Metric Design != Metric Behavior: Improving Metric Selection for the Unbiased Evaluation of Dimensionality Reduction

Jiyeon Bae[*]
Seoul National University

Hyeon Jeon[†]
Seoul National University

Jinwook Seo[‡]
Seoul National University

Abstract

Evaluating the accuracy of dimensionality reduction (DR) projections in preserving the structure of high-dimensional data is crucial for reliable visual analytics. Diverse evaluation metrics targeting different structural characteristics have thus been developed. However, evaluations of DR projections can become biased if highly correlated metrics—those measuring similar structural characteristics—are inadvertently selected, favoring DR techniques that emphasize those characteristics. To address this issue, we propose a novel workflow that reduces bias in the selection of evaluation metrics by clustering metrics based on their empirical correlations rather than on their intended design characteristics alone. Our workflow works by computing metric similarity using pairwise correlations, clustering metrics to minimize overlap, and selecting a representative metric from each cluster. Quantitative experiments demonstrate that our approach improves the stability of DR evaluation, which indicates that our workflow contributes to mitigating evaluation bias.

Index Terms: Dimensionality reduction, Evaluation metrics, Correlation analysis, Benchmarking, Visual analytics

1 Introduction

Dimensionality Reduction (DR) techniques play a central role in the visual analytics of high-dimensional data across diverse domains, including bioinformatics [10], HCI [5], and signal processing [32]. These techniques aim to project high-dimensional data into lower dimensions while preserving key structural characteristics, such as cluster formations or neighborhood relationships. However, inherent limitations in projecting complex high-dimensional structures to lower dimensions cause DR projections to inevitably emphasize certain structural properties at the expense of others—each having its own focused characteristics. Therefore, quantitatively evaluating DR projections to understand those focused characteristics is essential for interpreting their reliability and limitations before applying them in visual analytics tasks [6, 16, 15].

To quantify the accuracy of DR projections comprehensively, researchers have developed various evaluation metrics, each designed to assess distinct structural characteristics of data projections, e.g., local neighborhood preservation [38] or global distances between points [18, 19]. Although diverse metrics help capture different structural characteristics, the selection of evaluation metrics itself significantly influences evaluation outcomes. In particular, choosing highly correlated metrics—metrics assessing very similar structural properties—can bias evaluations toward DR techniques that specifically optimize for these characteristics, i.e., can erroneously favor certain DR techniques as consistently superior to others. For instance, predominantly using metrics sensitive to local neighborhood structures may disproportionately favor methods like t-SNE or UMAP, which optimize projections focusing on local relationships (see Appendix A). Currently, a common practice to avoid evaluation bias is selecting metrics based on their stated design goals (e.g., optimizing local, cluster-level, or global). However, such an approach can be biased because metrics designed with different intentions may still behave similarly in practice and vice versa. Thus, there still remains a need for an empirically driven approach that reduces bias in the selection of DR evaluation metrics.

To address this problem, we introduce a workflow that selects a subset of evaluation metrics according to their empirical behavior. Our workflow consists of three main steps: (1) computing empirical correlations among metrics across diverse DR projections; (2) clustering these metrics using their correlation-based similarity; and (3) selecting representative metrics from each cluster to minimize redundancy and bias. Quantitative experiments show that our approach improves stability of DR evaluation, i.e., the consistency of the rankings of DR techniques across different metric sets, outperforming baseline methods. The results thus verify that our workflow contributes in reducing evaluation bias.

2 Background and Related Work

We review prior studies on the evaluation of DR techniques and discuss existing approaches to evaluation metric selection.

2.1 Evaluating Dimensionality Reduction Techniques

Researchers have developed diverse evaluation metrics to measure the accuracy of DR projections. These metrics are broadly categorized into three classes: *local*, *cluster-level*, and *global* metrics [13].

Local metrics. These metrics focus on evaluating how well DR projections preserve local neighborhood structure. For example, *Trustworthiness & Continuity (T&C)* [38] and *Mean Relative Rank Error (MRRE)* [22] penalize a projection in which the nearest neighbors in the original space are no longer neighbors in the projection, or vice versa.

Cluster-level metrics. These metrics evaluate whether projections accurately preserve the cluster structure of the original data. For example, *Distance Consistency* [34] or clustering validation measures like *Silhouette* [17] measure how well the labeled classes stay separated in projections, based on the assumption that these classes are well separated in the high-dimensional space.

