CLAMS: A Cluster Ambiguity Measure for Estimating Perceptual Variability in Visual Clustering

Hyeon Jeon*, Ghulam Jilani Quadri*, Hyunwook Lee, Paul Rosen, Danielle Albers Szafir, and Jinwook Seo

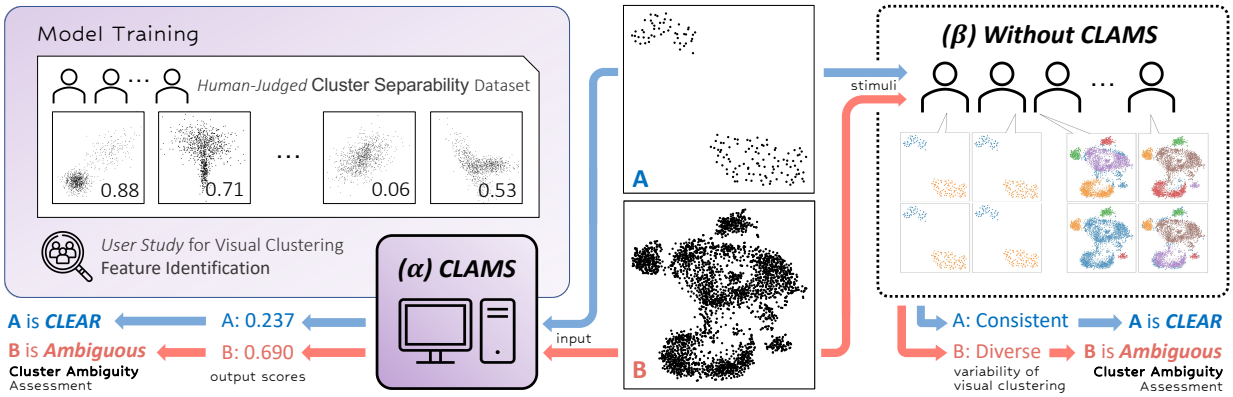


Fig. 1: The comparison between the way of estimating the perceptual variability in conducting visual clustering, i.e., cluster ambiguity, of monochrome scatterplots. Without CLAMS (β), we need to ask people to conduct visual clustering (colored scatterplots with β; color labels depict visual clustering results) and check whether the results are consistent (i.e., clear) or not (i.e., ambiguous). This approach requires extensive time and human resources. In contrast, CLAMS (α), which is constructed over perceptual data and a feature engineering based on a user study, automatically produces a score representing cluster ambiguity of an input scatterplot, where low and high scores correspond to clear and ambiguous cluster structure.

Abstract— Visual clustering is a common perceptual task in scatterplots that supports diverse analytics tasks (e.g., cluster identification). However, even with the same scatterplot, the ways of perceiving clusters (i.e., conducting visual clustering) can differ due to the differences among individuals and ambiguous cluster boundaries. Although such perceptual variability casts doubt on the reliability of data analysis based on visual clustering, we lack a systematic way to efficiently assess this variability. In this research, we study perceptual variability in conducting visual clustering, which we call *Cluster Ambiguity*. To this end, we introduce *CLAMS*, a data-driven visual quality measure for automatically predicting cluster ambiguity in monochrome scatterplots. We first conduct a qualitative study to identify key factors that affect the visual separation of clusters (e.g., proximity or size difference between clusters). Based on study findings, we deploy a regression module that estimates the human-judged separability of two clusters. Then, CLAMS predicts cluster ambiguity by analyzing the aggregated results of all pairwise separability between clusters that are generated by the module. CLAMS outperforms widely-used clustering techniques in predicting ground truth cluster ambiguity. Meanwhile, CLAMS exhibits performance on par with human annotators. We conclude our work by presenting two applications for optimizing and benchmarking data mining techniques using CLAMS. The **interactive demo** of CLAMS is available at `clusterambiguity.dev`.

Index Terms—Cluster, scatterplot, perception, cluster analysis, cluster ambiguity, visual quality measure.

1 Introduction

Clustering is a key analytic task in scatterplots [57, 81, 82]. It occurs when we infer structure in data by identifying groups (i.e., clusters) based on the pairwise proximity between data points [4]. Visual clustering occurs when people infer these groups visually, such as finding neighborhoods of points in a scatterplot. It is among the most common perceptual tasks people conduct with scatterplots [64]. Diverse domains, such as bioinformatics [69, 70] and machine learning [34], leverage visual clustering for data analysis. These applications include the identification of ground truth clusters for benchmarking data mining techniques, such as automatic clustering [7, 81] and dimensionality reduction techniques [19, 82].

While visual clustering supports a range of applications, *cluster ambiguity*—the intrinsic perceptual variability in visual clustering due to unclear cluster boundaries—can introduce uncertainty or even errors in applications relying on visual clustering. For example, cluster ambiguity can make data analysis unreliable. If a scatterplot is highly ambiguous, it is easy to make multiple conclusions about cluster structure within the data. Ambiguity also reduces the reliability of establishing ground truth clusters based on human perception for benchmarking data mining techniques: if each person perceives cluster structure differently, we cannot know which structure is "correct."

However, most studies and models of visual clustering focus on how people perceive clusters in general [2, 78, 81]. For example, Gestalt principles [78] explain how people "commonly" or "generally" group visual objects within a complex scene [54]. Studies examining average cluster perception support visualization designers in building effective systems that most people can use [42,50,77]. However, they do not offer solutions to deal with the challenges in data analysis and benchmarking imparted by cluster ambiguity. Designers thus need a more systematic

- *Hyeon Jeon and Jinwook Seo are with Seoul National University. E-mail: hj@hcil.snu.ac.kr, jseo@snu.ac.kr*
- *Ghulam Jilani Quadri and Danielle Albers Szafir are with the University of North Carolina, Chapel Hill. E-mail: {jiquad, danielle.szafir}@cs.unc.edu*
- *Hyunwook Lee is with UNIST. E-mail: gusdnr0916@unist.ac.kr*
- *Paul Rosen is with the University of Utah. E-mail: prosen@sci.utah.edu*
- *Hyeon Jeon and Ghulam Jilani Quadri equally contributed to the work.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

way to assess cluster ambiguity.

A reliable method to evaluate the ambiguity of a scatterplot is directly measuring perceptual variability via human experiments [26]. However, this process is costly and not scalable. As an alternative, we can use clustering techniques to mimic human perception [6, 67] (Sect. 2.2). Assuming that a clustering technique with a specific hyperparameter setting represents a human's perception, this method mimics human variability by running the technique under various hyperparameter settings. This approach is scalable but at the cost of accuracy due to the lack of human input (Sect. 5.2).

This research presents a scalable and accurate method to evaluate cluster ambiguity through a visual quality measure (VQM) called *CLAMS*. We design CLAMS based on a dataset gathered from human input about the separability of clusters [2]. We construct CLAMS in two steps: first, we conduct a user study to investigate important factors that influence visual clustering. Second, based on the study findings, we train a regression module estimating how human subjects separate clusters. Given a scatterplot, CLAMS computes the separability of every pair of identified clusters using the regression module. The measure then aggregates the computed pairwise separabilities of clusters to predict how ambiguously the clusters are portrayed by human subjects.

Our quantitative experiments show that CLAMS is more accurate than existing models in predicting cluster ambiguity. First, an ablation study verifies the accuracy of the regression module, validating our user-study-driven approach in constructing the module. Moreover, we find that the ranking of scatterplots set by CLAMS has a strong correlation with the ground truth ranking constructed from 20 participants in our study. Furthermore, CLAMS outperforms an average human annotator in estimating cluster ambiguity. We also present two applications of CLAMS in optimizing and benchmarking data mining algorithms. First, we propose AmbReducer, an optimization system that reduces the ambiguity of dimensionality reduction embeddings while maintaining accuracy. AmbReducer helps analysts effectively interpret high-dimensional data by informing cluster ambiguity. Second, we show how our measure can help select reliable benchmark datasets for comparing different clustering techniques. Findings from our experiments and applications open up discussions on leveraging perceptual variability in visualization research.

2 Background and Related Work

We survey past work about visual clustering in scatterplots, visual quality measures, and perceptual variability. These works collectively illustrate the necessity of CLAMS.

2.1 Cluster Perception in Scatterplots

Given the broad applications for visual clustering, several previous works have concentrated on understanding and modeling cluster perception. For example, eye-tracking was used to detect what aspects of a dataset people use for cluster identification, highlighting the role of Gestalt principles, especially proximity and closure [20]. ScatterNet captures perceptual similarities between scatterplots to emulate human clustering decisions [43]. Scagnostics identify scatterplot patterns, including cluster structure [15], but cannot reliably reproduce human perception [1]. In ClustMe [2], Abbas et al. computationally modeled human perception in judging the complexity of the cluster structure in scatterplots, which affects by the number of clusters and the nontriviality of patterns. Quadri & Rosen studied various factors that influence the perception of clusters. Relying on these factors, they built explainable models of how humans perceive cluster separation based on merge tree data structures from topological data analysis [57].

These studies aim to characterize cluster perception processes, the factors that influence cluster perception, and the relation between these factors and scatterplot design. However, visual clustering in scatterplots lacks any ground truth; we cannot always categorically determine which group of points forms a cluster. Underlying data characteristics lead to ambiguity in cluster structure (i.e., cluster ambiguity), causing intrinsic variability in the clusters people detect. We thus develop a VQM that automatically estimates cluster ambiguity for a wide range of cluster patterns. We then explore and rank the cluster ambiguity of scatterplots.

2.2 Visual Quality Measures on Cluster Perception

The visualization community has proposed and developed various *Visual Quality Measures* (VQMs) [9, 10] that metricize the specific characteristics of cluster patterns in scatterplots. These metrics enable analysts to rapidly evaluate their scatterplots in terms of supporting general pattern exploration [11] or a specific perceptual task [4], and can also be used to optimize scatterplots [56].

