Toward Perception-Based Evaluation of Clustering Techniques for Visual Analytics

Michaël Aupetit* QCRI, HBKU, Doha
Michael Sedlmair† VISUS, University of Stuttgart
Mostafa M. Abbas‡ QCRI, HBKU, Doha
Abdelkader Baggag§ QCRI, HBKU, Doha
Halima Bensmail¶ QCRI, HBKU, Doha

ABSTRACT

Automatic clustering techniques play a central role in Visual Analytics by helping analysts to discover interesting patterns in high-dimensional data. Evaluating these clustering techniques, however, is difficult due to the lack of universal ground truth. Instead, clustering approaches are usually evaluated based on a subjective visual judgment of low-dimensional scatterplots of different datasets. As clustering is an inherent human-in-the-loop task, we propose a more systematic way of evaluating clustering algorithms based on quantification of human perception of clusters in 2D scatterplots. The core question we are asking is in how far existing clustering techniques align with clusters perceived by humans. To do so, we build on a dataset from a previous study [1], in which 34 human subjects labeled 1000 synthetic scatterplots in terms of whether they could see one or more than one cluster. Here, we use this dataset to benchmark state-of-the-art clustering techniques in terms of how far they agree with these human judgments. More specifically, we assess 1437 variants of K-means, Gaussian Mixture Models, CLIQUE, DBSCAN, and Agglomerative Clustering techniques on these benchmarks data. We get unexpected results. For instance, CLIQUE and DBSCAN are at best in slight agreement on this basic cluster counting task, while model-agnostic Agglomerative clustering can be up to a substantial agreement with human subjects depending on the variants. We discuss how to extend this perception-based clustering benchmark approach, and how it could lead to the design of perception-based clustering techniques that would better support more trustworthy and explainable models of cluster patterns.

Index Terms: H.1.2 [User/Machine Systems]: Human factors; I.5.3 [Clustering]: Algorithms; I.5.2 [Design Methodology]: Pattern analysis

1 INTRODUCTION

Cluster analysis is a pivotal task for domain experts to categorize and abstract concepts from raw signals. A cluster can be loosely defined as a group of data more similar to each other than they are to data from any other groups. A clustering technique produces such groups based on specific definitions of data similarity and modeling assumptions [17, 36, 38].

Despite a long history of research in this area, understanding the outcome of clustering techniques and evaluating their quality are still challenging problems [4, 19, 22]. Evaluating clustering techniques is difficult due to the lack of a universal ground truth [35]. Thus, the most typical approach has been by subjectively judging their results in 2D scatterplots. However, this approach is not systematic and does not generalize well across different datasets and users. To overcome these issues, researchers have proposed to evaluate clustering techniques based on heuristics and mathematical frameworks [30]. While this is good in terms of objectivity, it does not involve human judgments in any way. Yet, clustering is essentially a task that heavily relates to the perception of what humans deem as similar and dissimilar. Without a good alignment of clustering results with human expectations the understanding and trust in these automatic techniques will remain low [32], and the use of the best techniques might be impeded (Kmeans is still widely used despite its known limitations).

The goal of our paper is to propose a new way of benchmarking clustering algorithms based on human perception, and in doing so open the way for the design of clustering techniques that better align with the perception of the human analyst. The basic idea is to gather a large and reliable set of human judgments of clusters in 2D scatterplots. Such judgments could be gathered through controlled user studies or online studies. Such data would on the one hand reflect the human nature of the clustering task. On the other hand, however, it rules out individual biases due the number of participants that provide judgments for the same clustering patterns. Clustering techniques can then be benchmarked by computing their performance on the same 2D scatterplots and check their consistency with the group of human raters.

To illustrate this approach, we use data collected in a recent study on class separability in 2D scatterplots [1]. In this data, 34 human subjects were tasked to decide if they can see one or more than one clusters in 1000 monochrome scatterplots. While the previous work used this data for developing a new visual quality measure of grouping patterns in monochrome scatterplots, we use it here to benchmark existing clustering approaches.