Global metrics. Finally, global metrics examine whether global relationships like pairwise distances between data points or clusters of the original data remain consistent in the low-dimensional projection. For instance, *Stress* [18, 19] measures discrepancies between the distance matrices of the original and projected spaces, whereas *Kullback–Leibler (KL) Divergence* [9] evaluates differences in how the probability distributions vary across these two spaces.

Our contribution. Although various DR evaluation metrics have been developed, the community lacks systematic approaches for selecting the optimal set of metrics. Evaluations of DR projections might thus inadvertently emphasize certain structural characteristics disproportionately, potentially misguiding users with biased projections. We address this problem by introducing a workflow that selects evaluation metrics that have dissimilar behavior.

[*] e-mail: bjy7266@gmail.com
[†] e-mail: hj@hcil.snu.ac.kr
[‡] e-mail: jseo@snu.ac.kr, corresponding author

2.2 Existing Approaches for Metric Selection

Although the community lacks a standardized way to select DR evaluation metrics, researchers commonly select metrics considering their intended target characteristics (subsection 2.1). For example, Espadoto et al. [6] leverage both local and global metrics to evaluate the accuracy of DR techniques comprehensively. Similarly, prior research proposing new DR techniques [12, 26, 1, 39] employ both local and global metrics, while benchmark studies compare DR techniques across diverse structural levels [4, 2, 3].

However, several studies show that such selection approaches cannot ensure a fair assessment of projection quality. Thrun et al. [37] identify bias in unsupervised DR evaluation metrics by using graph theory, demonstrating that each metric fails to evaluate structure preservation correctly when the input data violates its own structural assumptions. Machado et al. [24] find that adversarially optimized projections can inflate correlated quality scores and construct a guardrail set of DR metrics through correlation analysis and clustering to address this issue.

Our contribution. Our proposed workflow empirically clusters metrics according to their actual observed behaviors, ensuring the selected metrics offer complementary rather than redundant evaluations. This is done by evaluating the behavior of DR metrics across the projections generated by 96 datasets and 40 DR techniques. The recommended compact yet diverse set of metrics for DR evaluation also reduces unnecessary computational costs. Our approach systematically addresses the critical shortcomings of existing DR evaluation practices while complementing them.

3 The Workflow

We propose a workflow for selecting DR evaluation metrics that minimizes redundancy and bias by focusing on their empirical behavior. Our workflow begins by computing pairwise correlations between metrics as a proxy for similarity. We then cluster the metrics according to this similarity to identify optimal groups. Finally, we select representative metrics with minimal pairwise similarity. Please refer to Appendix B for the evaluation metrics and parameters we use.

(Step 1) Computing pairwise correlations. We construct a similarity matrix in which each row and column corresponds to an evaluation metric, and each cell stores the similarity between each pair of metrics. We aim to reflect how closely two metrics behave through their similarity. To obtain it, we first prepare 96 high-dimensional datasets that vary in size, dimensionality, and distribution [14]. For each dataset, we then produce 300 diverse projections exhibiting different visual patterns. We achieve this by repeatedly selecting one of the 40 DR techniques at random and sampling its hyperparameters from predefined ranges (see Appendix C). For every dataset, we quantify the quality of each projection using the chosen evaluation metrics. We subsequently rank the 300 projections for each metric, compute Spearman's rank correlation coefficient (ρ) between all pairs of rankings within each dataset, and define the similarity between two metrics as the average of these correlations across all datasets. We employ Spearman's ρ because it compares the ranks of metric values, enabling it to capture monotonic relationships even under nonlinear distribution of metric scores [35].

(Step 2) Clustering metrics. We cluster the evaluation metrics to avoid selecting measures that emphasize identical structural characteristics. We convert the similarity matrix to a distance matrix by subtracting each entry from 1. We then apply hierarchical clustering [28, 30] to the metrics, using this distance matrix as input and employing average linkage [27, 20] for merging clusters. Hierarchical clustering is chosen for its robustness to noise and stability across multiple runs [11].