A naïve approach in designing VQMs is to reuse existing algorithms (e.g., clustering techniques), where the performance of the algorithm is evaluated against the results of human experiments. Aupetit et al. investigated how well clustering techniques can mimic (i.e., model) visual clustering by testing the extent to which 1,400 variants of clustering techniques can reproduce the human separability judgments [7]. Sedlmair & Aupetit proposed a framework for evaluating class separability measures based on human-judged class separability data and used the framework to test 15 class separation measures [67]. The framework was later extended to test a large set of class separability measures, which were constructed as combinations of 17 neighborhood graph functions and 14 class purity functions [6]. However, reusing existing algorithms does not directly reflect human perception, thus often performing poorly in imitating human perception [2] (Sect. 5.2).

VQMs can be instead designed to target a specific pattern or a perceptual task. For example, Quadri & Rosen [57] developed a threshold plot based on their modeling of cluster perception. Threshold plots tell the extent to which clusters are salient in an input scatterplot. Designers can use these plots to enhance the effectiveness and efficiency of a scatterplot for visual clustering [56]. ClustMe [2] also produced a VQM simulating human perception in judging the complexity of cluster patterns. Specifically, ClustMe models human perception based on experimental data on perceived cluster separability judged by multiple subjects. We apply this modeling approach and the perceptual data from ClustMe to guide our model. However, ClustMe regards the judgments of multiple subjects as a binary decision (i.e., separated or not separated), estimating a *general* complexity perceived. In contrast, CLAMS summarizes the judgments made by subjects as a continuous probabilistic function for the sake of estimating the *variability* of the judgments (i.e., cluster ambiguity). Estimating variability helps us understand the distribution of human visual percepts, enabling related applications to more accurately reflect perception.

2.3 Perceptual Variability

Due to perceptual variability among individuals, the perception of clusters can differ even with the same scatterplot. Perceptual variability refers to individuals' "traits or stable tendencies to respond to certain stimuli or situations in predictable ways" [18]. Prior work on perceptual variability demonstrated that people exhibit observable differences in task-solving and behavioral patterns [79]. Psychology and vision research have shown that people exhibit substantial variability on tasks such as pattern identification, judgments, or adjustments [28, 29, 51]. Given the significance of perceptual variability in information visualization [16, 86], recent research explores when, where, and how perceptual variability influences visualization use (see Liu et al. [41] for a survey).

Perceptual variability can influence the reliability of any generalized model of perception [84]. The Axiom of Perceptual Variability suggests that the success of such generalized perceptual theories depends on their ability to account for variance in perceptual representations [5]. Moreover, the perceptual variability of a certain task may degrade the credibility of the applications relying on visual clustering (Sect. 1). Visual clustering is an ill-posed problem, where there is no "ground truth" for clusters (i.e., it is not always possible to determine a "correct" clustering). Generalized models and applications of visual clustering are thus likely to be more vulnerable to perceptual variability. We believe that our modeling of perceptual variability in visual clustering (i.e., cluster ambiguity) not only enhances our understanding of cluster perception but also resolves the vulnerabilities of such models and applications.

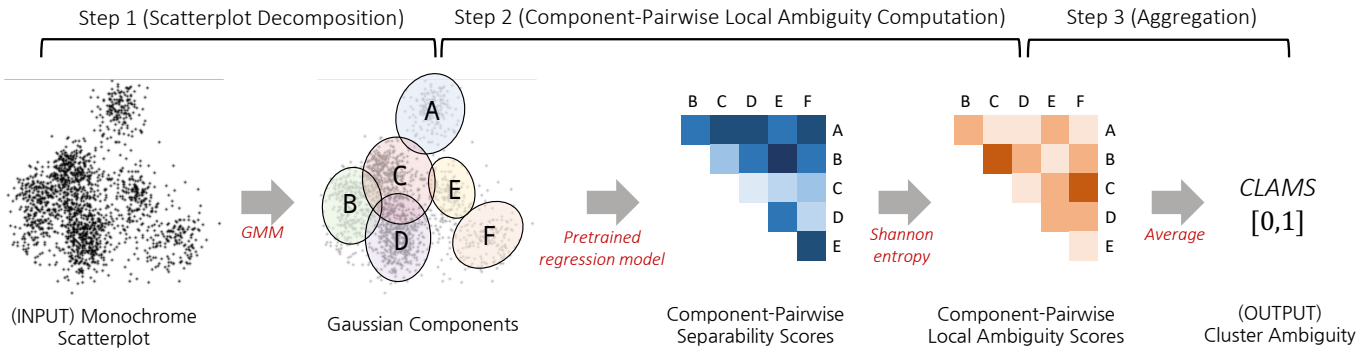


Fig. 2: The CLAMS pipeline (Sect. 3.2). (Step 1) Given a scatterplot, we first apply a Gaussian Mixture Model (GMM) to abstract a scatterplot into a set of Gaussian components. (Step 2) We estimate component-pairwise separability scores by applying a predefined regression module (Sect. 4) and convert the scores into local ambiguity scores by applying the binary Shannon entropy function. (Step 3) Finally, we obtain the cluster ambiguity score of the input scatterplot by averaging the local ambiguity scores.

3 CLAMS

We introduce CLAMS, a VQM for estimating the cluster ambiguity of a monochrome scatterplot. Trained over perceptual experiment data, our approach acts as a proxy for human perception. Moreover, by decomposing an input scatterplot with a Gaussian mixture model (GMM) and measuring ambiguity in a component-pairwise manner, CLAMS can deal with a wide range of cluster patterns while maintaining scalability.

3.1 Design Considerations

Based on a thorough examination of the related work on cluster perception (Sect. 2.1) and VQMs (Sect. 2.2), we set three design considerations in estimating cluster ambiguity.

- **(C1) Provide a proxy for human perception:** To assess the variability of people's perception, the measure should accurately reflect how people identify and analyze clusters in practice [2, 7, 57]. Refer to Step 2 in Sect. 3.2 for how we achieved this consideration.
- **(C2) Work with a wide range of scatterplot patterns:** Scatterplots have a wide range of cluster patterns; they can have clusters with diverse characteristics (e.g., shape, density, size) [32] and varying numbers of clusters [57]. CLAMS should properly estimate the ambiguity of an arbitrary scatterplot and should properly estimate cluster ambiguity across a diverse array of patterns [2, 57]. Check Step 1 in Sect. 3.2 to see how our measure design regards this consideration.
- **(C3) Be scalable:** To be readily used in practice (Sect. 6), CLAMS should scale to large numbers of data points and complex patterns [2]. Throughout the entire pipeline (Steps 1, 2, and 3), we maintain scalability as a key consideration.

3.2 CLAMS Pipeline

We draw on previous approaches (e.g., VQM [2], clustering [25, 60], dimensionality reduction [36, 75]) to design our measure to (Step 1) decompose a given scatterplot into smaller components, (Step 2) compute component-pairwise local ambiguity scores, and (Step 3) estimate the global ambiguity score by aggregating the local scores (see Fig. 2). As the diversity of local components is inherently much lower than that of the full scatterplot, this decomposition helps CLAMS to readily work with a wide range of scatterplot patterns (C2).

(Step 1) Scatterplot Decomposition

CLAMS starts by decomposing an input scatterplot into smaller components. We use the Gaussian Mixture Model (GMM), which decomposes a given dataset as a mixture of multidimensional Gaussian distributions. We select GMM due to several advantages. First, GMM does not require hyperparameter selection [2]. The number of Gaussian distributions (components) can be automatically determined based on fixed statistics, such as Bayesian information criteria (BIC) [66]. Moreover, as each component is represented as a Gaussian distribution, complex patterns can be abstracted into a concise statistical summary (e.g., mean, covariance matrix; C2). Note that GMM decomposition has been shown to be applicable to visual identification problems [2, 3], accurately representing a wide range of smooth cluster patterns. Finally, the complexity of GMM for 2D scatterplots is $O(NK)$, where N and K denote the number of points and components, respectively, ensuring that the technique is highly scalable (C3).

In contrast, a conventional approach for scatterplot decomposition, which involves using clustering techniques (e.g., K-Means [38], HDBSCAN [45]), falls short in enabling CLAMS to satisfy our target considerations. First, clustering results can vary significantly due to the sensitivity of the outcomes to changes in hyperparameters. We can find an optimal hyperparameter setting using clustering validation measures [83], but the optimal setting may depend on the selection criteria [40]. Clustering techniques also provide a partition of data points, meaning they neither abstract nor simplify clustering patterns. Lastly, clustering techniques that are widely known to be able to capture complex patterns (e.g., density-based clustering [45], density-peak clustering [39]) often suffer from scalability issues.

Note that while applying GMM to the input scatterplot, we determine the optimal number of Gaussian components based on BIC scores and the elbow rule. The elbow is found using the Kneedle [65] algorithm.

(Step 2) Component-Pairwise Local Ambiguity Computation

We predict local ambiguity for every pair of Gaussian components so that CLAMS can consider every possible interaction between components. This process is as follows:

Dataset: We use the human-judged cluster separability dataset $\mathbf{X}$ from the ClustMe study [2] (C1). $\mathbf{X}$ contains 1,000 scatterplots $\{X_1, X_2, \cdots, X_{1000}\}$, where each scatterplot consists of a pair of Gaussian components with diverse statistics (e.g., mean, covariance). The separability scores $\mathbf{S} = \{S_1, S_2, \cdots, S_{1000}\}$ are computed by aggregating the judgments of 34 participants on each scatterplot gathered by the ClustMe study [2]. These scores reflect the judgments of the participants regarding whether they saw one or multiple clusters in each scatterplot. S_i represents the proportion of participants who perceived more than one cluster in X_i.