Our claim is that this basic counting task can be used as a first filter to assess clustering techniques by comparing the number of clusters they output to the one found by the group of human raters for the same data. Techniques that do not pass this test would likely be not trustworthy for non-expert users to be used on more complex cluster patterns in 2 dimensions. The remaining techniques would further pass a set of cascading filters, getting more credit the further they go into these refined perception-based evaluations, possibly considering higher-dimensional spaces as well. Building such a cascading evaluation pipeline is an interesting challenge for the future.

2 RELATED WORK

We review different approaches for evaluating clustering techniques and discuss briefly how they relate to our proposition to rely on human perception.

Clustering validation indices. Many clustering validation indices exist either internal or external, quantifying the within and between clusters similarities, or the cluster stability in various ways based on different mathematical frameworks [4, 19, 22, 30, 35]. Our general approach and the specific benchmark we propose are the first to con-

*e-mail: maupetit@hbku.edu.qa
†e-mail: Michael.Sedlmair@visus.uni-stuttgart.de
‡e-mail: mohamza@hbku.edu.qa
§e-mail: abaggag@hbku.edu.qa
¶e-mail: hbensmail@hbku.edu.qa

IEEE VIS 2019: Short Papers
20-25 October, Vancouver, BC, Canada
978-1-7281-4941-7/19/$31.00 ©2019 IEEE

sider human perception in a quantitative way to evaluate clustering techniques.

Subjective eyeballing evaluation. Another very common, but only qualitative non-systematic approach, is to ask reviewers and readers of scientific reports of novel clustering techniques, to eyewitness the quality of the clustering results by looking at a handle of class color-coded scatterplots using the *Match Clusters and Classes in Map* described in [10, 28].

In a practical setting, users often resort to their own subjective visual judgment based on low-dimensional scatterplot views of the data, solely or in complement to existing quality metrics. This type of evaluation is common to evaluate dimension reduction techniques through the way they represent cluster patterns [5]. For lack of a consensus on what is a good quality clustering technique, we propose to evaluate clustering techniques based on perceptual judgments more consistent with the way they are finally evaluated in practice, but benefiting the robust framework of controlled user studies to get a more objective benchmark.

Benchmark datasets Benchmark clustering datasets labeled by humans exist to support quantitative analysis of automatic clustering techniques [12, 15]; but the class labels are usually not reliable for clustering tasks because, for synthetic data, they are assigned by few designers with no collective validation [15]. And for real data typically used in supervised classification tasks [12], the labels come from external knowledge rather than the intrinsic cluster structure of the data. In our benchmark, we rely on a simpler cluster counting task where scatterplots are labeled by human subjects.

Collecting data of perception of patterns in scatterplots A set of works developed around the concept of Visual Quality Metrics [8].Several of them attempt to exploit the collected perceptual data to improve modeling techniques, either to model a similarity metric between scatterplots [3, 26] or to design new measures for class separation [7, 33]. A similar approach [1] tackles the case of cluster patterns using Gaussian Mixture Models and merging techniques. Our work builds on the data collected from this previous study to define a perception-based benchmark for clustering techniques.

3 A PERCEPTION-BASED BENCHMARK FOR CLUSTERING TASK

In this section, we discuss how the human subject experiment conducted in a previous study [1] can be used to define a perception-based benchmark for clustering techniques.

3.1 Human subject experiment

In the experiment described in full details in [1], 34 subjects were tasked to judge whether each of 1000 monochrome scatterplots, generated from a mixture of 2 Gaussian distributions with various parameters, displays one or more than one cluster. Five of these scatterplots are shown in Figure 1 with the corresponding percentage of human raters judging that these scatterplots display more than one cluster.

3.2 Assessing clustering techniques with perception-based data

We propose to run the clustering technique to be evaluated, on the data from each of the 1000 scatterplots, and to compute the number of clusters it detects, then to compare this number to the one counted by the human subjects. Below, we argue that these data and tasks are relevant for setting up such a perception-based benchmark for clustering techniques.

Basic cluster analysis task

The basic counting task of that human-subject experiment fits well with our objective of benchmarking clustering techniques.