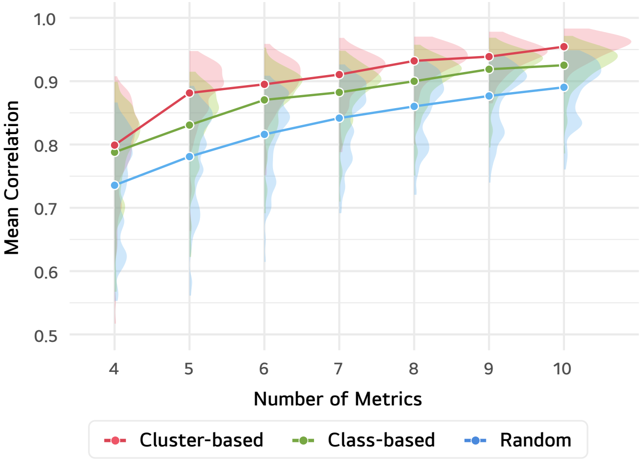

Figure 1: Effect of metric selection strategy on rank stability of DR techniques. The cluster-based approach yields consistently higher stability for $k \geq 5$, indicating that this strategy mitigates evaluation bias more effectively than random or class-balanced sampling.

(Step 3) Selecting representative metrics. For each cluster, we compute the average similarity of every metric to all other metrics in the cluster and select the metric with the highest average similarity as the representative. This strategy ensures that the selected metrics best represent the characteristics of each cluster while minimizing overlap across the full set.

In summary, our workflow aims to alleviate bias in selecting a diverse set of evaluation metrics by analyzing their behavior and clustering them based on similarity to minimize redundancy.

4 Evaluation

We evaluate the effectiveness of our workflow in selecting a set of evaluation metrics that minimizes bias in DR evaluation.

4.1 Objectives and Design

Our goal is to verify whether our workflow mitigates bias compared to baseline strategies for metric selection. We compare three strategies for selecting DR evaluation metrics: **Random** selection, **Class-based** selection, and **Cluster-based** selection. Random selection refers to a strategy in which evaluation metrics are randomly drawn from the available metrics. Class-based selection distributes an equal share of metrics across each category (local, cluster-level, and global) on average. Cluster-based selection randomly picks a single metric from each cluster produced by our workflow.

Procedure. We assess each selection strategy's bias by measuring the variability of evaluation outcomes across multiple executions. The rationale is that if a strategy is biased, the structural characteristics emphasized by the chosen metrics vary widely across runs, causing the overall ranking of projections to fluctuate.

We first prepare 96 datasets and generate 300 projections by randomly sampling DR techniques and hyperparameter settings. This process is identical to the procedure for generating projections in Step 1 of our workflow (section 3). Second, for each selection strategy, we execute it 200 times, yielding 200 distinct sets of evaluation metrics. Finally, for each dataset, we quantify how much the rankings of the 300 projections vary across metric sets by computing the pairwise Spearman correlation ρ between the rankings produced by each metric set. The final correlation for each dataset is obtained by averaging all pairwise correlations. We then average these dataset-level correlations across all datasets to obtain the rank stability for each selection strategy. We repeat this process while increasing the number of clusters from 4 to 10.