Deriving Cluster Ambiguity from Separation Score: We compute the cluster ambiguity of each Gaussian components pair with separation score S as: $A(S) = -S\log S - (1-S)\log(1-S)$. Theoretically, A is defined as the Shannon entropy of a binomial distribution representing the probabilistic event of counting the number of clusters. We use entropy as it is an efficient and mathematically grounded function representing the level of "uncertainty" of the associated event [44] (C3). A is minimized when S is 0 or 1 (i.e., every participant gave the same answer, meaning there was no variability in visual clustering), and is maximized when S is 0.5 (i.e., half of the participants saw a single cluster and the other half saw more than one cluster, maximizing variability). As the scatterplots and original separability scores have proven to be effective in modeling cluster perception [2], we can also expect the ambiguity scores to reliably represent human perception.

Note that our evaluation validates this expectation (Sect. 5.2; C1).

Predicting Ambiguity: To predict local ambiguity for each Gaussian component, we train a regression module f estimating the separability score $f(X)$ of an arbitrary pair of Gaussian components X, using $\mathbf{X}$ and $\mathbf{S}$. For an arbitrary pair of Gaussian components X, the local ambiguity score of a pair is computed as $A(f(X))$. Although training the regression module requires heavy computation (Sect. 4), inferring local ambiguity scores can be done in constant time regardless of the number of Gaussian components, making Step 2 scalable (C3). Please refer to Sect. 4 for a detailed explanation of how we designed and implemented the model.

(Step 3) Aggregating Local Ambiguity

Finally, we aggregate the component-pairwise ambiguity scores by averaging the scores. Note that the final score can be interpreted as the joint entropy of a set of events corresponding to each local ambiguity score with an assumption that the events are mutually independent.

3.3 Computational Complexity

As the time complexity of GMM on a 2D scatterplot is $O(NK)$, the time complexity of Step 1 is $O(NK_{max}^2)$, where K_{max} is the maximum number of components we consider in finding the optimal number of components and N is the number of points. The complexity for both Steps 2 and 3 is $O(K_{opt}^2)$, where K_{opt} denotes the optimal number of components found in Step 1. Thus, the computational complexity of CLAMS is $O(NK_{max}^2 + K_{opt}^2)$. As $N \gg K_{max}$ and $N \gg K_{opt}$, the effect of K_{max} and K_{opt} is negligible, making CLAMS a scalable measure that is only linear to the number of points. Refer to Appendix B for the quantitative experiment demonstrating the scalability of CLAMS.

4 Estimating Human-Judged Cluster Separability

We constructed a regression module to predict the separability of a given Gaussian components pair. We conducted a user study exploring the factors affecting visual clustering to ground our model in human reasoning processes. We trained the model to estimate ambiguity from the features extracted from each pair, where the features were engineered based on the study findings.

4.1 Factor Exploration Study

To design features that relate to the human-judged cluster separability, we first explored factors that affect visual clustering via a qualitative experiment. In the study, participants completed a series of visual clustering tasks and reported on factors that affected task results.

4.1.1 Study Design

Procedure and Tasks: The experiment began with informed consent and a basic demographic survey. Participants completed the rest of the study in three phases: (1) identifying clusters in a single scatterplot, (2) performing pairwise comparisons between scatterplots, and (3) participating in an interview to elicit core factors influencing visual clustering. In the first phase, we randomly sampled 12 scatterplots from $\mathbf{X}$ and showed them in sequence. For each scatterplot, we first asked participants to select whether there existed one or more than one cluster, following the original ClustMe [2] study. We additionally asked participants about their confidence in their response using a Likert scale and asked them to describe the reasoning process both for the number of clusters and their confidence.

In the second phase, we randomly sampled 12 pairs of scatterplots (24 in total) from $\mathbf{X}$. For each trial, we asked participants to report which scatterplot in a given pair was more separated. As in the first phase, we asked for the participants' confidence and had them report the reasoning behind their choices and confidence. After the two phases of the experiment, we conducted a semi-structured post hoc interview, asking the participants about the salient factors they felt were affected by cluster separation. All participants finished the experiment within one hour.

Participants: We recruited 10 participants from a local university (six males and four females, aged 19–28 [23.5 ± 2.6]). Six of the participants were undergraduates, three were graduate students, and one had just completed their Bachelor's degree. We selected only participants who had experience in data analysis using scatterplots to better ground our results in real-world data analysis and to ensure that they understood the concept of clusters. Three participants reported that they are at a novice level in data analysis, which meant that they did not regularly conduct analyses but have some experience. Five and two participants reported themselves as at intermediate and expert levels, respectively. Half of the participants reported being at the novice level of data analysis using a scatterplot and the other half reported themselves as intermediate. We also confirmed that participants had not read three papers [2, 3, 7] that incorporate ClustMe data (i.e., $\mathbf{X}$). Participants were compensated with the equivalent of $20.

Apparatus: The experiment was conducted over Zoom, and the sessions were recorded. We developed a website in which participants could see scatterplots and make their selection (number of clusters and confidence) with a mouse click. We fixed the stimuli size to 700px $\times$ 700px and constrained participants to use a laptop or desktop screen to minimize the impact of the display on study results. They were asked to access the website and share their screens so that the experimenter could monitor and guide the experiment.

4.1.2 Results

We analyzed the responses from the main experiment and the post hoc interview using axial coding done by two authors, one of whom works in machine learning while the other's primary expertise is in visualization. The coders individually developed a separate codebook in the first stage. Two codebooks were then merged, resulting in 13 total extracted factors. Finally, based on two stages of discussions, the coders agreed to categorize the codes into six main factors (see Table 1)—proximity, clarity of gap, density difference, size difference, ellipticity difference, and direction.

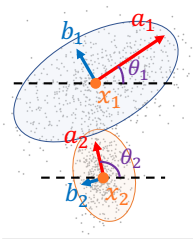
The most important factor was the *proximity* between the clusters, which was mentioned by seven out of ten participants in the interviews. The perception of proximity was mainly affected by the *clarity of a gap* between clusters. Participants reported that if the gap between clusters was bigger and less dense than the clusters, the clusters were perceived to be more distinct. Five out of ten participants explicitly mentioned the gap between clusters as an important factor in the interviews.

Participants tended to perceive more than one cluster when two Gaussian components had noticeably *different densities*, even if there was no gap between them or if they overlapped (i.e., when they had high proximity). However, two participants explicitly noted that if the density of a point group was too low, the points within the group were perceived as outliers rather than a cluster. We found that the *size difference* between clusters similarly affects visual clustering: if a cluster was too small, participants often wanted to interpret the cluster as a set of outliers. In the interviews, 30% (three out of 10) of participants noted that the differences in density were salient features that affected their choices, whereas one noted that differences in cluster size played a role in their decisions.

Ellipticity difference between clusters also affected perceived separability. Participants mentioned that if the ellipticity of a cluster was high, it could be fit by linear regression where the regression line accorded with the major axis of the ellipse, thus would be more likely to be perceived as a single cluster. We also found that if the two clusters had high ellipticity, their *direction* (i.e., the direction of the first principal axis) also played a crucial role in their separation. Participants noted that two clusters with high ellipticity and different directions were likely to fit into two independent regression lines, thus more likely to be perceived as two independent clusters. In the interviews, five out of ten participants mentioned ellipticity and direction as salient factors, while four additionally mentioned linear regression or correlation lines.

4.2 Feature Engineering

We designed features representing the characteristic of a pair of Gaussian components based on the six factors from the study (Sect. 4.1). To maintain the simplicity and explainability of the model, we aimed to keep the features as simple and few as possible. We thus designed the features to be (1) be bijective to the factors and (2) directly computed



Factors	Features	Formula	Used
Proximity	Distance between centers (DC)	$\| x_1 - x_2 \|$	○
Clarity of a gap	Distance-size ratio (DSR)	$\| x_1 - x_2 \| / (\sqrt{\| a_1 \|^2 + \| b_1 \|^2} + \sqrt{\| a_2 \|^2 + \| b_2 \|^2})$	○
Density Difference	Density difference (DD)	$\| (n_1/(2 \cdot \| a_1 \| \cdot \| b_1 \|)) - (n_2/(2 \cdot \| a_2 \| \cdot \| b_2 \|)) \|$	×
Size Difference	Size difference (SD)	$\| \sqrt{\| a_1 \|^2 + \| b_1 \|^2} - \sqrt{\| a_2 \|^2 + \| b_2 \|^2} \|$	○
Ellipticity Difference	Ellipticity difference (ED)	$\| (\| a_1 \| / \| b_1 \| - \| a_2 \| / \| b_2 \|) \|$	○
Direction	Angle between components (AC)	$\min(\Delta_\theta, 2\pi - \Delta_\theta)$ where $\Delta_\theta = \| \theta_1 - \theta_2 \|$	○

a_1, a_2 & b_1, b_2: Standard deviation along the major & minor axis of Gaussian components / x_1, x_2: the center of Gaussian components
θ_1, θ_2: Angle between the major axis of Gaussian components and the horizontal line / n_1, n_2: number of points in Gaussian components

Table 1: The visual clustering factors revealed by our preliminary user study (Sect. 4.1; first column) and the corresponding features designed for the regression module predicting the human-judged separability of Gaussian components (Sect. 4.2; second column). The third column depicts how we compute each feature, and the fourth column represents whether or not the feature is used in the final regression module (Sect. 5.1). Empowered by the designed features, the module succeeds in precisely estimating the separability scores.

from the statistical summary of the Gaussian components pair (see Table 1). The features are as follows:

Distance between Centers (DC): For *proximity*, we selected the distance between the centers of Gaussian components. We used the distance between centers as it is the sole metric representing proximity that can be directly derived from the population statistics (i.e., mean) and is also widely used for measuring the proximity between clusters [40].

Distance-Size Ratio (DSR): The gap between two Gaussian components generally becomes bigger if (1) the distance between the center of two Gaussian components increases or (2) the size of the components becomes smaller. We used this property to construct a feature corresponding to the *clarity of the gap* as the ratio of the distance between centers of Gaussian components over the sum of the size of two components. Note that we defined the size of a component as its standard deviation—the root sum square of the standard deviations along the first and second principal axes.