Indeed, the number of clusters is a typical parameter of clustering techniques, and it is a primary characterization of cluster patterns in scatterplots [34]. In partition-based techniques like K-means [23] or model-based like Gaussian Mixture Models [17], this number is set directly as one of the parameters, while in hierarchical [21] or density-based techniques like DBSCAN [14] and Mean-Shift [11], thresholds or scaling parameters control the number of clusters indirectly. The best number of clusters can be selected as one that makes the clustering technique produce a partition which optimizes some quality measure.

Figure 1: Subset of the 1000 scatterplots judged by the 34 human subjects with the percentage of them judging they display more than one cluster.

In short, the correct number of clusters is a necessary by-product of a good clustering with respect to some measure of quality, and so an incorrect number of clusters can prove a bad clustering. Therefore counting clusters can be viewed as a proxy for the clustering quality.

Low dimension and low pattern complexity

The cluster patterns displayed in the 1000 scatterplots are only 2-dimensional, generated by a parametric family of Gaussian Mixtures with 2 Gaussian components.

However, this setting allows generating a large variety of non trivial patterns (Figure 1) that could serve as a first base filtering step for clustering techniques: the ones which do not get sufficient agreement with human subjects, would be assumed not trustworthy to provide good quality clustering on more complex cluster patterns.

4 BENCHMARKING CLUSTERING TECHNIQUES

We first describe the clustering techniques, then how we use the human judgment data to benchmark them.

4.1 Clustering techniques

We selected 6 clustering techniques from available `R` packages and consider a total of 1437 variants summarized in the table 1. We focused on techniques for which the number of clusters is set automatically. `Gmeans` [20] and `Xmeans` [29] are variants of `Kmeans` [23] which search for clusters with convex shapes and equal density. We considered the Gaussian Mixture Models (model-based clustering) from the `Rmixmod` package which offers multiple criteria to find the parameters and the optimal number of components (clusters). We also tested the 7 non-parametric merging technique [22] which take a decision to merge two components of a GMM when they overlap too much. We considered the ground truth (GT) merging techniques used in the ClustMe study [1]: the 7 merging techniques applied to the parameters of the GMM which generated the 1000 datasets used in our benchmark. In all other cases, the GMM parameters and other clustering parameters are inferred from the data point actual coordinates in the 2D space (We do not consider the pixels of the scatterplot image graphically representing that 2D space). We also tested the CLIQUE [2] and DBSCAN [13] clustering techniques for different settings of their parameters. CLIQUE assumes clusters are made of union of high density rectangular areas, while DBSCAN considers clusters formed around core points having minimal *minPts* number of neighbors closer than a Euclidean distance *epsilon*. We aim to compare them with the GMM approach to investigate the impact of the model assumption in model-based clustering techniques.

At last we used the `R` package `NbClust` which offers to combine `agglomerative` clustering (model-agnostic clustering) with a large choice of metrics and aggregation methods, and propose 30 different selection criteria to select the optimal number of clusters.

We set the default range for the grid search of the optimal number of clusters between 1 and 5 for all techniques having a selection option, then assign rate 1 for clustering techniques (machine raters) finding a single cluster, rate 2 when more than one cluster is found,

and rate 0 when the technique did not provide a solution in reasonable time. Note that a clustering technique with a specific set of meta-parameters as listed in Table 1, is called a technique for short in the sequel. Some of the 624 variants for agglomerative clustering did not provide a result in reasonable time, ending up to 576 of them remaining in the final set.

We evaluate the usefulness of this benchmark using two criteria:
C1: We rank the clustering techniques based on the agreement value of each isolated one with the full set of 34 human raters, using the multinomial Vanbelle's Kappa κ_V index [31]. This index allows us to handle cases where the clustering technique gives rate 0 in disagreement with any human judgment on a scatterplot, without discarding that scatterplot. So all techniques are compared on the same basis against all the scatterplots. This benchmark will prove useful as *a way to rank the clustering techniques* if the κ_V indices are not equally distributed for the different techniques.
C2: We make 7 hypotheses based on predictions informed by the technical characteristics of the clustering techniques, as known by the authors. Our perception-based benchmark will prove useful if at least some of these hypotheses are not supported, showing a mismatch between the results we expect from the technical design of clustering techniques, and the one we observe when applied on the perception-based data.

