "Quality metrics for information visualization," in *Computer Graphics Forum*, vol. 37, no. 3. Wiley Online Library, 2018, pp. 625–662.

[13] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, "A taxonomy of visual cluster separation factors," in *Computer Graphics Forum*, vol. 31, no. 3pt4. Wiley Online Library, 2012, pp. 1335–1344.

[14] J. Heer and M. Bostock, "Crowdsourcing graphical perception: using mechanical turk to assess visualization design," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 203–212.

[15] J. Lewis, M. Ackerman, and V. de Sa, "Human cluster evaluation and formal quality measures: A comparative study," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, no. 34, 2012.

[16] M. Sedlmair and M. Aupetit, "Data-driven evaluation of visual quality measures," in *Computer Graphics Forum*, vol. 34, no. 3. Wiley Online Library, 2015, pp. 201–210.

[17] A. Tatu, P. Bak, E. Bertini, D. Keim, and J. Schneidewind, "Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data," in *Proceedings of the International Conference on Advanced Visual Interfaces*, 2010, pp. 49–56.

[18] J. Lewis, L. Van der Maaten, and V. de Sa, "A behavioral investigation of dimensionality reduction," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 34, no. 34, 2012.

[19] J. Mackinlay, "Automating the design of graphical presentations of relational information," *Acm Transactions On Graphics (Tog)*, vol. 5, no. 2, pp. 110–141, 1986.

[20] Y. Wang, X. Chen, T. Ge, C. Bao, M. Sedlmair, C.-W. Fu, O. Deussen, and B. Chen, "Optimizing color assignment for perception of class separability in multiclass scatterplots," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 820–829, 2018.

[21] J. Li, J. J. van Wijk, and J.-B. Martens, "Evaluation of symbol contrast in scatterplots," in *2009 IEEE Pacific visualization symposium*. IEEE, 2009, pp. 97–104.

[22] L. Tremmel, "The visual separability of plotting symbols in scatterplots," *Journal of Computational and Graphical Statistics*, vol. 4, no. 2, pp. 101–112, 1995.

[23] J. Fuchs, P. Isenberg, A. Bezerianos, F. Fischer, and E. Bertini, "The influence of contour on similarity perception of star glyphs," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2251–2260, 2014.

[24] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri, "Perception of average value in multiclass scatterplots," *IEEE transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2316–2325, 2013.

[25] J. A. Lee and M. Verleysen, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing*, vol. 72, no. 7-9, pp. 1431–1443, 2009.

[26] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen, "Glyph-based visualization: Foundations, design guidelines, techniques and applications." in *Eurographics (STARs)*, 2013, pp. 39–63.

[27] M. O. Ward, "Multivariate data glyphs: Principles and practice," in *Handbook of data visualization*. Springer, 2008, pp. 179–198.

[28] J. Fuchs, P. Isenberg, A. Bezerianos, and D. Keim, "A systematic review of experimental studies on data glyphs," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 7, pp. 1863–1879, 2016.

[29] M. D. Lee, R. E. Reilly, and M. E. Butavicius, "An empirical evaluation of chernoff faces, star glyphs, and spatial visualizations for binary data," in *Proceedings of the Asia-Pacific symposium on Information visualisation-Volume 24*. Australian Computer Society, Inc., 2003, pp. 1–10.

[30] A. Klippel, F. Hardisty, and C. Weaver, "Star plots: How shape characteristics influence classification tasks," *Cartography and Geographic Information Science*, vol. 36, no. 2, pp. 149–163, 2009.

[31] A. Klippel, F. Hardisty, R. Li, and C. Weaver, "Colour-enhanced star plot glyphs: Can salient shape characteristics be overcome?" *Cartographica: The International Journal for Geographic Information and Geovisualization*, vol. 44, no. 3, pp. 217–231, 2009.

[32] M. Miller, X. Zhang, J. Fuchs, and M. Blumenschein, "Evaluating ordering strategies of star glyph axes," in *2019 IEEE Visualization Conference (VIS)*. IEEE, 2019, pp. 91–95.

[33] N. Cao, Y.-R. Lin, D. Gotz, and F. Du, "Z-glyph: Visualizing outliers in multivariate data," *Information Visualization*, vol. 17, no. 1, pp. 22–40, 2018.

[34] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[35] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[36] M. Harrower and C. A. Brewer, "Colorbrewer. org: an online tool for selecting colour schemes for maps," *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, 2003.

[37] W. H. Tedford Jr, S. Bergquist, and W. E. Flynn, "The size-color illusion," *The Journal of General Psychology*, vol. 97, no. 1, pp. 145–149, 1977.

[38] W. S. Cleveland and R. McGill, "A color-caused optical illusion on a statistical graph," *The American Statistician*, vol. 37, no. 2, pp. 101–105, 1983.

[39] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE transactions on visualization and computer graphics*, vol. 17, no. 12, pp. 2301–2309, 2011.

[40] J. Li, J.-B. Martens, and J. J. van Wijk, "A model of symbol size discrimination in scatterplots," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2010, pp. 2553–2562.

[41] R. M. Pickett and G. G. Grinstein, "Iconographic displays for visualizing multidimensional data," in *Proceedings of the 1988 IEEE Conference on Systems, Man, and Cybernetics*, vol. 514, 1988, p. 519.

[42] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 2000, pp. 93–104.

Christian van Onzenoodt received his Master's degree in Media Informatics from Ulm University in 2017 and is now working as a research associate at the Visual Computing Group at Ulm University. His current research interests are information visualization focusing on the perception of visualizations.

Pere-Pau Vázquez is an associate professor in the Computer Science Department at the Universitat Politècnica de Catalunya in Barcelona. He is member of the Research Center for Visualization, Virtual Reality, and Graphics Interaction (ViRVIG). His current interests are mostly related to visualization of large data sets, perception, and interaction in Virtual Reality environments. After graduating in Computer Science (1999), he obtained a Ph.D. in Software (2003) at Universitat Politècnica de Catalunya.

Timo Ropinski He is a professor at Ulm University, heading the Visual Computing Group. Before moving to Ulm, he was Professor in Interactive Visualization at Linköping University in Sweden, heading the Scientific Visualization Group. He received his Ph.D. in computer science in 2004 from the University of Münster, where he completed his Habilitation in 2009.