Density Difference (DD): The *density difference* factor can be directly represented by computing the density of a component as the ratio of the number of points over the area covered by the component. We defined the area of a component as that of an ellipse where the major and minor axes' length is identical to the standard deviation along the first and second principal axes, respectively.

Size Difference (SD): *Size difference* can be also directly represented as a feature. We defined the size of the component as its standard deviation, as we do for the distance-size ratio feature.

Ellipticity Difference (ED): We directly used *ellipticity difference* as a feature, defining the ellipticity of a component as the one of an ellipse having the standard deviation along the first and second principal axes as major and minor axes' lengths, respectively.

Angle between Components (AC): Participants reported *direction* as salient only if they think two components are heading toward different directions. We therefore define direction as the angle between components, which can be represented as the angle between the first principal axes of two components.

4.3 Modeling Training

We trained a regression module for estimating human-judged separability of Gaussian components pair using $\mathbf{X}$ and $\mathbf{S}$. We used a regression module as we aimed to predict continuous scores. For a given pair of Gaussian components, the model first extracts the features as defined above and then predicts a separability score from the features. We implemented our module using AutoML [27] supported by the `auto-sklearn` [21] library. We report the performance of our model and the significance of the extracted features in Sect. 5.1.

5 Quantitative Evaluation

We conducted two evaluations to demonstrate the validity and effectiveness of CLAMS. We first examined the performance of our regression module and the significance of its features through an ablation study (Sect. 5.1). We then evaluated the accuracy (Sect. 5.2) of CLAMS by comparing it against both computational methods for estimating cluster ambiguity and human annotators.

		Excluded feature					
	orig.	DC	DSR	DD	SD	ED	AC
R^2	.9106	.9075	.8574	.9205	.8790	.8758	.9019
Change	0%	-0.34%	-5.84%	+1.08%	-3.47%	-3.82%	-0.95%

(a) Model performance with entire (*orig.*) features or without a single feature

		Excluded feature					
		DC	DSR	DD	SD	ED	AC
Excluded Feature	DC		.5817	.8751	.8917	.8495	.8333
	DSR	-36.1%		.8521	.8985	.8560	.8902
	DD	-3.89%	-6.42%		.8925	.8790	.8721
	SD	-2.08%	-1.33%	-1.99%		.8750	.9041
	ED	-6.71%	-6.00%	-3.47%	-3.91%		.8888
	AC	-8.49%	-2.24%	-4.23%	-0.71%	-2.40%	

(b) Model performance without a pair of features (in corresponding row and column)

Table 2: The result of an ablation study (Sect. 5.1) analyzing the performance of a regression module estimating human-judged separability scores. We seek how the performance of the module varies as we remove a single (a) or a pair (b) of features, reporting the R^2 score (first row in (a), upper right half in (b)), and the change rate compared to the model using entire features (second row in (a), lower left half in (b)).

5.1 Ablation Study for the Regression Module

We report the results of the ablation study examining (1) the performance of our regression module (Sect. 4) and (2) the significance of the features we designed (Table 1; Sect. 4.2).

5.1.1 Study Design

Objectives and Design: The study aimed to achieve two objectives: (1) to investigate the accuracy of our regression module in estimating human-judged separability and (2) to analyze how much each feature contributed to the model performance (i.e., the significance of features). For the first objective, we evaluated the performance of our model as it is (i.e., used all features we discovered). To accomplish the second objective, we switched off each feature individually and examined the extent to which the performance deteriorated. Additionally, to take into account the interplay between features, we repeated the process by disabling feature pairs.

Measurement: We conducted a five-fold cross-validation to examine the performance of the model while using $\mathbf{X}$ and $\mathbf{S}$ as input and target, respectively. We used R^2 for the performance metric as it is interpretable [13] and, moreover, unbiased if the number of points is fixed (as in our case).

5.1.2 Results and Discussions

Model Performance: As in Table 2 (a), the model achieved R^2 score over 0.9 with all features (which we mark as *orig.*). This result validates

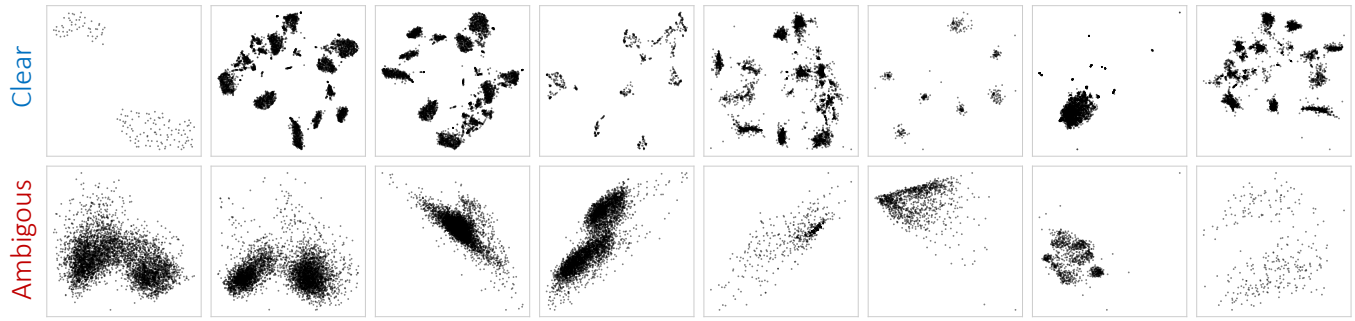


Fig. 3: The top eight clear scatterplots (low cluster ambiguity; first row) and the top eight ambiguous scatterplots (high cluster ambiguity; second row) based on CLAMS score, which are picked within 60 scatterplots we used in the accuracy evaluation (Sect. 5.2).

the effectiveness of the features and also the reliability of our study-driven approach to designing features.

Significance of the Features: According to Table 2 (a), DSR caused the biggest degradation of the performance when removed. ED was next, and SD was third. Meanwhile, removing DC, DD, and AC made only a subtle decrement ($< 1\%$) in the performance or even made it better, indicating that these three features barely explained visual clustering alone. However, we cannot conclude that the bottom-three features (DC, DD, AC) have less significance than the top-three features (DSR, ED, SD) due to the existence of interplay between features (see Table 2 (b)). For example, degradation resulting from switching off AC and DC together was substantially higher than the sum of degradation caused by switching off the features individually and was also higher than the largest degradation possible by removing a single feature (DSR). The same phenomenon occurs for ED and DC or DC and DSR. The interplay between DC and DSR was especially large compared to other combinations, validating the influence of the proximity factor in visual clustering. The significance of feature pairs suggests that future studies on visual clustering factors should focus on examining feature interplay in detail rather than analyzing features individually.

5.2 Accuracy Evaluation of CLAMS

We evaluated the accuracy of CLAMS by (1) constructing the ground truth cluster ambiguity ranking of scatterplots and (2) checking how well the ranking made by our measure matched the ground truth. The ground truth ambiguity of scatterplots was estimated based on a user study with 18 participants. Our results showed that CLAMS precisely estimated ground truth cluster ambiguity, outperforming the tested alternatives (e.g., human annotations, variability of clustering techniques).

5.2.1 User Study for Constructing Ground Truth Ambiguity

Objectives and Tasks: Our user study aimed to construct the ground truth cluster ambiguity of scatterplots. We formulated two experimental tasks relevant to our research questions.

(T1) Lasso the clusters in the given scatterplot using the mouse.
(T2) Subjectively determine the ambiguity of the scatterplot.

T1 aims to directly collect the visual clustering results of participants so that we could compute ground truth cluster ambiguity. We additionally asked participants to conduct T2 to check how well the subjective ambiguity annotated by individuals matches the ground truth ambiguity compared to CLAMS.

Procedure: The entire study consisted of three phases. We first collected the participants' non-identifying demographics. We then asked the participants to conduct the two tasks listed above, viewing 60 selected stimuli (i.e., scatterplots) in random order. Finally, participants answered the post-study interview questions: (1) *"Which characteristics of the scatterplots do you think are most important in determining separation?"* (2) *"What makes scatterplots ambiguous or clear based upon your response on the shown scatterplots?"*

Datasets and Preprocessing: Our primary consideration in generating scatterplot stimuli was maximizing the diversity of cluster patterns (related to C2 in Sect. 3.1). For this purpose, we first generated a large number of scatterplots with a high diversity of patterns. This was done by applying eight dimensionality reduction (DR) techniques (t-SNE [74], UMAP [46], Densmap [49], Isomap [73], LLE [61], MDS [35], PCA [52], and Random Projection) to 96 high-dimensional datasets [31]. Except for PCA and MDS, we generated 20 scatterplots while randomly adjusting the hyperparameters to further diversify the patterns. We obtained $(20 \cdot 6 + 2) \cdot 96 = 11,712$ scatterplots in total.

We consecutively applied a stratified sampling, following Pandey et al. [1] and Abbas et al. [2], to remove scatterplots with similar patterns [2]. To do so, we first grouped candidate scatterplots based on their number of clusters. We set the number as the optimal number of Gaussian components computed by GMM. We then computed Scagnostics [15] of scatterplots, representing each scatterplot as a vector consisting of Scagnostics scores (i.e., a vector abstracting the pattern of a scatterplot). Finally, we applied K-Means [38] to vectors representing each group of scatterplots and sampled scatterplots that were closest to centroids of resulting clusters. We set K as 12. If the number of scatterplots in a group was less than 12, we sampled all scatterplots from the group. K value is set to limit the number of sampled scatterplots to be around 100, making it possible to manually investigate each by eye. 114 scatterplots were sampled. We manually sampled these scatterplots to minimize similar patterns, resulting in the final 60 scatterplots.