We make the following technical hypotheses:
H1: We expect GMM to be better on average than other techniques given the data are generated from a GMM.
H2: We expect merging techniques based on the parameters of the generative model (GT) should be among the best techniques.
H3: We expect model-based techniques like GMM with model matching the one generated in the data, should be better than model-based with non matching model like CLIQUE and non model-based DBSCAN.
H4: We expect Gmeans performing similarly to the best GMM without merging, because Gmeans assumes a Normal distribution of each final clusters
H5: We expect Xmeans performing similarly to the best GMM with CEM inference as Xmeans and Kmeans are equivalent to GMM+CEM
H6: We expect Agglomerative Clustering should not be as good as any GMM because it might not handle smooth density properly
H7: We expect CLIQUE and DBSCAN should be better than Xmeans and Gmeans as they can handle non convex clusters

4.2 Benchmarking results

The figure 2 shows the distribution of Vanbelle Kappa κ_V index for the main families of techniques. Gaussian Mixture Models (GMM) with dipUni and dipTantrum merging techniques, EM parameter inference method, and ICL number of cluster selection method (see also figure 3 right), come first with Kappa index indicating between substantial and almost perfect agreement with human raters, as per the scale proposed by Landis and Koch [24]. We notice that the Demp technique selected in [1] based on the GMM having generated the data, rather than a more realistic GMM with parameters inferred from the data, is not as good as dipUni and dipTantrum when that more realistic setting is used. Then come most of the GT techniques and quite surprisingly, several variants of the Agglomerative Clustering (AggloClust) techniques are in substantial agreement with human raters (see top 30 table in supplemental material).

Regarding our hypotheses:
H1 is not supported: Despite the data are generated from a GMM, not all GMM-based techniques are in at least substantial agreement with human raters (Range from 0 to 0.81, with median 0.27).
H2 is partly supported: the predictive and ridgeUni merging techniques get a lower agreement than expected.
H3 is supported: CLIQUE and DBSCAN are very bad for the parameters we tested. The rectangular grid approach of CLIQUE or the core points density-based approach of DBSCAN do not allow a good fit with the Gaussian distribution as per human perception, despite the fact that both techniques are based on reasonable assumptions to approximate density-based clusters.
H4 is partly supported: Gmeans is in moderate agreement with human raters similar to the best subset of GMM techniques. Gmeans being faster than GMM could be recommended as a first approach to count clusters.
H5 is not supported: Xmeans happens to always find more than one cluster, leading to a poor agreement (0 kappa index).
H6 is partly supported: the median is close to 0 as in many cases, the agglomerative techniques provide either always one, or always more than one clusters for all scatterplots leading to poor agreement with a Kappa near 0. But 16% of the variants are at least in moderate and even substantial agreement with humans. The best setting is shown in the figure 3 left, with Duda threshold selection, `Average` aggregation metric, and either `Maximum` (L_∞ norm)(0.74), `Manhattan` (L_1 norm) (0.77), or `Euclidean`(L_2 norm)(0.78) base metric. Single linkage and median aggregation perform poorly while the ward aggregation method assumed to be better for unequal cluster densities is not the best in our setting.
H7 is partly supported: CLIQUE and DBSCAN are not able to go beyond slight agreement with human raters, while Gmeans reaches moderate agreement, possibly because it uses a Normality test in its hierarchical splitting process.

Several of our technically-based hypotheses are partially or not supported by the perception-based benchmark, and the distribution of the κ_V agreement index is clearly heterogeneous for the different techniques, fulfilling the criteria C1 and C2, proposed to evaluate the usefulness of the benchmark.

Table 1: List of variants of clustering techniques and their meta-parameters.