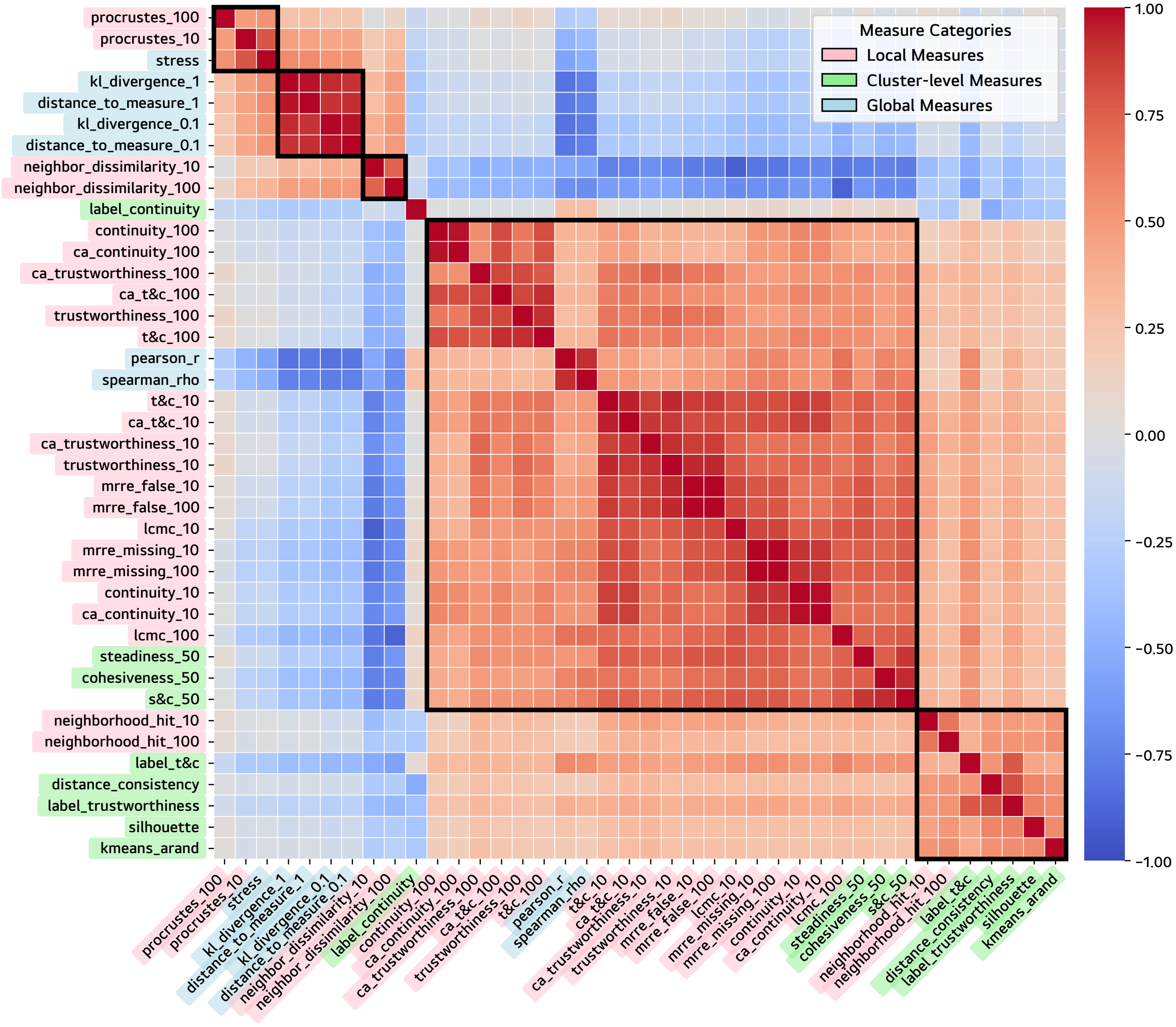

Figure 2: Heatmap of the Spearman correlation matrix among DR evaluation metrics. The order of metrics is determined by hierarchical clustering. Black borders denote the five optimal clusters obtained by our workflow. Singleton clusters are removed as outliers that could add noise to DR evaluation (see Appendix D). Row and column labels are color-coded according to their original design categories—local (pink), cluster-level (light green), and global (light blue)—highlighting that metrics with disparate design intentions can exhibit highly similar empirical behavior and thus group together. The numeric suffixes on local metrics indicate the number of k-nearest neighbors considered, whereas those on global metrics denote the bandwidth used to compute similarity between data points, commonly written as σ [13].

4.2 Results and Discussions

Our results (Figure 1) show that the cluster-based selection strategy achieves the highest rank stability across trials, outperforming the class-based and random baselines. We observe that this advantage persists across all examined cluster counts k. To test this observation, we conduct a one-way ANCOVA, treating k as a covariate and *selection strategy* as a fixed factor. The analysis reveals significant differences in rank stability among the three strategies ($F_{2,2012} = 315.84,\ p < 0.001$).

To further investigate this finding, we conduct a one-way ANOVA to compare rank stability among the three selection strategies for each k. The analysis reveals significant effects of selection strategy for $5 \leq k \leq 10$ ($p < 0.001$ for all). Bonferroni-corrected post-hoc comparisons indicate that the cluster-based selection strategy achieves higher rank stability than the other two strategies in these cases ($p < 0.001$ for all). However, no significant differences are observed at $k = 4$ ($F_{2,285} = 1.26,\ p = 0.285$).

These results show that our workflow mitigates bias in DR evaluation more effectively than the baselines, given the same number of metrics. This enables practitioners to obtain reliable assessments, reducing computational overhead.

5 Recommendations

In previous sections, we introduce a workflow that mitigates bias in the selection of DR evaluation metrics (section 3) and validate it empirically (section 4). Here, we apply this workflow to recommend a compact set of evaluation metrics. We first describe the selection procedure and then discuss the resulting recommendations.