Participants: CLAMS is designed for use in data analytics. We, thus, required participants to have experience in data analysis using scatterplots to ensure that they understood the concept of visual clustering. Based on snowball sampling [24], we recruited 18 participants from the data visualization, human-computer interaction, and data mining communities (13 males and five females, aged 23-33 [27.2 ± 2.7]). While 15 participants were graduate students, the remaining three were working professionals. Two participants self-reported as novice data analysts, two as intermediate, and 15 as experts. For familiarity with data analysis with scatterplots, 14 participants reported that they were at the expert level, and two participants said they were at an intermediate level. The remaining two participants reported themselves as novices. We compensated each participant with the equivalent of $20 for their time.

Apparatus: The experiment was conducted over a recorded Zoom session. We developed a website for the study participants. Participants were asked to access the website and share their screens so that the instructor could monitor and guide the experiment. As with the factor exploration study (Sect. 4.1), we fixed the stimuli size to 700px $\times$ 700px and constrained participants to use a laptop or desktop monitor. The studies were 45-60 minutes long. We instructed the participants on tutorials and experimental tasks, and conducted interviews at the end.

5.2.2 Extracting Ground Truth Cluster Ambiguity

We extracted ground truth cluster ambiguity for the sampled scatterplots by measuring the extent to which visual clustering results (i.e., lassoing result) differed by participant. To compute the difference, we used external clustering validation measures (EVMs) [80], following previous research on visual clustering [17, 30]. As EVMs can be applied only to a pair of clusters, we computed EVM scores of visual

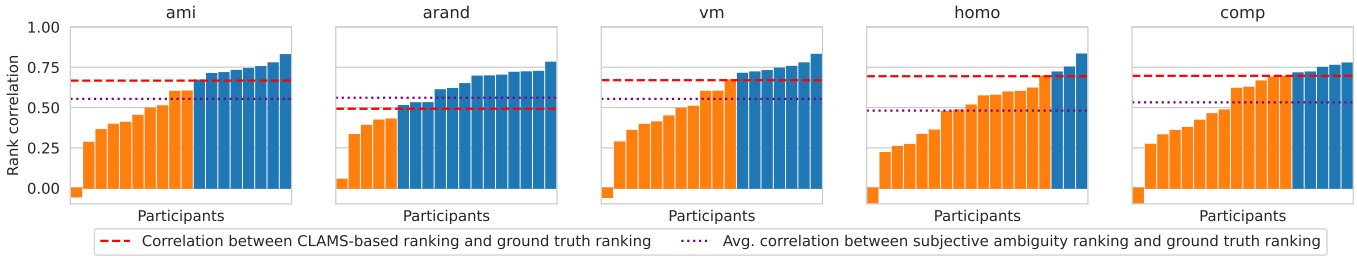


Fig. 4: The comparison of the performance of CLAMS and human annotators (i.e., participants in our user study; Sect. 5.2.1) in estimating ground truth cluster ambiguity ranking. Each bar represents the rank correlation between the cluster ambiguity ranking made by the subjective response of a single participant and the ground truth ambiguity ranking. The orange and blue colors denote the participants who showed less and more accurate performance compared to CLAMS (red dashed line), respectively. The purple dotted line depicts the average performance (i.e., rank correlation with ground truth ambiguity) of participants. We found that CLAMS's performance is better than the average performance of participants but fails to outperform the performance made by the best annotator.

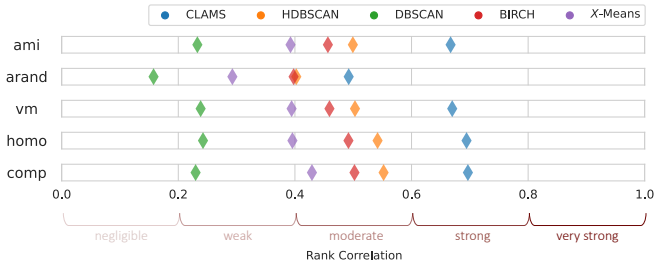


Fig. 5: The rank correlation between ground truth ambiguity extracted based on five EVMs (Sect. 5.2.2) and the rankings made by computational ways to measure cluster ambiguity (CLAMS, variability of clustering techniques). Overall, CLAMS outperformed the competitors regardless of the used EVM, objectively having a strong correlation ($\rho > 0.6$) with ground truth [55] for the majority of the cases.

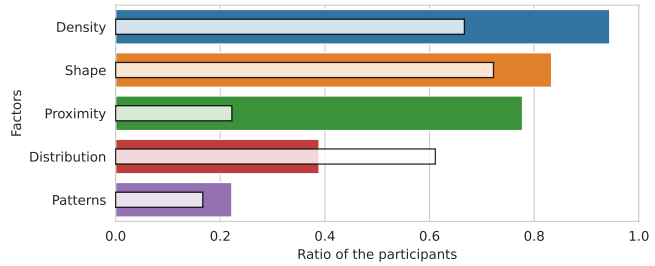


Fig. 6: We identified the factors that participants considered when lassoing clusters and rating the scatterplot on a scale of very ambiguous to very clear in our main study (Sect. 5.2). The length of the filled bars and white bars depicts the ratio of the participants who denote that the corresponding factor affects cluster separability and ambiguity, respectively. As the past studies [62, 68] have demonstrated, we observed density, proximity, and shape as main factors.

clustering results for each cluster pair and averaged the scores to get the final ground truth ambiguity. For EVMs, we used adjusted mutual information (`ami`) [76], adjusted rand index (`arand`) [72], v-measure (`vm`) [59], homogeneity (`homo`) [59], and completeness (`comp`) [59]. We chose these measures as they are widely adopted in clustering and data mining communities [31, 58]. As a result, we obtained five ground truth cluster ambiguity rankings of scatterplots, where each corresponds to an individual EVM. We only used EVMs that are "adjusted" so that the scores can be compared across different datasets [80].

5.2.3 Results and Discussions

Performance Analysis: The results validate the CLAMS's preciseness in estimating cluster ambiguity. As seen in Fig. 5, the rankings made by CLAMS generally showed a strong correlation with the ground truth ranking ($\rho > 0.6$) based on a criterion of Prion and Haerling [55]. The correlation was, however, low for `arand`. This is because, unlike other EVMs that interpret clustering assignments as a probabilistic event, `arand` discretely "counts" the disagreement of clustering results and thus generates results that do not align with our probabilistic approach. CLAMS also substantially outperformed clustering techniques in terms of predicting ground truth for all EVMs, verifying that CLAMS is the best of the tested computational options.

As depicted in Fig. 4, CLAMS showed better performance in estimating ground truth cluster ambiguity compared to the average human annotators. Except for the case of `arand`, CLAMS always showed a better correlation with ground truth (red dashed line) compared to the average correlation of human annotators (purple dotted line) regardless of the EVM choice. On average, 9.8 out of 18 participants (54%; the proportion of orange bars) made less accurate predictions compared to CLAMS. This result indicates that it is difficult to estimate ambiguity for the judgment of a single person. Therefore, our automatic solution offers increased value in accurately evaluating ambiguity.

Post-Study Interview: As described in Sect. 5.2.1, we interviewed participants to identify the reasoning behind their responses. We found that the main factors behind the lassoing (i.e., visual clustering) and determining the scatterplot ambiguity levels were density, proximity, shape, and distribution (as shown in Fig. 6). We elaborate on these findings in detail in Sect. 7.

6 Applications

We introduce two applications of CLAMS that validate the effectiveness, generalizability, and applicability of CLAMS. We present that CLAMS can further optimize nonlinear dimensionality reduction embeddings to reduce cluster ambiguity while maintaining accuracy (Sect. 6.1). We then demonstrate that the measure can be used to find reliable datasets for benchmarking clustering techniques (Sect. 6.2).

6.1 Reducing Cluster Ambiguity of Nonlinear Dimensionality Reduction Embeddings

6.1.1 Problem Statement and System Design

A common way to analyze the cluster structure of high-dimensional data is to use dimensionality reduction (DR) [50] techniques, e.g., t-SNE [74], Isomap [73], or UMAP [46], which synthesizes 2D representation (i.e., embedding) that preserves the original characteristics of the input high-dimensional data. The analysis is usually done by depicting the embedding as a scatterplot and conducting visual clustering. Recently, nonlinear DR techniques [37] that can reveal the complex data manifold have been widely developed [22, 33, 46, 74] and used for cluster analysis [8, 23].

Embedding varies by choice of DR technique or hyperparameter, where each can lead to significantly different analysis results. A typical way to address this problem is to optimize embeddings by assessing their accuracy in representing the original data. Diverse metrics to measure the accuracy of DR embeddings [32, 37, 50] have been developed for the purpose. However, an optimized embedding can still have an

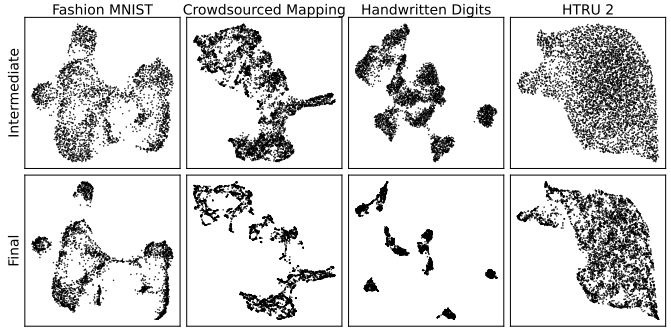


Fig. 7: Dimensionality reduction embeddings optimized solely based on the accuracy (top row; intermediate results of AmbReducer) and those optimized considering both accuracy and cluster ambiguity (bottom row; final results of AmbReducer). Our experiment shows that the accuracy of the final results has no significant difference compared to the intermediate results in terms of accuracy but has a substantially smaller amount of cluster ambiguity (Sect. 6.1.2, Fig. 8).