♯Var.	Tech.	Meta-parameters
1	`Gmeans`	None
1	`Xmeans`	None
72	`GMM`	Merg. {*dipUni*;*dipTantrum*;*demp*;*ridgeUni*;... ...*ridgeRatio*;*bhat*;*predictive*;*none*}
		Infer. 6m-1 GMM param. {*EM*;*SEM*;*CEM*}
		Select. m components {*BIC*;*ICL*;*NEC*}
7	`GT`	Merg. {*dipUni*;*dipTantrum*;*demp*;*ridgeUni*;... ...*ridgeRatio*;*bhat*;*predictive*}
		No inference from data
		No selection (m=2 components)
576	`Agglo.`	Metric {*euclidean*;*maximum*;*manhattan*}
		Aggreg. {*ward.D*;*ward.D2*;*single*;*complete*;... ...*average*;*mcquitty*;*median*;*centroid*}
		Select. {*all*} except {*frey*;*tau*;*gamma*;*gplus*}
380	`CLIQUE`	Grid {2;3;4;...;20}
		Density {0.05;0.1;0.15;...;1}
400	`DBSCAN`	minPts {3;4;...;10}
		epsilon {0.02;0.04;0.06;...;1}
1437		

5 Discussion

Our work serves mainly as a proof of concept, that we can obtain a quantitative assessment of clustering techniques on a human gold standard data. We further discuss the benefits and limitations of this work in the next section.

5.1 Is counting cluster enough?

The data we used [1] was encoded using a very basic class counting task, that is, deciding if there is one, or more than one cluster. By increasing the order of difficulty, naturally also the ecological validity of the study would increase, but at the same time also its

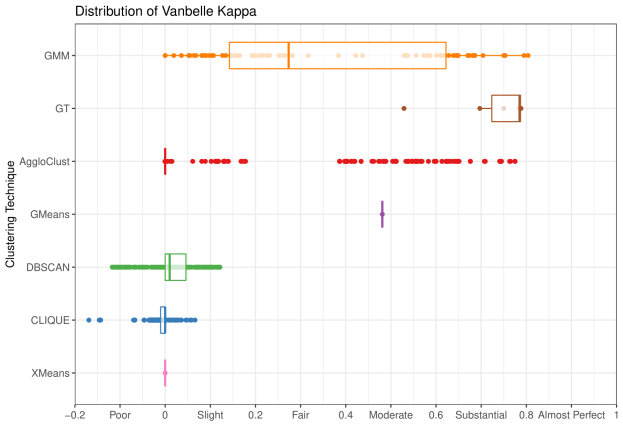

Figure 2: Box plots summarizing the distribution of Vanbelle Kappa index κ_V of all families of clustering techniques, ranked by decreasing maximum index from top to bottom. The index κ_V measures the agreement of each technique variant (dots) with the 34 human raters counting "one" or "more-than-one" clusters in the 1000 scatterplots.

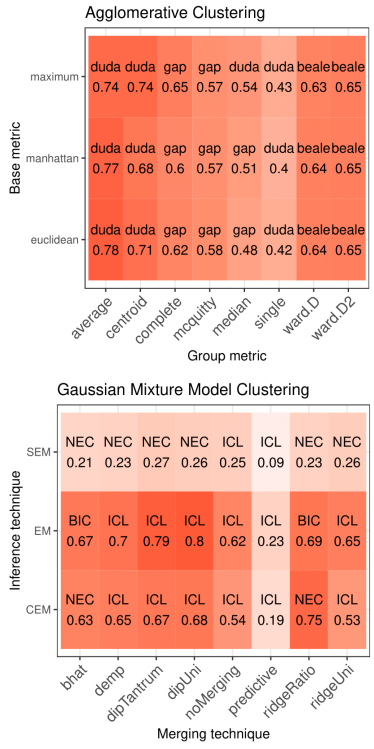

Figure 3: Variants of agglomerative clustering (top) and GMM (bottom) with the selection method leading to the maximum κ_V index (Redder color codes for higher agreement).

complexity and time needed. In the future it would be interesting to also investigate more complicated tasks such as counting the exact number of clusters, judging the absolute quality of a scatterplot, or interactively selecting clusters using a lasso.

While we deem counting clusters only as a baseline task, it resulted in several interesting findings. First, our baseline test revealed that not all clustering techniques can detect generic cluster patterns. Second, it also indicates that the techniques that failed this test (CLIQUE, DBSCAN, Xmeans, and some GMM and Agglomerative techniques) should not be considered as model-agnostic clustering techniques, while the others which are at least in substantial agreement with human raters ($\kappa_V > 0.6$) are trustworthy for these kind of cluster patterns, and candidates for being assessed against other sorts of cluster patterns (e.g non Gaussian, more clusters or dimensions).