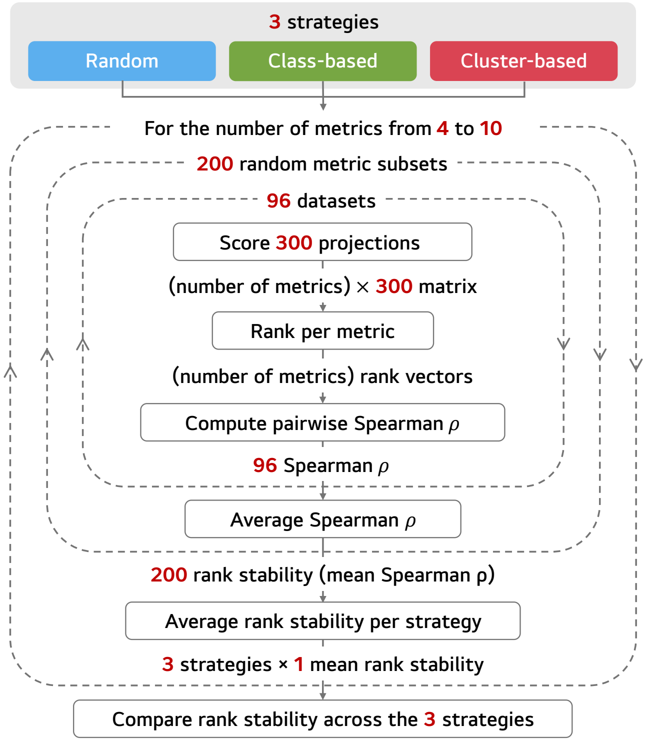
3 strategies
Random
Class-based
Cluster-based
For the number of metrics from 4 to 10
200 random metric subsets
96 datasets
Score 300 projections
(number of metrics) × 300 matrix
Rank per metric
(number of metrics) rank vectors
Compute pairwise Spearman ρ
96 Spearman ρ
Average Spearman ρ
200 rank stability (mean Spearman ρ)
Average rank stability per strategy
3 strategies × 1 mean rank stability
Compare rank stability across the 3 strategies

Figure 3: Procedure for evaluating rank stability across metric selection strategies. The results are depicted in Figure 1.

5.1 Procedure

We execute our workflow using the evaluation metrics employed in our evaluation (section 4). To determine the optimal metric count, we measure the diversity of the selected metrics as the cluster count increases. We then identify the optimal cluster count using the elbow method. Details of both procedures appear below.

Computing diversity. We define diversity as the minimum dissimilarity between each metric and its nearest neighbor in the selected set. First, for a set of k metrics $\{M_1, M_2, \ldots, M_k\}$, we quantify the degree to which each metric is distinct from the others:

$$\mathrm{Ind}(M_i) = \min_{j \neq i}\left(1 - R_{i,j}\right),$$

where $R_{i,j}$ denotes the Spearman correlation between M_i and M_j, so $\mathrm{Ind}(M_i)$ measures how far M_i lies from its nearest neighbor among the other representative metrics.

We then compute the diversity of the set as:

$$D = \sum_{i=1}^{k} \mathrm{Ind}(M_i).$$

As D typically grows with larger k, we use the normalized measure: $D_{\text{norm}} = D/k$.

Determining the optimal number of clusters. We observe that diversity tends to rise with the cluster count k but eventually levels off. We therefore apply the *elbow method* to identify the cluster count beyond which additional clusters yield negligible diversity gains. We employ Kneedle [33], which automatically detects the elbow. We choose the k just beyond the elbow that maximizes diversity among the representative metrics. This choice balances diversity against fragmentation, ensuring each representative metric captures a distinct structural perspective.

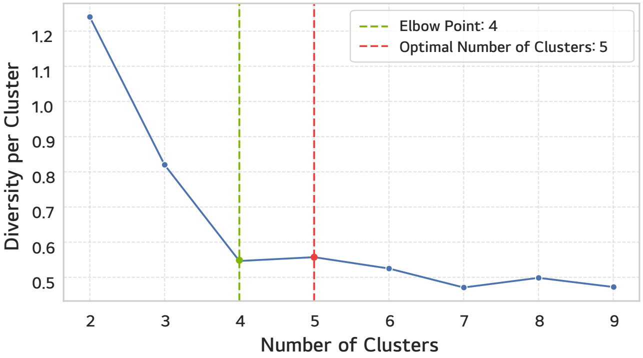

Figure 4: Elbow analysis for selecting the number k of metric clusters. We adopt $k = 5$, the largest cluster count beyond the elbow (red dashed line), as the optimal number of clusters (see Appendix E).