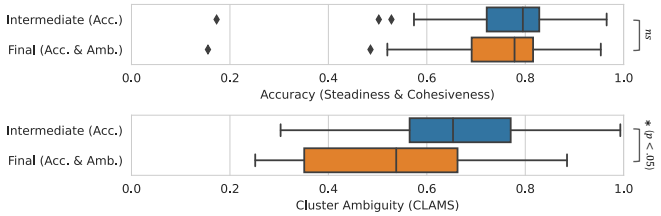


Fig. 8: The results of our evaluation on AmbReducer (Sect. 6.1.2). We compared the accuracy and cluster ambiguity of the intermediate (optimized solely on accuracy; blue) and final (optimized based on both accuracy and cluster ambiguity; orange) embeddings made by AmbReducer. We found that there is no significant difference (ns) between the final and intermediate results in terms of accuracy (top), while the cluster ambiguity is significantly reduced ($p < .05$) in the final results (bottom).

ambiguous cluster structure, which can potentially harm the reliability of the analysis despite high accuracy (Sect. 1). For the valid analysis of high-dimensional data, an embedding having both high accuracy and low cluster ambiguity is required.

Here, we introduce an optimization system that can discover DR embeddings satisfying both high accuracy and low ambiguity, which we call AmbReducer. Given the input high-dimensional data Z, DR technique t, and DR accuracy metric m, AmbReducer first finds the *intermediate* embedding that maximizes accuracy. This is done by searching a hyperparameter setting $h_1 = \arg\max_{h \in H} m(t(Z,h))$, where H denotes the set of every possible hyperparameter setting and $t(Z,h)$ represents the embedding made with t using hyperparameter setting h. We use Bayesian optimization [71] over H while maximizing $m(t(Z,h))$. Then, the system searches for the *final* embedding that maintains accuracy while reducing cluster ambiguity compared to the intermediate embedding. Formally, this is done by finding hyperparameter setting $h_2 = \arg\max_{h \in H} loss(h, h_1)$, where $loss$ is defined as

$$loss(h, h_1) = \begin{cases} CA(t(X,h)) & \text{if} \quad \mid m(t(Z,h_1)) - m(t(Z,h)) \mid < \tau \\ \infty & \text{if} \quad \mid m(t(Z,h_1)) - m(t(Z,h)) \mid > \tau \end{cases}.$$

Here, CA denotes CLAMS score, and τ denotes the threshold parameter, which can be set by users. τ adjusts the tradeoff between accuracy and cluster ambiguity. If we set small τ, the final embedding made by h_2 tends to have high accuracy but has fewer chances to reduce ambiguity. The proper value of τ highly depends on the target application and domain. For example, the more analysts involved, the greater τ should be to give more room to reduce ambiguity. Automatically determining τ is important future work that extends the applicability of AmbReducer.

6.1.2 Evaluation

Objectives and Design: We wanted to verify the effectiveness of AmbReducer. We hypothesized that AmbReducer substantially reduces cluster ambiguity while maintaining accuracy. To validate the hypothesis, we prepared 30 high-dimensional datasets, sampled from the datasets we used in our main study Sect. 5.2, and applied AmbReducer. We collected the intermediate (based on h_1) and final (based on h_2) embeddings, and checked how the accuracy and cluster ambiguity varied between the two groups. Fig. 7 depicts the subset of two groups.
Hyperparameter Setting: Our hyperparameter setting was as follows. For the DR technique, we used UMAP, as UMAP is widely used for cluster analysis in practice [8, 63]. We used Steadiness & Cohesiveness [32] (S&C) as accuracy measures, as they are specially developed to measure the accuracy in preserving cluster structure. We used the F1 score of S&C to represent the accuracy of DR, which ranges from 0 to 1. Finally, we arbitrarily set τ as 0.05.
Results and Discussions: We used a t-test to check whether there is a significant accuracy (measured by S&C) or cluster ambiguity (measured by CLAMS) difference between the final and intermediate embeddings. Although there was no significant difference in terms of accuracy ($t(58) = .455, p = .650$), we found that the cluster ambiguity of the final embeddings was significantly lower than the one of the intermediates ($t(58) = 2.55, p < .05$). The result verifies our hypothesis, verifying the effectiveness of AmbReducer and the applicability of CLAMS.

6.2 Finding Reliable Datasets for Clustering Benchmark

We demonstrate an experiment validating CLAMS's effectiveness in selecting reliable datasets for benchmarking clustering techniques [40]. We hypothesized that if the cluster ambiguity of a dataset (a scatterplot in this case) is high, then the clustering benchmark using the dataset is unreliable. Our assumption is that if the dataset is ambiguous, which means that there is no explicit cluster structure to be found, the performance of clustering techniques to "find the cluster" will be insufficiently compared. Thus, clustering benchmarking with the ambiguous dataset will likely provide an arbitrary ranking of techniques, which harms the credibility of the evaluation. To verify the hypothesis, we prepared three sets of datasets having different cluster ambiguity levels and checked the stability of the rankings made by different datasets in each set. The detailed explanation of the experiment is as follows:
Objectives and Design: We wanted to check whether the cluster ambiguity of benchmark datasets affects the stability of the ranking of clustering techniques, which we use as a proxy for the reliability of the evaluation. We prepared three sets of datasets with the same cardinality: $\mathbf{P}_h$, $\mathbf{P}_m$, and $\mathbf{P}_l$, with each having high, middle, and low cluster ambiguity, respectively. For each set, we obtained rankings of clustering techniques, where each was made based on an individual dataset. We then checked how the rankings correlate with each other by averaging the pairwise rank correlation computed by Spearman's ρ.
Datasets: We used 60 scatterplots we gathered in the main evaluation for CLAMS (see Sect. 5.2). We sorted the scatterplots based on CLAMS score in descending order, and then assigned top-20, middle-20, and bottom-20 scatterplots to $\mathbf{P}_h$, $\mathbf{P}_m$, and $\mathbf{P}_l$, respectively.
Clustering Techniques: We used eight techniques: HDBSCAN [14], DBSCAN, K-Means, X-Means [53], Birch [85], and Agglomerative clustering [48] with single, average, and complete linkage.
Measurements: We used the Silhouette coefficient [60] and Calinski-Harabasz index [12] to measure the performance of clustering techniques. To ensure that the evaluation reflected the maximum capability of clustering techniques, we ran Bayesian optimization and used the best score obtained while following the hyperparameter range setting of Jeon et al. [31]. Refer to Appendix C for detailed settings.
Results and Discussions: Fig. 9 depicts the results. We used a one-way ANOVA to analyze how the rank correlation scores vary due to the used set of datasets ($\mathbf{P}_l$, $\mathbf{P}_m$, $\mathbf{P}_h$). We used Tukey's HSD for post hoc analysis. For both the clustering metrics, the set of datasets had a significant effect (Silhouette: $F(3, 567) = 54.1, p < .001$; Calinski-Harabasz: $F(3, 567) = 44.0, p < .001$). Post-hoc analysis revealed that the rank correlation scores over $\mathbf{P}_l$ and $\mathbf{P}_m$ were significantly

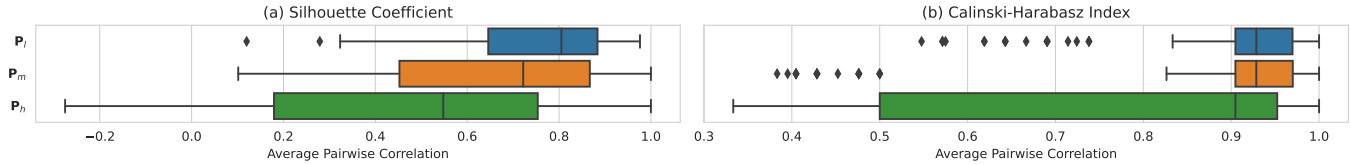


Fig. 9: Results of the experiment demonstrating CLAMS's effectiveness in picking reliable clustering benchmark datasets (Sect. 6.2). We first prepared a set of datasets (bottom-20 ($\mathbf{P}_l$; blue), middle-20 ($\mathbf{P}_m$; orange) or top-20 ($\mathbf{P}_h$; green) subset of 60 scatterplots we used in the main study based on CLAMS score) and clustering metrics (Silhouette Coefficient (a) or Calinski-Harabasz Index (b)). For each set, we compute the rankings of clustering techniques over individual datasets within the set based on clustering metrics. We then compute the within-set stability by assessing rank correlations of the rankings pairwisely. In general, the set consisting of datasets with less ambiguity showed better stability, confirming the hypothesis that datasets with less cluster ambiguity are more reliable for benchmarking clustering techniques.

higher than the ones over $\mathbf{P}_h$ for both metrics ($p < .001$ for every case). Meanwhile, scores over $\mathbf{P}_l$ were significantly higher than the ones over $\mathbf{P}_m$ for the Silhouette coefficient ($p < .001$), but not for the Calinski-Harabasz index ($p = 0.301$). Such results indicate that using datasets with less cluster ambiguity makes the clustering evaluation more stable, confirming our hypothesis and validating the effectiveness of CLAMS in picking reliable datasets for the clustering benchmarks.

7 Discussion

Our approach presents a system to identify and assess the ambiguity of cluster structure in scatterplots. We provided preliminary steps toward a comprehensive automation system to simulate human judgments in scatterplots. In this section, we discuss the implications of our measure design and experiments including opportunities for future work.

7.1 Proxy for Human Perception

In this work, we provided preliminary steps toward developing the model to identify and rank intrinsic variability in visual perception. Evaluating variability based on human participation is not always feasible, scalable, or robust considering the diversity of data characteristics, visualization designs, and other factors that might influence people's perceptions. There are simply too many sources of variance to account for. CLAMS, which automatically generates a score simulating human judgments, offers a scalable and robust alternative to an approach that attempts to account for every possible source of variance. Our work demonstrates the promise of using statistical modeling to infer patterns over a corpus of human estimates. Extending our approach to other perceptual tasks (e.g., outlier detection) would bring the notion of perceptual variability to more complex and practical applications [47].