5.2 Is 2D enough?

The data we used inherently came in 2D. We deem this a natural choice as 2D scatterplots as very standard charts for data scientist. Also, the Gestalt law of proximity [37] allows humans to detect cluster patterns pre-attentively in this sort of visualization. Nevertheless, it is an interesting question whether other dimensionalities might provide better results. Ideally, the evaluation of clustering techniques would take place in the actual high-dimensional space. However, here the human perception is the limiting factor. Going beyond 2D, 3D scatterplots, Scatterplot Matrices or parallel coordinate plots could be used, but would also require more expertise to understand [9] and interact with the visualization [6, 27]. In the other direction, 1-dimensional histograms might be used, but might loose much of the pattern complexity either due to over-plotting or to the smoothing introduced by the binning process. Extending the perception-based approach to higher dimensional spaces is clearly a challenge and we don't know yet how much it could be beneficial, but the surprising results we found in this simple case are a strong incentive to explore this approach further.

5.3 Beyond Gaussian Mixtures

Gaussian mixtures are used to model any continuous data distribution. We expect that covering the parameter space of this model enables generating a wide variety of cluster patterns. It seems we could follow this principle with more than 2 components to generate even more complex patterns still in 2 dimensions, but a linear increase of the number of parameters leads to an exponential increase of the parameter space to cover. Using some experiment design or active learning approach could be a way to explore such larger parameter space querying human subjects only where cluster number is not clear. Other generative models could be used to diversify the cluster patterns, for instance to generate manifold structures [16, 18].

5.4 Can we use crowdsourcing?

Another intersting question is how to gather the human data. The data used stemmed from a controlled experiment. Crowdsourcing would be a natural alternative, increasing the number of raters but at the same time lowering the reliability of the judgments [25]. Another intersting idea is to combine both approaches and, for instance, use lab data to validate crowdsourced data.

6 Conclusion and Future Work

As far as clustering techniques are designed to support human discovery, human beings are part of the decision process. Overall, we think that evaluating clustering techniques on perceptual data even if those are available only for low dimension space and low pattern complexity, could guide the design of more efficient clustering techniques for more complex patterns and possibly in higher dimension space. Moreover, it could be a way to improve trust in using such techniques because they would be designed based on human perception rather than heuristics or mathematical abstract concepts.

Beyond this first benchmark setting, we want to design perception-based benchmarks for more complex patterns and cluster analysis tasks, and possibly higher dimensional clusters, to determine which technique could be used to detect which type of cluster patterns *as perceived by humans*. Evaluating how this approach could improve non-experts trust in clustering techniques is also interesting for future research. This work starts paving the way in that direction.

Acknowledgments

M. Sedlmair was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project-ID 251654672 TRR 161.

REFERENCES

[1] 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 (Proc. EuroVis 2019)*, 2019.

[2] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. pp. 94–105. ACM Press, 1998.

[3] G. Albuquerque, M. Eisemann, and M. A. Magnor. Perception-based visual quality measures. In *IEEE VAST*, pp. 13–20. IEEE Computer Society, 2011.

[4] M. J. Amorim and M. G. M. S. Cardoso. Clustering stability and ground truth: Numerical experiments. In A. L. N. Fred, J. L. G. Dietz, D. Aveiro, K. Liu, and J. Filipe, eds., *KDIR*, pp. 259–264. SciTePress, 2015.

[5] M. Aupetit. Sanity check for class-coloring-based evaluation of dimension reduction techniques. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization, BELIV 2014, Paris, France, November 10, 2014*, pp. 134–141, 2014. doi: 10.1145/2669557.2669578

[6] M. Aupetit, N. Heulot, and J.-D. Fekete. A multidimensional brush for scatterplot data analytics. In M. Chen, D. S. Ebert, and C. North, eds., *IEEE VAST*, pp. 221–222. IEEE Computer Society, 2014.