5.2 Results and Recommendations

We find the optimal number of clusters is 5 (Figure 4). As shown in Figure 2, the clusters do not strictly align with the original, design-based classes of DR evaluation metrics: local, cluster-level, and global. For example, the largest cluster—although dominated by local metrics—also includes several cluster-level and global metrics (Figure 2). This result indicates that original classes of metrics do not accurately reflect their intended structural characteristics, reaffirming the value of our workflow for reducing bias in metric selection. The representative set comprises two local metrics (`neighbor_dissimilarity` and `t&c`), one cluster-level metric (`label_trustworthiness`), and two global metrics (`stress` and `kl_divergence`). This set spans a broad range of structural characteristics with minimal redundancy, enabling practitioners to alleviate bias in DR evaluations.

6 Conclusion and Future Work

In this work, we propose a workflow that mitigates bias in DR evaluations caused by highly correlated metrics through the selection of metrics that are maximally dissimilar. Quantitative experiments demonstrate that our workflow substantially reduces evaluation bias relative to baseline methods. Applying the workflow, we recommend a compact and diverse set of metrics to support more reliable and efficient DR evaluation in practice.

However, our workflow is limited in that it cannot capture the full range of real-world data distributions or explore every possible hyperparameter configuration. In future work, we will examine whether our findings can be generalized by utilizing a larger set of datasets and DR techniques.

We also aim to generalize our findings in other domains. Although we focus on DR evaluation, metric redundancy and bias also occur in other machine learning domains. For instance, in NLP domain, widely used machine translation metrics—such as BLEU [31, 25], ROUGE-L [23, 8], and METEOR [21]—are highly correlated and can bias system comparisons [7]. Similar issues arise in medical image segmentation evaluation [29, 36]. The Dice similarity coefficient and Jaccard index are widely used but are mathematically related and were shown to produce identical rankings of segmentation methods [36]. We plan to generalize our workflow to a broader range of machine learning and visualization evaluation tasks. Extending the approach to heterogeneous, task-specific domains poses challenges—e.g., balancing semantic preservation with fluency in text generation, or reconciling segmentation accuracy with boundary precision in medical imaging. Ultimately, we envision empirically driven metric selection playing a critical role in ensuring fair evaluation across diverse fields.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT) under Grant 2023R1A2C200520911, and in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP), funded by the Korean government (MSIT) under Grant RS-2021-II211343 [Artificial Intelligence Graduate School Program (Seoul National University)]. The ICT at Seoul National University provided research facilities for this study.

References

[1] E. Amid and M. K. Warmuth. Trimap: Large-scale dimensionality reduction using triplets, 2019. doi: 10.48550/ARXIV.1910.00204 2

[2] Atzberger et al. Large-scale evaluation of topic models and dimensionality reduction methods for 2d text spatialization. *IEEE Transactions on Visualization and Computer Graphics*, p. 1–11, 2023. doi: 10.1109/tvcg.2023.3326569 2

[3] D. Atzberger et al. Quantifying topic model influence on text layouts based on dimensionality reductions. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 1, GRAPP, HUCAPP and IVAPP, pp. 593–602, 2024. doi: 10.5220/0012391100003660 2

[4] D. Atzberger et al. A large-scale sensitivity analysis on latent embeddings and dimensionality reductions for text spatializations, Jan. 2025. doi: 10.1109/TVCG.2024.3456308 2

[5] M. Cavallo and c. Demiralp. A visual interaction framework for dimensionality reduction based data exploration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 1–13, 2018. doi: 10.1145/3173574.3174209 1

[6] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2021. doi: 10.1109/TVCG.2019.2944182 1, 2

[7] Fabbri et al. Summeval: Re-evaluating summarization evaluation, 04 2021. doi: 10.1162/tacl_a_00373 4

[8] K. Ganesan. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks, 2018. doi: 10.48550/arXiv.1803.01937 4

[9] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, NIPS'02, p. 857–864, 2002. 1

[10] G. Iván and V. Grolmusz. On dimension reduction of clustering results in structural bioinformatics. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1844(12):2277–2283, 2014. doi: 10.1016/j.bbapap.2014.08.015 1