Toward this end, we provide a demo interface that measures and explains cluster ambiguity. We believe that the interface will enhance the usability of CLAMS and moreover serve as a preliminary step toward future applications considering perceptual variability.

7.2 Credibility of the User Study-Based Design Process

While designing CLAMS, we identified several factors that play a vital role in cluster perception through a user study (Sect. 4.1), and built a regression module estimating human cluster perception based on study findings (Table 1). These factors support building a robust module that accurately reflects how people identify clusters in data analysis (Sect. 4). Our ablation study (Sect. 5.1) verifies the importance of these factors, validating the credibility of using human strategies to inform model parameters. Moreover, the main study (Sect. 5.2) showed that the measure built upon such a design process reliably estimates the ambiguity of scatterplots with a wide range of cluster patterns.

Our study asked participants about how the characteristics of datasets influence the ambiguity of a scatterplot (see Sect. 5.2). The findings from the interview validate the reliability of our design process. As shown in Fig. 6, participants identified the factors that we have considered in our model (see Table 1), such as density, proximity, shape, and how clusters are located and related to each other (i.e., distribution). Note that such results also match well with prior findings [57, 62, 67].

7.3 *Ambiguity* of Ambiguity

While our interview revealed common factors that influence cluster perception (Sect. 7.2), it also showed that the perceived influence of different factors on ambiguity varies among participants. In other words, the definition of cluster ambiguity is *ambiguous* to each participant. This observation depends on what factors are given more importance by different participants when determining a scatterplot's ambiguity. For example, P01 said, *"I find the shape of the group of points to be the main reason in identifying and separating clustering. If they have salient separable shapes, they are more clear"*, and P07 noted, *"Proximity and concentration of the points help me separate the clusters quickly"*. Fig. 6 also supports this finding, showing that no single factor received complete agreement from the participants (see the white transparent bars). Considering the variability across participants, we conclude that ambiguity cannot be determined by a single person but must instead reflect a population of individuals. The fact that the performance of human annotators (i.e., participants) in predicting ground truth ambiguity varied in our main study (Sect. 5.2) supports this claim. This observation indicates that, inevitably, multiple subjects are required for assessing ambiguity through human resources, thereby underscoring the importance of our automated solution.

7.4 Limitations and Future Work

CLAMS outperformed computational approaches and more than 50% of participants (Sect. 5.2) in precisely estimating cluster ambiguity. However, there are several opportunities to improve CLAMS. For example, we found that CLAMS erroneously considers scatterplots with more numbers to have less ambiguity, as seen in Fig. 3—the top eight clear scatterplots (top row) generally have more clusters than the top eight ambiguous ones (bottom row). We quantitatively demonstrate this bias in Appendix D. We can also improve the method of aggregating pairwise ambiguity scores, which is currently a naïve average. Considering cluster topology or their pairwise distances during the aggregation may better reflect human visual perception.

Another open direction is to generalize CLAMS. We can extend CLAMS to consider visual encoding (e.g., size, shape, and color of marks) of scatterplots by revising the features feed into our regression module. We may also improve our measure to deal with nested clusters by adopting hierarchical clustering algorithms [48] in place of GMM. We are also interested in broadening the concept of ambiguity to encompass general visualization. For example, we can model the perceptual variability that may arise when examining cluster structures of high-dimensional data via scatterplot matrices or parallel coordinates. Moreover, we aim to develop a model estimating ambiguity in various perceptual tasks, including outlier detection and trend analysis.

8 Conclusion

We introduce CLAMS, a VQM for estimating the cluster ambiguity of a monochrome scatterplot, which originally required expensive human resources to compute. To serve as a proxy for human perception across a wide range of cluster patterns, our measure is designed based on a qualitative user study and is trained over perceptual data. Through quantitative evaluations, we verified that CLAMS outperforms automatic competitors while showing competitive performance with human annotators. Our research findings not only demonstrate current applications but also invite discourse on potential future applications that capitalize on the concept of ambiguity. In summary, our work represents a significant advancement toward developing a comprehensive framework that elucidates the phenomenon of ambiguity in visualization.

Acknowledgments

This work was supported by NAVER Corporation (Cloud Data Box), by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2023R1A2C200520911), and by the National Science Foundation under grant No. 2127309 to the Computing Research Association for the CIFellows project, NSF IIS-2046725, NSF IIS-1764089, NSF IIS-2320920, NSF IIS-2233316, and NSF III-2316496. The authors wish to thank Sungbok Shin, Yun-Hsin Kuo, Seokhyeon Park and SNU-HCIL / UNC-VisuaLab members for their valuable feedback. We would also like to appreciate Michaël Aupetit for providing ClustMe dataset.

References

[1] Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pp. 3659–3669, 2016. doi: 10.1145/2858036.2858155 2, 6

[2] M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail. Clustme: A visual quality measure for ranking monochrome scatterplots based on cluster patterns. *Computer Graphics Forum*, 38(3):225–236, 2019. doi: 10.1111/cgf.13684 1, 2, 3, 4, 6

[3] M. Abbas, E. Ullah, A. Baggag, H. Bensmail, M. Sedlmair, and M. Aupetit. Clustrank: a visual quality measure trained on perceptual data for sorting scatterplots by cluster patterns, 2021. 3, 4

[4] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization (InfoVis)*, 2005. doi: 10.1109/INFVIS.2005.1532136 1, 2

[5] F. G. Ashby and W. W. Lee. Perceptual variability as a fundamental axiom of perceptual science. In *Advances in Psychology*, vol. 99, pp. 369–399. Elsevier, 1993. doi: 10.1016/S0166-4115(08)62778-8 2

[6] M. Aupetit and M. Sedlmair. Sepme: 2002 new visual separation measures. In *IEEE Pacific Visualization Symposium (PacificVis)*, pp. 1–8, 2016. doi: 10.1109/PACIFICVIS.2016.7465244 2

[7] M. Aupetit, M. Sedlmair, M. Abbas, A. Baggag, and H. Bensmail. Toward perception-based evaluation of clustering techniques for visual analytics. In *IEEE Visualization Conference (VIS)*, pp. 141–145, 2019. doi: 10.1109/VISUAL.2019.8933620 1, 2, 3, 4

[8] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37(1):38–44, 2019. doi: 10.1038/nbt.4314 7, 8

[9] 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. *Computer Graphics Forum*, 37(3):625–662, 2018. doi: 10.1111/cgf.13446 2

[10] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE TVCG*, 17(12):2203–2212, 2011. doi: 10.1109/TVCG.2011.229 2

[11] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE TVCG*, 19(12):2376–2385, 2013. doi: 10.1109/TVCG.2013.124 2

[12] T. Caliński and J. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. 8

[13] A. C. Cameron and F. A. Windmeijer. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics*, 77(2):329–342, 1997. doi: 10.1016/S0304-4076(96)01818-0 5

[14] R. J. G. B. Campello, D. Moulavi, and J. Sander. Density-based clustering based on hierarchical density estimates. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 160–172, 2013. doi: 10.1007/978-3-642-37456-2_14 8

[15] T. N. Dang and L. Wilkinson. Transforming scagnostics to reveal hidden features. *IEEE TVCG*, 2014. doi: 10.1109/TVCG.2014.2346572 2, 6

[16] R. Davis, X. Pu, Y. Ding, B. D. Hall, K. Bonilla, M. Feng, M. Kay, and L. Harrison. The risks of ranking: Revisiting graphical perception to model individual differences in visualization performance. *IEEE Transactions on Visualization and Computer Graphics*, 2022. doi: 10.1109/TVCG.2022.3226463 2

[17] M. desJardins, J. MacGlashan, and J. Ferraioli. Interactive visual clustering. In *International Conference on Intelligent User Interfaces*, IUI '07, p. 361–364, 2007. doi: 10.1145/1216295.1216367 6

[18] A. Dillon and C. Watson. User analysis in hci—the historical lessons from individual differences research. *International Journal of Human-Computer Studies*, 45(6):619–637, 1996. 2

[19] 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 Transactions on Visualization and Computer Graphics*, 21(1):81–94, 2014. doi: 10.1109/TVCG.2014.2330617 1

[20] R. Etemadpour, B. Olk, and L. Linsen. Eye-tracking investigation during visual analysis of projected multidimensional data with 2d scatterplots. In *Information Visualization Theory and Applications (IVAPP)*, pp. 233–246, 2014. doi: 10.5220/0004675802330246 2

[21] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, vol. 28, 2015. 5

[22] C. Fu, Y. Zhang, D. Cai, and X. Ren. Atsne: Efficient and robust visualization on gpu through hierarchical optimization. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, p. 176–186, 2019. doi: 10.1145/3292500.3330834 7

[23] T. Fujiwara, Y.-H. Kuo, A. Ynnerman, and K.-L. Ma. Feature learning for nonlinear dimensionality reduction toward maximal extraction of hidden patterns, 2022. doi: 10.48550/ARXIV.2206.13891 7

[24] L. A. Goodman. Snowball sampling. *The Annals of Mathematical Statistics*, 32(1):148–170, 1961. 6

[25] M. Halkidi and M. Vazirgiannis. Clustering validity assessment: finding the optimal partitioning of a data set. In *IEEE International Conference on Data Mining*, pp. 187–194, 2001. doi: 10.1109/ICDM.2001.989517 3

[26] S. Hartwig, C. van Onzenoodt, P. Hermosilla, and T. Ropinski. Clusternet: A perception-based clustering model for scattered data. *arXiv preprint arXiv:2304.14185*, 2023. 2

[27] X. He, K. Zhao, and X. Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, 212:106622, 2021. doi: 10.1016/j.knosys.2020.106622 5

[28] R. Hoskin, C. Berzuini, D. Acosta-Kane, W. El-Deredy, H. Guo, and D. Talmi. Sensitivity to pain expectations: A bayesian model of individual differences. *Cognition*, 182:127–139, 2019. doi: 10.1016/j.cognition.2018.08.022 2