[7] M. Aupetit and M. Sedlmair. Sepme: 2002 new visual separation measures. In *2016 IEEE Pacific Visualization Symposium, PacificVis 2016, Taipei, Taiwan, April 19-22, 2016*, pp. 1–8, 2016. doi: 10.1109/PACIFICVIS.2016.7465244

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

[9] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete. A principled way of assessing visualization literacy. *IEEE Trans. Vis. Comput. Graph.*, 20(12):1963–1972, 2014.

[10] M. Brehmer, M. Sedlmair, S. Ingram, and T. Munzner. Visualizing dimensionally-reduced data: interviews with analysts and a characterization of task sequences. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization, BELIV 2014, Paris, France, November 10, 2014*, pp. 1–8, 2014. doi: 10.1145/2669557.2669559

[11] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(5):603–619, 2002.

[12] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.

[13] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. pp. 226–231. AAAI Press, 1996.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and*, pp. 226–231, 1996.

[15] P. F. et al. Clustering basic benchmark, 2015.

[16] B. Fischer, V. Roth, and J. M. Buhmann. Clustering with the connectivity kernel. In S. Thrun, L. K. Saul, and B. Schlkopf, eds., *NIPS*, pp. 89–96. MIT Press, 2003.

[17] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631, 2002.

[18] P. Gaillard, M. Aupetit, and G. Govaert. Learning topology of a labeled data set with the supervised generative gaussian graph. *Neurocomputing*, 71(7-9):1283–1299, 2008. doi: 10.1016/j.neucom.2007.12.028

[19] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *Journal of Intelligent Information Systems*, 17:107–145, 2001.

[20] G. Hamerly and C. Elkan. Learning the k in k-means. In *Advances in Neural Information Processing Systems*, vol. 17, 2003.

[21] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer Verlag, 2009.

[22] C. Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53 – 62, 2015. Philosophical Aspects of Pattern Recognition. doi: 10.1016/j.patrec.2015.04.009

[23] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):881–892, 2002.

[24] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, pp. 159–174, 1977.

[25] Q. Liu, J. Peng, and A. T. Ihler. Variational inference for crowdsourcing. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds., *Advances in Neural Information Processing Systems 25*, pp. 692–700. Curran Associates, Inc., 2012.

[26] Y. Ma, A. K. H. 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. 1–1, 2018. doi: 10.1109/TVCG.2018.2875702

[27] E. Müller, I. Assent, R. Krieger, T. Jansen, and T. Seidl. Morpheus: interactive exploration of subspace clustering. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pp. 1089–1092. ACM, New York, NY, USA, 2008. doi: 10.1145/1401890.1402026

[28] 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*, pp. 1–1, 2018. doi: 10.1109/TVCG.2018.2846735

[29] D. Pelleg and A. Moore. X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In *Proc. 17th Int. Conf. Mach. Learn.*, pp. 727–734. Citeseer, 2000.

[30] R. Rabbany and O. R. Zaane. A general clustering agreement index: For comparing disjoint and overlapping clusters. In S. P. Singh and S. Markovitch, eds., *AAAI*, pp. 2492–2498. AAAI Press, 2017.

[31] V. S. and A. A. Agreement between an isolated rater and a group of raters. *Statistica Neerlandica*, 63(1):82–100. doi: 10.1111/j.1467-9574.2008.00412.x

[32] D. Sacha, H. Senaratne, B. C. Kwon, G. P. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Trans. Vis. Comput. Graph.*, 22(1):240–249, 2016. doi: 10.1109/TVCG.2015.2467591

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

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

[35] U. von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2(3):235–274, 2009. doi: 10.1561/2200000008

[36] F. Wang, B. Zhao, and C. Zhang. Linear time maximum margin clustering. *IEEE Trans. Neural Networks*, 21(2):319–332, 2010.

[37] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, San Francisco, 2004.

[38] M. Wu and B. Schölkopf. A local learning approach for clustering. In B. Schölkopf, J. C. Platt, and T. Hoffman, eds., *Advances in Neural Information Processing Systems 19*, pp. 1529–1536. MIT Press, 2007.