[11] A. K. o. Jain. Data clustering: a review. *ACM Comput. Surv.*, 31(3):264–323, Sept. 1999. doi: 10.1145/331499.331504 2

[12] Jeon et al. Uniform manifold approximation with two-phase optimization. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 80–84, 2022. doi: 10.1109/VIS54862.2022.00025 2

[13] H. Jeon et al. Zadu: A python library for evaluating the reliability of dimensionality reduction embeddings. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 196–200, 2023. doi: 10.1109/VIS54172.2023.00048 1, 3

[14] H. Jeon et al. Measuring the validity of clustering validation datasets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2025. doi: 10.1109/TPAMI.2025.3548011 2

[15] H. Jeon, H. Lee, Y.-H. Kuo, T. Yang, D. Archambault, S. Ko, T. Fujiwara, K.-L. Ma, and J. Seo. Unveiling high-dimensional backstage: A survey for reliable visual analytics with dimensionality reduction. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. Association for Computing Machinery, New York, NY, USA, 2025. doi: 10.1145/3706598.3713551 1

[16] H. Jeon, J. Park, S. Shin, and J. Seo. Stop misusing t-sne and umap for visual analytics, 2025. doi: 10.48550/arXiv.2506.08725 1

[17] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2563–2571, 2011. doi: 10.1109/TVCG.2011.220 1

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

[19] J. B. Kruskal. Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29(2):115–129, 1964. doi: 10.1007/BF02289694 1

[20] E. S. Laber and M. Bastista. On the cohesion and separability of average-link for hierarchical agglomerative clustering, 2024. doi: 10.48550/arXiv.2411.05097 2

[21] A. Lavie and A. Agarwal. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, p. 228–231, 2007. 4

[22] J. A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7):1431–1443, 2009. Advances in Machine Learning and Computational Intelligence. doi: 10.1016/j.neucom.2008.12.017 1

[23] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain, July 2004. 4

[24] A. Machado, M. Behrisch, and A. Telea. Necessary but not sufficient: Limitations of projection quality metrics. *Computer Graphics Forum*, 2025. doi: 10.1111/cgf.70101 2

[25] N. Mathur, T. Baldwin, and T. Cohn. Tangled up in bleu: Reevaluating the evaluation of automatic machine translation evaluation metrics, 01 2020. doi: 10.18653/v1/2020.acl-main.448 4

[26] Moor et al. Topological autoencoders. In *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, pp. 7045–7054, 2020. 2

[27] B. Moseley and J. R. Wang. Approximation bounds for hierarchical clustering: average linkage, bisecting k-means, and local search. *J. Mach. Learn. Res.*, 24(1), Jan. 2023. 2

[28] F. Murtagh and P. Contreras. Algorithms for hierarchical clustering: an overview. *WIREs Data Mining and Knowledge Discovery*, 2:86–97, 2011. doi: 10.1002/widm.53 2

[29] D. Müller et al. Towards a guideline for evaluation metrics in medical image segmentation, 2022. doi: 10.48550/arXiv.2202.05273 4

[30] F. Nielsen. *Hierarchical Clustering*, pp. 195–211. Springer International Publishing, Cham, 2016. doi: 10.1007/978-3-319-21903-5_8 2

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, p. 311–318, 2002. doi: 10.3115/1073083.1073135 4

[32] L. Rui, H. Nejati, and N.-M. Cheung. Dimensionality reduction of brain imaging data using graph signal processing. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 1329–1333, 2016. doi: 10.1109/ICIP.2016.7532574 1

[33] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pp. 166–171, 2011. doi: 10.1109/ICDCSW.2011.20 4

[34] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting Good Views of High-dimensional Data using Class Consistency. *Computer Graphics Forum*, 2009. doi: 10.1111/j.1467-8659.2009.01467.x 1

[35] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 100(3/4):441–471, 1987. 2

[36] A. A. Taha and A. Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15, 2015. doi: 10.1186/s12880-015-0068-x 4

[37] M. C. Thrun, J. Märte, and Q. Stier. Analyzing quality measurements for dimensionality reduction. *Machine Learning and Knowledge Extraction*, 5(3):1076–1118, 2023. doi: 10.3390/make5030056 2

[38] 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 1

[39] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization, 12 2020. doi: 10.48550/arXiv.2012.04456 2