[29] J. Huang and R. Sekuler. Distortions in recall from visual memory: Two classes of attractors at work. *Journal of Vision*, 10(2):24–24, 2010. doi: 10.1167/10.2.24 2

[30] H. Jeon, M. Aupetit, S. Lee, H.-K. Ko, Y. Kim, and J. Seo. Distortion-aware brushing for interactive cluster analysis in multidimensional projections, 2022. doi: 10.48550/ARXIV.2201.06379 6

[31] H. Jeon, M. Aupetit, D. Shin, A. Cho, S. Park, and J. Seo. Sanity check for external clustering validation benchmarks using internal validation measures. *arXiv preprint*, 2022. 6, 7, 8

[32] H. Jeon, H.-K. Ko, J. Jo, Y. Kim, and J. Seo. Measuring and explaining the inter-cluster reliability of multidimensional projections. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):551–561, 2022. doi: 10.1109/TVCG.2021.3114833 3, 7, 8

[33] H. Jeon, H.-K. Ko, S. Lee, J. Jo, and J. Seo. Uniform manifold approximation with two-phase optimization. In *IEEE Visualization and Visual Analytics (VIS)*, pp. 80–84, 2022. doi: 10.1109/VIS54862.2022.00025 7

[34] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. Chau. Activis: Visual exploration of industry-scale deep neural network models. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):88–97, 2018. doi: 10.1109/TVCG.2017.2744718 1

[35] J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964. doi: 10.1007/BF02289565 6

[36] J. Lee and M. Verleysen. *Nonlinear dimensionality reduction*, vol. 1. Springer, 2007. doi: 10.1007/978-0-387-39351-3 3

[37] J. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction*. Springer-Verlag New York, 2007. doi: 10.1007/978-0-387-39351-3 7

[38] A. Likas, N. Vlassis, and J. J. Verbeek. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461, 2003. Biometrics. doi: 10.1016/S0031-3203(02)00060-2 3, 6

[39] R. Liu, H. Wang, and X. Yu. Shared-nearest-neighbor-based clustering by fast search and find of density peaks. *Information Sciences*, 450:200–226, 2018. doi: 10.1016/j.ins.2018.03.031 3

[40] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu. Understanding of internal clustering validation measures. In *IEEE International Conference on Data Mining*, pp. 911–916, 2010. doi: 10.1109/ICDM.2010.35 3, 5, 8

[41] Z. Liu, R. J. Crouser, and A. Ottley. Survey on individual differences in visualization. *Computer Graphics Forum*, 39(3):693–712, 2020. doi: 10.1111/cgf.14033 2

[42] M. Lu, S. Wang, J. Lanir, N. Fish, Y. Yue, D. Cohen-Or, and H. Huang. Winglets: Visualizing association with uncertainty in multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):770–779, 2020. doi: 10.1109/TVCG.2019.2934811 1

[43] Y. Ma, A. K. Tung, W. Wang, X. Gao, Z. Pan, and W. Chen. Scatternet: A deep subjective similarity model for visual analysis of scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, pp:0, 2018. doi: 10.1109/TVCG.2018.2875702 2

[44] H. Maassen and J. B. M. Uffink. Generalized entropic uncertainty relations. *Physical Review Letters*, 60:1103–1106, Mar 1988. doi: 10.1103/PhysRevLett.60.1103 3

[45] L. McInnes and J. Healy. Accelerated hierarchical density based clustering. In *IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 33–42, 2017. doi: 10.1109/ICDMW.2017.12 3

[46] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. 6, 7

[47] D. Moritz, C. Wang, G. L. Nelson, H. Lin, A. M. Smith, B. Howe, and J. Heer. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE TVCG*, 25(1):438–448, 2018. doi: 10.1109/TVCG.2018.2865240 9

[48] D. Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint*, 2011. 8, 9

[49] A. Narayan, B. Berger, and H. Cho. Assessing single-cell transcriptomic variability through density-preserving data visualization. *Nature Biotechnology*, 39(6):765–774, 2021. doi: 10.1038/s41587-020-00801-7 6

[50] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, 2019. doi: 10.1109/TVCG.2018.2846735 1, 7

[51] B. Ons, W. De Baene, and J. Wagemans. Subjectively interpreted shape dimensions as privileged and orthogonal axes in mental shape space. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2):422, 2011. doi: 10.1037/a0020405 2

[52] K. F. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. 6

[53] D. Pelleg, A. Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *International Conference on Machine Learning (ICML)*, vol. 1, pp. 727–734, 2000. 8

[54] B. Pinna. New gestalt principles of perceptual organization: An extension from grouping to shape and meaning. *Gestalt Theory*, 2010. 1

[55] S. Prion and K. A. Haerling. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clinical Simulation in Nursing*, 10(10):535–536, 2014. doi: 10.1016/j.ecns.2014.07.005 7

[56] G. J. Quadri, J. A. Nieves, B. Wiernik, and P. Rosen. Automatic scatterplot design optimization for clustering identification. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–16, 2022. doi: 10.1109/TVCG.2022.3189883 2

[57] G. J. Quadri and P. Rosen. Modeling the influence of visual density on cluster perception in scatterplots using topology. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1829–1839, 2021. doi: 10.1109/TVCG.2020.3030365 1, 2, 3, 9

[58] M. Rezaei and P. Fränti. Set matching measures for external cluster validity. *IEEE Transactions on Knowledge and Data Engineering*, 28(8):2173–2186, 2016. doi: 10.1109/TKDE.2016.2551240 7

[59] A. Rosenberg and J. Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420, 2007. 7

[60] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. doi: 10.1016/0377-0427(87)90125-7 3, 8

[61] S. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. doi: 10.1126/science.290.5500.2323 6

[62] Y. Sadahiro. Cluster perception in the distribution of point objects. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 34(1):49–62, 1997. 7, 9

[63] M. Sánchez-Rico and J. M. Alvarado. Dimensionality reduction techniques as a preliminary step to cluster analysis: A comparison between pca, t-sne and umap. In *European Congress of Methodology*, 2020. 8

[64] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 2018. doi: 10.1109/TVCG.2017.2744184 1

[65] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *International Conference on Distributed Computing Systems Workshops*, pp. 166–171, 2011. doi: 10.1109/ICDCSW.2011.20 3

[66] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, pp. 461–464, 1978. doi: 10.1214/aos/1176344136 3

[67] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Computer Graphics Forum*, 34(3):201–210, 2015. doi: 10.1111/cgf.12632 2, 9

[68] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Computer Graphics Forum*, 2012. doi: 10.1111/j.1467-8659.2012.03125.x 7

[69] W. Shannon, R. Culverhouse, and J. Duncan. Analyzing microarray data using cluster analysis. *Pharmacogenomics*, 4(1):41–52, 2003. doi: 10.1517/phgs.4.1.41.22581 1

[70] J. Shi and Z. Luo. Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples. *Computers in Biology and Medicine*, 40(8):723–732, 2010. doi: 10.1016/j.compbiomed.2010.06.007 1

[71] J. Snoek, H. Larochelle, and R. Adams. Practical bayesian optimization of machine learning algorithms. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds., *Advances in Neural Information Processing Systems*, vol. 25. Curran Associates, Inc., 2012. 8

[72] D. Steinley. Properties of the hubert-arable adjusted rand index. *Psychological Methods*, (3):386, 2004. doi: 10.1037/1082-989X.9.3.386 7

[73] J. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319 6, 7

[74] L. van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. 6, 7

[75] J. Venna and S. Kaski. Local multidimensional scaling. *Neural Networks*, 19(6):889–899, 2006. Advances in Self Organising Maps - WSOM'05. doi: 10.1016/j.neunet.2006.05.014 3

[76] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *J. of Machine Learning Research*, pp. 2837–2854, 2010. 7

[77] 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*, 25(1):820–829, 2019. doi: 10.1109/TVCG.2018.2864912 1

[78] M. Wertheimer. Untersuchungen zur lehre von der gestalt. *Psychologische Forschung*, 1(1):47–58, 1922. doi: 10.1515/gth-2017-0007 1

[79] L. Woolhouse and R. Bayne. Personality and the use of intuition: Individual differences in strategy and performance on an implicit learning task. *European Journal of Personality*, 14(2):157–169, 2000. 2

[80] J. Wu, H. Xiong, and J. Chen. Adapting the right measures for k-means clustering. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, p. 877–886, 2009. doi: 10.1145/1557019.1557115 6, 7

[81] J. Xia, W. Lin, G. Jiang, Y. Wang, W. Chen, and T. Schreck. Visual clustering factors in scatterplots. *IEEE Computer Graphics and Applications*, 41(5):79–89, 2021. doi: 10.1109/MCG.2021.3098804 1

[82] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu. Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):529–539, 2022. doi: 10.1109/TVCG.2021.3114694 1

[83] H. Xiong and Z. Li. Clustering validation measures. In *Data Clustering*, pp. 571–606. 2018. doi: 10.1201/9781315373515-23 3

[84] J. Zaman, A. Chalkia, A.-K. Zenses, A. S. Bilgin, T. Beckers, B. Vervliet, and Y. Boddez. Perceptual variability: Implications for learning and generalization. *Psychonomic Bulletin & Review*, 28(1):1–19, 2021. doi: 10.3758/s13423-020-01780-1 2

[85] T. Zhang, R. Ramakrishnan, and M. Livny. Birch: An efficient data clustering method for very large databases. *SIGMOD Record*, 25(2):103–114, jun 1996. doi: 10.1145/235968.233324 8

[86] C. Ziemkiewicz and R. Kosara. Preconceptions and individual differences in understanding visual metaphors. *Computer Graphics Forum*, 28(3):911–918, 2009. doi: 10.1111/j.1467-8659.2009.01442.x 2